A Near-Optimal Primal-Dual Method for Off-Policy Learning in CMDP

Fan Chen
School of Mathematics
Peking University
chern@pku.edu.cn

Junyu Zhang[*]
Department of Industrial Systems Engineering and Management
National University of Singapore
junyuz@nus.edu.sg

Zaiwen Wen
Beijing International Center for Mathematical Research, Center For Machine Learning Research
Peking University
wenzw@pku.edu.cn

Abstract

As an important framework for safe Reinforcement Learning, the Constrained Markov Decision Process (CMDP) has been extensively studied in the recent literature. However, despite the rich results under various on-policy learning settings, there still lacks some essential understanding of the offline CMDP problems, in terms of both the algorithm design and the information theoretic sample complexity lower bound. In this paper, we focus on solving the CMDP problems where only offline data are available. By adopting the concept of the single-policy concentrability coefficient C^*, we establish an $\Omega\left(\frac{\min\{|\mathcal{S}||\mathcal{A}|,|\mathcal{S}|+I\}C^*}{(1-\gamma)^3\epsilon^2}\right)$ sample complexity lower bound for the offline CMDP problem, where I stands for the number of constraints. By introducing a simple but novel deviation control mechanism, we propose a near-optimal primal-dual learning algorithm called DPDL. This algorithm provably guarantees zero constraint violation and its sample complexity matches the above lower bound except for an $\tilde{\mathcal{O}}((1-\gamma)^{-1})$ factor. Comprehensive discussion on how to deal with the unknown constant C^* and the potential asynchronous structure on the offline dataset are also included.

1 Introduction

Reinforcement Learning (RL) is an important tool for modeling the real world tasks that involve sequential decision making. Such RL problems are often mathematically described as a Markov Decision Process (MDP) that maximizes a cumulative sum of rewards. The safe reinforcement learning, on the other hand, not only cares the reward maximization, but also attempts to ensure a reasonable system performance with respect to certain safety constraints. Such safety constrained RL problems are often formulated as the Constrained Markov Decision Process (CMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, u, \gamma, \rho_0)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\gamma \in (0,1)$ is the discount factor, $\mathbb{P}(s' \mid s, a)$ stands for the transition probability from s to s' under the action a for $\forall (s, a, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and $r : \mathcal{S}\times\mathcal{A} \to [-1, 1]$ is the reward function, $(u_i : \mathcal{S}\times\mathcal{A} \to [-1, 1])_{i\in[I]}$ is a set of I utility functions, ρ_0 is the initial state distribution over $\mathcal{S}$. The goal of CMDP is to find

[*] Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

an optimal policy π to maximize the cumulative reward while satisfying a group of constraints:

$$\max_\pi \quad J(\pi) := \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \cdot r\left(s_t, a_t\right) \,\Big|\, s_0 \sim \rho_0, \pi\right] \tag{1}$$

$$\text{s.t.} \quad J_i^u(\pi) := \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \cdot u_i\left(s_t, a_t\right)\right] \geq 0, \text{ for } i \in [I] = \{1, 2, ..., I\}\,.$$

For the CMDP problem, there has been plenty of on-policy algorithms, see [7, 8, 20, etc.]. However, in real world applications such as training physical robots, where safety is an important measure of performance, the real time on-policy interaction with the environment may suffer from the potential damages to the robots. Besides, in many non-simulating environments, the on-policy data collection may also be time-consuming. Therefore, it is crucial to design an off-policy algorithm to solve the CDMP problems, where plenty of historical data are already accumulated while real time interactions are limited. To our best knowledge, offline CMDP algorithms are rare [12, 27, 29], and the sample complexity guarantees are limited. In particular, a strong uniform concentrability assumption is required in [12], and the model-based method [27] mainly considers the case an empirical model is known. Thus it is still not clear how to efficiently solve offline CMDPs with model-free approaches, and there lacks essential understanding of the information theoretic lower bound on the sample complexity of the offline CMDP.

In this paper, we propose a Deviation-controlled Primal-Dual Learning (DPDL) method to solve problem (1). We adopt the primal-dual strategy developed in [4, 16, 26, 35, etc.] as the main algorithmic framework while several non-trivial contributions have been made beyond the existing results. Unlike the aforementioned literatures that exclusively rely on the accessibility of a generative model, DPDL utilizes the offline data, where the distribution shift difficulties of the offline data is tackled by a novel and effective adaptive deviation control mechanism. If the considered CMDP instance has a finite (but potentially unknown) *concentrability coefficient*, DPDL provably finds a policy with $\mathcal{O}(\epsilon)$-optimal reward and zero constraint violation. An information theoretical lower bound on the sample complexity of offline CMDP is also derived in this paper, which indicates that our deviation control mechanism achieves a minimax optimal complexity dependence on $I, |\mathcal{S}|, |\mathcal{A}|, C^*$.

Main Contribution. We summarize the contributions in details as follows.

- We propose the DPDL algorithm to solve the CMDP problem (1). Suppose the CMDP instance satisfies the Slater's condition and certain prior knowledge on the concentrability coefficient C^* is given, DPDL provably finds an ϵ-optimal policy with zero constraint violation using $\tilde{\mathcal{O}}\left(\frac{\min\{|\mathcal{S}||\mathcal{A}|, |\mathcal{S}|+I\}C^*}{(1-\gamma)^4\epsilon^2}\right)$ offline samples.
- We establish an information theoretic sample complexity lower bound of $\Omega\left(\frac{\min\{|\mathcal{S}||\mathcal{A}|, |\mathcal{S}|+I\}C^*}{(1-\gamma)^3\epsilon^2}\right)$ for the offline CMDPs, indicating that DPDL is near optimal up to an $\tilde{\mathcal{O}}((1-\gamma)^{-1})$ factor. The necessity of the Slater's condition for achieving zero constraint violation is also established. The sharp dependence on the number of constraints is mainly captured by our careful construction of the correlated actions.
- In order to handle the practical situation where C^* is unknown, an adaptive version of DPDL is designed with the same sample complexity as DPDL.
- Our analysis of DPDL also extends to the asynchronous case, where the offline dataset consists of a sample trajectory generated by certain behavior policy. In this situation, the sample complexity of DPDL is shown to be $\tilde{\mathcal{O}}\left(\frac{t_{\text{mix}}^2 \min\{|\mathcal{S}||\mathcal{A}|, |\mathcal{S}|+I\}C^*}{(1-\gamma)^4\epsilon^2}\right)$. Our handling of the correlated gradient estimators with large variance can also be beneficial to other algorithms under the asynchronous setting.

Related Work. Recently, considerable efforts have been devoted to the online learning of CMDP. Under the episodic and tabular setting, several works [7, 8, 20] have achieved the $\tilde{\mathcal{O}}\big(\sqrt{|\mathcal{S}|^2|\mathcal{A}|T}\big)$ regret and cumulative constraint violation, with different dependence on the episode length H omitted. Under proper assumptions, zero or bounded cumulative constraint violation can be achieved [1, 17]. In terms of the number of constraints I, MOMA proposed in [34] achieves an $\tilde{\mathcal{O}}\big(\sqrt{\min\{|\mathcal{S}|, I\}I|\mathcal{S}||\mathcal{A}|/T}\big)$ convergence on both average reward gap and constraint violation.

Nevertheless, all the above results adopt the model-based approaches. Except for [34], they either consider the cases where $I = 1$ or completely ignore the influence of I in the sample complexity. Therefore, both deriving an efficient model-free method and obtaining the optimal dependence on I remain open.

Another approach closely related to our paper is the primal-dual method in RL, see [4, 11, 25, 26, 35, etc.]. Given the access to a generative model, the model-free primal-dual method developed in [4] achieves an $\tilde{\mathcal{O}}\big(\frac{I|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\big)$ sample complexity to find an ϵ-optimal safe policy. The deviation control mechanism we develop enables the primal-dual approach to extend beyond the generative model.

Finally, we mention a few related works in the offline RL and safe RL. Previous offline RL algorithms with sample efficiency guarantees typically assume the *uniform concentrability* [12, 18, etc.] or lower bounded *minimum visitation* $\mu_{\min}$ [32, 33, etc.]. Recently, under the less restrictive assumption of the *single-policy concentrability coefficient* C^*, a minimax optimal sample complexity lower bound of $\Omega\big(\frac{|\mathcal{S}|C^*}{(1-\gamma)^3\epsilon^2}\big)$ for discounted offline MDPs is derived in [21]. A similar $\Omega\big(\frac{H^3|\mathcal{S}|C^*}{\epsilon^2}\big)$ lower bound is also derived for the episodic setting in [28]. Under both settings, offline algorithms with $\tilde{\mathcal{O}}(|\mathcal{S}|C^*\epsilon^{-2})$ sample complexity (with different $(1-\gamma)^{-1}$ or H factors omitted) have been discovered with either model-based [15, 21, 28, 31] or model-free approaches [22, 30]. In terms of the offline CMDP problem, the only existing results are [12, 27, 29], where [29] only provides asymptotic convergence, [12] relies on a much stronger uniform concentrability assumption, and [27] is a model based method that potentially suffers an $\mathcal{O}((C^*)^2)$ dependence. Compared to these works, our method is model-free and has an optimal $\mathcal{O}(C^*)$ dependence on the concentrability coefficient.

2 Problem setup

2.1 LP formulation of CMDP problem

For any policy π, the (unnormalized) state-action occupancy measure is defined as

$$\nu^\pi(s,a) := \sum_{t=0}^{+\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \mid s_0 \sim \rho_0, \pi\right), \quad \text{for } \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{2}$$

Given any occupancy measure ν^π, the policy π that generates ν^π can be recovered as

$$\pi(a|s) = \frac{\nu^\pi(s,a)}{\sum_{a'} \nu^\pi(s,a')}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{3}$$

According to [2], it is well known that the set of all state-action occupancy measures form a polyhedron $\big\{\nu \in \mathbb{R}_{\geq 0}^{|\mathcal{S}|\times|\mathcal{A}|} : \sum_{a\in\mathcal{A}} (\mathbf{I} - \gamma\mathbb{P}_a)\nu_a = \rho_0\big\}$, where $\nu_a := (\nu(s,a))_{s\in\mathcal{S}}$ is an $|\mathcal{S}|$-dimensional column vector, $\mathbf{I}$ is the $|\mathcal{S}|\times|\mathcal{S}|$ identity matrix, and $\mathbb{P}_a := (\mathbb{P}(s'|s,a))_{s',s}$ is an $|\mathcal{S}|\times|\mathcal{S}|$ transition matrix, see also [26]. Therefore, combined with the fact that $J(\pi) = \langle \nu^\pi, r\rangle$ and $J_i^u(\pi) = \langle \nu^\pi, u_i\rangle$, the CMDP problem (1) can be reformulated as an LP problem with $|\mathcal{S}|+I$ constraints:

$$\max_{\nu \in \mathbb{R}_{\geq 0}^{|\mathcal{S}|\times|\mathcal{A}|}} \quad \langle \nu, r\rangle \quad \text{s.t.} \quad \sum_{a\in\mathcal{A}} (\mathbf{I} - \gamma\mathbb{P}_a)\nu_a = \rho_0, \quad \langle \nu, u_i\rangle \geq 0,\ \forall i \in [I]. \tag{4}$$

Due to the fundamental theorem of LP, see e.g. [5], problem (4) has an optimal basic feasible solution with at most $|\mathcal{S}| + I$ positive entries, which indicates the following proposition.

Proposition 2.1. *For the CMDP problem* (1) *with I constraints, there is an optimal policy* π^* *such that* $|\mathrm{supp}(\nu^{\pi^*})| \leq \mathcal{N} := \min\{|\mathcal{S}|+I, |\mathcal{S}||\mathcal{A}|\}$*, where* $\mathrm{supp}(\cdot)$ *denotes the support of a vector.*

This result captures the potential sparse structure of the optimal policy when I is not as large as $|\mathcal{S}||\mathcal{A}|$, and is the key to deriving a tight complexity dependence on the number of constraints I.

2.2 Off-policy learning from demonstration

In this work, we consider the offline CMDP problems where the agent cannot interact with the environment. Instead, the optimization is conducted using a fixed offline dataset. To standardize the discussion, we make the following assumption on the offline dataset, see e.g. [21].

Assumption 2.2 (Independent batch dataset). *The batch dataset $\mathcal{D}$ consists of independent tuples $(s, a, s', r, \mathbf{u})$, such that $(s, a) \sim \mu$, $\mathbb{E}[r | s, a] = r(s, a)$, $\mathbb{E}[\mathbf{u}_i | s, a] = u_i(s, a)$, and $s' \sim \mathbb{P}(\cdot | s, a)$, where μ is called the reference distribution.*

To characterize the distribution shift of an arbitrary occupancy measure ν^π from the reference distribution μ, we introduce the following notion of the deviation: $D^\pi := \max_{s,a} \frac{(1-\gamma)\nu^\pi(s,a)}{\mu(s,a)}$, where the $(1-\gamma)$-factor normalizes ν^π to be a distribution. In offline RL, it is natural to assume that the deviation D^{π^*} of the optimal policy is finite. That is, the reference distribution μ fully covers $\mathrm{supp}(\pi^*)$. Otherwise, no optimality can be guaranteed. Combining the sparse nature of the optimal solution of (1), we introduce the following finite concentrability assumption for our problem.

Assumption 2.3. *For $\forall \psi \geq 1$, denote the ψ-deviated policy class as $\Pi(\psi) := \{\pi : \nu^\pi \in D(\psi)\}$ where*

$$D(\psi) := \left\{ \nu \in \mathbb{R}_{\geq 0}^{|\mathcal{S}||\mathcal{A}|} : \max_{s,a} \frac{(1-\gamma)\nu(s,a)}{\mu(s,a)} \leq \psi, \sum_{s,a} \frac{(1-\gamma)\nu(s,a)}{\mu(s,a)} \leq \mathcal{N}\psi \right\}. \tag{5}$$

We assume there exists a finite ψ such that some optimal policy π^* is contained in $\Pi(\psi)$. Let C^* be the minimum of such ψ. We call this constant C^* the (single-policy) concentrability coefficient.

The above assumption includes a sparsity induced constraint as a result of Proposition 2.1, its counterpart in the definition of single-policy concentrability of offline MDP [21] is the deterministic optimal policy. The explicit dependence on $\mathcal{N}$ in $D(\psi)$ facilitates the derivation of the information theoretic lower bound as well as a near-optimal algorithm.

A second remark is that if we know any upper bound ψ of the coefficient C^*, then it will be sufficient to only consider the policies in $\Pi(\psi)$. When C^* is unknown, ψ control the risk of distribution shift. Consequently, in this paper, we propose to solve the LP formulation (4) with a tighter feasible region introduced by $D(\psi)$. This will allow us to properly control the variance of the off-policy sampling when some of $\mu(s, a)$ is extremely small or even zero. We call this strategy *deviation control*.

2.3 Conservatism toward constraints

We say policy π is safe if it satisfies all constraints in (1), and we say π is ϵ-safe if $J_i^u(\pi) \geq -\epsilon$, for $\forall i \in [I]$. Most of the existing online CMDP algorithms guarantee $\mathcal{O}\big(1/\sqrt{T}\big)$ average safeness. To ensure the true safeness (zero constraint violation) in this work, we assume the Slater's condition to hold throughout this paper. In fact, in Section 5, we will show that the Slater's condition is the necessary condition for any offline CMDP algorithm to obtain zero constraint violation.

Assumption 2.4. *There exists $\varphi > 0$ and a policy π such that $J_i^u(\pi) \geq \frac{\varphi}{1-\gamma}$, $\forall i \in [I]$.*

A prior knowledge of such a constant φ is assumed throughout our discussion, and we also assume the Slater's condition holds for $\Pi' := \Pi(C^*)$. Given Assumption 2.4, we leverage the idea of conservative constraints proposed in [4]. Namely, instead of $J_i^u(\pi) \geq 0$, we consider the conservative constraints $J_i^u(\pi) \geq \kappa$ when solving the CMDP problem, where $\kappa > 0$ is a properly chosen parameter that controls the level of conservatism in the constraints. In order to keep the form of the constraints in problem (1), we adopt a shifted utility function u_i^κ defined by $u_i^\kappa(s,a) := u^i(s,a) - (1-\gamma)\kappa$ for $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, $\forall i \in [I]$. Therefore, $J_i^u(\pi) \geq \kappa$ is then equivalent to $J_i^{u^\kappa}(\pi) \geq 0$. It can be shown that a properly selected κ will facilitate a high probability of preserving zero constraint violation, while only introducing an extra $\mathcal{O}\big(\frac{\kappa}{\varphi}\big)$ sub-optimality gap in the reward.

3 The Deviation-controlled Primal Dual Learning (DPDL) algorithm

To solve CMDP with offline samples, we transform its LP formulation (4) to a saddle point form

$$\max_{\nu \in D(\psi)} \min_{\lambda \geq 0, V} \mathcal{L}(V, \lambda, \nu) := \langle r, \nu \rangle + \left\langle V, \rho_0 - \sum_a (\mathbf{I} - \gamma \mathbb{P}_a)\nu_a \right\rangle + \langle \lambda, U_\kappa \nu \rangle , \tag{6}$$

where $D(\psi)$ is defined by (5), $V \in \mathbb{R}^{|\mathcal{S}|}$, $\lambda \in \mathbb{R}^I$ are Lagrangian multipliers, and the matrix U_κ is defined as $U_\kappa := [u_1^\kappa, \cdots, u_I^\kappa]^\top \in \mathbb{R}^{I \times |\mathcal{S}||\mathcal{A}|}$ with u_i^κ being the shifted utility defined in Section 2.3.

Given the reference distribution μ, the objective function can be rewritten as an expectation:

$$\mathcal{L}(V,\lambda,\nu) = \mathop{\mathbb{E}}_{s_0\sim\rho_0}[V(s_0)] + \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mu\\ s'\sim\mathbb{P}(\cdot|s,a)}}\left[\frac{\nu(s,a)}{\mu(s,a)}\left(r(s,a) - (V(s)-\gamma V(s'))+\sum_i \lambda_i u_i^\kappa(s,a)\right)\right].$$

If the reference distribution μ is known, we can directly sample a stochastic gradient of $\mathcal{L}$. However, when the reference distribution μ is unknown in practice, then the importance sampling weight $\frac{\nu(s,a)}{\mu(s,a)}$ is also unknown. To tackle this issue, let $\hat{\mu}$ be a proper estimation of the reference distribution μ, we introduce the weights $w(s,a) = \frac{\mu(s,a)}{\hat{\mu}(s,a)}$, and the diagonal matrix $W = \operatorname{diag}(w(s,a))$. Then we apply a change of variables $x = W^{-1}\nu$, in other words, we set $\frac{x(s,a)}{\hat{\mu}(s,a)} = \frac{\nu(s,a)}{\mu(s,a)}$ for $\forall s, a$ to enable sampling. From now on, we will focus on the following reweighted problem

$$\min_{\lambda\in\Lambda, V\in\mathcal{V}} \max_{x\in\mathcal{X}} \mathcal{L}_w(V,\lambda,x) := \mathcal{L}(V,\lambda,Wx), \tag{7}$$

where the feasible regions are defined as

$$\begin{aligned}
\mathcal{X} &:= \left\{x\in\mathbb{R}_{\geq 0}^{|\mathcal{S}||\mathcal{A}|} : \max_{s,a}\frac{x(s,a)}{\hat{\mu}(s,a)} \leq \frac{\psi}{1-\gamma}, \sum_{s,a}\frac{x(s,a)}{\hat{\mu}(s,a)} \leq \frac{\mathcal{N}\psi}{1-\gamma}, \sum_{s,a} x(s,a) \leq \frac{4}{1-\gamma}\right\},\\
\mathcal{V} &:= \left\{V\in\mathbb{R}^{|\mathcal{S}|} : \|V\|_\infty \leq \frac{8}{1-\gamma}(1+\frac{2}{\varphi})\right\} \quad \text{and} \quad \Lambda =: \left\{\lambda\in\mathbb{R}_{\geq 0}^{I} : \|\lambda\|_1 \leq \frac{8}{\varphi}\right\}.
\end{aligned} \tag{8}$$

The sets $\mathcal{X}$, $\mathcal{V}$ and Λ are chosen to be large enough so that they contain the optimal solution of the problem (6), see detailed discussion in Appendix E. Given a sample $\zeta = (s_0, s, a, s', r, \mathbf{u}) \sim \rho_0\times\mathcal{D}$, and a point $Z := (V,\lambda,x)$, we construct the unbiased gradient estimators for $\mathcal{L}_w(\cdot)$ as

$$\begin{aligned}
\widehat{g}_V(Z;\zeta) &:= \mathbb{I}_{s_0} + \frac{x(s,a)}{\hat{\mu}(s,a)}\left(\gamma\mathbb{I}_{s'} - \mathbb{I}_s\right),\\
\widehat{g}_\lambda(Z;\zeta) &:= \frac{x(s,a)}{\hat{\mu}(s,a)}\mathbf{u}^\kappa,\\
\widehat{g}_x(Z;\zeta) &:= \frac{r+\gamma V(s) - V(s') + \langle\mathbf{u}^\kappa,\lambda\rangle}{\hat{\mu}(s,a)}\mathbb{I}_{s,a},
\end{aligned} \tag{9}$$

where $\mathbb{I}_s$ is the $|\mathcal{S}|$-dimensional unit vector with the s-th element being one, $\mathbb{I}_{s,a}$ is the $|\mathcal{S}||\mathcal{A}|$-dimensional unit vector with the (s,a)-th element being one, and $\mathbf{u}^\kappa = \mathbf{u} - \kappa(1-\gamma)\mathbf{1} \in \mathbb{R}^I$ is the shifted utility vector. Based on these estimators, we propose a stochastic mirror descent ascent approach to solve problem (7), as stated in Algorithm 1.

The algorithm starts from a feasible solution Z^1, which, for example, can be easily chosen as $V^1 = \mathbf{0}$, $\lambda^1 = \frac{\mathbf{1}}{\varphi I}$, $x^1 = \frac{\mathcal{N}}{|\mathcal{S}||\mathcal{A}|}\frac{\hat{\mu}}{1-\gamma}$. In each iteration, an offline sample ζ^t is used to construct the unbiased gradient estimators g_V^t, g_λ^t and g_x^t. A stochastic mirror descent ascent step (11) is then used to update the solution Z^t, where $\operatorname{Proj}_{\mathcal{V}}(\cdot)$ denotes the Euclidean projection to the set $\mathcal{V}$, and $\operatorname{KL}(Y\|Y') := \sum_i Y_i \log\frac{Y_i}{Y_i'} - \sum_i Y_i + \sum_i Y_i'$ denotes the generalized KL divergence. Simple closed form solutions are available to the V^{t+1} and λ^{t+1} updates. By taking the advantage of the special structure of g_x^t and the fact that $x^t \in \mathcal{X}$ is feasible, the x^{t+1} subproblem can be reduced to the root finding of a 1-dimensional monotone function, which can be solved efficiently, see details in Appendix A.

Finally, it is worth noting that $\overline{x}$ is the approximate optimal solution to the reweighted problem. And $W\overline{x}$ will be the approximate solution to the original problem (6) before the change of variable. Therefore, ideally, we should have output the policy $\overline{\pi}_w(a|s) = \frac{w(s,a)\bar{x}(s,a)}{\sum_{a'} w(s,a')\bar{x}(s,a')}$, which is inaccessible in practice without knowing the reference distribution μ. In order to overcome such dilemma, we show that by properly constructing the estimated distribution $\hat{\mu}$, the $\overline{\pi}$ output by Algorithm 1 will be close enough to the ideal output $\overline{\pi}_w$.

4 The sample complexity of DPDL

4.1 Main results of DPDL

For the DPDL algorithm, the convergence and performance guarantee of the output policy $\overline{\pi}$ are summarized as the following theorem.

```
Algorithm 1: Deviation-controlled Primal-Dual Learning algorithm (DPDL)
input : Tolerance $\epsilon > 0$, confidential level $\delta > 0$, conservatism level $\kappa > 0$, stepsize $\eta_t > 0$,
        constants $\alpha_V, \alpha_\lambda, \alpha_x, N_e, \varsigma > 0$, and initial feasible solution $Z^1 = [V^1; \lambda^1; x^1]$.
Obtain $N_e$ samples from $\mathcal{D}$, let $N(s,a)$ be the times that the pair $(s,a)$ appears. Compute
    $$\hat{\mu}(s,a) = \max\left(\frac{N(s,a)}{N_e}, \varsigma\right), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \qquad (10)$$
for $t = 1, \cdots, T-1$ do
    Sample $\zeta_t = (s_t^0, s_t, a_t, s_t', r_t, \mathbf{u}_t)$ from $\rho_0 \times \mathcal{D}$;
    Compute stochastic gradients $g_V^t := \widehat{g}_V(Z^t;\zeta^t), g_\lambda^t := \widehat{g}_\lambda(Z^t;\zeta^t)$, and $g_x^t := \widehat{g}_x(Z^t;\zeta^t)$;
    Compute the stochastic mirror descent ascent update
        $$\begin{aligned} V^{t+1} &= \mathrm{Proj}_{\mathcal{V}}\left(V^t - \eta_t \alpha_V^{-1} g_V^t\right), \\ \lambda^{t+1} &= \operatorname*{arg\,min}_{\lambda \in \Lambda}\left(\left\langle g_\lambda^t, \lambda - \lambda^t \right\rangle + \frac{\alpha_\lambda}{\eta_t}\mathrm{KL}(\lambda \parallel \lambda^t)\right), \\ x^{t+1} &= \operatorname*{arg\,min}_{x \in \mathcal{X}}\left(-\left\langle g_x^t, x - x^t \right\rangle + \frac{\alpha_x}{\eta_t}\mathrm{KL}(x \parallel x^t)\right), \end{aligned} \qquad (11)$$
Compute the average iterate $\overline{x} = \frac{1}{T}\sum_{t=1}^T x^t, \overline{V} = \frac{1}{T}\sum_{t=1}^T V^t, \overline{\lambda} = \frac{1}{T}\sum_{t=1}^T \lambda^t$;
Compute $\overline{\pi}(a|s) = \frac{\overline{x}(s,a)}{\sum_{a'} \overline{x}(s,a')}$, for all $(s,a)$;
output: Policy $\overline{\pi}$ and the approximate solution $\overline{x}$.
```

Theorem 4.1. *Suppose that Algorithm 1 runs with $\eta_t \equiv \frac{1}{\sqrt{T}}$, $\kappa = 5\varphi\epsilon$, $\alpha_\lambda = \frac{1}{1-\gamma}\sqrt{\frac{\psi}{\log I}}$, $\alpha_V = \varphi\sqrt{\frac{\psi}{|\mathcal{S}|}}$, $\alpha_x = \frac{1}{\varphi(1-\gamma)}\sqrt{\frac{\mathcal{N}\psi}{\log\psi}}$, and $\psi \geq C^*$. Then for any fixed $\epsilon \in \left(0, \frac{1}{10(1-\gamma)}\right]$, and $T \geq c_o\frac{\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4\epsilon^2}$, where $\iota = \log\left(\frac{\psi|\mathcal{S}||\mathcal{A}|I}{\delta}\right)$ and c_o is a universal constant, the output policy $\overline{\pi}$ of DPDL satisfies the following with probability at least $1-\delta$*

$$J(\pi^*) - J(\overline{\pi}) \leq \mathcal{O}(\epsilon), \quad \textit{and} \quad J_i^u(\overline{\pi}) \geq 0, \forall i \in [I].$$

When $\psi = \mathcal{O}(C^*)$, DPDL needs at most $\tilde{\mathcal{O}}\left(\frac{\mathcal{N}C^*}{\varphi^2(1-\gamma)^4\epsilon^2}\right)$ samples to find a safe $\mathcal{O}(\epsilon)$-optimal policy.

Remark 4.2. *When the prior knowledge of C^* is not available, and the selected parameter $\psi < C^*$ but the Slater's condition for $\Pi(\psi)$ still holds, the output policy $\overline{\pi}$ of DPDL satisfies that*

$$J(\overline{\pi}) \geq \max_{\pi \in \Pi(\psi) \cap \mathfrak{S}} J(\pi) - \mathcal{O}(\epsilon) \qquad \textit{and} \qquad J_i^u(\overline{\pi}) \geq -\epsilon_{\text{approx}}, \forall i \in [I],$$

where $\mathfrak{S}$ denotes the set of safe policies, and $\epsilon_{\text{approx}}(\psi) := J(\pi^*) - \max_{\pi \in \Pi(\psi) \cap \mathfrak{S}} J(\pi)$ in some sense measures the "sub-optimality" of the policy class $\Pi(\psi)$. In case a fixed sub-optimality gap ϵ is given, such difficulty of unknown C^* also appears in the guarantees provided in previous works [15, 21, 22, 28, 30, 31].

