Proceedings of Machine Learning Research vol XX:1–57, 2023

Provably Efficient Generalized Lagrangian Policy Optimization for Safe Multi-Agent Reinforcement Learning

Dongsheng Ding DONGSHED@SEAS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104, USA

Xiaohan Wei UBIMETEOR@FB.COM
Meta, Menlo Park, CA 94025 USA

Zhuoran Yang ZHUORAN.YANG@YALE.EDU
Yale University, New Haven, CT 06511, USA

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Northwestern University, Evanston, IL 60208, USA

Mihailo R. Jovanović MIHAILO@USC.EDU
University of Southern California, Los Angeles, CA 90089, USA

Abstract

We examine online safe multi-agent reinforcement learning using constrained Markov games in which agents compete by maximizing their expected total rewards under a constraint on expected total utilities. Our focus is confined to an episodic two-player zero-sum constrained Markov game with independent transition functions that are unknown to agents, adversarial reward functions, and stochastic utility functions. For such a Markov game, we employ an approach based on the occupancy measure to formulate it as an online constrained saddle-point problem with an explicit constraint. We extend the Lagrange multiplier method in constrained optimization to handle the constraint by creating a generalized Lagrangian with minimax decision primal variables and a dual variable. Next, we develop an upper confidence reinforcement learning algorithm to solve this Lagrangian problem while balancing exploration and exploitation. Our algorithm updates the minimax decision primal variables via online mirror descent and the dual variable via projected gradient step and we prove that it enjoys sublinear rate $O((|X|+|Y|)L\sqrt{T(|A|+|B|)})$ for both regret and constraint violation after playing T episodes of the game. Here, L is the horizon of each episode, $(|X|,|A|)$ and $(|Y|,|B|)$ are the state/action space sizes of the min-player and the max-player, respectively. To the best of our knowledge, we provide the first provably efficient online safe reinforcement learning algorithm in constrained Markov games.

Keywords: safe multi-agent reinforcement learning, constrained Markov game, upper confidence reinforcement learning, generalized Lagrange multiplier method, online mirror descent

1. Introduction

Safe Reinforcement Learning (RL) studies how a single agent learns to maximize its expected total reward subject to safety-concerned constraints by interacting with an unknown environment over time (Garcıa and Fernández, 2015; Thomas, 2015; Amodei et al., 2016). The constrained Markov decision processes (MDPs) provide a standard class of constraint critical environment models (Altman, 1999) that are utilized in autonomous robots (Feyzabadi, 2017; Fisac et al., 2018), personalized medicine (Girard, 2018), online advertising (Boutilier and Lu, 2016), and financial management (Abe et al., 2010). General constrained MDPs for two or more agents are often formulated as constrained Markov games (MGs) in which agents compete under constraints (Altman and Shwartz, 2000; Altman et al., 2005, 2008), providing an effective model for safe multi-agent RL (Nguyen et al., 2014; Shalev-Shwartz et al., 2016; Zhang et al., 2021).

Considerable recent progress has been made in single-agent safe RL, especially for solving constrained MDP problems with constraint satisfaction guarantees (Efroni et al., 2020; Brantley et al., 2020; Bai et al., 2020a; Ding et al., 2021; Chen et al., 2021; Singh et al., 2022; Ding et al., 2022). In these references, Lagrangian-based methods have been combined with the optimistic exploration to address exploration-exploitation trade-off under constraints. These constrained MDP learning algorithms are sample-efficient (in achieving both low regret and low constraint violation) and they effectively enhance classical RL methods to attain safety requirements. However, most of these algorithms are limited to the single-agent setting and it is an open question how to balance the exploration-exploitation trade-off under constraints for multiple agents. Another motivation for our work comes from recent advances on the efficient competitive RL algorithms in MGs (Wei et al., 2017; Bai and Jin, 2020; Bai et al., 2020b; Xie et al., 2020).

In this work, we take initial steps towards developing provably efficient safe multi-agent RL algorithms. We examine the most basic safe multi-agent RL setup that involves a two-player zero-sum constrained MG with independent state transitions (Altman and Shwartz, 2000; Altman et al., 2005, 2008; Singh and Hemachandra, 2014). This problem represents a generalization of constrained MDPs to the two-player case with coupled constraints. In such a constrained MG, the two players follow two independent state transitions, take actions simultaneously, and observe the reward and utility functions while competing against each other by maximizing/minimizing the reward while both are restrained by the constraint regarding some utility for safety reasons. The decision-coupling that arises from the constraint is often encountered in multi-agent systems (Rosen, 1965; Li and Marden, 2014; Kulkarni, 2011, 2017; De Nijs, 2019). More specifically, we aim to design an online RL algorithm for solving episodic two-player zero-sum constrained MGs. Here, the two players do not know the transition models and have no access to a generative model, but can play the game for multiple episodes using arbitrary policies. The goal is to find an approximate constrained Nash equilibrium of the game in hindsight, a generalization of Nash equilibrium to characterize violating constraints if any unilateral deviations occur. We utilize a notion of regret to quantify the approximation error of the constrained Nash equilibrium and employ a constraint dissatisfaction (which results from violation of any utility constraints) to evaluate the constraint violation.

Contribution. We develop the first provably efficient algorithm for a constrained Markov game (MG) with $O(\sqrt{T})$ regret and $O(\sqrt{T})$ constraint violation. Specifically, we introduce an episodic constrained MG with unknown independent transition functions and decision-couplings that come from both adversarial reward functions and coupled stochastic constraints on utility functions. We use the occupancy measure approach to formulate such a MG as a constrained saddle-point problem with an explicit constraint. We extend the Lagrange method in constrained optimization to deal with the constraint by creating a generalized Lagrangian with minimax decision primal variables and a dual variable. We develop an upper confidence reinforcement learning algorithm – an Upper Confidence Bound Constrained SAddle-Point Optimization (UCB-CSAPO) algorithm – to solve this Lagrangian problem while balancing exploration and exploitation. Our algorithm updates the minimax decision primal variables via optimistic mirror descent and the dual variable via projected gradient step and we prove that it enjoys sublinear rate $O((|X|+|Y|)L\sqrt{T(|A|+|B|)}))$ for both regret and constraint violation after playing T episodes of the game. Here, L is the horizon of each episode, $(|X|, |A|)$ and $(|Y|, |B|)$ are the state/action space sizes of the min-player and max-player, respectively.

Related Work. We briefly review the most-related work; see Appendix 6 for details. Our work is closely related to safe multi-agent RL in constrained MGs. The Nash equilibrium for constrained MGs have been studied in Altman and Shwartz (2000); Gómez-Ramırez et al. (2003); Altman et al. (2005); Alvarez-Mena and Hernández-Lerma (2006); Altman et al. (2007, 2008); Altman and Solan (2009); Singh and Hemachandra (2014) using the notion of *constrained Nash equilibrium* (which generalizes the concept of *generalized Nash equilibrium* in static games (Arrow and Debreu, 1954) to MGs); see more studies in Yaji and Bhatnagar (2015); Zhang (2019); Wei (2020, 2021); Zhang and Zou (2021). These results are not applicable to the RL setting that assumes unknown models. Recently, asymptotic convergence in learning constrained MGs was examined in Hakami and Dehghan (2015); Jiang et al. (2020) but sample efficiency and exploration were not addressed, except for learning general equilibria (Chen et al., 2022b). Our work fills this gap by adding built-in exploration mechanisms under constraints and proving the first non-asymptotic convergence.

Our work is also pertinent to a rich RL literature on learning constrained MDPs (Zheng and Ratliff, 2020; Qiu et al., 2020; Kalagarla et al., 2020; Bai et al., 2020a; Chow et al., 2017; Tessler et al., 2019; Ding et al., 2020, 2021, 2022; Wachi and Sui, 2020; Efroni et al., 2020; Brantley et al., 2020; Chen et al., 2021; Liu et al., 2021a; Ying et al., 2022; Liu et al., 2021b; Bai et al., 2022; Zhao and You, 2021; Li et al., 2021; Chen et al., 2022a). While these results provide provably efficient algorithms regarding regret and constraint satisfaction in the single-agent setting, they are not applicable to our multi-agent game being played under constraints, because of the *non-convexity* nauture of constrained multi-agent policy optimization and the *non-stationary* environment each agent is facing. An extended line of work on constrained MDPs focuses on cooperative multi-agent learning under constraints and most efforts study the case where multiple agents have independent MDPs with a coupled budget/resource constraint (Meuleau et al., 1998; Boutilier and Lu, 2016; Wei et al., 2018; de Nijs and Stuckey, 2020; Gagrani and Nayyar, 2020). All these results assume knowing transition models or system dynamics. Only a few studies considered the shared MDP case (Diddigi et al., 2019; Lu et al., 2020; Parnika et al., 2021; Gu et al., 2021), but they lack theoretical guarantees and do not handle exploration. In contrast, our work focuses on the MG setting with unknown models and attacks the exploration challenge directly.

2. Problem Setup

In this section, we introduce zero-sum Markov games (MGs) with constraints, which are categorized as constrained Markov/stochastic games (Altman and Shwartz, 2000; Altman et al., 2005, 2008).

In an episodic constrained MG there are two players; a *min-player* – (X, A, P_1, r, g, T), which minimizes the reward, and a *max-player* – (Y, B, P_2, r, h, T), which maximizes the reward, while adhering to a coupled utility constraint. Here, T is the number of episodes, X and Y are finite state spaces, A and B are finite action spaces, P_1 and P_2 are transition probability measures where $P_1(\cdot \,|\, x, a)$ is a distribution over X if the min-player takes action a in state x and $P_2(\cdot \,|\, y, b)$ is a distribution over Y if the max-player takes action b in state y, $r := \{r^t\}_{t=1}^T$ is a collection of players' reward functions $r^t\colon X \times Y \times A \times B \to [0, 1]$, whereas $g := \{g^t\}_{t=1}^T$ and $h := \{h^t\}_{t=1}^T$ are collections of players' utility functions $g^t\colon X \times A \to [0, 1]$, $h^t\colon Y \times B \to [0, 1]$. For two independent transitions, players are coupled via the reward function and a constraint on their utility functions.

We utilize layered Markov decision processes to model the environment dynamics. For each player, e.g., the min-player, we assume that the state space X has $L + 1$ layers and that it satisfies the loop-free property: (i) $X := X_0 \cup \cdots \cup X_L$ and $X_{\ell_1} \cap X_{\ell_2} = \emptyset$ for $\ell_1 \neq \ell_2$; (ii) $X_0 = \{x_0\}$ and $X_L = \{x_L\}$; (iii) if $P_1(x' \,|\, x, a) > 0$, then $x' \in X_{\ell+1}$ and $x \in X_\ell$ for some $\ell \in \{0, 1, \cdots, L\}$. This assumption is common in loop-free stochastic shortest path problems (György et al., 2007; Jaksch et al., 2010; Neu et al., 2010; Rosenberg and Mansour, 2019; Jin et al., 2020); it is often used to simplify notation/analysis since any episodic MDPs can be reduced to be loop-free.

The min/max players interact with the environment in episode t as follows. At the beginning, the environment determines the reward function r^t and the utility functions g^t and h^t. Meanwhile, two players decide their policies π^t: $X \times A \to [0,1]$ and μ^t: $Y \times B \to [0,1]$, where $\pi^t(\cdot \,|\, x)$ and $\mu^t(\cdot \,|\, y)$ are probability distributions over their action spaces A and B, respectively. Then, given initial states x_0 and y_0, both players execute their own policies π^t or μ^t for L steps. At step $\ell \in \{0, \ldots, L-1\}$, each player only observes its own state $x_\ell \in X$ or $y_\ell \in Y$, takes action a_ℓ or b_ℓ following its own policy π^t or μ^t, transits to next state $x_{\ell+1}$ or $y_{\ell+1}$ according to its own transition $P_1(\cdot \,|\, x_\ell, a_\ell)$ or $P_2(\cdot \,|\, y_\ell, b_\ell)$, and observes reward r^t and local utility g^t or h^t. Assume there is no dependence between functions r^t, g^t, and h^t and they are independent of the underlying MDPs.

To define the learning objective, for the min-player in episode t we introduce the occupancy measure q_1^t: $X \times A \times X \to [0,1]$ by $q_1^t(x, a, x') := \text{Prob}(x_\ell = x, a_\ell = a, x_{\ell+1} = x')$ for $x \in X_\ell$, describing the marginal probability of visiting (x, a, x') when executing policy π^t under the transition P_1. Similarly, we introduce the occupancy measure q_2^t: $Y \times B \times Y \to [0,1]$ for the max-player. We recall that a function q: $X \times A \times X \to [0,1]$ is an occupancy measure associated with policy π and transition P if and only if it satisfies two conditions (Altman, 1999): (i) $\sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} q(x, a, x') = 1$ for $\ell \in \{0, \ldots, L-1\}$; (ii) $\sum_{x \in X_{\ell-1}} \sum_{a \in A} q(x, a, x') = \sum_{a \in A} \sum_{x'' \in X_{\ell+1}} q(x', a, x'')$ for $x' \in X_\ell$ and $\ell \in \{1, \ldots, L-1\}$. We denote by $\Delta(P)$ a set of valid occupancy measures under P,

$$\Delta(P) \;:=\; \big\{ q \colon X \times A \times X \to [0,1] \;|\; q \text{ satisfies (i) and (ii) as shown above} \big\}.$$

It is worth noting that the occupancy measure set is convex and compact for finite MDPs (Altman, 1999). Using an occupancy measure q, we can express associated transition P and policy π as

$$P(x' \,|\, x, a) \;=\; \frac{q(x, a, x')}{\sum_{x'' \in X_{\ell+1}} q(x, a, x'')} \;\;\text{and}\;\; \pi(a \,|\, x) \;=\; \frac{\sum_{x' \in X_{\ell+1}} q(x, a, x')}{\sum_{a \in A} \sum_{x'' \in X_{\ell+1}} q(x, a, x'')} \tag{1}$$

where $x \in X_\ell$. Slightly extending the notation q, we use it to represent the probability of visiting (x, a), i.e., $q(x, a) = \sum_{x' \in X_{\ell+1}} q(x, a, x')$ for $x \neq x_L$. These properties imply that the problem of learning a policy equals learning the associated occupancy measure (Zimin and Neu, 2013).

In episode t, given a min-policy π^t and a max-policy μ^t, we introduce the expected total reward,

$$\begin{aligned} &\mathbb{E}_{P_1, P_2, \pi^t, \mu^t} \left[\sum_{\ell=0}^{L-1} r^t(x_\ell, y_\ell, a_\ell, b_\ell) \right] \;=\; \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell, y \in Y_\ell} \sum_{a \in A, b \in B} q_1^t(x, a) q_2^t(y, b) r^t(x, y, a, b) \\ &:= \; \left\langle q_1^t \cdot q_2^t, r^t \right\rangle \end{aligned} \tag{2}$$

where the expectation $\mathbb{E}$ is taken over the random state-action sequence $\{(x_\ell, y_\ell, a_\ell, b_\ell)\}_{\ell=0}^{L-1}$; the action a_ℓ follows the policy $\pi^t(\cdot \,|\, x_\ell)$ in the state x_ℓ and the next state $x_{\ell+1}$ follows the transition $P_1(\cdot \,|\, x_\ell, a_\ell)$; the action b_ℓ follows the policy $\mu^t(\cdot \,|\, y_\ell)$ in the state y_ℓ and the next state $y_{\ell+1}$ follows the transition $P_2(\cdot \,|\, y_\ell, b_\ell)$. Similarly, we can define the expected total utilities as

$$\mathbb{E}_{P_1,\pi^t}\left[\sum_{\ell=0}^{L-1} g_x^t(x_\ell,a_\ell)\right] \;=\; \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q_1^t(x,a)g^t(x,a) \;:=\; \left\langle q_1^t, g^t\right\rangle \tag{3a}$$

$$\mathbb{E}_{P_2,\mu^t}\left[\sum_{\ell=0}^{L-1} h^t(y_\ell,a_\ell)\right] \;=\; \sum_{\ell=0}^{L-1}\sum_{y\in Y_\ell}\sum_{b\in B} q_2^t(y,b)h^t(y,b) \;:=\; \left\langle q_2^t, h^t\right\rangle. \tag{3b}$$

In general, reward function r^t and utility functions g^t and h^t all can change arbitrarily, i.e., being adversarial. However, even if we fix the opponent's policy, there is no algorithm for the player to achieve sublinear regret and constraint violation at the same time when the constraints are changing adversarially (Mannor et al., 2009). Hence, we restrict the utility functions to be stochastic: $g^t(x,a) \;:=\; g(x,a;\xi^t)$, $h^t(y,b) \;:=\; h(y,b;\xi^t)$ with $\mathbb{E}\left[g^t(x,a)\right] \;=\; g(x,a)$ and $\mathbb{E}\left[h^t(y,b)\right] \;=\; h(y,b)$, for any $x\in X$, $a\in A$ and $y\in Y$, $b\in B$, where ξ^t is an independent random variable.

Learning Performance. We now define the underlying constrained optimization problem and the solution concept for learning constrained MGs. Using the notion of occupancy measure, we formulate a constrained minimax problem in which the objective function is a sum of the expected total rewards over T episodes and the constraint is on a sum of two agent' expected total utilities,

$$\underset{q_1\in\Delta(P_1)}{\text{minimize}}\;\underset{q_2\in\Delta(P_2)}{\text{maximize}}\;\; \sum_{t=0}^{T-1}\left\langle q_1\cdot q_2, r^t\right\rangle \quad \text{subject to} \quad \langle q_1,g\rangle \,+\, \langle q_2,h\rangle \;\le\; b \tag{4}$$

where we take $b\in(0,2L]$ to avoid trivial cases since we note that $\langle q_1,g\rangle$, $\langle q_2,h\rangle\in[0,L]$. The coupled constraint is used to model the limited use of budget/resource for two players; multi-agent problems with a common constraint are often called *weakly-coupled* or *non-orthogonal* in the literature on CMDPs (Meuleau et al., 1998; Boutilier and Lu, 2016; Wei et al., 2018; Salemi Parizi, 2018; Gagrani and Nayyar, 2020) and constrained MGs (Altman et al., 2008; Altman and Solan, 2009; Kulkarni, 2011; Singh and Hemachandra, 2014; Kulkarni, 2017). We can generalize it to account for multiple constraints or local side constraints, e.g., $\langle q_1,g\rangle\le b_1$ or $\langle q_2,h\rangle\le b_2$. When transitions P_1 and P_2 are known, the occupancy measure sets $\Delta(P_1)$ and $\Delta(P_2)$ define convex polytopes on q_1 and q_2.

Let $(q_1^\star,q_2^\star)$ be a solution to Problem (4) in hindsight. The existence of $(q_1^\star,q_2^\star)$ follows from compactness of the constraint sets (Neumann, 1928; Rosen, 1965). It is standard to define an intuitive solution – constrained Nash equilibrium – via two conditions (Altman and Shwartz, 2000; Daskalakis et al., 2021):

(i) $\displaystyle\sum_{t=0}^{T-1}\langle q_1^\star\cdot q_2^\star, r^t\rangle \;\le\; \sum_{t=0}^{T-1}\langle q_1\cdot q_2^\star, r^t\rangle$ for any $q_1\in\Delta(P_1)$ satisfying $\langle q_1,g\rangle+\langle q_2^\star,h\rangle\le b$;

(ii) $\displaystyle\sum_{t=0}^{T-1}\langle q_1^\star\cdot q_2, r^t\rangle \;\le\; \sum_{t=0}^{T-1}\langle q_1^\star\cdot q_2^\star, r^t\rangle$ for any $q_2\in\Delta(P_2)$ satisfying $\langle q_1^\star,g\rangle+\langle q_2,h\rangle\le b$.

Any unilateral deviation from the constrained Nash equilibrium will either break the constraint or if it is not, then there is no benefit for this player. With this solution concept, we define the regret for any algorithm that plays the game for T episodes by

$$\text{Regret}(T) \;=\; \sum_{t=0}^{T-1} \left(\left\langle q_1^t \cdot q_2^\star, r^t \right\rangle - \left\langle q_1^\star \cdot q_2^t, r^t \right\rangle \right) \tag{5}$$

which adds two side optimality gaps, $\sum_{t=0}^{T-1} \left\langle q_1^t \cdot q_2^\star, r^t \right\rangle - \left\langle q_1^\star \cdot q_2^\star, r^t \right\rangle$ for the min-player and $\sum_{t=0}^{T-1} \langle q_1^\star \cdot q_2^\star, r^t \rangle - \langle q_1^\star \cdot q_2^t, r^t \rangle$ for the max-player, and two players take policies π^t and μ^t in episode t and they define occupancy measures q_1^t and q_2^t under the true transitions P_1 and P_2. This regret works in a notion of weak regret (Brafman and Tennenholtz, 2002; Bai and Jin, 2020; Xie et al., 2020) instead of the single-agent type regret (Tian et al., 2020; Bai et al., 2020b) which is statistically and computationally hard to bound sublinearly.

To measure the constraint satisfaction, we introduce the violation as a non-negative part of accumulated constraint violations $\langle q_1^t, g \rangle + \langle q_2^t, h \rangle - b$ over T episodes,

$$\text{Violation}(T) \;=\; \left[\sum_{t=0}^{T-1} \left(\left\langle q_1^t, g^t \right\rangle + \left\langle q_2^t, h^t \right\rangle - b \right) \right]_+ . \tag{6}$$

We next make an assumption that guarantees the existence of constrained Nash equilibrium (Altman and Shwartz, 2000).

Assumption 1 (Feasibility) *There exists a joint policy $(\bar{\pi}, \bar{\mu})$ associated to the occupancy measure $(\bar{q}_1, \bar{q}_2)$ and a constant $\xi > 0$ such that $\langle \bar{q}_1, g \rangle + \langle \bar{q}_2, h \rangle + \xi \leq b$.*

Having defined the learning performance, we will work with the occupancy measure in the online learning setting where the two players do not know the transition functions, only observe reward/utility functions at the end of each episode, repeatedly play the game for a fixed number of episodes to learn the constrained Nash equilibrium in hindsight.

3. Proposed Algorithm

We present a variant of upper confidence reinforcement learning in Algorithm 1 – an Upper Confidence Bound Constrained SAddle-Point Optimization (UCB-CSAPO) algorithm – for learning constrained MGs. Conceptually, the algorithm works as the primal-dual policy optimization (Efroni et al., 2020; Ding et al., 2021; Chen et al., 2021) in the Lagrangian-based framework, which makes it a simple policy optimization algorithm. However, our primal update exploits the structure of constrained MGs to maintain two players' occupancy measures. The domain set of occupancy measures builds on the upper confidence bound exploration or optimism (Jaksch et al., 2010) regarding the estimated transition models using past trajectories. The dual update determines the penalty weight by collecting the possible constraint violation already acquired. In each episode, our algorithm has two key stages: (i) The generalized Lagrangian mirror descent step for updating the occupancy measures with optimism; (ii) The estimation of confidence sets on the occupancy measures.

The Generalized Lagrangian Mirror Descent Step. The main idea of this step is to apply the online primal-dual mirror descent – an algorithmic generalization of online mirror descent to the constrained problems (Wei et al., 2020) – to the constrained MG setting (Altman and Shwartz, 2000; Altman et al., 2005, 2008; Singh and Hemachandra, 2014). Let us recall that the occupancy measures q_1^t for the min-player and q_2^t for the max-player are defined over the true transitions P_1 and P_2 in

Algorithm 1 Upper Confidence Bound Constrained SAddle-Point Optimization (UCB-CSAPO)

Input: State/action spaces (X, A) and (Y, B), episode T, parameters V, η, θ, and $p \in (0, 1)$.

Initialization: The min-player: $\widehat{q}_1^0(x, a, x') = \frac{1}{|X^\ell||A||X^{\ell+1}|}$, $\forall (x, a, x') \in X^\ell \times A \times X^{\ell+1}, \ell \in [0, L-1]$; $n_1^1(x, a) = N_1^1(x, a) = 0, \forall (x, a)$; $m_1^1(x, a, x') = M_1^1(x, a, x') = \bar{P}_1^1(x' \,|\, x, a) = 0$, $\forall (x, a, x')$.
The max-player: $\widehat{q}_2^0(y, b, y') = \frac{1}{|Y^\ell||B||Y^{\ell+1}|}$, $\forall (y, b, y') \in Y^\ell \times B \times Y^{\ell+1}, \ell \in [0, L-1]$; $n_2^1(y, b) = N_2^1(y, b) = 0, \forall (y, b)$; $m_2^1(y, b, y') = M_2^1(y, b, y') = \bar{P}_2^1(y' \,|\, y, b) = 0, \forall (y, b, y')$.
Let r^0, g^0, h^0 be zero functions, λ^0 be zero, and $k_1^1 = k_2^1 = 1$.

for episode $t = 1, \ldots, T$ **do**

  Update the primal variable $\widehat{q}^{\,t}$ via (8) and the dual variable λ^t via (10).

  Compute the min-policy π^t and the max-policy μ^t via (7). Execute them for L steps and record trajectories $(x^0, a^0, x^1, \cdots, a^{L-1}, x^{L-1})$ and $(y^0, b^0, y^1, \cdots, b^{L-1}, y^{L-1})$, and reward/utility functions r^t, g^t, and h^t.

  Update local visitation counters at visited trajectories,

$$n_1^{k_1^t}(x^\ell, a^\ell) \,\leftarrow\, n_1^{k_1^t}(x^\ell, a^\ell) + 1 \;\text{ and }\; m_1^{k_1^t}(x^\ell, a^\ell, x^{\ell+1}) \,\leftarrow\, m_1^{k_1^t}(x^\ell, a^\ell, x^{\ell+1}) + 1$$

$$n_2^{k_2^t}(y^\ell, b^\ell) \,\leftarrow\, n_2^{k_2^t}(y^\ell, b^\ell) + 1 \;\text{ and }\; m_2^{k_2^t}(y^\ell, b^\ell, y^{\ell+1}) \,\leftarrow\, m_2^{k_2^t}(y^\ell, b^\ell, y^{\ell+1}) + 1.$$

  if $n_1^{k_1^t}(x, a) \geq N_1^{k_1^t}(x, a)$ or $n_2^{k_2^t}(y, b) \geq N_2^{k_2^t}(y, b)$ for some $(x, a) \in X \times A$ or $(y, b) \in Y \times B$ **then**

    Increase epoch counter by one, $k_1^{t+1} \leftarrow k_1^t + 1$ or $k_2^{t+1} \leftarrow k_2^t + 1$, and update global visitation counters,

$$N_1^{k_1^{t+1}}(x, a) \,\leftarrow\, N_1^{k_1^t}(x, a) \,+\, n_1^{k_1^t}(x, a) \;\text{ or }\; N_2^{k_2^{t+1}}(y, b) \,\leftarrow\, N_2^{k_2^t}(y, b) \,+\, n_2^{k_2^t}(y, b)$$

$$M_1^{k_1^{t+1}}(x, a, x') \leftarrow M_1^{k_1^t}(x, a, x') + m_1^{k_1^t}(x, a, x') \;\text{ or }\; M_2^{k_2^{t+1}}(y, b, y') \leftarrow M_2^{k_2^t}(y, b, y') + m_2^{k_2^t}(y, b, y').$$

    Update the confidence bounds for $\Delta(k_1^t)$ or $\Delta(k_2^t)$ in (11), and set $n_1^{k_1^{t+1}}(x, a) = m_1^{k_1^{t+1}}(x, a, x') = 0$ for all (x, a) and (x, a, x') or $n_2^{k_2^{t+1}}(y, b) = m_2^{k_2^{t+1}}(y, b, y') = 0$ for all (y, b) and (y, b, y').

  else

    Set either $k_1^{t+1} = k_1^t$ or $k_2^{t+1} = k_2^t$.

  end if

end for

episode t. The primal update of our algorithm maintains two occupancy measures $\widehat{q}_1^{\,t}$, $\widehat{q}_2^{\,t}$ to estimate q_1^t, q_2^t, separately. Although $\widehat{q}_1^{\,t}$, $\widehat{q}_2^{\,t}$ do not necessarily come from the true transitions P_1, P_2, they propose a min-policy π^t for the min-player and a max-policy μ^t for the max-player according to the occupancy measure's property (1), i.e., for all $(x, a) \in X \times A$ and $(y, b) \in Y \times B$,

$$\pi^t(a \,|\, x) \;=\; \frac{\sum_{x'} \widehat{q}_1^{\,t}(x, a, x')}{\sum_{a, x''} \widehat{q}_1^{\,t}(x, a, x'')} \;\text{ and }\; \mu^t(b \,|\, y) \;=\; \frac{\sum_{y'} \widehat{q}_2^{\,t}(y, b, y')}{\sum_{b, y''} \widehat{q}_2^{\,t}(y, b, y'')}. \tag{7}$$

We describe our Lagrangian-based design to update estimates $\widehat{q}_1^t$ and $\widehat{q}_2^t$ in an online fashion. Assume that the transitions P_1 and P_2 are known. We consider a one-episode constrained minimax problem based on reward/utility functions: r^{t-1}, g^{t-1}, h^{t-1}, revealed at the end of episode $t-1$,

$$\underset{q_1 \in \Delta(P_1)}{\text{minimize}} \ \underset{q_2 \in \Delta(P_2)}{\text{maximize}} \quad \left\langle q_1 \cdot q_2, r^{t-1} \right\rangle \quad \text{subject to} \quad \left\langle q_1, g^{t-1} \right\rangle + \left\langle q_2, h^{t-1} \right\rangle \ \leq \ b$$

where $\Delta(P_1)$ and $\Delta(P_2)$ are sets of valid occupancy measures under P_1 and P_2, respectively.

