Proc. ACM Meas. Anal. Comput. Syst.

Online Learning in Weakly Coupled Markov Decision Processes: A Convergence Time Study

Xiaohan Wei*, Hao Yu* and Michael J. Neely*†

Abstract: We consider multiple parallel Markov decision processes (MDPs) coupled by global constraints, where the time varying objective and constraint functions can only be observed *after* the decision is made. Special attention is given to how well the decision maker can perform in T slots, starting from any state, compared to the best feasible randomized stationary policy in hindsight. We develop a new distributed online algorithm where each MDP makes its own decision each slot after observing a multiplier computed from past information. While the scenario is significantly more challenging than the classical online learning context, the algorithm is shown to have a tight $O(\sqrt{T})$ regret and constraint violations simultaneously. To obtain such a bound, we combine several new ingredients including ergodicity and mixing time bound in weakly coupled MDPs, a new regret analysis for online constrained optimization, a drift analysis for queue processes, and a perturbation analysis based on Farkas' Lemma.

ACM reference format: Xiaohan Wei, Hao Yu, and Michael J. Neely. 2018. Online Learning in Weakly Coupled Markov Decision Processes: A Convergence Time Study. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 1, Article 12 (March 2018).

Keywords and phrases: Stochastic programming, Constrained programming, Markov decision processes.

1. Introduction

This paper considers online constrained Markov decision processes (OCMDP) where both the objective and constraint functions can vary each time slot after the decision is made. We assume a slotted time scenario with time slots $t \in \{0, 1, 2, \ldots\}$. The OCMDP consists of K parallel Markov decision processes with indices $k \in \{1, 2, \ldots, K\}$. The k-th MDP has state space $\mathcal{S}^{(k)}$, action space $\mathcal{A}^{(k)}$, and transition probability matrix $P_a^{(k)}$ which depends on the chosen action $a \in \mathcal{A}^{(k)}$. Specifically, $P_a^{(k)} = (P_a^{(k)}(s, s'))$ where

$$P_a^{(k)}(s, s') = Pr\left(s_{t+1}^{(k)} = s' \;\middle|\; s_t^{(k)} = s,\ a_t^{(k)} = a\right),$$

where $s_t^{(k)}$ and $a_t^{(k)}$ are the state and action for system k on slot t. We assume that both the state space and the action space are finite for all $k \in \{1, 2, \cdots, K\}$.

After each MDP $k \in \{1, \ldots, K\}$ makes the decision at time t (and assuming the current state is $s_t^{(k)} = s$ and the action is $a_t^{(k)} = a$), the following information is revealed:

1. The next state $s_{t+1}^{(k)}$.
2. A penalty function $f_t^{(k)}(s, a)$ that depends on the current state s and the current action a.
3. A collection of m constraint functions $g_{1,t}^{(k)}(s, a), \ldots, g_{m,t}^{(k)}(s, a)$ that depend on s and a.

The functions $f_t^{(k)}$ and $g_{i,t}^{(k)}$ are all bounded mappings from $\mathcal{S}^{(k)} \times \mathcal{A}^{(k)}$ to $\mathbb{R}$ and represent different types of costs incurred by system k on slot t (depending on the current state and

*Department of Electrical Engineering, University of Southern California

†Michael J. Neely was partially supported by the National Science Foundation under grant CCF-1718477.

action). Note that in our model, the functions $f_t^{(k)}$ are *arbitrary* time-varying processes with no assumed probability structure. The constraint functions $g_{i,t}^{(k)}$ are time-varying but are assumed to be i.i.d. over slots with unknown distributions.

One concrete example of the above model is a multi-server data center, where the different systems $k \in \{1, \ldots, K\}$ can represent different servers, the penalty function for a particular server k can represent monetary expenditure for the power on that server, whose per unit price can change arbitrarily over time, and the constraints can represent service rate requirements on these servers to balance the job arrivals. Coupling among the server systems comes from using all of them to collectively support a common stream of arriving jobs. We will detail this example in Section 1.1.

A key aspect of this general problem is that the functions $f_t^{(k)}$ and $g_{i,t}^{(k)}$ are unknown until after the slot t decision is made. Thus, the precise costs incurred by each system are only known at the end of the slot. For any fixed time horizon of T slots, the overall penalty and constraint accumulation resulting from a policy $\mathscr{P}$ is:

$$F_T(d_0, \mathscr{P}) := \mathbb{E}\left(\sum_{t=1}^{T} \sum_{k=1}^{K} f_t^{(k)} \left(a_t^{(k)}, s_t^{(k)} \right) \middle| \ d_0, \mathscr{P} \right), \tag{1}$$

and

$$G_{i,T}(d_0, \mathscr{P}) := \mathbb{E}\left(\sum_{t=1}^{T} \sum_{k=1}^{K} g_{i,t}^{(k)} \left(a_t^{(k)}, s_t^{(k)} \right) \middle| \ d_0, \mathscr{P} \right),$$

where d_0 represents a given distribution on the initial joint state vector $(s_0^{(1)}, \cdots, s_0^{(K)})$. Note that $(a_t^{(k)}, s_t^{(k)})$ denotes the state-action pair of the kth MDP, which is a pair of random variables determined by d_0 and $\mathscr{P}$. Define a constraint set

$$\mathcal{G} := \{(\mathscr{P}, d_0) : \ G_{i,T}(d_0, \mathscr{P}) \leq 0, \ i = 1, 2, \cdots, m\}. \tag{2}$$

Define the regret of a policy $\mathscr{P}$ with respect to a particular joint randomized stationary policy Π along with an arbitrary starting state distribution d_0 as:

$$F_T(d_0, \mathscr{P}) - F_T(d_0, \Pi),$$

The goal of OCMDP is to choose a policy $\mathscr{P}$ so that both the regret and constraint violations grow sublinearly with respect to T, where regret is measured against all feasible joint randomized stationary policies Π. An important feature of this "weakly coupled" MDP structure is that, while the total state space $(s^{(1)}, \cdots, s^{(K)})$ grows exponentially in the number of subsystems K, our solution can be implemented separately at each system $i \in \{1, ..., K\}$ with complexity that depends only on the size of the individual system state $s^{(i)}$, rather than the product of sizes across all systems.

1.1. A motivating example

Consider a data center with a central controller and K servers (see Fig. 1). Jobs arrive randomly and are stored in a queue to await service. The system operates in slotted time $t \in \{0, 1, 2, \ldots\}$ and each server $k \in \{1, \ldots, K\}$ is modeled as a 3-state MDP with states *active*, *idle*, and *setup*:

- Active: In this state the server is available to serve jobs. Server k incurs a time varying electricity cost on every active slot, regardless of whether or not there are jobs to serve. It has a control option to stay active or transition to the idle state.

- Idle: In this state no jobs can be served. This state has multiple sleep modes as control options, each with different per-slot costs and setup times required for transitioning from idle to active.
- Setup: This is a transition state between idle and active. No jobs can be served and there are no control options. The setup costs and durations are (possibly constant) random variables depending on the preceding chosen sleep mode.

The goal is to minimize the overall electricity cost subject to the constraint that the expected service amount should be no less than the expected number of arrivals over any fixed time horizon T.

In a typical data center scenario, the performance of each server on a given slot is governed by the current electricity price, which can be an arbitrary time-varying sequence that is unknown beforehand, and the service rate, which can depend on the server state, service decision, and unknown noise factors affecting service. This problem is challenging because:

- If one server is currently in a setup state, it has zero service rate and cannot make another decision until it reaches the active state (which typically takes more than one slot), whereas other active servers can make decisions during this time. Thus, servers are acting asynchronously.
- The electricity price exhibits variation across time, location, and utility providers. Its behavior is irregular and can be difficult to predict. As an example, Fig. 2 plots the average per 5 minute spot market price (between 05/01/2017 and 05/10/2017) at New York zone CENTRL ([1]). Servers in different locations can have different price offerings, and this piles up the uncertainty across the whole system.

Despite these difficulties, this problem fits into the formulation of this paper: The electricity price acts as the global penalty function, and stability of the queue can be treated as a global constraint that the expected total number of arrivals is less than the expected service rate.

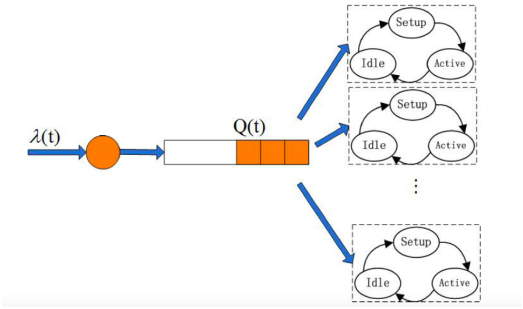

FIG 1. *Illustration of a data center server scheduling model.*

A review on data server provision can be found in [2] and references therein. Prior data center analysis often assumes the system has up-to-date information on service rates and electricity costs (see, for example, [3] and [4]). On the other hand, work that treats outdated information (such as [5], [6]) generally does not consider the potential Markov structure of the problem. The current paper treats the Markov structure of the problem and allows rate and price information to be unknown and outdated.

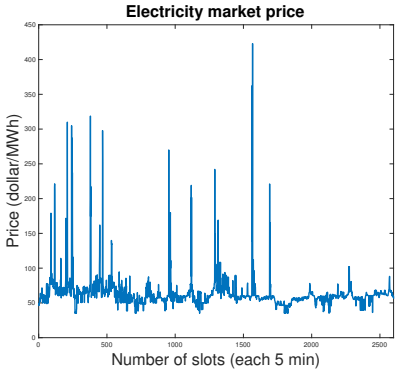

FIG 2. *A typical trace of electricity market price.*

1.2. Related work

- **Online convex optimization (OCO)**: This concerns multi-round cost minimization with arbitrarily-varying convex loss functions. Specifically, on each slot t the decision maker chooses decisions $x(t)$ within a convex set $\mathcal{X}$ (before observing the loss function $f^t(x)$) in order to minimize the total *regret* compared to the best fixed decision in hindsight, expressed as:

$$\text{regret}(T) = \sum_{t=1}^{T} f^t(\mathbf{x}(t)) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f^t(\mathbf{x}).$$

  See [7] for an introduction to OCO. Zinkevich introduced OCO in [8] and shows that an online projection gradient descent (OGD) algorithm achieves $O(\sqrt{T})$ regret. This $O(\sqrt{T})$ regret is proven to be the best in [9], although improved performance is possible if all convex loss functions are *strongly* convex. The OGD decision requires to compute a projection of a vector onto a set $\mathcal{X}$. For complicated sets $\mathcal{X}$ with functional equality constraints, e.g., $\mathcal{X} = \{x \in \mathcal{X}_0 : g_k(\mathbf{x}) \leq 0, k \in \{1, 2, \ldots, m\}\}$, the projection can have high complexity. To circumvent the projection, work in [10, 11, 12, 13] proposes alternative algorithms with simpler per-slot complexity and that satisfy the inequality constraints in the long term (rather than on every slot). Recently, new primal-dual type algorithms with low complexity are proposed in [14, 15] to solve more challenging OCO with time-varying functional inequality constraints. In particular, [15] also treats online convex optimization with stochastic i.i.d. constraints but without any Markov structure. Thus, in the degenerate scenario where there is only one state in the state-space $\mathcal{S}^{(k)}$, $\forall k$, our problem (1)-(2) can be solved via the method proposed in [15]. However, there is no prior work that addresses the general constrained online MDP problem.
- **Online Markov decision processes**: This extends OCO to allow systems with a more complex Markov structure. This is similar to the setup of the current paper of minimizing the expression (1), but does not have the constraint set (2). Unlike traditional OCO, the current penalty depends not only on the current action and the current (unknown) penalty function, but on the current system state (which depends on the history of previous actions). Further, the number of policies can grow exponentially with the sizes of the state and action spaces, so that solutions can be computationally intensive. The work [16] develops an algorithm in this context with $\mathcal{O}(\sqrt{T})$ regret. Extended algorithms and regularization methods are developed in [17][18][19] to reduce complexity and improve

dependencies on the number of states and actions. Online MDP under bandit feedback (where the decision maker can only observe the penalty corresponding to the chosen action) is considered in [20][19].

- **Constrained MDPs**: This aims to solve classical MDP problems with *known* cost functions but subject to additional constraints on the budget or resources. Linear programming methods for MDPs are found, for example, in [21], and algorithms beyond LP are found in [22][23]. Formulations closest to our setup appear in recent work on weakly coupled MDPs in [24][25] that have *known* cost and resource functions.

- **Reinforcement Learning (RL)**: This concerns MDPs with some unknown parameters (such as unknown functions and transition probabilities). The conventional setup of RL is different from constrained online MDP considered in this paper. Typically, RL considers decision making in an unknown but fixed probability structure (formulated as an MDP with unknown state spaces and/or unknown transmission probabilities). For example, prior work may assume the same expected penalty is incurred whenever we have the same state and same action. Methods for RL are developed in [26][27][28][29][30]. In contrast, the constrained online MDP studied in this paper assumes that state spaces and the transmission probabilities of the underlying MDPs are *known* to us, and deals with unknown and arbitrarily varying penalty functions for which there is no assumed probability structure.

1.3. Our contributions

The current paper proposes a new framework for online MDPs with time varying constraints. Further, it considers multiple MDP systems that are weakly coupled. While the scenario is significantly more challenging than the original Zinkevich OGD context as well as other classical online learning scenarios, the algorithm is shown to achieve tight $O(\sqrt{T})$ regret in both the objective function and the constraints, which ties the optimal $O(\sqrt{T})$ regret for those simpler unconstrained OCO problems. Along the way, we show the bound grows polynomially with the number of MDPs and linearly with respect to the number of states and actions in each MDP (Theorem 5.1).

The rest of the paper is organized as follows: In Section 2 we provide preliminary assumptions, facts and give some intuitions on the algorithm design (Section 2.5). In Section 3, we present our new algorithm along with the intuitions and roadmap of the analysis. In Section 4, we prove the regret and constraint violation bounds with respect to all randomized stationary policies starting from their stationary state distributions. Section 5 extends the result in the previous section by considering all randomized stationary policies starting from arbitrary states. Finally, we conclude the paper in Section 6.

2. Preliminaries

2.1. Basic Definitions

Throughout this paper, given an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, a *policy* $\mathscr{P}$ defines a (possibly probabilistic) method of choosing actions $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ based on the past information. We start with some basic definitions of important classes of policies:

Definition 2.1. *For an MDP, a* ***randomized stationary policy*** π *defines an algorithm which, whenever the system is in state $s \in \mathcal{S}$, chooses an action $a \in \mathcal{A}$ according to a fixed conditional probability function $\pi(a|s)$, defined for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$.*

Definition 2.2. *For an MDP, a* ***pure policy*** π *is a randomized stationary policy with all probabilities equal to either 0 or 1. That is, a pure policy is defined by a deterministic mapping between states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$. Whenever the system is in a state $s \in \mathcal{S}$, it always chooses a particular action $a_s \in \mathcal{A}$ (with probability 1).*

Note that if an MDP has a finite state and action space, the set of all pure policies is also finite. Consider the MDP associated with a particular system $k \in \{1, \ldots, K\}$. For any randomized stationary policy π, it holds that $\sum_{a \in \mathcal{A}^{(k)}} \pi(a|s) = 1$ for all $s \in \mathcal{S}^{(k)}$. Define the transition probability matrix $\mathbf{P}_\pi^{(k)}$ under policy π to have components as follows:

$$P_\pi^{(k)}(s, s') = \sum_{a \in \mathcal{A}^{(k)}} \pi(a|s) P_a^{(k)}(s, s'), \quad s, s' \in \mathcal{S}^{(k)}. \tag{3}$$

It is easy to verify that $\mathbf{P}_\pi^{(k)}$ is indeed a *stochastic matrix*, that is, it has rows with nonnegative components that sum to 1. Let $d_0^{(k)} \in [0,1]^{|\mathcal{S}^{(k)}|}$ be an (arbitrary) initial distribution for the k-th MDP. Define the state distribution at time t under π as $d_{\pi,t}^{(k)}$. By the Markov property of the system, we have $d_{\pi,t}^{(k)} = d_0^{(k)} \left(\mathbf{P}_\pi^{(k)}\right)^t$. A transition probability matrix $\mathbf{P}_\pi^{(k)}$ is *ergodic* if it gives rise to a Markov chain that is irreducible and aperiodic. Since the state space is finite, an ergodic matrix $\mathbf{P}_\pi^{(k)}$ has a unique stationary distribution denoted $d_\pi^{(k)}$, so that $d_\pi^{(k)}$ is the unique probability vector solving $d = d\mathbf{P}_\pi^{(k)}$.

Assumption 2.1 (Unichain model)**.** *There exists a universal integer $\widehat{r} \geq 1$ such that for any integer $r \geq \widehat{r}$ and every $k \in \{1, \ldots, K\}$, we have the product $\mathbf{P}_{\pi_1}^{(k)} \mathbf{P}_{\pi_2}^{(k)} \cdots \mathbf{P}_{\pi_r}^{(k)}$ is a transition matrix with strictly positive entries for any sequence of pure policies $\pi_1, \pi_2, \cdots, \pi_r$ associated with the kth MDP.*

Remark 2.1. *Assumption 2.1 implies that each MDP $k \in \{1, \ldots, K\}$ is ergodic under any pure policy. This follows by taking $\pi_1, \pi_2, \cdots, \pi_r$ all the same in Assumption 2.1. Since the transition matrix of any randomized stationary policy can be formed as a convex combination of those of pure policies, any randomized stationary policy results in an ergodic MDP for which there is a unique stationary distribution. Assumption 2.1 is easy to check via the following simple sufficient condition.*

Proposition 2.1. *Assumption 2.1 holds if, for every $k \in \{1, \ldots, K\}$, there is a fixed ergodic matrix $\mathbf{P}^{(k)}$ (i.e., a transition probability matrix that defines an irreducible and aperiodic Markov chain) such that for any pure policy π on MDP k we have the decomposition*

$$\mathbf{P}_\pi^{(k)} = \delta_\pi \mathbf{P}^{(k)} + (1 - \delta_\pi) \mathbf{Q}_\pi^{(k)},$$

where $\delta_\pi \in (0, 1]$ depends on the pure policy π and $\mathbf{Q}_\pi^{(k)}$ is a stochastic matrix depending on π.