A simple approach to resolve the difficulty of an unknown C^* is discussed later in Section 6.

4.2 The analysis of DPDL

We break down the analysis of Theorem 4.1 into the following steps. First of all, we provide a proper choice of N_e and ς so that $\hat{\mu}$ is close enough to μ. See proof in Appendix B.

Proposition 4.3. *Denote $\epsilon_e = \frac{\epsilon}{100}$, and let $\varsigma = \frac{\varphi(1-\gamma)^2\epsilon_e}{2\mathcal{N}\psi}$, and $N_e \geq \frac{512\mathcal{N}\psi}{\varphi^2(1-\gamma)^4\epsilon_e^2} \cdot \log\left(\frac{6|\mathcal{S}||\mathcal{A}|}{\delta}\right)$. Then with probability at least $1 - \delta/3$, the estimated reference distribution $\hat{\mu}$ defined by* (10) *satisfies the following properties simultaneously:* ***(1).*** *$\frac{\mu(s,a)}{\hat{\mu}(s,a)} \leq 2$, and $\hat{\mu}(s,a) \geq \varsigma$, for all s, a;* ***(2).*** *For any $\pi \in \Pi(\psi)$, $W^{-1}\nu^\pi \in \mathcal{X}$;* ***(3).*** *For any $x \in \mathcal{X}$, $\|Wx - x\|_1 \leq \varphi(1-\gamma)\epsilon_e$.*

All the rest of our analyses are all conditioning on the success of Proposition 4.3. It is worth noting that in Proposition 4.3, (3) clarifies the validity of constructing the output policy $\overline{\pi}$ with $\overline{x}$ instead of $W\overline{x}$; (2) explains why the feasible region $\mathcal{X}$ is defined as (8); and (1), combined with the carefully specified feasible domains, provides the proper upper bounds on the magnitude and variance of the unbiased gradient estimators in (9). A very detailed discussion is provided in Appendix C. In particular, for the $\widehat{g}_x(\cdot)$ estimator, an explicit $\mathcal{O}(\mathcal{N})$ dependence has been established for both the magnitude and variance, which plays a crucial role in deriving the optimal $\mathcal{O}(\min\{|\mathcal{S}||\mathcal{A}|, |\mathcal{S}|+I\})$ dependence on $|\mathcal{S}|$, $|\mathcal{A}|$ and I. Let us define the following gap to measure the performance of the output $\bar{x}$ w.r.t. problem (7):

$$\mathrm{Gap}(\overline{x}) := \max_{x\in\mathcal{X}} \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,x) - \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,\overline{x}). \tag{12}$$

Based on the properly bounded gradient estimators, a high probability bound for $\mathrm{Gap}(\overline{x})$ is established in the following theorem. Its proof is detailed in Appendix D.

Theorem 4.4. *Suppose the constants η_t, α_V, α_λ, α_x and κ are chosen the same as Theorem 4.1. Then there is a universal constant c_o such that, as long as $T \geq c_o\frac{\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4\epsilon^2}$, the output $\overline{x}$ satisfies $\mathrm{Gap}(\overline{x}) \leq \frac{\epsilon}{2}$ with probability at least $1-\delta/3$.*

Given Theorem 4.4, we finalize the proof of Theorem 4.1 by properly transforming the bound on $\mathrm{Gap}(\overline{x})$ to the expected reward gap and the constraint violation on the original CMDP problem (1), which is discussed in details in Appendix E.

4.3 Extension to asynchronous setting

In some situations, an independent dataset that satisfies Assumption 2.2 may not be available. Instead, the dataset may have the following asynchronous structure.

Assumption 4.5. *The asynchronous dataset $\mathcal{D}_{async}$ is a single sample trajectory generated by some behavior policy π_b. Namely, what we observe is a sequence $\{s_t, a_t, r_t, \mathbf{u}_t\}_{t\geq 1}$ generated under π_b. We assume the Markov Chain $\{(s_t,a_t)\}_{t\geq 1}$ is irreducible, aperiodic and uniformly ergodic, with the stationary distribution μ and the mixing time $t_{\mathrm{mix}} < +\infty$.*

The asynchronous data structure introduced here is frequently considered in RL, for example, the asynchronous Q-learning [14]. However, to our best knowledge, this type of offline data has yet been considered under the assumption of a finite single-policy concentrability. In this situation, we set $\zeta_t = (s_t^0, s_t, a_t, s_{t+1}, r_t, \mathbf{u}_t)$ in the DPDL method (Algorithm 1), where $s_t^0 \sim \rho_0$ and $(s_t, a_t, s_{t+1}, r_t, \mathbf{u}_t)$ is the tuple in the t-th time step of the asynchronous dataset. The sample complexity of the DPDL Algorithm under Assumption 4.5 is established as follows.

Theorem 4.6. *Under Assumption 4.5, we follow the choice of constants in Theorem 4.1. Then given any fixed $\epsilon \in \left(0, \frac{1}{10(1-\gamma)}\right]$, $\psi \geq C^*$, and $T \geq c_o'\frac{t_{\mathrm{mix}}^2\mathcal{N}\psi\iota^3}{\varphi^2(1-\gamma)^4\epsilon^2}$, the output policy $\overline{\pi}$ of DPDL satisfies the following with probability at least $1-\delta$*

$$J(\pi^*) - J(\overline{\pi}) \leq \epsilon \qquad \textit{and} \qquad J_i^u(\overline{\pi}) \geq 0, \forall i \in [I].$$

Here $\iota = \log\left(T|\mathcal{S}||\mathcal{A}|I/\delta\right)$ and c_o' is a universal constant. Therefore, when $\psi = \mathcal{O}(C^*)$, DPDL needs at most $\tilde{\mathcal{O}}\left(\frac{t_{\mathrm{mix}}^2\mathcal{N}C^*}{\varphi^2(1-\gamma)^4\epsilon^2}\right)$ samples to find a safe ϵ-optimal policy.

The main framework for proving Theorem 4.6 is similar to that in Section 4.2, thus we present the proof in the Appendix H. However, compared to the synchronous setting, a key difficulty here is that the gradient estimators $\widehat{g}_V(Z^t;\zeta_t)$, $\widehat{g}_\lambda(Z^t;\zeta_t)$, and $\widehat{g}_x(Z^t;\zeta_t)$ are no longer unbiased, because the samples $\{\zeta_t\}_{t=1}^T$ are obtained from a sample path. This brings further difficulties in the analysis because the variance of the estimators can be amplified by the correlation between samples.

The basic idea to deal with this difficulty is to leverage the mixing property of the uniformly ergodic Markov chain. Take the $\widehat{g}_x(\cdot)$ estimator for example, the bias can be well controlled as long as T is selected larger than the mixing time t_{mix} of the sample path, which can be illustrated by the following decomposition

$$\begin{aligned}\widehat{g}_x(Z^t;\zeta_t) - \nabla_x\mathcal{L}_w(Z^t) \;=\;& \underbrace{\widehat{g}_x(Z^t;\zeta_t) - \widehat{g}_x(Z^{t-\tau};\zeta_t) + \nabla_x\mathcal{L}_w(Z^{t-\tau}) - \nabla_x\mathcal{L}_w(Z^t)}_{\text{order } \mathcal{O}(\tau\eta)} \\ &+ \underbrace{\widehat{g}_x(Z^{t-\tau};\zeta_t) - \mathbb{E}\left[\widehat{g}_x(Z^{t-\tau};\zeta_t)\middle|\, Z^{t-\tau}\right]}_{\text{zero mean}} + \underbrace{\mathbb{E}\left[\widehat{g}_x(Z^{t-\tau};\zeta_t)\middle|\, Z^{t-\tau}\right] - \nabla_x\mathcal{L}_w(Z^{t-\tau})}_{\text{order } \mathcal{O}(\exp(-\tau/t_{\mathrm{mix}}))}.\end{aligned} \tag{13}$$

When $t = \tilde{\Omega}\left(t_{\text{mix}}\right)$, one can bound the bias of $\widehat{g}_x(Z^t;\zeta_t)$ by $\tilde{\mathcal{O}}\left(t_{\text{mix}}\eta\right)$ with suitably chosen τ.

5 Lower Bound of Sample Complexity for Learning CMDP

In this section we will discuss whether the DPDL Algorithm is the near-optimal and whether the Slater's condition (Assumption 2.4) is necessary in achieving zero constraint violation. We answer these questions affirmatively by establishing the following theorems.

Theorem 5.1. *Suppose $S \geq 4$, $A \geq 3$, $I \geq 8$, $C \geq 2$, $\gamma \in [\frac{1}{2}, 1)$, $N \geq 1$. For any learning algorithm $\mathfrak{A}$, there exists a CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, (u_i)_{i\in[I]}, \gamma, \rho_0)$ and a reference distribution μ, such that the following hold true.*

(1) $|\mathcal{S}| \leq 4S + 1$, $|\mathcal{A}| \leq A$, and the concentrability coefficient C^* for $\mathcal{M}$ and μ satisfies $C^* \leq C$.

(2) Let $\hat{\pi}$ be the policy output by $\mathfrak{A}$ given N offline samples from μ, and let π^* be the optimal policy, then at least one of the following two inequalities hold true:

$$\mathbb{E}_{\mathcal{M},\mathfrak{A}}[J(\pi^*) - J(\hat{\pi})] \gtrsim \min\left\{\frac{1}{1-\gamma}, \sqrt{\frac{\min\{SA, S+I\}\,C}{(1-\gamma)^3 N}}\right\}, \quad \textit{and} \quad \mathbb{E}_{\mathcal{M},\mathfrak{A}}\big[\text{violation}(\hat{\pi})\big] \gtrsim 1,$$

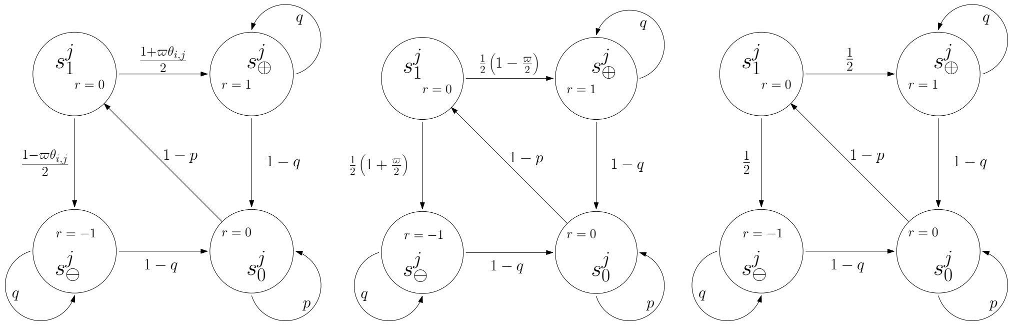
where $\text{violation}(\hat{\pi}) := \sum_{i=1}^{I}\left[J_i^u(\hat{\pi})\right]_-$*, and J_i^u is the utility w.r.t. the constraints $J_i^u \geq 0, \forall i \in [I]$.*

For DPDL, the constraint violation is guaranteed to be zero with high probability, then only the first inequality is valid for our method, which indicates an $\Omega\big(\frac{\mathcal{N}C^*}{(1-\gamma)^3\epsilon^2}\big)$ sample complexity lower bound. Therefore, the complexity of DPDL is nearly optimal up to an $\tilde{\mathcal{O}}\big(\frac{1}{1-\gamma}\big)$ factor. Besides the lower bound, we also establish the necessity of the Slater's condition in ensuring zero violation.

Theorem 5.2. *Let S, A, C, γ be the same as Theorem 5.1. For any algorithm $\mathfrak{A}$, there exists a CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, (u_i)_{i\in[I]}, \gamma, \rho_0)$ with $I = 1$, $|\mathcal{S}| \leq S$, $|\mathcal{A}| \leq A$ and a reference distribution μ with $C^* \leq C$, such that $\mathbb{E}_{\mathcal{M},\mathfrak{A}}[\text{violation}(\hat{\pi})] \gtrsim \min\left\{\frac{1}{1-\gamma}, \sqrt{\frac{SC}{(1-\gamma)^3 N}}\right\}$, where $\hat{\pi}$ is the output policy of $\mathfrak{A}$ given N samples from μ.*

Theorem 5.2 is obtained by utilizing the same idea as Theorem 5.1. Thus we only discuss the derivation of Theorem 5.1, while moving all the details to Appendix F.

For offline CMDPs, the fixed data distribution μ fully dominates the frequency of exploring the state-action pairs. Therefore, intuitively, the hard CMDP instances will be the ones with a large support $\text{supp}(\nu^{\pi^*})$ that widely spreads across the less frequently visited station-action pairs of μ. Based on this intuition, we design a basic block of CMDP presented in Fig. 1, which is essentially a constrained bandit with $2K+1$ arms. The instance $\mathcal{M}$ will be S replicas of the basic blocks, plus an extra "null" state s_{-1} to control C^*. In this discussion, we only consider the case where $I \simeq KS$, the more general construction that cover full range of I is presented in the appendix.


s_1^j
$r=0$
$\frac{1+\varpi\theta_{i,j}}{2}$
$s_\oplus^j$
$r=1$
q
$\frac{1-\varpi\theta_{i,j}}{2}$
$1-p$
$1-q$
$r=-1$
$s_\ominus^j$
$1-q$
$r=0$
s_0^j
q
p
$\frac{1}{2}\left(1-\frac{\varpi}{2}\right)$
$\frac{1}{2}\left(1+\frac{\varpi}{2}\right)$
$\frac{1}{2}$
$\frac{1}{2}$

(a) (s_1^j, a_i) (b) (s_1^j, b_i) (c) (s_1^j, e)

Figure 1: Transition dynamics of the jth replica under different actions, $i \in [K]$.

State, action and transition. At the states $s_\oplus^j, s_\ominus^j, s_0^j$, there is no action to be taken. At each state s_1^j, there are $2K+1$ actions $a_1, b_1, \cdots, a_K, b_K, e$. The transition dynamics of the jth replica under different actions are illustrated in Fig. 1 where the directed arcs and the numbers associated with them are the transitions and the corresponding probabilities, where $p = \frac{1}{2-\gamma}$ and $q = 2 - \frac{1}{\gamma}$ are some constants, while ϖ and $\theta_{i,j} \in \{-1, 1\}, \forall i, j$ are parameters to be designed.

Constraints and Reward. By carefully selecting the u_i's, one can construct a set of $I = 2SK$ constraints that indicate $\pi(a_i|s_1^j) \le \pi(b_i|s_1^j) \le \frac{1}{4K}, \forall i, j$. For the reward, we set $r(s_1^j) = r(s_0^j) = 0$, $r(s_\oplus^j) = 1$, and $r(s_\ominus^j) = -1$, regardless of the actions. At any replica j, we can view a_i, b_i, and e as bandit arms with (cumulative) reward $c\varpi\theta_{i,j}$, $-\frac{c\varpi}{2}$, and 0 respectively, for some $c > 0$. When $\theta_{i,j} = –1$, one would rather pick e . But when $\theta_{i,j} = 1$, due to the constraint $\pi(a_i|s_1^j) \le \pi(b_i|s_1^j) \le \frac{1}{4K}$, picking a_i and b_i with equal probability $\frac{1}{4K}$ will be optimal. In fact, this $\frac{1}{4K}$ upper bound forces the support of the optimal policy to widely spread across the (i, j)'s where $\theta_{i,j} = 1$, and the task of learning is essentially determining whether $\theta_{i,j} = 1$ for each (i, j).

Optimal policy. Based on the above discussion, it is not hard to see that the unique optimal policy is $\pi^{*,\theta}(a_i|s_1^j) = \pi^{*,\theta}(b_i|s_1^j) = \frac{\mathbb{I}\{\theta_{i,j}=1\}}{4K}$ and $\pi^{*,\theta}(e|s_1^j) = 1 - \frac{1}{2K}\sum_{i=1}^K \mathbb{I}\{\theta_{i,j} = 1\}$.

Finally, with the above $\pi^{*,\theta}$ and a proper initial distribution ρ_0, the occupancy measure can be explicitly computed and a reference distribution μ with concentrability coefficient $C^* \le C$ can be designed. Moreover, for any policy $\hat{\pi}$, we consider $\hat{\theta}_{i,j}(\hat{\pi}) := 8K\hat{\pi}(a_i|s_1^j) - 1$, then

$$\mathcal{L}(\hat{\pi}; \theta) := \left[J(\pi^{*,\theta}; \theta) - J(\hat{\pi}; \theta)\right]_+ + \frac{\gamma\varpi}{1-\gamma}\text{violation}(\hat{\pi}; \theta) \ge \frac{\gamma^2\varpi\|\hat{\theta}(\hat{\pi}) - \theta\|_1}{64KS(1-\gamma)}.$$

Namely, if $\hat{\theta}(\hat{\pi})$ is not close enough to the underlying parameter θ, the policy $\hat{\pi}$ will incur a considerable reward gap or constraint violation. By setting $\varpi = \min\left\{\sqrt{\frac{(SK-3)C}{16(1-\gamma)N}}, \frac{1}{2}\right\}$ to be a small enough number, any two CMDP instances with different θ parameters will be non-distinguishable, given N samples from μ. According to [9] and [24], there exists a subset $\Theta \subseteq \{-1, 1\}^{SK}$ such that $|\Theta| \ge \exp(SK/8)$, and $\|\theta - \theta'\|_1 \ge \frac{SK}{2}$ for any pair of different $\theta, \theta' \in \Theta$. In other words, there will be at least $\exp(SK/8)$ CMDP instances with different enough θ parameters while being non-distinguishable under N samples. Then the rest of the arguments will follow by applying the generalized Fano's inequality [3]. A detailed proof is provided in Appendix F.

6 Adaptive deviation-control framework of DPDL

We should notice that in both Theorems 4.1 and 4.6, it has been explicitly emphasized that a prior belief $\psi \ge C^*$ is required. Otherwise, both the reward and the constraints will suffer an extra loss of $\epsilon_{\text{approx}}(\psi)$. In this section, we propose an adaptive deviation-control framework (Algorithm 2) to handle the practical situation where no such prior knowledge is available.

Algorithm 2: The Adaptive-DPDL framework

```
input : Sub-optimality $\epsilon$, confidence level $\delta$.
Initialize $\psi_1$, default $J^K \equiv -\infty$, for $K = 0, 1, 2, ...$;
for $K = 1, 2, \cdots$ do
    Call DPDL with $\psi = \psi_K$, obtain an approximate solution $x^{(K)}$ and the policy $\pi^{(K)}$;
    if VERIFY $\left(x^{(K)}; \epsilon, \delta\right)$ == TRUE then
        Compute $\widehat{J}(\pi^{(K)})$ as an $\mathcal{O}(\epsilon)$-accurate estimator of $J(\pi^{(K)})$, set $J^K = \widehat{J}(\pi^{(K)})$;
    if $-\infty < J^K \le J^{K-1} + \mathcal{O}(\epsilon)$ then Terminate;
    Set $\psi_{K+1} = 2\psi_K$;
output: Policy $\pi^{(K)}$.
```

At a high-level, Algorithm 2 consists of the following steps.

Verification For the output $\overline{x}$ of the DPDL, we develop a verification method VERIFY$(\overline{x}; \epsilon, \delta)$ that, with probability at least $1 - \delta$, returns TRUE only when following two statements hold: (1). The vector $\overline{\nu} := W\overline{x}$ satisfies $\|\sum_a(\mathbf{I} - \gamma\mathbb{P}_a)\overline{\nu}_a - \rho_0\|_1 = \mathcal{O}(\epsilon)$, which essentially checks whether $\overline{\nu}$ is approximately a valid occupancy measure; (2). The policy $\overline{\pi}$ induced by $\overline{x}$ is safe. At step K, if any one of the two statements does not hold, we immediately know $\psi_K < C^*$ due to the analysis of Theorem 4.1. Consequently, we to double the coefficient $\psi_{K+1} \leftarrow 2\psi_K$ in the next iteration.

Certifying performance improvement When VERIFY$(\overline{x}; \epsilon, \delta)$ returns TRUE, then it holds that $j_0(\psi) = J(\pi^{(K)}) + \mathcal{O}(\epsilon)$, where $j_0(\psi)$ denotes the optimal value of problem (7) with $\kappa = 0$. That is, one can estimate $j_0(\psi)$ with $\widehat{J}(\pi^{(K)})$ if VERIFY$(\overline{x}; \epsilon, \delta) =$ TRUE. As long as VERIFY returns TRUE for two consecutive runs, and the performance improvement is small, i.e., $j_0(\psi_K) - j_0(\psi_{K-1}) = \mathcal{O}(\epsilon)$, then Lemma 6.1 guarantees that the safe policy $\pi^{(K)}$ is $\mathcal{O}(\frac{C^*}{\psi_K}\epsilon)$-optimal.

Lemma 6.1. *The function $j_0(\cdot)$ is strictly increasing in the range $\psi \in [1, C^*]$, and $j_0(\psi) = J(\pi^*)$ for $\psi \geq C^*$. For any $\psi < \psi' \leq C^*$, it holds that*

$$J(\pi^*) - j_0(\psi') \leq \frac{C^* - \psi}{\psi' - \psi}\left(j_0(\psi') - j_0(\psi)\right).$$

Detailed descriptions of VERIFY and Adaptive-DPDL are presented in Appendix G, and so does the proof of the following theorem.

Theorem 6.2. *Fixed $\epsilon \in \left(0, \frac{1}{10(1-\gamma)}\right], \delta \in (0,1)$. Then with probability at least $1-\delta$, Adaptive-DPDL stops at step K such that $\psi_K \leq 4C^*$ and outputs the safe policy $\pi^{(K)}$ with sub-optimality gap $J(\pi^*) - J(\pi^{(K)}) \leq \mathcal{O}\left(\frac{C^*}{\psi_K}\epsilon\right)$. Moreover, there exists a (problem dependent) constant $\epsilon_0(\mathcal{M})$ such that, if $\epsilon \leq \epsilon_0(\mathcal{M})$, then it must hold that $\psi_K \in [C^*, 2C^*)$ and $\pi^{(K)}$ is ϵ-optimal.*

Intuitively, the Adaptive-DPDL will quickly terminate within $\mathcal{O}(\log_2 C^*)$ calls of DPDL, resulting in a total samples complexity of $\tilde{\mathcal{O}}\left(\frac{\mathcal{N}C^*}{(1-\gamma)^4\epsilon^2}\right)$.

Acknowledgement

J. Zhang is supported by Ministry of Education, Singapore, grant R-266-000-158-133. Z. Wen is supported in part by the NSFC grant 11831002. F. Chen is supported in part by the elite undergraduate training program of School of Mathematical Sciences in Peking University.

References

[1] Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *arXiv preprint arXiv:2109.05439*, 2021.

[2] Eitan Altman. *Constrained Markov decision processes*. PhD thesis, INRIA, 1995.

[3] Bin Yu Assouad. Fano, and le cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1996.

[4] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. *arXiv preprint arXiv:2109.06332*, 2021.

[5] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[6] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

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

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

[9] Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31 (3):504–522, 1952.

[10] Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

[11] Angeliki Kamoutsi, Goran Banjac, and John Lygeros. Efficient performance bounds for primal-dual reinforcement learning from demonstrations. In *International Conference on Machine Learning*, pages 5257–5268. PMLR, 2021.

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

[13] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[14] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.

[15] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

[16] Yongfeng Li, Mingming Zhao, Weijie Chen, and Zaiwen Wen. A stochastic composite augmented Lagrangian method for reinforcement learning. *arXiv preprint arXiv:2105.09716*, 2021.

[17] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[19] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

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

[21] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

[23] Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[24] Rom Rubenovich Varshamov. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117:739–741, 1957.

[25] Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

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

[27] Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. *Advances in Neural Information Processing Systems*, 34, 2021.

[28] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

[29] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized Q-learning for safe offline reinforcement learning. *arXiv preprint arXiv:2107.09003*, 2021.

[30] Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

[31] Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

[32] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*, 2020.

[33] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34, 2021.

[34] Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. In *International Conference on Machine Learning*, pages 12167–12176. PMLR, 2021.

[35] Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Cautious reinforcement learning via distributional risk in the dual domain. *IEEE Journal on Selected Areas in Information Theory*, 2(2):611–626, 2021.

A Efficiently solving the subproblems of DPDL

In this section, we describe how to efficiently solve the subproblems (11) in the DPDL Algorithm. In the following discussion, at most $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}| + I)$ flops are needed to compute the update.

A.1 Closed form solution for the V-update

The dual variable V is updated by the formula $V^{t+1} = \mathrm{Proj}_{\mathcal{V}}\left(V^t - \eta_t \alpha_V^{-1} g_V^t\right)$, where $\mathcal{V}$ is an ℓ_∞ normal ball defined as $\mathcal{V} := \left\{V \in \mathbb{R}^{|\mathcal{S}|} : \|V\|_\infty \le R_{\mathcal{V}}\right\}$, $R_{\mathcal{V}} = \frac{8}{1-\gamma}(1 + \frac{2}{\varphi})$. For any vector $V \in \mathbb{R}^{|\mathcal{S}|}$, the Euclidean projection $V_+ = \mathrm{Proj}_{\mathcal{V}}(V)$ can be written as a simple truncation

$$V_+(s) = \begin{cases} -R_{\mathcal{V}}, & \text{if } V(s) < -R_{\mathcal{V}}, \\ V(s), & \text{if } -R_{\mathcal{V}} \le V(s) \le +R_{\mathcal{V}}, \\ +R_{\mathcal{V}}, & \text{if } V(s) > +R_{\mathcal{V}}, \end{cases} \qquad \text{for} \qquad \forall s \in \mathcal{S}.$$

This update will need $\mathcal{O}(1)$ flops due to the special structure of g_V^t.

A.2 Closed form solution for the λ-update

The dual variable λ is updated by the formula $\lambda^{t+1} = \arg\min_{\lambda \in \Lambda} \left(\langle g_\lambda^t, \lambda - \lambda^t\rangle + \frac{\alpha_\lambda}{\eta_t}\mathrm{KL}(\lambda || \lambda^t)\right)$, where Λ is the nonnegative part of an ℓ_1 norm ball $\Lambda = \left\{\lambda \in \mathbb{R}^I_{\ge 0} : \|\lambda\|_1 \le R_\Lambda\right\}$, $R_\Lambda = \frac{8}{\varphi}$. The solution to this subproblem has the following closed form formula

$$\lambda^{t+1} = \lambda^{t+\frac{1}{2}} \min\left\{\frac{R_\Lambda}{\left\|\lambda^{t+\frac{1}{2}}\right\|_1}, 1\right\},$$

where $\lambda^{t+\frac{1}{2}} = \lambda^t \exp(-\frac{\eta_t}{\alpha_\lambda} g_\lambda^t)$ is an intermediate point. This update will need $\mathcal{O}(I)$ flops.

A.3 Efficient implementation of the x-update

Compared to the previous two updates, the subproblem for x-update does not have a closed form solution. By carefully discussing the KKT condition of the problem and utilizing the special structure of $\mathcal{X}$ and g_x^t, we reduce the problem to finding the root of a monotonically decreasing 1-dimensional function. If the bisection method is applied to find the root, then in total $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|)$ flops are needed. We present the details as follows. For notational simplicity, we rewrite the subproblem as follows.