It is standard to use the method of Lagrange multipliers (Bertsekas, 2014) to handle constraints by adding penalty terms, if any constraint violation appears, into the original objective, and formulate an unconstrained problem. This is found in constrained games with separate side constraints (Pearsall, 1976) and multiple MDPs with coupled constraints (Boutilier and Lu, 2016; Wei et al., 2018). However, for constrained MGs either player can contribute to constraint violation $\langle q_1, g^{t-1}\rangle + \langle q_2, h^{t-1}\rangle - b$. It is important to specify which player should get such penalty terms (Altman and Solan, 2009; Dai and Zhang, 2020). We employ an attitude that the two players are jointly against the constraint while competing for rewards (Altman and Solan, 2009). As a result, both would sacrifice their rewards to satisfy the constraint if any violation occurs. We approximate the violation for each player as: $\langle q_1, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b$ for the min-player, and $\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle q_2, h^{t-1}\rangle - b$ for the max-player. We formulate a generalized Lagrangian-type function,

$$\begin{aligned} L^t(q_1, q_2; \lambda) \quad := \quad & \langle q_1 \cdot q_2, r^{t-1} \rangle \\ & + \lambda\big(\langle q_1, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big) \ - \ \lambda\big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle q_2, h^{t-1}\rangle - b\big) \end{aligned}$$

where q_1 is the first primal variable for the min-player, q_2 is the second primal variable for the max-player, and $\lambda \geq 0$ works as the Lagrange multiplier or the dual variable in penalizing the min-player/max-player via the first/second λ-term. Once we update $\lambda = \lambda^{t-1}$ from the last episode, we reach a constrained saddle-point problem, $\text{minimize}_{q_1 \in \Delta(P_1)} \ \text{maximize}_{q_2 \in \Delta(P_2)} \ L^t(q_1, q_2; \lambda^{t-1})$.

However, it is not feasible to take the domains $\Delta(P_1)$ and $\Delta(P_2)$ since the true transitions P_1 and P_2 are unknown. Instead, by the optimism in the face of uncertainty, we use their optimistic estimates $\Delta(k_1^t)$ and $\Delta(k_2^t)$ in sense that $q_1^t \in \Delta(k_1^t)$ and $q_2^t \in \Delta(k_2^t)$ hold with high probability in Lemma 1, where $\Delta(k_1^t)$ and $\Delta(k_2^t)$ are given by (11). Let $\widehat{q}^t := (\widehat{q}_1^t, \widehat{q}_2^t)$ and $D(p\,|\,q) := \sum_i p_i \ln \frac{p_i}{q_i} - \sum_i (p_i - q_i)$ that is the unnormalized Kullback-Leibler (KL) divergence between two distributions p, q. By a linear approximation of $L^t(q_1, q_2; \lambda^{t-1})$ at the previous iterate (q_1^{t-1}, q_2^{t-1}), we update the primal variable via an online mirror descent step over the domains of q_1 and q_2,

$$\begin{aligned} \widehat{q}^t \ \leftarrow \ \underset{q_1 \in \Delta(k_1^t)}{\text{argmin}} \ \underset{q_2 \in \Delta(k_2^t)}{\text{argmax}} \Big(& V \left\langle q_1 \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot q_2, r^{t-1} \right\rangle \\ & + \lambda^{t-1}\big(\langle q_1, g^{t-1}\rangle - \langle q_2, h^{t-1}\rangle\big) \ + \ \eta^{-1} D\big(q \,|\, \widetilde{q}^{t-1}\big)\Big) \end{aligned} \tag{8}$$

where $V > 0$ provides the tradeoff between the minimax objective and the constraint, $\eta > 0$ is the learning rate, $D(\cdot\,|\,\cdot)$ is the unnormalized Kullback-Leibler divergence with a slightly abuse in a way that $D(q\,|\,q') := D(q_1\,|\,q_1') - D(q_2\,|\,q_2')$, $\widetilde{q}_1^{t-1}$ and $\widetilde{q}_2^{t-1}$ are mixing policies, e.g.,

$$\widetilde{q}_1^{t-1}(x, a) \ = \ (1 - \theta)\, \widehat{q}_1^{t-1}(x, a) \ + \ \theta\, \frac{1}{|X_\ell||A|} \tag{9}$$

for $(x, a) \in X_\ell \times A$, $\ell \in \{0, 1, \ldots, L-1\}$, $\theta \in (0, 1]$. The mixing step ensures the uniform boundedness of KL divergence and also adds extra exploration into policy search (Wei et al., 2020). Moreover, we offer an efficient implementation of (8) as solving a convex program in Appendix 8.

Once we obtain $\widehat{q}^{\,t}$, we next perform the dual update. If we treat two λ-related regularization terms in $L^t(\widehat{q}_1^{\,t}, \widehat{q}_2^{\,t}; \lambda)$ separately, then gradient ascent/descent over either λ leads to the same update rule using the constraint violation $\langle \widehat{q}_1^{\,t}, g^{t-1} \rangle + \langle \widehat{q}_2^{\,t}, h^{t-1} \rangle - b$. Hence, the dual update works in the usual way by adding up all past constraint violations,

$$\lambda^t \;=\; \max\big(\lambda^{t-1} + (\langle \widehat{q}_1^{\,t}, g^{t-1} \rangle + \langle \widehat{q}_2^{\,t}, h^{t-1} \rangle - b\,),\, 0\big). \tag{10}$$

The dual update (10) increases λ^{t-1} when $\widehat{q}^{\,t}$ violates the approximate constraint $\langle q_1, g^{t-1} \rangle + \langle q_2, h^{t-1} \rangle \le b$. It penalizes both players by yielding individual gains to the constraint satisfaction. The dual update finds uses in constrained optimization (Wei et al., 2020) and constrained MDP problems (Efroni et al., 2020; Ding et al., 2021).

The Estimation of Confidence Sets. To deal with unknown transitions P_1 and P_2, we employ the upper confidence bound (Jaksch et al., 2010; Neu et al., 2010) to estimate occupancy measure sets $\Delta(P_1)$, $\Delta(P_2)$. We exploit players' history trajectories to estimate their true transitions: P_1, P_2, and describe estimation uncertainty as confidence sets. The estimation proceeds in epochs as follows.

Let the epoch index for the min-player be $k_1 \in \{1, 2, \dots\}$ and the epoch index for the max-player be $k_2 \in \{1, 2, \dots\}$. We may represent them by k_1^t and k_2^t for showing the dependence on episode t. The epoch counters work in the following way. For each player, e.g., the min-player, we denote by $N_1^{k_1}(x,a)$ and $M_1^{k_1}(x,a,x')$ the total numbers of visitations to (x,a) and (x,a,x') before epoch k_1, respectively; we represent the total numbers of visitations to (x,a) and (x,a,x') in epoch k_1 by $n_1^{k_1}(x,a)$ and $m_1^{k_1}(x,a,x')$, respectively; If there exists (x,a) such that $n_1^{k_1}(x,a) \ge N_1^{k_1}(x,a)$, then we set a new epoch by increasing k_1 by one. Similarly, we define $N_2^{k_2}(y,b)$, $M_2^{k_2}(y,b,y')$, $n_2^{k_2}(y,b)$, and $m_2^{k_2}(y,b,y')$ for the max-player. Using the defined epoch and visitation counters, we empirically estimate the true transitions P_1 or P_2 in epoch k_1 or k_2 by

$$\bar{P}_1^{k_1}(x' \,|\, x,a) \;=\; \frac{M_1^{k_1}(x,a,x')}{\max(1, N_1^{k_1}(x,a))} \;\text{ and }\; \bar{P}_2^{k_2}(y' \,|\, y,b) \;=\; \frac{M_2^{k_2}(y,b,y')}{\max(1, N_2^{k_2}(y,b))}$$

for all $(x,a,x') \in X \times A \times X$ and $(y,b,y') \in Y \times B \times Y$.

Let the confidence set of epoch k_1 for the min-player be $\mathcal{P}_1^{k_1}$ and the confidence set of epoch k_2 for the max-player be $\mathcal{P}_2^{k_2}$. We take $\mathcal{P}_1^{k_1}$ and $\mathcal{P}_2^{k_2}$ as collections of transitions that deviate from the empirical ones at most $\epsilon_1^{k_1}$ and $\epsilon_2^{k_2}$,

$$\begin{aligned} \mathcal{P}_1^{k_1} \;&=\; \big\{ \widehat{P}_1 \,\big|\, \|\widehat{P}_1(\cdot\,|\,x,a) - \bar{P}_1^{k_1}(\cdot\,|\,x,a)\|_1 \le \epsilon_1^{k_1}, \forall (x,a) \big\} \\ \mathcal{P}_2^{k_2} \;&=\; \big\{ \widehat{P}_2 \,\big|\, \|\widehat{P}_2(\cdot\,|\,y,b) - \bar{P}_2^{k_2}(\cdot\,|\,y,b)\|_1 \le \epsilon_2^{k_2}, \forall (y,b) \big\} \end{aligned}$$

where we take $\epsilon_1^{k_1}(x,a) = \sqrt{\frac{2|X_{\ell(x)+1}|\log(T|A||X|/\delta)}{\max(1, N_1^{k_1}(x,a))}}$ and $\epsilon_2^{k_2}(y,b) = \sqrt{\frac{2|Y_{\ell(y)+1}|\log(T|B||Y|/\delta)}{\max(1, N_2^{k_2}(y,b))}}$, $\ell(x)$ and $\ell(y)$ are the layers that certain states belong to, and $\delta \in (0,1)$. We recall the occupancy measure sets $\Delta(P_1)$ or $\Delta(P_2)$ that are induced by the true transitions P_1 or P_2. We generalize this notion to define $\Delta(\mathcal{P}_1^{k_1^t})$ or $\Delta(\mathcal{P}_2^{k_2^t})$ as collections of all possible occupancy measures that are induced by the estimated transitions $\widehat{P}_1 \in \mathcal{P}_1^k$ or $\widehat{P}_2 \in \mathcal{P}_2^k$,

$$\Delta(k_1^t) := \Delta(\mathcal{P}_1^{k_1^t}) \;\text{ or }\; \Delta(k_2^t) := \Delta(\mathcal{P}_2^{k_2^t});\;\; \text{see (17) in Appendix 8 for explicit forms.} \tag{11}$$

Lemma 1 *Fix $\delta \in (0,1)$. With probability $1-\delta$, $\Delta(P_1) \subset \Delta(\mathcal{P}_1^{k_1})$ and $\Delta(P_2) \subset \Delta(\mathcal{P}_2^{k_2})$ for all $k_1, k_2 \in \{1, 2, \dots\}$.*

The proof of Lemma 1 follows the confidence bound construction; we provide it in Appendix 9. For all epoch k_1^t or k_2^t (episode t), the true transitions P_1 and P_2 are contained in $\mathcal{P}_1^{k_1^t}$ and $\mathcal{P}_2^{k_2^t}$, respectively, with high probability. This supports the primal update (8) such that both players are optimistically searching solutions in a large but tractable domain.

4. Performance Guarantees

In Theorem 2, we present our main theoretical result on the regret and the constraint violation for Algorithm 1. We recall the total number of games played by the algorithm T, the size of state/action spaces of the min-player $|X|$, $|A|$, and the size of state/action spaces of the max-player $|Y|$, $|B|$.

Theorem 2 (Regret Bound and Constraint Violation) *Let Assumption 1 hold. Fix $p \in (0,1)$ and $T \geq \max(|X||A|, |B||Y|)$. In Algorithm 1, we set $V = L\sqrt{T}$, $\eta = 1/(TL)$, and $\theta = 1/T$. Then, with probability $1-p$, the regret (5) and the constraint violation (6) satisfy*

$$\text{Regret}(T),\ \text{Violation}(T) \quad \leq \quad \widetilde{O}\big(\,(|X|+|Y|)\,L\sqrt{T(|A|+|B|)}\,\big)$$

where $\widetilde{O}(\cdot)$ hides the logarithmic factor $\log\frac{1}{p}$.

In Theorem 2, we prove that UCB-CSAPO enjoys $O(\sqrt{T})$ regret and $O(\sqrt{T})$ constraint violation using appropriate algorithm parameters $\{V, \eta, \theta, p\}$ and Assumption 1; see Appendix 7 for proof. Our bounds have the optimal dependence on the total number of episodes T up to some logarithmic factors. The $\sqrt{|A|+|B|}$ dependence matches the existing lower bound for the single-player case (Bai and Jin, 2020). The only suboptimal dependence comes from $|X|$, $|Y|$ that also exists in existing unconstrained loop-free stochastic shortest path problems (Rosenberg and Mansour, 2019). It is straightforward to remove knowledge of T by using the doubling trick while not altering our bounds up to logarithmic factors (Rakhlin and Sridharan, 2013).

We see that Assumption 1 does not impose any restrictions on rewards. Hence, UCB-CSAPO is robust against adversarial reward functions. Moreover, Theorem 2 carries to other settings, e.g., constrained MGs with side constraints; see Appendix 14.

5. Concluding Remarks

We have examined an episodic two-player zero-sum constrained Markov game (MG) with independent transition functions. In our setup, transition functions are unknown to agents, reward functions are adversarial, and utility functions are stochastic. We have proposed the first provably efficient algorithm for playing constrained MGs with $O(\sqrt{T})$ regret and constraint violation. Our algorithm provides a principled extension of the upper confidence reinforcement learning to deal with coupled constraints in constrained MGs. We also remark that the developed algorithmic framework can be readily applied to learning other constrained MGs, e.g., the ones that involve a single controller.

Our work opens up many interesting directions for future work, such as sharper algorithms with sample complexity lower bounds, constrained rational algorithms, and how to perform safe exploration in other models of constrained MGs.

References

Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 75–84, 2010.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, volume 70, pages 22–31, 2017.

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

Eitan Altman and Adam Shwartz. Constrained markov games: Nash equilibria. In *Advances in Dynamic Games and Applications*, pages 213–221. Birkhäuser Boston, 2000.

Eitan Altman and Eilon Solan. Constrained games: The impact of the attitude to adversary's constraints. *IEEE Transactions on Automatic Control*, 54(10):2435–2440, 2009.

Eitan Altman, Konstantin Avrachenkov, Richard Marquez, and Gregory Miller. Zero-sum constrained stochastic games with independent state processes. *Mathematical Methods of Operations Research*, 62(3):375–386, 2005.

Eitan Altman, Saswati Sarkar, and Eilon Solan. Constrained Markov games with transition probabilities controlled by a single player. In *International Conference on Performance Evaluation Methodologies and Tools*, pages 1–6, 2007.

Eitan Altman, Konstantin Avrachenkov, Nicolas Bonneau, Merouane Debbah, Rachid El-Azouzi, and Daniel Sadoc Menasche. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36(2):160–164, 2008.

Jorge Alvarez-Mena and Onésimo Hernández-Lerma. Existence of Nash equilibria for constrained stochastic games. *Mathematical Methods of Operations Research*, 63(2):261–285, 2006.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Kenneth J Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290, 1954.

Qinbo Bai, Vaneet Aggarwal, and Ather Gattami. Model-free algorithm and regret analysis for MDPs with long-term constraints. *arXiv preprint arXiv:2006.05961*, 2020a.

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

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020b.

Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic Press, 2014.

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

Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, pages 52–61, 2016.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, pages 183–221, 2010.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward Markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270, 2022a.

Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Finding correlated equilibrium of constrained Markov game: A primal-dual approach. In *Advances in Neural Information Processing Systems*, 2022b.

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

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Yu-HOng Dai and Liwei Zhang. Optimality conditions for constrained minimax optimization. *arXiv preprint arXiv:2004.09730*, 2020.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.

Frits De Nijs. *Resource-constrained multi-agent Markov decision processes*. PhD thesis, Delft University of Technology, 2019.

Frits de Nijs and Peter J Stuckey. Risk-aware conditional replanning for globally constrained multi-agent sequential decision making. In *International Conference on Autonomous Agents and MultiAgent Systems*, pages 303–311, 2020.

Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, Shalabh Bhatnagar, et al. Actor-critic algorithms for constrained multi-agent reinforcement learning. *arXiv preprint arXiv:1905.02907*, 2019.

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

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312, 2021.

Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *arXiv preprint arXiv:2206.02346*, 2022.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, pages 2826–2836, 2021.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

Seyedshams Feyzabadi. *Robot Planning with Constrained Markov Decision Processes*. PhD thesis, UC Merced, 2017.

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

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Mukul Gagrani and Ashutosh Nayyar. Weakly coupled constrained Markov decision processes in Borel spaces. In *2020 American Control Conference*, pages 2790–2795, 2020.

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

Cory Jay Girard. *STRUCTURAL RESULTS FOR CONSTRAINED MARKOV DECISION PROCESSES*. PhD thesis, Cornell University, 2018.

E Gómez-Ramırez, K Najim, and AS Poznyak. Saddle-point calculation for constrained finite Markov chains. *Journal of Economic Dynamics and Control*, 27(10):1833–1853, 2003.

Shangding Gu, Jakub Grudzien Kuba, Munning Wen, Ruiqing Chen, Ziyan Wang, Zheng Tian, Jun Wang, Alois Knoll, and Yaodong Yang. Multi-agent constrained policy optimisation. *arXiv preprint arXiv:2110.02793*, 2021.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10), 2007.

Vesal Hakami and Mehdi Dehghan. Learning stationary correlated equilibria in constrained general-sum stochastic games. *IEEE transactions on cybernetics*, 46(7):1640–1654, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Xiaofeng Jiang, Shuangwu Chen, Jian Yang, Han Hu, and Zhenliang Zhang. Finding the equilibrium for continuous constrained Markov games under the average criteria. *IEEE Transactions on Automatic Control*, 65(12):5399–5406, 2020.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning– A simple, efficient, decentralized algorithm for multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022a.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279, 2022b.

Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *AAAI Conference on Artificial Intelligence*, 2020.

Ankur A Kulkarni. *Generalized Nash games with shared constraints: existence, efficiency, refinement and equilibrium constraints*. PhD thesis, University of Illinois at Urbana-Champaign, 2011.

Ankur A Kulkarni. Games and teams with shared constraints. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375(2100):20160302, 2017.

Na Li and Jason R Marden. Decoupling coupled constraints through utility design. *IEEE Transactions on Automatic Control*, 59(8):2289–2294, 2014.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained Markov decision process. *arXiv preprint arXiv:2110.10351*, 2021.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 157–163, 1994.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021a.

Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Fast global convergence of policy optimization for constrained MDPs. *arXiv preprint arXiv:2111.00552*, 2021b.

Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Basar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2020.

Shie Mannor, John N Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(3), 2009.

Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. Solving very large weakly coupled Markov decision processes. In *AAAI/IAAI*, pages 165–172, 1998.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pages 231–243, 2010.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813, 2012.

J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Duc Thien Nguyen, William Yeoh, Hoong Chuin Lau, Shlomo Zilberstein, and Chongjie Zhang. Decentralized multi-agent reinforcement learning in average-reward dynamic DCOPs. In *AAAI conference on artificial intelligence*, 2014.

Afshin OroojlooyJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

P Parnika, Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, and Shalabh Bhatnagar. Attention actor-critic algorithm for multi-agent constrained co-operative reinforcement learning. *arXiv preprint arXiv:2101.02349*, 2021.

Edward S Pearsall. A Lagrange multiplier method for certain constrained min-max problems. *Operations Research*, 24(1):70–91, 1976.

Alexei B Piunovskiy and Xuerong Mao. Constrained Markovian decision processes: the dynamic programming approach. *Operations research letters*, 27(3):119–126, 2000.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2212–2221, 2019.

Mahshid Salemi Parizi. *Approximate dynamic programming for weakly coupled Markov decision processes with perfect and imperfect information*. PhD thesis, The University of Washington, 2018.

Lukas M Schmidt, Johanna Brosig, Axel Plinge, Bjoern M Eskofier, and Christopher Mutschler. An introduction to multi-agent reinforcement learning and review of its application to autonomous mobility. *arXiv preprint arXiv:2203.07676*, 2022.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.

Rahul Singh, Abhishek Gupta, and Ness Shroff. Learning in Markov decision processes under constraints. *IEEE Transactions on Control of Network Systems*, 2022.

Vikas Vikram Singh and N Hemachandra. A characterization of stationary Nash equilibria of constrained stochastic games with independent state processes. *Operations Research Letters*, 42 (1):48–52, 2014.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2021.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Philip S Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Provably efficient online agnostic learning in Markov games. *arXiv preprint arXiv:2010.15020*, 2020.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *URL http://www. math. washington. edu/˜ tseng/papers/apgm. pdf*, 2009.

Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained Markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806, 2020.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4994–5004, 2017.

Qingda Wei. Discrete-time constrained stochastic games with the expected average payoff criteria. *Optimization*, pages 1–32, 2020.

Qingda Wei. Constrained expected average stochastic games for continuous-time jump processes. *Applied Mathematics & Optimization*, 83(3):1277–1309, 2021.

Xiaohan Wei, Hao Yu, and Michael J Neely. Online learning in weakly coupled Markov decision processes: A convergence time study. *ACM on Measurement and Analysis of Computing Systems*, 2(1):1–38, 2018.

Xiaohan Wei, Hao Yu, and Michael J Neely. Online primal-dual mirror descent under stochastic constraints. In *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2020.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682, 2020.

Vinayaka G Yaji and Shalabh Bhatnagar. Necessary and sufficient conditions for optimality in constrained general sum stochastic games. *Systems & Control Letters*, 85:8–15, 2015.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained Markov decision processes with entropy regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 1887–1909, 2022.

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pages 1428–1438, 2017.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

Wenzhao Zhang. Discrete-time constrained average stochastic games with independent state processes. *Mathematics*, 7(11):1089, 2019.

Wenzhao Zhang and Xiaolong Zou. Constrained average stochastic games with continuous-time independent state processes. *Optimization*, pages 1–24, 2021.

Feiran Zhao and Keyou You. Primal-dual learning for the model-free risk-constrained linear quadratic regulator. In *Learning for Dynamics and Control*, pages 702–714, 2021.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Conference on Learning for Dynamics and Control*, volume 120, pages 620–629, 2020.

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems 26*, 2013.

Supplementary Materials for "Provably Efficient Generalized Lagrangian Policy Optimization for Safe Multi-Agent Reinforcement Learning"

6. Related Work

Safety constraints have gained increasing attention in the literature on multi-agent reinforcement learning (RL); see surveys (Busoniu et al., 2008; Buşoniu et al., 2010; Zhang et al., 2021; Oroojlooy-Jadid and Hajinezhad, 2019; Yang and Wang, 2020; Schmidt et al., 2022). We first discusss some related work in framework of Markov games (MGs) (Shapley, 1953; Littman, 1994).

Constrained MGs. Our work is closely related to safe multi-agent RL in constrained MGs. The constrained MGs generalize constrained MDPs (Altman, 1999) to multiple agents and Markov/stochastic games (Shapley, 1953; Littman, 1994) to account for constraints. The Nash equilibrium for constrained MGs have been studied in Altman and Shwartz (2000); Gómez-Ramırez et al. (2003); Altman et al. (2005); Alvarez-Mena and Hernández-Lerma (2006); Altman et al. (2007, 2008); Altman and Solan (2009); Singh and Hemachandra (2014) using the notion of *constrained Nash equilibrium* (which generalizes the concept of *generalized Nash equilibrium* in static games (Arrow and Debreu, 1954) to MGs) by assuming some particular transition models and constraints on reward/utility functions *a priori*. More general studies include Yaji and Bhatnagar (2015); Zhang (2019); Wei (2020, 2021); Zhang and Zou (2021). These results are not applicable to the RL setting where transition models and reward/utility functions are unknown, and only a finite number of samples are available. Recently, asymptotic convergence in learning constrained MGs was examined in Hakami and Dehghan (2015); Jiang et al. (2020) but sample efficiency, constraint satisfaction, and exploration were not addressed, except for learning general equilibria (Chen et al., 2022b). Our development fills this gap by adding built-in exploration mechanisms under constraints and proving the first non-asymptotic convergence.

Constrained MDPs. Our work is also pertinent to a rich RL literature on learning unknown constrained MDPs (Zheng and Ratliff, 2020; Qiu et al., 2020; Kalagarla et al., 2020; Bai et al., 2020a; Chow et al., 2017; Tessler et al., 2019; Ding et al., 2020, 2021, 2022; Wachi and Sui, 2020; Efroni et al., 2020; Brantley et al., 2020; Chen et al., 2021; Liu et al., 2021a; Ying et al., 2022; Liu et al., 2021b; Bai et al., 2022; Zhao and You, 2021; Li et al., 2021; Chen et al., 2022a). While these results provide provably efficient algorithms regarding regret and constraint satisfaction in the single-agent setting, they are not applicable to our multi-agent game being played under constraints, because of the *non-convexity* nauture of constrained multi-agent policy optimization and the *non-stationary* environment each agent is facing. An extended line of work on constrained MDPs focuses on cooperative multi-agent learning under constraints and most efforts study the case where multiple agents have independent MDPs with a coupled budget/resource constraint (Meuleau et al., 1998; Boutilier and Lu, 2016; Wei et al., 2018; de Nijs and Stuckey, 2020; Gagrani and Nayyar, 2020). All these results assume that transition models or system dynamics are known. Only a few studies considered the shared MDP case (Diddigi et al., 2019; Lu et al., 2020; Parnika et al., 2021; Gu et al., 2021), but they either lack of theoretical guarantees or do not handle exploration. In contrast, our work focuses on the MG setting with unknown transition models, and attacks the exploration challenge directly.