Proof. Fix $k \in \{1, \ldots, K\}$ and assume every pure policy on MDP k has the above decomposition. Since there are only finitely many pure policies, there exists a lower bound $\delta_{\min} > 0$ such that $\delta_\pi \geq \delta_{\min}$ for every pure policy π. Since $\mathbf{P}^{(k)}$ is an ergodic matrix, there exists an integer $r^{(k)} > 0$ large enough such that $(\mathbf{P}^{(k)})^r$ has strictly positive components for all $r \geq r^{(k)}$. Fix $r \geq r^{(k)}$ and let $\pi_1, \ldots, \pi_r$ be any sequence of r pure policies on MDP k. Then

$$\mathbf{P}_{\pi_1}^{(k)} \cdots \mathbf{P}_{\pi_r}^{(k)} \geq \delta_{\min} \left(\mathbf{P}^{(k)}\right)^r > 0,$$

where inequality is treated entrywise. The universal integer r can be taken as the maximum integer $r^{(k)}$ over all $k \in \{1, \ldots, K\}$. □

Definition 2.3. *A **joint randomized stationary policy** Π on K parallel MDPs defines an algorithm which chooses a joint action $\mathbf{a} := \left(a^{(1)},\ a^{(2)},\ \cdots,\ a^{(K)}\right) \in \mathcal{A}^{(1)} \times \mathcal{A}^{(2)} \cdots \times \mathcal{A}^{(K)}$ given the joint state $\mathbf{s} := \left(s^{(1)},\ s^{(2)},\ ,\cdots,s^{(K)}\right) \in \mathcal{S}^{(1)} \times \mathcal{S}^{(2)} \cdots \times \mathcal{S}^{(K)}$ according to a fixed conditional probability $\Pi\left(\mathbf{a}\,|\mathbf{s}\right)$.*

The following special class of *separable* policies can be implemented separately over each of the K MDPs and plays a role in both algorithm design and performance analysis.

Definition 2.4. *A joint randomized stationary policy π is **separable** if the conditional probabilities $\pi := \left(\pi^{(1)},\ \pi^{(2)},\ \cdots,\ \pi^{(K)}\right)$ decompose as a product*

$$\pi\left(\mathbf{a}\,|\mathbf{s}\right) = \prod_{k=1}^{K} \pi^{(k)}\left(a^{(k)}|s^{(k)}\right)$$

for all $\mathbf{a} \in \mathcal{A}^{(1)} \times \cdots \times \mathcal{A}^{(K)}$, $\mathbf{s} \in \mathcal{S}^{(1)} \cdots \times \mathcal{S}^{(K)}$.

2.2. Technical assumptions

The functions $f_t^{(k)}$ and $g_{i,t}^{(k)}$ are determined by random processes defined over $t = 0, 1, 2, \cdots$. Specifically, let Ω be a finite dimensional vector space. Let $\{\omega_t\}_{t=0}^{\infty}$ and $\{\mu_t\}_{t=0}^{\infty}$ be two sequences of random vectors in Ω. Then for all $a \in \mathcal{A}^{(k)}$, $s \in \mathcal{S}^{(k)}$, $i \in \{1, 2, \cdots, m\}$ we have

$$g_{i,t}^{(k)}(a, s) = \hat{g}_i^{(k)}\left(a, s, \omega_t\right),$$
$$f_t^{(k)}(a, s) = \hat{f}^{(k)}\left(a, s, \mu_t\right)$$

where $\hat{g}_i^{(k)}$ and $\hat{f}^{(k)}$ formally define the time-varying functions in terms of the random processes ω_t and μ_t. It is assumed that the processes $\{\omega_t\}_{t=0}^{\infty}$ and $\{\mu_t\}_{t=0}^{\infty}$ are generated at the start of slot 0 (before any control actions are taken), and revealed gradually over time, so that functions $g_{i,t}^{(k)}$ and $f_t^{(k)}$ are only revealed at the end of slot t.

Remark 2.2. *The functions generated at time 0 in this way are also called oblivious functions because they are not influenced by control actions. Such an assumption is commonly adopted in previous unconstrained online MDP works (e.g. [16], [19] and [17]). Further, it is also shown in [19] that without this assumption, one can choose a sequence of objective functions against the decision maker in a specifically designed MDP scenario so that one never achieves the sublinear regret.*

The functions are also assumed to be bounded by a universal constant Ψ, so that:

$$|\hat{g}_i^{(k)}(a, s, \omega)| \leq \Psi, |\hat{f}^{(k)}(a, s, \mu)| \leq \Psi \quad , \forall k \in \{1, \ldots, K\}, \forall a \in \mathcal{A}^{(k)},\ s \in \mathcal{S}^{(k)},\ \forall \omega, \mu \in \Omega. \quad (4)$$

It is assumed that $\{\omega_t\}_{t=0}^{\infty}$ is independent, identically distributed (i.i.d.) and independent of $\{\mu_t\}_{t=0}^{\infty}$. Hence, the constraint functions can be arbitrarily correlated on the same slot, but appear i.i.d. over different slots. On the other hand, no specific model is imposed on $\{\mu_t\}_{t=0}^{\infty}$. Thus, the functions $f_t^{(k)}$ can be arbitrarily time varying. Let $\mathcal{H}_t$ be the system information up to time t, then, for any $t \in \{0, 1, 2, \cdots\}$, $\mathcal{H}_t$ contains state and action information up to time t, i.e. $\mathbf{s}_0, \cdots, \mathbf{s}_t$, $\mathbf{a}_0, \cdots, \mathbf{a}_t$, and $\{\omega_t\}_{t=0}^{\infty}$ and $\{\mu_t\}_{t=0}^{\infty}$. Throughout this paper, we make the following assumptions.

Assumption 2.2 (Independent transition)**.** *For each MDP, given the state $s_t^{(k)} \in \mathcal{S}^{(k)}$ and action $a_t^{(k)} \in \mathcal{A}^{(k)}$, the next state $s_{t+1}^{(k)}$ is independent of all other past information up to time t as well as the state transition $s_{t+1}^{(j)},\ \forall j \neq k$, i.e., for all $s \in \mathcal{S}^{(k)}$ it holds that*

$$Pr\left(s_{t+1}^{(k)} = s | \mathcal{H}_t, s_{t+1}^{(j)},\ \forall j \neq k\right) = Pr\left(s_{t+1}^{(k)} = s | s_t^{(k)}, a_t^{(k)}\right)$$

where $\mathcal{H}_t$ contains all past information up to time t.

Intuitively, this assumption means that all MDPs are running independently in the joint probability space and thus the only coupling among them comes from the constraints, which reflects the notion of *weakly coupled MDPs* in our title. Furthermore, by definition of $\mathcal{H}_t$, given $s_t^{(k)}, a_t^{(k)}$, the next transition $s_{t+1}^{(k)}$ is also independent of function paths $\{\omega_t\}_{t=0}^{\infty}$ and $\{\mu_t\}_{t=0}^{\infty}$.

The following assumption states the constraint set is strictly feasible.

Assumption 2.3 (Slater's condition)**.** *There exists a real value $\eta > 0$ and a fixed separable randomized stationary policy $\widetilde{\pi}$ such that*

$$\mathbb{E}\left[\sum_{k=1}^{K} g_{i,t}^{(k)}\left(a_t^{(k)}, s_t^{(k)}\right) \,\middle|\, d_{\widetilde{\pi}}, \widetilde{\pi}\right] \leq -\eta,\ \forall i \in \{1, 2, \cdots, m\},$$

where the initial state is $d_{\widetilde{\pi}}$ and is the unique stationary distribution of policy $\widetilde{\pi}$, and the expectation is taken with respect to the random initial state and the stochastic function $g_{i,t}^{(k)}(a, s)$ (i.e., ω_t).

Slater's condition is a common assumption in convergence time analysis of constrained convex optimization (e.g. [31], [32]). Note that this assumption readily implies the constraint set $\mathcal{G}$ can be achieved by the above randomized stationary policy. Specifically, take $d_0^{(k)} = d_{\widetilde{\pi}^{(k)}}$ and $\mathscr{P} = \widetilde{\pi}$, then, we have

$$G_{i,T}(d_0, \widetilde{\pi}) = \sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{k=1}^{K} g_{i,t}^{(k)}\left(a_t^{(k)}, s_t^{(k)}\right) \,\middle|\, d_{\widetilde{\pi}}, \widetilde{\pi}\right] \leq -\eta T < 0.$$

2.3. The state-action polyhedron

In this section, we recall the well-known linear program formulation of an MDP (see, for example, [21] and [33]). Consider an MDP with a state space $\mathcal{S}$ and an action space $\mathcal{A}$. Let $\Delta \subseteq \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a probability simplex, i.e.

$$\Delta = \left\{\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \theta(s,a) = 1,\ \theta(s,a) \geq 0\right\}.$$

Given a randomized stationary policy π with stationary state distribution d_π, the MDP is a Markov chain with transition matrix $\mathbf{P}_\pi$ given by (3). Thus, it must satisfy the following balance equation:

$$\sum_{s\in\mathcal{S}} d_\pi(s) P_\pi(s, s') = d_\pi(s'),\ \forall s' \in \mathcal{S}.$$

Defining $\theta(a, s) = \pi(a|s) d_\pi(s)$ and substituting the definition of transition probability (3) into the above equation gives

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \theta(s,a) P_a(s, s') = \sum_{a\in\mathcal{A}} \theta(s', a),\quad \forall s' \in \mathcal{S}.$$

The variable $\theta(a,s)$ is often interpreted as a stationary probability of being at state $s \in \mathcal{S}$ and taking action $a \in \mathcal{A}$ under some randomized stationary policy. The state action polyhedron Θ is then defined as

$$\Theta := \left\{ \theta \in \Delta : \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \theta(s,a) P_a(s,s') = \sum_{a\in\mathcal{A}} \theta(s',a), \quad \forall s' \in \mathcal{S} \right\}.$$

Given any $\theta \in \Theta$, one can recover a randomized stationary policy π at any state $s \in \mathcal{S}$ as

$$\pi(a|s) = \begin{cases} \frac{\theta(a,s)}{\sum_{a\in\mathcal{A}} \theta(a,s)}, & \text{if } \sum_{a\in\mathcal{A}} \theta(a,s) \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Given any fixed penalty function $f(a,s)$, the best policy minimizing the penalty (without constraint) is a randomized stationary policy given by the solution to the following linear program (LP):

$$\min \quad \langle \mathbf{f}, \theta \rangle, \quad s.t. \quad \theta \in \Theta. \tag{6}$$

where $\mathbf{f} := [f(a,s)]_{a\in\mathcal{A},\ s\in\mathcal{S}}$. Note that for any policy π given by the state-action pair θ according to (5),

$$\langle \mathbf{f}, \theta \rangle = \mathbb{E}_{s\sim d_\pi, a\sim\pi(\cdot|s)} \left[f(a,s) \right],$$

Thus, $\langle \mathbf{f}, \theta \rangle$ is often referred to as the stationary state penalty of policy π.

It can also be shown that any state-action pair in the set Θ can be achieved by a convex combination of state-action vectors of pure policies, and thus all corner points of the polyhedron Θ are from pure policies. As a consequence, the best randomized stationary policy solving (6) is always a pure policy.

2.4. Preliminary results on MDPs

In this section, we give preliminary results regarding the properties of our weakly coupled MDPs under randomized stationary policies. The proofs can be found in Appendix A.1. We start with a lemma on the uniform mixing of MDPs.

Lemma 2.1. *Suppose Assumption 2.1 and 2.2 hold. There exists a positive integer r and a constant $\tau \geq 1$ such that for any two state distributions d_1 and d_2,*

$$\sup_{\pi_1^{(k)},\cdots,\pi_r^{(k)}} \left\| \left(d_1^{(k)} - d_2^{(k)} \right) \mathbf{P}_{\pi_1^{(k)}}^{(k)} \mathbf{P}_{\pi_2^{(k)}}^{(k)} \cdots \mathbf{P}_{\pi_r^{(k)}}^{(k)} \right\|_1 \leq e^{-1/\tau} \left\| d_1^{(k)} - d_2^{(k)} \right\|_1, \ \forall k \in \{1,2,\cdots,K\}$$

*where the supremum is taken with respect to **any** sequence of r randomized stationary policies $\left\{ \pi_1^{(k)}, \cdots, \pi_r^{(k)} \right\}$.*

For the k-th MDP, let $\Theta^{(k)}$ be its state-action polyhedron according to the definition in Section 2.3. For any joint randomized stationary policy, let $\theta^{(k)}$ be the marginal state-action probability vector on the k-th MDP, i.e. for any joint state-action distribution $\Phi(\mathbf{a},\mathbf{s})$ where $\mathbf{a} \in \mathcal{A}^{(1)} \times \cdots \times \mathcal{A}^{(K)}$ and $\mathbf{s} \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$, we have $\theta^{(k)}(a^{(k)}, s^{(k)}) = \sum_{a^{(j)},s^{(j)},\ j\neq k} \Phi(\mathbf{a},\mathbf{s})$.

We have the following lemma:

Lemma 2.2. *Suppose Assumption 2.1 and 2.2 hold. Consider the product MDP with product state space $\mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$ and action space $\mathcal{A}^{(1)} \times \cdots \times \mathcal{A}^{(K)}$. Then, for any joint randomized stationary policy, the following hold:*

1. *The product MDP is irreducible and aperiodic.*
2. *The marginal stationary state-action probability vector $\theta^{(k)} \in \Theta^{(k)},\ \forall k \in \{1, 2, \cdots, K\}$.*

An immediate conclusion we can draw from this lemma is that given any penalty and constraint functions $\mathbf{f}^{(k)}$ and $\mathbf{g}_i^{(k)}$, $k = 1, 2, \cdots, K$, the stationary penalty and constraint value of any joint randomized stationary policy can be expressed as

$$\sum_{k=1}^{K} \left\langle \mathbf{f}^{(k)}, \theta^{(k)} \right\rangle, \ \sum_{k=1}^{K} \left\langle \mathbf{g}_i^{(k)}, \theta^{(k)} \right\rangle, \quad i = 1, 2, \cdots, m,$$

with $\theta^{(k)} \in \Theta^{(k)}$. This in turn implies such stationary state-action probabilities $\{\theta^{(k)}\}_{k=1}^K$ can also be realized via a separable randomized stationary policy π with

$$\pi^{(k)}(a|s) = \frac{\theta^{(k)}(a, s)}{\sum_{a \in \mathcal{A}^{(k)}} \theta^{(k)}(a, s)},\ a \in \mathcal{A}^{(k)},\ s \in \mathcal{S}^{(k)}, \tag{7}$$

and the corresponding stationary penalty and constraint value can also be achieved via this policy. This fact implies that when considering the stationary state performance only, the class of separable randomized stationary policies is large enough to cover all possible stationary penalty and constraint values.

In particular, let $\tilde{\pi} = \left(\tilde{\pi}^{(1)}, \cdots, \tilde{\pi}^{(K)}\right)$ be the separable randomized stationary policy associated with the Slater condition (Assumption 2.3). Using the fact that the constraint functions $\mathbf{g}_{i,t}^{(k)}, k = 1, 2, \cdots, K$ (i.e. w_t) are i.i.d.and Assumption 2.2 on independence of probability transitions, we have the constraint functions $g_{i,t}^{(k)}$ and the state-action pairs at any time t are mutuallly independent. Thus,

$$\mathbb{E}\left[\sum_{k=1}^{K} g_{i,t}^{(k)}\left(a_t^{(k)}, s_t^{(k)}\right) \,\middle|\, d_{\tilde{\pi}}, \tilde{\pi}\right] = \sum_{k=1}^{K} \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \tilde{\theta}^{(k)} \right\rangle,$$

where $\tilde{\theta}^{(k)}$ corresponds to $\tilde{\pi}$ according to (7).

Then, Slater's condition can be translated to the following: There exists a sequence of state-action probabilities $\{\tilde{\theta}^{(k)}\}_{k=1}^K$ from a separable randomized stationary policy such that $\tilde{\theta}^{(k)} \in \Theta^{(k)},\ \forall k$, and

$$\sum_{k=1}^{K} \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \tilde{\theta}^{(k)} \right\rangle \leq -\eta, \quad i = 1, 2, \cdots, m, \tag{8}$$

The assumption on separability does not lose generality in the sense that if there is no separable randomized stationary policy that satisfies (8), then, there is no *joint* randomized stationary policy that satisfies (8) either.

2.5. The blessing of slow-update property in online MDPs

The current state of an MDP depends on previous states and actions. As a consequence, the slot t penalty not only depends on the current penalty function and current action, but also on the system history. This complication does not arise in classical online convex optimization

([7],[8]) as there is no notion of "state" and the slot t penalty depends only on the slot t penalty function and action.

Now imagine a virtual system where, on each slot t, a policy π_t is chosen (rather than an action). Further imagine the MDP immediately reaching its corresponding stationary distribution d_{π_t}. Then the states and actions on previous slots do not matter and the slot t performance depends only on the chosen policy π_t and on the current penalty and constraint functions. This imaginary system now has a structure similar to classical online convex optimization as in the Zinkevich scenario [8].

A key feature of online convex optimization algorithms as in [8] is that they update their decision variables slowly. For a fixed time scale T over which $\mathcal{O}(\sqrt{T})$ regret is desired, the decision variables are typically changed no more than a distance $\mathcal{O}(1/\sqrt{T})$ from one slot to the next. An important insight in prior (unconstrained) MDP works(e.g. [17], [16], and [19]) is that such slow updates also guarantee the "approximate" convergence of an MDP to its stationary distribution. As a consequence, one can design the decision policies under the imaginary assumption that the system instantly reaches its stationary distribution, and later bound the error between the true system and the imaginary system. If the error is on the same order as the desired $\mathcal{O}(\sqrt{T})$ regret, then this approach works. This idea serves as a cornerstone of our algorithm design of the next section, which treats the case of multiple weakly coupled systems with both objective functions and constraint functions.

3. OCMDP algorithm

Our proposed algorithm is distributed in the sense that each time slot, each MDP solves its own subproblem and the constraint violations are controlled by a simple update of global multipliers called "virtual queues" at the end of each slot. Let $\Theta^{(1)}$, $\Theta^{(2)}$, $\cdots$, $\Theta^{(K)}$ be the state-action polyhedra of K MDPs, respectively. Let $\theta_t^{(k)} \in \Theta^{(k)}$ be a state-action vector at time slot t. At $t = 0$, each MDP chooses its initial state-action vector $\theta_0^{(k)}$ resulting from any *separable* randomized stationary policy $\pi_0^{(k)}$. For example, one could choose a uniform policy $\pi^{(k)}(a|s) = 1/\left|\mathcal{A}^{(k)}\right|$, $\forall s \in \mathcal{S}^{(k)}$, solve the equation $d_{\pi_0^{(k)}} = d_{\pi_0^{(k)}}\mathbf{P}_{\pi_0^{(k)}}^{(k)}$ to get a probability vector $d_{\pi_0^{(k)}}$, and obtain $\theta_0^{(k)}(a,s) = d_{\pi_0^{(k)}}(s)/\left|\mathcal{A}^{(k)}\right|$. For each constraint $i \in \{1,2,\cdots,m\}$, let $Q_i(t)$ be a *virtual queue* defined over slots $t = 0,1,2,\cdots$ with the initial condition $Q_i(0) = Q_i(1) = 0$, and update equation:

$$Q_i(t+1) = \max\left\{Q_i(t) + \sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t\right\rangle,\ 0\right\},\ \forall t \in \{1,2,3,\cdots\}. \tag{9}$$

Our algorithm uses two parameters $V > 0$ and $\alpha > 0$ and makes decisions as follows: At the start of each slot $t \in \{1,2,3,\cdots\}$,

- The k-th MDP observes $Q_i(t)$, $i = 1,2,\cdots,m$ and chooses $\theta_t^{(k)}$ to solve the following subproblem:

$$\theta_t^{(k)} = \text{argmin}_{\theta\in\Theta^{(k)}}\left\langle V\mathbf{f}_{t-1}^{(k)} + \sum_{i=1}^{m} Q_i(t)\mathbf{g}_{i,t-1}^{(k)}, \theta\right\rangle + \alpha\left\|\theta - \theta_{t-1}^{(k)}\right\|_2^2. \tag{10}$$

- Construct the randomized stationary policy $\pi_t^{(k)}$ according to (5) with $\theta = \theta_t^{(k)}$, and choose the action $a_t^{(k)}$ at k-th MDP according to the conditional distribution $\pi_t^{(k)}\left(\cdot|s_t^{(k)}\right)$.
- Update the virtual queue $Q_i(t)$ according to (9) for all $i = 1,2,\cdots,m$.