Problem. *Given a set $\mathcal{Y}$ defined by the linear constraints*

$$\mathcal{Y} := \Big\{y \in \mathbb{R}^n : 0 \le y_i \le a_i, \sum_{i=1}^n y_i \le B_1, \sum_{i=1}^n c_i y_i \le B_2\Big\},$$

where $B_1, B_2 > 0$, and $c_i > 0$ are some constants. Let $y^0 \in \mathcal{Y}$, $y^0 > 0$, and let $g \in \mathbb{R}^n$ be a vector that has at most 1 non-zero entry. Then the goal is to solve

$$y^* = \arg\min_{y \in \mathcal{Y}} \left(\langle y, g\rangle + \mathrm{KL}(y \parallel y^0)\right). \tag{14}$$

Without loss of generality, we assume $g_2 = \cdots = g_n = 0$. For problem (14), we introduce two Lagrangian multipliers to the coupling constraints $\sum_{i=1}^n y_i \le B_1, \sum_{i=1}^n c_i y_i \le B_2$, while remaining the coordinately separable constraints $0 \le y_i \le a_i$ in the problem. Thus we get the following Lagrangian function:

$$L(y, \alpha, \beta) := y_1 g_1 + \mathrm{KL}(y||y^0) + \alpha\big(\sum_i y_i - B_1\big) + \beta\big(\sum_i c_i y_i - B_2\big). \tag{15}$$

By the strong convexity of KL divergence, there is a unique KKT point (y^*, α^*, β^*) of problem (14). Note that $y^* = \arg\min_{y_i \in [0, a_i], \forall i} L(y, \alpha^*, \beta^*)$. Because $y_0^i > 0$, we know

$$\lim_{y_i \to 0+} \nabla_{y_i} L(y, \alpha^*, \beta^*) = \lim_{y_i \to 0+} g_1 \cdot \mathbb{I}\{i = 1\} + \alpha^* + c_i\beta^* + \log y_i - \log y_i^0 = -\infty,$$

we know y_i^* will not be 0. Thus we can write the KKT condition for problem (14) as

$$\begin{cases} \nabla_{y_i} L(y^*, \alpha^*, \beta^*) \leq 0, & \text{if } y_i^* = a_i, \quad \forall i \in [n], \\ \nabla_{y_i} L(y^*, \alpha^*, \beta^*) = 0, & \text{if } y_i^* \in (0, a_i), \quad \forall i \in [n], \\ \alpha^* \big(\sum_i y_i^* - B_1 \big) = 0, & \beta^* \big(\sum_i c_i y_i^* - B_2 \big) = 0, \\ y^* \in \mathcal{Y}, \alpha^* \geq 0, \beta^* \geq 0. \end{cases} \tag{16}$$

For $i = 2, ..., n$, the condition $\nabla_{y_i} L(y^*, \alpha^*, \beta^*) \leq 0$ implies that $y_i^* \leq y_i^0 \exp(-\alpha^* - c_i \beta^*)$. Note that $\alpha^*, \beta^* \geq 0, c_i > 0$, $y_i^0 \leq a_i$. If $y_i^* < a_i$, then $\nabla_{y_i} L(y^*, \alpha^*, \beta^*) = 0$ indicates that $y_i^* = y_i^0 \exp(-\alpha^* - c_i \beta^*)$. If $y_i^* = a_i$, then the only possibility is $y_i^0 = a_i$ happen to hold and $\alpha^* = \beta^* = 0$, in this case, we still have $y_i^* = y_i^0 \exp(-\alpha^* - c_i \beta^*)$. A similar formula can also be derived for y_1^*. Therefore, utilizing the feasibility of the point y^0, we solve the first two rows of the KKT condition and get

$$\begin{cases} y_1^*(\alpha^*, \beta^*) = \min \left\{ y_1^0 \exp(-g_1 - \alpha^* - c_1 \beta^*), a_1 \right\}, \\ y_i^*(\alpha^*, \beta^*) = y_i^0 \cdot \exp\{-\alpha^* - c_i \beta^*\}, \quad \text{for } i = 2, ..., n. \end{cases} \tag{17}$$

Here, we write y_i^* as functions of α^*, β^* for the ease of later discussion. Next, we solve the third row of the KKT condition (16) by considering the following cases.

Case 1: $\beta^* = 0, \alpha^* = 0$**.** In this case, if $y^*(0,0), \alpha^* = 0, \beta^* = 0$ satisfies (16), then $y^*(0,0)$ is the solution to (14). Otherwise we conclude that $\alpha^* = \beta^* = 0$ is not true.

Case 2: $\beta^* = 0, \alpha^* > 0$**.** In this case, the KKT condition tells us that $\sum_i y_i^* = B_1$. Together with (17), we have the following two possible solutions to α^*

$$\begin{cases} \alpha_1 = \ln \left(\frac{y_2^0 + \cdots + y_n^0}{B_1 - a_1} \right), & \text{corresponds to } y_1^* = a_1, \\ \alpha_2 = \ln \left(\frac{e^{-g_1} \cdot y_1^0 + y_2^0 + \cdots + y_n^0}{B_1} \right), & \text{corresponds to } y_1^* = y_1^0 \exp(-g_1 - \alpha^*). \end{cases}$$

Then if $y^*(\alpha_1, 0), \alpha^* = \alpha_1, \beta^* = 0$ satisfies (16), we conclude that $y^*(\alpha_1, 0)$ is the solution to (14). If $y^*(\alpha_2, 0) \in \mathcal{Y}, \alpha^* = \alpha_2, \beta^* = 0$ satisfies (16), we conclude that $y^*(\alpha_2, 0)$ is the solution to (14). Otherwise, we know $\alpha^* > 0, \beta^* = 0$ is not possible.

Case 3: $\beta^* > 0, \alpha^* = 0$**.** In this case, the KKT condition tells us that $\sum_i c_i y_i^* = B_2$. Denote $\hat{y}_1^0 = y_1^0 \exp(-g_1)$, $\hat{y}_i^0 = y_i^0, i = 2, ..., n$. In this case, depending on the value of y_1^* we set

$$\begin{cases} \beta_1 = \mathrm{Root}_{\beta > 0} \big\{ \sum_{i=2}^n c_i \hat{y}_i^0 \exp(-c_i \beta) = B_2 - c_1 a_1 \big\}, \\ \beta_2 = \mathrm{Root}_{\beta > 0} \big\{ \sum_{i=1}^n c_i \hat{y}_i^0 \exp(-c_i \beta) = B_2 \big\}. \end{cases}$$

Note that in both cases, the problem is finding the positive root of a 1-dimensional monotonically decreasing function, which can be solved efficiently. These equations should either have one unique positive solution or no positive solution at all. If there is no positive root, then $\mathrm{Root}_{\beta>0}$ will return FALSE. One can easily determine whether there is a positive solution. For example, due to the monotonicity, the first equation will have a positive solution if and only if $\sum_{i=2}^n \hat{c}_i y_i^0 > B_2 - c_1 a_1$.

Similar to case 2, we check the feasibility of $\{y^*(0, \beta_1), \alpha^* = 0, \beta^* = \beta_1\}$ and $\{y^*(0, \beta_2), \alpha^* = 0, \beta^* = \beta_2\}$ w.r.t. (16). If any one of them is feasible to the KKT condition, then it will be the solution to (14). Otherwise, we know $\alpha^* = 0, \beta^* > 0$ is not possible.

Case 4: $\beta^* > 0, \alpha^* > 0$**.** In this case, the KKT condition implies that $\sum_i c_i y_i^* = B_1$, $\sum_i c_i y_i^* = B_2$. Let us inherit the $\hat{y}$ notation from Case 3. Then we need to solve the following group of equations

$$\begin{cases} \sum_{i=2}^n \hat{y}_i^0 \exp(-\alpha_3 - c_i \beta_3) = B_1 - a_1, \\ \sum_{i=2}^n c_i \hat{y}_i^0 \exp(-\alpha_3 - c_i \beta_3) = B_2 - c_1 a_1 \end{cases} \quad \text{or} \quad \begin{cases} \sum_{i=1}^n \hat{y}_i^0 \exp(-\alpha_4 - c_i \beta_4) = B_1, \\ \sum_{i=1}^n c_i \hat{y}_i^0 \exp(-\alpha_4 - c_i \beta_4) = B_2 \end{cases}$$

We should notice that in both cases, as soon as we determine the value of β, then α will have a closed form formula given β. To demonstrate how to determine β, let us take the second group of equations for example. Taking the quotient between the two equations cancels α_4, we get the following equation of β_4

$$f(\beta_4) := \frac{\sum_{i=1}^n c_i \hat{y}_i^0 \exp(-c_i \beta_4)}{\sum_{i=1}^n \hat{y}_i^0 \exp(-c_i \beta_4)} = \frac{B_2}{B_1}. \tag{18}$$

By Cauchy's inequality, we know $f'(\beta) < 0$ holds for $\forall \beta \in \mathbb{R}$ if $c_i \neq c_j$ for some i, j. In details

$$f'(\beta) = \frac{\left(\sum_{i=1}^n c_i \hat{y}_i^0 \exp(-c_i\beta)\right)^2 - \left(\sum_{i=1}^n \hat{y}_i^0 \exp(-c_i\beta)\right)\left(\sum_{i=1}^n c_i^2 \hat{y}_i^0 \exp(-c_i\beta)\right)}{\left(\sum_{i=1}^n \hat{y}_i^0 \exp(-c_i\beta)\right)^2} < 0.$$

Hence, f is again a monotonically decreasing function, and finding its positive root can be implemented efficiently. After finding β_4, one immediately know $\alpha_4 = \ln\left(\frac{\sum_{i=1}^n \hat{y}_i^0 \exp(-c_i\beta_4)}{B_1}\right)$.

Finally, we need to check the feasibility of $\{y^*(\alpha_3, \beta_3), \alpha^* = \alpha_3, \beta^* = \beta_3\}$ and $\{y^*(\alpha_4, \beta_4), \alpha^* = \alpha_4, \beta^* = \beta_4\}$ w.r.t. (16). If any one of them is feasible to the KKT condition, then it will be the solution to (14). Otherwise, we know $\alpha^* > 0, \beta^* > 0$ is not possible. Due to the existence of a KKT pair, at least one of the 4 cases will return us a solution.

B Proof of Proposition 4.3

For the analysis of Proposition 4.3 and later results, let us first introduce a vector version of the Bernstein's inequality, which is a direct specification of the Freedman's inequality of matrix martingale [23]. To prove the current proposition, we only need the scalar case of the following lemma.

Lemma B.1 (Vector Bernstein Inequality). *Assume that $\{x_i\}_{i=1}^n$ is a sequence of random vectors in $\mathbb{R}^d$, and it forms a martingale difference sequence with respect to $(\mathcal{F}_t)$ (i.e. $\mathbb{E}\left[x_t \middle| \mathcal{F}_{t-1}\right] = 0$ and x_t is $\mathcal{F}_t$-measurable). If $\mathbb{E}\left[\|x_t\|^2 \middle| \mathcal{F}_{t-1}\right] \leq \sigma^2$ and $\|x_t\| \leq M$ a.s., then with probability at least $1 - \delta$,*

$$\left\|\sum_{i=1}^n x^i\right\| \leq 2\sigma\sqrt{n \log\left(\frac{d+1}{\delta}\right)} + 2M \log\left(\frac{d+1}{\delta}\right).$$

When the ℓ_2 norm is replaced by the ℓ_∞ norm, i.e., $\{x_i\}_{i=1}^n$ satisfies $\mathbb{E}\left[\|x_t\|_\infty^2 \middle| \mathcal{F}_{t-1}\right] \leq \sigma^2$ and $\|x_t\|_\infty \leq M$,

$$\left\|\sum_{i=1}^n x^i\right\|_\infty \leq 2\sigma\sqrt{n \log\left(\frac{2d}{\delta}\right)} + 2M \log\left(\frac{2d}{\delta}\right)$$

holds with probability at least $1 - \delta$.

To prove Proposition 4.3, we consider $\hat{\mu}_0(s, a) = \frac{N(s,a)}{N_e}$, then it is clear that $\hat{\mu}(s, a) = \max(\hat{\mu}_0(s, a), \varsigma)$. Now, according to the Bernstein's inequality, we construct the "failure event"

$$\Omega := \bigcup_{s,a} \left\{ |\mu(s, a) - \hat{\mu}_0(s, a)| > \sqrt{\mu(s, a)\frac{\ell}{N_e}} + \frac{\ell}{N_e} \right\},$$

where $\ell \geq 4 \log\left(\frac{6|\mathcal{S}||\mathcal{A}|}{\delta}\right)$ is a mild logarithmic term. We next prove the three properties listed in Proposition 4.3 one by one.

Proof of Proposition 4.3 (1). In fact, we only need to show that $\mathbb{P}(\Omega) \leq \frac{\delta}{3}$, and the event Ω^c implies that $\mu(s, a) \leq 2\hat{\mu}(s, a), \forall s, a$, as long as our choice of batch size satisfies $N_e \geq \frac{128\mathcal{N}\psi\ell}{\varphi^2(1-\gamma)^4\epsilon_e^2} \geq \frac{32\ell\mathcal{N}\psi}{\varphi(1-\gamma)^2\epsilon_e} = \frac{32\ell}{\varsigma}$.

By Bernstein's inequality, it holds that

$$\mathbb{P}\left(|\mu(s, a) - \hat{\mu}_0(s, a)| > \sqrt{\mu(s, a)\frac{\ell}{N_e}} + \frac{\ell}{N_e}\right) \leq \frac{\delta}{3|\mathcal{S}||\mathcal{A}|}, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Then $\mathbb{P}(\Omega) \leq \frac{\delta}{3}$ follows directly from the union bound. Conditioning on Ω^c, we have

$$|\mu(s, a) - \hat{\mu}_0(s, a)| \leq \sqrt{\mu(s, a)\frac{\ell}{N_e}} + \frac{\ell}{N_e} \leq \sqrt{\mu(s, a)\frac{\varsigma}{32}} + \frac{\varsigma}{32} \leq \frac{\mu(s, a)}{4} + \frac{\varsigma}{16}. \tag{19}$$

Hence, it holds that

$$\mu(s,a) \leq \frac{4}{3}\hat{\mu}_0(s,a) + \frac{\varsigma}{12} \leq \frac{3}{2}\max(\hat{\mu}_0(s,a), \varsigma) \leq 2\hat{\mu}(s,a).$$

From now on, the argument is all conditioning on Ω^c.

Proof of Proposition 4.3 (2). Given a $\pi \in \Pi(\psi)$, we have to prove that $W^{-1}\nu^\pi \in \mathcal{X}$.

Let $\nu = \nu^\pi$, $x = W^{-1}\nu$. Then due to $\pi \in \Pi(\psi)$, we have

$$\max_{s,a} \frac{x(s,a)}{\hat{\mu}(s,a)} = \max_{s,a} \frac{\nu(s,a)}{\mu(s,a)} \leq \frac{\psi}{1-\gamma},$$
$$\sum_{s,a} \frac{x(s,a)}{\hat{\mu}(s,a)} = \sum_{s,a} \frac{\nu(s,a)}{\mu(s,a)} \leq \frac{\mathcal{N}\psi}{1-\gamma}.$$

Now it remains to show $\sum_{s,a} x(s,a) \leq \frac{4}{1-\gamma}$. Note that (19) also implies

$$\hat{\mu}_0(s,a) \leq \frac{5}{4}\mu(s,a) + \frac{\varsigma}{16}.$$

Hence if $\mu(s,a) \leq \frac{1}{2}\hat{\mu}(s,a)$, then it must hold that $\hat{\mu}_0(s,a) < \hat{\mu}(s,a) \Rightarrow \hat{\mu}_0(s,a) < \varsigma, \hat{\mu}(s,a) = \varsigma$. We define $\mathfrak{S} := \{(s,a) \in \mathcal{S} \times \mathcal{A} : \hat{\mu}(s,a) = \varsigma\}$, then for $(s,a) \notin \mathfrak{S}$, it holds that $\mu(s,a) \geq \frac{1}{2}\hat{\mu}(s,a)$. Thus, we have

$$\begin{aligned}\sum_{s,a} x(s,a) &= \sum_{(s,a)\in\mathfrak{S}} \hat{\mu}(s,a)\frac{\nu(s,a)}{\mu(s,a)} + \sum_{(s,a)\notin\mathfrak{S}} \frac{\hat{\mu}(s,a)}{\mu(s,a)}\nu(s,a) \\ &\leq \varsigma\frac{\mathcal{N}\psi}{1-\gamma} + \sum_{(s,a)\notin\mathfrak{S}} 2\nu(s,a) \\ &\leq \frac{3}{1-\gamma}.\end{aligned}$$

The last inequality holds as long as $\varsigma \leq \frac{1}{\mathcal{N}\psi}$.

Proof of Proposition 4.3 (3). We decompose the quantity $\|Wx - x\|_1$ as

$$\begin{aligned}\|Wx - x\|_1 &= \sum_{s,a} |\mu(s,a) - \hat{\mu}(s,a)| \frac{x(s,a)}{\hat{\mu}(s,a)} \\ &= \sum_{(s,a)\in\mathfrak{S}} |\mu(s,a) - \hat{\mu}(s,a)| \frac{x(s,a)}{\hat{\mu}(s,a)} + \sum_{(s,a)\notin\mathfrak{S}} |\mu(s,a) - \hat{\mu}(s,a)| \frac{x(s,a)}{\hat{\mu}(s,a)}.\end{aligned}$$

From our definition of $\mathfrak{S}$, we see if $(s,a) \in \mathfrak{S}$, then $\hat{\mu}(s,a) = \varsigma \geq \hat{\mu}_0(s,a)$, and from (19) we have $\mu(s,a) \leq 2\varsigma \Rightarrow |\mu(s,a) - \hat{\mu}(s,a)| \leq \varsigma$. Thus, the first part can be bounded as

$$\sum_{(s,a)\in\mathfrak{S}} |\mu(s,a) - \hat{\mu}(s,a)| \frac{x(s,a)}{\hat{\mu}(s,a)} \leq \sum_{s,a} \varsigma\frac{x(s,a)}{\hat{\mu}(s,a)} \leq \varsigma\frac{\mathcal{N}\psi}{1-\gamma}.$$

As for the second part, we have

$$\begin{aligned}&\sum_{(s,a)\notin\mathfrak{S}} |\mu(s,a) - \hat{\mu}(s,a)| \frac{x(s,a)}{\hat{\mu}(s,a)} \\ &= \sum_{(s,a)\notin\mathfrak{S}} |\mu(s,a) - \hat{\mu}_0(s,a)| \frac{x(s,a)}{\hat{\mu}(s,a)} \\ &\leq \sum_{(s,a)\notin\mathfrak{S}} \left(\sqrt{\mu(s,a)\frac{\ell}{N_e}} + \frac{\ell}{N_e}\right) \frac{x(s,a)}{\hat{\mu}(s,a)} \\ &= \sqrt{\frac{\ell}{N_e}} \sum_{(s,a)\notin\mathfrak{S}} \sqrt{\frac{\mu(s,a)}{\hat{\mu}(s,a)}} \sqrt{x(s,a)\cdot\frac{x(s,a)}{\hat{\mu}(s,a)}} + \frac{\ell}{N_e} \sum_{(s,a)\notin\mathfrak{S}} \frac{x(s,a)}{\hat{\mu}(s,a)}\end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{\leq} \sqrt{\frac{2\ell}{N_e}}\sum_{s,a}\sqrt{x(s,a)\cdot\frac{x(s,a)}{\hat{\mu}(s,a)}}+\frac{\ell}{N_e}\sum_{s,a}\frac{x(s,a)}{\hat{\mu}(s,a)}\\
&\overset{(b)}{\leq} \sqrt{\frac{2\ell}{N_e}}\sqrt{\sum_{s,a}x(s,a)\sum_{s,a}\frac{x(s,a)}{\hat{\mu}(s,a)}}+\frac{\ell}{N_e}\sum_{s,a}\frac{x(s,a)}{\hat{\mu}(s,a)}\\
&\overset{(c)}{\leq} \frac{2}{1-\gamma}\sqrt{\frac{2\mathcal{N}\psi\ell}{N_e}}+\frac{\mathcal{N}\psi\ell}{(1-\gamma)N_e},
\end{aligned}$$

where the inequality (a) comes from the fact $\frac{\mu(s,a)}{\hat{\mu}(s,a)} \leq 2$; (b) is due to Cauchy's inequality, and (c) is due to $\sum_{s,a}\frac{x(s,a)}{\hat{\mu}(s,a)} \leq \frac{\mathcal{N}\psi}{1-\gamma}$ and $\sum_{s,a}x(s,a) \leq \frac{4}{1-\gamma}$. Therefore, because we set $\varsigma = \frac{\varphi(1-\gamma)^2\epsilon_e}{2\mathcal{N}\psi}$, and $N_e \geq \frac{128\mathcal{N}\psi\ell}{\varphi^2(1-\gamma)^4\epsilon_e^2}$, we have $\|Wx - x\|_1 \leq \varphi(1-\gamma)\epsilon_e, \forall x \in \mathcal{X}$.

C The magnitude and variance of the gradient estimators

Proposition C.1. *For any sample $\zeta \sim \rho_0 \times \mathcal{D}$, and any feasible solution $Z = [V;\lambda;x]$, the stochastic gradient estimators constructed in* (9) *are unbiased, and they satisfy the following bounds:*[2]

$$\begin{cases}\mathbb{E}\left[\widehat{g}_V(Z;\zeta)\right] = \nabla_V\mathcal{L}_w(Z)\\ \|\widehat{g}_V(Z;\zeta)\| \leq \mathcal{O}\left(\frac{\psi}{1-\gamma}\right)\\ \mathbb{E}\left[\|\widehat{g}_V(Z;\zeta)\|^2\right] \leq \mathcal{O}\left(\frac{\psi}{(1-\gamma)^2}\right)\end{cases}
\begin{cases}\mathbb{E}\left[\widehat{g}_\lambda(Z;\zeta)\right] = \nabla_\lambda\mathcal{L}_w(Z)\\ \|\widehat{g}_\lambda(Z;\zeta)\|_\infty \leq \mathcal{O}\left(\frac{\psi}{1-\gamma}\right)\\ \mathbb{E}\left[\|\widehat{g}_\lambda(Z;\zeta)\|_\infty^2\right] \leq \mathcal{O}\left(\frac{\psi}{(1-\gamma)^2}\right)\end{cases}
\begin{cases}\mathbb{E}\left[\widehat{g}_x(Z;\zeta)\right] = \nabla_x\mathcal{L}_w(Z)\\ \|\widehat{g}_x(Z;\zeta)\|_{x'}^2 \leq \mathcal{O}\left(\frac{\psi^2\mathcal{N}}{\varphi^3(1-\gamma)^5\epsilon_e}\right)\\ \mathbb{E}\left[\|\widehat{g}_x(Z;\zeta)\|_{x'}^2\right] \leq \mathcal{O}\left(\frac{\mathcal{N}\psi}{\varphi^2(1-\gamma)^3}\right)\end{cases}$$

where $x' \in \mathcal{X}$ is an arbitrary vector.

For any sample $\zeta = (s_0, s, a, s', r, \mathbf{u}) \sim \rho_0 \times \mathcal{D}$, it is not hard to see that the estimators constructed in (9) are unbiased. Next, we provide the bound on the norm and variance of these estimators.

For the estimator $\widehat{g}_V(Z;\zeta) := \mathbb{I}_{s_0} + \frac{x(s,a)}{\hat{\mu}(s,a)}\left(\gamma\mathbb{I}_{s'} - \mathbb{I}_s\right)$, we have

$$\|\widehat{g}_V(Z;\zeta)\| \leq 1 + \frac{x(s,a)}{\hat{\mu}(s,a)}(1+\gamma) \overset{(a)}{\leq} 1 + \frac{2\psi}{1-\gamma},$$

$$\begin{aligned}
\mathbb{E}\left[\|g_V(Z;\zeta)\|^2\right] \quad &\leq \quad \sum_{s,a}\mu(s,a)\cdot 2\left(1 + 4\cdot\frac{x(s,a)^2}{\hat{\mu}(s,a)^2}\right),\\
&\leq \quad 2 + 8\cdot\sum_{s,a}\frac{\mu(s,a)}{\hat{\mu}(s,a)}\frac{x(s,a)}{\hat{\mu}(s,a)}x(s,a) \overset{(b)}{\leq} 2 + \frac{64\psi}{(1-\gamma)^2}.
\end{aligned}$$

Here (a) is due to $x \in \mathcal{X}$, which indicates that $\frac{x(s,a)}{\hat{\mu}(s,a)} \leq \frac{\psi}{1-\gamma}$ for all (s,a). The inequality (b) is due to $\frac{\mu(s,a)}{\hat{\mu}(s,a)} \leq 2$ established in Proposition 4.3, and $\sum_{s,a}x(s,a) \leq \frac{4}{1-\gamma}$.

Similarly, for the estimator $\widehat{g}_\lambda(Z;\zeta) := \frac{x(s,a)}{\hat{\mu}(s,a)}\mathbf{u}^\kappa$, we have

$$\|\widehat{g}_\lambda(Z;\zeta)\|_\infty \leq \left\|\frac{x(s,a)}{\hat{\mu}(s,a)}\mathbf{u}^\kappa\right\|_\infty \overset{(a)}{\leq} \frac{x(s,a)}{\hat{\mu}(s,a)}(1+(1-\gamma)\kappa) \overset{(b)}{\leq} \frac{2\psi}{1-\gamma},$$

$$\begin{aligned}
\mathbb{E}\left[\|g_\lambda(Z;\zeta)\|_\infty^2\right] \quad &\leq \quad \sum_{s,a}\mu(s,a)\cdot 4\frac{x(s,a)^2}{\hat{\mu}(s,a)^2},\\
&= \quad 4\sum_{s,a}\frac{\mu(s,a)}{\hat{\mu}(s,a)}\frac{x(s,a)}{\hat{\mu}(s,a)}x(s,a) \overset{(c)}{\leq} \frac{32\psi}{(1-\gamma)^2}.
\end{aligned}$$

Here (a) follows from $\|\mathbf{u}^\kappa\|_\infty \leq \|\mathbf{u}\|_\infty + (1-\gamma)\kappa$, and (b) is due to $(1-\gamma)\kappa = 5\varphi\epsilon(1-\gamma) < 1$, and (c) is similar to the argument of the bound on $\mathbb{E}\left[\|g_V(Z;\zeta)\|^2\right]$.

[2] For vectors $u, v \in \mathbb{R}^n$, we write $\|u\|_v^2 := \sum_{i=1}^n v_i u_i^2$ for simplicity.

Finally, for the estimator $\widehat{g}_x(Z;\zeta) := \frac{r+\gamma V(s)-V(s')+\langle \mathbf{u}^\kappa,\lambda\rangle}{\hat{\mu}(s,a)}\mathbb{I}_{s,a}$, we have

$$\begin{aligned}
\|\widehat{g}_x(Z;\zeta)\|_{x'}^2 &= \frac{x'(s,a)}{\hat{\mu}(s,a)^2}\cdot\left|r+\gamma V(s)-V(s')+\langle \mathbf{u}^\kappa,\lambda\rangle\right|^2 \\
&\le \frac{x'(s,a)}{\hat{\mu}(s,a)^2}\left(1+\frac{16}{1-\gamma}(1+\frac{2}{\varphi})+\frac{8(1+\kappa)}{\varphi}\right)^2 \\
&\le \frac{\psi}{(1-\gamma)\varsigma}\cdot\frac{64^2}{\varphi^2(1-\gamma)^2} \\
&= \mathcal{O}\left(\frac{\psi^2\mathcal{N}}{\varphi^3(1-\gamma)^5\epsilon_e}\right),
\end{aligned}$$

and as long as ζ is independent of $x'\in\mathcal{X}$,

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{g}_x(Z;\zeta)\|_{x'}^2\right] &\le \sum_{s,a}\frac{\mu(s,a)x'(s,a)}{\hat{\mu}(s,a)^2}\cdot\left(1+\frac{16}{1-\gamma}(1+\frac{2}{\varphi})+\frac{8(1+\kappa)}{\varphi}\right)^2 \\
&\le \sum_{s,a}\frac{\mu(s,a)}{\hat{\mu}(s,a)}\frac{x'(s,a)}{\hat{\mu}(s,a)}\cdot\frac{64^2}{\varphi^2(1-\gamma)^2} \\
&\le \mathcal{O}\left(\frac{\mathcal{N}\psi}{\varphi^2(1-\gamma)^3}\right).
\end{aligned}$$

This completes the proof of Proposition C.1.

A few notational definitions. We should notice that the above bounds on the gradient estimators are notationally very complicated. Therefore, Let us conveniently write the above bounds as

$$\begin{cases}\|g_V(Z^t;\zeta_t)\|\le M_V,\\ \|g_\lambda(Z^t;\zeta_t)\|_\infty\le M_\lambda,\\ \|g_x(Z^t;\zeta_t)\|_{x'}\le M_x\sqrt{D_{x,1}},\end{cases}\qquad\text{and}\qquad\begin{cases}\mathbb{E}\left[\|g_V(Z;\zeta)\|^2\right]\le\sigma_V^2,\\ \mathbb{E}\left[\|g_\lambda(Z;\zeta)\|_\infty^2\right]\le\sigma_\lambda^2,\\ \mathbb{E}\left[\|g_x(Z^t;\zeta_t)\|_{x'}^2\right]\le\sigma_x^2 D_{x,1},\end{cases}$$

where the constants $\sigma_V,\sigma_\lambda,\sigma_x$ and M_V,M_λ,M_x are

$$\sigma_V^2=\Theta\left(\frac{\psi}{(1-\gamma)^2}\right),\qquad \sigma_\lambda^2=\Theta\left(\frac{\psi}{(1-\gamma)^2}\right),\qquad \sigma_x^2=\Theta\left(\frac{\mathcal{N}\psi}{\varphi^2(1-\gamma)^2}\right),\tag{20}$$

$$M_V=\Theta\left(\frac{\psi}{1-\gamma}\right),\qquad M_\lambda=\Theta\left(\frac{\psi}{1-\gamma}\right),\qquad M_x=\Theta\left(\frac{\psi}{\varphi(1-\gamma)^2}\sqrt{\frac{\mathcal{N}}{\varphi\epsilon_e}}\right),\tag{21}$$

and $D_{x,1}$ is a suitable upper bound on the diameter of $\mathcal{X}$, namely we choose $D_{x,1}=\Theta\left(\frac{1}{1-\gamma}\right)$ such that $D_{x,1}\ge\sup_{x,x'\in\mathcal{X}}\|x'-x\|_1$. Similarly, we define $D_{\lambda,1}:=\sup_{\lambda,\lambda'\in\Lambda}\|\lambda'-\lambda\|_1=\Theta\left(\frac{1}{\varphi}\right)$.