Single-agent RL in MDPs & multi-agent RL in MGs. A considerable literature has provided sample-efficient online RL methods in single-agent and multi-agent unconstrained RL settings; see recent summaries in Foster et al. (2021); Du et al. (2021); Jin et al. (2021) for single-agent RL and Jin et al. (2022b,a); Song et al. (2021) for multi-agent RL. However, it is largely open to extend those sample-efficient online RL methods to constrained MGs due to several technical challenges. First, since the Bellman optimality fails even in constrained MDPs (Piunovskiy and Mao, 2000; Borkar, 2005) and the optimal constrained policy is often stochastic (Altman, 1999), value-based RL methods are not suitable. Second, applying policy-based RL methods often warrants solving constrained policy optimization problems that are not convex (Achiam et al., 2017; Ding et al., 2020), not mentioning multi-agent policy optimization problems. Third, designing a sample-efficient online RL algorithm for constrained MGs has to deal with the fundamental exploitation/exploration tradeoff under constraints (Efroni et al., 2020; Brantley et al., 2020; Ding et al., 2021). Despite some recent progress in dealing with each technical issue individually, it is crucial to address them together for multi-agent RL in constrained MGs. In this work, we offer the first positive answer by identifying a class of zero-sum constrained MGs, establishing a new policy optimization algorithm with online exploration for learning such games, and proving near-optimal sample efficiency.

7. Proof Sketch of Theorem 2

Regret Analysis. We recall that our algorithm maintains the occupancy measures $(\widehat{q}_1^t, \widehat{q}_2^t)$ for estimating policies (π^t, μ^t) and Problem (4) defines the comparison solution $(q_1^\star, q_2^\star)$ in hindsight. Naturally, we decompose the regret (5) into two side regrets for both players by inserting $\langle q_1^\star \cdot q_2^\star, r^t \rangle$. By the occupancy measures (q_1^t, q_2^t) associated with (π^t, μ^t) under the true transitions P_1 and P_2, we further decompose two side regrets into two terms by inserting $\langle \widehat{q}_1^t \cdot q_2^\star, r^t \rangle$ and $\langle q_1^\star \cdot \widehat{q}_2^t, r^t \rangle$, individually. Specifically, we have

$$\text{Regret}(T) \;=\; \underbrace{\sum_{t=0}^{T-1} \left\langle \widehat{q}_1^t \cdot q_2^\star - q_1^\star \cdot \widehat{q}_2^t, r^t \right\rangle}_{\widehat{\text{Regret}}(T)} + \underbrace{\sum_{t=0}^{T-1} \left\langle (q_1^t - \widehat{q}_1^t) \cdot q_2^\star, r^t \right\rangle}_{\text{Error}_1} + \underbrace{\sum_{t=0}^{T-1} \left\langle q_1^\star \cdot (\widehat{q}_2^t - q_2^t), r^t \right\rangle}_{\text{Error}_2}$$

where $\widehat{\text{Regret}}(T)$ depicts a regret of an online primal-dual mirror descent problem, Error_1 is the error of using $\widehat{q}_1^t$ for the min-player, and Error_2 is the error of using $\widehat{q}_2^t$ for the max-player.

We begin with a relatively standard lemma on estimation errors of $\widehat{q}_1^t$, $\widehat{q}_2^t$; we prove it in Appendix 10.

Lemma 3 *Fix $\delta \in (0,1)$. Then, with probability $1 - 2\delta$,*

$$\begin{aligned}
\sum_{t=0}^{T-1} \left\| \widehat{q}_1^t - q_1^t \right\|_1 \;&\leq\; O\Big(L|X| \sqrt{T|A| \log \tfrac{T|X||A|}{\delta}} \Big) \\
\sum_{t=0}^{T-1} \left\| \widehat{q}_2^t - q_2^t \right\|_1 \;&\leq\; O\Big(L|Y| \sqrt{T|B| \log \tfrac{T|Y||B|}{\delta}} \Big).
\end{aligned}$$

We note that $r^t \in [0,1]$, $q_2^\star$ is a probability distribution, and $\text{Error}_1 = \sum_{t=0}^{T-1} \langle (q_1^t - \widehat{q}_1^t) \cdot q_2^\star, r^t \rangle \le \sum_{t=0}^{T-1} \| q_1^t - \widehat{q}_1^t \|_1$. Application of Lemma 3 yields the following bounds on Error_1 and Error_2.

Lemma 4 *Fix $\delta \in (0,1)$. Then, with probability $1-2\delta$,*

$$\text{Error}_1 \;\le\; O\Big(L|X|\sqrt{T|A|\log\tfrac{T|X||A|}{\delta}}\Big) \;\text{ and }\; \text{Error}_2 \;\le\; O\Big(L|Y|\sqrt{T|B|\log\tfrac{T|Y||B|}{\delta}}\Big).$$

We next bound $\widehat{\text{Regret}}(T)$ by establishing an upper bound in Lemma 5 first that is crucial to our regret analysis. The proof idea of Lemma 5 is similar to the analysis of online constrained convex optimization (Yu et al., 2017; Wei et al., 2020). A distinction is that we analyze the primal update (8) via a new property of KL divergence for the minimax objective; see it in Appendix 11.

Lemma 5 *Fix $\delta \in (0,1)$. Then, with probability $1-\delta$,*

$$\begin{aligned}\widehat{\text{Regret}}(T) \;\;\le\;\; & V^{-1}\sum_{t=0}^{T-1}\lambda^t\big(\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\big) \\ & + (\eta V)^{-1}L(1+\theta T)\big(\log(|X||A|)+\log(|Y||B|)\big) \;+\; (2V^{-1}L + 4\theta + \eta V)LT.\end{aligned}$$

Lemma 5 establishes an upper bound relying on a stochastic process of duals $\{\lambda^t, t \ge 0\}$. To analyze this bound, we establish the boundedness of λ^t in Lemma 6 first. Then, we apply a general Azuma-Hoeffding inequality for supermartingales in Lemma 7. We delay their proofs to Appendix 12.

Lemma 6 *Let Assumption 1 hold. Fix $\delta \in (0,1)$. For any integer $t_0 > 0$, with probability $1 - T\delta$,*

$$\lambda^t \;\le\; \Theta + 2t_0 L + t_0\frac{64L^2}{\xi}\log\left(\frac{128L^2}{\xi}\right) + t_0\frac{64L^2}{\xi}\log\frac{1}{\delta}$$

for all $t = 1,\dots,T$, where $\xi > 0$ and

$$\Theta \;\;:=\;\; t_0\left(\tfrac{1}{2}\xi + 2L\right) \;+\; \tfrac{4L^2+(8\theta+2\eta V+2)VL}{\xi} \;+\; \tfrac{2L(\log(|X||A|/\theta)+\log(|Y||B|/\theta))}{t_0\xi\eta}.$$

Lemma 7 *Let Assumption 1 hold. Fix $\delta \in (0,1)$. For any integer $t_0 > 0$, with probability $1 - 2T\delta$,*

$$\sum_{t=0}^{T-1}\lambda^t\big(\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\big) \;\le\; \sqrt{2Tc^2\log(1/(\delta T))}$$

where $c := 2\Theta L + 4t_0L^2 + \frac{128 t_0 L^3}{\xi}\left(\log\left(\frac{128L^2}{\xi}\right) + \log\frac{1}{\delta}\right)$ and $\xi > 0$.

We now ready to conclude a bound on $\widehat{\text{Regret}}(T)$ by combining Lemma 7 and Lemma 5.

Theorem 8 *Let Assumption 1 hold. Fix $T \ge \max(|X||A|, |B||Y|)$. Let $V = L\sqrt{T}$, $\eta = 1/(TL)$, $t_0 = \sqrt{T}$, and $\theta = 1/T$. Then, with probability $1-2T\delta$ it holds that*

$$\widehat{\text{Regret}}(T) \;\le\; \widetilde{O}\big((|X|+|Y|)L\sqrt{T}\big).$$

Proof Using the given parameters V, η, t_0, and θ for Lemma 5, $\widehat{\text{Regret}}(T)$ is upper bounded by $\frac{1}{L\sqrt{T}}\sum_{t=0}^{T-1}\lambda^t\big(\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\big) + \widetilde{O}(L\sqrt{T})$ with probability $1-\delta$. We note that $\Theta \leq \widetilde{O}(L^2\sqrt{T})$ and $T \geq \max(|X||A|, |B||Y|)$. Using parameters in Lemma 7, with probability $1-2T\delta$,

$$\sum_{t=0}^{T-1}\lambda^t\big(\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\big) \;\leq\; \widetilde{O}(L^3 T).$$

We complete the proof by noting $L \leq |X| + |Y|$. ■

We conclude the regret bound in Theorem 2 by combining Lemma 4 and Theorem 8, and $\delta = p/(2T)$.

Constraint Violation Analysis. We begin with a decomposition using the auxiliary occupancy measures (q_1^t, q_2^t). By inserting $\langle \widehat{q}_1^t, g^t\rangle$ and $\langle \widehat{q}_2^t, h^t\rangle$ into Violation(T), we have

$$\text{Violation}(T) \;=\; \underbrace{\left[\sum_{t=0}^{T-1}\left(\left\langle \widehat{q}_1^t, g^t\right\rangle + \left\langle \widehat{q}_2^t, h^t\right\rangle - b\right)\right]_+}_{\widehat{\text{Violation}}(T)} + \underbrace{\sum_{t=0}^{T-1}\left\langle q_1^t - \widehat{q}_1^t, g^t\right\rangle}_{\text{Error}_3} + \underbrace{\sum_{t=0}^{T-1}\left\langle q_2^t - \widehat{q}_2^t, h^t\right\rangle}_{\text{Error}_4}.$$

Similar to Lemma 4, we can prove the following bounds on Error_3 and Error_4.

Lemma 9 *Fix $\delta \in (0,1)$. Then, with probability $1-2\delta$,*

$$\text{Error}_3 \;\leq\; O\Big(L|X|\sqrt{T|A|\log\tfrac{T|X||A|}{\delta}}\Big) \;\; and \;\; \text{Error}_4 \;\leq\; O\Big(L|Y|\sqrt{T|B|\log\tfrac{T|Y||B|}{\delta}}\Big).$$

We next bound $\widehat{\text{Violation}}(T)$ by applying the epoch property (Jaksch et al., 2010); see a proof in Appendix 13.

Theorem 10 *Let $V = L\sqrt{T}$, $\eta = 1/(TL)$, $t_0 = \sqrt{T}$, and $\theta = 1/T$. Then,*

$$\widehat{\text{Violation}}(T) \;\;\leq\;\; \lambda^T + \frac{2}{T-1}\sum_{t=1}^{T}\lambda^{t-1} + \widetilde{O}\big(L\sqrt{T(|X||A|+|Y||B|)}\big).$$

To get the violation bound, we apply Lemma 6 to Theorem 10, use Lemma 9, and take $\delta = p/(2T)$.

8. Efficient Implementation of (8)

In this section, we provide an efficient implementation for the primal update (8).

Since the minimax objective in the primal update (8) is separable for two players, it is equivalent to update two occupancy measures individually via

$$\widehat{q}_1^t \;=\; \underset{q_1\,\in\,\Delta(k_1^t)}{\text{argmin}}\; V\left\langle q_1\cdot\widehat{q}_2^{t-1}, r^{t-1}\right\rangle + \lambda^{t-1}\langle q_1, g^{t-1}\rangle + \eta^{-1}D\big(q_1\,|\,\widetilde{q}_1^{t-1}\big) \tag{12a}$$

$$\widehat{q}_2^t \;=\; \underset{q_2\,\in\,\Delta(k_2^t)}{\text{argmax}}\; V\left\langle \widehat{q}_1^{t-1}\cdot q_2, r^{t-1}\right\rangle - \lambda^{t-1}\langle q_2, h^{t-1}\rangle - \eta^{-1}D\big(q_2\,|\,\widetilde{q}_2^{t-1}\big). \tag{12b}$$

Note that $\langle q_1 \cdot \widehat{q}_2^{\,t-1}, r^{t-1} \rangle = \langle q_1, \widehat{q}_2^{\,t-1} \cdot r^{t-1} \rangle$ and $\langle \widehat{q}_1^{\,t-1} \cdot q_2, r^{t-1} \rangle = \langle q_2, \widehat{q}_1^{\,t-1} \cdot r^{t-1} \rangle$. Let

$$\phi_1^{t-1} \; := \; V\,\widehat{q}_2^{\,t-1} \cdot r^{t-1} \, + \, \lambda^{t-1} g^{t-1} \;\; \text{and} \;\; \phi_2^{t-1} \; := \; -V\,\widehat{q}_1^{\,t-1} \cdot r^{t-1} \, + \, \lambda^{t-1} h^{t-1}.$$

We can express (12) in a more compact form,

$$\widehat{q}_1^{\,t} \;=\; \mathop{\mathrm{argmin}}_{q_1 \,\in\, \Delta(k_1^t)} \; \eta \, \langle q_1, \phi_1^{t-1} \rangle \,+\, D\big(q_1 \,|\, \widetilde{q}_1^{\,t-1}\big) \tag{13a}$$

$$\widehat{q}_2^{\,t} \;=\; \mathop{\mathrm{argmin}}_{q_2 \,\in\, \Delta(k_2^t)} \; \eta \, \langle q_2, \phi_2^{t-1} \rangle \,+\, D\big(q_2 \,|\, \widetilde{q}_2^{\,t-1}\big) \tag{13b}$$

where we flip the argmax in (12b) to write argmin in (13b) and scale both objectives by multiplying $\eta > 0$.

Now, we state an efficient implementation for the primal update (8) by solving convex optimization problems. The proof is based on the method of Lagrange multipliers and the Lagrange duality theory; they also find uses in the literature (Zimin and Neu, 2013; Rosenberg and Mansour, 2019; Jin et al., 2020).

Lemma 11 (Efficient Implementation) *The primal update (8) is equivalent to*

$$\widehat{q}_1^{\,t}(x,a) \;=\; \frac{\widetilde{q}_1^{\,t}(x,a)}{Z_{1,\ell}^t(\beta_1^t, \mu_1^{+,t}, \mu_1^{-,t})}\, \mathrm{e}^{-B_{1,t}^{\beta_1^t, \mu_1^{+,t}, \mu_1^{-,t}}(x,a,x')} \tag{14a}$$

$$\widehat{q}_2^{\,t}(y,b) \;=\; \frac{\widetilde{q}_2^{\,t}(x,a)}{Z_{2,\ell}^t(\beta_2^t, \mu_2^{+,t}, \mu_2^{-,t})}\, \mathrm{e}^{-B_{2,t}^{\beta_2^t, \mu_2^{+,t}, \mu_2^{-,t}}(y,b,y')} \tag{14b}$$

where $B_{1,t}^{\beta_1, \mu_1^+, \mu_1^-}(x,a,x')$ *and* $B_{2,t}^{\beta_2, \mu_2^+, \mu_2^-}(y,b,y')$ *are given by*

$$\begin{aligned} B_{1,t}^{\beta_1, \mu_1^+, \mu_1^-}(x,a,x') \;\; &:= \;\; \beta_1(x') \,-\, \beta_1(x) \,+\, \eta \phi_1^{t-1} \\ &\quad + (1 - \epsilon_1^{k_1}(x,a)) \mu_1^+(x,a,x') \,-\, (1 + \epsilon_1^{k_1}(x,a)) \mu_1^-(x,a,x') \\ &\quad + \sum_{x'' \,\in\, X_{\ell+1}} \bar{P}_1^{k_1}(x'' \,|\, x,a) (\mu_1^-(x,a,x'') - \mu_1^+(x,a,x'')) \end{aligned}$$

$$\begin{aligned} B_{2,t}^{\beta_2, \mu_2^+, \mu_2^-}(y,b,y') \;\; &:= \;\; \beta_2(y') \,-\, \beta_2(y) \,+\, \eta \phi_2^{t-1} \\ &\quad + (1 - \epsilon_2^{k_2}(y,b)) \mu_2^+(y,b,y') \,-\, (1 + \epsilon_2^{k_2}(y,b)) \mu_2^-(y,b,y') \\ &\quad + \sum_{y'' \,\in\, Y_{\ell+1}} \bar{P}_2^{k_2}(y'' \,|\, y,b) (\mu_2^-(y,b,y'') - \mu_2^+(y,b,y'')) \end{aligned}$$

and $Z_{1,\ell}^t(\beta_1, \mu_1^+, \mu_1^-)$ *and* $Z_{2,\ell}^t(\beta_2, \mu_2^+, \mu_2^-)$ *are given by*

$$Z_{1,\ell}^t(\beta_1, \mu_1^+, \mu_1^-) \;=\; \sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} \widetilde{q}_1^{\,t}(x,a)\, \mathrm{e}^{-B_{1,t}^{\beta_1, \mu_1^+, \mu_1^-}(x,a,x')}$$

$$Z_{2,\ell}^t(\beta_2, \mu_2^+, \mu_2^-) \;=\; \sum_{y \in Y_\ell} \sum_{b \in B} \sum_{y' \in Y_{\ell+1}} \widetilde{q}_2^{\,t}(y,b)\, \mathrm{e}^{-B_{2,t}^{\beta_2, \mu_2^+, \mu_2^-}(y,b,y')}$$

and the dual variables $\beta_1^t(x)$*,* $\mu_1^{+,t}(x,a,x')$*,* $\mu_1^{-,t}(x,a,x')$ *and* $\beta_2^t(y)$*,* $\mu_2^{+,t}(y,b,y')$*,* $\mu_2^{-,t}(y,b,y')$ *are the solutions to*

$$\beta_1^t, \mu_1^{+,t}, \mu_1^{-,t} \;=\; \operatorname*{argmin}_{\beta_1,\,\mu_1^+,\,\mu_1^- \,\geq\, 0} \; \sum_{\ell\,=\,0}^{L-1} \ln Z_{1,\ell}^t(\beta_1, \mu_1^+, \mu_1^-)$$

$$\beta_2^t, \mu_2^{+,t}, \mu_2^{-,t} \;=\; \operatorname*{argmin}_{\beta_2,\,\mu_2^+,\,\mu_2^- \,\geq\, 0} \; \sum_{\ell\,=\,0}^{L-1} \ln Z_{2,\ell}^t(\beta_2, \mu_2^+, \mu_2^-).$$

Proof In (13), we have two standard mirror descent problems. Since two problems enjoy the same structure, we only prove an efficient solution to the first problem (13a).

By the online mirror descent optimization (Zimin and Neu, 2013), Problem (13a) is equivalent to

$$\bar{q}_1^t \;=\; \operatorname*{argmin}_{q_1} \; \eta\,\langle q_1, \phi_1^{t-1} \rangle \,+\, D\big(q_1 \,|\, \widetilde{q}_1^{\,t-1}\big) \;\; \text{and} \;\; \widehat{q}_1^{\,t} \;=\; \operatorname*{argmin}_{q_1 \,\in\, \Delta(k_1^t)} \; D\big(q_1 \,|\, \bar{q}_1^t\big) \tag{15}$$

where $\bar{q}_1^t$ is a solution to an unconstrained problem and $\widehat{q}_1^{\,t}$ simply takes the projection of $\bar{q}_1^t$ to the domain $\Delta(k_1^t)$ in the unnormalized Kullback-Leibler divergence.

It is straightforward to compute a closed-form solution for the unconstrained problem,

$$\bar{q}_1^t(x,a) \;=\; \widetilde{q}_1^{\,t}(x,a)\, \mathrm{e}^{-\eta \phi_1^{t-1}(x,a)}, \;\; \text{for all } (x,a) \in X \times A. \tag{16}$$

To compute the projection of $\bar{q}_1^t$, we recall that the domain set $\Delta(k_1^t)$ explicitly takes the following linear constraints on q_1: $X \times A \to [0,1]$,

$$\Delta(k_1^t) \;:=\; \{q_1 : X \times A \to [0,1] \,|\, q_1 \text{satisfies the following (i), (ii), (iii), (iv)}\} \tag{17}$$

(i) $q_1(x,a) = \sum_{x' \,\in\, X_{\ell+1}} q_1(x,a,x')$ for $(x,a) \in X_\ell \times A$ and $\ell \in \{0,1,\cdots,L-1\}$;

(ii) $\sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} q_1(x,a,x') = 1$ for $\ell \in \{0,1,\cdots,L-1\}$;

(iii) $\sum_{x \in X_{\ell-1}} \sum_{a \in A} q_1(x,a,x') = \sum_{a \in A} \sum_{x'' \in X_{\ell+1}} q_1(x',a,x'')$ for $x' \in X_\ell$ and $\ell \in \{1,\cdots,L-1\}$;

(iv) $q_1(x,a,x') - \bar{P}_1^{k_1}(x' \,|\, x,a) \sum_{x'' \,\in\, X_{\ell+1}} q_1(x,a,x'') \leq \epsilon(x,a,x')$,

$\bar{P}_1^{k_1}(x' \,|\, x,a) \sum_{x'' \,\in\, X_{\ell+1}} q_1(x,a,x'') - q_1(x,a,x') \leq \epsilon(x,a,x')$,

and $\sum_{x' \,\in\, X_{\ell+1}} \epsilon(x,a,x') \leq \epsilon_1^{k_1}(x,a) \sum_{x' \,\in\, X_{\ell+1}} q_1(x,a,x')$ for $(x,a,x') \in X_\ell \times A \times X_{\ell+1}$ and $\ell \in \{0,1,\cdots,L-1\}$.

where (ii) and (iii) follow the occupancy measure's property and (iv) displays the confidence set condition for $q_1 \in \Delta(k_1^t)$,

$$\left\| \frac{q_1(x,a,\cdot)}{\sum_{x'' \,\in\, X_{\ell+1}} q_1(x,a,x'')} - \bar{P}_1^{k_1}(\cdot \,|\, x,a) \right\|_1 \;\leq\; \epsilon_1^{k_1}(x,a), \;\; \text{for all } (x,a) \in X \times A$$

and we also introduce ϵ: $X \times A \times X \to [0, \infty)$ additionally. Therefore, the projection problem is a convex optimization with the linear constraints. By the method of Lagrange multipliers, we have the following Lagrangian $\mathcal{L}(q_1, \epsilon; \alpha, \lambda, \beta, \mu^+, \mu^-, \mu)$,

$$
\begin{aligned}
&\mathcal{L}(q_1, \epsilon; \alpha, \lambda, \beta, \mu^+, \mu^-, \mu) \\
&= D\big(q_1 \,|\, \bar{q}_1^t\big) + \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \alpha(x,a) \left(q_1(x,a) - \sum_{x' \in X_{\ell+1}} q_1(x,a,x') \right) \\
&+ \sum_{\ell=0}^{L-1} \lambda_\ell \left(\sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} q_1(x,a,x') - 1 \right) \\
&+ \sum_{\ell=1}^{L-1} \sum_{x' \in X_\ell} \beta(x') \left(\sum_{x \in X_{\ell-1}} \sum_{a \in A} q_1(x,a,x') - \sum_{a \in A} \sum_{x'' \in X_{\ell+1}} q_1(x',a,x'') \right) \\
&+ \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} \mu^+(x,a,x') \left(q_1(x,a,x') - \bar{P}_1^{k_1}(x' \,|\, x,a) \sum_{x'' \in X_{\ell+1}} q_1(x,a,x'') - \epsilon(x,a,x') \right) \\
&+ \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} \mu^-(x,a,x') \left(\bar{P}_1^{k_1}(x' \,|\, x,a) \sum_{x'' \in X_{\ell+1}} q_1(x,a,x'') - q_1(x,a,x') - \epsilon(x,a,x') \right) \\
&+ \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \mu(x,a) \left(\sum_{x' \in X_{\ell+1}} \epsilon(x,a,x') - \epsilon_1^{k_1}(x,a) \sum_{x' \in X_{\ell+1}} q_1(x,a,x') \right)
\end{aligned}
$$

where $\alpha(x,a)$, λ_ℓ, $\beta(x)$, $\mu^+(x,a,x') \geq 0$, $\mu^-(x,a,x') \geq 0$, and $\mu(x,a,x') \geq 0$ for $(x,a,x') \in X_\ell \times A \times X_{\ell+1}$ are Lagrange multipliers associated to the linear constraints.

By the Lagrange duality theory, the strong duality holds. To find the optimal solution to the projection problem in (15), it suffices to check the first-order stationary conditions. We first take the derivative over $\epsilon(x,a,x')$ for $(x,a,x') \in X_\ell \times A \times X_{\ell+1}$,

$$
\frac{\partial \mathcal{L}}{\partial \epsilon(x,a,x')} = -\,\mu^+(x,a,x') - \mu^-(x,a,x') + \mu(x,a)
$$

which is zero if we take $\mu(x,a) = \mu^+(x,a,x') + \mu^-(x,a,x')$. Using this stationary condition, we simplify the Lagrangian $\mathcal{L}(q_1, \epsilon; \alpha, \lambda, \beta, \mu^+, \mu^-, \mu)$ by eliminating μ and ϵ into,

$$
\begin{aligned}
&\mathcal{L}(q_1; \alpha, \lambda, \beta, \mu^+, \mu^-) \\
&= D\big(q_1 \,|\, \bar{q}_1^t\big) + \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \alpha(x,a) \left(q_1(x,a) - \sum_{x' \in X_{\ell+1}} q_1(x,a,x') \right) \\
&+ \sum_{\ell=0}^{L-1} \lambda_\ell \left(\sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} q_1(x,a,x') - 1 \right) \\
&+ \sum_{\ell=1}^{L-1} \sum_{x' \in X_\ell} \beta(x') \left(\sum_{x \in X_{\ell-1}} \sum_{a \in A} q_1(x,a,x') - \sum_{a \in A} \sum_{x'' \in X_{\ell+1}} q_1(x',a,x'') \right) \\
+ &\sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} \mu^+(x,a,x') \left((1 - \epsilon_1^{k_1}(x,a)) q_1(x,a,x') - \bar{P}_1^{k_1}(x' \,|\, x,a) \sum_{x'' \in X_{\ell+1}} q_1(x,a,x'') \right) \\
+ &\sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \sum_{x' \in X_{\ell+1}} \mu^-(x,a,x') \left(\bar{P}_1^{k_1}(x' \,|\, x,a) \sum_{x'' \in X_{\ell+1}} q_1(x,a,x'') - (1 + \epsilon_1^{k_1}(x,a)) q_1(x,a,x') \right).
\end{aligned}
$$

For the notational simplicity, we take $\beta(x_0) = \beta(x_L) = 0$. We next check the first-order stationary conditions of $\mathcal{L}(q_1; \alpha, \lambda, \beta, \mu^+, \mu^-)$ and solve them for the stationary point. We first take the derivative over $q_1(x,a,x')$ and $q_1(x,a)$ for $(x,a,x') \in X_\ell \times A \times X_{\ell+1}$, respectively,

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial q_1(x,a,x')} \;=\; & -\alpha(x,a) + \lambda_\ell + \beta(x') - \beta(x) \\
& + (1 - \epsilon_1^{k_1}(x,a)) \mu^+(x,a,x') - (1 + \epsilon_1^{k_1}(x,a)) \mu^-(x,a,x') \\
& + \sum_{x'' \in X_{\ell+1}} \bar{P}_1^{k_1}(x'' \,|\, x,a) (\mu^-(x,a,x'') - \mu^+(x,a,x''))
\end{aligned}
$$

$$
\frac{\partial \mathcal{L}}{\partial q_1(x,a)} \;=\; \ln q_1(x,a) - \ln \bar{q}_1^t(x,a) + \alpha(x,a).
$$

By setting the second derivative above to be zero, we have $\alpha(x,a) = -\ln q_1(x,a) + \ln \bar{q}_1^t(x,a)$. Then, substituting it into the first zero-derivative by eliminating $\alpha(x,a)$ yields,

$$
\ln q_1(x,a) \;=\; \ln \bar{q}_1^t(x,a) + \eta \phi_1^{t-1}(x,a) - \lambda_\ell - B^t(x,a,x')
$$

$$
\begin{aligned}
B^t(x,a,x') \;=\; & \beta(x') - \beta(x) + \eta \phi_1^{t-1}(x,a) \\
& + (1 - \epsilon_1^{k_1}(x,a)) \mu^+(x,a,x') - (1 + \epsilon_1^{k_1}(x,a)) \mu^-(x,a,x') \\
& + \sum_{x'' \in X_{\ell+1}} \bar{P}_1^{k_1}(x'' \,|\, x,a) (\mu^-(x,a,x'') - \mu^+(x,a,x'')).
\end{aligned}
$$

The solution $q_1^\star(x,a)$ leads to an explicit formula for $\widehat{q}_1^t$,

$$
\widehat{q}_1^t(x,a) \;=\; q_1^\star(x,a) \;=\; \bar{q}_1^t(x,a)\, \mathrm{e}^{\eta \phi_1^{t-1}(x,a) - \lambda_\ell - B^t(x,a,x')} \;=\; \widetilde{q}_1^t(x,a)\, \mathrm{e}^{-\lambda_\ell - B^t(x,a,x')} \tag{18}
$$

where the last equality is due to (16) and $x \neq x_L$. We note that it is not unique to determine $\alpha(x,a)$ since it takes the form $\alpha^\star(x,a) = -\eta\phi_1^{t-1}(x,a) + \lambda_\ell + B^t(x,a,x')$ for some x'. It remains to determine the optimal β, μ^+, and μ^-.