Remark 3.1. *Note that for any slot $t \geq 1$, this algorithm gives a separable randomized stationary policy, so that each MDP chooses its own policy based on its own function $\mathbf{f}_{t-1}^{(k)}$, $\mathbf{g}_{i,t-1}^{(k)}, i \in \{1, 2, \cdots, m\}$, and a common multiplier $\mathbf{Q}(t) := (Q_1(t), \cdots, Q_m(t))$. Furthermore, note that (10) is a convex quadratic program (QP). Standard theory of QP (e.g. [34]) shows that the computation complexity solving (10) is poly $\left(\left|\mathcal{S}^{(k)}\right|\left|\mathcal{A}^{(k)}\right|\right)$ for each k. Thus, the total computation complexity over all MDPs during each round is poly $\left(K\left|\mathcal{S}^{(k)}\right|\left|\mathcal{A}^{(k)}\right|\right)$.*

Remark 3.2. *The quadratic term $\alpha\left\|\theta - \theta_{t-1}^{(k)}\right\|_2^2$ in (10) penalizes the deviation of θ from the previous decision variable $\theta_{t-1}^{(k)}$. Thus, under proper choice of α, the distance between $\theta_t^{(k)}$ and $\theta_{t-1}^{(k)}$ would be very small, which is the slow update condition we need according to Section 2.5.*

The next lemma shows that solving (10) is in fact a projection onto the state-action polyhedron. For any set $\mathcal{X} \in \mathbb{R}^n$ and a vector $\mathbf{y} \in \mathbb{R}^n$, define the projection operator $\mathcal{P}_{\mathcal{X}}(\mathbf{y})$ as

$$\mathcal{P}_{\mathcal{X}}(\mathbf{y}) = \text{arginf}_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x} - \mathbf{y}\|_2.$$

Lemma 3.1. *Fix an $\alpha > 0$ and $t \in \{1, 2, 3, \cdots\}$. The θ_t that solves (10) is*

$$\theta_t^{(k)} = \mathcal{P}_{\Theta^{(k)}}\left(\theta_{t-1}^{(k)} - \frac{\mathbf{w}_t^{(k)}}{2\alpha}\right),$$

where $\mathbf{w}_t^{(k)} = V\mathbf{f}_{t-1}^{(k)} + \sum_{i=1}^m Q_i(t)\mathbf{g}_{i,t-1}^{(k)} \in \mathbb{R}^{|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}$.

Proof. By definition, we have

$$\begin{aligned}
\theta_t^{(k)} =& \text{argmin}_{\theta\in\Theta^{(k)}}\left\langle \mathbf{w}_t^{(k)}, \theta\right\rangle + \alpha\left\|\theta - \theta_{t-1}^{(k)}\right\|_2^2 \\
=& \text{argmin}_{\theta\in\Theta^{(k)}}\left\langle \mathbf{w}_t^{(k)}, \theta - \theta_{t-1}^{(k)}\right\rangle + \alpha\left\|\theta - \theta_{t-1}^{(k)}\right\|_2^2 \\
&+ \left\langle \mathbf{w}_t^{(k)}, \theta_{t-1}^{(k)}\right\rangle \\
=& \text{argmin}_{\theta\in\Theta^{(k)}}\ \alpha\cdot\left(\left\langle \mathbf{w}_t^{(k)}\Big/\alpha, \theta - \theta_{t-1}^{(k)}\right\rangle + \left\|\theta - \theta_{t-1}^{(k)}\right\|_2^2\right) \\
&+ \left\langle \mathbf{w}_t^{(k)}, \theta_{t-1}^{(k)}\right\rangle \\
=& \text{argmin}_{\theta\in\Theta^{(k)}}\ \alpha\cdot\left\|\theta - \theta_{t-1}^{(k)} + \mathbf{w}_t^{(k)}\Big/2\alpha\right\|_2^2 \\
=& \mathcal{P}_{\Theta^{(k)}}\left(\theta_{t-1}^{(k)} - \mathbf{w}_t^{(k)}\Big/2\alpha\right),
\end{aligned}$$

finishing the proof. □

3.1. Intuition of the algorithm and roadmap of analysis

The intuition of this algorithm follows from the discussion in Section 2.5. Instead of the Markovian regret (1) and constraint set (2), we work on the imaginary system that after the decision maker chooses any joint policy Π_t and the penalty/constraint functions are revealed, the K parallel Markov chains reach stationary state distribution right away, with state-action probability vectors $\left\{\theta_t^{(k)}\right\}_{k=1}^K$ for K parallel MDPs. Thus there is no Markov state in such a system

anymore and the corresponding stationary penalty and constraint function value at time t can be expressed as $\sum_{k=1}^{K}\left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)}\right\rangle$ and $\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t}^{(k)}, \theta_t^{(k)}\right\rangle$, $i = 1, 2, \cdots, m$, respectively. As a consequence, we are now facing a relatively easier task of minimizing the following regret:

$$\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\Big(\Big\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)}\Big\rangle\Big) - \sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\Big(\Big\langle \mathbf{f}_t^{(k)}, \theta_*^{(k)}\Big\rangle\Big), \tag{11}$$

where $\left\{\theta_*^{(k)}\right\}_{k=1}^{K}$ are the state-action probabilities corresponding to the best fixed joint randomized stationary policy within the following stationary constraint set

$$\overline{\mathcal{G}} := \left\{\theta^{(k)} \in \Theta^{(k)},\ k \in \{1, 2, \cdots, K\} : \sum_{k=1}^{K}\left\langle \mathbb{E}\Big(\mathbf{g}_{i,t}^{(k)}\Big), \theta^{(k)}\right\rangle \le 0,\ i = 1, 2, \cdots, m\right\}, \tag{12}$$

with the assumption that Slater's condition (8) holds.

To analyze the proposed algorithm, we need to tackle the following two major challenges:

- Whether or not the policy decision of the proposed algorithm would yield $\mathcal{O}(\sqrt{T})$ regret and constraint violation on the imaginary system that reaches steady state instantaneously on each slot.
- Whether the error between the imaginary and true systems can be bounded by $\mathcal{O}(\sqrt{T})$.

In the next section, we answer these questions via a multi-stage analysis piecing together the results of MDPs from Section 2.4 with multiple ingredients from convex analysis and stochastic queue analysis. We first show the $\mathcal{O}(\sqrt{T})$ regret and constraint violation in the imaginary online linear program incorporating a new regret analysis procedure with a stochastic drift analysis for queue processes. Then, we show if the benchmark randomized stationary algorithm always starts from its stationary state, then, the discrepancy of regrets between the imaginary and true systems can be controlled via the slow-update property of the proposed algorithm together with the properties of MDPs developed in Section 2.4. Finally, for the problem with arbitrary non-stationary starting state, we reformulate it as a perturbation on the aforementioned stationary state problem and analyze the perturbation via Farkas' Lemma.

4. Convergence time analysis

4.1. Stationary state performance: An online linear program

Let $\mathbf{Q}(t) := [Q_1(t),\ Q_2(t),\ \cdots,\ Q_m(t)]$ be the virtual queue vector and $L(t) = \frac{1}{2}\|\mathbf{Q}(t)\|_2^2$. Define the drift $\Delta(t) := L(t+1) - L(t)$.

4.1.1. Sample-path analysis

This section develops a couple of bounds given a sequence of penalty functions $f_0^{(k)}, f_1^{(k)}, \cdots, f_{T-1}^{(k)}$ and constraint functions $g_{i,0}^{(k)}, g_{i,1}^{(k)}, \cdots, g_{i,T-1}^{(k)}$. The following lemma provides bounds for virtual queue processes:

Lemma 4.1. *For any $i \in \{1, 2, \cdots, m\}$ at $T \in \{1, 2, \cdots\}$, the following holds under the virtual queue update (9),*

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_{t-1}^{(k)}\right\rangle \le Q_i(T+1) - Q_i(1) + \Psi\sum_{t=1}^{T}\sum_{k=1}^{K}\sqrt{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|}\left\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\right\|_2,$$

where $\Psi > 0$ is the constant defined in (4).

Proof. By the queue updating rule (9), for any $t \in \mathbb{N}$,

$$\begin{aligned}
&Q_i(t+1)\\
=&\max\left\{Q_i(t)+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle,0\right\}\\
\geq&Q_i(t)+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle\\
=&Q_i(t)+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_{t-1}^{(k)}\right\rangle+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\rangle\\
\geq&Q_i(t)+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_{t-1}^{(k)}\right\rangle-\sum_{k=1}^{K}\left\|g_{i,t-1}^{(k)}\right\|_2\left\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\|_2,
\end{aligned}$$

Note that the constraint functions are deterministically bounded,

$$\left\|g_{i,t-1}^{(k)}\right\|_2^2\leq\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi^2.$$

Substituting this bound into the above queue bound and rearranging the terms finish the proof. □

The next lemma provides a bound for the drift $\Delta(t)$.

Lemma 4.2. *For any slot $t \geq 1$, we have*

$$\Delta(t)\leq\frac{1}{2}mK^2\Psi^2+\sum_{i=1}^{m}Q_i(t)\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle.$$

Proof. By definition, we have

$$\begin{aligned}
\Delta(t)=&\frac{1}{2}\|\mathbf{Q}(t+1)\|_2^2-\frac{1}{2}\|\mathbf{Q}(t)\|_2^2\\
\leq&\frac{1}{2}\sum_{i=1}^{m}\left(\left(Q_i(t)+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle\right)^2-Q_i(t)^2\right)\\
=&\sum_{i=1}^{m}Q_i(t)\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle+\frac{1}{2}\sum_{i=1}^{m}\left(\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle\right)^2.
\end{aligned}$$

Note that by the queue update (9), we have

$$\left|\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)},\theta_t^{(k)}\right\rangle\right|\leq K\left\|\mathbf{g}_{i,t-1}^{(k)}\right\|_\infty\left\|\theta_t^{(k)}\right\|_1\leq K\Psi.$$

Substituting this bound into the drift bound finishes the proof. □

Consider a convex set $\mathcal{X}\subseteq\mathbb{R}^n$. Recall that for a fixed real number $c>0$, a function $h:\mathcal{X}\to\mathbb{R}$ is said to be *c-strongly convex*, if $h(x)-\frac{c}{2}\|x\|_2^2$ is convex over $x\in\mathcal{X}$. It is easy to see that if $q:\mathcal{X}\to\mathbb{R}$ is convex, $c>0$ and $b\in\mathbb{R}^n$, the function $q(x)+\frac{c}{2}\|x-b\|_2^2$ is c-strongly convex.

Furthermore, if the function h is c-strongly convex that is minimized at a point $x_{\min} \in \mathcal{X}$, then (see, e.g., Corollary 1 in [35]):

$$h(x_{\min}) \leq h(y) - \frac{c}{2}\|y - x_{\min}\|_2^2, \quad \forall y \in \mathcal{X}. \tag{13}$$

The following lemma is a direct consequence of the above strongly convex result. It also demonstrates the key property of our minimization subproblem (10).

Lemma 4.3. *The following bound holds for any $k \in \{1, 2, \cdots, K\}$ and any fixed $\theta_*^{(k)} \in \Theta^{(k)}$:*

$$\begin{aligned} &V\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \sum_{i=1}^{m} Q_i(t) \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t^{(k)} \right\rangle + \alpha\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \\ &\leq V\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \sum_{i=1}^{m} Q_i(t) \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle + \alpha\|\theta_*^{(k)} - \theta_{t-1}^{(k)}\|_2^2 - \alpha\|\theta_*^{(k)} - \theta_t^{(k)}\|_2^2. \end{aligned} \tag{14}$$

This lemma follows easily from the fact that the proposed algorithm (10) gives $\theta_t^{(k)} \in \Theta^{(k)}$ minimizing the left hand side, which is a strongly convex function, and then, applying (13), with

$$h\left(\theta_*^{(k)}\right) = V\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \sum_{i=1}^{m} Q_i(t) \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle + \alpha \left\|\theta_*^{(k)} - \theta_{t-1}^{(k)}\right\|_2^2$$

Combining the previous two lemmas gives the following "drift-plus-penalty" bound.

Lemma 4.4. *For any fixed $\{\theta_*^{(k)}\}_{k=1}^K$ such that $\theta_*^{(k)} \in \Theta^{(k)}$ and $t \in \mathbb{N}$, we have the following bound,*

$$\begin{aligned} &\Delta(t) + V\sum_{k=1}^{K}\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \alpha\sum_{k=1}^{K}\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \\ &\qquad \leq \frac{3}{2}mK^2\Psi^2 + V\sum_{k=1}^{K}\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \sum_{i=1}^{m} Q_i(t-1) \\ &\qquad\qquad \cdot \sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle + \alpha\sum_{k=1}^{K}\|\theta_*^{(k)} - \theta_{t-1}^{(k)}\|_2^2 - \alpha\sum_{k=1}^{K}\|\theta_*^{(k)} - \theta_t^{(k)}\|_2^2 \end{aligned} \tag{15}$$

Proof. Using Lemma 4.2 and then Lemma 4.3, we obtain

$$\begin{aligned} &\Delta(t) + V\sum_{k=1}^{K}\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \alpha\sum_{k=1}^{K}\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \\ \leq &\frac{1}{2}mK^2\Psi^2 + \sum_{i=1}^{m} Q_i(t)\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t^{(k)} \right\rangle + V\sum_{k=1}^{K}\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \alpha\sum_{k=1}^{K}\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \\ \leq &\frac{1}{2}mK^2\Psi^2 + \sum_{k=1}^{K}\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \sum_{i=1}^{m} Q_i(t)\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle + \alpha\sum_{k=1}^{K}\|\theta_*^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \\ &- \alpha\sum_{k=1}^{K}\|\theta_*^{(k)} - \theta_t^{(k)}\|_2^2. \end{aligned} \tag{16}$$

Note that by the queue updating rule (9), we have for any $t \geq 2$,

$$|Q_i(t) - Q_i(t-1)| \leq \left| \sum_{k=1}^{K} \left\langle \mathbf{g}_{i,t-2}^{(k)}, \theta_{t-1}^{(k)} \right\rangle \right| \leq K \left\| \mathbf{g}_{i,t-2}^{(k)} \right\|_\infty \left\| \theta_{t-1}^{(k)} \right\|_1 \leq K\Psi,$$

and for $t = 1$, $Q_i(t) - Q_i(t-1) = 0$ by the initial condition of the algorithm. Also, we have for any $\theta_*^{(k)} \in \Theta^{(k)}$,

$$\left| \sum_{k=1}^{K} \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle \right| \leq K \left\| \mathbf{g}_{i,t-2}^{(k)} \right\|_\infty \left\| \theta_*^{(k)} \right\|_1 \leq K\Psi.$$

Thus, we have

$$\sum_{i=1}^{m} Q_i(t) \sum_{k=1}^{K} \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle \leq \sum_{i=1}^{m} Q_i(t-1) \sum_{k=1}^{K} \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle + mK^2\Psi^2.$$

Substituting this bound into (16) finishes the proof. □

4.1.2. Objective bound

Theorem 4.1. *For any $\{\theta_*^{(k)}\}_{k=1}^K$ in the constraint set (12) and any $T \in \{1, 2, 3, \cdots\}$, the proposed algorithm has the following stationary state performance bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)} \right\rangle \right) \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{f}_t^{(k)}, \theta_*^{(k)} \right\rangle \right) + \frac{2\alpha K}{TV} + \frac{mK^2\Psi^2}{T} + \frac{V\Psi^2}{2\alpha} \sum_{k=1}^{K} \left|\mathcal{S}^{(k)}\right| \left|\mathcal{A}^{(k)}\right| + \frac{3}{2}\frac{mK^2\Psi^2}{V},$$

In particular, choosing $\alpha = T$ and $V = \sqrt{T}$ gives the $\mathcal{O}(\sqrt{T})$ regret

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)} \right\rangle \right) \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{f}_t^{(k)}, \theta_*^{(k)} \right\rangle \right) + \left(2K + \frac{\Psi^2}{2} \sum_{k=1}^{K} \left|\mathcal{S}^{(k)}\right| \left|\mathcal{A}^{(k)}\right| + \frac{5}{2} mK^2\Psi^2 \right) \frac{1}{\sqrt{T}}.$$

Proof. First of all, note that $\{\mathbf{g}_{i,t-1}^{(k)}\}_{k=1}^K$ is i.i.d. and independent of all system history up to $t-1$, and thus independent of $Q_i(t-1)$, $i = 1, 2, \cdots, m$. We have

$$\mathbb{E}\left(Q_i(t-1) \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle \right) = \mathbb{E}(Q_i(t-1)) \mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_*^{(k)} \right\rangle \right) \leq 0 \tag{17}$$

where the last inequality follows from the assumption that $\{\theta_*^{(k)}\}_{k=1}^K$ is in the constraint set (12). Substituting $\theta_*^{(k)}$ into (15), taking expectation with respect to both sides and using (17) give

$$\mathbb{E}(\Delta(t)) + V\mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle \right) + \alpha\mathbb{E}\left(\sum_{k=1}^{K} \|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \right)$$
$$\leq \frac{3}{2} mK^2\Psi^2 + V\mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} - \theta_{t-1}^{(k)} \right\rangle \right) + \alpha\mathbb{E}\left(\sum_{k=1}^{K} \|\theta_*^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \right) - \alpha\mathbb{E}\left(\sum_{k=1}^{K} \|\theta_*^{(k)} - \theta_t^{(k)}\|_2^2 \right),$$

where the second inequality follows from (17). Note that for any k, completing the squares gives

$$V\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \alpha\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \geq \left\| \sqrt{\frac{\alpha}{2}} \left(\theta_t^{(k)} - \theta_{t-1}^{(k)}\right) + \frac{V}{2\sqrt{\alpha/2}} \mathbf{f}_{t-1}^{(k)} \right\|_2^2 - \frac{V^2\Psi^2 \left|\mathcal{S}^{(k)}\right| \left|\mathcal{A}^{(k)}\right|}{2\alpha}.$$