Furthermore, we also introduce the diameters of the feasible domains w.r.t. the initial solution V^1,λ^1,x^1. Recall that the initial point of Algorithm 1 is chosen as

$$V^1=\mathbf{0}\in\mathcal{V},\qquad \lambda^1=\frac{\mathbf{1}}{\varphi I}\in\Lambda,\qquad x^1=\frac{c_x\hat{\mu}}{1-\gamma}\in\mathcal{X},$$

where $c_x=\frac{\mathcal{N}}{|\mathcal{S}||\mathcal{A}|}$ ensures that $x^1\in\mathcal{X}$. Then, we can take D_V,D_λ,D_x as

$$\begin{aligned}
D_V^2 &:= \sup_{V'\in\mathcal{V}}\left\|V'-V^1\right\|^2=\Theta\left(\frac{|\mathcal{S}|}{\varphi^2(1-\gamma)^2}\right),\\
D_\lambda &:= \sup_{\lambda'\in\Lambda}\mathrm{KL}(\lambda'\parallel\lambda^1)=\Theta\left(\frac{\log I}{\varphi}\right),\\
D_x &\ge \sup_{x'\in\mathcal{X}}\mathrm{KL}(x'\parallel x^1),\qquad D_x=\Theta\left(\frac{\log\psi}{1-\gamma}\right).
\end{aligned}$$

Remark C.2. *It is worth noting that, Proposition C.1 directly implies* $\mathbb{E}\left[\left\|\widehat{g}_V(Z^t;\zeta_t)\right\|^2\middle| Z^t\right] \leq \sigma_V^2$, $\mathbb{E}\left[\left\|\widehat{g}_\lambda(Z^t;\zeta_t)\right\|_\infty^2\middle| Z^t\right] \leq \sigma_\lambda^2$ *and* $\mathbb{E}\left[\left\|\widehat{g}_x(Z^t;\zeta_t)\right\|_{x^t}^2\middle| Z^t\right] \leq D_{x,1}\sigma_x^2$ *for each step t.*

Remark C.3. *The reason why we bound the term* $\left\|g_x(Z^t;\zeta_t)\right\|_{x^t}$ *instead of* $\left\|g_x(Z^t;\zeta_t)\right\|_\infty$ *is that,*

$$\left\|g_x(Z^t;\zeta_t)\right\|_\infty \lesssim \frac{1}{\varphi(1-\gamma)}\frac{1}{\hat{\mu}(s_t,a_t)} \leq \frac{1}{\varphi(1-\gamma)\varsigma}.$$

Thus, we have to take $M_{x,\infty} = \Theta\left(\frac{1}{\varphi(1-\gamma)\varsigma}\right)$ *to ensure a uniformly bound as*

$$\left\|g_x(Z^t;\zeta_t)\right\|_\infty \leq M_{x,\infty}. \tag{22}$$

D Proof of Theorem 4.4

To bound $\mathrm{Gap}(\overline{x})$, let us denote

$$(V',\lambda') = \underset{V\in\mathcal{V},\lambda\in\Lambda}{\arg\min}\, \mathcal{L}_w(V,\lambda,\overline{x}), \qquad x' = \underset{x\in\mathcal{X}}{\arg\max} \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,x), \tag{23}$$

and denote $Z' = [V';\lambda';x']$. It is worth mentioning that V',λ' are random variables that depend on $\overline{x}$ while x' is deterministic. For the ease of notation, we define

$$\mathcal{G}(Z) := \begin{bmatrix} +\nabla_V\mathcal{L}_w(Z) \\ +\nabla_\lambda\mathcal{L}_w(Z) \\ -\nabla_x\mathcal{L}_w(Z) \end{bmatrix} \qquad \text{and} \qquad \widehat{g}(Z;\zeta) := \begin{bmatrix} +\widehat{g}_V(Z;\zeta) \\ +\widehat{g}_\lambda(Z;\zeta) \\ -\widehat{g}_x(Z;\zeta) \end{bmatrix}.$$

Then, by the definition of V',λ',x' and the bi-linearity of $\mathcal{L}_w(\cdot)$, we have

$$\begin{aligned}
\mathrm{Gap}(\overline{x}) &= \max_{x\in\mathcal{X}} \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,x) - \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,\overline{x}) \\
&= \mathcal{L}_w(\overline{V},\overline{\lambda},x') - \mathcal{L}_w(V',\lambda',\overline{x}) \\
&= \frac{1}{T}\sum_{t=1}^T \Big(\mathcal{L}_w(V^t,\lambda^t,x') - \mathcal{L}_w(V',\lambda',x^t)\Big) \qquad (24) \\
&= \frac{1}{T}\sum_{t=1}^T \left\langle \mathcal{G}(Z^t), Z^t - Z'\right\rangle \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^T \left\langle \widehat{g}(Z^t;\zeta_t), Z^t - Z'\right\rangle}_{S_1} + \underbrace{\frac{1}{T}\sum_{t=1}^T \left\langle \mathcal{G}(Z^t) - \widehat{g}(Z^t;\zeta_t), Z^t - Z'\right\rangle}_{S_2}.
\end{aligned}$$

Then with the estimations in Appendix C, the S_1 and S_2 terms can be bounded by

$$\begin{aligned}
S_1 \quad &\lesssim \quad \frac{\alpha_V D_V^2 + \alpha_\lambda D_\lambda + \alpha_x D_x}{\eta T} + \eta\left(\frac{\sigma_V^2}{\alpha_V} + \frac{\sigma_\lambda^2 D_{\lambda,1}}{\alpha_\lambda} + \frac{\sigma_x^2 D_{x,1}}{\alpha_x}\right) \qquad (25) \\
&\qquad + \frac{\eta\iota}{T}\left(\frac{M_V^2}{\alpha_V} + \frac{M_\lambda^2 D_{\lambda,1}}{\alpha_\lambda} + \frac{M_x^2 D_{x,1}}{\alpha_x}\right)
\end{aligned}$$

and

$$S_2 \lesssim \left(D_V\sigma_V + D_{\lambda,1}\sigma_\lambda + D_{x,1}\sigma_x\right)\sqrt{\frac{\iota}{T}} + \left(D_V M_V + D_{\lambda,1}M_\lambda + D_{x,1}M_x\right)\frac{\iota}{T} \tag{26}$$

with probability at least $1 - {}^{\delta}/_{10}$ respectively, as long as the stepsize satisfies

$$\eta \leq \frac{1}{2}\min\left(\frac{\alpha_\lambda}{M_\lambda}, \frac{\alpha_x}{M_{x,\infty}}\right). \tag{27}$$

Due to the sophistication of the proof, we move the analysis of (25) and (26) to Appendix D.2 and D.3 respectively.

Finally, combining the inequalities (24), (25) and (26), and requiring that (27) holds true for $\eta = 1/\sqrt{T}$, we have with probability at least $1 - {}^{\delta}/_{3}$

$$\begin{aligned}
\mathrm{Gap}(\overline{x}) \lesssim & \frac{\alpha_V D_V^2 + \alpha_\lambda D_\lambda + \alpha_x D_x}{\eta T} + \eta \left(\frac{\sigma_V^2}{\alpha_V} + \frac{\sigma_\lambda^2 D_{\lambda,1}}{\alpha_\lambda} + \frac{\sigma_x^2 D_{x,1}}{\alpha_x} \right) \\
& + \sqrt{\frac{\iota}{T}} \left(D_V \sigma_V + D_{\lambda,1} \sigma_\lambda + D_{x,1} \sigma_x \right) \\
& + \frac{\iota}{T} \left(D_V M_V + D_{\lambda,1} M_\lambda + D_{x,1} M_x \right) \\
& + \frac{\eta \iota}{T} \left(\frac{M_V^2}{\alpha_V} + \frac{M_\lambda^2 D_{\lambda,1}}{\alpha_\lambda} + \frac{M_x^2 D_{x,1}}{\alpha_x} \right).
\end{aligned}$$

Note that the normalizing constants are chosen as $\alpha_V = \varphi \sqrt{\frac{\psi}{|\mathcal{S}|}} = \Theta\left(\frac{\sigma_V}{D_V}\right)$, $\alpha_\lambda = \frac{1}{1-\gamma} \sqrt{\frac{\psi}{\log I}} = \Theta\left(\sigma_\lambda \sqrt{\frac{D_{\lambda,1}}{D_\lambda}}\right)$, $\alpha_x = \frac{1}{\varphi(1-\gamma)} \sqrt{\frac{\mathcal{N}\psi}{\log \psi}} = \Theta\left(\sigma_x \sqrt{\frac{D_{x,1}}{D_x}}\right)$. Then (27) holds true for the stepsize $\eta = \frac{1}{\sqrt{T}}$ with $T \gtrsim \frac{\mathcal{N}\psi\iota}{\varphi^2 (1-\gamma)^4 \epsilon_e^2}$, and we can plug in the values of the constants α, D, M, then with probability at least $1 - {}^{\delta}/_{3}$ it holds that

$$\mathrm{Gap}(\overline{x}) \lesssim \sqrt{\frac{\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4 T}} \left(1 + \frac{\iota}{T} \cdot \frac{\psi}{\varphi(1-\gamma)^2 \epsilon_e} \right) \lesssim \sqrt{\frac{\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4 T}}.$$

Choosing c_o to ensure $\mathrm{Gap}(\overline{x}) \le \frac{\epsilon}{2}$ completes the proof of Theorem 4.4.

D.1 A few supporting lemmas

For the proof in the following parts of Appendix D, we introduce a few supporting lemmas.

Lemma D.1. *Let $\{Y^k\}_{k=1}^T$ be generated by $Y^{k+1} = \mathrm{argmin}_{Y \in \mathcal{Y}} \left(\eta \left\langle Y - Y^k, g^k \right\rangle + \mathrm{KL}(Y \parallel Y^k) \right)$, where $\eta \le \frac{1}{2 \max_k \|g_k\|_\infty}$ and $\mathcal{Y}$ is some convex set. Then for all $Y' \in \mathcal{Y}$, it holds that*

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \left\langle Y^t - Y', g^t \right\rangle \le & \frac{\mathrm{KL}(Y' \parallel Y^1)}{\eta T} + \frac{4\eta}{T} \sum_{t=1}^{T} \left\| g^k \right\|_{Y^k}^2 \\
\le & \frac{\mathrm{KL}(Y' \parallel Y^1)}{\eta T} + \frac{4\eta D_{Y,1}}{T} \sum_{t=1}^{T} \left\| g^k \right\|_\infty^2.
\end{aligned}$$

where $D_{Y,1}$ can be any upper bound of $\max_{Y \in \mathcal{Y}} \|Y\|_1$.

The proof of Lemma D.1 is presented in Appendix D.4.

Lemma D.2. *Let $\{Y^k\}_{k=1}^T$ be generated by $Y^{k+1} = \mathrm{Proj}_{\mathcal{Y}} \left(Y^k - \eta g^k \right)$, where $\mathcal{Y}$ is some convex set. Then for all $Y' \in \mathcal{Y}$, it holds that*

$$\frac{1}{T} \sum_{t=1}^{T} \left\langle Y^t - Y', g^t \right\rangle \le \frac{\left\| Y' - Y^1 \right\|^2}{2\eta T} + \frac{\eta}{T} \sum_{t=1}^{T} \left\| g^k \right\|^2.$$

The proof of Lemma D.2 is similar but a lot simpler than that of Lemma D.1, and is hence omitted.

Proposition D.3 (Corollary of Bernstein's inequality)**.** *For a sequence of random variables $X_1, \cdots, X_N$ adapted to $(\mathcal{F}_n)$, and $\mathbb{E}\left[|X_i| \,\middle|\, \mathcal{F}_{i-1} \right] \le c$, $|X_i| \le M$, we have with probability at least $1 - \delta$,*

$$\left| \frac{1}{N} \sum_{i=1}^{N} X_i \right| \le 2c + 3M \frac{\log(2/\delta)}{N}.$$

Proof. Notice that $\mathbb{E}\left[X_i^2 \,\middle|\, \mathcal{F}_{i-1} \right] \le cM$, and by Bernstein's inequality

$$\left| \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \mathbb{E}\left[X_i \,\middle|\, \mathcal{F}_{i-1} \right] \right) \right| \le \sqrt{\frac{2cM \log(2/\delta)}{N}} + 2M \frac{\log(2/\delta)}{N},$$

$$\Rightarrow \left|\sum_{i=1}^{N} X_i\right| \leq cN + \sqrt{2cMN\log(2/\delta)} + 2M\log(2/\delta),$$

holds with probability at least $1-\delta$. By the AM-GM inequality, $\sqrt{2cMN\log(2/\delta)} \leq \frac{1}{2}cN + M\log(2/\delta)$, which completes the proof. □

D.2 Bounding the term S_1

First, by definition of $\widehat{g}(\cdot)$, we have

$$S_1 = \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left\langle \widehat{g}_V(Z^t;\zeta_t), V^t - V'\right\rangle}_{S_{1,V}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left\langle \widehat{g}_\lambda(Z^t;\zeta_t), \lambda^t - \lambda'\right\rangle}_{S_{1,\lambda}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left\langle -\widehat{g}_x(Z^t;\zeta_t), x^t - x'\right\rangle}_{S_{1,x}}.$$

Applying Lemma D.2 with $Y^t = V^t$, $g^t = \widehat{g}_V(Z^t;\zeta_t)$ yields

$$S_{1,V} \leq \frac{\alpha_V \left\|V' - V^1\right\|^2}{2\eta T} + \frac{\eta}{\alpha_V T}\sum_{t=1}^{T}\left\|\widehat{g}_V(Z^t;\zeta_t)\right\|^2.$$

Applying Lemma D.1 with $Y^t = \lambda^t$, $g^t = \widehat{g}_\lambda(Z^t;\zeta_t)$, we have

$$S_{1,\lambda} \leq \frac{\alpha_\lambda \operatorname{KL}(\lambda' \parallel \lambda^1)}{\eta T} + \frac{4\eta D_{\lambda,1}}{\alpha_\lambda T}\sum_{t=1}^{T}\left\|\widehat{g}_\lambda(Z^t;\zeta_t)\right\|_\infty^2,$$

as long as $\|\widehat{g}_\lambda(Z^t;\zeta_t)\|_\infty \leq \frac{\alpha_\lambda}{2\eta}$ holds for all t, and $\frac{1}{\eta} \geq \frac{2M_\lambda}{\alpha_\lambda}$ suffices.

Finally, applying Lemma D.1 with $Y^t = x^t$, $g^t = -\widehat{g}_x(Z^t;\zeta_t)$, we obtain

$$S_{1,x} \leq \frac{\alpha_x \operatorname{KL}(x' \parallel x^1)}{\eta T} + \frac{4\eta}{\alpha_x T}\sum_{t=1}^{T}\left\|\widehat{g}_x(Z^t;\zeta_t)\right\|_{x^t}^2,$$

as long as $\|\widehat{g}_x(Z^t;\zeta_t)\|_\infty \leq \frac{\alpha_x}{2\eta}$ holds for all t, and $\frac{1}{\eta} \geq \frac{2M_{x,\infty}}{\alpha_x}$ suffices.

Combining all the estimations above, as long as the stepsize η satisfies (27), we have

$$\begin{aligned} S_1 \leq & \frac{\alpha_V \left\|V' - V^1\right\|^2 + \alpha_\lambda \operatorname{KL}(\lambda' \parallel \lambda^1) + \alpha_x \operatorname{KL}(x' \parallel x^1)}{\eta T} \\ & + \frac{4\eta}{T}\sum_{t=1}^{T}\left(\frac{\|\widehat{g}_V(Z^t;\zeta_t)\|^2}{\alpha_V} + \frac{D_{\lambda,1}\|\widehat{g}_\lambda(Z^t;\zeta_t)\|_\infty^2}{\alpha_\lambda} + \frac{\|\widehat{g}_x(Z^t;\zeta_t)\|_{x^t}^2}{\alpha_x}\right). \end{aligned} \tag{28}$$

For the second term of S_1 in (28), with the variance and magnitude bounds provided in Proposition C.1, applying Proposition D.3 to the sequences $\{\|\widehat{g}_V(Z^t;\zeta_t)\|^2\}_{t=1}^T$, $\{\|\widehat{g}_\lambda(Z^t;\zeta_t)\|_\infty^2\}_{t=1}^T$ and $\{\|\widehat{g}_x(Z^t;\zeta_t)\|_{x^t}^2\}_{t=1}^T$ proves the inequality (25) with probability at least $1 - \delta/10$.

D.3 Bounding the term S_2

For the term S_2, we introduce the martingale difference sequences

$$\begin{aligned} \Delta_V^t &:= \widehat{g}_V(Z^t;\zeta_t) - \nabla_V \mathcal{L}_w(V^t,\lambda^t,x^t), \\ \Delta_\lambda^t &:= \widehat{g}_\lambda(Z^t;\zeta_t) - \nabla_\lambda \mathcal{L}_w(V^t,\lambda^t,x^t), \\ \Delta_x^t &:= \widehat{g}_x(Z^t;\zeta_t) - \nabla_x \mathcal{L}_w(V^t,\lambda^t,x^t), \end{aligned}$$

Then S_2 can be decomposed as

$$\begin{aligned} S_2 = & \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\left\langle \Delta_V^t, V' - V^1\right\rangle + \left\langle \Delta_\lambda^t, \lambda' - \lambda^1\right\rangle\right)}_{S_{2,c}} \\ & + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\left\langle \Delta_V^t, V^1 - V^t\right\rangle + \left\langle \Delta_\lambda^t, \lambda^1 - \lambda^t\right\rangle + \left\langle -\Delta_x^t, x' - x^t\right\rangle\right)}_{S_{2,m}}. \end{aligned}$$

Note that the martingale part $S_{2,m}$ has expectation zero. However, for the first part, V' and λ' are random variables depending on $\bar{x}$. Thus the correlated part $S_{2,c}$ may not have zero mean.

Bounding the term $S_{2,c}$ For the correlated part $S_{2,c}$, the sequence Δ_V^t and Δ_λ^t are (vector-valued) martingale difference sequences, and hence

$$\begin{aligned} S_{2,c} &= \left\langle \frac{1}{T}\sum_{t=1}^{T}\Delta_V^t, V' - V^1\right\rangle + \left\langle \frac{1}{T}\sum_{t=1}^{T}\Delta_\lambda^t, \lambda' - \lambda^1\right\rangle \\ &\le \left\|V' - V^1\right\| \cdot \frac{1}{T}\left\|\sum_{t=1}^{T}\Delta_V^t\right\| + \left\|\lambda' - \lambda^1\right\|_1 \cdot \frac{1}{T}\left\|\sum_{t=1}^{T}\Delta_\lambda^t\right\|_\infty. \end{aligned}$$

The quantity $\left\|\sum_{t=1}^{T}\Delta_V^t\right\|$ and $\left\|\sum_{t=1}^{T}\Delta_\lambda^t\right\|_\infty$ both can be bounded by applying Lemma B.1. More specifically, with probability at least $1 - \delta/20$, it holds that

$$\begin{aligned} \left\|\frac{1}{T}\sum_{t=1}^{T}\Delta_V^t\right\| &\lesssim \sigma_V\sqrt{\frac{\log(|\mathcal{S}|/\delta)}{T}} + M_V\frac{\log(|\mathcal{S}|/\delta)}{T}, \\ \left\|\frac{1}{T}\sum_{t=1}^{T}\Delta_\lambda^t\right\|_\infty &\lesssim \sigma_\lambda\sqrt{\frac{\log(I/\delta)}{T}} + M_\lambda\frac{\log(I/\delta)}{T}. \end{aligned}$$

Therefore, we have

$$S_{2,c} \lesssim \left(D_V\sigma_V + D_{\lambda,1}\sigma_\lambda\right)\sqrt{\frac{\iota}{T}} + \left(D_V M_V + D_{\lambda,1}M_\lambda\right)\frac{\iota}{T}.$$

Bounding the term $S_{2,m}$ In order to bound the martingale part $S_{2,m}$, we have to consider martingales difference sequences[3] $\overline{\Delta}_V^t := \left\langle \Delta_V^t, V^1 - V^t\right\rangle$, $\overline{\Delta}_\lambda^t := \left\langle \Delta_\lambda^t, \lambda^1 - \lambda^t\right\rangle$, $\overline{\Delta}_x^t := \left\langle \Delta_x^t, x^t - x'\right\rangle$. We estimate the variance and magnitude as

$$\begin{aligned} &\left|\overline{\Delta}_V^t\right| \le 2D_V M_V, \quad \mathbb{E}\left[\left(\overline{\Delta}_V^t\right)^2\middle|\mathcal{F}_t\right] \le \mathbb{E}\left[\left\|V^1 - V'\right\|^2\left\|\Delta_V^t\right\|^2\middle|\mathcal{F}_t\right] \le D_V^2\sigma_V^2, \\ &\left|\overline{\Delta}_\lambda^t\right| \le 2D_{\lambda,1}M_\lambda, \quad \mathbb{E}\left[\left(\overline{\Delta}_\lambda^t\right)^2\middle|\mathcal{F}_t\right] \le \mathbb{E}\left[\left\|\lambda^1 - \lambda^t\right\|_1^2\left\|\Delta_\lambda^t\right\|_\infty^2\middle|\mathcal{F}_t\right] \le D_{\lambda,1}^2\sigma_\lambda^2, \\ &\left|\overline{\Delta}_x^t\right| \le 2D_{x,1}M_x, \quad \mathbb{E}\left[\left(\overline{\Delta}_x^t\right)^2\middle|\mathcal{F}_t\right] \le \mathbb{E}\left[\left\|\frac{x' - x^t}{\sqrt{x' + x^t}}\right\|^2\left\|\Delta_x^t\right\|_{x'+x^t}^2\middle|\mathcal{F}_t\right] \le 2D_{x,1}^2\sigma_x^2. \end{aligned}$$

Thus, by the Bernstein's Inequality, the following holds with probability at least $1 - \delta/20$:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\overline{\Delta}_V^t &\lesssim D_V\sigma_V\sqrt{\frac{\log(1/\delta)}{T}} + \frac{D_V M_V\log(1/\delta)}{T}, \\ \frac{1}{T}\sum_{t=1}^{T}\overline{\Delta}_\lambda^t &\lesssim D_{\lambda,1}\sigma_\lambda\sqrt{\frac{\log(1/\delta)}{T}} + \frac{D_{\lambda,1}M_\lambda\log(1/\delta)}{T}, \\ \frac{1}{T}\sum_{t=1}^{T}\overline{\Delta}_x^t &\lesssim D_{x,1}\sigma_x\sqrt{\frac{\log(1/\delta)}{T}} + \frac{D_{x,1}M_x\log(1/\delta)}{T}. \end{aligned}$$

[3] They are martingale difference sequences w.r.t. the filtration $(\mathcal{F}_t)$ defined by $\mathcal{F}_t = \sigma(\zeta_1, \cdots, \zeta_{t-1})$.

Therefore, with probability at least $1 - \delta/20$,

$$S_{2,m} \lesssim (D_V \sigma_V + D_{\lambda,1}\sigma_\lambda + D_{x,1}\sigma_x)\sqrt{\frac{\iota}{T}} + (D_V M_V + D_{\lambda,1} M_\lambda + D_{x,1} M_x)\frac{\iota}{T}.$$

Bounding the term S_2 Finally, combining the bounds on $S_{2,m}$ and $S_{2,c}$ proves the inequality (26).

D.4 Basics of mirror descent

Before we provide the proof of Lemma D.1, we state a basic property of the mirror descent (see e.g. [6]).

Lemma D.4. *Under the same assumption in Lemma D.1, it holds that for any $Y' \in \mathcal{Y}$,*

$$\eta \left\langle Y^{k+1} - Y', g^k \right\rangle \le \mathrm{KL}(Y' \parallel Y^k) - \mathrm{KL}(Y' \parallel Y^{k+1}) - \mathrm{KL}(Y^{k+1} \parallel Y^k).$$

In particular,

$$\mathrm{KL}(Y^k \parallel Y^{k+1}) + \mathrm{KL}(Y^{k+1} \parallel Y^k) \le \eta \left\langle Y^k - Y^{k+1}, g^k \right\rangle.$$

Proof of Lemma D.1. By the fact that $(x-y)\log\frac{x}{y} \ge \frac{(x-y)^2}{\max(x,y)}$, we have

$$\mathrm{KL}(Y^k \parallel Y^{k+1}) + \mathrm{KL}(Y^{k+1} \parallel Y^k) = \left\langle Y^k - Y^{k+1}, \log Y^k - \log Y^{k+1}\right\rangle \ge \sum_i \frac{(Y_i^k - Y_i^{k+1})^2}{\max(Y_i^k, Y_i^{k+1})}.$$

Together with Lemma D.4, the estimation above yields

$$\left\| \frac{Y^k - Y^{k+1}}{\sqrt{Y^k + Y^{k+1}}} \right\|^2 \le \mathrm{KL}(Y^k \parallel Y^{k+1}) + \mathrm{KL}(Y^{k+1} \parallel Y^k) \le \eta \left\langle Y^k - Y^{k+1}, g^k \right\rangle.$$

By Cauchy inequality, $\left\langle Y^k - Y^{k+1}, g^k \right\rangle \le \left\| \frac{Y^k - Y^{k+1}}{\sqrt{Y^k + Y^{k+1}}} \right\| \left\| g^k \sqrt{Y^k + Y^{k+1}} \right\|$, and hence

$$\begin{aligned}
\left\| \frac{Y^k - Y^{k+1}}{\sqrt{Y^k + Y^{k+1}}} \right\| &\le \eta \left\| g^k \sqrt{Y^k + Y^{k+1}} \right\| = \eta \left\| g^k \right\|_{Y^k + Y^{k+1}}, \\
\left\langle Y^k - Y^{k+1}, g^k \right\rangle &\le \left\| \frac{Y^k - Y^{k+1}}{\sqrt{Y^k + Y^{k+1}}} \right\| \left\| g^k \sqrt{Y^k + Y^{k+1}} \right\| \le \eta \left\| g^k \right\|_{Y^k + Y^{k+1}}^2.
\end{aligned}$$

To further bound $\left\| g^k \right\|_{Y^k + Y^{k+1}}$ in terms of $\left\| g^k \right\|_{Y^k}$, we estimate it as

$$\begin{aligned}
\left\| g^k \right\|_{Y^k + Y^{k+1}}^2 &= \sum_i (Y_i^k + Y_i^{k+1})(g_i^k)^2 \\
&\le 2\left\| g^k \right\|_{Y^k}^2 + \sum_i \left| Y_i^{k+1} - Y_i^k \right| (g_i^k)^2 \\
&\le 2\left\| g^k \right\|_{Y^k}^2 + \max_i \left| g_i^k \right| \left\| \frac{Y^k - Y^{k+1}}{\sqrt{Y^k + Y^{k+1}}} \right\| \left\| g^k \right\|_{Y^k + Y^{k+1}} \\
&\le 2\left\| g^k \right\|_{Y^k}^2 + \eta \left\| g^k \right\|_\infty \left\| g^k \right\|_{Y^k + Y^{k+1}}^2.
\end{aligned}$$

Thus, as long as $\eta \le \frac{1}{2\left\| g^k \right\|_\infty}$, it holds that $\left\| g^k \right\|_{Y^k + Y^{k+1}} \le 2\left\| g^k \right\|_{Y^k}$. Therefore, for all $Y' \in \mathcal{Y}$,

$$\begin{aligned}
\left\langle Y^k - Y', g^k \right\rangle &\le \frac{1}{\eta}\left[\mathrm{KL}(Y' \parallel Y^k) - \mathrm{KL}(Y' \parallel Y^{k+1})\right] + \left\langle Y^k - Y^{k+1}, g^k \right\rangle \\
&\le \frac{1}{\eta}\left[\mathrm{KL}(Y' \parallel Y^k) - \mathrm{KL}(Y' \parallel Y^{k+1})\right] + 4\eta \left\| g^k \right\|_{Y^k}^2.
\end{aligned}$$

Summing over $k = 1, \cdots, T$ completes the proof. □

Corollary D.5. *Under the same assumption in Lemma D.1, it holds that for each k,*

$$\begin{aligned}
&\left\| \frac{Y^k - Y^{k+1}}{\sqrt{Y^k + Y^{k+1}}} \right\| \le 2\eta \left\| g^k \right\|_{Y^k}, \\
&\left\| Y^{k+1} - Y^k \right\|_1 \le 4\eta\sqrt{D_{Y,1}} \left\| g^k \right\|_{Y^k} \le 4\eta D_{Y,1} \left\| g^k \right\|_\infty.
\end{aligned}$$

Proof. From the proof of Lemma D.1 above, we see

$$J(Y^k, Y^{k+1}) = \mathrm{KL}(Y^k \parallel Y^{k+1}) + \mathrm{KL}(Y^{k+1} \parallel Y^k) \le \eta \left\langle Y^k - Y^{k+1}, g^k \right\rangle \le 4\eta^2 \left\|g^k\right\|_{Y^k}^2 .$$

Then by Lemma D.6 we have

$$\left\|Y^{k+1} - Y^k\right\|_1 \le \left(\sqrt{\left\|Y^k\right\|_1} + \sqrt{\left\|Y^{k+1}\right\|_1}\right) \sqrt{J(Y^k, Y^{k+1})} \le 4\eta\sqrt{D_{Y,1}} \left\|g^k\right\|_{Y^k} . \qquad \square$$

Lemma D.6 (Generalized Pinsker's Inequality). *For $y, y' \in \mathbb{R}^n_{>0}$, we consider the generalized Jeffery divergence between them:*

$$J(y, y') := \mathrm{KL}(y \parallel y') + \mathrm{KL}(y' \parallel y) = \sum_i (y_i - y'_i) \log \frac{y_i}{y'_i}.$$

Then it holds that

$$\|y - y'\|_1 \le \left(\sqrt{\|y\|_1} + \sqrt{\|y'\|_1}\right) \sqrt{J(y, y')}.$$

Proof. Denote $J = J(y, y')$, $Y = \|y\|_1$, $Y' = \|y'\|_1$. We consider two (normalized) distributions $\overline{y} = \frac{y}{Y}$ and $\overline{y}' = \frac{y'}{Y'}$, then

$$\begin{aligned} J(y, y') &= \sum_i (y_i - y'_i) \log \frac{y_i}{y'_i} \\ &= \sum_i \left(Y\overline{y}_i - Y'\overline{y}'_i\right) \left(\log \frac{\overline{y}_i}{\overline{y}'_i} + \log \frac{Y}{Y'}\right) \\ &= Y\,\mathrm{KL}(\overline{y} \parallel \overline{y}') + Y'\,\mathrm{KL}(\overline{y}' \parallel \overline{y}) + (Y - Y') \log \frac{Y}{Y'} \\ &\ge (Y + Y') \cdot \frac{1}{2} \left\|\overline{y} - \overline{y}'\right\|_1^2 + \frac{|Y - Y'|^2}{\max(Y, Y')}, \end{aligned}$$

where the last inequality is due to Pinsker's inequality and the fact $(x - y) \log \frac{x}{y} \ge \frac{(x-y)^2}{\max(x,y)}$. Therefore, w.l.o.g. $Y < Y'$, then $|Y - Y'| \le \sqrt{Y'J}$, and

$$\sqrt{\frac{2J}{Y + Y'}} \ge \|\overline{y} - \overline{y}'\|_1 = \left\|\frac{y}{Y} - \frac{y'}{Y'}\right\|_1 = \left\|\frac{y - y'}{Y} + \frac{y'}{Y'}\left(\frac{Y'}{Y} - 1\right)\right\|_1 .$$

Hence, we have

$$\begin{aligned} \|y - y'\|_1 &\le \left\|\frac{y'}{Y'}\,(Y' - Y)\right\|_1 + Y\sqrt{\frac{J}{Y + Y'}} \\ &= |Y' - Y| + Y\sqrt{\frac{J}{Y + Y'}} \\ &\le \sqrt{Y'J} + \sqrt{YJ}. \end{aligned} \qquad \square$$

E Proof of Theorem 4.1

In this section, we provide the proof of Theorem 4.1 and Remark 4.2. We should notice that if $\psi \ge C^*$, then $\epsilon_{\mathrm{approx}}(\psi)$ reduces to 0, and the result in Remark 4.2 actually agrees with Theorem 4.1. Thus we handle them simultaneously. The key to the analysis is controlling the reward sub-optimality gap and the constraint violation in terms of the duality gap $\mathrm{Gap}(\overline{x})$ that is bounded in Theorem 4.4. Before presenting the proof, let us introduce a few notations and lemmas.