Before showing the optimal β, μ^+, and μ^-, we take another derivative over λ_ℓ at $q_1 = \widehat{q}_1^t$ and set it to be zero,

$$\sum_{x\in X_\ell}\sum_{a\in A}\sum_{x'\in X_{\ell+1}} \widehat{q}_1^t(x,a,x') \;=\; 1$$

or, equivalently,

$$\mathrm{e}^{\lambda_\ell} \;=\; \sum_{x\in X_\ell}\sum_{a\in A}\sum_{x'\in X_{\ell+1}} \widetilde{q}_1^t(x,a)\,\mathrm{e}^{-B^t(x,a,x')} \;:=\; Z_\ell^t$$

which shows that $\lambda_\ell^\star = \ln Z_\ell^t$. It also leads to $\alpha^\star(x,a) = -\eta\phi_1^{t-1}(x,a) + \lambda_\ell^\star + B^t(x,a,x')$.

We note that

$$\begin{aligned}
&\mathcal{L}(q_1;\alpha,\lambda,\beta,\mu^+,\mu^-)\\
&= D\big(q_1\,|\,\bar{q}_1^t\big) + \sum_{\ell=0}^{L-1}\sum_{x\,\in\,X_\ell}\sum_{a\,\in\,A}\sum_{x'\,\in\,X_{\ell+1}}\left(\frac{\partial\mathcal{L}}{\partial q_1(x,a,x')} + \alpha(x,a)\right)q_1(x,a,x') - \sum_{\ell=0}^{L-1}\lambda_\ell\\
&= D\big(q_1\,|\,\bar{q}_1^t\big) + \sum_{\ell=0}^{L-1}\sum_{x\,\in\,X_\ell}\sum_{a\,\in\,A}\sum_{x'\,\in\,X_{\ell+1}}\frac{\partial\mathcal{L}}{\partial q_1(x,a,x')}q_1(x,a,x')\\
&\quad + \sum_{\ell=0}^{L-1}\sum_{x\,\in\,X_\ell}\sum_{a\,\in\,A}\left(\frac{\partial\mathcal{L}}{\partial q_1(x,a)} - \ln q_1(x,a) + \ln\bar{q}_1^t(x,a)\right)q_1(x,a) - \sum_{\ell=0}^{L-1}\lambda_\ell\\
&= \sum_{\ell=0}^{L-1}\sum_{x\,\in\,X_\ell}\sum_{a\,\in\,A}\sum_{x'\,\in\,X_{\ell+1}}\frac{\partial\mathcal{L}}{\partial q_1(x,a,x')}q_1(x,a,x')\\
&\quad + \sum_{\ell=0}^{L-1}\sum_{x\,\in\,X_\ell}\sum_{a\,\in\,A}\left(\left(\frac{\partial\mathcal{L}}{\partial q_1(x,a)} - 1\right)q_1(x,a) + \ln\bar{q}_1^t(x,a)\right) - \sum_{\ell=0}^{L-1}\lambda_\ell.
\end{aligned}$$

We now collect all previously determined optimal dual variables and apply the strong duality,

$$\begin{aligned}
\beta^\star,\mu^{+,\star},\mu^{-,\star} &= \underset{\beta,\,\mu^+,\,\mu^-\,\geq\,0}{\operatorname{argmax}}\;\underset{\alpha,\,\lambda}{\text{maximize}}\;\underset{q_1}{\text{minimize}}\;\mathcal{L}(q_1;\alpha,\lambda,\beta,\mu^+,\mu^-)\\
&= \underset{\beta,\,\mu^+,\,\mu^-\,\geq\,0}{\operatorname{argmax}}\;\mathcal{L}(q_1^\star;\alpha^\star,\lambda^\star,\beta,\mu^+,\mu^-)\\
&= \underset{\beta,\,\mu^+,\,\mu^-\,\geq\,0}{\operatorname{argmax}}\;-L + \sum_{\ell=0}^{L-1}\sum_{x\,\in\,X_\ell}\sum_{a\,\in\,A}\ln\bar{q}_1^t(x,a) - \sum_{\ell=0}^{L-1}\lambda_\ell^\star\\
&= \underset{\beta,\,\mu^+,\,\mu^-\,\geq\,0}{\operatorname{argmin}}\;\sum_{\ell=0}^{L-1}\ln Z_\ell^t
\end{aligned}$$

where the third equality is due to: $\frac{\partial\mathcal{L}}{\partial q_1(x,a,x')}\big|_{q_1^\star(x,a,x')} = 0$ and $\frac{\partial\mathcal{L}}{\partial q_1(x,a)}\big|_{q_1^\star(x,a)} = 0$, and we ignore all constants that are independent of β, μ^+, and μ^- for the last equality; we note that this minimization problem is a convex optimization problem over the nonnegative orthant. Hence, we have proved the update (14a) as an efficient update (18). Similarly, we have an efficient update (14b) for the second problem (13b) and the proof is complete.

■

9. Proof of Lemma 1

For any $q_1 \in \Delta(P_1)$ and $q_2 \in \Delta(P_2)$, we estimate

$$\widehat{P}_1(\cdot \,|\, x, a) \;=\; \frac{q_1(x,a,\cdot)}{\sum_{x' \in X_{\ell+1}} q_1(x,a,x')} \;\text{ and }\; \widehat{P}_2(\cdot \,|\, y, b) \;=\; \frac{q_2(y,b,\cdot)}{\sum_{y' \in Y_{\ell+1}} q_2(y,b,y')}.$$

Consequently,

$$\begin{aligned}
&\left\| \frac{q_1(x,a,\cdot)}{\sum_{x' \in X_{\ell+1}} q_1(x,a,x')} - \bar{P}_1^{k_1}(\cdot \,|\, x,a) \right\|_1 \\
\leq\; & \left\| \frac{q_1(x,a,\cdot)}{\sum_{x' \in X_{\ell+1}} q_1(x,a,x')} - \widehat{P}_1(\cdot \,|\, x,a) \right\|_1 + \left\| \widehat{P}_1(\cdot \,|\, x,a) - \bar{P}_1^{k_1}(\cdot \,|\, x,a) \right\|_1 \\
=\; & \left\| \widehat{P}_1(\cdot \,|\, x,a) - \bar{P}_1^{k_1}(\cdot \,|\, x,a) \right\|_1
\end{aligned}$$

which implies that $q_1 \in \Delta(P_1^{k_1})$. Similarly, we have $q_2 \in \Delta(P_2^{k_2})$. Therefore, $\Delta(P_1) \subset \Delta(\mathcal{P}_1^{k_1})$ and $\Delta(P_2) \in \Delta(\mathcal{P}_2^{k_2})$. The probability argument follows Lemma 1 (Neu et al., 2012) or its original version, Lemma 17 (Jaksch et al., 2010): with probability $1-\delta$ it holds that

$$\|\widehat{P}_1(\cdot \,|\, x,a) - \bar{P}_1^{k_1}(\cdot \,|\, x,a)\|_1 \leq \epsilon_1^{k_1} \;\text{ and }\; \|\widehat{P}_2(\cdot \,|\, y,b) - \bar{P}_2^{k_2}(\cdot \,|\, y,b)\|_1 \leq \epsilon_2^{k_2}$$

for all $(x,a) \in X \times A$, $(y,b) \in Y \times B$, and all epochs k_1 and k_2.

10. Proof of Lemma 3

We recall the occupancy measures induced by the empirical transitions $\widehat{P}_1$ and $\widehat{P}_2$,

$$\begin{aligned}
\widehat{q}_1^{\,t}(x,a,x') \;&=\; \widehat{d}_1^{\,t}(x)\pi^t(a \,|\, x)\widehat{P}_1^{k_1}(x' \,|\, x,a) \;\text{ and }\; \widehat{q}_2^{\,t}(y,b,y') \;=\; \widehat{d}_2^{\,t}(y)\mu^t(b \,|\, y)\widehat{P}_2^{k_2}(y' \,|\, y,b) \\
\widehat{q}_1^{\,t}(x,a) \;&=\; \sum_{x' \in X_{\ell+1}} \widehat{q}_1^{\,t}(x,a,x') \;\text{ and }\; \widehat{q}_2^{\,t}(y,b) \;=\; \sum_{y' \in Y_{\ell+1}} \widehat{q}_2^{\,t}(y,b,y')
\end{aligned}$$

where $\widehat{d}_1^{\,t}(x)$ and $\widehat{d}_2^{\,t}(x)$ are the stationary state visitation probabilities, and the occupancy measures induced by the true transitions P_1 and P_2,

$$\begin{aligned}
q_1^t(x,a,x') \;&=\; d_1^t(x)\pi^t(a \,|\, x)P_1(x' \,|\, x,a) \;\text{ and }\; q_2^t(y,b,y') \;=\; d_2^t(y)\mu^t(b \,|\, y)P_2(y' \,|\, y,b) \\
q_1^t(x,a) \;&=\; \sum_{x' \in X_{\ell+1}} q_1^t(x,a,x') \;\text{ and }\; q_2^t(y,b) \;=\; \sum_{y' \in Y_{\ell+1}} q_2^t(y,b,y')
\end{aligned}$$

where $d_1^t(x)$ and $d_2^t(x)$ are the stationary state visitation probabilities. We denote by ℓ the layer that x or y belongs to.

We first present a useful property on how the transition estimation errors affect the mismatch of occupancy measures.

Lemma 12 *Let $\widehat{q}_1^t$, $\widehat{q}_2^t$, q_1^t, and q_2^t be generated by Algorithm 1. Then,*

$$\left\|\widehat{q}_1^t - q_1^t\right\|_1 \;\leq\; \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\pi^t(a\,|\,x)\widehat{d}_1^t(x)\left\|\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - P_1(\cdot\,|\,x,a)\right\|_1 \tag{19a}$$

$$\left\|\widehat{q}_2^t - q_2^t\right\|_1 \;\leq\; \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{y\in Y_\ell}\sum_{b\in B}\mu^t(b\,|\,y)\widehat{d}_2^t(y)\left\|\widehat{P}_2^{k_2}(\cdot\,|\,y,b) - P_2(\cdot\,|\,y,b)\right\|_1. \tag{19b}$$

Proof Since two players have the independent transitions, it suffices to just prove one of two players. We next prove (19a) for the min-player. By the definitions, we can bound $\left\|\widehat{q}_1^t - q_1^t\right\|_1$ by

$$\begin{aligned}
\left\|\widehat{q}_1^t - q_1^t\right\|_1 &= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\left|\sum_{x'\in X_{\ell+1}}\widehat{q}_1^t(x,a,x') - \sum_{x'\in X_{\ell+1}} q_1^t(x,a,x')\right| \\
&\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\left\|\widehat{q}_1^t(x,a,\cdot) - q_1^t(x,a,\cdot)\right\|_1 \\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\pi^t(a\,|\,x)\left\|\widehat{d}_1^t(x)\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - d_1^t(x)P_1(\cdot\,|\,x,a)\right\|_1
\end{aligned} \tag{20}$$

where we apply the triangle inequality to obtain the inequality. We add and subtract $\widehat{d}_1^t(x)P_1(\cdot\,|\,x,a)$ into the norm $\|\widehat{d}_1^t(x)\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - d_1^t(x)P_1(\cdot\,|\,x,a)\|_1$, and apply the triangle inequality again,

$$\begin{aligned}
&\left\|\widehat{d}_1^t(x)\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - d_1^t(x)P_1(\cdot\,|\,x,a)\right\|_1 \\
&\leq \left\|\widehat{d}_1^t(x)\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - \widehat{d}_1^t(x)P_1(\cdot\,|\,x,a)\right\|_1 + \left\|\widehat{d}_1^t(x)P_1(\cdot\,|\,x,a) - d_1^t(x)P_1(\cdot\,|\,x,a)\right\|_1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\left\|\widehat{q}_1^t(x,a,\cdot) - q_1^t(x,a,\cdot)\right\|_1 \\
&\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\pi^t(a\,|\,x)\widehat{d}_1^t(x)\left\|\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - P_1(\cdot\,|\,x,a)\right\|_1 \\
&\quad + \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\pi^t(a\,|\,x)\left|\widehat{d}_1^t(x) - d_1^t(x)\right|\left\|P_1(\cdot\,|\,x,a)\right\|_1.
\end{aligned} \tag{21}$$

We can further simplify the upper bound in (21). Using $\|P_1(\cdot\,|\,x,a)\|_1 = 1$ and $\sum_{a\in A}\pi^t(a\,|\,x) = 1$, we have

$$\sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\pi^t(a\,|\,x)\left|\widehat{d}_1^t(x) - d_1^t(x)\right|\left\|P_1(\cdot\,|\,x,a)\right\|_1 \;=\; \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\left|\widehat{d}_1^t(x) - d_1^t(x)\right|.$$

By the definitions, $\widehat{d}_1^t(x) = d_1^t(x) = 1$ for $x \in X_0$, and $\widehat{d}_1^t(x) = \sum_{x^\circ \in X_{\ell-1}} \sum_{a \in A} \widehat{q}_1^t(x^\circ, a, x)$ and $d_1^t(x) = \sum_{x^\circ \in X_{\ell-1}} \sum_{a \in A} q_1^t(x^\circ, a, x)$ for $x \in X_\ell$. Thus,

$$\begin{aligned}
& \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \left| \widehat{d}_1^t(x) - d_1^t(x) \right| \\
&= \sum_{\ell=1}^{L-1} \sum_{x \in X_\ell} \left| \widehat{d}_1^t(x) - d_1^t(x) \right| \\
&= \sum_{\ell=1}^{L-1} \sum_{x \in X_\ell} \left| \sum_{x^\circ \in X_{\ell-1}} \sum_{a \in A} \widehat{q}_1^t(x^\circ, a, x) - \sum_{x^\circ \in X_{\ell-1}} \sum_{a \in A} q_1^t(x^\circ, a, x) \right| \\
&\leq \sum_{\ell=1}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \sum_{x^\circ \in X_{\ell-1}} \left| \widehat{q}_1^t(x^\circ, a, x) - q_1^t(x^\circ, a, x) \right| \\
&= \sum_{\ell=1}^{L-1} \sum_{a \in A} \sum_{x^\circ \in X_{\ell-1}} \left\| \widehat{q}_1^t(x^\circ, a, \cdot) - q_1^t(x^\circ, a, \cdot) \right\|_1 \\
&= \sum_{\ell=0}^{L-2} \sum_{x \in X_\ell} \sum_{a \in A} \left\| \widehat{q}_1^t(x, a, \cdot) - q_1^t(x, a, \cdot) \right\|_1 .
\end{aligned}$$

We now return back to (21),

$$\begin{aligned}
& \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left\| \widehat{q}_1^t(x, a, \cdot) - q_1^t(x, a, \cdot) \right\|_1 \\
&\leq \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \pi^t(a \,|\, x) \widehat{d}_1^t(x) \left\| \widehat{P}_1^{k_1}(\cdot \,|\, x, a) - P_1(\cdot \,|\, x, a) \right\|_1 \\
&\quad + \sum_{\ell=0}^{L-2} \sum_{x \in X_\ell} \sum_{a \in A} \left\| \widehat{q}_1^t(x, a, \cdot) - q_1^t(x, a, \cdot) \right\|_1
\end{aligned} \tag{22}$$

which is a recursive formula for $\sum_{\ell=0}^{j} \sum_{x \in X_\ell} \sum_{a \in A} \left\| \widehat{q}_1^t(x, a, \cdot) - q_1^t(x, a, \cdot) \right\|_1$ over $j \in \{0, 1, \ldots, L-1\}$. By the recursion,

$$\begin{aligned}
& \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left\| \widehat{q}_1^t(x, a, \cdot) - q_1^t(x, a, \cdot) \right\|_1 \\
&\leq \sum_{j=0}^{L-1} \sum_{\ell=0}^{j} \sum_{x \in X_\ell} \sum_{a \in A} \pi^t(a \,|\, x) \widehat{d}_1^t(x) \left\| \widehat{P}_1^{k_1}(\cdot \,|\, x, a) - P_1(\cdot \,|\, x, a) \right\|_1 .
\end{aligned}$$

Finally, we complete the proof by using (20). ■

With Lemma 12 in place, we are ready to prove Lemma 3.

Proof [Proof of Lemma 3] The proof is based on Lemma 12. By (19a),

$$
\begin{aligned}
&\left\|\widehat{q}_1^t - q_1^t\right\|_1 \\
&\leq \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) - \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\left\|\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - P_1(\cdot\,|\,x,a)\right\|_1 \\
&\quad + \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\left\|\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - P_1(\cdot\,|\,x,a)\right\|_1
\end{aligned}
$$

where $\mathbb{I}\{\cdot\}$ is the indicator function that is 1 with probability $\pi^t(a\,|\,x)\widehat{d}_1^t(x)$ and 0 otherwise.

Let $\rho_1^t(x,a) := \|\widehat{P}_1^{k_1}(\cdot\,|\,x,a) - P_1(\cdot\,|\,x,a)\|_1$. Clearly, $\rho_1^t(x,a) \leq 2$. Summing $\left\|\widehat{q}_1^t - q_1^t\right\|_1$ from $t=0$ to $t=T-1$ leads to,

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\left\|\widehat{q}_1^t - q_1^t\right\|_1 \\
&\leq \sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) - \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\rho_1^t(x,a) \\
&\quad + \sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\rho_1^t(x,a)
\end{aligned}
\tag{23}
$$

where the layer ℓ depends on episode t implicitly. We next apply the martingale concentration and Lemma 1 to the right-hand side of (23).

Let $\mathcal{F}_1^t$ be an σ-algebra that is generated by the state-action sequence, reward/utility functions for the min-player up to episode t. By the definition of epoch $k_1 := k_1^t$, $\rho_1^t(x,a)$ defines over $\mathcal{F}_1^{t-1}$ only and thus,

$$
\mathbb{E}\left[\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) - \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\rho_1^t(x,a)\,\middle|\,\mathcal{F}_1^{t-1}\right] \;=\; 0.
$$

Meanwhile, it is easy to see that

$$
\begin{aligned}
&\left|\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) - \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\rho_1^t(x,a)\right| \\
&\leq\; 2\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) + \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)
\end{aligned}
$$

which is bounded by 4 since the summands are probability distributions. Hence, $\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) - \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\rho_1^t(x,a)$ is a martingale difference sequence that adapts to the filtration $\{\mathcal{F}_1^t\}_{t\geq 0}$. By the Azuma-Hoeffding inequality, with probability $1-\delta/L$ it holds that

$$
\sum_{t=0}^{T-1}\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x) - \mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\rho_1^t(x,a) \;\leq\; 4\sqrt{2T\log\frac{L}{\delta}}
\tag{24}
$$

where $\delta \in (0,1)$. By the union bound, (24) holds with probability $1-\delta$ for all $\ell \in \{0,1,\ldots,L-1\}$. Thus, with probability $1-\delta$, we have

$$\sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\left(\pi^t(a\,|\,x)\widehat{d}_1^t(x)-\mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\right)\rho_1^t(x,a) \;\le\; 2L^2\sqrt{2T\log\frac{L}{\delta}}. \quad (25)$$

For the rest, we apply Lemma 1. By the definition of epoch $k_1 := k_1^t$, we have $N_1^{k_1^t}(x,a) = \sum_{k=0}^{k_1^t-1} n_1^k(x,a)$. An application of Lemma 24 yields

$$\sum_{k=1}^{k_1^t}\frac{n_1^k(x,a)}{\max(1,\sqrt{N_1^k(x,a)})} \;\le\; 2\sqrt{N_1^{k_1^t}(x,a)}. \quad (26)$$

We note that $\sum_{x\in X_\ell}\sum_{a\in A}\mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\rho_1^t(x,a) = \|\widehat{P}_1^{k_1}(\cdot\,|\,x_\ell,a_\ell)-P_1(\cdot\,|\,x_\ell,a_\ell)\|_1$. By Lemma 1, with probability $1-\delta$ it holds that

$$\begin{aligned}
&\sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\rho_1^t(x,a)\\
=\;& \sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\|\widehat{P}_1^{k_1}(\cdot\,|\,x_\ell,a_\ell)-P_1(\cdot\,|\,x_\ell,a_\ell)\|_1\\
\le\;& \sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sqrt{\frac{2|X_{\ell+1}|\log(T|A||X|/\delta)}{\max(1,N_1^{k_1}(x_\ell,a_\ell))}}.
\end{aligned}$$

By the definition of $N_1^{k_1} := N_1^{k_1^t}$, using (26) it is convenient to have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sqrt{\frac{2|X_{\ell+1}|\log(T|A||X|/\delta)}{\max(1,N_1^{k_1}(x_\ell,a_\ell))}}\\
\le\;& \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{k=0}^{k_1^T}\sum_{x\in X_\ell}\sum_{a\in A} n_1^k(x,a)\sqrt{\frac{2|X_{\ell+1}|\log(T|A||X|/\delta)}{\max(1,N_1^k(x,a))}}\\
\le\;& \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A} 2\sqrt{2N_1^{k^T}(x,a)|X_{\ell+1}|\log\frac{T|A||X|}{\delta}}.
\end{aligned}$$

Furthermore, we can make the following simplifications. By the Jensen's inequality,

$$\begin{aligned}
&\sum_{x\in X_\ell}\sum_{a\in A} 2\sqrt{2N_1^{k^T}(x,a)|X_{\ell+1}|\log\frac{T|A||X|}{\delta}}\\
\le\;& 2\sqrt{2\sum_{x\in X_\ell}\sum_{a\in A}N_1^{k^T}(x,a)|X_{\ell+1}||X_\ell||A|\log\frac{T|A||X|}{\delta}}.
\end{aligned}$$

We also note that $\sum_{x\in X_\ell}\sum_{a\in A} N_1^{k^T}(x,a) \leq T$ and $\sqrt{|X_{\ell+1}||X_\ell|} \leq (|X_{\ell+1}|+|X_\ell|)/2$. Thus,

$$\begin{aligned}
&\sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sqrt{\frac{2|X_{\ell+1}|\ln(T|A||X|/\delta)}{\max(1,N_1^{k_1}(x_\ell,a_\ell))}}\\
\leq\ & \sum_{j=0}^{L-1}\sum_{\ell=0}^{j} 2\sqrt{2T|X_{\ell+1}||X_\ell||A|\log\frac{T|A||X|}{\delta}}\\
\leq\ & \sum_{j=0}^{L-1}\sum_{\ell=0}^{j}(|X_{\ell+1}|+|X_\ell|)\sqrt{2T|A|\log\frac{T|A||X|}{\delta}}\\
\leq\ & L|X|\sqrt{2T|A|\log\tfrac{T|A||X|}{\delta}}.
\end{aligned}$$

Therefore, with probability $1-\delta$ it holds that

$$\sum_{t=0}^{T-1}\sum_{j=0}^{L-1}\sum_{\ell=0}^{j}\sum_{x\in X_\ell}\sum_{a\in A}\mathbb{I}\{(x_\ell,a_\ell)=(x,a)\}\rho_1^t(x,a) \quad\leq\quad L|X|\sqrt{2T|A|\log\tfrac{T|A||X|}{\delta}}. \tag{27}$$

Finally, we take a union of (25) and (27) and substitute it into (23) to conclude the proof. ■

11. Proof of Lemma 5

We first present a basic property of the Kullback-Leibler divergence that generalizes similar properties in the literature (Nemirovski et al., 2009; Tseng, 2009; Wei et al., 2020) to the convex-concave minimax problems. For this purpose, we set some standard notations. Let $\mathcal{X}\subset\mathbb{R}^d$ be a convex set with non-empty interior, $\mathcal{X}^{\text{int}}\neq\emptyset$. Let ϕ: $\mathcal{X}\to\mathbb{R}$ be a function that is is continuously differentiable on $\mathcal{X}^{\text{int}}$. Let $\Delta_x\subset\mathcal{X}$ be a compact convex set containing the origin. Denote $\Delta_x^o = \Delta\cap\mathcal{X}^{\text{ int}}$ and let $\Delta_x^o\neq\emptyset$. We define the Kullback-Leibler divergence, D: $\Delta_x\times\Delta_x^o\to\mathbb{R}$,

$$D(x,x') \;:=\; \phi(x)-\phi(x')-\langle\nabla\phi(x'),x-x'\rangle.$$

An interesting case is when Δ_x becomes a probability simplex. If $\phi(x)=\sum_{i=1}^d(x_i\log x_i - x_i)$, then $D(x,x')=\sum_{i=1}^d x_i\log(x_i/x_i')-\sum_{i=1}^d(x_i-x_i')$ defines the unnormalized Kullback-Leibler divergence (Cover, 1999; Boyd et al., 2004). This is the setup we will discuss later.

Lemma 13 *Let $f(x,y)$: $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be a continuous differentiable function that is convex in x and concave in y, where $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets in $\mathbb{R}^d$. Suppose for some $x'\in\Delta_x^o$ and $y'\in\Delta_y^o$,*

$$(x^\star,y^\star)\;\in\;\operatorname*{argminimax}_{x\in\Delta_x,\,y\in\Delta_y}\; f(x,y)+\eta^{-1}D(x\,|\,x')-\eta^{-1}D(y\,|\,y')$$

and $x^\star\in\Delta_x^o$ and $y^\star\in\Delta_y^o$, where $\eta>0$. Then, for any $x\in\Delta_x$ and $y\in\Delta_y$,

$$f(x^\star,y)+\eta^{-1}\big(D(x^\star,x')+D(y^\star,y')\big)\;\leq\; f(x,y^\star)+\eta^{-1}\big(D(x,x')+D(y,y')-D(x,x^\star)-D(y,y^\star)\big).$$

Proof For the smooth convex-concave function f, it is necessary to have the first-order stationary condition on $(x^\star, y^\star)$. There exist $\nabla_x f(x^\star, y^\star)$ and $\nabla_y(x^\star, y^\star)$ such that

$$\left\langle \nabla_x f(x^\star, y^\star) + \eta^{-1}\big(\phi(x^\star) - \phi(x')\big), x - x^\star \right\rangle \;\geq\; 0, \;\; \text{for any } x \in \Delta_x \tag{28a}$$

$$\left\langle -\nabla_y f(x^\star, y^\star) + \eta^{-1}\big(\phi(y^\star) - \phi(y')\big), y - y^\star \right\rangle \;\geq\; 0, \;\; \text{for any } y \in \Delta_y. \tag{28b}$$

By the definition of $D(\cdot \,|\, \cdot)$,

$$\begin{aligned}
&\eta^{-1}\big(D(x,x') - D(x,x^\star)\big) \\
&= \;\; \eta^{-1}\big(\phi(x^\star) - \phi(x') - \langle \nabla\phi(x'), x - x'\rangle + \langle \nabla\phi(x^\star), x - x^\star\rangle\big) \\
&= \;\; \eta^{-1}\big(\phi(x^\star) - \phi(x') - \eta^{-1}\langle \nabla\phi(x'), x^\star - x'\rangle\big) - \langle \nabla_x f(x^\star, y^\star), x - x^\star\rangle \\
&\qquad + \langle \nabla_x f(x^\star, y^\star) + \eta^{-1}\big(\nabla\phi(x^\star) - \nabla\phi(x')\big), x - x^\star\rangle \\
&= \;\; \eta^{-1} D(x^\star, x') - \langle \nabla_x f(x^\star, y^\star), x - x^\star\rangle \\
&\qquad + \langle \nabla_x f(x^\star, y^\star) + \eta^{-1}\big(\nabla\phi(x^\star) - \nabla\phi(x')\big), x - x^\star\rangle.
\end{aligned}$$

Application of (28a) leads to

$$\begin{aligned}
\eta^{-1}\big(D(x,x') - D(x,x^\star)\big) \;\;&\geq\;\; \eta^{-1} D(x^\star, x') - \langle \nabla_x f(x^\star, y^\star), x - x^\star\rangle \\
&\geq\;\; \eta^{-1} D(x^\star, x') + f(x^\star, y^\star) - f(x, y^\star)
\end{aligned} \tag{29}$$

where the last inequality is due to the convexity fo $f(x,y^\star)$ in x: $f(x,y^\star) \geq f(x^\star,y^\star) + \langle \nabla_x f(x^\star, y^\star), x - x^\star\rangle$.