Substituting this inequality into the previous bound and rearranging the terms give

$$V\mathbb{E}\left(\sum_{k=1}^K \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_{t-1}^{(k)} \right\rangle\right) \leq V\mathbb{E}\left(\sum_{k=1}^K \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} \right\rangle\right) - \mathbb{E}(\Delta(t)) + \frac{V^2 \sum_{k=1}^K \Psi^2 \left|\mathcal{S}^{(k)}\right| \left|\mathcal{A}^{(k)}\right|}{2\alpha} + \frac{3}{2} mK^2\Psi^2 + \alpha\mathbb{E}\left(\sum_{k=1}^K \|\theta_*^{(k)} - \theta_{t-1}^{(k)}\|_2^2\right) - \alpha\mathbb{E}\left(\sum_{k=1}^K \|\theta_*^{(k)} - \theta_t^{(k)}\|_2^2\right).$$

Taking telescoping sums from 1 to T and dividing both sides by TV gives,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\sum_{k=1}^K \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_{t-1}^{(k)} \right\rangle\right) \leq & \mathbb{E}\left(\sum_{k=1}^K \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} \right\rangle\right) + \frac{L(0) - L(T+1)}{VT} + \frac{V\sum_{k=1}^K \Psi^2 \left|\mathcal{S}^{(k)}\right| \left|\mathcal{A}^{(k)}\right|}{2\alpha} \\ & + \frac{3}{2}\frac{mK^2\Psi^2}{V} + \frac{\alpha\mathbb{E}\left(\sum_{k=1}^K \|\theta_*^{(k)} - \theta_{T-1}^{(k)}\|_2^2\right) - \alpha\mathbb{E}\left(\sum_{k=1}^K \|\theta_*^{(k)} - \theta_T^{(k)}\|_2^2\right)}{VT} \\ \leq & \mathbb{E}\left(\sum_{k=1}^K \left\langle \mathbf{f}_{t-1}^{(k)}, \theta_*^{(k)} \right\rangle\right) + \frac{V\sum_{k=1}^K \Psi^2 \left|\mathcal{S}^{(k)}\right| \left|\mathcal{A}^{(k)}\right|}{2\alpha} + \frac{3}{2}\frac{mK^2\Psi^2}{V} + \frac{2\alpha K}{VT},\end{aligned}$$

where we use the fact that $L(0) = 0$ and $\|\theta_*^{(k)} - \theta_{T-1}^{(k)}\|_2^2 \leq \|\theta_*^{(k)} - \theta_{T-1}^{(k)}\|_1 \leq 2$. □

4.1.3. A drift lemma and its implications

From Lemma 4.1, we know that in order to get the constraint violation bound, we need to look at the size of the virtual queue $Q_i(T+1),\ i = 1, 2, \cdots, m$. The following drift lemma serves as a cornerstone for our goal.

Lemma 4.5 (Lemma 5 of [15]). *Let $\{\Omega, \mathcal{F}, P\}$ be a probability space. Let $\{Z(t), t \geq 1\}$ be a discrete time stochastic process adapted to a filtration $\{\mathcal{F}_{t-1}, t \geq 1\}$ with $Z(1) = 0$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Suppose there exist integer $t_0 > 0$, real constants $\lambda \in \mathbb{R}$, $\delta_{\max} > 0$ and $0 < \zeta \leq \delta_{\max}$ such that*

$$|Z(t+1) - Z(t)| \leq \delta_{\max}, \tag{18}$$

$$\mathbb{E}[Z(t+t_0) - Z(t)|\mathcal{F}_{t-1}] \leq \begin{cases} t_0\delta_{\max}, & \text{if } Z(t) < \lambda \\ -t_0\zeta, & \text{if } Z(t) \geq \lambda \end{cases}. \tag{19}$$

hold for all $t \in \{1, 2, \ldots\}$. Then, the following holds:

$$\mathbb{E}[Z(t)] \leq \lambda + t_0\delta_{\max} + t_0\frac{4\delta_{\max}^2}{\zeta}\log\left[\frac{8\delta_{\max}^2}{\zeta^2}\right], \forall t \in \{1, 2, \ldots\}.$$

Note that a special case of above drift lemma for $t_0 = 1$ dates back to the seminal paper of Hajek ([36]) bounding the size of a random process with strongly negative drift. Since then, its

power has been demonstrated in various scenarios ranging from steady state queue bound ([37]) to feasibility analysis of stochastic optimization ([38]). The current generalization to a multi-step drift is first considered in [15].

This lemma is useful in the current context due to the following lemma, whose proof can be found in Appendix A.2.

Lemma 4.6. *Let $\mathcal{F}_t$, $t \geq 1$ be the system history functions up to time t, including $f_0^{(k)}, \cdots, f_{t-1}^{(k)}$, $g_{0,i}^{(k)}, \cdots, g_{t-1,i}^{(k)}$, $i = 1, 2, \cdots, m$, $k = 1, 2, \cdots, K$, and $\mathcal{F}_0$ is a null set. Let t_0 be an arbitrary positive integer, then, we have*

$$\big|\|\mathbf{Q}(t+1)\|_2 - \|\mathbf{Q}(t)\|_2\big| \leq \sqrt{m}K\Psi,$$

$$\mathbb{E}[\|\mathbf{Q}(t+t_0)\|_2 - \|\mathbf{Q}(t)\|_2 \big| \mathcal{F}(t-1)] \leq \begin{cases} t_0\sqrt{m}K\Psi, & \text{if } \|\mathbf{Q}(t)\| < \lambda \\ -t_0\frac{\eta}{2}, & \text{if } \|\mathbf{Q}(t)\| \geq \lambda \end{cases}$$

where $\lambda = \frac{8VK\Psi + 3mK^2\Psi^2 + 4K\alpha + t_0(t_0-1)m\Psi + 2mK\Psi\eta t_0 + \eta^2 t_0^2}{\eta t_0}$.

Combining the previous two lemmas gives the virtual queue bound as

$$\mathbb{E}(\|\mathbf{Q}(t)\|_2) \leq \frac{8VK\Psi + 3mK^2\Psi^2 + 4K\alpha + t_0(t_0-1)m\Psi + 2mK\Psi\eta t_0 + \eta^2 t_0^2}{\eta t_0} + t_0\sqrt{m}K\Psi + \frac{4t_0 mK^2\Psi^2}{\eta}\log\big[\frac{8mK^2\Psi^2}{\eta^2}\big].$$

We then choose $t_0 = \sqrt{T}$, $V = \sqrt{T}$ and $\alpha = T$, which implies that

$$\mathbb{E}(\|\mathbf{Q}(t)\|_2) \leq C(m, K, \Psi, \eta)\sqrt{T}, \tag{20}$$

where $C(m, K, \Psi, \eta) = \frac{8K\Psi}{\eta} + \frac{3mK^2\Psi^2}{\eta^2} + \frac{4K+m\Psi}{\eta} + 2mK\Psi + \eta + \sqrt{m}K\Psi + \frac{4mK^2\Psi^2}{\eta}\log\big[\frac{8mK^2\Psi^2}{\eta^2}\big]$.

4.1.4. The slow-update condition and constraint violation

In this section, we prove the slow-update property of the proposed algorithm, which not only implies the the $\mathcal{O}(\sqrt{T})$ constraint violation bound, but also plays a key role in Markov analysis.

Lemma 4.7. *The sequence of state-action vectors $\theta_t^{(k)}$, $t \in \{1, 2, \cdots, T\}$ satisfies*

$$\mathbb{E}\Big(\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2\Big) \leq \frac{\sqrt{m|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi\mathbb{E}(\|\mathbf{Q}(t)\|_2)}{2\alpha} + \frac{\sqrt{|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi V}{2\alpha}.$$

In particular, choosing $V = \sqrt{T}$ and $\alpha = T$ gives a slow-update condition

$$\mathbb{E}\Big(\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2\Big) \leq \frac{\sqrt{|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi + C\sqrt{m|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi}{2\sqrt{T}}, \tag{21}$$

where $C = C(m, K, \Psi, \eta)$ is defined in (20).

Proof of Lemma 4.7. First, choosing $\theta = \theta_{t-1}$ in (14) gives

$$V\left\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)} - \theta_{t-1}^{(k)} \right\rangle + \sum_{i=1}^m Q_i(t)\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t^{(k)} \right\rangle + \alpha\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\|_2^2 \leq \sum_{i=1}^m Q_i(t)\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_{t-1}^{(k)}\rangle - \alpha\|\theta_{t-1}^{(k)} - \theta_t^{(k)}\|_2^2.$$

Rearranging the terms gives

$$\begin{aligned}2\alpha\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\|_2^2 \leq & -V\langle \mathbf{f}_{t-1}^{(k)}, \theta_t^{(k)}-\theta_{t-1}^{(k)}\rangle - \sum_{i=1}^m Q_i(t)\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t^{(k)}-\theta_{t-1}^{(k)}\rangle \\ \leq & V\|\mathbf{f}_{t-1}^{(k)}\|_2\cdot\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\|_2 + \sum_{i=1}^m Q_i(t)\|\mathbf{g}_{i,t-1}^{(k)}\|_2\cdot\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\|_2 \\ \leq & V\|\mathbf{f}_{t-1}\|_2\cdot\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\|_2 + \|\mathbf{Q}(t)\|_2\sqrt{\sum_{i=1}^m \|\mathbf{g}_{i,t-1}^{(k)}\|_2^2}\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\|_2,\end{aligned}$$

where the second and third inequality follow from Cauchy-Schwarz inequality. Thus, it follows

$$\left\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\|_2 \leq \frac{V\|\mathbf{f}_{t-1}^{(k)}\|_2 + \|\mathbf{Q}(t)\|_2\cdot\sqrt{\sum_{i=1}^m\|\mathbf{g}_{i,t-1}^{(k)}\|_2^2}}{2\alpha}.$$

Applying the fact that $\|\mathbf{f}_{t-1}^{(k)}\|_2 \leq \sqrt{|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi$, $\|\mathbf{g}_{i,t-1}^{(k)}\|_2 \leq \sqrt{|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi$ and taking expectation from both sides give the first bound in the lemma. The second bound follows directly from the first bound by further substituting (20). □

Theorem 4.2. *The proposed algorithm has the following stationary state constraint violation bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left(\sum_{k=1}^K\left\langle \mathbf{g}_{i,t}^{(k)}, \theta_t^{(k)}\right\rangle\right) \leq \frac{1}{\sqrt{T}}\left(C + \sum_{k=1}^K\sqrt{m|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi C + \sum_{k=1}^K|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|\Psi^2\right),$$

where $C = C(m, K, \Psi, \eta)$ is defined in (20).

Proof. Taking expectation from both sides of Lemma 4.1 gives

$$\sum_{t=1}^T\mathbb{E}\left(\sum_{k=1}^K\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_{t-1}^{(k)}\right\rangle\right) \leq \mathbb{E}(Q_i(T+1)) + \Psi\sum_{t=1}^T\sum_{k=1}^K\sqrt{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|}\mathbb{E}\left(\left\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\|_2\right).$$

Substituting the bounds (20) and (21) in to the above inequality gives the desired result. □

4.2. Markov analysis

So far, we have shown that our algorithm achieves an $\mathcal{O}(\sqrt{T})$ regret and constraint violation simultaneously regarding the stationary online linear program (11) with constraint set given by (12) in the imaginary system. In this section, we show how these stationary state results lead to a tight performance bound on the original true online MDP problem (1) and (2) comparing to any joint randomized stationary algorithm starting from its stationary state.

4.2.1. Approximate mixing of MDPs

Let $\mathcal{F}_t$, $t \geq 1$ be the set of system history functions up to time t, including $f_0^{(k)}, \cdots, f_{t-1}^{(k)}$, $g_{0,i}^{(k)}, \cdots, g_{t-1,i}^{(k)}$, $i = 1, 2, \cdots, m$, $k = 1, 2, \cdots, K$, and $\mathcal{F}_0$ is a null set. Let $d_{\pi_t^{(k)}}$ be the stationary state distribution at k-th MDP under the randomized stationary policy $\pi_t^{(k)}$ in the proposed algorithm. Let $v_t^{(k)}$ be the true state distribution at time slot t under the proposed algorithm given

the function path $\mathcal{F}_T$ and starting state $d_0^{(k)}$, i.e. for any $s \in \mathcal{S}^{(k)}$, $v_t^{(k)}(s) := Pr\left(s_t^{(k)} = s|\mathcal{F}_T\right)$ and $v_0^{(k)} = d_0^{(k)}$.

The following lemma provides a key estimate on the distance between stationary distribution and true distribution at each time slot t. It builds upon the slow-update condition (Lemma 4.7) of the proposed algorithm and uniform mixing bound of general MDPs (Lemma 2.1).

Lemma 4.8. *Consider the proposed algorithm with $V = \sqrt{T}$ and $\alpha = T$. For any initial state distribution $\{d_0^{(k)}\}_{k=1}^K$ and any $t \in \{0, 1, 2, \cdots, T-1\}$, we have*

$$\mathbb{E}\left(\left\|d_{\pi_t^{(k)}} - v_t^{(k)}\right\|_1\right) \leq \tau r \left(\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi + C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi\right)\Big/ 2\sqrt{T} + 2e^{-\frac{t}{\tau r}+1},$$

where τ and r are mixing parameters defined in Lemma 2.1 and C is an absolute constant defined in (20).

Proof of Lemma 4.8. By Lemma 4.7 we know that for any $t \in \{1, 2, \cdots, T\}$,

$$\mathbb{E}\left(\left\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\right\|_2\right) \leq \frac{\sqrt{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|}\Psi + C\sqrt{m\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|}\Psi}{2\sqrt{T}},$$

Thus,

$$\mathbb{E}\left(\left\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\right\|_1\right) \leq \frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi + C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}},$$

Since for any $s \in \mathcal{S}^{(k)}$, $\left|d_{\pi_t^{(k)}}(s) - d_{\pi_{t-1}^{(k)}}(s)\right| = \left|\sum_{a\in\mathcal{A}^{(k)}} \theta_t^{(k)}(a,s) - \theta_{t-1}^{(k)}(a,s)\right| \leq \sum_{a\in\mathcal{A}^{(k)}} \left|\theta_t^{(k)}(a,s) - \theta_{t-1}^{(k)}(a,s)\right|$, it then follows

$$\mathbb{E}\left(\left\|d_{\pi_t^{(k)}} - d_{\pi_{t-1}^{(k)}}\right\|_1\right) \leq \mathbb{E}\left(\left\|\theta_t^{(k)} - \theta_{t-1}^{(k)}\right\|_1\right) \leq \frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi + C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}}. \quad (22)$$

Now, we use the above relation to bound $\mathbb{E}\left(\left\|d_{\pi_t^{(k)}} - v_t^{(k)}\right\|_1\right)$ for any $t \geq r$.

$$\begin{aligned}
\mathbb{E}\left(\left\|d_{\pi_t^{(k)}} - v_t^{(k)}\right\|_1\right) \leq& \mathbb{E}\left(\left\|d_{\pi_t^{(k)}} - d_{\pi_{t-1}^{(k)}}\right\|_1\right) + \mathbb{E}\left(\left\|d_{\pi_{t-1}^{(k)}} - v_t^{(k)}\right\|_1\right) \\
\leq& \frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi + C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}} + \mathbb{E}\left(\left\|d_{\pi_{t-1}^{(k)}} - v_t^{(k)}\right\|_1\right) \\
=& \frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi + C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}} + \mathbb{E}\left(\left\|\left(d_{\pi_{t-1}^{(k)}} - v_{t-1}^{(k)}\right)\mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)}\right\|_1\right),
\end{aligned} \quad (23)$$

where the second inequality follows from the slow-update condition (22) and the final equality follows from the fact that given the function path $\mathcal{F}_T$, the following holds

$$d_{\pi_{t-1}^{(k)}} - v_t^{(k)} = \left(d_{\pi_{t-1}^{(k)}} - v_{t-1}^{(k)}\right)\mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)}. \quad (24)$$

To see this, note that from the proposed algorithm, the policy $\pi_t^{(k)}$ is determined by $\mathcal{F}_T$. Thus, by definition of stationary distribution, given $\mathcal{F}_T$, we know that $d_{\pi_{t-1}^{(k)}} = d_{\pi_{t-1}^{(k)}} \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)}$, and it is enough to show that given $\mathcal{F}_T$,

$$v_t^{(k)} = v_{t-1}^{(k)} \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)}.$$

First of all, the state distribution $v_t^{(k)}$ is determined by $v_{t-1}^{(k)}$, $\pi_{t-1}^{(k)}$ and probability transition from s_{t-1} to s_t, which are in turn determined by $\mathcal{F}_T$. Thus, given $\mathcal{F}_T$, for any $s \in \mathcal{S}^{(k)}$,

$$v_t^{(k)}(s) = \sum_{s' \in \mathcal{S}^{(k)}} Pr(s_t = s | s_{t-1} = s', \mathcal{F}_T) v_{t-1}^{(k)}(s'),$$

and

$$\begin{aligned} Pr(s_t = s | s_{t-1} = s', \mathcal{F}_T) &= \sum_{a \in \mathcal{A}^{(k)}} Pr(s_t = s | a_t = a, s_{t-1} = s', \mathcal{F}_T) Pr(a_t = a | s_{t-1} = s', \mathcal{F}_T) \\ &= \sum_{a \in \mathcal{A}^{(k)}} P_a(s', s) Pr(a_t = a | s_{t-1} = s', \mathcal{F}_T) \\ &= \sum_{a \in \mathcal{A}^{(k)}} P_a(s', s) \pi_{t-1}^{(k)}(a | s') = P_{\pi_{t-1}^{(k)}}(s', s), \end{aligned}$$

where the second inequality follows from the Assumption 2.2, the third equality follows from the fact that $\pi_{t-1}^{(k)}$ is determined by $\mathcal{F}_T$, thus, for any t,

$$\pi_t^{(k)}(a|s') = Pr(a_t = a | s_{t-1} = s', \mathcal{F}_T), \ \forall a \in \mathcal{A}^{(k)}, \ s' \in \mathcal{S}^{(k)},$$

and the last equality follows from the definition of transition probability (3). This gives

$$v_t^{(k)}(s) = \sum_{s' \in \mathcal{S}^{(k)}} P_{\pi_{t-1}^{(k)}}(s', s) v_{t-1}^{(k)}(s'),$$

and thus (24) holds.