E.1 Notations and supporting lemmas

In this proof, we will view ν as vectors in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and we define a matrix A as

$$A := \left[\mathbb{1}_{\{s'=s\}} - \gamma\mathbb{P}\left(s'|s,a\right)\right]_{(s,a),s'} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}. \tag{29}$$

Given the matrix A, we conveniently write $\sum_a(\mathbf{I}-\gamma\mathbb{P}_a)\nu_a$ as $A^\top\nu$. For the reweighted saddle point problem (7), one can easily partially minimize over V and λ since their domains are simple normal balls. Therefore, we define

$$\mathcal{J}_\kappa(x) := \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,x) = r^TWx - R_\mathcal{V}\left\|A^\top Wx - \rho_0\right\|_1 - R_\Lambda \left\|[U_\kappa Wx]_-\right\|_\infty, \tag{30}$$

where we denote $R_\mathcal{V} = \frac{8}{1-\gamma}\left(1+\frac{2}{\varphi}\right)$, $R_\Lambda = \frac{8}{\varphi}$. We also define

$$j(\psi) := \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,x) = \max_{x\in\mathcal{X}} \mathcal{J}_\kappa(x) \tag{31}$$

as the optimal value of problem (7). Then $j(\psi)$ has an implicit dependence on κ due to the term $\|[U_\kappa Wx]_-\|_\infty$. In particular, we will write $j_0(\psi)$ for the case where $\kappa = 0$. Finally, we define π^*_κ as the optimal policy with κ conservative constraints. That is,

$$\pi^*_\kappa = \arg\max_\pi J(\pi) \text{ s.t. } J^u_i(\pi) \geq \kappa, \ \forall i \in [I].$$

Then the following lemmas hold true.

Lemma E.1. *Let π^* be the optimal policy, and let π^*_κ be defined above, then it holds that*

$$J(\pi^*) \geq J(\pi^*_\kappa) \geq J(\pi^*) - \frac{2\kappa}{\varphi}.$$

Proof. The inequality $J(\pi^*) \geq J(\pi^*_\kappa)$ follows from definition. For the other inequality, we fix a "baseline" policy $\tilde{\pi}$ satisfying the Slater's condition, namely $J^u_i(\tilde{\pi}) \geq \frac{\varphi}{1-\gamma}$. Let $s = \frac{(1-\gamma)\kappa}{\varphi}$, we interpolate $\nu_s := s\nu^{\pi^*} + (1-s)\nu^{\tilde{\pi}}$. ν_s is still an occupancy measure such that $\langle u_i, \nu_s\rangle \geq s\left\langle u_i, \nu^{\tilde{\pi}}\right\rangle \geq \kappa$ for $\forall i \in [I]$, and

$$\langle r, \nu_s\rangle = \left\langle r, \nu^{\pi^*}\right\rangle - s(\left\langle r, \nu^{\pi^*}\right\rangle - \left\langle r, \nu^{\tilde{\pi}}\right\rangle) \geq \left\langle r, \nu^{\pi^*}\right\rangle - \frac{2s}{1-\gamma} = J(\pi^*) - \frac{2\kappa}{\varphi}.$$

We complete the proof by noticing $J(\pi^*_\kappa) \geq \langle r, \nu_s\rangle$. □

The next lemma discusses the property of $j(\cdot)$.

Lemma E.2. *Suppose the policy class $\Pi(\psi)$ satisfies Slater's condition, then it holds that*

$$j(\psi) \geq \max_{\pi\in\Pi(\psi)\cap\mathfrak{S}} J(\pi) - \frac{2\kappa}{\varphi} = J(\pi^*) - \epsilon_{\text{approx}}(\psi) - \frac{2\kappa}{\varphi}.$$

Proof. Similar to the proof of Lemma E.1, we fix a $\hat{\pi} = \arg\max_{\pi\in\Pi(\psi)\cap\mathfrak{S}} J(\pi)$ and a "baseline" policy $\tilde{\pi} \in \Pi(\psi)$ satisfying the Slater's condition. Let $\hat{\nu} := \nu^{\hat{\pi}}$ and $\tilde{\nu} := \nu^{\tilde{\pi}}$ be the corresponding occupancy measures. Let $s = \frac{(1-\gamma)\kappa}{\varphi}$, then $\nu_s := s\tilde{\nu} + (1-s)\hat{\nu}$ is still an occupancy measure for which the corresponding policy belongs to $\Pi(\psi)$. For $i \in [I]$, $\langle u_i, \nu_s\rangle \geq s\langle u_i, \tilde{\nu}\rangle \geq \kappa$, and

$$\langle r, \nu_s\rangle = \langle r, \hat{\nu}\rangle - s(\langle r, \hat{\nu}\rangle - \langle r, \tilde{\nu}\rangle) \geq \langle r, \hat{\nu}\rangle - \frac{2s}{1-\gamma} = \langle r, \tilde{\nu}\rangle - \frac{2\kappa}{\varphi}.$$

Now $W^{-1}\nu_s \in \mathcal{X}$ by Proposition 4.3, and

$$j(\psi) \geq \mathcal{J}_\kappa(W^{-1}\nu_s) = \langle r, \nu_s\rangle \geq \langle r, \hat{\nu}\rangle - \frac{2\kappa}{\varphi} = \max_{\pi\in\Pi(\psi)\cap\mathfrak{S}} J(\pi) - \frac{2\kappa}{\varphi} = J(\pi^*) - \epsilon_{\text{approx}}(\psi) - \frac{2\kappa}{\varphi}. \ \square$$

The following result is obtained from [4, Lemma 3], by replacing λ and (v, u) in [4, Lemma 3] with our notation $(1-\gamma)\nu$ and (V, λ), respectively.

Lemma E.3. *For any dual optimal solution $(V_\kappa^*, \lambda_\kappa^*)$ of the problem* (4)*, where the constraint utilities u_i is replaced with the shifted utilities u_i^κ, we have*

$$\|\lambda_\kappa^*\|_1 \le \frac{2}{\varphi} \qquad \textit{and} \qquad \|V_\kappa^*\| \le \frac{1}{1-\gamma}\left(1+\frac{2}{\varphi}\right).$$

For any $\nu \in \mathbb{R}_{\ge 0}^{|\mathcal{S}||\mathcal{A}|}$ and any $\Delta > 0$, the inequality $J(\pi_\kappa^*) - \mathcal{J}(W^{-1}\nu) \le \Delta$ immediately implies that

$$J(\pi_\kappa^*) - \langle r, \nu\rangle \le \Delta, \quad \left\|A^\top \nu - \rho_0\right\|_1 \le \frac{2\Delta}{R_{\mathcal{V}}}, \quad \textit{and} \quad \|[U_\kappa \nu]_-\|_\infty \le \frac{2\Delta}{R_\Lambda}$$

as long as $R_{\mathcal{V}} \ge 2\left\|V_\kappa^*\right\|_\infty, R_\Lambda \ge 2\left\|\lambda_\kappa^*\right\|_1$.

Finally, we introduce the last lemma that is needed in this proof.

Lemma E.4. *For any vector $\tilde{\nu} \in \mathbb{R}_{\ge 0}^{|\mathcal{S}||\mathcal{A}|}$ that is an approximate visitation measure, consider its associate policy $\tilde{\pi}$ defined by $\tilde{\pi}(a|s) = \frac{\tilde{\nu}(s,a)}{\sum_{a'} \tilde{\nu}(s,a')}$. Let $\nu^{\tilde{\pi}}$ be the true visitation measure of $\tilde{\pi}$, then*

$$\left\|\tilde{\nu} - \nu^{\tilde{\pi}}\right\|_1 \le \frac{1}{1-\gamma}\left\|A^\top \tilde{\nu} - \rho_0\right\|_1.$$

Proof. For policy π, we consider its state visitation measure ν_π defined by $\nu_\pi(s) = \sum_a \nu^\pi(s,a)$. Then $\nu^\pi(s,a) = \pi(a|s)\nu_\pi(s)$. With the transition matrix $\mathbb{P}_\pi(s'|s) = \sum_a \pi(a|s)\mathbb{P}(s'|s,a)$, then the constraint $A^\top \nu^\pi = \rho_0$ is equivalent to $(I - \gamma\mathbb{P}_\pi)\nu_\pi = \rho_0$.

Let $\tilde{\pi}$ induced by $\tilde{\nu}$, then $\nu_{\tilde{\pi}}$ satisfies $(I - \gamma\mathbb{P}_{\tilde{\pi}})\nu_{\tilde{\pi}} = \rho_0$. Let $\tilde{\nu}'$ be defined by $\tilde{\nu}'(s) = \sum_a \tilde{\nu}(s,a)$, then $\tilde{\nu}(s,a) = \tilde{\pi}(a|s)\tilde{\nu}'(s)$, and hence $(I - \gamma\mathbb{P}_{\tilde{\pi}})\tilde{\nu}' = A\tilde{\nu}$. Therefore,

$$\|\nu_{\tilde{\pi}} - \tilde{\nu}'\|_1 = \left\|(I - \gamma\mathbb{P}_{\tilde{\pi}})^{-1}(\rho_0 - A^\top\tilde{\nu})\right\|_1 \le \left\|(I - \gamma\mathbb{P}_{\tilde{\pi}})^{-1}\right\|_1 \left\|A^\top\tilde{\nu} - \rho_0\right\|_1 \le \frac{1}{1-\gamma}\left\|A^\top\tilde{\nu} - \rho_0\right\|_1.$$

We finalize the proof by the following equality

$$\left\|\nu^{\tilde{\pi}} - \tilde{\nu}\right\|_1 = \sum_{s,a}\left|\tilde{\pi}(a|s)(\nu_{\tilde{\pi}}(s) - \tilde{\nu}'(s))\right| = \sum_s \Big(\sum_a \tilde{\pi}(a|s)\Big)|\nu_{\tilde{\pi}}(s) - \tilde{\nu}'(s)| = \|\nu_{\tilde{\pi}} - \tilde{\nu}'\|_1. \quad \square$$

E.2 Analysis

Now we are ready to present the proof of Remark 4.2 and Theorem 4.1.

Proof. By definition of $\text{Gap}(\overline{x})$, we have

$$\text{Gap}(\overline{x}) = \max_{x\in\mathcal{X}} \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,x) - \min_{V\in\mathcal{V},\lambda\in\Lambda} \mathcal{L}_w(V,\lambda,\overline{x}) = j(\psi) - \mathcal{J}_\kappa(\overline{x}). \tag{32}$$

Define $\overline{\nu} = W\overline{x}$, and define $\Delta := J(\pi_\kappa^*) - \mathcal{J}_\kappa(W^{-1}\overline{\nu})$, then we have

$$\Delta = \text{Gap}(\overline{x}) + J(\pi_\kappa^*) - j(\psi) \overset{(i)}{\le} \text{Gap}(\overline{x}) + J(\pi^*) - j(\psi) \overset{(ii)}{\le} \text{Gap}(\overline{x}) + \epsilon_{\text{approx}}(\psi) + \frac{2\kappa}{\varphi}, \tag{33}$$

where (i) is because $J(\pi_\kappa^*) \le J(\pi^*)$ and (ii) is due to Lemma E.2. Now, let $\nu^{\overline{\pi}}$ be the true visitation measure of $\overline{\pi}$, where $\overline{\pi}(a|s) := \frac{\overline{x}(s,a)}{\sum_{a'}\overline{x}(s,a')}$. Then Lemma E.4 immediately indicates that

$$\begin{aligned}
\left\|\nu^{\overline{\pi}} - \overline{x}\right\|_1 &\le \frac{1}{1-\gamma}\left\|A^\top\overline{x} - \rho_0\right\|_1 \\
&\le \frac{1}{1-\gamma}\Big(\|A^\top(\overline{x} - W\overline{x})\|_1 + \|A^\top W\overline{x} - \rho_0\|_1\Big) \\
&\le \frac{1}{1-\gamma}\Big(2\left\|\overline{x} - \overline{\nu}\right\|_1 + \left\|A^\top\overline{\nu} - \rho_0\right\|_1\Big)
\end{aligned}$$

which further gives

$$\left\|\nu^{\overline{\pi}} - \overline{\nu}\right\|_1 \le \frac{1}{1-\gamma}\big(\left\|A^\top\overline{\nu} - \rho_0\right\|_1 + 3\left\|\overline{x} - \overline{\nu}\right\|_1\big). \tag{34}$$

Consequently, we have

$$
\begin{aligned}
& J(\pi_\kappa^*) - \mathcal{J}_\kappa(W^{-1}\nu^{\overline{\pi}}) \\
\overset{(i)}{=} \quad & J(\pi_\kappa^*) - \left\langle r, \nu^{\overline{\pi}} \right\rangle + R_\Lambda \left\| \, \left[U_\kappa \nu^{\overline{\pi}} \right]_- \right\|_\infty \\
\overset{(ii)}{\leq} \quad & J(\pi_\kappa^*) - \langle r, \overline{\nu} \rangle + R_\Lambda \left\| [U_\kappa \overline{\nu}]_- \right\|_\infty + \frac{1 + \frac{3}{2} R_\Lambda}{1-\gamma} \left(\left\| A^\top \overline{\nu} - \rho_0 \right\|_1 + 3 \left\| \overline{x} - \overline{\nu} \right\|_1 \right) \\
\overset{(iii)}{\leq} \quad & J(\pi_\kappa^*) - \mathcal{J}_\kappa(W^{-1}\overline{\nu}) + \frac{5(R_\Lambda + 1)}{1-\gamma} \left\| \overline{x} - \overline{\nu} \right\|_1 \\
\overset{(iv)}{\leq} \quad & \Delta + 45\epsilon_e,
\end{aligned}
$$

where (i) is because $\|A^\top \nu^{\overline{\pi}} - \rho_0\|_1 = 0$, (ii) is due to the fact that $\left| \left\langle r, \nu^{\overline{\pi}} \right\rangle - \langle r, \overline{\nu} \rangle \right| \leq \left\| \nu^{\overline{\pi}} - \overline{\nu} \right\|_1$ and $\left| \|[U_\kappa \nu^{\overline{\pi}}]_-\|_\infty - \|[U_\kappa \overline{\nu}]_-\|_\infty \right| \leq \frac{3}{2} \left\| \nu^{\overline{\pi}} - \overline{\nu} \right\|_1$, (iii) is because of $\frac{1+\frac{3}{2}R_\Lambda}{1-\gamma} \leq R_\mathcal{V}$, and (iv) is because of $\left\| \overline{x} - \overline{\nu} \right\|_1 \leq \varphi(1-\gamma)\epsilon_e$ by Proposition 4.3. Finally, applying Lemma E.3 to $\nu^{\overline{\pi}}$ yields

$$
J(\pi_\kappa^*) - \left\langle r, \nu^{\overline{\pi}} \right\rangle \leq \Delta + 45\epsilon_e, \qquad \left\| \left[U_\kappa \nu^{\overline{\pi}} \right]_- \right\|_\infty \leq \frac{\varphi}{4} \left(\Delta + 45\epsilon_e \right).
$$

By Lemma E.2, we have

$$
\begin{aligned}
& J(\pi^*) - \left\langle r, \nu^{\overline{\pi}} \right\rangle \leq J(\pi^*) - j(\psi) + \operatorname{Gap}(\overline{x}) + 45\epsilon_e \leq \operatorname{Gap}(\overline{x}) + \epsilon_{\text{approx}}(\psi) + \frac{2\kappa}{\varphi} + 45\epsilon_e, \\
& J_i^u(\overline{\pi}) \geq \kappa - \left\| \, \left[U_\kappa \nu^{\overline{\pi}} \right]_- \right\|_\infty \geq \frac{\kappa}{2} - \frac{\varphi}{4} \left(\operatorname{Gap}(\overline{x}) + \epsilon_{\text{approx}}(\psi) \right) - 12\varphi\epsilon_e.
\end{aligned}
\tag{35}
$$

Combining the above inequality with the fact that $\epsilon_e = \frac{\epsilon}{100}$, $\kappa = 5\varphi\epsilon$, $\operatorname{Gap}(\overline{x}) \leq \epsilon/2$ completes the proof. □

Finally, we point out a by-product of the above analysis, which is useful for the VERIFY method.

Corollary E.5. *Under the same assumption of Theorem 4.1, with probability at least* $1 - 2\delta/3$*, it holds that*

$$
\left\| A^\top \overline{\nu} - \rho_0 \right\|_1 \leq \frac{11}{8} \varphi(1-\gamma)\epsilon, \qquad \left\| [U_\kappa \overline{\nu}]_- \right\|_\infty \leq \frac{11}{4} \varphi\epsilon \tag{36}
$$

for $\overline{\nu} := W\overline{x}$.

Proof. Due to $\psi \geq C^*$ and Lemma E.2, we have

$$
J(\pi_\kappa^*) - \mathcal{J}_\kappa(\overline{x}) \leq j(\psi) - \mathcal{J}_\kappa(\overline{x}) + \frac{2\kappa}{\varphi} = \operatorname{Gap}(\overline{x}) + \frac{2\kappa}{\varphi} \leq 11\epsilon.
$$

Applying Lemma E.3 yields

$$
\left\| A^\top \overline{\nu} - \rho_0 \right\|_1 \leq \frac{2 \cdot 11\epsilon}{R_\mathcal{V}} \leq \frac{11}{8} \varphi(1-\gamma)\epsilon, \qquad \|[U_\kappa \overline{\nu}]_-\|_\infty \leq \frac{2 \cdot 11\epsilon}{R_\Lambda} = \frac{11}{4} \varphi\epsilon. \qquad \square
$$

F Proofs for Section 5

F.1 Proof of Theorem 5.1

In this section, we provide the complete version of the construction illustrated in Section 5. Let us define

$$
K := \min\left(\left\lfloor \frac{I}{2} \right\rfloor, \left\lfloor \frac{A-1}{2} \right\rfloor \right), \quad S_c = \min\left(\left\lfloor \frac{I}{2K} \right\rfloor, S \right), \quad S_u = \begin{cases} S - S_c, & \text{if } S_c < S - 3, \\ 0, & \text{otherwise.} \end{cases}
$$

The CMDP instance $\mathcal{M}$ that we construct consists of two groups of basic blocks. The first group includes S_c replicas of the basic block characterized in Fig. 1, each with actions $\{a_1, b_1, ..., a_k, b_k, e\}$ and $2K$ constraints. The second group includes S_u replicas of the basic blocks characterized by Fig. 1 (a) and Fig. 1(c), each basic block only has two actions $\{a, e\}$ and no constraint. In fact the construction of the second group (“unconstrained part”) is similar to the hard MDP constructed in [21]. The transition kernel $\mathbb{P}_\theta$ of $\mathcal{M}$ is parametrized by $\theta = (\theta_c, \theta_u) \in \Theta := \{-1, +1\}^{S_c K} \times \{-1, +1\}^{S_u}$ and $\varpi_c, \varpi_u \in (0, \frac{1}{2}]$. The details of $\mathcal{M}$ are listed as follows.

States and actions The state space $\mathcal{S}$ consists of $S_c + S_u$ 4-state basic blocks, plus an extra “null” state s_{-1}. The first S_c basic blocks are exactly what we described in Section 5, we write $\mathcal{S}_c = \bigsqcup_{j=1}^{S_c} \{s_0^j, s_1^j, s_\oplus^j, s_\ominus^j\}$. The next S_u basic blocks will be described below, we write $\mathcal{S}_u = \bigsqcup_{j=S_c+1}^{S_c+S_u} \{s_0^j, s_1^j, s_\oplus^j, s_\ominus^j\}$. By default, $\mathcal{S}_u = \emptyset$ if $S_u = 0$. Then $\mathcal{S} = \mathcal{S}_c \bigsqcup \mathcal{S}_u \bigsqcup \{s_{-1}\}$. Next, we describe the detailed information of each block j.

- At s_0^j, $s_\oplus^j$ and $s_\ominus^j$, there is no action, and the transition does not depend on θ:

$$\begin{aligned} \mathbb{P}\left(s_0^j \middle| s_0^j\right) &= p, \quad \mathbb{P}\left(s_1^j \middle| s_0^j\right) = 1 - p, \\ \mathbb{P}\left(s_\oplus^j \middle| s_\oplus^j\right) &= q, \quad \mathbb{P}\left(s_0^j \middle| s_\oplus^j\right) = 1 - q, \\ \mathbb{P}\left(s_\ominus^j \middle| s_\ominus^j\right) &= q, \quad \mathbb{P}\left(s_0^j \middle| s_\ominus^j\right) = 1 - q, \end{aligned} \tag{37}$$

  where $p = \frac{1}{2-\gamma}$ and $q = 2 - \frac{1}{\gamma}$. We assign reward as $r(s_\oplus^j) = 1$, $r(s_\ominus^j) = -1$.

- **Constrained state** At $s_1^j \in \mathcal{S}_c$, there are $2K + 1$ actions $a_1, b_1, \cdots, a_K, b_K, e$ such that

$$\begin{aligned} &\mathbb{P}_\theta\left(s_\oplus^j \middle| s_1^j, a_i\right) = \frac{1 + \varpi_c \theta_{i,j}}{2}, && \mathbb{P}_\theta\left(s_\ominus^j \middle| s_1^j, a_i\right) = \frac{1 - \varpi_c \theta_{i,j}}{2}, \\ &\mathbb{P}\left(s_\oplus^j \middle| s_1^j, b_i\right) = \frac{1}{2}\left(1 - \frac{\varpi_c}{2}\right), && \mathbb{P}\left(s_\ominus^j \middle| s_1^j, a_i\right) = \frac{1}{2}\left(1 + \frac{\varpi_c}{2}\right), \\ &\mathbb{P}\left(s_\oplus^j \middle| s_1^j, e\right) = \frac{1}{2}, && \mathbb{P}\left(s_\ominus^j \middle| s_1^j, e\right) = \frac{1}{2}. \end{aligned}$$

  Here we use subscript θ to emphasize the dependency of $\mathbb{P}_\theta$ on θ.[4]

- **Unconstrained state** At $s_1^j \in \mathcal{S}_u$, there are two actions a, e such that

$$\begin{aligned} &\mathbb{P}_\theta\left(s_\oplus^j \middle| s_1^j, a\right) = \frac{1 + \varpi_u \theta_j}{2}, && \mathbb{P}_\theta\left(s_\ominus^j \middle| s_1^j, a\right) = \frac{1 - \varpi_u \theta_j}{2}, \\ &\mathbb{P}\left(s_\oplus^j \middle| s_1^j, e\right) = \frac{1}{2}, && \mathbb{P}\left(s_\ominus^j \middle| s_1^j, e\right) = \frac{1}{2}. \end{aligned}$$

- The null state s_{-1} has no action or reward, and it always transits to itself.

Initial distribution In the initial distribution, $\rho_0(s_{-1}) = \rho_0(s_1^j) = \rho_0(s_\oplus^j) = \rho_0(s_\ominus^j) = 0, \forall j$. The nonzero probabilities only spread across the $\{s_0^j\}$. In the case $S_u > 0$, we choose ρ_0 to be

$$\rho_0(s_0^j) = \begin{cases} \frac{\mathbb{I}\{s_0^j \in \mathcal{S}_c\}}{2S_c} + \frac{\mathbb{I}\{s_0^j \in \mathcal{S}_u\}}{2S_u}, & \text{if } \mathcal{S}_u \neq \emptyset, \\ \frac{1}{S_c}, & \text{otherwise.} \end{cases}$$

Without loss of generality, we will only deal with the case where $\mathcal{S}_u \neq \emptyset$.

Constraints At each constrained block in S_c, for each pair of actions (a_i, b_i) at the state $s_0^j \in \mathcal{S}_c$, we introduce two constraints defined by the utilities

$$u_{i,j}(s_1^j, a_i) = -1, \qquad u_{i,j}(s_1^j, b_i) = 1, \qquad \tilde{u}_{i,j}(s_1^j, b_i) = -1.$$

At all the other state and actions, $u_{i,j}$ and $\tilde{u}_{i,j}$ returns 0. Then we set the constraints to be

$$J_{i,j}^u(\pi) := \langle \nu^\pi, u_{i,j} \rangle \geq 0, \qquad \text{and} \qquad \tilde{J}_{i,j}^u(\pi) := \langle \nu^\pi, \tilde{u}_{i,j} \rangle \geq -\frac{\rho_c v_1}{4K},$$

[4] Here we view $\theta_c \in \{-1, 1\}^{S_c K}$ as a vector indexed by $(i, j) \in [K] \times [S_c]$, and $\theta_{i,j}$ stands for the (i, j)-th component of θ_c. Similarly, we view $\theta_u \in \{-1, 1\}^{S_u}$ as a vector indexed by j with $S_c + 1 \leq j \leq S_c + S_u$, and θ_j stands for the j-th component of θ_u.

where ρ_c and v_1 are constants specified later in (38). After suitable shifting we can make sure that each constraint has the form $J^u \geq 0$. Basically, these two constraints are equivalent to $\pi(a_i|s_1^j) \leq \pi(b_i|s_1^j) \leq \frac{1}{4K}$. We remark that there are in total $S_c K \leq I$ constraints.