Similarly, we work on $\eta^{-1}\big(D(y,y') - D(y,y^\star)\big)$ and (28b).

$$\eta^{-1}\big(D(y,y') - D(y,y^\star)\big) \;\;\geq\;\; \eta^{-1} D(y^\star, y') + f(x^\star, y) - f(x^\star, y^\star). \tag{30}$$

Finally, we conclude the proof by adding (29) to (30) from both sides. ■

Before the proof of Lemma 5, we next show some useful bounds on the unnormalized Kullback-Leibler divergence.

Lemma 14 *Let $q(x,a,x')$ and $q'(x,a,x')$ be two occupancy measures, and $q(x,a)$ and $q'(x,a)$ be the associated state-action visitation probability distributions. Then,*

$$D(q,q') \;\geq\; \frac{1}{2L}\left\|q - q'\right\|_1^2.$$

Proof We recall $q(x,a)$ and $q'(x,a)$,

$$q(x,a) \;=\; \sum_{x' \in X_\ell} q(x,a,x') \;\;\text{and}\;\; q'(x,a) \;=\; \sum_{x' \in X_\ell} q'(x,a,x')$$

where ℓ is the layer that x belongs to. We note that $q(x,a)$ and $q'(x,a)$ define probability laws for each $\ell \in \{0,1,\ldots,L-1\}$, and $\sum_{x \in X_\ell}\sum_{a \in A} q(x,a) = \sum_{x \in X_\ell}\sum_{a \in A} q'(x,a) = 1$.

By the definition,

$$
\begin{aligned}
D(q,q') &= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\log\frac{q(x,a)}{q'(x,a)} - \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\big(q(x,a)-q'(x,a)\big)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\log\frac{q(x,a)}{q'(x,a)}\\
&\geq \frac{1}{2}\sum_{\ell=0}^{L-1}\big\|q(x,a)-q'(x,a)\big\|_1^2\\
&\geq \frac{1}{2L}\left(\sum_{\ell=0}^{L-1}\big\|q(x,a)-q'(x,a)\big\|_1\right)^2\\
&= \frac{1}{2L}\big\|q-q'\big\|_1^2
\end{aligned}
$$

where we apply the Pinsker's inequality to $\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\log\frac{q(x,a)}{q'(x,a)}$ in the first inequality. ■

Lemma 15 *Let $q(x,a,x')$ and $q'(x,a,x')$ be two occupancy measures, and $q(x,a)$ and $q'(x,a)$ be the associated state-action visitation probability laws. Define $\widetilde{q}\,'(x,a) = (1-\theta)q'(x,a) + \theta\frac{1}{|X_\ell||A|}$ for $(x,a)\in X_\ell\times A$, $\ell\in\{0,1,\ldots,L-1\}$, and $\theta\in(0,1]$. Then,*

$$
D(q,\widetilde{q}\,') - D(q,q') \;\leq\; \theta L\log(|X||A|) \;\; \textit{and} \;\; D(q,\widetilde{q}\,') \;\leq\; L\log(|X||A|/\theta).
$$

Proof By the definition,

$$
\begin{aligned}
&D(q,\widetilde{q}\,') - D(q,q')\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\left(\log\frac{q(x,a)}{\widetilde{q}\,'(x,a)} - \log\frac{q(x,a)}{q'(x,a)}\right) - \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\big(q'(x,a)-\widetilde{q}\,'(x,a)\big)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\big(\log q'(x,a) - \log\widetilde{q}\,'(x,a)\big)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\left(\log q'(x,a) - \log\left((1-\theta)q'(x,a)+\theta\frac{1}{|X_\ell||A|}\right)\right).
\end{aligned}
$$

By the Jensen's inequality,

$$
\begin{aligned}
&D(q,\widetilde{q}\,') - D(q,q')\\
&\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\left(\log q'(x,a) - (1-\theta)\log q'(x,a) - \theta\log\frac{1}{|X_\ell||A|}\right)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} \theta q(x,a)\big(\log q'(x,a) + \log|X_\ell||A|\big)\\
&\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} \theta q(x,a)\log|X_\ell||A|\\
&\leq \theta L\log|X||A|
\end{aligned}
$$

where the second inequality is due to that a negative entropy is non-positive.

We next prove the second inequality. By the definition,

$$
\begin{aligned}
D(q,\widetilde{q}') &= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\log\frac{q(x,a)}{\widetilde{q}'(x,a)} - \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\big(q(x,a)-\widetilde{q}'(x,a)\big)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\big(\log q(x,a)-\log\widetilde{q}'(x,a)\big)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\left(\log q(x,a)-\log\left((1-\theta)q'(x,a)+\theta\frac{1}{|X_\ell||A|}\right)\right)\\
&\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} -q(x,a)\log\left((1-\theta)q'(x,a)+\theta\frac{1}{|X_\ell||A|}\right)
\end{aligned}
$$

where the last inequality is due to that a negative entropy is non-positive. We note that $-\log(\cdot)$ is a non-increasing function. We can simplify the upper bound on $D(q,\widetilde{q}')$ above by,

$$
\begin{aligned}
D(q,\widetilde{q}') &\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} -q(x,a)\log\left(\theta\frac{1}{|X_\ell||A|}\right)\\
&= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\log\frac{|X_\ell||A|}{\theta}\\
&\leq \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A} q(x,a)\log\frac{|X||A|}{\theta}\\
&= L\log\frac{|X||A|}{\theta}.
\end{aligned}
$$

■

We now are ready to prove Lemma 5.

Proof [Proof of Lemma 5] By Lemma 1, with probability $1-\delta$ it holds that

$$
\Delta(P_1)\ \subset\ \cap_{t=0}^{T-1}\Delta(k_1^t)\ \text{ and }\ \Delta(P_2)\ \subset\ \cap_{t=0}^{T-1}\Delta(k_2^t).
$$

We note that the solution $(q_1^\star,q_2^\star)$ in hindsight to Problem (4) satisfies $q_1^\star\in\Delta(P_1)$ and $q_2^\star\in\Delta(P_2)$. Hence, $q_1^\star\in\cap_{t=0}^{T-1}\Delta(k_1^t)$ and $q_2^\star\in\Delta(P_2)\cap_{t=0}^{T-1}\Delta(k_2^t)$ with probability $1-\delta$. For episode t, we apply Lemma 13 to the primal update (8) with

$$
f(x,y)|_{x=q_1,\,y=q_2}\ =\ V\left\langle q_1\cdot\widehat{q}_2^{t-1}+\widehat{q}_1^{t-1}\cdot q_2, r^{t-1}\right\rangle\ +\ \lambda^{t-1}\langle q_1,g^{t-1}\rangle\ -\ \lambda^{t-1}\langle q_2,h^{t-1}\rangle
$$

and $x^\star=\widehat{q}_1^t$, $y^\star=\widehat{q}_2^t$, $x'=\widetilde{q}_1^{t-1}$, $y'=\widetilde{q}_2^{t-1}$, $x=q_1^\star$, and $y=q_2^\star$. Thus, with probability $1-\delta$ it holds for any t that

$$
\begin{aligned}
&V\left\langle \widehat{q}_1^t\cdot\widehat{q}_2^{t-1}+\widehat{q}_1^{t-1}\cdot q_2^\star, r^{t-1}\right\rangle\ +\ \lambda^{t-1}\langle\widehat{q}_1^t,g^{t-1}\rangle\ -\ \lambda^{t-1}\langle q_2^\star,h^{t-1}\rangle\\
&+\eta^{-1}\big(D(\widehat{q}_1^t,\widetilde{q}_1^{t-1})+D(\widehat{q}_2^t,\widetilde{q}_2^{t-1})\big)\\
\leq\ & V\left\langle q_1^\star\cdot\widehat{q}_2^{t-1}+\widehat{q}_1^{t-1}\cdot\widehat{q}_2^t, r^{t-1}\right\rangle\ +\ \lambda^{t-1}\langle q_1^\star,g^{t-1}\rangle\ -\ \lambda^{t-1}\langle\widehat{q}_2^t,h^{t-1}\rangle\\
&\quad+\eta^{-1}\big(D(q_1^\star,\widetilde{q}_1^{t-1})\ +\ D(q_2^\star,\widetilde{q}_2^{t-1})\ -\ D(q_1^\star,\widehat{q}_1^t)\ -\ D(q_2^\star,\widehat{q}_2^t)\big)
\end{aligned}
$$

or, equivalently,

$$\begin{aligned} & V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1} \right\rangle + \lambda^{t-1} \langle \widehat{q}_1^t, g^{t-1} \rangle + \lambda^{t-1} \langle \widehat{q}_2^t, h^{t-1} \rangle \\ & + \eta^{-1} \big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) \big) \\ \leq \;\; & V\left\langle q_1^\star \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1} \right\rangle + \lambda^{t-1} \langle q_1^\star, g^{t-1} \rangle + \lambda^{t-1} \langle q_2^\star, h^{t-1} \rangle \\ & \quad + \eta^{-1} \big(D(q_1^\star, \widetilde{q}_1^{t-1}) + D(q_2^\star, \widetilde{q}_2^{t-1}) - D(q_1^\star, \widehat{q}_1^t) - D(q_2^\star, \widehat{q}_2^t) \big). \end{aligned} \tag{31}$$

Let $\Delta^t := \frac{1}{2}\left((\lambda^t)^2 - (\lambda^{t-1})^2\right)$ be the drift of the consecutive dual updates. Then,

$$\begin{aligned} \Delta^t \;&=\; \frac{1}{2}\left((\lambda^t)^2 - (\lambda^{t-1})^2\right) \\ &=\; \frac{1}{2}\left(\max{}^2\Big(\lambda^{t-1} + \big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big),\; 0\Big) - (\lambda^{t-1})^2\right) \\ &\leq\; \lambda^{t-1}\big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big) + \frac{1}{2}\big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big)^2 \\ &\leq\; \lambda^{t-1}\big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big) + 2L^2 \end{aligned} \tag{32}$$

where the first inequality is due to $\max^2(x,0) \leq x^2$ and we apply $\langle \widehat{q}_1^t, g^{t-1}\rangle$, $\langle \widehat{q}_2^t, h^{t-1}\rangle$, $b \in [0, L]$ in the last inequality. Adding (32) to (31) from both sides of the inequalities without changing the inequality direction yields

$$\begin{aligned} & V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1} \right\rangle + \Delta^t + \eta^{-1} \big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) \big) \\ \leq \;\; & V\left\langle q_1^\star \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1} \right\rangle + \lambda^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) + 2L^2 \\ & + \eta^{-1} \big(D(q_1^\star, \widetilde{q}_1^{t-1}) + D(q_2^\star, \widetilde{q}_2^{t-1}) - D(q_1^\star, \widehat{q}_1^t) - D(q_2^\star, \widehat{q}_2^t) \big). \end{aligned} \tag{33}$$

However,

$$\begin{aligned} & V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1} \right\rangle + \eta^{-1} \big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) \big) \\ = \;\; & V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widetilde{q}_1^{t-1} \cdot \widehat{q}_2^{t-1}, r^{t-1} \right\rangle + V\left\langle \widetilde{q}_1^{t-1} \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^{t-1}, r^{t-1} \right\rangle \\ & + V\left\langle \widehat{q}_1^{t-1} \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widetilde{q}_2^{t-1}, r^{t-1} \right\rangle + V\left\langle \widehat{q}_1^{t-1} \cdot \widetilde{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1} \right\rangle \\ & + \eta^{-1} D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + \eta^{-1} D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) \\ \geq \;\; & -V \left\| \widehat{q}_2^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1 - V \left\| \widehat{q}_2^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widetilde{q}_1^{t-1} - \widehat{q}_1^{t-1} \right\|_1 \\ & - V \left\| \widehat{q}_1^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widehat{q}_2^{t-1} - \widetilde{q}_2^{t-1} \right\|_1 - V \left\| \widehat{q}_1^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widetilde{q}_2^{t-1} - \widehat{q}_2^t \right\|_1 \\ & + (2\eta L)^{-1} \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1^2 + (2\eta L)^{-1} \left\| \widehat{q}_2^t - \widetilde{q}_2^{t-1} \right\|_1 \\ \geq \;\; & -V \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1 - 2\theta V L + (2\eta L)^{-1} \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1^2 \\ & - 2\theta V L - V \left\| \widetilde{q}_2^{t-1} - \widehat{q}_2^t \right\|_1 + (2\eta L)^{-1} \left\| \widehat{q}_2^t - \widetilde{q}_2^{t-1} \right\|_1 \\ \geq \;\; & -4\theta V L - \eta V^2 L \end{aligned}$$

where we apply the Hölder's inequality and Lemma 14 in the first inequality, the second inequality is due to that

$$
\begin{aligned}
\|\widetilde{q}_1^{t-1} - \widehat{q}_1^{t-1}\|_1 &= \sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\left|(1-\theta)\widehat{q}_1^{t-1}(x,a) + \theta\frac{1}{|X_\ell||A|} - \widehat{q}_1^{t-1}(x,a)\right| \\
&\leq \theta\sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\widehat{q}_1^{t-1}(x,a) + \theta\sum_{\ell=0}^{L-1}\sum_{x\in X_\ell}\sum_{a\in A}\frac{1}{|X_\ell||A|} \\
&= 2\theta L
\end{aligned}
$$

and $\|\widetilde{q}_2^{t-1} - \widehat{q}_2^{t-1}\|_1 \leq 2\theta L$ that can be proved similarly, and the last inequality is due to $-bx + ax^2 \geq -b^2/(4a)$ for any $a, b > 0$. Therefore, we take the lower bound above for the left-hand side of (33),

$$
\begin{aligned}
&\Delta^t \,-\, 4\theta VL \,-\, \eta V^2 L \\
&\leq V\left\langle q_1^\star\cdot\widehat{q}_2^{t-1} - \widehat{q}_1^{t-1}\cdot q_2^\star, r^{t-1}\right\rangle \,+\, \lambda^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) \,+\, 2L^2 \\
&\quad + \eta^{-1}\big(D(q_1^\star,\widetilde{q}_1^{t-1}) + D(q_2^\star,\widetilde{q}_2^{t-1}) - D(q_1^\star,\widehat{q}_1^{t}) - D(q_2^\star,\widehat{q}_2^{t})\big).
\end{aligned} \tag{34}
$$

By Lemma 15,

$$
\begin{aligned}
D(q_1^\star,\widetilde{q}_1^{t-1}) - D(q_1^\star,\widehat{q}_1^{t}) &= D(q_1^\star,\widetilde{q}_1^{t-1}) - D(q_1^\star,\widehat{q}_1^{t-1}) + D(q_1^\star,\widehat{q}_1^{t-1}) - D(q_1^\star,\widehat{q}_1^{t}) \\
&\leq \theta L\log(|X||A|) + D(q_1^\star,\widehat{q}_1^{t-1}) - D(q_1^\star,\widehat{q}_1^{t})
\end{aligned}
$$

and, similarly,

$$
D(q_2^\star,\widetilde{q}_2^{t-1}) - D(q_2^\star,\widehat{q}_2^{t}) \;\leq\; \theta L\log(|Y||B|) + D(q_2^\star,\widehat{q}_2^{t-1}) - D(q_2^\star,\widehat{q}_2^{t}).
$$

We now simplify (34) into

$$
\begin{aligned}
\Delta^t \;\leq\; & V\left\langle q_1^\star\cdot\widehat{q}_2^{t-1} - \widehat{q}_1^{t-1}\cdot q_2^\star, r^{t-1}\right\rangle \,+\, \lambda^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) \\
& + \eta^{-1}\big(D(q_1^\star,\widehat{q}_1^{t-1}) + D(q_2^\star,\widehat{q}_2^{t-1}) - D(q_1^\star,\widehat{q}_1^{t}) - D(q_2^\star,\widehat{q}_2^{t})\big) \\
& + \eta^{-1}\theta L\big(\log(|X||A|) + \log(|Y||B|)\big) \,+\, 2L^2 \,+\, 4\theta VL \,+\, \eta V^2 L
\end{aligned}
$$

which leads to the desired result by summing it up from $t = 1$ to T,

$$
\begin{aligned}
\sum_{t=1}^T \Delta^t \;\leq\; & V\sum_{t=1}^T\left\langle q_1^\star\cdot\widehat{q}_2^{t-1} - \widehat{q}_1^{t-1}\cdot q_2^\star, r^{t-1}\right\rangle \,+\, \sum_{t=1}^T\lambda^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) \\
& + \eta^{-1}\sum_{t=1}^T\big(D(q_1^\star,\widehat{q}_1^{t-1}) + D(q_2^\star,\widehat{q}_2^{t-1}) - D(q_1^\star,\widehat{q}_1^{t}) - D(q_2^\star,\widehat{q}_2^{t})\big) \\
& + \eta^{-1}\theta LT\big(\log(|X||A|) + \log(|Y||B|)\big) \,+\, 2L^2T \,+\, 4\theta VLT \,+\, \eta V^2 LT \\
\leq\; & V\sum_{t=1}^T\left\langle q_1^\star\cdot\widehat{q}_2^{t-1} - \widehat{q}_1^{t-1}\cdot q_2^\star, r^{t-1}\right\rangle \,+\, \sum_{t=1}^T\lambda^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) \\
& + \eta^{-1}\big(D(q_1^\star,\widehat{q}_1^{0}) + D(q_2^\star,\widehat{q}_2^{0})\big) \\
& + \eta^{-1}\theta LT\big(\log(|X||A|) + \log(|Y||B|)\big) \,+\, 2L^2T \,+\, 4\theta VLT \,+\, \eta V^2 LT
\end{aligned}
$$

which leads to the desired result by noting that

$$D(q_1^\star, \widehat{q}_1^{\,0}) \;\leq\; L\log(|X||A|),\; D(q_2^\star, \widehat{q}_2^{\,0}) \;\leq\; L\log(|Y||B|),\;\; \text{and} \;\; \sum_{t=1}^{T} \Delta^t \;\geq\; 0.$$

■

12. Proofs of Lemma 6 and Lemma 7

We first present the boundedness of the dual update λ^t in Lemma 6. Our proof is based on a new drift analysis in Lemma 22 that has been established in Yu et al. (2017) for providing a high probability bound for stochastic processes.

Proof [Proof of Lemma 6] Let $\mathcal{F}^t$ be an σ-algebra that is generated by the state-action sequence, reward/utility functions for both players up to episode t. At the beginning, $\mathcal{F}^0 = \{\emptyset, \Omega\}$. We have a discrete-time random process $\{\lambda^t, t \geq 0\}$ that adapts to $\mathcal{F}^t$. It suffices to check all assumptions in Lemma 22.

By the dual update (10),

$$\begin{aligned} \left|\lambda^{t+1} - \lambda^t\right| \;&=\; \left|\max\Big(\lambda^t \,+\, \big(\langle \widehat{q}_1^{\,t+1}, g^t\rangle + \langle \widehat{q}_2^{\,t+1}, h^t\rangle - b\big),\; 0\Big) - \lambda^t\right| \\ &\leq\; \left|\langle \widehat{q}_1^{\,t+1}, g^t\rangle + \langle \widehat{q}_2^{\,t+1}, h^t\rangle - b\right| \\ &\leq\; 2L \end{aligned}$$

where the first inequality is clear from two cases for $\max(\cdot)$ and the second inequality is due to $\langle \widehat{q}_1^{\,t+1}, g^t\rangle$, $\langle \widehat{q}_2^{\,t+1}, h^t\rangle \in [0, L]$, $b \in [0, 2L]$. Consequently,

$$\lambda^{t+t_0} - \lambda^t \;=\; \sum_{s=t}^{t+t_0-1} \big(\lambda^{s+1} - \lambda^s\big) \;\leq\; \sum_{s=t}^{t+t_0-1} \left|\lambda^{s+1} - \lambda^s\right| \;\leq\; 2t_0 L \tag{35}$$

which leads to $\mathbb{E}[\, \lambda^{t+t_0} - \lambda^t \,|\, \mathcal{F}^t \,] \leq 2t_0 L$. It is convenient to take $\delta_{\max} = 2L$ in Lemma 22.

We next determine the validity of other assumptions in Lemma 22. Let us denote the event in Lemma 1 by $\mathcal{E}_{\text{good}}$ and we have $P(\mathcal{E}_{\text{good}}) \geq 1 - \delta$. We recall that the proof of Lemma 5 remains to be valid if we replace $q_1^\star$ by $\bar{q}_1$ and $q_2^\star$ by $\bar{q}_2$ starting from (31). By doing so, it is ready to obtain a similar result as (34): under the good event $\mathcal{E}_{\text{good}}$ it holds for any τ that

$$\begin{aligned} &\Delta^\tau \,-\, 4\theta V L \,-\, \eta V^2 L \\ &\leq\; V\left\langle \bar{q}_1 \cdot \widehat{q}_2^{\,\tau-1} - \widehat{q}_1^{\,\tau-1} \cdot \bar{q}_2, r^{\tau-1}\right\rangle \,+\, \lambda^{\tau-1}\big(\langle \bar{q}_1, g^{\tau-1}\rangle + \langle \bar{q}_2, h^{\tau-1}\rangle - b\big) \,+\, 2L^2 \\ &\qquad + \eta^{-1}\big(D(\bar{q}_1, \widetilde{q}_1^{\,\tau-1}) + D(\bar{q}_2, \widetilde{q}_2^{\,\tau-1}) - D(\bar{q}_1, \widehat{q}_1^{\,\tau}) - D(\bar{q}_2, \widehat{q}_2^{\,\tau})\big) \end{aligned}$$

or, equivalently,

$$\begin{aligned} &(\lambda^\tau)^2 \,-\, (\lambda^{\tau-1})^2 \\ &\leq\; 2V\left\langle \bar{q}_1 \cdot \widehat{q}_2^{\,\tau-1} - \widehat{q}_1^{\,\tau-1} \cdot \bar{q}_2, r^{\tau-1}\right\rangle \,+\, 2\lambda^{\tau-1}\big(\langle \bar{q}_1, g^{\tau-1}\rangle + \langle \bar{q}_2, h^{\tau-1}\rangle - b\big) \,+\, 4L^2 \\ &\qquad + 2\eta^{-1}\big(D(\bar{q}_1, \widetilde{q}_1^{\,\tau-1}) + D(\bar{q}_2, \widetilde{q}_2^{\,\tau-1}) - D(\bar{q}_1, \widehat{q}_1^{\,\tau}) - D(\bar{q}_2, \widehat{q}_2^{\,\tau})\big) \,+\, 8\theta V L \,+\, 2\eta V^2 L. \end{aligned} \tag{36}$$

We note that $|\langle \bar{q}_1 \cdot \widehat{q}_2^\tau - \widehat{q}_1^\tau \cdot \bar{q}_2, r^\tau \rangle| \leq L$. By summing both sides of (36) from $\tau = t+1$ to $\tau = t + t_0$,

$$
\begin{aligned}
(\lambda^{t+t_0})^2 \,-\, (\lambda^t)^2 \;\;\leq\;\; & 2t_0 V L \,+\, \sum_{\tau=t}^{t+t_0-1} 2\lambda^\tau\big(\langle \bar{q}_1, g^\tau\rangle + \langle \bar{q}_2, h^\tau\rangle - b\big) \,+\, 4t_0 L^2 \\
& + 2\eta^{-1}\big(D(\bar{q}_1, \widetilde{q}_1^t) + D(\bar{q}_2, \widetilde{q}_2^t)\big) \,+\, 8t_0\theta V L \,+\, 2t_0\eta V^2 L
\end{aligned}
$$

where we omit two non-positive terms. Taking the conditional expectation given $\mathcal{F}^t$ and $\mathcal{E}_{\text{good}}$ yields,

$$
\begin{aligned}
& \mathbb{E}\left[(\lambda^{t+t_0})^2 \,-\, (\lambda^t)^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
\leq\;\; & 2t_0 V L \,+\, \sum_{\tau=t}^{t+t_0-1} 2\mathbb{E}\left[\lambda^\tau\big(\langle \bar{q}_1, g^\tau\rangle + \langle \bar{q}_2, h^\tau\rangle - b\big) \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,+\, 4t_0 L^2 \\
& + 2\eta^{-1}\mathbb{E}\left[D(\bar{q}_1, \widetilde{q}_1^t) + D(\bar{q}_2, \widetilde{q}_2^t) \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,+\, 8t_0\theta V L \,+\, 2t_0\eta V^2 L \\
\leq\;\; & 2t_0 V L \,-\, 2\xi \sum_{\tau=t}^{t+t_0-1} \mathbb{E}\left[\lambda^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,+\, 4t_0 L^2 \\
& + 2\eta^{-1} L (\log(|X||A|/\theta) + \log(|Y||B|/\theta)) \,+\, 8t_0\theta V L \,+\, 2t_0\eta V^2 L \\
\leq\;\; & 2t_0 V L \,-\, 2\xi t_0 \mathbb{E}\left[\lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,+\, 2\xi t_0(t_0-1)L \,+\, 4t_0 L^2 \\
& + 2\eta^{-1} L (\log(|X||A|/\theta) + \log(|Y||B|/\theta)) \,+\, 8t_0\theta V L \,+\, 2t_0\eta V^2 L
\end{aligned}
\tag{37}
$$

where the second inequality is due to Lemma 15 and the fact: by the law of total expectation, for any $\tau \geq t$, $\mathcal{F}^t \subset \mathcal{F}^\tau$ and

$$
\begin{aligned}
\mathbb{E}\left[\lambda^\tau\big(\langle \bar{q}_1, g^\tau\rangle + \langle \bar{q}_2, h^\tau\rangle - b\big) \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;&=\; \mathbb{E}\left[\mathbb{E}\left[\lambda^\tau\big(\langle \bar{q}_1, g^\tau\rangle + \langle \bar{q}_2, h^\tau\rangle - b\big) \,|\, \mathcal{F}^\tau\right] \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
&=\; \mathbb{E}\left[\lambda^\tau \mathbb{E}\left[\langle \bar{q}_1, g^\tau\rangle + \langle \bar{q}_2, h^\tau\rangle - b\right] \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
&=\; \mathbb{E}\left[\langle \bar{q}_1, g^\tau\rangle + \langle \bar{q}_2, h^\tau\rangle - b\right] \mathbb{E}\left[\lambda^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
&\leq\; -\xi\, \mathbb{E}\left[\lambda^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]
\end{aligned}
$$

where the inequality is due to the strict feasibility assumption on $(\bar{q}_1, \bar{q}_2)$; the last inequality is due to that

$$
\sum_{\tau=t}^{t+t_0-1} \mathbb{E}\left[\lambda^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;\geq\; \sum_{\tau=t}^{t+t_0-1} \mathbb{E}\left[\lambda^t - 2(\tau - t)L \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;=\; \sum_{\tau=0}^{t_0-1} \mathbb{E}\left[\lambda^t - 2\tau L \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]
$$

which follows the fact $\lambda^\tau \geq \lambda^t - 2(\tau - t)L$ for any $\tau \geq t \geq 0$ if we note that $|\lambda^{t+1} - \lambda^t| \leq 2L$. Hence, we can simplify (37) as

$$\begin{aligned}
& \mathbb{E}\left[(\lambda^{t+t_0})^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
\leq\ & \mathbb{E}\left[(\lambda^t)^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,-\, 2\xi t_0 \mathbb{E}\left[\lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,+\, 2\xi t_0^2 L \,+\, 4t_0 L^2 \,+\, 2t_0 VL \\
& + 2\eta^{-1} L(\log(|X||A|/\theta) + \log(|Y||B|/\theta)) \,+\, 8t_0\theta VL \,+\, 2t_0\eta V^2 L \\
\leq\ & \mathbb{E}\left[(\lambda^t)^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,-\, \xi t_0 \mathbb{E}\left[\lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,-\, \xi t_0 \Theta \,+\, 2\xi t_0^2 L \,+\, 4t_0 L^2 \,+\, 2t_0 VL \\
& + 2\eta^{-1} L(\log(|X||A|/\theta) + \log(|Y||B|/\theta)) \,+\, 8t_0\theta VL \,+\, 2t_0\eta V^2 L \\
=\ & \mathbb{E}\left[(\lambda^t)^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,-\, \xi t_0 \mathbb{E}\left[\lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,-\, \frac{1}{2}\xi^2 t_0^2 \\
\leq\ & \left(\mathbb{E}\left[\lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \frac{1}{2}\xi t_0\right)^2
\end{aligned}$$

where we apply $\lambda^t \geq \Theta$ for the second inequality and we take Θ in Lemma 22,

$$\Theta \;=\; \frac{1}{2}\xi t_0 + 2t_0 L \,+\, \frac{4L^2 + 8\theta VL + 2\eta V^2 L + 2VL}{\xi} \,+\, \frac{2L(\log(|X||A|/\theta) + \log(|Y||B|/\theta))}{t_0 \xi \eta}.$$

Taking the square root and applying the Jensen's inequality yield

$$\mathbb{E}\left[\lambda^{t+t_0} \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;\leq\; \sqrt{\mathbb{E}\left[(\lambda^{t+t_0})^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]} \;\leq\; \mathbb{E}\left[\lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \frac{1}{2}\xi t_0$$

which shows that $\mathbb{E}\left[\lambda^{t+t_0} - \lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \leq -\frac{1}{2}\xi t_0$. Application of law of total expectation to this inequality and (35) with $\delta < \frac{1}{12}$ yields

$$\begin{aligned}
\mathbb{E}\left[\lambda^{t+t_0} - \lambda^t \,|\, \mathcal{F}^t\right] \;&=\; P(\mathcal{E}_{\text{good}})\mathbb{E}\left[\lambda^{t+t_0} - \lambda^t \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \,+\, P(\bar{\mathcal{E}}_{\text{good}})\mathbb{E}\left[\lambda^{t+t_0} - \lambda^t \,|\, \mathcal{F}^t, \bar{\mathcal{E}}_{\text{good}}\right] \\
&\leq\; -\frac{1}{2}\xi t_0 \times (1-\delta) \,+\, 2t_0 L \times \delta \\
&\leq\; -\frac{1}{4}\xi t_0
\end{aligned}$$

which verifies the assumption of Lemma 22 if we take $\zeta = \xi/4$.