We can iteratively apply the procedure (23) r times as follows

$$\begin{aligned} &\mathbb{E}\left(\left\| d_{\pi_t^{(k)}} - v_t^{(k)} \right\|_1 \right) \\ \leq &\frac{\left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi + C\sqrt{m} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi}{2\sqrt{T}} + \mathbb{E}\left(\left\| \left(d_{\pi_{t-1}^{(k)}} - d_{\pi_{t-2}^{(k)}} \right) \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)} \right\|_1 \right) + \mathbb{E}\left(\left\| \left(d_{\pi_{t-2}^{(k)}} - v_{t-1}^{(k)} \right) \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)} \right\|_1 \right) \\ \leq &2 \cdot \frac{\left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi + C\sqrt{m} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi}{2\sqrt{T}} + \mathbb{E}\left(\left\| \left(d_{\pi_{t-2}^{(k)}} - v_{t-1}^{(k)} \right) \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)} \right\|_1 \right) \\ = &2 \cdot \frac{\left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi + C\sqrt{m} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi}{2\sqrt{T}} + \mathbb{E}\left(\left\| \left(d_{\pi_{t-2}^{(k)}} - v_{t-2}^{(k)} \right) \mathbf{P}_{\pi_{t-2}^{(k)}}^{(k)} \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)} \right\|_1 \right) \\ \leq &\cdots \leq r \cdot \frac{\left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi + C\sqrt{m} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \Psi}{2\sqrt{T}} + \mathbb{E}\left(\left\| \left(d_{\pi_{t-r}^{(k)}} - v_{t-r}^{(k)} \right) \mathbf{P}_{\pi_{t-r}^{(k)}}^{(k)} \cdots \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)} \right\|_1 \right), \end{aligned}$$

where the second inequality follows from the nonexpansive property in ℓ_1 norm of the stochastic matrix $\mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)}$ that

$$\left\| \left(d_{\pi_{t-1}^{(k)}} - d_{\pi_{t-2}^{(k)}} \right) \mathbf{P}_{\pi_{t-1}^{(k)}}^{(k)} \right\|_1 \leq \left\| d_{\pi_{t-1}^{(k)}} - d_{\pi_{t-2}^{(k)}} \right\|_1,$$

and then using the slow-update condition (22) again. By Lemma 2.1, we have

$$\mathbb{E}\left(\left\|d_{\pi_t^{(k)}}-v_t^{(k)}\right\|_1\right)\leq r\cdot\frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi+C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}}+e^{-1/\tau}\mathbb{E}\left(\left\|d_{\pi_{t-r}^{(k)}}-v_{t-r}^{(k)}\right\|_1\right).$$

Iterating this inequality down to $t=0$ gives

$$\begin{aligned}\mathbb{E}\left(\left\|d_{\pi_t^{(k)}}-v_t^{(k)}\right\|_1\right)&\leq\sum_{j=0}^{\lfloor t/\tau\rfloor}e^{-j/\tau}\cdot r\cdot\frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi+C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}}\\&\quad+\mathbb{E}\left(\left\|d_{\pi_0^{(k)}}-v_0^{(k)}\right\|_1\right)e^{-\lfloor t/r\rfloor/\tau}\\&\leq\sum_{j=0}^{\lfloor t/\tau\rfloor}e^{-j/\tau}\cdot r\cdot\frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi+C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}}+2e^{-\lfloor t/r\rfloor/\tau}\\&\leq\int_{x=0}^{\infty}e^{-x/\tau}dx\cdot r\cdot\frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi+C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}}+2e^{-\frac{t}{r\tau}+1}\\&\leq\tau r\cdot\frac{\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi+C\sqrt{m}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\Psi}{2\sqrt{T}}+2e^{-\frac{t}{r\tau}+1}\end{aligned}$$

finishing the proof. □

4.2.2. Benchmarking against policies starting from stationary state

Combining the results derived so far, we have the following regret bound regarding any randomized stationary policy Π starting from its stationary state distribution d_Π such that (d_Π,Π) in the constraint set $\mathcal{G}$ defined in (2).

Theorem 4.3. *Let $\mathscr{P}$ be the sequence of randomized stationary policies resulting from the proposed algorithm with $V=\sqrt{T}$ and $\alpha=T$. Let d_0 be the starting state of the proposed algorithm. For any randomized stationary policy Π starting from its stationary state distribution d_Π such that $(d_\Pi,\Pi)\in\mathcal{G}$, we have*

$$F_T(d_0,\mathscr{P})-F_T(d_\Pi,\Pi)\leq\mathcal{O}\left(m^{3/2}K^2\sum_{k=1}^{K}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\cdot\sqrt{T}\right),$$

$$G_{i,T}(d_0,\mathscr{P})\leq\mathcal{O}\left(m^{3/2}K^2\sum_{k=1}^{K}\left|\mathcal{A}^{(k)}\right|\left|\mathcal{S}^{(k)}\right|\cdot\sqrt{T}\right),\ i=1,2,\cdots,m.$$

Proof of Theorem 4.3. First of all, by Lemma 2.2, for any randomized stationary policy Π, there exists some stationary state-action probability vectors $\{\theta_*^{(k)}\}_{k=1}^K$ such that $\theta_*^{(k)}\in\Theta^{(k)}$,

$$F_T(d_\Pi,\Pi)=\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle\mathbb{E}(\mathbf{f}_t),\theta_*^{(k)}\right\rangle,$$

and $G_{i,T}(d_\Pi,\Pi)=\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle\mathbb{E}(\mathbf{g}_{i,t}),\theta_*^{(k)}\right\rangle$. As a consequence, $(d_\Pi,\Pi)\in\mathcal{G}$ implies $G_{i,T}(d_\Pi,\Pi)=\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle\mathbb{E}(\mathbf{g}_{i,t}),\theta_*^{(k)}\right\rangle\leq 0,\ \forall i\in\{1,2,\cdots,m\}$ and it follows $\{\theta_*^{(k)}\}_{k=1}^K$ is in the imaginary constraint set $\overline{\mathcal{G}}$ defined in (12). Thus, we are in a good shape applying Theorem 4.1 from imaginary systems.

We then split $F_T(d_0, \mathscr{P}) - F_T(d_\Pi, \Pi)$ into two terms:

$$F_T(d_0, \mathscr{P}) - F_T(d_0, \Pi) \leq \underbrace{\left| \mathbb{E}\left(\sum_{t=0}^{T-1} \sum_{k=1}^{K} f_t^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| \ d_0, \mathscr{P} \right) - \sum_{t=0}^{T-1} \sum_{k=1}^{K} \mathbb{E}\left(\left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)} \right\rangle \right) \right|}_{\text{(I)}}$$
$$+ \underbrace{\sum_{t=0}^{T-1} \sum_{k=1}^{K} \left(\mathbb{E}\left(\left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)} \right\rangle \right) - \left\langle \mathbb{E}(\mathbf{f}_t), \theta_*^{(k)} \right\rangle \right)}_{\text{(II)}}.$$

By Theorem 4.1, we get

$$\text{(II)} \leq \left(2K + \frac{\Psi^2}{2} \sum_{k=1}^{K} \left| \mathcal{S}^{(k)} \right| \left| \mathcal{A}^{(k)} \right| + \frac{5}{2} m K^2 \Psi^2 \right) \sqrt{T}. \tag{25}$$

We then bound (I). Consider each time slot $t \in \{0, 1, \cdots, T-1\}$. We have

$$\mathbb{E}\left(\left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)} \right\rangle \right) = \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} \mathbb{E}\left(d_{\pi_t^{(k)}}(s) \pi_t^{(k)}(a|s) f_t^{(k)}(a, s) \right)$$

$$\mathbb{E}\left(f_t^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| \ d_0, \mathscr{P} \right) = \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} \mathbb{E}\left(v_t^{(k)}(s) \pi_t^{(k)}(a|s) f_t^{(k)}(a, s) \right),$$

where the first equality follows from the definition of $\theta_t^{(k)}$ and the second equality follows from the following: Given a specific function path $\mathcal{F}_T$, the policy $\pi_t^{(k)}$ and the true state distribution $v_t^{(k)}$ are fixed. Thus, we have,

$$\mathbb{E}\left(f_t^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| \ d_0, \mathscr{P}, \mathcal{F}_T \right) = \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} v_t^{(k)}(s) \pi_t^{(k)}(a|s) f_t^{(k)}(a, s).$$

Taking the full expectation regarding the function path gives the result. Thus,

$$\begin{aligned}
&\left| \mathbb{E}\left(f_t^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| \ d_0, \mathscr{P} \right) - \mathbb{E}\left(\left\langle \mathbf{f}_t^{(k)}, \theta_t^{(k)} \right\rangle \right) \right| \\
\leq &\left| \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} \mathbb{E}\left(\left(v_t^{(k)}(s) - d_{\pi_t^{(k)}}(s) \right) \pi_t^{(k)}(a|s) \right) \right| \Psi \\
\leq &\mathbb{E}\left(\left\| v_t^{(k)} - d_{\pi_t^{(k)}} \right\|_1 \right) \Psi \\
\leq &\frac{\tau r \left(1 + C\sqrt{m}\right) \left| \mathcal{A}^{(k)} \right| \left| \mathcal{S}^{(k)} \right| \Psi^2}{2\sqrt{T}} + 2e^{-\frac{t}{\tau r} + 1} \Psi
\end{aligned}$$

where the last inequality follows from Lemma 4.8. Thus, it follows,

$$\begin{aligned}
\text{(I)} \leq &\sum_{t=0}^{T-1} \sum_{k=1}^{K} \left(\frac{\tau r \left(1 + C\sqrt{m}\right) \left| \mathcal{A}^{(k)} \right| \left| \mathcal{S}^{(k)} \right| \Psi^2}{2\sqrt{T}} + 2e^{-\frac{t}{\tau r} + 1} \Psi \right) \\
\leq &\sum_{k=1}^{K} \left(\tau r \left(1 + C\sqrt{m}\right) \left| \mathcal{A}^{(k)} \right| \left| \mathcal{S}^{(k)} \right| \Psi^2 \right) \sqrt{T} + 2\Psi K \int_{t=0}^{T-1} e^{-\frac{x}{\tau r} + 1} dx \\
\leq &\tau r \Psi^2 \left(1 + C\sqrt{m}\right) \sum_{k=1}^{K} \left| \mathcal{A}^{(k)} \right| \left| \mathcal{S}^{(k)} \right| \cdot \sqrt{T} + 2e \Psi K \tau r.
\end{aligned} \tag{26}$$

Overall, combining (25),(26) and substituting the constant $C = C(m, K, \Psi, \eta)$ defined in (20) gives the objective regret bound.

For the constraint violation, we have

$$G_{i,T}(d_0, \mathscr{P}) = \underbrace{\mathbb{E}\left(\sum_{t=0}^{T-1} \sum_{k=1}^{K} g_{i,t}^{(k)}(a_t, s_t) \middle| \ d_0, \mathscr{P} \right) - \sum_{t=1}^{T} \sum_{k=1}^{K} \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta_t \right\rangle}_{\text{(IV)}} + \underbrace{\sum_{t=1}^{T} \sum_{k=1}^{K} \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta_t \right\rangle}_{\text{(V)}}.$$

The term (V) can be readily bounded using Theorem 4.2 as

$$\sum_{t=0}^{T-1} \mathbb{E}\left(\sum_{k=1}^{K} \left\langle \mathbf{g}_{i,t}^{(k)}, \theta_t^{(k)} \right\rangle \right) \leq \left(C + \sum_{k=1}^{K} \sqrt{m|\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|}\Psi C + \sum_{k=1}^{K} |\mathcal{A}^{(k)}||\mathcal{S}^{(k)}|\Psi^2 \right) \sqrt{T}.$$

For the term (IV), we have

$$\mathbb{E}\left(\left\langle \mathbf{g}_{i,t}^{(k)}, \theta_t^{(k)} \right\rangle \right) = \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} \mathbb{E}\left(d_{\pi_t^{(k)}}(s) \pi_t^{(k)}(a|s) g_{i,t}^{(k)}(a, s) \right)$$

$$\mathbb{E}\left(g_{i,t}^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| \ d_0, \mathscr{P} \right) = \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} \mathbb{E}\left(v_t^{(k)}(s) \pi_t^{(k)}(a|s) g_{i,t}^{(k)}(a, s) \right),$$

where the first equality follows from the definition of $\theta_t^{(k)}$ and the second equality follows from the following: Given a specific function path $\mathcal{F}_T$, the policy $\pi_t^{(k)}$ and the true state distribution $v_t^{(k)}$ are fixed. Thus, we have,

$$\mathbb{E}\left(g_t^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| \ d_0, \mathscr{P}, \mathcal{F}_T \right) = \sum_{s \in \mathcal{S}^{(k)}} \sum_{a \in \mathcal{A}^{(k)}} v_t^{(k)}(s) \pi_t^{(k)}(a|s) g_t^{(k)}(a, s).$$

Taking the full expectation regarding the function path gives the result. Then, repeat the same proof as that of (26) gives

$$\text{(IV)} \leq \tau r \Psi^2 \left(1 + C\sqrt{m}\right) \sum_{k=1}^{K} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \cdot \sqrt{T} + 2e\Psi K \tau r.$$

This finishes the proof of constraint violation. □

5. A more general regret bound against policies with arbitrary starting state

Recall that Theorem 4.3 compares the proposed algorithm with any randomized stationary policy Π starting from its stationary state distribution d_Π, so that $(d_\Pi, \Pi) \in \mathcal{G}$. In this section, we generalize Theorem 4.3 and obtain a bound of the regret against all $(d_0, \Pi) \in \mathcal{G}$ where d_0 is an arbitrary starting state distribution (not necessarily the stationary state distribution). The main technical difficulty doing such a generalization is as follows: For any randomized stationary policy Π such that $(d_0, \Pi) \in \mathcal{G}$, let $\{\theta_*^{(k)}\}_{k=1}^K$ be the stationary state-action probabilities such that $\theta_*^{(k)} \in \Theta^{(k)}$ and $G_{i,T}(d_\Pi, \Pi) = \sum_{t=0}^{T-1} \sum_{k=1}^{K} \left\langle \mathbb{E}(\mathbf{g}_{i,t}), \theta_*^{(k)} \right\rangle$. For some finite horizon T, there might exist some "low-cost" starting state distribution d_0 such that $G_{i,T}(d_0, \Pi) < G_{i,T}(d_\Pi, \Pi)$ for some $i \in \{1, 2, \cdots, m\}$. As a consequence, one coud have

$$G_{i,T}(d_0, \Pi) \leq 0, \text{ and } \sum_{t=0}^{T-1} \sum_{k=1}^{K} \left\langle \mathbb{E}(\mathbf{g}_{i,t}), \theta_*^{(k)} \right\rangle > 0.$$

This implies although (d_0, Π) is feasible for our true system, its stationary state-action probabilities $\{\theta_*^{(k)}\}_{k=1}^K$ can be *infeasible* with respect to the imaginary constraint set (12), and all our analysis so far fails to cover such randomized stationary policies.

To resolve this issue, we have to "enlarge" the imaginary constraint set (12) so as to cover all state-action probabilities $\{\theta_*^{(k)}\}_{k=1}^K$ arising from any randomized stationary policy Π such that $(d_0, \Pi) \in \mathcal{G}$. But a perturbation of constraint set would result in a perturbation of objective in the imaginary system also. Our main goal in this section is to bound such a perturbation and show that the perturbation bound leads to the final $\mathcal{O}(\sqrt{T})$ regret bound.

5.0.1. A relaxed constraint set

We begin with a supporting lemma on the uniform mixing time bound over all joint randomized stationary policies. The proof is given in Appendix A.3.

Lemma 5.1. *Consider any randomized stationary policy Π in (2) with arbitrary starting state distribution $d_0 \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$. Let $\mathbf{P}_\Pi$ be the corresponding transition matrix on the product state space. Then, the following holds*

$$\left\| (d_0 - d_\Pi) (\mathbf{P}_\Pi)^t \right\|_1 \leq 2e^{(r_1 - t)/r_1}, \forall t \in \{0, 1, 2, \cdots\}, \tag{27}$$

where r_1 is fixed positive constant independent of Π.

The following lemma shows a relaxation of $\mathcal{O}(1/T)$ on the imaginary constraint set (12) is enough to cover all the $\{\theta_*^{(k)}\}_{k=1}^K$ discussed at the beginning of this section. The proof is given in Appendix A.3.

Lemma 5.2. *For any $T \in \{1, 2, \cdots\}$ and any randomized stationary policies Π in (2), with arbitrary starting state distribution $d_0 \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$ and stationary state-action probability $\{\theta_*^{(k)}\}_{k=1}^K$,*

$$\sum_{t=0}^{T-1} \left| \mathbb{E}\left(\sum_{k=1}^K f_t^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| d_0, \Pi \right) - \sum_{k=1}^K \left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)} \right), \theta_*^{(k)} \right\rangle \right| \leq C_1 K \Psi \tag{28}$$

$$\sum_{t=0}^{T-1} \left| \mathbb{E}\left(\sum_{k=1}^K g_{i,t}^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| d_0, \Pi \right) - \sum_{k=1}^K \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)} \right), \theta_*^{(k)} \right\rangle \right| \leq C_1 K \Psi \tag{29}$$

where C_1 is an absolute constant. In particular, $\{\theta_*^{(k)}\}_{k=1}^K$ is contained in the following relaxed constraint set

$$\overline{\mathcal{G}}^+ := \left\{ \theta^{(k)} \in \Theta^{(k)},\ k = 1, 2, \cdots, K : \sum_{k=1}^K \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)} \right), \theta^{(k)} \right\rangle \leq \frac{C_1 K \Psi}{T}, i = 1, 2, \cdots, m \right\}.$$

5.0.2. Best stationary performance over the relaxed constraint set

Recall that the best stationary performance in hindsight over all randomized stationary policies in the constraint set $\overline{\mathcal{G}}$ can be obtained as the minimum achieved by the following linear program.

$$\min \quad \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)}\right), \theta^{(k)}\right\rangle \tag{30}$$

$$s.t. \quad \sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta^{(k)}\right\rangle \leq 0, \quad i=1,2,\cdots,m. \tag{31}$$

On the other hand, if we consider all the randomized stationary policies contained in the original constraint set (2), then, By Lemma 5.2, the relaxed constraint set $\overline{\mathcal{G}}$ contains all such policies and the best stationary performance over this relaxed set comes from the minimum achieved by the following perturbed linear program:

$$\min \quad \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)}\right), \theta^{(k)}\right\rangle \tag{32}$$

$$s.t. \quad \sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta^{(k)}\right\rangle \leq \frac{C_1 K\Psi}{T}, \quad i=1,2,\cdots,m. \tag{33}$$

We aim to show that the minimum achieved by (32)-(33) is not far away from that of (30)-(31). In general, such a conclusion is not true due to the unboundedness of Lagrange multipliers in constrained optimization. However, since Slater's condition holds in our case, the perturbation can be bounded via the following well-known Farkas' lemma ([32]):

Lemma 5.3 (Farkas' Lemma)**.** *Consider a convex program with objective $f(x)$ and constraint function $g_i(x)$, $i=1,2,\cdots,m$:*

$$\min \quad f(x), \tag{34}$$

$$s.t. \quad g_i(x) \leq b_i, \quad i=1,2,\cdots,m, \tag{35}$$

$$x \in \mathcal{X}, \tag{36}$$

for some convex set $\mathcal{X} \subseteq \mathbb{R}^n$. Let x^* be one of the solutions to the above convex program. Suppose there exists $\widetilde{x} \in \mathcal{X}$ such that $g_i(\widetilde{x}) < 0$, $\forall i \in \{1,2,\cdots,m\}$. Then, there exists a separation hyperplane parametrized by $(1,\mu_1,\mu_2,\cdots,\mu_m)$ such that $\mu_i \geq 0$ and

$$f(x)+\sum_{i=1}^{m}\mu_i g_i(x) \geq f(x^*)+\sum_{i=1}^{m}\mu_i b_i, \quad \forall x \in \mathcal{X}.$$

The parameter $\mu=(\mu_1,\mu_2,\cdots,\mu_m)$ is usually referred to as a Lagrange multiplier. From the geometric perspective, Farkas' Lemma states that if Slater's condition holds, then, there exists a non-vertical separation hyperplane supported at $\left(f(x^*),b_1,\cdots,b_m\right)$ and contains the set $\left\{\left(f(x),g_1(x),\cdots,g_m(x)\right),\ x\in\mathcal{X}\right\}$ on one side. Thus, in order to bound the perturbation of objective with respect to the perturbation of constraint level, we need to bound the slope of the supporting hyperplane from above, which boils down to controlling the magnitude of the Lagrange multiplier. This is summarized in the following lemma:

Lemma 5.4 (Lemma 1 of [31])**.** *Consider the convex program (34)-(36), and define the Lagrange dual function*

$$q(\mu)=\inf_{x\in\mathcal{X}}\left\{f(x)+\sum_{i=1}^{m}\mu_i(g_i(x)-b_i)\right\}.$$

Suppose there exists $\widetilde{x} \in \mathcal{X}$ *such that* $g_i(\widetilde{x}) - b_i \leq -\eta,\ \forall i \in \{1, 2, \cdots, m\}$ *for some positive constant* $\eta > 0$. *Then, the level set* $\mathcal{V}_{\bar{\mu}} = \{\mu_1, \mu_2, \cdots, \mu_m \geq 0,\ q(\mu) \geq q(\bar{\mu})\}$ *is bounded for any nonnegative* $\bar{\mu}$. *Furthermore, we have*

$$\max_{\mu \in \mathcal{V}_{\bar{\mu}}} \|\mu\|_2 \leq \frac{1}{\min_{1 \leq i \leq m} \{-g_i(\widetilde{x}) + b_i\}} \left(f(\widetilde{x}) - q(\bar{\mu})\right).$$

The technical importance of these two lemmas in the current context is contained in the following corollary.