Optimal policy First, let us calculate the visitation measure of any given policy π. According to the proof of Lemma E.4, we set ν_π be the state visitation measure and let $\mathbb{P}_\pi$ be the state transition matrix under policy π, then ν_π will be the unique solution to $(I - \gamma\mathbb{P}_\pi)\nu_\pi = \rho_0$. Note that the $S_c + S_u$ basic blocks are in fact independent blocks, i.e., there are no transitions between different blocks. The matrix $(I - \gamma\mathbb{P}_\pi)$ is in fact a block-diagonal with $S_c + S_u$ 4 by 4 blocks and a 1 by 1 block, and we can solve the ν_π block by block. Define the constants

$$v_0 = \frac{2}{(2+\gamma)}, \quad v_1 = \frac{2\gamma}{(2+\gamma)(2-\gamma)}, \quad v = \frac{\gamma^2}{(2+\gamma)(2-\gamma)}, \quad \rho_c = \frac{1}{2S_c}, \quad \rho_u = \frac{1}{2S_u}, \tag{38}$$

and we consider

$$r_j(\pi) = \begin{cases} \sum_i \left(\theta_{i,j}\pi(a_i|s_1^j) - \frac{1}{2}\pi(b_i|s_1^j)\right), & s_1^j \in \mathcal{S}_c, \\ \theta_j \pi(a|s_1^j), & s_1^j \in \mathcal{S}_u. \end{cases}$$

By a direct computation, the state visitation measure of π is given by

$$\nu_\pi(s_\oplus^j) = \frac{v}{1-\gamma}\frac{1+\varpi_\diamond r_j(\pi)}{2}, \qquad \nu_\pi(s_0^j) = \frac{\rho_\diamond v_0}{1-\gamma},$$
$$\nu_\pi(s_\ominus^j) = \frac{v}{1-\gamma}\frac{1-\varpi_\diamond r_j(\pi)}{2}, \qquad \nu_\pi(s_1^j) = \frac{\rho_\diamond v_1}{1-\gamma},$$

where $\diamond$ stands for c if the block j belongs to $\mathcal{S}_c$, and $\diamond$ stands for u if the block j belongs to $\mathcal{S}_u$. Consequently, the cumulative reward and the utilities are

$$\begin{aligned} J(\pi;\theta) &= \sum_j \left(\nu_\pi(s_\oplus^j) - \nu_\pi(s_\ominus^j)\right) = \frac{v}{1-\gamma}\Big(\rho_c\varpi_c \sum_{j:s_1^j\in\mathcal{S}_c} r_j(\pi) + \rho_u\varpi_u \sum_{j:s_1^j\in\mathcal{S}_u} r_j(\pi)\Big), \\ J_{i,j}(\pi;\theta) &= \nu^\pi(s_1^j, b_i) - \nu^\pi(s_1^j, a_i) = \rho_c v_1\left(\pi(b_i|s_1^j) - \pi(a_i|s_1^j)\right), \\ \tilde{J}_{i,j}(\pi;\theta) &= -\nu^\pi(s_1^j, b_i) = -\rho_c v_1 \pi(b_i|s_1^j). \end{aligned} \tag{39}$$

Therefore, π being safe is equivalent to requiring $\pi(a_i|s_1^j) \leq \pi(b_i|s_1^j) \leq \frac{1}{4K}$ for all the constrained block j in $\mathcal{S}_c$, and any $1 \leq i \leq K$. With the above explicit expression of $J(\pi;\theta)$, we know that the (unique) optimal policy $\pi^{*,\theta}$ under the transition dynamic $\mathbb{P}_\theta$ is

$$\begin{aligned} &\pi^{*,\theta}(a_i|s_1^j) = \pi^{*,\theta}(b_i|s_1^j) = \frac{\mathbb{I}\{\theta_{i,j}=1\}}{4K}, && s_1^j \in \mathcal{S}_c, \\ &\pi^{*,\theta}(a|s_1^j) = \mathbb{I}\{\theta_j = 1\}, && s_1^j \in \mathcal{S}_u. \end{aligned} \tag{40}$$

Denote $J_\theta^* := J(\pi^{*,\theta};\theta)$ the optimal safe reward and $\tilde{\theta} = \frac{\theta+1}{2}$, then

$$J_\theta^* = J(\pi^{*,\theta};\theta) = \frac{v}{1-\gamma}\left(\varpi_c\rho_c \sum_{j:s_1^j\in\mathcal{S}_c}\sum_{i=1}^K \frac{\tilde{\theta}_{i,j}}{8K} + \varpi_u\rho_u \sum_{j:s_1^j\in\mathcal{S}_u}\tilde{\theta}_j\right). \tag{41}$$

Reference distribution Finally, we set the reference distribution μ as

$$\mu(s_0^j) = \frac{v_0}{C}\rho_\diamond, \qquad \mu(s_\oplus^j) = \frac{3}{4}\frac{v}{C}\rho_\diamond, \qquad \mu(s_\ominus^j) = \frac{1}{2}\frac{v}{C}\rho_\diamond, \qquad \mu(s_1^j, e) = \frac{v_1(1-\gamma)}{C}\rho_\diamond,$$
$$\begin{cases} \mu(s_1^j, a_i) = \mu(s_1^j, b_i) = \frac{\rho_c v_1(1-\gamma)}{4KC}, \ i \in [I] & s_1^j \in \mathcal{S}_c, \\ \mu(s_1^j, a) = \frac{\rho_u v_1(1-\gamma)}{C}, & s_1^j \in \mathcal{S}_u, \end{cases}$$
$$\mu(s_{-1}) = 1 - \sum_j \left(\mu(s_0^j) + \mu(s_1^j) + \mu(s_\oplus^j) + \mu(s_\ominus^j)\right).$$

As long as $C \geq 2$, $\mu(s_{-1})$ defined above is positive. Also, for any θ, it holds that

$$\max_{s,a}\frac{\nu^{\pi^{*,\theta}}(s,a)}{\mu(s,a)} \leq \frac{C}{1-\gamma}, \quad \sum_{s,a}\frac{\nu^{\pi^{*,\theta}}(s,a)}{\mu(s,a)} \leq \frac{(|\mathcal{S}|+I)C}{1-\gamma}.$$

We denote $\mu_\theta = \mu \otimes \mathbb{P}_\theta$ as the probability measures of the transition pair $\zeta = (s, a, s')$ generated from the reference distribution μ.

Output policy as an estimator of θ Assume that an algorithm $\mathfrak{A}$ consumes N samples generated from μ_θ, and outputs a policy $\hat{\pi}$ that is possibly dependent on the internal randomness of $\mathfrak{A}$. Consider the corresponding random vector $\hat{\pi}_c := \left(\hat{\pi}(a_i|s_1^j)\right)_{i,j}$ and $\hat{\pi}_u := \left(\hat{\pi}(a|s_1^j)\right)_j$. Then, $4K\hat{\pi}_c$ can be viewed as an estimator of $\tilde{\theta}_c$, and $\hat{\pi}_u$ can be viewed as an estimator of $\tilde{\theta}_u$. We establish the following lemma to characterize the error for "misspecifying" the parameter θ.

Lemma F.1. *For any policy* π*, we define*

$$\begin{aligned}\mathcal{L}(\pi;\theta) :=& [J_\theta^* - J(\hat{\pi};\theta)]_+ + \frac{\gamma\varpi_c}{1-\gamma}\sum_{i,j}\left([J_{i,j}(\hat{\pi};\theta)]_- + \left[\tilde{J}_{i,j}(\hat{\pi};\theta) - \frac{v_1}{4SK}\right]_-\right) \\ =& [J_\theta^* - J(\hat{\pi};\theta)]_+ + \frac{\gamma\varpi_c}{1-\gamma}\text{violation}(\hat{\pi};\theta).\end{aligned} \tag{42}$$

Then it holds that

$$\mathcal{L}(\pi;\theta) \geq \frac{v\rho_c\varpi_c}{8K(1-\gamma)}\left\|4K\hat{\pi}_c - \tilde{\theta}_c\right\|_1 + \frac{v\rho_u\varpi_u}{1-\gamma}\left\|\hat{\pi}_u - \tilde{\theta}_u\right\|. \tag{43}$$

Proof. The description of J_θ^* in (41) gives

$$\begin{aligned}\mathcal{L}(\hat{\pi};\theta) =& \frac{v}{1-\gamma}\left[\rho_c\varpi_c\sum_{i,j}\left(\frac{\tilde{\theta}_{i,j}}{8K} - \theta_{i,j}\pi(a_i|s_1^j) + \frac{\pi(b_i|s_1^j)}{2}\right) + \rho_u\varpi_u\sum_j\left(\tilde{\theta}_j - \theta_j\pi(a|s_1^j)\right)\right]_+ \\ &+ \frac{\gamma v_1\rho_c\varpi_c}{1-\gamma}\sum_{i,j}\left(\left[\pi(b_i|s_1^j) - \pi(a_i|s_1^j)\right]_- + \left[\frac{1}{4K} - \pi(b_i|s_1^j)\right]_-\right) \\ \geq& \frac{v}{1-\gamma}\left(\rho_c\varpi_c\sum_{i,j}\delta_{i,j} + \rho_u\varpi_u\sum_j\left(\tilde{\theta}_j - \theta_j\pi(a|s_1^j)\right)\right),\end{aligned}$$

where we use the fact $\gamma v_1 = 2v$, and denote

$$\delta_{i,j} = \frac{\tilde{\theta}_{i,j}}{8K} - \theta_{i,j}\pi(a_i|s_1^j) + \frac{\pi(b_i|s_1^j)}{2} + 2\left[\pi(b_i|s_1^j) - \pi(a_i|s_1^j)\right]_- + 2\left[\frac{1}{4K} - \pi(b_i|s_1^j)\right]_-.$$

Clearly $\tilde{\theta}_j - \theta_j\pi(a|s_1^j) \geq \left|\tilde{\theta}_j - \pi(a|s_1^j)\right|$ for all $s_1^j \in \mathcal{S}_u$. As for $s_1^j \in \mathcal{S}_c$, we consider the case $\theta_{i,j} = 1$ and $\theta_{i,j} = -1$ separately.

Case 1, $\theta_{i,j} = -1$. Directly $\delta_{i,j} \geq \pi(a_i|s_1^j) = \left|\pi(a_i|s_1^j) - \frac{\tilde{\theta}_{i,j}}{4K}\right|$.

Case 2, $\theta_{i,j} = 1$. By the fact that

$$\frac{z}{2} - x + \frac{y}{2} + 2[y-x]_- + 2[z-y]_- \geq \frac{z-x}{2} + \frac{3}{2}[z-x]_- \geq \frac{|z-x|}{2} \quad \forall x, y, z,$$

we can plug in $x = \pi(a_i|s_1^j)$, $y = \pi(b_i|s_1^j)$ and $z = \frac{1}{4K}$ and derive

$$\delta_{i,j} \geq \frac{1}{2}\left|\frac{1}{4K} - \pi(a_i|s_1^j)\right|.$$

Consequently, (43) is established by combining the above inequalities. □

We now invoke the following lemma due to [9] and [24].

Lemma F.2. *For any integer* $n \geq 1$*, there exists a subset* Θ_n *of* $\{-1, 1\}^n$ *such that* $|\Theta_n| \geq \exp(n/8)$*, and for any pair of different* $\theta, \theta' \in \Theta_n$*, one has* $\|\theta - \theta'\|_1 \geq \frac{n}{2}$.

Fix a Θ_c with $n = S_cK$ and a Θ_u with $n = S_u$, we consider the family of CMDPs $\mathfrak{M} := \{\mathcal{M}_\theta\}_{\theta \in \Theta_c \times \Theta_u}$. Intuitively, CMDPs from this family are hard to distinguish according to samples. This idea can be shown mathematically by the following generalized version of Fano's inequality from [3, Lemma 3].

Lemma F.3 (Generalized Fano's inequality)**.** *Let $r \geq 2$ be an integer and let $\mathcal{P}$ be a set of r probability measures on $(\Omega, \mathcal{F})$. Assume that $\theta(\mathbb{P})$ is the parameter of interest with values in a pseudo-metric space $(\mathcal{D}, d)$. Let $\hat{\theta} = \hat{\theta}(X)$ be an estimator of $\theta(\mathbb{P})$ based on a sample X from a distribution $\mathbb{P} \in \mathcal{P}$. Assume that*

$$d\left(\theta(\mathbb{P}), \theta(\mathbb{P}')\right) \geq \alpha, \quad \forall \mathbb{P}, \mathbb{P}' \in \mathcal{P},$$

and

$$\mathrm{KL}(\mathbb{P} \parallel \mathbb{P}') = \int_{\Omega} \log\left(\frac{d\mathbb{P}}{d\mathbb{P}'}\right) d\mathbb{P} \leq \beta.$$

Then it holds that

$$\max_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} d\left(\hat{\theta}, \theta\left(\mathbb{P}\right)\right) \geq \frac{\alpha}{2}\left(1 - \frac{\beta + \log 2}{\log r}\right).$$

It is worth noting that the estimator needs not to belong to $\{\theta(\mathbb{P})\}_{\mathbb{P} \in \mathcal{P}}$. In our problem, the underlying space $(\Omega, \mathcal{F})$ depends on the internal randomness of $\mathfrak{A}$, and the probability measure on $(\Omega, \mathcal{F})$ is the extension of $\mu_\theta^{\otimes N}$ ($\mu_\theta^{\otimes N}$ is the probability measure on $\Omega_0 = (\mathcal{S} \times \mathcal{A} \times \mathcal{S})^N$, the space of the N-tuple of samples $(\zeta_1, \cdots, \zeta_N)$).

The proof of Theorem 5.1 We have already demonstrated that $4K\hat{\pi}_c$ can be viewed as an estimator of $\tilde{\theta}_c$ in Lemma F.1, and hence $8K\hat{\pi}_c - 1$ can be viewed as an estimator of θ_c. We fix a $\theta_u \in \varTheta_u$, then Fano's inequality (Lemma F.3) yields

$$\begin{aligned}\max_{\theta_c \in \varTheta_c} \mathbb{E}_{(\theta_c, \theta_u)} \left\|8K\hat{\pi}_a - 1 - \theta_c\right\|_1 &\geq \frac{S_cK}{2}\left(1 - \frac{\max_{\theta_c, \theta_c' \in \varTheta_c} \mathrm{KL}(\mu_{(\theta_c,\theta_u)}^{\otimes N} \parallel \mu_{(\theta_c',\theta_u)}^{\otimes N}) + \log 2}{\log |\varTheta_c|}\right) \\ &= \frac{S_cK}{2}\left(1 - \frac{N \max_{\theta_c, \theta_c' \in \varTheta_c} \mathrm{KL}(\mu_{(\theta_c,\theta_u)} \parallel \mu_{(\theta_c',\theta_u)}) + \log 2}{\log |\varTheta_c|}\right).\end{aligned}$$

For any $\theta_c, \theta_c' \in \varTheta_c$, we have

$$\begin{aligned}\mathrm{KL}(\mu_{(\theta_c,\theta_u)} \parallel \mu_{(\theta_c',\theta_u)}) &= \sum_{i,j} \mu(s_1^j, a_i)\, \mathrm{KL}\left(\frac{1 + \theta_{i,j}\varpi_c}{2} \,\middle\|\, \frac{1 + \theta_{i,j}'\varpi_c}{2}\right) \\ &\leq \sum_{i,j} \mu(s_1^j, a_i) \frac{4\varpi_c^2}{1 - \varpi_c^2} = \frac{(1-\gamma)v_1}{2C} \frac{\varpi_c^2}{1 - \varpi_c^2}.\end{aligned}$$

Then, taking $\varpi_c = \min\left\{\sqrt{\frac{(S_cK - 3)C}{8(1-\gamma)N}}, \frac{1}{2}\right\}$ is enough to ensure

$$\frac{N \max_{\theta_c, \theta_c' \in \varTheta_c} \mathrm{KL}(\mu_{(\theta_c,\theta_u)} \parallel \mu_{(\theta_c',\theta_u)}) + \log 2}{\log |\varTheta_c|} \leq \frac{5}{6},$$

which further gives $\max_{\theta_c \in \varTheta_c} \mathbb{E}_{(\theta_c,\theta_u)} \left\|8K\hat{\pi}_a - 1 - \theta_c\right\|_1 \geq \frac{S_cK}{12}$, and hence

$$\max_{\theta_c \in \varTheta_c} \mathbb{E}_{(\theta_c,\theta_u)} \left[\frac{1}{S_cK} \left\|4K\hat{\pi}_a - \tilde{\theta}_c\right\|_1\right] \geq \frac{1}{24}.$$

Similarly, we can take $\varpi_u = \min\left\{\sqrt{\frac{(S_u - 3)C}{8(1-\gamma)N}}, \frac{1}{2}\right\}$ to ensure that for any fixed $\theta_c \in \varTheta_c$,

$$\max_{\theta_u \in \varTheta_u} \mathbb{E}_{(\theta_c,\theta_u)} \left[\frac{1}{S_u} \left\|\hat{\pi}_u - \tilde{\theta}_u\right\|_1\right] \geq \frac{1}{24}$$

Therefore, we obtain

$$\begin{aligned}&\max_{\theta \in \varTheta} \mathbb{E}_\theta \mathcal{L}(\hat{\pi}; \theta) \\ &\geq \max_{\theta \in \varTheta} \mathbb{E}_\theta \left[\frac{v\rho_c\varpi_c}{8K(1-\gamma)} \left\|4K\hat{\pi}_c - \tilde{\theta}_c\right\|_1 + \frac{v\rho_u\varpi_u}{1-\gamma} \left\|\hat{\pi}_u - \tilde{\theta}_u\right\|\right] \\ &= \frac{v}{2(1-\gamma)} \max_{\theta_c \in \varTheta_c} \max_{\theta_u \in \varTheta_u} \left\{\frac{\varpi_c}{8} \mathbb{E}_{(\theta_c,\theta_u)} \left[\frac{1}{S_cK} \left\|4K\hat{\pi}_c - \tilde{\theta}_c\right\|_1\right] + \varpi_u \mathbb{E}_{(\theta_c,\theta_u)} \left[\frac{1}{S_u} \left\|\hat{\pi}_u - \tilde{\theta}_u\right\|\right]\right\}\end{aligned}$$

$$\geq \frac{v}{2(1-\gamma)}\left(\frac{\varpi_c}{192}+\frac{\varpi_u}{24}\right) \gtrsim \min\left\{\frac{1}{1-\gamma}, \sqrt{\frac{S_cK+S_u}{(1-\gamma)^3N}}\right\} \gtrsim \min\left\{\frac{1}{1-\gamma}, \sqrt{\frac{\min\{SA, S+I\}}{(1-\gamma)^3N}}\right\}.$$

In conclusion, for a fixed algorithm $\mathfrak{A}$, there exists some $\theta \in \Theta_c \times \Theta_u$, such that for the policy $\hat{\pi}$ output by $\mathfrak{A}$ on $\mathcal{M}_\theta$, either

$$\mathbb{E}_{\mathcal{M}_\theta}\left[J_\theta^* - J(\hat{\pi})\right] \gtrsim \min\left\{\frac{1}{1-\gamma}, \sqrt{\frac{\min\{SA, S+I\}}{(1-\gamma)^3N}}\right\},$$

or

$$\mathbb{E}_{\mathcal{M}_\theta}\left[\text{violation}(\hat{\pi})\right] \gtrsim 1.$$

This completes the proof of Theorem 5.1.

Remark F.4. *The family $(\mathcal{M}_\theta)$ constructed here does not satisfy the Slater's condition with $\varphi = \Theta(1)$, but a small modification can be made to ensure a φ with constant order. Namely, at each s_0^j we add two extra arms e, e', such that $r(s_0^j, e) = 0, r(s_0^j, e') = -1$ and all utilities of e' is 1. The transition at s_0^j is not affected by e, e'. We omit this construction in the argument above for the sake of cleanness and simplicity.*

F.2 Proof of Theorem 5.2

We further extend the idea of construction in Appendix F.1 to show that, when the Slater's condition does not hold, no zero constraint violation can be ensured. Intuitively, we can directly include an extra constraint $J(\pi) \geq J^*$ in the previous construction. However, the subtlety in such a transfer is that, the constraint will leak information of the underlying parameters θ, ϖ. Thus, rather than making ad hoc adaption from Appendix F.1, we present a more interesting construction for the case $I = 1$, as follows.

States and actions We take the state space $\mathcal{S} = \{s_{-1}, s_0, s_\oplus, s_\ominus\} \bigsqcup_{j=1}^{S} \{s^j\}$, with actions and transition dynamic specified as follows. Here we merge the states $s_0^j, s_\oplus^j, s_\ominus^j$ in Appendix F.1 for notational simplicity. The transition dynamic is parametrized by $\theta \in \{0,1\}^S$ and $\varpi \in (0, \frac{1}{2}]$, as follows.

- At s_0, $s_\oplus$ and $s_\ominus$, there is no action, and the transition does not depend on θ:

$$\begin{aligned} &\mathbb{P}\left(s_0 | s_0\right) = p, \quad \mathbb{P}\left(s^j | s_0\right) = \frac{1-p}{S}, \ j \in [S], \\ &\mathbb{P}\left(s_\oplus | s_\oplus\right) = q, \quad \mathbb{P}\left(s_0 | s_\oplus\right) = 1-q, \\ &\mathbb{P}\left(s_\ominus | s_\ominus\right) = q, \quad \mathbb{P}\left(s_0 | s_\ominus\right) = 1-q, \end{aligned} \tag{44}$$

  where $p = \frac{1}{2-\gamma}$ and $q = 2 - \frac{1}{\gamma}$.

- At s^j, there are two actions a, b such that

$$\begin{aligned} \mathbb{P}_\theta\left(s_\oplus | s^j, a\right) &= \frac{1-\varpi\theta_j}{2}, & \mathbb{P}_\theta\left(s_\ominus | s^j, a\right) &= \frac{1+\varpi\theta_j}{2}, \\ \mathbb{P}_\theta\left(s_\oplus | s^j, b\right) &= \frac{1-\varpi(1-\theta_j)}{2}, & \mathbb{P}_\theta\left(s_\ominus | s^j, b\right) &= \frac{1+\varpi(1-\theta_j)}{2}. \end{aligned}$$

- The null state s_{-1} always transits to itself.

Utilities and rewards We assign $u(s_\oplus) = +1, u(s_\ominus) = -1$, and $u(s_0) = u(s^j) = 0$. No reward is assigned to $\mathcal{M}$, namely the only goal in $\mathcal{M}$ is to fulfill the constraint: $J^u(\pi) \geq 0$. Basically, this constraint requires us to determine whether $\theta_i = 1$ for each i.

Optimal policy For any policy π, we define

$$r_j(\pi) = \theta_j \pi(a|s_1^j) + (1-\theta_j)\pi(b|s_1^j), \qquad \overline{r}(\pi) = \frac{1}{S}\sum_{j=1}^{S} r_j(\pi).$$

Then by exactly the same calculation as in Appendix F.1, we have

$$
\begin{aligned}
\nu_\pi(s_0) &= \frac{v_0}{1-\gamma}, & \nu_\pi(s^j) &= \frac{v_1}{S}, \\
\nu_\pi(s_\oplus) &= \frac{1-\varpi\overline{r}(\pi)}{2}\frac{v}{1-\gamma}, & \nu_\pi(s_\ominus) &= \frac{1+\varpi\overline{r}(\pi)}{2}\frac{v}{1-\gamma}.
\end{aligned}
$$

Therefore, it holds that

$$
J^u(\pi) = -\frac{v}{1-\gamma}\overline{r}(\pi) = -\frac{v}{S(1-\gamma)}\left\|\pi_b-\theta\right\|_1, \tag{45}
$$

where we denote $\pi_b = \left(\pi(b|s_1^j)\right)_j$ for a policy π. Hence, there is a unique safe policy $\pi^{*,\theta}$ in $\mathcal{M}_\theta$ that can be specified by

$$
\pi^{*,\theta}(a|s^j) = 1-\theta_j, \qquad \pi^{*,\theta}(b|s^j) = \theta_j, \qquad j\in[S].
$$

The formula (45) also indicates that, for $\hat{\pi}$ outputed by an algorithm $\mathfrak{A}$ after consuming N samples, the vector $\hat{\pi}_b$ can be viewed as an estimator of θ.

Reference distribution We take $\rho_0(s_0)=1$. The reference distribution μ is chosen similar to Appendix F.1, namely

$$
\begin{aligned}
\mu(s_0) &= \frac{v_0}{C}, & \mu(s^j,a) &= \mu(s^j,b) = \frac{v_1(1-\gamma)}{SC}, \\
\mu(s_\oplus) = \mu(s_\ominus) &= \frac{v}{C}, & \mu(s_{-1}) &= 1-\mu(s_0)-\mu(s_\oplus)-\mu(s_\ominus)-\sum_j \mu(s^j).
\end{aligned}
$$

As long as $C\geq 2$, $\mu(s_{-1})$ defined above is positive. Also, for any θ, it holds that

$$
\max_{s,a}\frac{\nu^{\pi^{*,\theta}}(s,a)}{\mu(s,a)} \leq \frac{C}{1-\gamma}, \quad \sum_{s,a}\frac{\nu^{\pi^{*,\theta}}(s,a)}{\mu(s,a)} \leq \frac{(|\mathcal{S}|+1)C}{1-\gamma}.
$$

Lower bound Still, we take a subset Θ of $\{0,1\}^S$ such that $|\Theta|\geq\exp(S/8)$, and for any pair of different $\theta,\theta'\in\Theta$ it holds $\left\|\theta-\theta'\right\|_1\geq\frac{S}{4}$. We next consider the family of CMDPs $\mathfrak{M} := \{\mathcal{M}_\theta\}_{\theta\in\Theta}$, with the reference $\mu_\theta = \mu\otimes\mathbb{P}_\theta$.

By Fano's inequality (Lemma F.3), it holds that

$$
\max_{\theta\in\Theta}\mathbb{E}_\theta\left\|\hat{\pi}_b-\theta\right\|_1 \geq \frac{S}{4}\left(1-\frac{N\max_{\theta,\theta'\in\Theta}\mathrm{KL}(\mu_\theta\parallel\mu_{\theta'})+\log 2}{\log|\Theta|}\right).
$$

We also have $\max_{\theta,\theta'\in\Theta}\mathrm{KL}(\mu_\theta\parallel\mu_{\theta'}) \leq \frac{2\varpi^2(1-\gamma)}{C}$ by a simple calculation. Therefore, taking $\varpi=\min\left\{\sqrt{\frac{(S-3)C}{16(1-\gamma)N}},\frac{1}{2}\right\}$ is enough to ensure $\max_{\theta\in\Theta}\mathbb{E}_\theta\left\|\hat{\pi}_b-\theta\right\|_1\geq\frac{S}{24}$. Hence, we obtain

$$
\max_{\theta\in\Theta}\mathbb{E}_\theta\left[J^u(\hat{\pi};\theta)\right]_- = \max_{\theta\in\Theta}\mathbb{E}_\theta\left[\frac{v\varpi}{S(1-\gamma)}\left\|\hat{\pi}_b-\theta\right\|_1\right] \geq \frac{v\varpi}{24(1-\gamma)} \gtrsim \min\left\{\sqrt{\frac{SC}{(1-\gamma)^3N}},\frac{1}{1-\gamma}\right\}.
$$

G The Adaptive-DPDL framework

G.1 The verification method

First, let us provide the details of the VERIFY$(\cdot)$ method that is used in Algorithm 2.

As a remark, $\widehat{\Delta}_p$ is an estimator of the residual $A^\top W\overline{x}-\rho_0$, where A is defined in (29), that is $\mathbb{E}_{\mathcal{D}}\big[\widehat{\Delta}_p\big] = A^\top W\overline{x}-\rho_0$. By a direct computation, we also know $\mathbb{E}_{\mathcal{D}}\big[\widehat{J}(\overline{\pi})\big] = r^\top W\overline{x}$ and $\mathbb{E}_{\mathcal{D}}\big[\widehat{J}^{u^\kappa}(\pi)\big] = U_\kappa W\overline{x}$. Intuitively, when $\|\widehat{\Delta}_p\|_1$ is small, then $W\overline{x}$ is a good approximation of $\nu^{\overline{\pi}}$ and thus $\widehat{J}(\overline{\pi}),\widehat{J}^{u^\kappa}(\overline{\pi})$ are good approximations of $J(\overline{\pi}),J^{u^\kappa}(\overline{\pi})$. With this in mind, we present the following proposition that characterizes the VERIFY method, whose proof is moved to Appendix G.3.