We now have verified all assumptions of Lemma 22 with appropriate parameters Θ, $\delta_{\max}$, ζ. For episode t, with probability $1 - \delta$ it holds that

$$\lambda^t \;\;\leq\;\; \Theta + t_0\delta_{\max} + t_0 \frac{4\delta_{\max}^2}{\zeta} \log\left(\frac{8\delta_{\max}^2}{\zeta}\right) + t_0 \frac{4\delta_{\max}^2}{\zeta} \log\frac{1}{\delta}.$$

We complete the proof by taking a union bound over $t = 1, \cdots, T$. ■

With Lemma 6 in place, we are ready to prove Lemma 7.

Proof [Proof of Lemma 7]

Let $Z^t := \sum_{\tau=0}^{t-1} \lambda^\tau \big(\langle q_1^\star, g^\tau\rangle + \langle q_2^\star, h^\tau\rangle - b\big)$. We note that

$$
\begin{aligned}
&\mathbb{E}\left[Z^t \,|\, \mathcal{F}^{t-1}\right] \\
=\;& \mathbb{E}\left[\sum_{\tau=0}^{t-1} \lambda^\tau \big(\langle q_1^\star, g^\tau\rangle + \langle q_2^\star, h^\tau\rangle - b\big) \,\middle|\, \mathcal{F}^{t-1}\right] \\
=\;& \mathbb{E}\left[\sum_{\tau=0}^{t-2} \lambda^\tau \big(\langle q_1^\star, g^\tau\rangle + \langle q_2^\star, h^\tau\rangle - b\big) \,\middle|\, \mathcal{F}^{t-1}\right] + \lambda^{t-1}\mathbb{E}\left[\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) \,|\, \mathcal{F}^{t-1}\right] \\
\leq\;& \mathbb{E}\left[\sum_{\tau=0}^{t-2} \lambda^\tau \big(\langle q_1^\star, g^\tau\rangle + \langle q_2^\star, h^\tau\rangle - b\big) \,\middle|\, \mathcal{F}^{t-1}\right] \\
=\;& \mathbb{E}\left[Z^{t-1}\right]
\end{aligned}
$$

where the inequality is because of $\mathbb{E}\left[\big(\langle q_1^\star, g^{t-1}\rangle + \langle q_2^\star, h^{t-1}\rangle - b\big) \,|\, \mathcal{F}^{t-1}\right] = \langle q_1^\star, g\rangle + \langle q_2^\star, h\rangle - b \leq 0$. Hence, $\{Z^t, t \geq 0\}$ a supermartingale.

We also note that $\left|Z^{t+1} - Z^t\right| = \lambda^t \left|\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\right| \leq 2\lambda^t L$. Thus, if $|Z^{t+1} - Z^t| > c$ for some $c \in \mathbb{R}^+$, then $\lambda^t > c/(2L)$. Let $Y^t := \lambda^t - c/(2L)$. Therefore,

$$\{|Z^{t+1} - Z^t| > c\} \;\subset\; \{Y^t > 0\}.$$

By Lemma 23,

$$P\left(\sum_{t=0}^{T-1} \lambda^t \big(\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\big) \geq z\right) \;\leq\; \mathrm{e}^{-z^2/(2c^2T)} \;+\; \sum_{\tau=0}^{T-1} P\left(\lambda^t > \frac{c}{2L}\right). \tag{38}$$

By Lemma 6, with probability $1-\delta$ it holds for any t that

$$\lambda^t \;\leq\; \Theta + 2t_0 L + t_0 \frac{64L^2}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{64L^2}{\xi} \log\frac{1}{\delta}$$

or, equivalently,

$$P\left(\lambda^t \;\geq\; \Theta + 2t_0 L + t_0 \frac{64L^2}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{64L^2}{\xi} \log\frac{1}{\delta}\right) \;\leq\; \delta.$$

If we take

$$c \;=\; 2\Theta L + 4t_0 L^2 + t_0 \frac{128L^3}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{128L^3}{\xi} \log\frac{1}{\delta} \;\text{ and }\; z \;=\; \sqrt{2Tc^2 \log(1/(\delta T))}$$

then (38) becomes

$$P\left(\sum_{t=0}^{T-1} \lambda^t \big(\langle q_1^\star, g^t\rangle + \langle q_2^\star, h^t\rangle - b\big) \geq z\right) \;\leq\; 2\delta T$$

which proves the desired result. ■

13. Proof of Theorem 10

By the dual update (10),

$$
\begin{aligned}
\lambda^t &= \max\Big(\lambda^{t-1} + \big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big),\ 0\Big) \\
&\geq \lambda^{t-1} + \big(\langle \widehat{q}_1^t, g^{t-1}\rangle + \langle \widehat{q}_2^t, h^{t-1}\rangle - b\big) \\
&= \lambda^{t-1} + \big(\langle \widehat{q}_1^{t-1}, g^{t-1}\rangle + \langle \widehat{q}_2^{t-1}, h^{t-1}\rangle - b\big) + \langle \widehat{q}_1^t - \widehat{q}_1^{t-1}, g^{t-1}\rangle + \langle \widehat{q}_2^t - \widehat{q}_2^{t-1}, h^{t-1}\rangle \\
&\geq \lambda^{t-1} + \big(\langle \widehat{q}_1^{t-1}, g^{t-1}\rangle + \langle \widehat{q}_2^{t-1}, h^{t-1}\rangle - b\big) - \big\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\big\|_1 - \big\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\big\|_1
\end{aligned}
\tag{39}
$$

where the last inequality is due to: $\langle \widehat{q}_1^t - \widehat{q}_1^{t-1}, g^{t-1}\rangle \leq \|\widehat{q}_1^t - \widehat{q}_1^{t-1}\|_1 \|g^{t-1}\|_\infty$, $\langle \widehat{q}_2^t - \widehat{q}_2^{t-1}, h^{t-1}\rangle \leq \|\widehat{q}_2^t - \widehat{q}_2^{t-1}\|_1 \|h^{t-1}\|_\infty$, and $\|g^{t-1}\|_\infty, \|h^{t-1}\|_\infty \in [0,1]$. We note that $\lambda^0 = 0$ from the initialization. Summing up both sides of (39) from $t=1$ to $t=T$ leads to

$$
\sum_{t=0}^{T-1} \big(\langle \widehat{q}_1^t, g^t\rangle + \langle \widehat{q}_2^t, h^t\rangle - b\big) \ \leq\ \lambda^T + \sum_{t=1}^{T} \big(\big\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\big\|_1 + \big\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\big\|_1\big). \tag{40}
$$

We recall $\widehat{q}_1^t \in \Delta(k_1^t)$, $\widehat{q}_2^t \in \Delta(k_2^t)$ in the primal update (8) and $\Delta(k_1^t)$ and $\Delta(k_2^t)$ in the confidence sets (11). To bound $\big\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\big\|_1 + \big\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\big\|_1$, we consider two cases: (i) $k_1^t = k_1^{t-1}$ and $k_2^t = k_2^{t-1}$; (ii) either $k_1^t \neq k_1^{t-1}$ or $k_2^t \neq k_2^{t-1}$.

Case (i). In this case, we have: $\widehat{q}_1^t, \widehat{q}_1^{t-1} \in \Delta(k_1^t)$, $\widehat{q}_2^t, \widehat{q}_2^{t-1} \in \Delta(k_2^t)$. We begin with the primal update (8) and apply Lemma 13 with,

$$
f(x,y)|_{x=q_1,\, y=q_2} \ =\ V\left\langle q_1 \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1}\cdot q_2, r^{t-1}\right\rangle + \lambda^{t-1}\langle q_1, g^{t-1}\rangle - \lambda^{t-1}\langle q_2, h^{t-1}\rangle
$$

and $x^\star = \widehat{q}_1^t$, $y^\star = \widehat{q}_2^t$, $x' = \widetilde{q}_1^{t-1}$, $y' = \widetilde{q}_2^{t-1}$, $x = \widetilde{q}_1^{t-1}$, and $y = \widetilde{q}_2^{t-1}$. Thus,

$$
\begin{aligned}
& V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1}\cdot \widetilde{q}_2^{t-1}, r^{t-1}\right\rangle + \lambda^{t-1}\langle \widehat{q}_1^t, g^{t-1}\rangle - \lambda^{t-1}\langle \widetilde{q}_2^{t-1}, h^{t-1}\rangle \\
&\quad + \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \\
\leq\ & V\left\langle \widetilde{q}_1^{t-1} \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1}\cdot \widehat{q}_2^t, r^{t-1}\right\rangle + \lambda^{t-1}\langle \widetilde{q}_1^{t-1}, g^{t-1}\rangle - \lambda^{t-1}\langle \widehat{q}_2^t, h^{t-1}\rangle \\
&\quad - \eta^{-1}\big(D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t)\big).
\end{aligned}
$$

or, equivalently,

$$
\begin{aligned}
& \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) + \eta^{-1}\big(D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t)\big) \\
\leq\ & V\left\langle (\widetilde{q}_1^{t-1} - \widehat{q}_1^t)\cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1}\cdot(\widehat{q}_2^t - \widetilde{q}_2^{t-1}), r^{t-1}\right\rangle \\
&\quad + \lambda^{t-1}\langle \widetilde{q}_1^{t-1} - \widehat{q}_1^t, g^{t-1}\rangle + \lambda^{t-1}\langle \widetilde{q}_2^{t-1} - \widehat{q}_2^t, h^{t-1}\rangle.
\end{aligned}
\tag{41}
$$

We note that $\langle (\widetilde{q}_1^{t-1} - \widehat{q}_1^t)\cdot \widehat{q}_2^{t-1}, r^{t-1}\rangle \leq \|(\widetilde{q}_1^{t-1} - \widehat{q}_1^t)\cdot \widehat{q}_2^{t-1}\|_1 \|r^{t-1}\|_\infty \leq \|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\|_1$, and, similarly, $\langle \widehat{q}_1^{t-1}\cdot(\widehat{q}_2^t - \widetilde{q}_2^{t-1}), r^{t-1}\rangle \leq \|\widehat{q}_2^t - \widetilde{q}_2^{t-1}\|_1$. Thus, we can reduce (41) into

$$
\begin{aligned}
& \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) + \eta^{-1}\big(D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t)\big) \\
\leq\ & (V + \lambda^{t-1})\big(\big\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\big\|_1 + \big\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\big\|_1\big)
\end{aligned}
$$

where the left-hand side can be lower bounded by Lemma 14,

$$
D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) \ \geq\ L^{-1}\big\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\big\|_1^2
$$

$$D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t) \;\geq\; L^{-1} \left\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\right\|_1^2.$$

Then, we apply the inequality $(x+y)^2 \leq 2(x^2+y^2)$ and cancel a non-negative term to obtain

$$\left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 + \left\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\right\|_1 \;\leq\; 2\eta L(V + \lambda^{t-1}). \tag{42}$$

By the definition of $\widetilde{q}_1^{t-1}$ and $\widetilde{q}_2^{t-1}$,

$$\begin{aligned}
\left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 &= \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left|(1-\theta)\widehat{q}_1^{t-1}(x,a) + \theta \frac{1}{|X_\ell||A|} - \widehat{q}_1^t(x,a)\right| \\
&\geq \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left((1-\theta)\left|\widehat{q}_1^{t-1}(x,a) - \widehat{q}_1^t(x,a)\right| - \theta\left(\frac{1}{|X_\ell||A|} + \widehat{q}_1^t(x,a)\right)\right) \\
&= (1-\theta)\left\|\widehat{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 - 2\theta L.
\end{aligned}$$

Similarly, we have $\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\|_1 \leq (1-\theta)\|\widehat{q}_2^{t-1} - \widehat{q}_2^t\|_1 - 2\theta L$. Thus, we can further reduce (42) into

$$\left\|\widehat{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 + \|\widehat{q}_2^{t-1} - \widehat{q}_2^t\|_1 \;\leq\; 2\eta(1-\theta)^{-1}L(V + \lambda^{t-1}) + 4\theta(1-\theta)^{-1}L. \tag{43}$$

Case (ii). In this case, either $\widehat{q}_1^t$, $\widehat{q}_1^{t-1}$ or $\widehat{q}_2^t$, $\widehat{q}_2^{t-1}$ might not have the same domain. For instance, when $k_1^t > k_1^{t-1}$, it is possible that $\Delta(k_1^t)$ becomes different from $\Delta(k_1^{t-1})$. We note that $k_1^t > k_1^{t-1}$ only happens when episode t is the first one that belongs to epoch k_1^t. By Lemma 25, $k_1^T \leq \sqrt{T|X||A|}\log(8T/(|X||A|))$ and $k_2^T \leq \sqrt{T|Y||B|}\log(8T/(|Y||B|))$ if we are given $T \geq \max(|X||A|, |Y||B|)$.

We now combine two cases above for (40),

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \\
&= \sum_{\substack{1 \leq t \leq T \\ k_1^t = k_1^{k-1} \wedge k_2^t = k_2^{k-1}}} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \\
&\quad + \sum_{\substack{1 \leq t \leq T \\ k_1^t = k_1^{k-1} \vee k_2^t = k_2^{k-1}}} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \\
&\leq \sum_{\substack{1 \leq t \leq T \\ k_1^t = k_1^{k-1} \wedge k_2^t = k_2^{k-1}}} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \;+\; 2L(k_1^T + k_2^T) \\
&\leq 2\eta(1-\theta)^{-1}L\sum_{t=1}^{T}(V + \lambda^{t-1}) + 4\theta(1-\theta)^{-1}LT + 2L(k_1^T + k_2^T)
\end{aligned}$$

where the first inequality is due to: $\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 \leq 2L$ and $\left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1 \leq 2L$, and we apply (43) from the case (i) for the last inequality. Using the bounds on k_1^T, k_2^T in the case (ii), we conclude the desired bound for (40),

$$
\begin{aligned}
& \sum_{t=0}^{T-1} \left(\langle \widehat{q}_1^t, g^t \rangle + \langle \widehat{q}_2^t, h^t \rangle - b \right) \\
\leq \;\; & \lambda^T + \frac{2\eta L}{1-\theta} \sum_{t=1}^{T} \lambda^{t-1} + \frac{2\eta V + 4\theta}{1-\theta} LT \\
& + 2L \left(\sqrt{T|X||A|} \log(8T/(|X||A|)) + \sqrt{T|Y||B|} \log(8T/(|Y||B|)) \right).
\end{aligned}
$$

We complete the proof by noting $\lambda^0 = 0$, $V = L\sqrt{T}$, $\eta = 1/(TL)$, and $\theta = 1/T$.

14. Constrained MGs with Side Constraints

In this section, we present a special case of Problem (4) that is described as a zero-sum MG with side constraint (Singh and Hemachandra, 2014). Having defined episodic MDPs and occupancy measures in Section 2, we can formulate a constrained minimax problem in which the objective function is a sum of the expected total rewards over T episodes and the constraint is on two agent' expected total utilities,

$$
\begin{array}{rl}
\underset{q_1 \in \Delta(P_1)}{\text{minimize}} \; \underset{q_2 \in \Delta(P_2)}{\text{maximize}} & \displaystyle\sum_{t=0}^{T-1} \langle q_1 \cdot q_2, r^t \rangle \\
\text{subject to} & \langle q_1, g \rangle \; \leq \; b_1 \;\; \text{and} \;\; \langle q_2, h \rangle \; \leq \; b_2
\end{array}
\tag{44}
$$

where we take $b_1, b_2 \in (0, L]$ to avoid trivial cases since we note that $\langle q_1, g\rangle$, $\langle q_2, h\rangle \in [0, L]$. The side constraint corresponds to the limited use of budget/resource for each player. It is straightforward to generalize it to account for multiple constraints. When the transitions P_1 and P_2 are known, the occupancy measure sets $\Delta(P_1)$ and $\Delta(P_2)$ define convex polytopes on q_1 and q_2.

Let $(q_1^\star, q_2^\star)$ be a solution to Problem (44) in hindsight. The existence of $(q_1^\star, q_2^\star)$ is well-known under compactness of the constraint sets (Neumann, 1928; Rosen, 1965). Since two constraints are decoupled, it is natural to define the usual Nash equilibrium via two conditions (Altman and Shwartz, 2000; Daskalakis et al., 2021): (i) $\sum_{t=0}^{T-1} \langle q_1^\star \cdot q_2^\star, r^t \rangle \leq \sum_{t=0}^{T-1} \langle q_1 \cdot q_2^\star, r^t \rangle$ for any $q_1 \in \Delta(P_1)$ satisfying $\langle q_1, g \rangle \leq b_1$; (ii) $\sum_{t=0}^{T-1} \langle q_1^\star \cdot q_2, r^t \rangle \leq \sum_{t=0}^{T-1} \langle q_1^\star \cdot q_2^\star, r^t \rangle$ for any $q_2 \in \Delta(P_2)$ satisfying $\langle q_2, h \rangle \leq b_2$. With this solution concept, we define the regret for any algorithm that plays the game for T episodes by

$$
\text{Regret}(T) \;=\; \sum_{t=0}^{T-1} \left(\langle q_1^t \cdot q_2^\star, r^t \rangle - \langle q_1^\star \cdot q_2^t, r^t \rangle \right) \tag{45}
$$

where two players take policies π^t and μ^t in episode t and they define occupancy measures q_1^t and q_2^t under the true transitions P_1 and P_2.

To measure the constraint satisfaction, we introduce the violation as a non-negative part of accumulated constraint violations $\langle q_1^t, g \rangle - b_1$ and $\langle q_2^t, h \rangle - b_2$ over T episodes,

$$
\text{Violation}_1(T) \;=\; \left[\sum_{t=0}^{T-1} \left(\langle q_1^t, g^t \rangle - b_1 \right) \right]_+ \;\; \text{and} \;\; \text{Violation}_2(T) \;=\; \left[\sum_{t=0}^{T-1} \left(\langle q_2^t, h^t \rangle - b_2 \right) \right]_+ . \tag{46}
$$

We next make an assumption that guarantees the existence of constrained Nash equilibrium (Altman and Shwartz, 2000).

Assumption 2 (Feasibility) *There exists a joint policy $(\bar{\pi}, \bar{\mu})$ associated to the occupancy measure $(\bar{q}_1, \bar{q}_2)$ and $\xi > 0$ such that $\langle \bar{q}_1, g\rangle + \xi \leq b_1$ and $\langle \bar{q}_2, h\rangle + \xi \leq b_1$.*

14.1. Algorithm and Performance Guarantees

We now are ready to specialize Algorithm 1 to Problem (44). The only change is to replace the primal-dual update (8) and (10) by the following optimistic primal-dual mirror descent step.

Let us recall that the occupancy measures q_1^t for the min-player and q_2^t for the max-player are defined over the true transitions P_1 and P_2 in episode t. The primal update of our algorithm maintains two occupancy measures $\widehat{q}_1^t$, $\widehat{q}_2^t$ to estimate q_1^t, q_2^t, separately. Although $\widehat{q}_1^t$, $\widehat{q}_2^t$ do not necessarily come from the true transitions P_1, P_2, they propose a min-policy π^t for the min-player and a max-policy μ^t for the max-player given by (7).

We can revise our Lagrangian-based design to update estimates $\widehat{q}_1^t$ and $\widehat{q}_2^t$ as follows. Assume that the transitions P_1 and P_2 are known. We consider a one-episode constrained minimax problem based on reward/utility functions: r^{t-1}, g^{t-1}, h^{t-1}, revealed at the end of episode $t-1$,

$$\begin{array}{rl} \underset{q_1 \in \Delta(P_1)}{\text{minimize}} \; \underset{q_2 \in \Delta(P_2)}{\text{maximize}} & \left\langle q_1 \cdot q_2, r^{t-1} \right\rangle \\ \text{subject to} & \left\langle q_1, g^{t-1} \right\rangle \;\leq\; b_1 \;\text{ and }\; \left\langle q_2, h^{t-1} \right\rangle \;\leq\; b_2 \end{array}$$

where $\Delta(P_1)$ and $\Delta(P_2)$ are sets of valid occupancy measures under P_1 and P_2, respectively. We apply the method of Lagrange multipliers (Bertsekas, 2014) to deal with constraints by formulating a generalized Lagrangian-based function,

$$L^t(q_1, q_2; \lambda_1, \lambda_2) \;\; := \;\; \langle q_1 \cdot q_2, r^{t-1} \rangle \,+\, \lambda_1\big(\langle q_1, g^{t-1}\rangle - b_1\big) \,-\, \lambda_2\big(\langle q_2, h^{t-1}\rangle - b_2\big)$$

where q_1 is the first primal variable for the min-player, q_2 is the second primal variable for the max-player, and λ_1, $\lambda_2 \geq 0$ work as the Lagrange multiplier or the dual variable in penalizing the min-player/max-player via the first/second λ-term. Once we update $\lambda_1 = \lambda_1^{t-1}$ and $\lambda_2 = \lambda_2^{t-1}$ from the last episode, we reach a constrained saddle-point problem,

$$\underset{q_1 \in \Delta(P_1)}{\text{minimize}} \; \underset{q_2 \in \Delta(P_2)}{\text{maximize}} \;\; L^t(q_1, q_2; \lambda_1^{t-1}, \lambda_2^{t-1}).$$

However, it is not feasible to take the domains $\Delta(P_1)$ and $\Delta(P_2)$ since the true transitions P_1 and P_2 are unknown. Instead, we use their optimistic estimates $\Delta(k_1^t)$ and $\Delta(k_2^t)$ in sense that $q_1^t \in \Delta(k_1^t)$ and $q_2^t \in \Delta(k_2^t)$ hold with high probability; see Lemma 1. Denote $\widehat{q}^{\,t} := (\widehat{q}_1^t, \widehat{q}_2^t)$. By the linear approximation of $L^t(q_1, q_2; \lambda^{t-1})$ at the previous iterate (q_1^{t-1}, q_2^{t-1}), we update the primal variable via an online mirror descent step over the optimistic domains of q_1 and q_2,

$$\begin{aligned} \widehat{q}^{\,t} \leftarrow \underset{q_1 \in \Delta(k_1^t)}{\text{argmin}} \; \underset{q_2 \in \Delta(k_2^t)}{\text{argmax}} \Big(& V \left\langle q_1 \cdot \widehat{q}_2^{\,t-1} + \widehat{q}_1^{\,t-1} \cdot q_2, r^{t-1} \right\rangle \\ & + \lambda_1^{t-1} \langle q_1, g^{t-1}\rangle \,-\, \lambda_2^{t-1} \langle q_2, h^{t-1}\rangle \,+\, \eta^{-1} D\big(q \,|\, \widetilde{q}^{\,t-1}\big) \Big) \end{aligned} \tag{47}$$

where V, $\eta > 0$ are some regularization parameters, $D(\cdot\,|\,\cdot)$ is the unnormalized Kullback-Leibler divergence with a slightly abuse in a way that $D(q\,|\,q') := D(q_1\,|\,q_1') - D(q_2\,|\,q_2')$, $\widetilde{q}_1^{\,t-1}$ and $\widetilde{q}_2^{\,t-1}$ are mixing policies given by (9). The unnormalized Kullback-Leibler (KL) divergence between two distributions p, q is defined by $D(p\,|\,q) := \sum_i p_i \ln \frac{p_i}{q_i} - \sum_i (p_i - q_i)$. Moreover, (47) has an efficient update that is similar as the one in Appendix 8.

Once we obtain $\widehat{q}^{\,t}$, we next perform the dual update. We treat two λ-related regularization terms in $L^t(\widehat{q}_1^t, \widehat{q}_2^t; \lambda_1, \lambda_2)$, separately. The dual update works for each player in the usual way by adding up all past constraint violations,

$$\lambda_1^t \;=\; \max\big(\lambda_1^{t-1} + (\langle \widehat{q}_1^t, g^{t-1}\rangle - b_1\,),\, 0\big) \;\;\text{and}\;\; \lambda_2^t \;=\; \max\big(\lambda_2^{t-1} + (\langle \widehat{q}_2^t, h^{t-1}\rangle - b_2\,),\, 0\big). \quad (48)$$

The dual update (48) increases λ_1^{t-1} when $\widehat{q}_1^t$ violates the approximate constraint $\langle q_1, g^{t-1}\rangle \leq b_1$; it is similar for λ_2^{t-1}. Once we replace the primal-dual update (8) and (10) in line 4 of Algorithm 1 by (47) and (48), we obtain a new version of Algorithm 1 for Problem (44).

Similar to Theorem 2, we have the following bounds on the regret and the constraint violation.