Corollary 5.1. *Let* $\left\{\theta_*^{(k)}\right\}_{k=1}^K$ *and* $\left\{\overline{\theta}_*^{(k)}\right\}_{k=1}^K$ *be solutions to* (30)-(31) *and* (32)-(33), *respectively. Then, the following holds*

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)}\right), \overline{\theta}_*^{(k)}\right\rangle \geq \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta_*^{(k)}\right\rangle - \frac{C_1 K^2\sqrt{m}\Psi^2}{\eta T}$$

where η *is the constant defined in Assumption 2.3.*

Proof of Corollary 5.1. Take

$$\begin{aligned}
f\left(\theta^{(1)}, \cdots, \theta^{(K)}\right) &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta^{(k)}\right\rangle,\\
g_i\left(\theta^{(1)}, \cdots, \theta^{(K)}\right) &= \sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta^{(k)}\right\rangle,\\
\mathcal{X} &= \Theta^{(1)} \times \Theta^{(2)} \times \cdots \times \Theta^{(K)},
\end{aligned}$$

and $b_i = 0$ in Farkas' Lemma and we have the following display

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta^{(k)}\right\rangle + \sum_{i=1}^{m}\mu_i\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta^{(k)}\right\rangle \geq \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta_*^{(k)}\right\rangle,$$

for any $\left(\theta^{(1)}, \cdots, \theta^{(K)}\right) \in \mathcal{X}$ and some $\mu_1, \mu_2, \cdots, \mu_m \geq 0$. In particular, substituting $\left(\overline{\theta}_*^{(1)}, \cdots, \overline{\theta}_*^{(K)}\right)$ into the above display gives

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \overline{\theta}_*^{(k)}\right\rangle &\geq \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta_*^{(k)}\right\rangle - \sum_{i=1}^{m}\mu_i\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \overline{\theta}_*^{(k)}\right\rangle\\
&\geq \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta_*^{(k)}\right\rangle - \frac{C_1 K\Psi}{T}\sum_{i=1}^{m}\mu_i, \qquad (37)
\end{aligned}$$

where the final inequality follows from the fact that $\left(\overline{\theta}_*^{(1)}, \cdots, \overline{\theta}_*^{(K)}\right)$ satisfies the relaxed constraint $\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \overline{\theta}_*^{(k)}\right\rangle \leq \frac{C_1 K\Psi}{T}$ and $\mu_i \geq 0,\ \forall i \in \{1, 2, \cdots, m\}$. Now we need to bound the magnitude of Lagrange multiplier $(\mu_1, \cdots, \mu_m)$. Note that in our scenario,

$$\left|f\left(\theta^{(1)}, \cdots, \theta^{(K)}\right)\right| = \left|\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta^{(k)}\right\rangle\right| \leq \Psi K,$$

and the Lagrange multiplier μ is the solution to the maximization problem

$$\max_{\mu_i \geq 0, i \in \{1,2,\cdots,m\}} q(\mu),$$

where $q(\mu)$ is the dual function defined in Lemma 5.4. thus, it must be in any super level set $\mathcal{V}_{\bar{\mu}} = \{\mu_1, \mu_2, \cdots, \mu_m \geq 0,\ q(\mu) \geq q(\bar{\mu})\}$. In particular, taking $\bar{\mu} = 0$ in Lemma 5.4 and using Slater's condition (8), we have there exists $\widetilde{\theta}^{(1)}, \cdots, \widetilde{\theta}^{(K)}$ such that

$$\sum_{i=1}^{m} \mu_i \leq \sqrt{m}\|\mu\|_2 \leq \frac{\sqrt{m}}{\eta}\left(f\left(\widetilde{\theta}^{(1)}, \cdots, \widetilde{\theta}^{(K)}\right) - \inf_{\left(\theta^{(1)}, \cdots, \theta^{(K)}\right) \in \mathcal{X}} f\left(\theta^{(1)}, \cdots, \theta^{(K)}\right)\right) \leq \frac{2\sqrt{m}\Psi K}{\eta},$$

where the final inequality follows from the deterministic bound of $|f(\theta^{(1)}, \cdots, \theta^{(K)})|$ by ΨK. Substituting this bound into (37) gives the desired result. □

As a simple consequence of the above corollary, we have our final bound on the regret and constraint violation regarding any $(d_0, \Pi) \in \mathcal{G}$.

Theorem 5.1. *Let $\mathscr{P}$ be the sequence of randomized stationary policies resulting from the proposed algorithm with $V = \sqrt{T}$ and $\alpha = T$. Let d_0 be the starting state of the proposed algorithm. For any randomized stationary policy Π starting from the state d_0 such that $(d_0, \Pi) \in \mathcal{G}$, we have*

$$F_T(d_0, \mathscr{P}) - F_T(d_0, \Pi) \leq \mathcal{O}\left(m^{3/2}K^2 \sum_{k=1}^{K} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \cdot \sqrt{T}\right),$$

$$G_{i,T}(d_0, \mathscr{P}) \leq \mathcal{O}\left(m^{3/2}K^2 \sum_{k=1}^{K} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \cdot \sqrt{T}\right),\ i = 1, 2, \cdots, m.$$

Proof. Let Π_* be the randomized stationary policy corresponding to the solution $\{\theta_*^{(k)}\}_{k=1}^K$ to (30)-(31) and let Π be any randomized stationary policy such that $(d_0, \Pi) \in \mathcal{G}$. Since $G_{i,T}(d_{\Pi_*}, \Pi_*) = \sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}(\mathbf{g}_{i,t}), \theta_*^{(k)}\right\rangle \leq 0$, it follows $(d_{\Pi_*}, \Pi_*) \in \mathcal{G}$. By Theorem 4.3, we know that

$$F_T(d_0, \mathscr{P}) - F_T(d_{\Pi_*}, \Pi_*) \leq \mathcal{O}\left(m^{3/2}K^2 \sum_{k=1}^{K} \left|\mathcal{A}^{(k)}\right| \left|\mathcal{S}^{(k)}\right| \cdot \sqrt{T}\right),$$

and $G_{i,T}(d_0, \mathscr{P})$ satisfies the bound in the statement. It is then enough to bound $F_T(d_{\Pi_*}, \Pi_*) - F_T(d_0, \Pi)$. We split it in to two terms:

$$F_T(d_{\Pi_*}, \Pi_*) - F_T(d_0, \Pi) \leq \underbrace{F_T(d_{\Pi_*}, \Pi_*) - F_T(d_\Pi, \Pi)}_{\text{(I)}} + \underbrace{F_T(d_\Pi, \Pi) - F_T(d_0, \Pi)}_{\text{(II)}}.$$

By (28) in Lemma 5.2, the term (II) is bounded by $C_1 K\Psi$. It remains to bound the first term. Since $(d_0, \Pi) \in \mathcal{G}$, by Lemma 5.2, the corresponding state-action probabilities $\{\theta^{(k)}\}_{k=1}^K$ of Π satisfies $\sum_{k=1}^K \left\langle \mathbb{E}(\mathbf{g}_{i,t}), \theta^{(k)}\right\rangle \leq C_1 K\Psi/T$ and $\{\theta^{(k)}\}_{k=1}^K$ is feasible for (32)-(33). Since $\{\overline{\theta}_*^{(k)}\}_{k=1}^K$ is the solution to (32)-(33), we must have

$$F_T(d_\Pi, \Pi) = \sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)}\right), \theta^{(k)}\right\rangle \geq \sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)}\right), \overline{\theta}_*^{(k)}\right\rangle$$

On the other hand, by Corollary 5.1,

$$\sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}_t^{(k)}\right), \overline{\theta}_*^{(k)}\right\rangle \geq \sum_{t=0}^{T-1}\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{f}^{(k)}\right), \theta_*^{(k)}\right\rangle - \frac{C_1K^2\sqrt{m}\Psi^2}{\eta} = F_T(d_{\Pi_*}, \Pi_*) - \frac{C_1K^2\sqrt{m}\Psi^2}{\eta}.$$

Combining the above two displays gives (I) $\leq \frac{C_1K^2\sqrt{m}\Psi^2}{\eta}$ and the proof is finished. □

6. Conclusion

This paper considers online learning over weakly coupled MDPs where the coupling comes from the global constraint functions, and the time varying objective and constraint functions can only be observed after the decision is made. We develop a new algorithm along with a new framework for analysis guaranteeing $\mathcal{O}(\sqrt{T})$ regret and constraint violation simultaneously. The analysis proceeds by first proving $\mathcal{O}(\sqrt{T})$ regret and constraint violation on an imaginary system where stationary distribution is reached instantly every time slot after the decision is made, and then bounding the error between the true system and the imaginary system via an slow-update property of the algorithm.

Note that the current algorithm and analysis assume the full knowledge of the transition probabilities of underlying MDPs and the condition that the decision maker can observe the entire objective and constraint functions over all state-action pairs each slot after the decision is made. It would be interesting if one can relax the above assumptions, and develop algorithms with competitive regret and constraint violation bounds. Specifically, the following two scenarios are worth exploring:

- Bandit setting: The decision maker can only observe the objective and constraint functions' values corresponding to the actions on the current MDP state as oppose to those of all MDP states.
- MDP with unknown parameters: The decision maker has no knowledge on the state space and/or transition probabilities corresponding to different actions of the underlying MDP.

References

[1] New york iso open access pricing data. *http://www.nyiso.com/*.

[2] Anshul Gandhi. *Dynamic server provisioning for data center power management*. PhD thesis, Carnegie Mellon University, 2013.

[3] Anshul Gandhi, Sherwin Doroudi, Mor Harchol-Balter, and Alan Scheller-Wolf. Exact analysis of the m/m/k/setup class of markov chains via recursive renewal reward. In *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 153–166. ACM, 2013.

[4] Xiaohan Wei and Michael Neely. Data center server provision: Distributed asynchronous control for coupled renewal systems. *IEEE/ACM Transactions on Networking*, PP(99):1–15, 2017.

[5] Minghong Lin, Adam Wierman, Lachlan LH Andrew, and Eno Thereska. Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Transactions on Networking (TON)*, 21(5):1378–1391, 2013.

[6] Rahul Urgaonkar, Bhuvan Urgaonkar, Michael J Neely, and Anand Sivasubramaniam. Optimal power cost management using stored energy in data centers. *ACM SIGMETRICS Performance Evaluation Review*, 39(1):181–192, 2011.

[7] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[8] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

[9] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

[10] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13(Sep):2503–2528, 2012.

[11] Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pages 402–411, 2016.

[12] Hao Yu and Michael J Neely. A low complexity algorithm with $o(\sqrt{T})$ regret and finite constraint violations for online convex optimization with long term constraints. *arXiv preprint arXiv:1604.02218*, 2016.

[13] Tianyi Chen, Qing Ling, and Georgios B Giannakis. An online convex optimization approach to dynamic network resource allocation. *arXiv preprint arXiv:1701.03974*, 2017.

[14] Michael J Neely and Hao Yu. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783*, 2017.

[15] Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, pages 1427–1437, 2017.

[16] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

[17] Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 512–520, 2014.

[18] Peng Guan, Maxim Raginsky, and Rebecca M Willett. Online markov decision processes with kullback–leibler control cost. *IEEE Transactions on Automatic Control*, 59(6):1423–1438, 2014.

[19] Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

[20] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1804–1812, 2010.

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

[22] Constantine Caramanis, Nedialko B Dimitrov, and David P Morton. Efficient algorithms for budget-constrained markov decision processes. *IEEE Transactions on Automatic Control*, 59(10):2813–2817, 2014.

[23] Michael J Neely. Online fractional programming for markov decision systems. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 353–360. IEEE, 2011.

[24] Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained markov decision processes. In *UAI*, 2016.

[25] Xiaohan Wei and Michael J Neely. On the theory and application of distributed asynchronous optimization over weakly coupled renewal systems. *arXiv preprint arXiv:1608.00195*, 2016.

[26] Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

[27] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific

Belmont, MA, 1995.
[28] Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.
[29] Tor Lattimore, Marcus Hutter, Peter Sunehag, et al. The sample-complexity of general reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*. Journal of Machine Learning Research, 2013.
[30] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
[31] Angelia Nedić and Asuman Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.
[32] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.
[33] Bennett Fox. Markov renewal programming by linear fractional programming. *SIAM Journal on Applied Mathematics*, 14(6):1418–1432, 1966.
[34] Yinyu Ye and Edison Tse. An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical programming*, 44(1):157–179, 1989.
[35] Hao Yu and Michael J. Neely. A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017.
[36] Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 14(3):502–525, 1982.
[37] Atilla Eryilmaz and R Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359, 2012.
[38] Xiaohan Wei and Michael J Neely. Online constrained optimization over time varying renewal systems: An empirical method. *arXiv preprint arXiv:1606.03463*, 2016.
[39] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

Appendix A: Additional proofs

A.1. Missing proofs in Section 2.4

We prove Lemma 2.1 and 2.2 in this section.

Proof of Lemma 2.1. For simplicity of notations, we drop the dependencies on k throughout this proof. We first show that for any $r \geq \widehat{r}$, where $\widehat{r}$ is specified in Assumption 2.1, $\mathbf{P}_{\pi_1}\mathbf{P}_{\pi_2}\cdots\mathbf{P}_{\pi_r}$ is a strictly positive stochastic matrix.

Since the MDP is finite state with a finite action set, the set of all pure policies (Definition 2.2) is finite. Let $\mathbf{P}_1,\ \mathbf{P}_2, \cdots,\ \mathbf{P}_N$ be probability transition matrices corresponding to these pure policies. Consider any sequence of randomized stationary policies $\pi_1, \cdots, \pi_r$. Then, it follows their transition matrices can be expressed as convex combinations of pure policies, i.e.

$$\mathbf{P}_{\pi_1} = \sum_{i=1}^{N} \alpha_i^{(1)}\mathbf{P}_i, \quad \mathbf{P}_{\pi_2} = \sum_{i=1}^{N} \alpha_i^{(2)}\mathbf{P}_i, \quad \cdots, \mathbf{P}_{\pi_r} = \sum_{i=1}^{N} \alpha_i^{(r)}\mathbf{P}_i,$$

where $\sum_{i=1}^{N}\alpha_i^{(j)} = 1,\ \forall j \in \{1,2,\cdots,r\}$ and $\alpha_i^{(j)} \geq 0$. Thus, we have the following display

$$\begin{aligned}\mathbf{P}_{\pi_1}\mathbf{P}_{\pi_2}\cdots\mathbf{P}_{\pi_r} &= \left(\sum_{i=1}^{N}\alpha_i^{(1)}\mathbf{P}_i\right)\left(\sum_{i=1}^{N}\alpha_i^{(2)}\mathbf{P}_i\right)\cdots\left(\sum_{i=1}^{N}\alpha_i^{(r)}\mathbf{P}_i\right)\\ &= \sum_{(i_1,\cdots,i_r)\in\mathcal{G}_r} \alpha_{i_1}^{(1)}\cdots\alpha_{i_r}^{(r)}\cdot\mathbf{P}_{i_1}\mathbf{P}_{i_2}\cdots\mathbf{P}_{i_r}, \end{aligned} \tag{38}$$

where $\mathcal{G}_r$ ranges over all N^r configurations.

Since $\left(\sum_{i=1}^N \alpha_i^{(1)}\right)\cdots\left(\sum_{i=1}^N \alpha_i^{(r)}\right)=1$, it follows (38) is a convex combination of all possible sequences $\mathbf{P}_{i_1}\mathbf{P}_{i_2}\cdots\mathbf{P}_{i_r}$. By assumption 2.1, we have $\mathbf{P}_{i_1}\mathbf{P}_{i_2}\cdots\mathbf{P}_{i_r}$ is strictly positive for any $(i_1,\cdots,i_r)\in\mathcal{G}_r$, and there exists a universal lower bound $\delta>0$ of all entries of $\mathbf{P}_{i_1}\mathbf{P}_{i_2}\cdots\mathbf{P}_{i_r}$ ranging over all configurations in $(i_1,\cdots,i_r)\in\mathcal{G}_r$. This implies $\mathbf{P}_{\pi_1}\mathbf{P}_{\pi_2}\cdots\mathbf{P}_{\pi_r}$ is also strictly positive with the same lower bound $\delta>0$ for any sequences of randomized stationary policies $\pi_1,\cdots,\pi_r$.