Algorithm 3: VERIFY$(\overline{x})$ = VERIFY$(\overline{x}; \epsilon, \delta)$

```
input : The output $\overline{x}$ and the parameters $\epsilon, \delta > 0$ in Algorithm 1.
Obtain $N_v$ offline samples $\big\{(s_t, a_t, s_t', r_t, \mathbf{u}_t)\big\}_{t=1}^{N_v}$ from $\mathcal{D}$;
Compute the estimators $\widehat{J}(\overline{\pi}), \widehat{J}^{u^\kappa}(\overline{\pi}) \in \mathbb{R}$ and $\widehat{\Delta}_p \in \mathbb{R}^{|\mathcal{S}|}$ as
    $\widehat{J}(\overline{\pi}) := \frac{1}{N_v}\sum_{t=1}^{N_v} r_t \frac{\overline{x}(s_t,a_t)}{\hat{\mu}(s_t,a_t)}, \qquad \widehat{J}^{u^\kappa}(\overline{\pi}) := \frac{1}{N_v}\sum_{t=1}^{N_v} \mathbf{u}_t^\kappa \frac{\overline{x}(s_t,a_t)}{\hat{\mu}(s_t,a_t)},$
    $\widehat{\Delta}_p(s') := \sum_a \frac{N(s',a)}{N_v}\frac{\overline{x}(s',a)}{\hat{\mu}(s',a)} - \gamma\sum_{s,a}\frac{N(s,a,s')}{N_v}\frac{\overline{x}(s,a)}{\hat{\mu}(s,a)} - \rho_0(s'), \quad \forall s' \in \mathcal{S},$
    where $N(s,a), N(s,a,s')$ are the times that $(s,a)$ and $(s,a,s')$ are observed in the $N_v$ samples.
if $\big\|\widehat{\Delta}_p\big\|_1 \le \frac{3}{2}\varphi(1-\gamma)\epsilon$ && $\big\|\widehat{J}^{u^\kappa}(\overline{\pi})\big\|_\infty \le 3\varphi\epsilon$ then
    Return VERIFY$(\overline{x})$ = TRUE, and return $\widehat{J}(\overline{\pi})$ as an estimate of $J(\overline{\pi})$;
else
    Return VERIFY$(x)$ = FALSE;
```

Proposition G.1. *For the VERIFY method, if we choose* $N_v \ge \frac{64|\mathcal{S}|\psi\ell}{\varphi^2(1-\gamma)^4\epsilon_{\rm ver}^2}$*, with* $\ell = 4\log\left(\frac{40|\mathcal{S}|I}{\delta}\right)$ *and* $\epsilon_{\rm ver} = \frac{\epsilon}{10}$*, then with probability at least* $1-\delta$*, it holds that:*

(1). If VERIFY$(\overline{x})$ *= FALSE, then* $\psi < C^*$*.*

(2). If VERIFY$(\overline{x})$ *= TRUE, then* $J^{u^\kappa}(\overline{\pi}) \ge 0$*, and* $j_0(\psi) - 400\epsilon \le \widehat{J}(\overline{\pi}) \le j_0(\psi) + 100\epsilon$*.*

Basically, this proposition states that if VERIFY$(\overline{x})$ = FALSE, then we know $\psi < C^*$ with high probability. If VERIFY$(\overline{x})$ = TRUE, then we know that $\overline{\pi}$ is safe, and $j_0(\psi) = \widehat{J}(\overline{\pi}) + \mathcal{O}(\epsilon)$. We can apply Lemma 6.1 to determine whether the current policy is good enough.

G.2 The adaptive-DPDL method

In this section we will discuss the details of Algorithm 2. The key to the analysis of this section is Lemma 6.1, whose proof is presented in Appendix G.4.

Setting of sub-routine We use ϵ' for the input sub-optimality of Adaptive-DPDL. At each step K, we call DPDL and VERIFY with $\epsilon = \frac{\epsilon'}{15}$ and $\delta_K := \frac{6\delta}{\pi^2 K^2}$. The δ_K is chosen so that $\sum_K \delta_K = \delta$.

Exit condition In Algorithm 2, line 4 to 6, we write the exit condition as $-\infty < J^K \le J^{K-1} + \mathcal{O}(\epsilon)$. More specifically, the exit condition can be equivalently stated as

$$\text{VERIFY}\big(x^{(K)}\big) \ \&\& \ \text{VERIFY}\big(x^{(K-1)}\big) \ \&\& \ \widehat{J}\big(\pi^{(K)}\big) - \widehat{J}\big(\pi^{(K-1)}\big) \le 500\epsilon. \tag{46}$$

Here the third condition only needs to be checked when both VERIFY$\big(x^{(K)}\big)$ and VERIFY$\big(x^{(K-1)}\big)$ return TRUE. The constant 500 is chosen to ensure that Adaptive-DPDL will exit for $\psi_K > 2C^*$, as will be demonstrated in the following proposition, whose proof is presented in Appendix G.5.

Proposition G.2. *Suppose Algorithm 2 exits at step* K*. Then with probability at least* $1-\delta$*, the following results hold. (1)* $\pi^{(K)}$ *is safe and* $\psi_K \le 4C^*$*. (2) It holds that* $J^* - J(\pi^{(K)}) \le \mathcal{O}\big(\frac{C^*}{\psi_K}\epsilon\big)$*. (3) There is a constant* $\epsilon_0(\mathcal{M})$ *such that for* $\epsilon' \le \epsilon_0(\mathcal{M})$*,* $\psi_K \ge C^*$*.*

As a remark, ϵ_0 is (up to a scalar factor) the minimum performance improvement by increasing $\psi \to 2\psi$, and the minimum of slope of j as a function of $\log\psi$ for $\psi \in [1, C^*]$. Therefore, when Adaptive-DPDL exits at some step K, the improvement that can be achieved by increase ψ grows as at most $\frac{\epsilon_0}{\psi_K}$. If in this case ψ_K is still far small from C^*, then the difficulty essentially comes from a prohibitively large C^*.

Sample complexity of Adaptive-DPDL At step K, the samples needed for DPDL are $\tilde{\mathcal{O}}\left(\frac{\mathcal{N}\psi_K}{\varphi^2(1-\gamma)^4\epsilon^2}\right)$, and the samples needed for verification are $\tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|\psi_K}{(1-\gamma)^4\epsilon^2}\right)$. There are at most

$\lceil \log_2(C^*/\psi^1) \rceil + 1$ outer steps and the ψ_K is twofold at each step, thus the total samples needed are $\tilde{\mathcal{O}}\left(\frac{\mathcal{N}\psi_K}{\varphi^2(1-\gamma)^4\epsilon^2}\right)$ if it exits at step K. Especially, as long as $\epsilon \le \epsilon_0(\mathcal{M})$, Adaptive-DPDL ends after consuming $\tilde{\mathcal{O}}\left(\frac{\mathcal{N}C^*}{\varphi^2(1-\gamma)^4\epsilon^2}\right)$ samples and outputs a policy which is safe and $\mathcal{O}(\epsilon)$-optimal.

G.3 Proof of Proposition G.1

Proof. First, we provide the following lemma for the estimators $\widehat{\Delta}_p$, $\widehat{J}(\overline{\pi})$ and $\widehat{J}^{u^\kappa}(\overline{\pi})$. The calculation of Lemma G.3 is very closed to Appendix B, and is thus omitted.

Lemma G.3. *Suppose that N_v and ϵ_{ver} are chosen according to Proposition G.1. Denote $\overline{\nu} = W\overline{x}$, then with probability at least $1 - \delta/3$, we have*

$$\max\left\{ \left\|\widehat{\Delta}_p - (A^\top\overline{\nu} - \rho_0)\right\|_1, \left|\widehat{J}(\overline{\pi}) - \langle r, \overline{\nu}\rangle\right|, \left\|\widehat{J}^{u^\kappa}(\overline{\pi}) - U_\kappa\overline{\nu}\right\|_\infty \right\} \le \varphi(1-\gamma)\epsilon_{ver}.$$

Proof of the case VERIFY$(\overline{x})$ = FALSE. By Corollary E.5, it holds that when $\psi \ge C^*$,

$$\|A^\top\overline{\nu} - \rho_0\|_1 \le \frac{11}{8}\varphi(1-\gamma)\epsilon, \quad \text{and} \quad \|\,[U_\kappa\overline{\nu}]_-\,\|_\infty \le \frac{11}{4}\varphi\epsilon.$$

Combining the above inequality with Lemma G.3 indicates that $\|\widehat{\Delta}_p\|_1 \le \frac{3}{2}\varphi(1-\gamma)\epsilon$ and $\|[\widehat{J}^{u^\kappa}(\overline{\pi})]_-\|_\infty \le 3\varphi\epsilon$. This contradicts the condition for returning FALSE. Therefore, we know that $\psi < C^*$.

Proof of the case VERIFY$(\overline{x})$ = TRUE. By the condition for returning TRUE, we know $\|\widehat{\Delta}_p\|_1 \le \frac{3}{2}\varphi(1-\gamma)\epsilon$ and $\|[\widehat{J}^{u^\kappa}(\overline{\pi})]_-\|_\infty \le 3\varphi\epsilon$. Together with Lemma G.3, we have

$$\|A^\top\overline{\nu} - \rho_0\|_1 \le 1.6\varphi(1-\gamma)\epsilon \qquad \text{and} \qquad \left\|[U_\kappa\overline{\nu}]_-\right\|_\infty \le 3.1\varphi\epsilon.$$

Similar to our analysis in Appendix E, we write $\nu^{\overline{\pi}}$ the true visitation measure of $\overline{\pi}$. Then by (34),

$$\left\|\nu^{\overline{\nu}} - \overline{\nu}\right\|_1 \le \frac{1}{1-\gamma}\left(\left\|A^\top\overline{\nu} - \rho_0\right\|_1 + 3\left\|\overline{x} - \overline{\nu}\right\|_1\right) \le 1.63\varphi\epsilon.$$

where the term $\|\overline{x} - \overline{\nu}\|_1$ is controlled by Proposition 4.3. Due to the fact that $\left|\left\|\,\left[U_\kappa\nu^{\overline{\pi}}\right]_-\,\right\|_\infty - \left\|\,[U_\kappa\overline{\nu}]_-\,\right\|_\infty\right| \le (1+5(1-\gamma)\varphi\epsilon)\left\|\nu^{\overline{\pi}} - \overline{\nu}\right\|_1 \le 1.1\left\|\nu^{\overline{\pi}} - \overline{\nu}\right\|_1$ for small $\epsilon \le \frac{1}{50(1-\gamma)}$, it holds that

$$\min_i J_i^u(\overline{\pi}) \ge \kappa - \left\|\,\left[U_\kappa\nu^{\overline{\pi}}\right]_-\,\right\|_\infty \ge \kappa - \left\|[U_\kappa\overline{\nu}]_-\right\|_\infty - 1.1\left\|\nu^{\overline{\pi}} - \overline{\nu}\right\|_1 \ge 0. \tag{47}$$

Moreover, the definition of $\mathrm{Gap}(\overline{x})$ gives

$$j(\psi) - \mathrm{Gap}(\overline{x}) = \langle r, \overline{\nu}\rangle - R_{\mathcal{V}}\left\|A^\top\overline{\nu} - \rho_0\right\|_1 - R_\Lambda\left\|[U_\kappa\overline{\nu}]_-\right\|_\infty \le j(\psi) \le j_0(\psi),$$

which yields

$$\langle r, \overline{\nu}\rangle \le j_0(\psi) + R_{\mathcal{V}}\left\|A^\top\overline{\nu} - \rho_0\right\|_1 + R_\Lambda\left\|[U_\kappa\overline{\nu}]_-\right\|_\infty \le j_0(\psi) + 100\epsilon,$$
$$\langle r, \overline{\nu}\rangle \ge j(\psi) - \mathrm{Gap}(\overline{x}) \ge j_0(\psi) - 400\epsilon,$$

where we use the fact that $0 \le j_0(\psi) - j_\kappa(\psi) \le \frac{64\kappa}{\varphi} = 320\epsilon$. The same bound for $\widehat{J}(\overline{\pi})$ can be derived by Lemma G.3. □

G.4 Proof of Lemma 6.1

Proof. First we show that, when $\psi < C^*$, $j_0(\psi) < J^*$. Otherwise, for $x_* = \arg\max_{x\in\mathcal{X}} \mathcal{J}_0(x)$, it holds that $\mathcal{J}_0(x_*) = j_0(\psi) \ge J^*$, i.e., for $\nu_* = Wx_*$,

$$J^* - \langle r, \nu_*\rangle + R_{\mathcal{V}}\left\|A^\top\nu_* - \rho_0\right\|_1 + R_\Lambda\left\|[U\nu_*]_-\right\|_\infty \le 0.$$

Applying Lemma E.3 gives $\left\|A^\top\nu_* - \rho_0\right\|_1 \le 0$, $\left\|[U\nu_*]_-\right\|_\infty \le 0$, $J^* - \langle r, \nu_*\rangle \le 0$. Thus, $\nu \in \mathfrak{V} \cap \mathfrak{S}$, and $\langle r, \nu_*\rangle \ge J^*$, which imply that ν_* is indeed an optimal solution of problem (4). However, $\nu_* \in W\mathcal{X} \Rightarrow \nu_* \in \mathfrak{V}(\psi) \Rightarrow \psi \ge C^*$, a contradiction.

Now the monotonicity is easy. We still fix an optimal $\nu_* \in \mathfrak{V}(C^*)$ and let $x_* := W^{-1}\nu_*$. For $1 \le \psi' < \psi$, we write $x_{\psi'} = \arg\max_{x\in\mathcal{X}(\psi')} \mathcal{J}_0(x)$, $c = \frac{\psi-\psi'}{C^*-\psi'}$, and we consider $x_\psi := cx_* + (1-c)x_{\psi'} \in \mathcal{X}(\psi)$. It holds that

$$j_0(\psi) \ge \mathcal{J}_0(x_\psi) \ge (1-c)\mathcal{J}_0(x_{\psi'}) + c\mathcal{J}_0(x_*) = j_0(\psi') + c(J^* - j_0(\psi)).$$

The proof is completed by reorganizing the above inequality. □

G.5 Proof of Proposition G.2

Proof. By Proposition G.1, if $\psi_K \geq 2C^*$, then $\psi_{K-1} \geq C^*$. By Proposition G.1, with probability at least $1-\delta$ it holds that $\text{VERIFY}\big(x^{(K)}\big) = \text{VERIFY}\big(x^{(K-1)}\big) = \text{TRUE}$, and

$$\begin{aligned} &\widehat{J}(\overline{\pi}^{(K)}), \widehat{J}(\overline{\pi}^{(K-1)}) \in [J(\pi^*) - 400\epsilon, J(\pi^*) + 100\epsilon], \\ \Rightarrow &\left|\widehat{J}(\overline{\pi}^{(K)}) - \widehat{J}(\overline{\pi}^{(K-1)})\right| \leq 500\epsilon, \end{aligned}$$

where we use the fact $j_0(\psi_K) = j_0(\psi_{K-1}) = j_0(C^*) = J(\pi^*)$ from Lemma 6.1. Therefore, if $\psi_K \geq 2C^*$, Adaptive-DPDL must exit at step K.

Now, we only need to consider the case that Adaptive-DPDL ends at some step K, but ψ_K might not be greater than C^*. Because $\text{VERIFY}\big(x^{(K)}\big) = \text{VERIFY}\big(x^{(K-1)}\big) = \text{TRUE}$, we combine the exit condition (46) with Proposition G.1 and derive $|j_0(\psi_K) - j_0(\psi_{K-1})| \leq 1000\epsilon$. Then by Lemma 6.1, we have

$$J(\pi^*) - j_0(\psi_K) \leq \frac{2(C^* - \psi_K)}{\psi_K}\left(j_0(\psi_K) - j_0(\psi_{K-1})\right).$$

Thus $J(\pi^*) - J(\pi^{(K)}) \lesssim \frac{C^*}{\psi_K}\epsilon$. Furthermore, we can define the following quantity

$$\epsilon_0 := \min_{1 \leq \psi \leq C^*}\left(j_0(\psi) - j_0(\frac{\psi}{2})\right) > 0.$$

Here $\epsilon_0 > 0$ is due to Lemma 6.1. If $j_0(\psi) - j(\frac{\psi}{2}) < \epsilon_0$ for some $\psi \geq 1$, then immediately we have $\psi \geq C^*$. If $\epsilon' = 15\epsilon \leq \epsilon_0/100 =: \epsilon_0(\mathcal{M})$, Adaptive-DPDL must exit at step K with $C^* \leq \psi_K \leq 4C^*$. By Theorem 4.1, the output policy $\pi^{(K)}$ is safe and $J(\pi^*) - J(\pi^{(K)}) \leq \epsilon'$. □

H Convergence Analysis in Asynchronous Setting

H.1 Mixing property of Markov chain

Under the setting of the asynchronous learning (Assumption 4.5), we can observe a sequence of state-action trajectory generated under the behavioral policy π_b, namely

$$s_1, a_1, s_2, a_2, s_3, \cdots, s_n, a_n, s_{n+1}, \cdots.$$

This sequence can be naturally viewed as a Markov chain $(X_t)_{t \geq 1}$ where $X_t = (s_t)$, plus a marginal component $a_t \in \mathcal{A}$. In the asynchronous setting, the reference distribution μ is the stationary distribution μ_{π_b} of this chain product with the policy π_b. As in the synchronous setting, we denote $\mathcal{F}_t$ for all the history information at time t. Actually, by the Markov property and our update rule, conditioning on $\mathcal{F}_t$ is equivalent to conditioning on s_t, Z^t. According to [13, Section 4], we define the mixing time of this Markov chain as

$$\begin{cases} \mathcal{E}(t) := \sup_{s \in \mathcal{S}} d_{\text{TV}}\left(\mu_{\pi_b}, \mathbb{P}^t_{\pi_b}(\cdot|s_0 = s)\right), \\ t_{\text{mix}} := \min\{t : \mathcal{E}(t) \leq \frac{1}{4}\}, \end{cases} \tag{48}$$

where $\mathbb{P}^t_{\pi_b}(\cdot|s_0 = s)$ denotes the distribution of s_t given $s_0 = s$ and policy π_b. By [13, Remark 4.12], it holds that

$$\mathcal{E}(t) \leq 2^{-\left\lfloor \frac{t}{t_{\text{mix}}} \right\rfloor}.$$

Given the concept of the mixing time, we modify the standard Bernstein inequality for Markov chain [10, 19, etc.] to cover the non-stationary Markov chains.

Proposition H.1. *Suppose that $(X_t)_{t \geq 1}$ is a Markov chain with invariant distribution π and mixing time $t_{\text{mix}} < +\infty$. Let f be a measurable function such that $\mathbb{E}_\pi[f(X)] = 0$, $|f(X)| \leq M$. Denote $\sigma^2 = \mathbb{E}_\pi\left[f(X)^2\right]$, then for $\delta \in (0,1)$, the following holds with probability at least $1-\delta$*

$$\left|\sum_{t=1}^{n} f(X_t)\right| \leq \sqrt{32 t_{\text{mix}} n \sigma^2 \log\frac{4}{\delta}} + 82 t_{\text{mix}} M \log\frac{4}{\delta}.$$

The difficulty of analyzing Markovian gradients is the correlation between updates and samples. As demonstrated in Section 4.3, in our analysis, we leverage the fact that $s_{t+\tau}$ is a sample "almost" from μ_{π_b} and "almost" independent of s_t, as long as $\tau \geq t_{\mathrm{mix}} \cdot \log$ factor. We further demonstrate this idea in the following proposition, by comparing $\mathbb{E}\left[\widehat{g}(\cdot;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right]$ and $\mathcal{G}(Z^t)$.

Proposition H.2 (Almost unbiased)**.** *For a $\mathcal{F}_t$-measurable random variable $Z \in \mathcal{Z} := \mathcal{V} \times \Lambda \times \mathcal{X}$, it holds that*

$$\left\|\mathbb{E}\left[\widehat{g}_V(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right] - \nabla_V \mathcal{L}_w(Z)\right\|_1 \leq \frac{2\psi}{1-\gamma}\mathcal{E}(\tau),$$

$$\left\|\mathbb{E}\left[\widehat{g}_\lambda(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right] - \nabla_\lambda \mathcal{L}_w(Z)\right\|_\infty \leq \frac{2\psi}{1-\gamma}\mathcal{E}(\tau),$$

$$\left\|\mathbb{E}\left[\widehat{g}_x(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right] - \nabla_x \mathcal{L}_w(Z)\right\|_\infty \leq \frac{64}{\varphi(1-\gamma)\varsigma}\mathcal{E}(\tau).$$

Furthermore, for any $Z' \in \mathcal{Z}$, we have

$$\left|\left\langle Z', \mathcal{G}(Z) - \mathbb{E}\left[\widehat{g}(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right]\right\rangle\right| \leq \frac{128\psi}{\varphi(1-\gamma)^2}\mathcal{E}(\tau).$$

The following proposition indicates that, the estimator $\widehat{g}(\cdot;\zeta_{t+\tau})$ is not only "nearly unbiased" conditional on $\mathcal{F}_t$, but it also has a well bounded moment.

Proposition H.3 (Bounded moment)**.** *For any $\mathcal{F}_t$-measurable random variable $Z \in \mathcal{Z}$, it holds that*

$$\mathbb{E}\left[\left\|\widehat{g}_V(Z;\zeta_{t+\tau})\right\|\middle| \mathcal{F}_t\right] \lesssim \frac{C(\tau)}{1-\gamma},$$

$$\mathbb{E}\left[\left\|\widehat{g}_\lambda(Z;\zeta_{t+\tau})\right\|_\infty \middle| \mathcal{F}_t\right] \lesssim \frac{C(\tau)}{1-\gamma},$$

$$\mathbb{E}\left[\left\|\widehat{g}_x(Z;\zeta_{t+\tau})\right\|_{x^t}^2 \middle| \mathcal{F}_t\right] \lesssim \frac{C(\tau)\mathcal{N}\psi}{\varphi^2(1-\gamma)^3},$$

where $C(\tau) = 2 + \frac{\mathcal{E}(\tau)}{\varsigma}$.

Therefore, there is a universal constant c_τ such that for $\tau \geq \lfloor c_\tau t_{\mathrm{mix}} \iota \rfloor$, we have $C(\tau) \leq 3$ and $\mathcal{E}(\tau) \leq \frac{1}{T}$ (the log factor ι and the range of T are specified in Theorem H.5). We denote $\tau_0 = \lfloor c_\tau t_{\mathrm{mix}} \iota \rfloor$.

H.2 Proof sketch of Theorem 4.6

Before our analysis of DPDL on $\mathcal{D}_{async}$, we have to first provide an analogue of Proposition 4.3. As in the synchronous setting, we set $\epsilon_e = \frac{\epsilon}{100}$ and $\varsigma = \frac{\varphi(1-\gamma)^2\epsilon_e}{2\mathcal{N}\psi}$.

Proposition H.4. *Given $N_e \geq c_e' \frac{t_{\mathrm{mix}}\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4\epsilon_e^2}$ samples from a trajectory generated by π_b, the $\hat{\mu}$ constructed in* (10) *satisfies the following properties with probability at least $1 - \delta/3$.*
(1) For all s, a, $\frac{\mu(s,a)}{\hat{\mu}(s,a)} \leq 2$, and $\hat{\mu}(s,a) \geq \varsigma$.
(2) For any $\pi \in \Pi(\psi)$, $W^{-1}\nu^\pi \in \mathcal{X}$.
(3) For any $x \in \mathcal{X}$, $\|Wx - x\|_1 \leq \varphi(1-\gamma)\epsilon_e$.

Now, we present the convergence guarantee of the duality gap $\mathrm{Gap}(\overline{x})$.

Theorem H.5. *Given $\epsilon \in \left(0, \frac{1}{1-\gamma}\right]$, $\delta \in \left(0, \frac{1}{2}\right)$, we denote $\iota = \log\left(T|\mathcal{S}||\mathcal{A}|I/\delta\right)$. Then as long as $T \gtrsim \frac{\tau_0^2\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4\epsilon_e^2}$, with probability at least $1 - \delta/3$ it holds*

$$\mathrm{Gap}(\overline{x}) \lesssim \frac{t_{\mathrm{mix}}}{\varphi(1-\gamma)^2}\sqrt{\frac{\mathcal{N}\psi\iota^3}{T}} \leq \epsilon.$$

Therefore, there is a universal constant c_o' such that $\mathrm{Gap}(\overline{x}) \leq \frac{\epsilon}{2}$ as long as $T \geq c_o' \frac{t_{\mathrm{mix}}^2\mathcal{N}\psi\iota^3}{\varphi^2(1-\gamma)^4\epsilon^2}$. Then the proof in Appendix E can be applied directly. In conclusion, the number of samples needed is

$$\tilde{\mathcal{O}}\left(\frac{t_{\mathrm{mix}}^2\mathcal{N}\psi}{\varphi^2(1-\gamma)^4\epsilon^2}\right).$$

We sketch the proof of Theorem H.5 as follows. The detailed proofs of propositions are organized by order in the rest of this section.

Decomposition of duality gap We define the auxiliary variables V', λ', x' as in Appendix D,

$$(V', \lambda') = \underset{V \in \mathcal{V}, \lambda \in \Lambda}{\arg\min}\, \mathcal{L}_w(V, \lambda, \overline{x}), \quad x' = \underset{x \in \mathcal{X}}{\arg\max} \min_{V \in \mathcal{V}, \lambda \in \Lambda} \mathcal{L}_w(V, \lambda, x), \quad Z' = [V'; \lambda'; x'].$$

Recall the decomposition (24), we have

$$\mathrm{Gap}(\overline{x}) = \underbrace{\frac{1}{T}\sum_{t=1}^{T} \left\langle \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle}_{S_1} + \underbrace{\frac{1}{T}\sum_{t=1}^{T} \left\langle \mathcal{G}(Z^t) - \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle}_{S_2}.$$

Bounding the term S_1 The proof in Appendix D.2 can be applied without change. Namely, as long as $\eta \leq \frac{1}{2}\min\left(\frac{\alpha_\lambda}{M_\lambda}, \frac{\alpha_x}{M_{x,\infty}}\right)$, it holds that

$$S_1 \lesssim \frac{\alpha_V D_V^2 + \alpha_\lambda D_\lambda + \alpha_x D_x}{\eta T} + \frac{\eta}{T}\sum_{t=1}^{T}\left(\frac{\|\widehat{g}_V(Z^t;\zeta_t)\|^2}{\alpha_V} + \frac{D_{\lambda,1}\|\widehat{g}_\lambda(Z^t;\zeta_t)\|_\infty^2}{\alpha_\lambda} + \frac{\|\widehat{g}_x(Z^t;\zeta_t)\|_{x^t}^2}{\alpha_x}\right).$$

Bounding the term S_2 In the asynchronous setting, $\zeta_1, \cdots, \zeta_T$ are no longer i.i.d samples. To deal with this issue, let us consider the following decomposition

$$\begin{aligned}
\Gamma^t := \left\langle \mathcal{G}(Z^t) - \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle = & \underbrace{\left\langle \mathcal{G}(Z^t), Z^t - Z' \right\rangle - \left\langle \mathcal{G}(Z^{t-\tau}), Z^{t-\tau} - Z' \right\rangle}_{\Gamma_1^t} \\
& + \underbrace{\left\langle \mathcal{G}(Z^{t-\tau}) - \mathbb{E}\left[\widehat{g}(Z^{t-\tau};\zeta_t)\middle|\, \mathcal{F}_{t-\tau}\right], Z^{t-\tau} - Z' \right\rangle}_{\Gamma_2^{t-\tau}} \\
& + \underbrace{\left\langle \mathbb{E}\left[\widehat{g}(Z^{t-\tau};\zeta_t)\middle|\, \mathcal{F}_{t-\tau}\right] - \widehat{g}(Z^{t-\tau};\zeta_t), Z^{t-\tau} - Z' \right\rangle}_{\Gamma_3^{t-\tau}} \\
& + \underbrace{\left\langle \widehat{g}(Z^{t-\tau};\zeta_t), Z^{t-\tau} - Z' \right\rangle - \left\langle \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle}_{\Gamma_4^t},
\end{aligned}$$

where $1 \leq \tau \leq \tau_0$ is a fixed integer. The quantity $\Gamma_2^{t-\tau}$ can be bounded by Proposition H.2, and $\Gamma_3^{t-\tau}$ can be bounded as in Appendix D.3. As of Γ_1^t, Γ_4^t, we bound it in terms of $Z^t - Z^{t-\tau}$. In conclusion, with probability at least $1 - \delta/5$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\Gamma^t \lesssim & \frac{\psi}{\varphi(1-\gamma)^2}\mathcal{E}(\tau) + \frac{1}{\varphi(1-\gamma)^2}\sqrt{\frac{\tau C(\tau)\mathcal{N}\psi\iota}{T}} + \frac{1}{\varphi(1-\gamma)}\sum_{t=\tau+1}^{T}\frac{\left|x^t - x^{t-\tau}\right|(s_t,a_t)}{\hat{\mu}(s_t,a_t)} \\
& + \frac{\eta}{T}\sum_{t=\tau+1}^{T}\left(\frac{\|\widehat{g}_V(Z^t;\zeta_t)\|^2}{\alpha_V} + \frac{D_{\lambda,1}\|\widehat{g}_\lambda(Z^t;\zeta_t)\|_\infty^2}{\alpha_\lambda}\right).
\end{aligned} \tag{49}$$

The detailed analysis is presented in Appendix H.7.