Theorem 16 (Regret Bound and Constraint Violation) *Let Assumption 2 hold. Fix $p \in (0,1)$ and $T \geq \max(|X||A|, |B||Y|)$. In Algorithm 1 with the primal-dual update (47) and (48), we set $V = L\sqrt{T}$, $\eta = 1/(TL)$, and $\theta = 1/T$. Then, the regret (5) and the constraint violation (6) satisfy*

$$\begin{aligned}
\text{Regret}(T) &\;\leq\; \widetilde{O}\big((|X|+|Y|)L\sqrt{T(|A|+|B|)}\big) \\
\text{Violation}_1(T), \text{Violation}_2(T) &\;\leq\; \widetilde{O}\big((|X|+|Y|)L\sqrt{T(|A|+|B|)}\big)
\end{aligned}$$

with probability $1-p$, where $\widetilde{O}(\cdot)$ hides factor $\log\frac{1}{p}$.

We analyze Algorithm 1 with the primal-dual update (47) and (48) by following the proof idea in Appendix 7. For completeness, we provide proof details in next two sections.

14.2. Regret Analysis

We recall that our algorithm maintains the occupancy measures $(\widehat{q}_1^t, \widehat{q}_2^t)$ for estimating policies (π^t, μ^t) and Problem (44) defines the comparison solution $(q_1^\star, q_2^\star)$ in hindsight. We decompose the regret (45) as follows,

$$\text{Regret}(T) \;=\; \underbrace{\sum_{t=0}^{T-1} \big\langle \widehat{q}_1^t \cdot q_2^\star - q_1^\star \cdot \widehat{q}_2^t, r^t \big\rangle}_{\widehat{\text{Regret}}(T)} + \underbrace{\sum_{t=0}^{T-1} \big\langle (q_1^t - \widehat{q}_1^t) \cdot q_2^\star, r^t \big\rangle}_{\text{Error}_1} + \underbrace{\sum_{t=0}^{T-1} \big\langle q_1^\star \cdot (\widehat{q}_2^t - q_2^t), r^t \big\rangle}_{\text{Error}_2}$$

where Error_1 is the error of using $\widehat{q}_1^t$ for the min-player and Error_2 is the error of using $\widehat{q}_2^t$ for the max-player. By the occupancy measures in Algorithm 1, Error_1 and Error_2 take the bounds in Lemma 4. However, we need to develop a new upper bound for $\widehat{\text{Regret}}(T)$ as follows.

Lemma 17 *Fix $\delta \in (0,1)$. Then, with probability $1-\delta$,*

$$\begin{aligned}
\widehat{\text{Regret}}(T) \;\;\leq\;\; & V^{-1}\sum_{t=0}^{T-1}\big(\lambda_1^t\big(\langle q_1^\star, g^t\rangle - b_1\big) \,+\, \lambda_2^t\big(\langle q_2^\star, h^t\rangle - b_2\big)\big) \\
& + (\eta V)^{-1} L(1+\theta T)\big(\log(|X||A|) + \log(|Y||B|)\big) \,+\, (2V^{-1}L + 4\theta + \eta V)LT.
\end{aligned}$$

Proof By Lemma 1, with probability $1-\delta$ it holds that

$$\Delta(P_1) \;\subset\; \cap_{t=0}^{T-1}\Delta(k_1^t) \;\;\text{and}\;\; \Delta(P_2) \;\subset\; \cap_{t=0}^{T-1}\Delta(k_2^t).$$

We note that the solution $(q_1^\star, q_2^\star)$ in hindsight to Problem (44) satisfies $q_1^\star \in \Delta(P_1)$ and $q_2^\star \in \Delta(P_2)$. Hence, $q_1^\star \in \cap_{t=0}^{T-1}\Delta(k_1^t)$ and $q_2^\star \in \Delta(P_2) \cap_{t=0}^{T-1} \Delta(k_2^t)$ with probability $1-\delta$. For episode t, we apply Lemma 13 to the primal update (47) with

$$f(x,y)|_{x=q_1,\,y=q_2} \;=\; V\left\langle q_1 \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot q_2, r^{t-1}\right\rangle \,+\, \lambda_1^{t-1}\langle q_1, g^{t-1}\rangle \,-\, \lambda_2^{t-1}\langle q_2, h^{t-1}\rangle$$

and $x^\star = \widehat{q}_1^t$, $y^\star = \widehat{q}_2^t$, $x' = \widetilde{q}_1^{t-1}$, $y' = \widetilde{q}_2^{t-1}$, $x = q_1^\star$, and $y = q_2^\star$. Thus, with probability $1-\delta$ it holds for any t that

$$\begin{aligned}
& V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1}\right\rangle \,+\, \lambda_1^{t-1}\langle \widehat{q}_1^t, g^{t-1}\rangle \,-\, \lambda_2^{t-1}\langle q_2^\star, h^{t-1}\rangle \\
& + \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \\
\leq \;\; & V\left\langle q_1^\star \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1}\right\rangle \,+\, \lambda_1^{t-1}\langle q_1^\star, g^{t-1}\rangle \,-\, \lambda_2^{t-1}\langle \widehat{q}_2^t, h^{t-1}\rangle \\
& \quad + \eta^{-1}\big(D(q_1^\star, \widetilde{q}_1^{t-1}) \,+\, D(q_2^\star, \widetilde{q}_2^{t-1}) \,-\, D(q_1^\star, \widehat{q}_1^t) \,-\, D(q_2^\star, \widehat{q}_2^t)\big)
\end{aligned}$$

or, equivalently,

$$\begin{aligned}
& V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1}\right\rangle \,+\, \lambda_1^{t-1}\langle \widehat{q}_1^t, g^{t-1}\rangle \,+\, \lambda_2^{t-1}\langle \widehat{q}_2^t, h^{t-1}\rangle \\
& + \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \\
\leq \;\; & V\left\langle q_1^\star \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1}\right\rangle \,+\, \lambda_1^{t-1}\langle q_1^\star, g^{t-1}\rangle \,+\, \lambda_2^{t-1}\langle q_2^\star, h^{t-1}\rangle \\
& \quad + \eta^{-1}\big(D(q_1^\star, \widetilde{q}_1^{t-1}) \,+\, D(q_2^\star, \widetilde{q}_2^{t-1}) \,-\, D(q_1^\star, \widehat{q}_1^t) \,-\, D(q_2^\star, \widehat{q}_2^t)\big).
\end{aligned} \tag{49}$$

Let $\Delta_1^t := \frac{1}{2}\left((\lambda_1^t)^2 - (\lambda_1^{t-1})^2\right)$ be the drift of the first consecutive dual updates. Then,

$$\begin{aligned}
\Delta_1^t \;&=\; \frac{1}{2}\left((\lambda_1^t)^2 - (\lambda_1^{t-1})^2\right) \\
&=\; \frac{1}{2}\left(\max{}^2\Big(\lambda_1^{t-1} \,+\, (\langle \widehat{q}_1^t, g^{t-1}\rangle - b_1),\; 0\Big) - (\lambda_1^{t-1})^2\right) \\
&\leq\; \lambda_1^{t-1}\big(\langle \widehat{q}_1^t, g^{t-1}\rangle - b_1\big) \,+\, \frac{1}{2}\big(\langle \widehat{q}_1^t, g^{t-1}\rangle - b_1\big)^2 \\
&\leq\; \lambda_1^{t-1}\big(\langle \widehat{q}_1^t, g^{t-1}\rangle - b_1\big) \,+\, L^2
\end{aligned} \tag{50}$$

where the first inequality is due to $\max^2(x, 0) \leq x^2$ and we apply $\langle \widehat{q}_1^t, g^{t-1}\rangle \in [0, L]$, $b_1 \in [0, L]$ in the last inequality. Similarly, if we let $\Delta_2^t := \frac{1}{2}\left((\lambda_2^t)^2 - (\lambda_2^{t-1})^2\right)$, then

$$\Delta_2^t \;\leq\; \lambda_2^{t-1}\big(\langle \widehat{q}_2^t, h^{t-1}\rangle - b_2\big) \,+\, L^2. \tag{51}$$

Adding (50) and (51) to (49) from both sides of the inequalities without changing the inequality direction yields

$$\begin{aligned}
& V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1}\right\rangle \,+\, \Delta_1^t \,+\, \Delta_2^t \,+\, \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \\
\leq \;\; & V\left\langle q_1^\star \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1}\right\rangle \,+\, \lambda_1^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) \,+\, \lambda_2^{t-1}\big(\langle q_2^\star, h^{t-1}\rangle - b_2\big) \,+\, 2L^2 \\
& \quad + \eta^{-1}\big(D(q_1^\star, \widetilde{q}_1^{t-1}) + D(q_2^\star, \widetilde{q}_2^{t-1}) \,-\, D(q_1^\star, \widehat{q}_1^t) \,-\, D(q_2^\star, \widehat{q}_2^t)\big).
\end{aligned} \tag{52}$$

However,

$$
\begin{aligned}
& V \left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1} \right\rangle \,+\, \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \\
=\;& V \left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} - \widetilde{q}_1^{t-1} \cdot \widehat{q}_2^{t-1}, r^{t-1} \right\rangle \,+\, V \left\langle \widetilde{q}_1^{t-1} \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^{t-1}, r^{t-1} \right\rangle \\
& + V \left\langle \widehat{q}_1^{t-1} \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widetilde{q}_2^{t-1}, r^{t-1} \right\rangle \,+\, V \left\langle \widehat{q}_1^{t-1} \cdot \widetilde{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1} \right\rangle \\
& + \eta^{-1} D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) \,+\, \eta^{-1} D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) \\
\geq\;& - V \left\| \widehat{q}_2^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1 \,-\, V \left\| \widehat{q}_2^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widetilde{q}_1^{t-1} - \widehat{q}_1^{t-1} \right\|_1 \\
& - V \left\| \widehat{q}_1^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widehat{q}_2^{t-1} - \widetilde{q}_2^{t-1} \right\|_1 \,-\, V \left\| \widehat{q}_1^{t-1} \cdot r^{t-1} \right\|_\infty \left\| \widetilde{q}_2^{t-1} - \widehat{q}_2^t \right\|_1 \\
& + (2\eta L)^{-1} \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1^2 \,+\, (2\eta L)^{-1} \left\| \widehat{q}_2^t - \widetilde{q}_2^{t-1} \right\|_1 \\
\geq\;& - V \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1 \,-\, 2\theta V L \,+\, (2\eta L)^{-1} \left\| \widehat{q}_1^t - \widetilde{q}_1^{t-1} \right\|_1^2 \\
& - 2\theta V L \,-\, V \left\| \widetilde{q}_2^{t-1} - \widehat{q}_2^t \right\|_1 \,+\, (2\eta L)^{-1} \left\| \widehat{q}_2^t - \widetilde{q}_2^{t-1} \right\|_1 \\
\geq\;& - 4\theta V L \,-\, \eta V^2 L
\end{aligned}
$$

where we apply the Hölder's inequality and Lemma 14 in the first inequality, the second inequality is due to that

$$
\begin{aligned}
\left\| \widetilde{q}_1^{t-1} - \widehat{q}_1^{t-1} \right\|_1 \;&=\; \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left| (1-\theta) \widehat{q}_1^{t-1}(x,a) + \theta \frac{1}{|X_\ell||A|} - \widehat{q}_1^{t-1}(x,a) \right| \\
&\leq\; \theta \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \widehat{q}_1^{t-1}(x,a) + \theta \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \frac{1}{|X_\ell||A|} \\
&=\; 2\theta L
\end{aligned}
$$

and $\|\widetilde{q}_2^{t-1} - \widehat{q}_2^{t-1}\|_1 \leq 2\theta L$ that can be proved similarly, and the last inequality is due to $-bx+ax^2 \geq -b^2/(4a)$ for any a, $b > 0$. Therefore, we take the lower bound above for the left-hand side of (52),

$$
\begin{aligned}
& \Delta_1^t \,+\, \Delta_2^t \,-\, 4\theta V L \,-\, \eta V^2 L \\
\leq\;& V \left\langle q_1^\star \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1} \right\rangle \,+\, \lambda_1^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) \,+\, \lambda_2^{t-1}\big(\langle q_2^\star, h^{t-1}\rangle - b_2\big) \,+\, 2L^2 \\
& + \eta^{-1}\big(D(q_1^\star, \widetilde{q}_1^{t-1}) + D(q_2^\star, \widetilde{q}_2^{t-1}) - D(q_1^\star, \widehat{q}_1^t) - D(q_2^\star, \widehat{q}_2^t)\big).
\end{aligned}
\tag{53}
$$

By Lemma 15,

$$
\begin{aligned}
D(q_1^\star, \widetilde{q}_1^{t-1}) - D(q_1^\star, \widehat{q}_1^t) \;&=\; D(q_1^\star, \widetilde{q}_1^{t-1}) - D(q_1^\star, \widehat{q}_1^{t-1}) + D(q_1^\star, \widehat{q}_1^{t-1}) - D(q_1^\star, \widehat{q}_1^t) \\
&\leq\; \theta L \log(|X||A|) + D(q_1^\star, \widehat{q}_1^{t-1}) - D(q_1^\star, \widehat{q}_1^t)
\end{aligned}
$$

and, similarly,

$$
D(q_2^\star, \widetilde{q}_2^{t-1}) - D(q_2^\star, \widehat{q}_2^t) \;\leq\; \theta L \log(|Y||B|) + D(q_2^\star, \widehat{q}_2^{t-1}) - D(q_2^\star, \widehat{q}_2^t).
$$

We now simplify (53) into

$$
\begin{aligned}
\Delta_1^t \,+\, \Delta_2^t \;\leq\;& V \left\langle q_1^\star \cdot \widehat{q}_2^{t-1} - \widehat{q}_1^{t-1} \cdot q_2^\star, r^{t-1} \right\rangle \,+\, \lambda_1^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) \,+\, \lambda_2^{t-1}\big(\langle q_2^\star, h^{t-1}\rangle - b_2\big) \\
& + \eta^{-1}\big(D(q_1^\star, \widehat{q}_1^{t-1}) + D(q_2^\star, \widehat{q}_2^{t-1}) - D(q_1^\star, \widehat{q}_1^t) - D(q_2^\star, \widehat{q}_2^t)\big) \\
& + \eta^{-1}\theta L\big(\log(|X||A|) + \log(|Y||B|)\big) \,+\, 2L^2 \,+\, 4\theta V L \,+\, \eta V^2 L
\end{aligned}
$$

which leads to the desired result by summing it up from $t = 1$ to T,

$$
\begin{aligned}
\sum_{t=1}^{T} \left(\Delta_1^t + \Delta_2^t\right) \;\; & \leq \;\; V \sum_{t=1}^{T} \left\langle q_1^\star \cdot \widehat{q}_2^{\,t-1} - \widehat{q}_1^{\,t-1} \cdot q_2^\star, r^{t-1} \right\rangle \\
& \quad + \sum_{t=1}^{T} \left(\lambda_1^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) + \lambda_2^{t-1}\big(\langle q_2^\star, h^{t-1}\rangle - b_2\big)\right) \\
& \quad + \eta^{-1} \sum_{t=1}^{T} \left(D(q_1^\star, \widehat{q}_1^{\,t-1}) + D(q_2^\star, \widehat{q}_2^{\,t-1}) - D(q_1^\star, \widehat{q}_1^{\,t}) - D(q_2^\star, \widehat{q}_2^{\,t}) \right) \\
& \quad + \eta^{-1}\theta L T \big(\log(|X||A|) + \log(|Y||B|) \big) \; + \; 2L^2 T \; + \; 4\theta V L T \; + \; \eta V^2 L T \\
& \leq \;\; V \sum_{t=1}^{T} \left\langle q_1^\star \cdot \widehat{q}_2^{\,t-1} - \widehat{q}_1^{\,t-1} \cdot q_2^\star, r^{t-1} \right\rangle \\
& \quad + \sum_{t=1}^{T} \left(\lambda_1^{t-1}\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) + \lambda_2^{t-1}\big(\langle q_2^\star, h^{t-1}\rangle - b_2\big)\right) \\
& \quad + \eta^{-1} \left(D(q_1^\star, \widehat{q}_1^{\,0}) + D(q_2^\star, \widehat{q}_2^{\,0}) \right) \\
& \quad + \eta^{-1}\theta L T \big(\log(|X||A|) + \log(|Y||B|) \big) \; + \; 2L^2 T \; + \; 4\theta V L T \; + \; \eta V^2 L T
\end{aligned}
$$

which leads to the desired result by noting that

$$
D(q_1^\star, \widehat{q}_1^{\,0}) \;\leq\; L \log(|X||A|), \;\; D(q_2^\star, \widehat{q}_2^{\,0}) \;\leq\; L \log(|Y||B|), \;\; \text{and} \;\; \sum_{t=1}^{T} \left(\Delta_1^t + \Delta_2^t\right) \;\geq\; 0.
$$

■

To analyze the bound in Lemma 17, in Lemma 18, we next utilize a new drift bound from Lemma 22 to establish the boundedness of $\lambda^t := (\lambda_1^t, \lambda_2^t)$ first. Then, we apply a general Azuma-Hoeffding inequality for supermartingales in Lemma 19.

Lemma 18 *Let Assumption 2 hold. Fix $\delta \in (0, 1)$. For any integer $t_0 > 0$, with probability $1 - T\delta$,*

$$
\left\| \lambda^t \right\| \;\leq\; \Theta + 2t_0 L + t_0 \frac{64L^2}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{64L^2}{\xi} \log \frac{1}{\delta}
$$

for all $t = 1, \ldots, T$, where $\xi > 0$ and

$$
\Theta \;\; := \;\; t_0\left(\tfrac{1}{2}\xi + 2L\right) \; + \; \tfrac{4L^2 + (8\theta + 2\eta V + 2)VL}{\xi} \; + \; \tfrac{2L(\log(|X||A|/\theta) + \log(|Y||B|/\theta))}{t_0 \xi \eta}.
$$

Proof Let $\mathcal{F}^t$ be an σ-algebra that is generated by the state-action sequence, reward/utility functions for both players up to episode t. At the beginning, $\mathcal{F}^0 = \{\emptyset, \Omega\}$. We have a discrete-time random process $\{\left\|\lambda^t\right\|, t \geq 0\}$ that adapts to $\mathcal{F}^t$. It suffices to check all assumptions in Lemma 22.

By the dual update (48),

$$
\begin{aligned}
\left|\lambda_1^{t+1} - \lambda_1^t\right| \;\; & = \;\; \left| \max\Big(\lambda_1^t \; + \; \big(\langle \widehat{q}_1^{\,t+1}, g^t\rangle - b_1\big), \; 0\Big) - \lambda_1^t \right| \\
& \leq \;\; \left| \langle \widehat{q}_1^{\,t+1}, g^t \rangle - b_1 \right| \\
& \leq \;\; L
\end{aligned}
$$

where the first inequality is clear from two cases for $\max(\cdot)$ and the second inequality is due to $\langle \widehat{q}_1^{t+1}, g^t\rangle \in [0,L]$, $b_1 \in [0,L]$. Similarly, $\left|\lambda_2^{t+1} - \lambda_2^t\right| \le L$. Hence,

$$\left|\left\|\lambda^{t+1}\right\| - \left\|\lambda^t\right\|\right| \;\le\; \left\|\lambda^{t+1} - \lambda^t\right\| \;=\; \sqrt{(\lambda_1^{t+1}-\lambda_1^t)^2 + (\lambda_2^{t+1}-\lambda_2^t)^2} \;\le\; 2L.$$

Consequently,

$$\|\lambda\|^{t+t_0} - \|\lambda\|^t \;=\; \sum_{s=t}^{t+t_0-1}\big(\,\|\lambda\|^{s+1} - \|\lambda\|^s\,\big) \;\le\; \sum_{s=t}^{t+t_0-1}\Big|\|\lambda\|^{s+1} - \|\lambda\|^s\Big| \;\le\; 2t_0L \tag{54}$$

which leads to $\mathbb{E}[\,\|\lambda\|^{t+t_0} - \|\lambda\|^t \mid \mathcal{F}^t\,] \le 2t_0L$. It is convenient to take $\delta_{\max} = 2L$ in Lemma 22.

We next determine the validity of other assumptions in Lemma 22. Let us denote the event in Lemma 1 by $\mathcal{E}_{\text{good}}$ and we have $P(\mathcal{E}_{\text{good}}) \ge 1-\delta$. Let $\Delta^t := \frac{1}{2}\big(\left\|\lambda^t\right\|^2 - \left\|\lambda^{t-1}\right\|^2\big)$. Clearly, $\Delta^t = \Delta_1^t + \Delta_2^t$. We recall that the proof of Lemma 5 remains to be valid if we replace $q_1^\star$ by $\bar{q}_1$ and $q_2^\star$ by $\bar{q}_2$ starting from (49). By doing so, it is ready to obtain a similar result as (53): under the good event $\mathcal{E}_{\text{good}}$ it holds for any τ that

$$\begin{aligned}
&\Delta^\tau \;-\; 4\theta VL \;-\; \eta V^2 L\\
&\le\;\; V\left\langle \bar{q}_1\cdot\widehat{q}_2^{\tau-1} - \widehat{q}_1^{\tau-1}\cdot\bar{q}_2, r^{\tau-1}\right\rangle \;+\; \lambda_1^{\tau-1}\big(\langle\bar{q}_1, g^{\tau-1}\rangle - b_1\big) \;+\; \lambda_2^{\tau-1}\big(\langle\bar{q}_2, h^{\tau-1}\rangle - b_2\big) \;+\; 2L^2\\
&\qquad +\eta^{-1}\big(D(\bar{q}_1,\widetilde{q}_1^{\tau-1}) + D(\bar{q}_2,\widetilde{q}_2^{\tau-1}) - D(\bar{q}_1,\widehat{q}_1^{\tau}) - D(\bar{q}_2,\widehat{q}_2^{\tau})\big).
\end{aligned}$$

or, equivalently,

$$\begin{aligned}
&\left\|\lambda^\tau\right\|^2 \;-\; \left\|\lambda^{\tau-1}\right\|^2\\
&\le\;\; 2V\left\langle \bar{q}_1\cdot\widehat{q}_2^{\tau-1} - \widehat{q}_1^{\tau-1}\cdot\bar{q}_2, r^{\tau-1}\right\rangle \;+\; 2\lambda_1^{\tau-1}\big(\langle\bar{q}_1, g^{\tau-1}\rangle - b_1\big) \;+\; 2\lambda_2^{\tau-1}\big(\langle\bar{q}_2, h^{\tau-1}\rangle - b_2\big) \;+\; 4L^2\\
&\qquad +2\eta^{-1}\big(D(\bar{q}_1,\widetilde{q}_1^{\tau-1}) + D(\bar{q}_2,\widetilde{q}_2^{\tau-1}) - D(\bar{q}_1,\widehat{q}_1^{\tau}) - D(\bar{q}_2,\widehat{q}_2^{\tau})\big) \;+\; 8\theta VL \;+\; 2\eta V^2 L.
\end{aligned} \tag{55}$$

We note that $|\langle \bar{q}_1\cdot\widehat{q}_2^\tau - \widehat{q}_1^\tau\cdot\bar{q}_2, r^\tau\rangle| \le L$. By summing both sides of (55) from $\tau = t+1$ to $\tau = t+t_0$,

$$\begin{aligned}
\left\|\lambda^{t+t_0}\right\|^2 \;-\; \left\|\lambda^t\right\|^2 \;\;\le\;\; & 2t_0VL \;+\; 2\sum_{\tau=t}^{t+t_0-1}\big(\lambda_1^\tau\big(\langle\bar{q}_1,g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle\bar{q}_2,h^\tau\rangle - b_2\big)\big) \;+\; 4t_0L^2\\
& +2\eta^{-1}\big(D(\bar{q}_1,\widetilde{q}_1^t) + D(\bar{q}_2,\widetilde{q}_2^t)\big) \;+\; 8t_0\theta VL \;+\; 2t_0\eta V^2 L
\end{aligned}$$

where we omit two non-positive terms. Taking the conditional expectation given $\mathcal{F}^t$ and $\mathcal{E}_{\text{good}}$ yields,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\|^2 \;-\; \left\|\lambda^t\right\|^2 \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]\\
&\le\;\; 2t_0VL \;+\; 2\sum_{\tau=t}^{t+t_0-1}\mathbb{E}\left[\lambda_1^\tau\big(\langle\bar{q}_1,g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle\bar{q}_2,h^\tau\rangle - b_2\big) \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;+\; 4t_0L^2\\
&\qquad +2\eta^{-1}\mathbb{E}\left[D(\bar{q}_1,\widetilde{q}_1^t) + D(\bar{q}_2,\widetilde{q}_2^t) \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;+\; 8t_0\theta VL \;+\; 2t_0\eta V^2 L\\
&\le\;\; 2t_0VL \;-\; 2\xi\sum_{\tau=t}^{t+t_0-1}\mathbb{E}\left[\|\lambda^\tau\| \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;+\; 4t_0L^2\\
&\qquad +2\eta^{-1}L(\log(|X||A|/\theta) + \log(|Y||B|/\theta)) \;+\; 8t_0\theta VL \;+\; 2t_0\eta V^2 L\\
&\le\;\; 2t_0VL \;-\; 2\xi t_0\mathbb{E}\left[\left\|\lambda^t\right\| \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;+\; 2\xi t_0(t_0-1)L \;+\; 4t_0L^2\\
&\qquad +2\eta^{-1}L(\log(|X||A|/\theta) + \log(|Y||B|/\theta)) \;+\; 8t_0\theta VL \;+\; 2t_0\eta V^2 L
\end{aligned} \tag{56}$$

where the second inequality is due to Lemma 15 and the fact: by the law of total expectation, for any $\tau \geq t$, $\mathcal{F}^t \subset \mathcal{F}^\tau$ and

$$\begin{aligned}
&\mathbb{E}\left[\lambda_1^\tau\big(\langle \bar{q}_1, g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle \bar{q}_2, h^\tau\rangle - b_2\big) \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
=\;& \mathbb{E}\left[\mathbb{E}\left[\lambda_1^\tau\big(\langle \bar{q}_1, g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle \bar{q}_2, h^\tau\rangle - b_2\big) \,|\, \mathcal{F}^\tau, \mathcal{E}_{\text{good}}\right] \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
=\;& \mathbb{E}\left[\lambda_1^\tau \mathbb{E}\left[\langle \bar{q}_1, g^\tau\rangle - b_1\right] \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] + \mathbb{E}\left[\lambda_2^\tau \mathbb{E}\left[\langle \bar{q}_2, h^\tau\rangle - b_2\right] \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
=\;& \mathbb{E}\left[\langle \bar{q}_1, g^\tau\rangle - b_1\right] \mathbb{E}\left[\lambda_1^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] + \mathbb{E}\left[\langle \bar{q}_2, h^\tau\rangle - b_2\right] \mathbb{E}\left[\lambda_2^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
\leq\;& -\xi\, \mathbb{E}\left[\lambda_1^\tau + \lambda_2^\tau \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
\leq\;& -\xi\, \mathbb{E}\left[\|\lambda^\tau\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]
\end{aligned}$$

where the inequality is due to the strict feasibility assumption on $(\bar{q}_1, \bar{q}_2)$; the last inequality is due to that

$$\begin{aligned}
\sum_{\tau=t}^{t+t_0-1} \mathbb{E}\left[\|\lambda^\tau\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;&\geq\; \sum_{\tau=t}^{t+t_0-1} \mathbb{E}\left[\left\|\lambda^t\right\| - 2(\tau - t)L \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
&=\; \sum_{\tau=0}^{t_0-1} \mathbb{E}\left[\left\|\lambda^t\right\| - 2\tau L \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]
\end{aligned}$$

which follows the fact $\|\lambda^\tau\| \geq \left\|\lambda^t\right\| - 2(\tau - t)L$ for any $\tau \geq t \geq 0$ if we note that $\left|\left\|\lambda^{t+1}\right\| - \left\|\lambda^t\right\|\right| \leq 2L$. Hence, we can simplify (56) as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\|^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
\leq\;& \mathbb{E}\left[\left\|\lambda^t\right\|^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - 2\xi t_0 \mathbb{E}\left[\left\|\lambda^t\right\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] + 2\xi t_0^2 L + 4t_0 L^2 + 2t_0 V L \\
&+ 2\eta^{-1} L(\log(|X||A|/\theta) + \log(|Y||B|/\theta)) + 8t_0\theta V L + 2t_0\eta V^2 L \\
\leq\;& \mathbb{E}\left[\left\|\lambda^t\right\|^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \xi t_0 \mathbb{E}\left[\left\|\lambda^t\right\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \xi t_0 \Theta + 2\xi t_0^2 L + 4t_0 L^2 + 2t_0 V L \\
&+ 2\eta^{-1} L(\log(|X||A|/\theta) + \log(|Y||B|/\theta)) + 8t_0\theta V L + 2t_0\eta V^2 L \\
=\;& \mathbb{E}\left[\left\|\lambda^t\right\|^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \xi t_0 \mathbb{E}\left[\left\|\lambda^t\right\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \frac{1}{2}\xi^2 t_0^2 \\
\leq\;& \left(\mathbb{E}\left[\left\|\lambda^t\right\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] - \frac{1}{2}\xi t_0\right)^2
\end{aligned}$$

where we apply $\lambda^t \geq \Theta$ for the second inequality and we take Θ in Lemma 22,

$$\Theta \;=\; \frac{1}{2}\xi t_0 + 2t_0 L + \frac{4L^2 + 8\theta V L + 2\eta V^2 L + 2V L}{\xi} + \frac{2L(\log(|X||A|/\theta) + \log(|Y||B|/\theta))}{t_0 \xi \eta}.$$

Taking the square root and applying the Jensen's inequality yield

$$\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\| \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \;\leq\; \sqrt{\mathbb{E}\left[\|\lambda^{t+t_0}\|^2 \,|\, \mathcal{F}^t, \mathcal{E}_{\text{good}}\right]} \;\leq\; \left\|\lambda^t\right\| - \frac{1}{2}\xi t_0$$

which shows that $\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\|-\left\|\lambda^t\right\| \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \leq -\frac{1}{2}\xi t_0$. Application of law of total expectation to this inequality and (54) with $\delta < \frac{1}{12}$ yields

$$
\begin{aligned}
\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\|-\left\|\lambda^t\right\| \mid \mathcal{F}^t\right] &= P(\mathcal{E}_{\text{good}})\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\|-\left\|\lambda^t\right\| \mid \mathcal{F}^t, \mathcal{E}_{\text{good}}\right] \\
&\quad + P(\bar{\mathcal{E}}_{\text{good}})\mathbb{E}\left[\left\|\lambda^{t+t_0}\right\|-\left\|\lambda^t\right\| \mid \mathcal{F}^t, \bar{\mathcal{E}}_{\text{good}}\right] \\
&\leq -\frac{1}{2}\xi t_0 \times (1-\delta) \,+\, 2t_0 L \times \delta \\
&\leq -\frac{1}{4}\xi t_0
\end{aligned}
$$

which verifies the assumption of Lemma 22 if we take $\zeta = \xi/4$.