Now, we proceed to prove the mixing bound. Choose $r=\widehat{r}$ and we can decompose any $\mathbf{P}_{\pi_1}\mathbf{P}_{\pi_2}\cdots\mathbf{P}_{\pi_r}$ as follows:

$$\mathbf{P}_{\pi_1}\cdots\mathbf{P}_{\pi_r}=\delta\mathbf{\Pi}+(1-\delta)\mathbf{Q},$$

where $\mathbf{\Pi}$ has each entry equal to $1/|\mathcal{S}|$ (recall that $|\mathcal{S}|$ is the number of states which equals the size of the matrix) and $\mathbf{Q}$ depends on $\pi_1,\cdots,\pi_r$. Then, $\mathbf{Q}$ is also a stochastic matrix (nonnegative and row sum up to 1) because both $\mathbf{P}_{\pi_1}\cdots\mathbf{P}_{\pi_r}$ and $\mathbf{\Pi}$ are stochastic matrices. Thus, for any two distribution vectors d_1 and d_2, we have

$$(d_1-d_2)\,\mathbf{P}_{\pi_1}\cdots\mathbf{P}_{\pi_r}=\delta\,(d_1-d_2)\,\mathbf{\Pi}+(1-\delta)\,(d_1-d_2)\,\mathbf{Q}=(1-\delta)\,(d_1-d_2)\,\mathbf{Q},$$

where we use the fact that for distribution vectors

$$(d_1-d_2)\,\mathbf{\Pi}=\frac{1}{|\mathcal{S}|}\mathbf{1}-\frac{1}{|\mathcal{S}|}\mathbf{1}=0.$$

Since $\mathbf{Q}$ is a stochastic matrix, it is non-expansive on ℓ_1-norm, namely, for any vector x, $\|x\mathbf{Q}\|_1\le\|x\|_1$. To see this, simply compute

$$\|x\mathbf{Q}\|_1=\sum_{j=1}^{|\mathcal{S}|}\left|\sum_{i=1}^{|\mathcal{S}|}x_iQ_{ij}\right|\le\sum_{j=1}^{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|}|x_iQ_{ij}|=\sum_{j=1}^{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|}|x_i|\,Q_{ij}=\sum_{i=1}^{|\mathcal{S}|}|x_i|=\|x\|_1. \tag{39}$$

Overall, we obtain,

$$\|(d_1-d_2)\,\mathbf{P}_{\pi_1}\cdots\mathbf{P}_{\pi_r}\|_1=(1-\delta)\,\|(d_1-d_2)\,\mathbf{Q}\|_1\le(1-\delta)\,\|d_1-d_2\|_1.$$

We can then take $\tau=-\frac{1}{\log(1-\delta)}$ to finish the proof. □

Proof of Lemma 2.2. Since the probability transition matrix of any randomized stationary policy is a convex combination of those of pure policies, it is enough to show that the product MDP is irreducible and aperiodic under any joint pure policy. For simplicity, let $\mathbf{s}_t=\left(s^{(1)},\cdots,s^{(K)}\right)$ and $\mathbf{a}_t=\left(a^{(1)},\cdots,a^{(K)}\right)$. Consider any joint pure policy Π which select a fixed joint action $\mathbf{a}\in\mathcal{A}^{(1)}\times\cdots\times\mathcal{A}^{(K)}$ given a joint state $\mathbf{s}\in\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}$, with probability 1. By Assumption 2.2, we have

$$\begin{aligned}
&Pr\left(s_{t+1}^{(1)},\cdots,s_{t+1}^{(K)}\middle|s_t^{(1)},\cdots,s_t^{(K)},a_t^{(1)},\cdots,a_t^{(K)}\right)\\
=&Pr\left(s_{t+1}^{(1)}\middle|s_t^{(1)},\cdots,s_t^{(K)},a_t^{(1)},\cdots,a_t^{(K)},s_{t+1}^{(2)},\cdots,s_{t+1}^{(K)}\right)\\
&\cdot Pr\left(s_{t+1}^{(2)},\cdots,s_{t+1}^{(K)}\middle|s_t^{(1)},\cdots,s_t^{(K)},a_t^{(1)},\cdots,a_t^{(K)}\right)\\
=&Pr\left(s_{t+1}^{(1)}\middle|s_t^{(1)},a_t^{(1)}\right)Pr\left(s_{t+1}^{(2)},\cdots,s_{t+1}^{(K)}\middle|s_t^{(1)},\cdots,s_t^{(K)},a_t^{(1)},\cdots,a_t^{(K)}\right)\\
=&\cdots=\prod_{k=1}^{K-1}Pr\left(s_{t+1}^{(k)}\middle|s_t^{(k)},a_t^{(k)}\right)\cdot Pr\left(s_{t+1}^{(K)}\middle|s_t^{(1)},\cdots,s_t^{(K)},a_t^{(1)},\cdots,a_t^{(K)}\right)\\
=&\prod_{k=1}^{K}Pr\left(s_{t+1}^{(k)}\middle|s_t^{(k)},a_t^{(k)}\right),
\end{aligned}\tag{40}$$

where the second equality follows from the independence relation in Assumption 2.2. Thus, we obtain the equality,

$$Pr(\mathbf{s}_{t+1}=\mathbf{s}'|\mathbf{s}_t=\mathbf{s},\mathbf{a}_t=\mathbf{a})=\prod_{k=1}^{K}Pr\left(s_{t+1}^{(k)}=\tilde{s}^{(k)}\left|s_t^{(k)}=s^{(k)},a_t^{(k)}=a^{(k)}\right.\right),$$

Then, the one step transition probability between any two states $\mathbf{s},\tilde{\mathbf{s}}\in\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}$ can be computed as

$$\begin{aligned}Pr(\mathbf{s}_{t+1}=\tilde{\mathbf{s}}|\mathbf{s}_t=\mathbf{s})&=\sum_{\mathbf{a}}Pr(\mathbf{s}_{t+1}=\tilde{\mathbf{s}}|\mathbf{s}_t=\mathbf{s},\mathbf{a}_t=\mathbf{a})\cdot Pr(\mathbf{a}_t=\mathbf{a}|\mathbf{s}_t=\mathbf{s})\\&=\sum_{\mathbf{a}}\prod_{k=1}^{K}Pr\left(s_{t+1}^{(k)}=\tilde{s}^{(k)}\left|s_t^{(k)}=s^{(k)},a_t^{(k)}=a^{(k)}\right.\right)\cdot Pr(\mathbf{a}_t=\mathbf{a}|\mathbf{s}_t=\mathbf{s})\\&=\prod_{k=1}^{K}P_{a^{(k)}(\mathbf{s})}\left(s^{(k)},\tilde{s}^{(k)}\right),\end{aligned}$$

where we can remove the summation on $\mathbf{a}$ due to the fact that $\mathbf{a}_t$ is a pure policy. The notation $a^{(k)}(\mathbf{s})$ denotes a fixed mapping from product state space $\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}$ to an individual action space $\mathcal{A}^{(k)}$ resulting from the pure policy, and $P_{a^{(k)}(\mathbf{s})}\left(s^{(k)},\tilde{s}^{(k)}\right)$ is the Markov transition probability from state $s^{(k)}$ to $\tilde{s}^{(k)}$ under the action $a^{(k)}(\mathbf{s})$. One can then further compute the r $(r\geq 2)$ step transition probability from between any two states $\mathbf{s},\tilde{\mathbf{s}}\in\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}$ as

$$\begin{aligned}Pr(\mathbf{s}_{t+r}=\tilde{\mathbf{s}}|\mathbf{s}_t=\mathbf{s})&=\sum_{\mathbf{s}_{t+r-1}}\cdots\sum_{\mathbf{s}_{t+1}}\prod_{k=1}^{K}P_{a^{(k)}(\mathbf{s})}\left(s^{(k)},s_{t+1}^{(k)}\right)\cdot\prod_{k=1}^{K}P_{a^{(k)}(\mathbf{s}_{t+1})}\left(s_{t+1}^{(k)},s_{t+2}^{(k)}\right)\\&\quad\cdots\prod_{k=1}^{K}P_{a^{(k)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)},\tilde{s}^{(k)}\right)\\&=\sum_{\mathbf{s}_{t+r-1}}\cdots\sum_{\mathbf{s}_{t+1}}\prod_{k=1}^{K}P_{a^{(k)}(\mathbf{s})}\left(s^{(k)},s_{t+1}^{(k)}\right)\cdot P_{a^{(k)}(\mathbf{s}_{t+1})}\left(s_{t+1}^{(k)},s_{t+2}^{(k)}\right)\\&\quad\cdots P_{a^{(k)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)},\tilde{s}^{(k)}\right).\end{aligned}\tag{41}$$

For any $k\in\{1,2,\cdots,K\}$, the term

$$P_{a^{(k)}(\mathbf{s})}\left(s^{(k)},s_{t+1}^{(k)}\right)\cdot P_{a^{(k)}(\mathbf{s}_{t+1})}\left(s_{t+1}^{(k)},s_{t+2}^{(k)}\right)\cdots P_{a^{(k)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)},\tilde{s}^{(k)}\right)$$

denotes the probability of moving from $s^{(k)}$ to $\tilde{s}^{(k)}$ along a certain path under a certain sequence of fixed decisions $a^{(k)}(\mathbf{s}),\ a^{(k)}(\mathbf{s}_{t+1}),\cdots,\ a^{(k)}(\mathbf{s}_{t+r-1})$. Let

$$\mathbf{s}^{(k)}=\left(s_{t+1}^{(k)},s_{t+2}^{(k)},\cdots,s_{t+r-1}^{(k)}\right)\in\mathcal{S}^{(k)}\times\cdots\times\mathcal{S}^{(k)},\ k\in\{1,2,\cdots,K\}$$

be the state path of k-th MDP. One can then change the order of summation in (41) and sum over state paths of each MDP as follows:

$$(41)=\sum_{\mathbf{s}^{(K)}}\cdots\sum_{\mathbf{s}^{(1)}}\prod_{k=1}^{K}P_{a^{(k)}(\mathbf{s})}\left(s^{(k)},s_{t+1}^{(k)}\right)\cdot P_{a^{(k)}(\mathbf{s}_{t+1})}\left(s_{t+1}^{(k)},s_{t+2}^{(k)}\right)\cdots P_{a^{(k)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)},\tilde{s}^{(k)}\right)$$

We would like to exchange the order of the product and the sums so that we can take the path sum over each individual MDP respectively. However, the problem is that the transition probabilities are coupled through the actions. The idea to proceed is to first apply a "hard" decoupling by taking the infimum of transition probabilities of each MDP over all pure policies, and use Assumption 2.1, to bound the transition probability from below uniformly. We have

$$\begin{aligned}(41) \geq & \inf_{\mathbf{s}^{(1)}} \sum_{\mathbf{s}^{(K)}} \cdots \sum_{\mathbf{s}^{(2)}} \prod_{k=2}^{K} P_{a^{(k)}(\mathbf{s})}\left(s^{(k)}, s_{t+1}^{(k)}\right) \cdots P_{a^{(k)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)}, \tilde{s}^{(k)}\right) \\ & \cdot \inf_{\mathbf{s}^{(j)},\ j\neq 1} \sum_{\mathbf{s}^{(1)}} P_{a^{(1)}(\mathbf{s})}\left(s^{(1)}, s_{t+1}^{(1)}\right) \cdots P_{a^{(1)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(1)}, \tilde{s}^{(1)}\right) \\ \geq & \inf_{\mathbf{s}^{(1)}} \sum_{\mathbf{s}^{(K)}} \cdots \sum_{\mathbf{s}^{(2)}} \prod_{k=2}^{K} P_{a^{(k)}(\mathbf{s})}\left(s^{(k)}, s_{t+1}^{(k)}\right) \cdots P_{a^{(k)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)}, \tilde{s}^{(k)}\right) \\ & \cdot \inf_{\pi_1^{(1)},\cdots,\pi_r^{(1)}} \sum_{\mathbf{s}^{(1)}} P_{\pi_1^{(1)}}\left(s^{(1)}, s_{t+1}^{(1)}\right) \cdots P_{\pi_r^{(1)}}\left(s_{t+r-1}^{(1)}, \tilde{s}^{(1)}\right),\end{aligned}$$

where $\pi_1^{(1)}, \cdots, \pi_r^{(1)}$ range over all pure policies, and the second inequality follows from the fact that *fix any path of other MDPs* (i.e. $\mathbf{s}^{(j)},\ j \neq 1$), the term

$$\sum_{\mathbf{s}^{(1)}} P_{a^{(1)}(\mathbf{s})}\left(s^{(1)}, s_{t+1}^{(1)}\right) \cdots P_{a^{(1)}(\mathbf{s}_{t+r-1})}\left(s_{t+r-1}^{(k)}, \tilde{s}^{(1)}\right)$$

is the probability of reaching $\tilde{s}^{(1)}$ from $s^{(1)}$ in r steps using a sequence of actions $a^{(1)}(\mathbf{s}^{(1)}), \cdots, a^{(1)}(\mathbf{s}_{t+r-1}^{(1)})$, where each action is a deterministic function of the previous state at the 1-st MDP only. Thus, it dominates the infimum over all sequences of pure policies $\pi_1^{(1)}, \cdots, \pi_r^{(1)}$ on this MDP. Similarly, we can decouple the rest of the sums and obtain the follow display:

$$\begin{aligned}(41) \geq & \prod_{k=1}^{K} \inf_{\pi_1^{(k)},\cdots,\pi_r^{(k)}} \sum_{\mathbf{s}^{(k)}} P_{\pi_1^{(k)}}\left(s^{(k)}, s_{t+1}^{(k)}\right) \cdots P_{\pi_r^{(k)}}\left(s_{t+r-1}^{(k)}, \tilde{s}^{(k)}\right) \\ = & \prod_{k=1}^{K} \inf_{\pi_1^{(k)},\cdots,\pi_r^{(k)}} P_{\pi_1^{(k)},\cdots,\pi_r^{(k)}}\left(s^{(k)}, \tilde{s}^{(k)}\right),\end{aligned}$$

where $P_{\pi_1^{(k)},\cdots,\pi_r^{(k)}}\left(s^{(k)}, \tilde{s}^{(k)}\right)$ denotes the $\left(s^{(k)}, \tilde{s}^{(k)}\right)$-th entry of the product matrix $\mathbf{P}_{\pi_1^{(k)}}^{(k)} \cdots \mathbf{P}_{\pi_r^{(k)}}^{(k)}$. Now, by Assumption 2.1, there exists a large enough integer $\widehat{r}$ such that $\mathbf{P}_{\pi_1^{(k)}}^{(k)} \cdots \mathbf{P}_{\pi_r^{(k)}}^{(k)}$ is a strictly positive matrix for any sequence of $r \geq \widehat{r}$ randomized stationary policy. As a consequence, the above probability is strictly positive and (41) is also strictly positive.

This implies, if we choose $\tilde{\mathbf{s}} = \mathbf{s}$, then, starting from any arbitrary product state $\mathbf{s} \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$, there is a positive probability of returning to this state after r steps for all $r \geq \widehat{r}$, which gives the aperiodicity. Similarly, there is a positive probability of reaching any other composite state after r steps for all $r \geq \widehat{r}$, which gives the irreducibility. This implies the product state MDP is irreducible and aperiodic under any joint pure policy, and thus, any joint randomized stationary policy.

For the second part of the claim, we consider any randomized stationary policy Π and the corresponding joint transition probability matrix $\mathbf{P}_\Pi$, there exists a stationary state-action probability vector $\Phi(\mathbf{a}, \mathbf{s}),\ \mathbf{a} \in \mathcal{A}^{(1)} \times \cdots \times \mathcal{A}^{(K)},\ \mathbf{s} \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$, such that

$$\sum_{\mathbf{a}} \Phi(\mathbf{a}, \tilde{\mathbf{s}}) = \sum_{\mathbf{s}} \sum_{\mathbf{a}} \Phi(\mathbf{a}, \mathbf{s}) P_{\mathbf{a}}(\mathbf{s}, \tilde{\mathbf{s}}),\ \forall \tilde{\mathbf{s}} \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}. \tag{42}$$

Then, the state-action probability of the k-th MDP is $\theta^{(k)}(a^{(k)},\tilde{s}^{(k)}) = \sum_{\tilde{s}^{(j)},a^{(j)},\ j\neq k}\Phi(\mathbf{a},\tilde{\mathbf{s}})$. Thus,

$$\begin{aligned}\sum_{a^{(k)}}\theta^{(k)}(a^{(k)},\tilde{s}^{(k)}) &= \sum_{\tilde{s}^{(j)},\ j\neq k}\sum_{\mathbf{a}}\Phi(\mathbf{a},\tilde{\mathbf{s}}) = \sum_{\mathbf{s}}\sum_{\mathbf{a}}\Phi(\mathbf{a},\mathbf{s})\sum_{\tilde{s}^{(j)},\ j\neq k}P_{\mathbf{a}}(\mathbf{s},\tilde{\mathbf{s}})\\ &= \sum_{\mathbf{s}}\sum_{\mathbf{a}}\Phi(\mathbf{a},\mathbf{s})\cdot Pr\left(\tilde{s}^{(k)}|\mathbf{a},\mathbf{s}\right) = \sum_{\mathbf{s}}\sum_{\mathbf{a}}\Phi(\mathbf{a},\mathbf{s})\cdot Pr\left(\tilde{s}^{(k)}|a^{(k)},s^{(k)}\right)\\ &= \sum_{a^{(k)}}\sum_{s^{(k)}}\theta^{(k)}(a^{(k)},\tilde{s}^{(k)})\cdot Pr\left(\tilde{s}^{(k)}|a^{(k)},s^{(k)}\right)\\ &= \sum_{a^{(k)}}\sum_{s^{(k)}}\theta^{(k)}(a^{(k)},\tilde{s}^{(k)})\cdot P_{a^{(k)}}\left(s^{(k)},\tilde{s}^{(k)}\right)\end{aligned}$$

where the third from the last inequality follows from Assumption 2.2. This finishes the proof. □

A.2. Missing proofs in Section 4.1

Proof of Lemma 4.6. Consider the state-action probabilities $\{\tilde{\theta}^{(k)}\}_{k=1}^K$ which achieves the Slater's condition in (8). First of all, note that $Q_i(t)\in\mathcal{F}_{t-1},\ \forall t\geq 1$. Then, using the assumption that $\{\mathbf{g}_{i,t-1}^{(k)}\}_{k=1}^K$ is i.i.d. and independent of all system information up to $t-1$, we have

$$\mathbb{E}\left(Q_i(t-1)\sum_{k=1}^K\left\langle\mathbf{g}_{i,t-1}^{(k)},\tilde{\theta}\right\rangle\ \middle|\ \mathcal{F}_{t-1}\right) = \mathbb{E}\left(\sum_{k=1}^K\left\langle\mathbf{g}_{i,t-1}^{(k)},\tilde{\theta}\right\rangle\right)Q_i(t-1)\leq -\eta Q_i(t-1). \tag{43}$$