We now have verified all assumptions of Lemma 22 with appropriate parameters Θ, $\delta_{\max}$, ζ. For episode t, with probability $1-\delta$ it holds that

$$
\left\|\lambda^t\right\| \;\leq\; \Theta + t_0\delta_{\max} + t_0\frac{4\delta_{\max}^2}{\zeta}\log\left(\frac{8\delta_{\max}^2}{\zeta}\right) + t_0\frac{4\delta_{\max}^2}{\zeta}\log\frac{1}{\delta}.
$$

We complete the proof by taking a union bound over $t = 1, \cdots, T$. ■

Lemma 19 *Let Assumption 2 hold. Fix $\delta \in (0,1)$. For any integer $t_0 > 0$, with probability $1-2T\delta$,*

$$
\sum_{t=0}^{T-1}\left(\lambda_1^t\big(\langle q_1^\star, g^t\rangle - b_1\big) + \lambda_2^t\big(\langle q_2^\star, h^t\rangle - b_2\big)\right) \;\leq\; \sqrt{2Tc^2\log(1/(\delta T))}
$$

where $c := 2\Theta L + 4t_0L^2 + \frac{128t_0L^3}{\xi}\left(\log\left(\frac{128L^2}{\xi}\right)+\log\frac{1}{\delta}\right)$ and $\xi > 0$.

Proof Let $Z^t := \sum_{\tau=0}^{t-1}\left(\lambda_1^\tau\big(\langle q_1^\star, g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle q_2^\star, h^\tau\rangle - b_2\big)\right)$. We note that

$$
\begin{aligned}
&\mathbb{E}\left[Z^t \mid \mathcal{F}^{t-1}\right] \\
&= \mathbb{E}\left[\sum_{\tau=0}^{t-1}\left(\lambda_1^\tau\big(\langle q_1^\star, g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle q_2^\star, h^\tau\rangle - b_2\big)\right) \,\middle|\, \mathcal{F}^{t-1}\right] \\
&= \mathbb{E}\left[\sum_{\tau=0}^{t-2}\left(\lambda_1^\tau\big(\langle q_1^\star, g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle q_2^\star, h^\tau\rangle - b_2\big)\right) \,\middle|\, \mathcal{F}^{t-1}\right] \\
&\quad + \lambda_1^{t-1}\mathbb{E}\left[\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) \mid \mathcal{F}^{t-1}\right] \,+\, \lambda_2^{t-1}\mathbb{E}\left[\big(\langle q_2^\star, h^{t-1}\rangle - b_2\big) \mid \mathcal{F}^{t-1}\right] \\
&\leq \mathbb{E}\left[\sum_{\tau=0}^{t-2}\left(\lambda_1^\tau\big(\langle q_1^\star, g^\tau\rangle - b_1\big) + \lambda_2^\tau\big(\langle q_2^\star, h^\tau\rangle - b_2\big)\right) \,\middle|\, \mathcal{F}^{t-1}\right] \\
&= \mathbb{E}\left[Z^{t-1}\right]
\end{aligned}
$$

where the inequality is because of $\mathbb{E}\left[\big(\langle q_1^\star, g^{t-1}\rangle - b_1\big) \mid \mathcal{F}^{t-1}\right] = \langle q_1^\star, g\rangle - b_1 \leq 0$ and $\mathbb{E}\left[\big(\langle q_2^\star, h^{t-1}\rangle - b_1\big) \mid \mathcal{F}^{t-1}\right] \leq \langle q_2^\star, h\rangle - b_2 \leq 0$. Hence, $\{Z^t, t \geq 0\}$ a supermartingale.

We also note that

$$
\left|Z^{t+1} - Z^t\right| \;=\; \lambda_1^t\left|\langle q_1^\star, g^t\rangle - b_1\right| + \lambda_2^t\left|\langle q_2^\star, h^t\rangle - b_2\right| \;\leq\; 2\left\|\lambda^t\right\| L
$$

Thus, if $|Z^{t+1} - Z^t| > c$ for some $c \in \mathbb{R}^+$, then $\|\lambda^t\| > c/(2L)$. Let $Y^t := \|\lambda^t\| - c/(2L)$. Therefore,

$$\{|Z^{t+1} - Z^t| > c\} \subset \{Y^t > 0\}.$$

By Lemma 23,

$$P\left(\sum_{t=0}^{T-1} \left(\lambda_1^t\big(\langle q_1^\star, g^t\rangle - b_1\big) + \lambda_2^t\big(\langle q_2^\star, h^t\rangle - b_2\big)\right) \geq z\right) \leq \mathrm{e}^{-z^2/(2c^2T)} + \sum_{\tau=0}^{T-1} P\left(\|\lambda^t\| > \frac{c}{2L}\right). \tag{57}$$

By Lemma 18, with probability $1-\delta$ it holds for any t that

$$\|\lambda^t\| \leq \Theta + 2t_0 L + t_0 \frac{64L^2}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{64L^2}{\xi} \log\frac{1}{\delta}$$

or, equivalently,

$$P\left(\|\lambda^t\| \geq \Theta + 2t_0 L + t_0 \frac{64L^2}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{64L^2}{\xi} \log\frac{1}{\delta}\right) \leq \delta.$$

If we take

$$c = 2\Theta L + 4t_0 L^2 + t_0 \frac{128L^3}{\xi} \log\left(\frac{128L^2}{\xi}\right) + t_0 \frac{128L^3}{\xi} \log\frac{1}{\delta} \text{ and } z = \sqrt{2Tc^2 \log(1/(\delta T))}$$

then (57) becomes

$$P\left(\sum_{t=0}^{T-1} \left(\lambda_1^t\big(\langle q_1^\star, g^t\rangle - b_1\big) + \lambda_2^t\big(\langle q_2^\star, h^t\rangle - b_2\big)\right) \geq z\right) \leq 2\delta T$$

which proves the desired result. ■

We now ready to conclude a bound on $\widehat{\text{Regret}}(T)$ by combining Lemma 19 and Lemma 17.

Theorem 20 *Let Assumption 2 hold. Fix $T \geq \max(|X||A|, |B||Y|)$. Let $V = L\sqrt{T}$, $\eta = 1/(TL)$, $t_0 = \sqrt{T}$, and $\theta = 1/T$. Then, with probability $1 - 2T\delta$ it holds that*

$$\widehat{\text{Regret}}(T) \leq \widetilde{O}\big((|X| + |Y|)L\sqrt{T}\big).$$

Proof Using the given parameters V, η, t_0, and θ for Lemma 17, $\widehat{\text{Regret}}(T)$ is upper bounded by $\frac{1}{L\sqrt{T}} \sum_{t=0}^{T-1} \left(\lambda_1^t\big(\langle q_1^\star, g^t\rangle - b_1\big) + \lambda_2^t\big(\langle q_2^\star, h^t\rangle - b_2\big)\right) + \widetilde{O}(L\sqrt{T})$ with probability $1-\delta$. We note that $\Theta \leq \widetilde{O}(L^2\sqrt{T})$ and $T \geq \max(|X||A|, |B||Y|)$. Using parameters in Lemma 19, with probability $1 - 2T\delta$,

$$\sum_{t=0}^{T-1} \left(\lambda_1^t\big(\langle q_1^\star, g^t\rangle - b_1\big) + \lambda_2^t\big(\langle q_2^\star, h^t\rangle - b_2\big)\right) \leq \widetilde{O}(L^3 T).$$

We complete the proof by noting $L \leq |X| + |Y|$. ■

We conclude the regret bound in Theorem 16 by combining Lemma 4 and Theorem 20, and $\delta = p/(2T)$.

14.3. Constraint Violation Analysis

We begin with a decomposition using the auxiliary occupancy measures (q_1^t, q_2^t). By inserting $\langle \widehat{q}_1^t, g^t \rangle$ and $\langle \widehat{q}_2^t, h^t \rangle$ into $\text{Violation}_1(T)$ and $\text{Violation}_2(T)$, we have

$$\text{Violation}_1(T) \;=\; \underbrace{\left[\sum_{t=0}^{T-1} \left(\langle \widehat{q}_1^t, g^t \rangle - b_1\right)\right]_+}_{\widehat{\text{Violation}}_1(T)} + \underbrace{\sum_{t=0}^{T-1} \langle q_1^t - \widehat{q}_1^t, g^t \rangle}_{\text{Error}_3}$$

$$\text{Violation}_2(T) \;=\; \underbrace{\left[\sum_{t=0}^{T-1} \left(\langle \widehat{q}_2^t, h^t \rangle - b_2\right)\right]_+}_{\widehat{\text{Violation}}_2(T)} + \underbrace{\sum_{t=0}^{T-1} \langle q_2^t - \widehat{q}_2^t, h^t \rangle}_{\text{Error}_4}.$$

For Error_3 and Error_4, we have the same bounds in Lemma 9. We next bound $\widehat{\text{Violation}}_1(T)$ and $\widehat{\text{Violation}}_2(T)$ by applying the epoch property (Jaksch et al., 2010); see a proof in Appendix 13.

Theorem 21 *Let $V = L\sqrt{T}$, $\eta = 1/(TL)$, $t_0 = \sqrt{T}$, and $\theta = 1/T$. Then,*

$$\widehat{\text{Violation}}_1(T), \widehat{\text{Violation}}_2(T) \;\leq\; \|\lambda^T\| + \frac{2}{T-1}\sum_{t=1}^{T} \|\lambda^{t-1}\| + \widetilde{O}\big(L\sqrt{T(|X||A| + |Y||B|)}\big).$$

Proof By the dual update (48),

$$\begin{aligned}
\lambda_1^t \;&=\; \max\Big(\lambda_1^{t-1} + \left(\langle \widehat{q}_1^t, g^{t-1} \rangle - b_1\right),\; 0\Big) \\
&\geq\; \lambda_1^{t-1} + \left(\langle \widehat{q}_1^t, g^{t-1} \rangle - b_1\right) \\
&=\; \lambda_1^{t-1} + \left(\langle \widehat{q}_1^{t-1}, g^{t-1} \rangle - b_1\right) + \langle \widehat{q}_1^t - \widehat{q}_1^{t-1}, g^{t-1} \rangle \\
&\geq\; \lambda_1^{t-1} + \left(\langle \widehat{q}_1^{t-1}, g^{t-1} \rangle - b_1\right) - \left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1
\end{aligned} \tag{58a}$$

where the last inequality is due to: $\langle \widehat{q}_1^t - \widehat{q}_1^{t-1}, g^{t-1} \rangle \leq \|\widehat{q}_1^t - \widehat{q}_1^{t-1}\|_1 \|g^{t-1}\|_\infty$, and $\|g^{t-1}\|_\infty \in [0,1]$. Similarly,

$$\lambda_2^t \;\geq\; \lambda_2^{t-1} + \left(\langle \widehat{q}_2^{t-1}, h^{t-1} \rangle - b_2\right) - \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1. \tag{58b}$$

We note that $\lambda_1^0 = \lambda_2^0 = 0$ from the initialization. Summing up both sides of (58a) from $t = 1$ to $t = T$ leads to

$$\sum_{t=0}^{T-1} \left(\langle \widehat{q}_1^t, g^t \rangle - b_1\right) \;\leq\; \lambda_1^T + \sum_{t=1}^{T} \left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1. \tag{59a}$$

Similarly,

$$\sum_{t=0}^{T-1} \left(\langle \widehat{q}_2^t, h^t \rangle - b_2\right) \;\leq\; \lambda_2^T + \sum_{t=1}^{T} \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1. \tag{59b}$$

Hence,

$$\widehat{\text{Violation}}_1(T), \widehat{\text{Violation}}_2(T) \;\leq\; \|\lambda^T\| + \sum_{t=1}^{T} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1\right). \tag{60}$$

We recall $\widehat{q}_1^t \in \Delta(k_1^t)$, $\widehat{q}_2^t \in \Delta(k_2^t)$ in the primal update (47) and $\Delta(k_1^t)$ and $\Delta(k_2^t)$ in the confidence sets (11). To bound $\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1$, we consider two cases: (i) $k_1^t = k_1^{t-1}$ and $k_2^t = k_2^{t-1}$; (ii) either $k_1^t \neq k_1^{t-1}$ or $k_2^t \neq k_2^{t-1}$.

Case (i). In this case, we have: $\widehat{q}_1^t, \widehat{q}_1^{t-1} \in \Delta(k_1^t)$, $\widehat{q}_2^t, \widehat{q}_2^{t-1} \in \Delta(k_2^t)$. We begin with the primal update (8) and apply Lemma 13 with,

$$f(x,y)|_{x\,=\,q_1,\,y\,=\,q_2} \;=\; V\left\langle q_1 \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot q_2, r^{t-1}\right\rangle \;+\; \lambda_1^{t-1}\langle q_1, g^{t-1}\rangle \;-\; \lambda_2^{t-1}\langle q_2, h^{t-1}\rangle$$

and $x^\star = \widehat{q}_1^t$, $y^\star = \widehat{q}_2^t$, $x' = \widetilde{q}_1^{t-1}$, $y' = \widetilde{q}_2^{t-1}$, $x = \widehat{q}_1^{t-1}$, and $y = \widetilde{q}_2^{t-1}$. Thus,

$$\begin{aligned}
& V\left\langle \widehat{q}_1^t \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot \widetilde{q}_2^{t-1}, r^{t-1}\right\rangle \;+\; \lambda_1^{t-1}\langle \widehat{q}_1^t, g^{t-1}\rangle \;-\; \lambda_2^{t-1}\langle \widetilde{q}_2^{t-1}, h^{t-1}\rangle \\
&\quad + \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \\
\leq\;\; & V\left\langle \widetilde{q}_1^{t-1} \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot \widehat{q}_2^t, r^{t-1}\right\rangle \;+\; \lambda_1^{t-1}\langle \widetilde{q}_1^{t-1}, g^{t-1}\rangle \;-\; \lambda_2^{t-1}\langle \widehat{q}_2^t, h^{t-1}\rangle \\
&\quad - \eta^{-1}\big(D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t)\big).
\end{aligned}$$

or, equivalently,

$$\begin{aligned}
& \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \;+\; \eta^{-1}\big(D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t)\big) \\
\leq\;\; & V\left\langle (\widetilde{q}_1^{t-1} - \widehat{q}_1^t) \cdot \widehat{q}_2^{t-1} + \widehat{q}_1^{t-1} \cdot (\widehat{q}_2^t - \widetilde{q}_2^{t-1}), r^{t-1}\right\rangle \\
&\quad + \lambda_1^{t-1}\langle \widetilde{q}_1^{t-1} - \widehat{q}_1^t, g^{t-1}\rangle \;+\; \lambda_2^{t-1}\langle \widetilde{q}_2^{t-1} - \widehat{q}_2^t, h^{t-1}\rangle.
\end{aligned} \tag{61}$$

We note that $\langle(\widetilde{q}_1^{t-1} - \widehat{q}_1^t) \cdot \widehat{q}_2^{t-1}, r^{t-1}\rangle \leq \|(\widetilde{q}_1^{t-1} - \widehat{q}_1^t) \cdot \widehat{q}_2^{t-1}\|_1 \|r^{t-1}\|_\infty \leq \|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\|_1$, and, similarly, $\langle \widehat{q}_1^{t-1} \cdot (\widehat{q}_2^t - \widetilde{q}_2^{t-1}), r^{t-1}\rangle \leq \|\widehat{q}_2^t - \widetilde{q}_2^{t-1}\|_1$. Thus, we can reduce (61) into

$$\begin{aligned}
& \eta^{-1}\big(D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widehat{q}_2^t, \widetilde{q}_2^{t-1})\big) \;+\; \eta^{-1}\big(D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t)\big) \\
\leq\;\; & (V + \lambda_1^{t-1}) \left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 \;+\; (V + \lambda_2^{t-1}) \left\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\right\|_1 \\
\leq\;\; & (V + \left\|\lambda^{t-1}\right\|)\left(\left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 + \left\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\right\|_1\right)
\end{aligned}$$

where the left-hand side can be lower bounded by Lemma 14,

$$D(\widehat{q}_1^t, \widetilde{q}_1^{t-1}) + D(\widetilde{q}_1^{t-1}, \widehat{q}_1^t) \;\geq\; L^{-1}\left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1^2$$

$$D(\widehat{q}_2^t, \widetilde{q}_2^{t-1}) + D(\widetilde{q}_2^{t-1}, \widehat{q}_2^t) \;\geq\; L^{-1}\left\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\right\|_1^2.$$

Then, we apply the inequality $(x+y)^2 \leq 2(x^2 + y^2)$ and cancel a non-negative term to obtain

$$\left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 + \left\|\widetilde{q}_2^{t-1} - \widehat{q}_2^t\right\|_1 \;\leq\; 2\eta L(V + \left\|\lambda^{t-1}\right\|). \tag{62}$$

By the definition of $\widetilde{q}_1^{t-1}$ and $\widetilde{q}_2^{t-1}$,

$$\begin{aligned}
\left\|\widetilde{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 \;&=\; \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left|(1-\theta)\widehat{q}_1^{t-1}(x,a) + \theta\frac{1}{|X_\ell||A|} - \widehat{q}_1^t(x,a)\right| \\
&\geq\; \sum_{\ell=0}^{L-1} \sum_{x \in X_\ell} \sum_{a \in A} \left((1-\theta)\left|\widehat{q}_1^{t-1}(x,a) - \widehat{q}_1^t(x,a)\right| - \theta\left(\frac{1}{|X_\ell||A|} + \widehat{q}_1^t(x,a)\right)\right) \\
&=\; (1-\theta)\left\|\widehat{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 - 2\theta L.
\end{aligned}$$

Similarly, we have $\|\widehat{q}_2^{t-1} - \widehat{q}_2^t\|_1 \leq (1-\theta)\|\widehat{q}_2^{t-1} - \widehat{q}_2^t\|_1 - 2\theta L$. Thus, we can further reduce (62) into

$$\left\|\widehat{q}_1^{t-1} - \widehat{q}_1^t\right\|_1 + \|\widehat{q}_2^{t-1} - \widehat{q}_2^t\|_1 \;\leq\; 2\eta(1-\theta)^{-1}L(V + \left\|\lambda^{t-1}\right\|) + 4\theta(1-\theta)^{-1}L. \tag{63}$$

Case (ii). In this case, either $\widehat{q}_1^t$, $\widehat{q}_1^{t-1}$ or $\widehat{q}_2^t$, $\widehat{q}_2^{t-1}$ might not have the same domain. For instance, when $k_1^t > k_1^{t-1}$, it is possible that $\Delta(k_1^t)$ becomes different from $\Delta(k_1^{t-1})$. We note that $k_1^t > k_1^{t-1}$ only happens when episode t is the first one that belongs to epoch k_1^t. By Lemma 25, $k_1^T \leq \sqrt{T|X||A|}\log(8T/(|X||A|))$ and $k_2^T \leq \sqrt{T|Y||B|}\log(8T/(|Y||B|))$ if we are given $T \geq \max(|X||A|, |Y||B|)$.

We now combine two cases above for (60),

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \\
&= \sum_{\substack{1\leq t\leq T\\ k_1^t = k_1^{k-1} \wedge k_2^t = k_2^{k-1}}} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \\
&\quad + \sum_{\substack{1\leq t\leq T\\ k_1^t = k_1^{k-1} \vee k_2^t = k_2^{k-1}}} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \\
&\leq \sum_{\substack{1\leq t\leq T\\ k_1^t = k_1^{k-1} \wedge k_2^t = k_2^{k-1}}} \left(\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 + \left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1\right) \;+\; 2L(k_1^T + k_2^T) \\
&\leq 2\eta(1-\theta)^{-1}L\sum_{t=1}^{T}(V + \left\|\lambda^{t-1}\right\|) + 4\theta(1-\theta)^{-1}LT + 2L(k_1^T + k_2^T)
\end{aligned}$$

where the first inequality is due to: $\left\|\widehat{q}_1^t - \widehat{q}_1^{t-1}\right\|_1 \leq 2L$ and $\left\|\widehat{q}_2^t - \widehat{q}_2^{t-1}\right\|_1 \leq 2L$, and we apply (63) from the case (i) for the last inequality. Using the bounds on k_1^T, k_2^T in the case (ii), we conclude the desired bound for (60),

$$\begin{aligned}
&\widehat{\text{Violation}}_1(T), \widehat{\text{Violation}}_2(T) \\
&\leq \left\|\lambda^T\right\| + \frac{2\eta L}{1-\theta}\sum_{t=1}^{T}\left\|\lambda^{t-1}\right\| + \frac{2\eta V + 4\theta}{1-\theta}LT \\
&\quad + 2L\left(\sqrt{T|X||A|}\log(8T/(|X||A|)) + \sqrt{T|Y||B|}\log(8T/(|Y||B|))\right).
\end{aligned}$$

We complete the proof by noting $\lambda_1^0 = \lambda_2^0 = 0$, $V = L\sqrt{T}$, $\eta = 1/(TL)$, and $\theta = 1/T$. ■

To get the violation bound, we apply Lemma 18 to Theorem 21, use Lemma 9, and take $\delta = p/(2T)$.

15. Supporting Lemmas

We collect some useful lemmas in literature for the convenience of reading our paper.

The following drift analysis of stochastic processes is useful in the constraint violation analysis.

Lemma 22 *(Yu et al., 2017) Let $\{Z^t, t \geq 0\}$ be a discrete-time stochastic process that is adapted to a filtration $\{\mathcal{F}^t, t \geq 0\}$ with $Z^0 = 0$ and $\mathcal{F}^0 = \{\emptyset, \Omega\}$. Assume that there exists $t_0 \in \mathbb{Z}^+$, $\Theta \in \mathbb{R}^+$, $\delta_{\max} \in \mathbb{R}^+$, and $\zeta \in (0, \delta_{\max}]$ such that for all $t \geq 1$,*

$$\left|Z^{t+1} - Z^t\right| \;\leq\; \delta_{\max} \;\text{ and }\; \mathbb{E}\left[Z^{t+t_0} - Z^t \,|\, \mathcal{F}^t\right] \;\leq\; \begin{cases} t_0\,\delta_{\max} & \text{when } Z^t \leq \Theta \\ -\,t_0\,\zeta & \text{otherwise } Z^t \geq \Theta. \end{cases}$$

Then, with probability $1-\delta$ it holds for any t that

$$Z^t \;\leq\; \Theta + t_0\delta_{\max} + t_0\frac{4\delta_{\max}^2}{\zeta}\log\left(\frac{8\delta_{\max}^2}{\zeta}\right) + t_0\frac{4\delta_{\max}^2}{\zeta}\log\frac{1}{\delta}.$$

A general Azuma-Hoeffding inequality for supermartingales with unbounded differences is given as follows.

Lemma 23 *(Yu et al., 2017) Let $\{Z^t, t \geq 0\}$ be a supermartingale that is adapted to a filtration $\{\mathcal{F}^t, t \geq 0\}$ with $Z^0 = 0$ and $\mathcal{F}^0 = \{\emptyset, \Omega\}$. Let $\{Y^t, t \geq 0\}$ be a discrete-time stochastic process that is adapted to a filtration $\{\mathcal{F}^t, t \geq 0\}$. Assume that there exists a constant $c \in \mathbb{R}^+$ such that $\{|Z^{t+1} - Z^t| > c\} \subset \{Y^t > 0\}$ for any $t \geq 0$. Then, for any $z \in \mathbb{R}^+$ and $t \geq 1$,*

$$P(Z^t \geq z) \;\leq\; \mathrm{e}^{-z^2/(2c^2t)} + \sum_{\tau=0}^{t-1} P(Y^t > 0).$$

The following two lemmas are useful in the epoch analysis.

Lemma 24 *(Jaksch et al., 2010) Let a sequence of positive numbers be $x_1, \ldots, x_n$. Assume that $0 \leq x_k \leq X_{k-1} := \max(1, \sum_{i=1}^{k-1} x_i)$ for $1 \leq k \leq n$. Then,*

$$\sum_{k=1}^n \frac{x_k}{\sqrt{X_{k-1}}} \;\leq\; (\sqrt{2}+1)\sqrt{X_n}.$$

Lemma 25 *(Jaksch et al., 2010) Assume that $T \geq \max(|X||A|, |Y||B|)$. Then, the epochs k_1^T and k_2^T for episode T*

$$k_1^T \;\leq\; |X||A|\log\left(\frac{8T}{|X||A|}\right) \;\leq\; \sqrt{T|X||A|}\log\left(\frac{8T}{|X||A|}\right)$$

$$k_2^T \;\leq\; |Y||B|\log\left(\frac{8T}{|Y||B|}\right) \;\leq\; \sqrt{T|Y||B|}\log\left(\frac{8T}{|Y||B|}\right).$$