Now, by the drift-plus-penalty bound (15), with $\theta^{(k)}=\tilde{\theta}^{(k)}$,

$$\begin{aligned}\Delta(t)\leq& -V\sum_{k=1}^K\left\langle\mathbf{f}_{t-1}^{(k)},\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\rangle-\alpha\sum_{k=1}^K\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\|_2^2+\frac{3}{2}mK^2\Psi^2+V\sum_{k=1}^K\left\langle\mathbf{f}_{t-1}^{(k)},\tilde{\theta}^{(k)}-\theta_{t-1}^{(k)}\right\rangle\\ &+\sum_{i=1}^m Q_i(t-1)\sum_{k=1}^K\left\langle\mathbf{g}_{i,t-1}^{(k)},\tilde{\theta}^{(k)}\right\rangle+\alpha\sum_{k=1}^K\|\tilde{\theta}^{(k)}-\theta_{t-1}^{(k)}\|_2^2-\alpha\sum_{k=1}^K\|\tilde{\theta}^{(k)}-\theta_t^{(k)}\|_2^2\\ \leq& 4VK\Psi+\frac{3}{2}mK^2\Psi^2+\sum_{i=1}^m Q_i(t-1)\sum_{k=1}^K\left\langle\mathbf{g}_{i,t-1}^{(k)},\tilde{\theta}^{(k)}\right\rangle\\ &+\alpha\sum_{k=1}^K\|\tilde{\theta}^{(k)}-\theta_{t-1}^{(k)}\|_2^2-\alpha\sum_{k=1}^K\|\tilde{\theta}^{(k)}-\theta_t^{(k)}\|_2^2\end{aligned}$$

where the second inequality follows from Holder's inequality that

$$\left|\left\langle\mathbf{f}_{t-1}^{(k)},\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\rangle\right|\leq\|\mathbf{f}_{t-1}^{(k)}\|_\infty\left\|\theta_t^{(k)}-\theta_{t-1}^{(k)}\right\|_1\leq 2\Psi.$$

Summing up the drift from t to $t+t_0-1$ and taking a conditional expectation $\mathbb{E}(\cdot|\mathcal{F}_{t-1})$ give

$$\begin{aligned}
&\mathbb{E}\Big(\|\mathbf{Q}(t+t_0)\|_2^2-\|\mathbf{Q}(t)\|_2^2\Big|\mathcal{F}_{t-1}\Big)\\
\leq& 8VK\Psi+3mK^2\Psi^2+2\sum_{i=1}^m\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}Q_i(\tau-1)\sum_{k=1}^K\left\langle\mathbf{g}_{i,\tau-1}^{(k)},\tilde{\theta}^{(k)}\right\rangle\Big|\mathcal{F}_{t-1}\right)\\
&+2\alpha\mathbb{E}\left(\sum_{k=1}^K\left(\|\tilde{\theta}^{(k)}-\theta_{t-1}^{(k)}\|_2^2-\|\tilde{\theta}^{(k)}-\theta_{t+t_0}^{(k)}\|_2^2\right)\Big|\mathcal{F}_{t-1}\right)\\
\leq& 8VK\Psi+3mK^2\Psi^2+4K\alpha+2\sum_{i=1}^m\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}Q_i(\tau-1)\sum_{k=1}^K\left\langle\mathbf{g}_{i,\tau-1}^{(k)},\tilde{\theta}^{(k)}\right\rangle\Big|\mathcal{F}_{t-1}\right).
\end{aligned}$$

Using the tower property of conditional expectations (further taking conditional expectations $\mathbb{E}\Big(\cdot\Big|\mathcal{F}_{t+t_0-1}\cdots\Big|\mathcal{F}_t\Big)$ inside the conditional expectation) and the bound (43), we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}Q_i(\tau-1)\sum_{k=1}^K\left\langle\mathbf{g}_{i,\tau-1}^{(k)},\tilde{\theta}^{(k)}\right\rangle\Big|\mathcal{F}_{t-1}\right)\\
&\leq-\eta\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}Q_i(\tau-1)\Big|\mathcal{F}_{t-1}\right)\\
&\leq-\eta t_0Q_i(t-1)+\frac{t_0(t_0-1)}{2}\Psi\leq-\eta t_0Q_i(t)+\frac{t_0(t_0-1)}{2}\Psi+\eta t_0K\Psi,
\end{aligned}$$

where the last inequality follows from the queue updating rule (9) that

$$|Q_i(t-1)-Q_i(t)|\leq\left|\sum_{k=1}^K\left\langle\mathbf{g}_{i,t-2}^{(k)},\theta_{t-1}^{(k)}\right\rangle\right|\leq K\|\mathbf{g}_{i,t-2}^{(k)}\|_\infty\|\theta_{t-1}^{(k)}\|_1\leq K\Psi.$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\Big(\|\mathbf{Q}(t+t_0)\|_2^2-\|\mathbf{Q}(t)\|_2^2\Big|\mathcal{F}_{t-1}\Big)\\
&\leq 8VK\Psi+3mK^2\Psi^2+4K\alpha+t_0(t_0-1)m\Psi+2mK\Psi\eta t_0-2\eta t_0\sum_{i=1}^mQ_i(t)\\
&\leq 8VK\Psi+3mK^2\Psi^2+4K\alpha+t_0(t_0-1)m\Psi+2mK\Psi\eta t_0-2\eta t_0\|Q_i(t)\|_2.
\end{aligned}$$

Suppose $\|Q_i(t)\|_2\geq\frac{8VK\Psi+3mK^2\Psi^2+4K\alpha+t_0(t_0-1)m\Psi+2mK\Psi\eta t_0+\eta^2t_0^2}{\eta t_0}$, then, it follows,

$$\mathbb{E}\Big(\|\mathbf{Q}(t+t_0)\|_2^2-\|\mathbf{Q}(t)\|_2^2\Big|\mathcal{F}_{t-1}\Big)\leq-\eta t_0\|Q_i(t)\|_2,$$

which implies

$$\mathbb{E}\Big(\|\mathbf{Q}(t+t_0)\|_2^2\Big|\mathcal{F}_{t-1}\Big)\leq\left(\|Q_i(t)\|_2-\frac{\eta t_0}{2}\right)^2$$

Since $\|Q_i(t)\|_2\geq\frac{\eta t_0}{2}$, taking square root from both sides using Jensen' inequality gives

$$\mathbb{E}\Big(\|\mathbf{Q}(t+t_0)\|_2\Big|\mathcal{F}_{t-1}\Big)\leq\|Q_i(t)\|_2-\frac{\eta t_0}{2}.$$

On the other hand, we always have

$$\left|\|\mathbf{Q}(t+1)\|_2 - \|\mathbf{Q}(t)\|_2\right| = \left|\sqrt{\sum_{i=1}^{m}\max\left\{Q_i(t)+\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t^{(k)}\right\rangle, 0\right\}^2} - \sqrt{\sum_{i=1}^{m} Q_i(t)^2}\right|$$
$$\leq \left(\sum_{i=1}^{m}\left(\sum_{k=1}^{K}\left\langle \mathbf{g}_{i,t-1}^{(k)}, \theta_t^{(k)}\right\rangle\right)^2\right)^{1/2} \leq \sqrt{m}K\Psi.$$

Overall, we finish the proof. □

A.3. Missing proofs in Section 5

Proof of Lemma 5.1. Consider any joint randomized stationary policy Π and a starting state probability d_0 on the product state space $\mathcal{S}^{(1)}\times\mathcal{S}^{(2)}\times\cdots\times\mathcal{S}^{(K)}$. Let $\mathbf{P}_\Pi$ be the corresponding transition matrix on the product state space. Let d_t be the state distribution at time t under Π and d_Π be the stationary state distribution. By Lemma 2.2, we know that this product state MDP is irreducible and aperiodic (ergodic) under any randomized stationary policy. In particular, it is ergodic under any pure policy. Since there are only finitely many pure policies, let $\mathbf{P}_{\Pi_1},\cdots,\mathbf{P}_{\Pi_N}$ be probability transition matrices corresponding to these pure policies. By Proposition 1.7 of [39] , for any $\Pi_i,\ i\in\{1,2,\cdots,N\}$, there exists integer $\tau_i>0$ such that $(\mathbf{P}_{\Pi_i})^t$ is strictly positive for any $t\geq\tau_i$. Let

$$\tau_1 = \max_i \tau_i,$$

then, it follows $(\mathbf{P}_{\Pi_i})^{\tau_1}$ is strictly positive uniformly for all Π_i's. Let $\delta>0$ be the least entry of $(\mathbf{P}_{\Pi_i})^{\tau_1}$ over all Π_i's. Following from the fact that the probability transition matrix $\mathbf{P}_\Pi$ is a convex combination of those of pure policies, i.e. $\mathbf{P}_\Pi=\sum_{i=1}^N\alpha_i\mathbf{P}_{\Pi_i},\ \alpha_i\geq 0,\ \sum_{i=1}^N\alpha_i=1$, we have $(\mathbf{P}_\Pi)^{\tau_1}$ is also strictly positive. To see this, note that

$$(\mathbf{P}_\Pi)^{\tau_1} = \left(\sum_{i=1}^N\alpha_i\mathbf{P}_{\Pi_i}\right)^{\tau_1} \geq \sum_{i=1}^N\alpha_i^{\tau_1}(\mathbf{P}_{\Pi_i})^{\tau_1} > 0,$$

where the inequality is taken to be entry-wise. Furthermore, the least entry of $(\mathbf{P}_\Pi)^{\tau_1}$ is lower bounded by δ/N^{τ_1-1} uniformly over all joint randomized stationary policies Π, which follows from the fact that the least entry of $\frac{1}{N}(\mathbf{P}_\Pi)^{\tau_1}$ is bounded as

$$\frac{1}{N}\sum_{i=1}^N\alpha_i^{\tau_1}\delta \geq \left(\frac{1}{N}\sum_{i=1}^N\alpha_i\right)^{\tau_1}\delta = \frac{\delta}{N^{\tau_1}}.$$

The rest is a standard bookkeeping argument following from the Markov chain mixing time theory (Theorem 4.9 of [39]). Let $\mathbf{D}_\Pi$ be a matrix of the same size as $\mathbf{P}_\Pi$ and each row equal to the stationary distribution d_Π. Let $\varepsilon=\delta/N^{\tau_1-1}$. We claim that for any integer $n>0$, and any Π,

$$\mathbf{P}_\Pi^{\tau_1 n} = (1-(1-\varepsilon)^n)\mathbf{D}_\Pi + (1-\varepsilon)^n\mathbf{Q}^n, \tag{44}$$

for some stochastic matrix $\mathbf{Q}$. We use induction to prove this claim. First of all, for $n=1$, from the fact that $(\mathbf{P}_\Pi)^{\tau_1}$ is a positive matrix and the least entry is uniformly lower bounded by ε over all policies Π, we can write $(\mathbf{P}_\Pi)^{\tau_1}$ as

$$(\mathbf{P}_\Pi)^{\tau_1} = \varepsilon\mathbf{D}_\Pi + (1-\varepsilon)\mathbf{Q},$$

for some stochastic matrix $\mathbf{Q}$, where we use the fact that $\varepsilon \in (0,1]$. Suppose (44) holds for $n = 1, 2, \cdots, \ell$, we show that it also holds for $n = \ell + 1$. Using the fact that $\mathbf{D}_\Pi \mathbf{P}_\Pi = \mathbf{D}_\Pi$ and $\mathbf{Q}\mathbf{D}_\Pi = \mathbf{D}_\Pi$ for any stochastic matrix $\mathbf{Q}$, we can write out $\mathbf{P}_\Pi^{\tau_1(\ell+1)}$:

$$\begin{aligned}
\mathbf{P}_\Pi^{\tau_1(\ell+1)} =& \mathbf{P}_\Pi^{\tau_1 \ell} \mathbf{P}_\Pi^{\tau_1} = \left(\left(1-(1-\varepsilon)^\ell\right)\mathbf{D}_\Pi + (1-\varepsilon)^\ell Q^\ell\right)\mathbf{P}_\Pi^{\tau_1} \\
=& \left(1-(1-\varepsilon)^\ell\right)\mathbf{D}_\Pi \mathbf{P}_\Pi^{\tau_1} + (1-\varepsilon)^\ell \mathbf{Q}^\ell \mathbf{P}_\Pi^{\tau_1} \\
=& \left(1-(1-\varepsilon)^\ell\right)\mathbf{D}_\Pi + (1-\varepsilon)^\ell \mathbf{Q}^\ell (\varepsilon \mathbf{D}_\Pi + (1-\varepsilon)\mathbf{Q}) \\
=& \left(1-(1-\varepsilon)^\ell\right)\mathbf{D}_\Pi + (1-\varepsilon)^\ell \mathbf{Q}^\ell ((1-(1-\varepsilon))\mathbf{D}_\Pi + (1-\varepsilon)\mathbf{Q}) \\
=& (1-(1-\varepsilon)^{\ell+1})\mathbf{D}_\Pi + (1-\varepsilon)^{\ell+1}\mathbf{Q}^{\ell+1}.
\end{aligned}$$

Thus, (44) holds. For any integer $t > 0$, we write $t = \tau_1 n + j$ for some integer $j \in [0, \tau_1)$ and $n \geq 0$. Then,

$$(\mathbf{P}_\Pi)^t - \mathbf{D}_\Pi = (\mathbf{P}_\Pi)^t - \mathbf{D}_\Pi = (1-\varepsilon)^n \left(Q^n \mathbf{P}_\Pi^j - \mathbf{D}_\Pi\right).$$

Let $\mathbf{P}_\Pi^t(i,\cdot)$ be the i-th row of $\mathbf{P}_\Pi^t$, then, we obtain

$$\max_i \|\mathbf{P}_\Pi^t(i,\cdot) - d_\Pi\|_1 \leq 2(1-\varepsilon)^n,$$

where we use the fact that the ℓ_1-norm of the row difference is bounded by 2. Finally, for any starting state distribution d_0, we have

$$\begin{aligned}
\left\|d_0 \mathbf{P}_\Pi^t - d_\Pi\right\|_1 =& \left\|\sum_i d_0(i)\left(\mathbf{P}_\Pi^t(i,\cdot) - d_\Pi\right)\right\|_1 \\
=& \sum_i d_0(i) \left\|\mathbf{P}_\Pi^t(i,\cdot) - d_\Pi\right\|_1 \leq \max_i \|\mathbf{P}_\Pi^t(i,\cdot) - d_\Pi\|_1 \leq 2(1-\varepsilon)^n.
\end{aligned}$$

Take $r_1 = \log \frac{1}{1-\varepsilon}$ finishes the proof. □

Proof of Lemma 5.2. Let $v_t \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(K)}$ be the joint state distribution at time t under policy Π. Using the fact that Π is a fixed policy independent of $\mathbf{g}_{i,t}^{(k)}$ and Assumption 2.2 that the probability transition is also independent of function path given any state and action, the function $\mathbf{g}_{i,t}^{(k)}$ and state-action pair $(a_t^{(k)}, s_t^{(k)})$ are mutually independent. Thus, for any $t \in \{0, 1, 2, \cdots, T-1\}$

$$\mathbb{E}\left(\sum_{k=1}^K g_{i,t}^{(k)}(a_t^{(k)}, s_t^{(k)}) \middle| d_0, \Pi\right) = \sum_{\mathbf{s}\in\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}} \sum_{\mathbf{a}\in\mathcal{A}^{(1)}\times\cdots\times\mathcal{A}^{(K)}} v_t(\mathbf{s})\Pi(\mathbf{a}|\mathbf{s}) \sum_{k=1}^K \mathbb{E}\left(g_{i,t}^{(k)}(a^{(k)}, s^{(k)})\right),$$

where $\mathbf{s} = [s^{(1)}, \cdots, s^{(K)}]$ and $\mathbf{a} = [a^{(1)}, \cdots, a^{(K)}]$ and the latter expectation is taken with respect to $\mathbf{g}_{i,t}^{(k)}$ (i.e. the random variable w_t). On the other hand, by Lemma 2.2, we know that for any randomized stationary policy Π, the corresponding stationary state-action probability can be expressed as $\{\theta_*^{(k)}\}_{k=1}^K$ with $\theta_*^{(k)} \in \Theta^{(k)}$. Thus,

$$\sum_{k=1}^K \left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta^{(k)} \right\rangle = \sum_{\mathbf{s}\in\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}} \sum_{\mathbf{a}\in\mathcal{A}^{(1)}\times\cdots\times\mathcal{A}^{(K)}} d_\Pi(\mathbf{s})\Pi(\mathbf{a}|\mathbf{s}) \sum_{k=1}^K \mathbb{E}\left(g_{i,t}^{(k)}(a^{(k)}, s^{(k)})\right).$$

Hence, we can control the difference:

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left|\mathbb{E}\left(\sum_{k=1}^{K} g_{i,t}^{(k)}(a_t^{(k)}, s_t^{(k)})\Big| d_0, \Pi\right) - \sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta_*^{(k)}\right\rangle\right| \\
\leq &\sum_{t=0}^{T-1}\left|\sum_{\mathbf{s}\in\mathcal{S}^{(1)}\times\cdots\times\mathcal{S}^{(K)}}\ \sum_{\mathbf{a}\in\mathcal{A}^{(1)}\times\cdots\times\mathcal{A}^{(K)}} \left(v_t(\mathbf{s}) - d_\Pi(\mathbf{s})\right)\Pi(\mathbf{a}|\mathbf{s})\right| K\Psi \\
\leq &K\Psi\sum_{t=0}^{T-1}\|v_t - d_\Pi\|_1 \leq 2K\Psi\sum_{t=0}^{T-1} e^{(r_1-t)/r_1} \leq 2eK\Psi\int_0^{T-1} e^{-t/r_1}dt = 2er_1K\Psi,
\end{aligned}$$

where the third inequality follows from Lemma 5.1. Taking $C_1 = 2er_1$ finishes the proof of (29) and (28) can be proved in a similar way.

In particular, we have for any randomized stationary policy Π that satisfies the constraint (2), we have

$$\begin{aligned}
T\cdot\sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta_*^{(k)}\right\rangle \leq &\sum_{t=0}^{T-1}\left|\mathbb{E}\left(\sum_{k=1}^{K} g_{i,t}^{(k)}(a_t^{(k)}, s_t^{(k)})\Big| d_0, \Pi\right) - \sum_{k=1}^{K}\left\langle \mathbb{E}\left(\mathbf{g}_{i,t}^{(k)}\right), \theta_*^{(k)}\right\rangle\right| \\
&+ \sum_{t=0}^{T-1}\mathbb{E}\left(\sum_{k=1}^{K} g_{i,t}^{(k)}(a_t^{(k)}, s_t^{(k)})\Big| d_0, \Pi\right) \leq 2er_1K\Psi + 0 = 2er_1K\Psi,
\end{aligned}$$

finishing the proof. □