Bounding the variance and magnitude of the updates It remains to bound $\|\widehat{g}_V(Z^t;\zeta_t)\|$, $\|\widehat{g}_\lambda(Z^t;\zeta_t)\|_\infty$, $\|\widehat{g}_x(Z^t;\zeta_t)\|_{x^t}$, and the term $|x^t(s_t,a_t) - x^{t-\tau}(s_t,a_t)|$. For any $x \in \mathbb{R}_{\geq 0}^{|\mathcal{S}||\mathcal{A}|}$, and any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we introduce the following abbreviation for the ease of notation

$$p(x;s,a) := \frac{x(s,a)}{\hat{\mu}(s,a)}, \quad q(x;s,a) := \frac{x(s,a)}{\hat{\mu}(s,a)^2}.$$

For any sample $\zeta = (s_0, s, a, s', r, \mathbf{u})$, we also reload the notations p, q as $p(x;\zeta) := p(x;s,a)$ and $q(x;\zeta) := q(x;s,a)$.

It is not hard to see that $p(x^t;\zeta_t)$ and $q(x^t;\zeta_t)$ dominate the variance of the gradient estimators (for detailed discussion, see Appendix H.5). More specifically, we have

$$\begin{aligned}
& \left\|\widehat{g}_V(Z^t;\zeta_t)\right\| \lesssim p(x^t;\zeta_t), \qquad \left\|\widehat{g}_\lambda(Z^t;\zeta_t)\right\|_\infty \lesssim p(x^t;\zeta_t), \\
& \left\|\widehat{g}_x(Z^t;\zeta_t)\right\|_{x^t} \lesssim \frac{1}{\varphi(1-\gamma)}\sqrt{q(x^t;\zeta_t)}.
\end{aligned}$$

Then, we only need to bound $\sum_{t=1}^{T} q(x^t;\zeta_t)$, $\sum_{t=\tau+1}^{T} p(\left|x^t - x^{t-\tau}\right|;\zeta_t)$ and $\sum_{t=1}^{T} p(x^t;\zeta_t)^2$. By leveraging the idea of the decomposition (13), we can derive the desired estimation, as follows.

Proposition H.6. *There is a universal constant c such that for $T \geq c\frac{\tau_0^2 \mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4\epsilon_e^2}$, the following holds for all $1 \leq \tau \leq \tau_0$ simultaneously, with probability at least $1 - \delta/10$:*

$$\frac{1}{T}\sum_{t=1}^{T} p(x^t;\zeta_t)^2 \lesssim \frac{\psi}{(1-\gamma)^2},$$
$$\frac{1}{T}\sum_{t=1}^{T} q(x^t;\zeta_t) \lesssim \frac{\mathcal{N}\psi}{1-\gamma},$$
$$\frac{1}{T}\sum_{t=\tau+1}^{T} p(\left|x^t - x^{t-\tau}\right|;\zeta_t) \lesssim \frac{\tau C(\tau)}{1-\gamma}\sqrt{\frac{\mathcal{N}\psi\iota}{T}}.$$

Conclusion Combining Proposition H.6 with the estimations of S_1 and S_2, we have with probability at least $1 - \delta/3$,

$$\mathrm{Gap}(\overline{x}) \lesssim \frac{\tau C(\tau)}{\varphi(1-\gamma)^2}\sqrt{\frac{\mathcal{N}\psi\iota}{T}} + \frac{\psi}{\varphi(1-\gamma)^2}\mathcal{E}(\tau). \tag{50}$$

Now, we can take $\tau = \tau_0 = \lfloor c_\tau t_{\mathrm{mix}}\iota \rfloor$. Then by the definition, it holds $C(\tau_0) \leq 3$ and $\epsilon(\tau_0) \leq \frac{1}{T}$, and hence with probability at least $1 - \delta/3$ we have

$$\mathrm{Gap}(\overline{x}) \lesssim \frac{t_{\mathrm{mix}}}{\varphi(1-\gamma)^2}\sqrt{\frac{\mathcal{N}\psi\iota^3}{T}}.$$

As a remark, if we have an (empirical) estimation $\hat{t}_{\mathrm{mix}}$ such that $\hat{t}_{\mathrm{mix}} \geq t_{\mathrm{mix}}$, then by taking $\eta = \frac{1}{\sqrt{\hat{t}_{\mathrm{mix}}T}}$, the final bound can be improved to $\mathrm{Gap}(\overline{x}) \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{\frac{\hat{t}_{\mathrm{mix}}\mathcal{N}\psi\iota^3}{T}}$, as long as $T \gtrsim \frac{t_{\mathrm{mix}}^2}{\hat{t}_{\mathrm{mix}}}\frac{\mathcal{N}\psi\iota^3}{\varphi^2(1-\gamma)^4\epsilon^2}$.

H.3 Proof of Proposition H.1

In order to prove Proposition H.1, we invoke the following standard version of the Bernstein's inequality. We also leverage the idea of the proof of [14, Lemma 8].

Theorem H.7 ([19, Theorem 3.9])**.** *Suppose $\{X_i\}_{i\geq 1}$ is a stationary Markov chain with invariant distribution π and pseudo spectral gap γ_{ps}. Let f be a measurable function such that $\mathbb{E}_\pi[f(X)] = 0$, $|f(X)| \leq M$. Denote $\sigma^2 = \mathbb{E}_\pi\left[f(X)^2\right]$, then for all $x \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} f(X_i)\right| \geq x\right) \leq 2\exp\left(-\frac{x^2 \cdot \gamma_{\mathrm{ps}}}{8\left(n + 1/\gamma_{\mathrm{ps}}\right)\sigma^2 + 20xM}\right).$$

In particular, for uniformly ergodic chains with mixing time t_{mix}, $\gamma_{\mathrm{ps}} \geq \frac{1}{2t_{\mathrm{mix}}}$.

Proof of Proposition H.1. Without loss of generality, we assume the Markov chain (X_t) has a finite state space $\mathcal{X}$. We fix integer τ and $x \geq 0$ to be specified later, and let π_n be the distribution of X_n. Theorem H.7 yields

$$\mathbb{P}\left(\left|\sum_{i=\tau+1}^{n} f(X_i)\right| \geq x \,\middle|\, X_1 \sim \pi\right) \leq 2\exp\left(-\frac{x^2}{16t_{\mathrm{mix}}\left(n + 2t_{\mathrm{mix}} - \tau\right)\sigma^2 + 40xt_{\mathrm{mix}}M}\right).$$

Let $\mathcal{B}_\tau$ be the event $\left\{\left|\sum_{i=\tau+1}^{n} f(X_i)\right| \geq x\right\}$, then

$$\begin{aligned}
&\left|\mathbb{P}\left(\mathcal{B}_\tau \middle| X_1 \sim \pi\right) - \mathbb{P}\left(\mathcal{B}_\tau \middle| X_1 \sim \pi_1\right)\right| \\
&= \left|\sum_{x\in\mathcal{X}} \mathbb{P}\left(\mathcal{B}_\tau \middle| X_{\tau+1} = x\right)\left(\mathbb{P}\left(X_{\tau+1} = x \middle| X_1 \sim \pi\right) - \mathbb{P}\left(X_{\tau+1} = x \middle| X_1 \sim \pi_1\right)\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&= \left|\sum_{x\in\mathcal{X}} \mathbb{P}\left(\mathcal{B}_\tau \middle| X_{\tau+1} = x\right)\left(\pi(x) - \pi_{\tau+1}(x)\right)\right| \\
&\le \max\left(\left\|[\pi - \pi_{\tau+1}]_+\right\|_1, \left\|[\pi - \pi_{\tau+1}]_-\right\|_1\right) \\
&= d_{\mathrm{TV}}(\pi, \pi_{\tau+1}) \le \mathcal{E}(\tau).
\end{aligned}
$$

Therefore, we can take $x = \sqrt{32 t_{\mathrm{mix}}(n - \tau + 2t_{\mathrm{mix}}) \log \frac{4}{\delta}} + 80 t_{\mathrm{mix}} M \log \frac{4}{\delta}$ and $\tau = \left\lceil \log_2 \frac{2}{\delta} \right\rceil t_{\mathrm{mix}}$, then

$$\mathbb{P}\left(\mathcal{B}_\tau \middle| X_1 \sim \pi_1\right) \le \mathcal{E}(\tau) + \mathbb{P}\left(\mathcal{B}_\tau \middle| X_1 \sim \pi\right) \le \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Hence with probability at least $1 - \delta$, it holds that

$$\left|\sum_{i=\tau+1}^{n} f(X_i)\right| \le \sqrt{32 t_{\mathrm{mix}}(n - \tau + 2t_{\mathrm{mix}}) \log \frac{4}{\delta}} + 80 t_{\mathrm{mix}} M \log \frac{4}{\delta}.$$

The proof is completed by noticing that $|\sum_{i=1}^{\tau} f(X_i)| \le \tau M \le 2t_{\mathrm{mix}} M \log \frac{4}{\delta}$ and $\tau \ge 2t_{\mathrm{mix}}$. □

H.4 Proof of Proposition H.2

Proof. Recall that the gradient estimators are constructed as

$$
\begin{aligned}
\widehat{g}_V(Z;\zeta) &:= \mathbb{I}_{s_0} + \frac{x(s,a)}{\hat{\mu}(s,a)}\left(\gamma \mathbb{I}_{s'} - \mathbb{I}_s\right), \\
\widehat{g}_\lambda(Z;\zeta) &:= \frac{x(s,a)}{\hat{\mu}(s,a)} \mathbf{u}^\kappa, \\
\widehat{g}_x(Z;\zeta) &:= \frac{r + \gamma V(s) - V(s') + \langle \mathbf{u}^\kappa, \lambda\rangle}{\hat{\mu}(s,a)} \mathbb{I}_{s,a}.
\end{aligned}
$$

Therefore, for $Z = [V; \lambda; x]$ that is $\mathcal{F}_t$ measurable, we have

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{g}_x(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right] \\
=&\mathbb{E}\left[\frac{r_{t+\tau} + \gamma V(s_{t+\tau}) - V(s_{t+\tau+1}) + \left\langle \mathbf{u}^\kappa_{t+\tau}, \lambda\right\rangle}{\hat{\mu}(s_{t+\tau}, a_{t+\tau})} \mathbb{I}_{s_{t+\tau}, a_{t+\tau}} \middle| s_t, Z\right] \\
=&\mathbb{E}\left[\frac{r(s_{t+\tau}, a_{t+\tau}) + \gamma V(s_{t+\tau}) - V(s_{t+\tau+1}) + \left\langle \mathbf{u}^\kappa(s_{t+\tau}, a_{t+\tau}), \lambda\right\rangle}{\hat{\mu}(s_{t+\tau}, a_{t+\tau})} \mathbb{I}_{s_{t+\tau}, a_{t+\tau}} \middle| s_t, Z\right] \\
=&\sum_{s,a,s'} \mathbb{P}\left(s_{t+\tau} = s, a_{t+\tau} = a, s_{t+\tau+1} = s' \middle| s_t\right) \frac{r(s,a) + \gamma V(s) - V(s') + \langle \mathbf{u}^\kappa(s,a), \lambda\rangle}{\hat{\mu}(s,a)} \mathbb{I}_{s,a} \\
=&\sum_{s,a} \frac{\mathbb{P}\left(s_{t+\tau} = s, a_{t+\tau} = a \middle| s_t\right)}{\hat{\mu}(s,a)} \left(r(s,a) + \gamma V(s) - \mathbb{E}_{s'|s,a}\left[V(s')\right] + \langle \mathbf{u}^\kappa(s,a), \lambda\rangle\right) \mathbb{I}_{s,a}.
\end{aligned}
$$

For the sake of simplicity, we denote

$$W^{\tau, s_t} := \mathrm{diag}\left(\frac{\mathbb{P}_{\pi_b}\left(s_{t+\tau} = s, a_{t+\tau} = a \middle| s_t\right)}{\hat{\mu}(s,a)}\right) = \mathrm{diag}\left(\frac{\mathbb{P}_{\pi_b}\left(s_{t+\tau} = s \middle| s_t\right) \pi_b(a|s)}{\hat{\mu}(s,a)}\right)_{s,a}, \quad (51)$$

and we follow the matrix notation introduced in Appendix E.1. Then

$$
\begin{aligned}
\mathbb{E}\left[\widehat{g}_x(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right] &= W^{\tau, s_t}(r - AV + U_\kappa^{\mathrm{T}} \lambda), \\
\mathbb{E}\left[\widehat{g}_V(Z;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right] &= \mathbb{E}\left[\mathbb{I}_{s_0} + \frac{x(s_{t+\tau}, a_{t+\tau})}{\hat{\mu}(s_{t+\tau}, a_{t+\tau})}\left(\gamma \mathbb{I}_{s_{t+\tau+1}} - \mathbb{I}_{s_{t+\tau}}\right) \middle| s_t, Z\right] \\
&= \rho_0 + \sum_{s,a} \mathbb{P}_{\pi_b}\left(s_{t+\tau} = s, a_{t+\tau} = a \middle| s_t\right) \frac{x(s,a)}{\hat{\mu}(s,a)} \left(\gamma \mathbb{E}_{s'|s,a}\left[\mathbb{I}_{s'}\right] - \mathbb{I}_s\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \rho_0 - A^{\mathrm{T}} W^{\tau,s_t} x,\\
\mathbb{E}\left[\widehat{g}_\lambda(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] &= \mathbb{E}\left[\left.\frac{x(s_{t+\tau},a_{t+\tau})}{\hat{\mu}(s_{t+\tau},a_{t+\tau})}\mathbf{u}^\kappa\right| s_t, Z\right]\\
&= \sum_{s,a}\mathbb{P}_{\pi_b}\left(s_{t+\tau}=s,a_{t+\tau}=a\middle|\,s_t\right)\frac{x(s,a)}{\hat{\mu}(s,a)}\mathbf{u}^\kappa(s,a)\\
&= U_\kappa W^{\tau,s_t} x.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\left\|\mathbb{E}\left[\widehat{g}_V(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] - \nabla_V \mathcal{L}_w(Z)\right\|_1 = \left\|A^{\mathrm{T}}\left(W^{\tau,s_t}-W\right)x\right\|_1 &\le 2\left\|\left(W^{\tau,s_t}-W\right)x\right\|_1\\
&\le 2\sum_{s,a}\left|\mathbb{P}_{\pi_b}\left(s_{t+\tau}=s\middle|\,s_t\right)-\mu_{\pi_b}(s)\right|\frac{\pi_b(a|s)x(s,a)}{\hat{\mu}(s,a)}\\
&\le \frac{2\psi}{1-\gamma} d_{\mathrm{TV}}\left(\mathbb{P}^\tau_{\pi_b}\left(\cdot|s_t\right),\mu_{\pi_b}\right) \le \frac{2\psi}{1-\gamma}\mathcal{E}(\tau),
\end{aligned}
$$

where $\mathbb{P}^\tau_{\pi_b}\left(\cdot|s_t\right)$ is the distribution of $s_{t+\tau}$ conditioning on s_t, and the last inequality is due to the definition of $\mathcal{E}(\cdot)$. Similarly, it holds that

$$
\begin{aligned}
\left\|\mathbb{E}\left[\widehat{g}_\lambda(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] - \nabla_\lambda \mathcal{L}_w(Z)\right\|_\infty &\le \frac{2\psi}{1-\gamma}\mathcal{E}(\tau),\\
\left\|\mathbb{E}\left[\widehat{g}_x(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] - \nabla_x \mathcal{L}_w(Z)\right\|_\infty &\le \frac{64}{\varphi(1-\gamma)\varsigma}\mathcal{E}(\tau).
\end{aligned}
$$

Furthermore, for any $Z' = [V';\lambda';x'] \in \mathcal{Z}$, we have

$$
\begin{aligned}
&\left|\left\langle Z', \mathcal{G}(Z) - \mathbb{E}\left[\widehat{g}(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right]\right\rangle\right|\\
\le &\left\|V'\right\|_\infty\left\|\mathbb{E}\left[\widehat{g}_V(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] - \nabla_V \mathcal{L}_w(Z)\right\|_1 + \left\|\lambda\right\|_1\left\|\mathbb{E}\left[\widehat{g}_\lambda(Z;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] - \nabla_\lambda \mathcal{L}_w(Z)\right\|_\infty\\
&+\left|\left\langle r - AV + U_\kappa^{\mathrm{T}}\lambda, \left(W^{\tau,s_t}-W\right)x\right\rangle\right|\\
\le &\frac{128\psi}{\varphi(1-\gamma)^2}\mathcal{E}(\tau).
\end{aligned}
$$

□

H.5 Proof of Proposition H.3

In fact, to prove Proposition H.3, let us prove a more general result stated as follows. Proposition H.3 will follow directly from the (2) and (3) of Proposition H.8. This proposition will also be useful for our later discussion. Recall that we introduce the notation $p(x;s,a) := \frac{x(s,a)}{\hat{\mu}(s,a)}$ and $q(x;s,a) := \frac{x(s,a)}{\hat{\mu}(s,a)^2}$, and the reloaded notation $p(x;\zeta) := p(x;s,a)$ and $q(x;\zeta) := q(x;s,a)$ for sample $\zeta = (s_0,s,a,s',r,\mathbf{u})$. Then the following proposition holds true.

Proposition H.8. ***(1).*** *For all $x \in \mathcal{X}$ and ζ, it holds that*

$$
p(x;\zeta) \le \frac{\psi}{1-\gamma}, \qquad q(x;\zeta) \le \frac{1}{\varsigma}p(x;\zeta) \le \frac{\psi}{(1-\gamma)\varsigma}.
$$

(2). *For all $Z = [V;\lambda;x]$ and ζ, it holds that*

$$
\begin{aligned}
&\left\|\widehat{g}_V(Z;\zeta)\right\| \le 3p(x;\zeta), \qquad \left\|\widehat{g}_\lambda(Z;\zeta)\right\|_\infty \le 2p(x;\zeta),\\
&\left\|\widehat{g}_x(Z;\zeta)\right\|_x \le \frac{64}{\varphi(1-\gamma)}\sqrt{q(x;\zeta)}.
\end{aligned}
$$

(3). *For $x \in \mathcal{X}$ a (possibly random) vector that is $\mathcal{F}_t$-measurable, the (asynchronous) moments of p,q can be bounded as*

$$
\mathbb{E}\left[p(x;\zeta_{t+\tau})\middle|\,\mathcal{F}_t\right] = \sum_{s,a}\frac{\mathbb{P}\left(s_{t+\tau}=s,a_{t+\tau}=a\middle|\,s_t\right)}{\hat{\mu}(s,a)}x(s,a) \le C(\tau)\frac{4}{1-\gamma},
$$

$$\mathbb{E}\left[q(x;\zeta_{t+\tau})\middle|\mathcal{F}_t\right] = \sum_{s,a} \frac{\mathbb{P}\left(s_{t+\tau}=s, a_{t+\tau}=a \middle| s_t\right)}{\hat{\mu}(s,a)} \frac{x(s,a)}{\hat{\mu}(s,a)} \leq C(\tau)\frac{\mathcal{N}\psi}{1-\gamma},$$

$$\mathbb{E}\left[p(x^t;\zeta_{t+\tau})^2\middle|\mathcal{F}_t\right] \leq \frac{\psi}{1-\gamma}\mathbb{E}\left[p(x^t;\zeta_{t+\tau})\middle|\mathcal{F}_t\right] \leq C(\tau)\frac{4\psi}{(1-\gamma)^2}.$$

Since each step of this proposition can be proved by a direct computation similar to the one in Appendix C, we omit the proof for succinctness.

H.6 Proof of Proposition H.4

Similar to the proof of Proposition 4.3, we consider $\hat{\mu}_0(s,a) = \frac{N(s,a)}{N_e}$ and the "failure event"

$$\Omega := \bigcup_{s,a} \left\{ |\mu(s,a) - \hat{\mu}_0(s,a)| > \sqrt{\mu(s,a)\frac{\ell}{N_e}} + \frac{\ell}{N_e} \right\},$$

where $\ell = 100 t_{\mathrm{mix}} \log\left(\frac{12|\mathcal{S}||\mathcal{A}|}{\delta}\right)$. Then by the Bernstein's inequality (Proposition H.1), it holds that

$$\mathbb{P}\left(|\mu(s,a) - \hat{\mu}_0(s,a)| > \sqrt{\mu(s,a)\frac{\ell}{N_e}} + \frac{\ell}{N_e} \right) \leq \frac{\delta}{3|\mathcal{S}||\mathcal{A}|}, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

which further gives $\mathbb{P}(\Omega) \leq \frac{\delta}{3}$. The proof is completed by exactly repeating the estimations in the proof of Proposition 4.3, conditioning on Ω^c.

H.7 Bounding the term S_2

By separately considering each term in the decomposition

$$\Gamma^t = \Gamma_1^t + \Gamma_2^{t-\tau} + \Gamma_3^{t-\tau} + \Gamma_4^t,$$

the following inequalities hold true. The detailed derivations are placed at the end of Appendix H.7.

$$\sum_{t=1}^{\tau} \Gamma^t + \sum_{t=\tau+1}^{T} \Gamma_1^t \lesssim \frac{\tau\psi}{\varphi(1-\gamma)^2}, \tag{52}$$

$$\sum_{t=1}^{T-\tau} \Gamma_3^t \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)\mathcal{N}\psi\iota} \qquad \textit{with probability at least } 1-\frac{\delta}{10}, \tag{53}$$

$$\left|\Gamma_4^t\right| \lesssim \frac{1}{\varphi(1-\gamma)} \frac{|x^t - x^{t-\tau}|(s_t,a_t)}{\hat{\mu}(s_t,a_t)} + \left(1 + \frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\right)\left(\left\|V^t - V^{t-\tau}\right\|_\infty + \left\|\lambda^t - \lambda^{t-\tau}\right\|_1\right). \tag{54}$$

As of Γ_2^t, by directly applying Proposition H.2 we have $|\Gamma_2^t| \lesssim \frac{\psi}{\varphi(1-\gamma)^2}\mathcal{E}(\tau)$. Thus, to estimate S_2, it remains to bound the sum of quantities $\left\|V^t - V^{t-\tau}\right\|_\infty$, $\left\|\lambda^t - \lambda^{t-\tau}\right\|_1$ and $\frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}$. For $\left\|V^t - V^{t-\tau}\right\|_\infty$ and $\left\|\lambda^t - \lambda^{t-\tau}\right\|_1$, as long as $\eta \leq \frac{\alpha_\lambda}{2M_\lambda}$, we have

$$\left\|V^{t+1} - V^t\right\|_\infty \leq \left\|V^{t+1} - V^t\right\| \leq \frac{\eta}{\alpha_V}\left\|\widehat{g}_V(Z^t;\zeta_t)\right\|,$$

$$\left\|\lambda^{t+1} - \lambda^t\right\|_1 \leq \frac{\eta D_{\lambda,1}}{\alpha_\lambda}\left\|\widehat{g}_\lambda(Z^t;\zeta_t)\right\|_\infty,$$

due to Corollary D.5. Therefore, it holds that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=\tau+1}^{T}\left(1 + \frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\right)\left(\left\|V^t - V^{t-\tau}\right\|_\infty + \left\|\lambda^t - \lambda^{t-\tau}\right\|_1\right) \\
\lesssim &\frac{\eta}{T}\sum_{t=\tau+1}^{T}\left(1 + \frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\right)\left(\frac{\left\|\widehat{g}_V(Z^t;\zeta_t)\right\|}{\alpha_V} + \frac{D_{\lambda,1}\left\|\widehat{g}_\lambda(Z^t;\zeta_t)\right\|_\infty}{\alpha_\lambda}\right) \\
\lesssim &\frac{\eta}{T}\sum_{t=\tau+1}^{T}\left(\frac{\left\|\widehat{g}_V(Z^t;\zeta_t)\right\|^2}{\alpha_V} + \frac{D_{\lambda,1}\left\|\widehat{g}_\lambda(Z^t;\zeta_t)\right\|_\infty^2}{\alpha_\lambda}\right) + \frac{\eta}{T}\left(\frac{1}{\alpha_V} + \frac{D_{\lambda,1}}{\alpha_\lambda}\right)\sum_{t=\tau+1}^{T}\left(1 + \frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\right)^2.
\end{aligned}$$

Finally, we apply Bernstein's inequality to bound the sequence $\left(\frac{x'(s_t,a_t)^2}{\hat{\mu}(s_t,a_t)^2}\right)_t$ as follows. Due to

$$\frac{x'(s,a)}{\hat{\mu}(s,a)} \le \frac{\psi}{1-\gamma}, \qquad \mathbb{E}_{s,a\sim\mu}\left[\frac{x'(s,a)^2}{\hat{\mu}(s,a)^2}\right] = \sum_{s,a} \frac{\mu(s,a)}{\hat{\mu}(s,a)} \frac{x'(s,a)}{\hat{\mu}(s,a)} x'(s,a) \le \frac{8\psi}{(1-\gamma)^2},$$

and Proposition H.1, with probability at least $1 - \delta/10$, it holds that

$$\sum_{t=\tau+1}^{T} \frac{x'(s_t,a_t)^2}{\hat{\mu}(s_t,a_t)^2} \lesssim T\frac{\psi}{(1-\gamma)^2} + t_{\mathrm{mix}} \frac{\psi^2}{(1-\gamma)^2} \log\frac{1}{\delta} \lesssim \frac{T\psi}{(1-\gamma)^2}.$$

Combining all the estimations above completes the proof of (49).

H.7.1 Derivation of inequality (52)

By definition, it holds that

$$\sum_{t=1}^{\tau} \Gamma^t + \sum_{t=\tau+1}^{T} \Gamma_1^t = \sum_{t=T-\tau+1}^{T} \left\langle \mathcal{G}(Z^t), Z^t - Z' \right\rangle - \sum_{t=1}^{\tau} \left\langle \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle.$$

For a sample $\zeta = (s_0, s, a, s', r, \mathbf{u})$, we denote

$$\widehat{\mathcal{L}}_\zeta(V,\lambda,x) := V(s_0) + \frac{x(s,a)}{\hat{\mu}(s,a)} \left(r - V(s) + \gamma V(s') + \langle \lambda, \mathbf{u}^\kappa \rangle\right).$$

Then, it holds that

$$\langle \widehat{g}(Z;\zeta), Z - Z' \rangle = \widehat{\mathcal{L}}_\zeta(V,\lambda,x') - \widehat{\mathcal{L}}_\zeta(V',\lambda',x). \tag{55}$$

Hence we have

$$\left|\left\langle \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle\right| \le \left|\widehat{\mathcal{L}}_{\zeta_t}(V^t,\lambda^t,x')\right| + \left|\widehat{\mathcal{L}}_{\zeta_t}(V',\lambda',x^t)\right| \le \frac{100\psi}{\varphi(1-\gamma)^2}.$$

Similarly, it holds that

$$\left|\left\langle \mathcal{G}(Z^t), Z^t - Z' \right\rangle\right| \le \left|\mathcal{L}_w(V^t,\lambda^t,x')\right| + \left|\mathcal{L}_w(V',\lambda',x^t)\right| \le \frac{512}{\varphi(1-\gamma)^2},$$

and we complete the proof by combining the estimations above.

H.7.2 Derivation of inequality (53)

As in Appendix D.3, we consider the sequences

$$\begin{aligned}
\Delta_V^t &:= \widehat{g}_V(Z^t;\zeta_{t+\tau}) - \mathbb{E}\left[\widehat{g}_V(Z^t;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right], \\
\Delta_\lambda^t &:= \widehat{g}_\lambda(Z^t;\zeta_{t+\tau}) - \mathbb{E}\left[\widehat{g}_\lambda(Z^t;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right], \\
\Delta_x^t &:= \widehat{g}_x(Z^t;\zeta_{t+\tau}) - \mathbb{E}\left[\widehat{g}_x(Z^t;\zeta_{t+\tau}) \middle| \mathcal{F}_t\right].
\end{aligned}$$

They are no longer martingale difference sequences, because $\mathbb{E}\left[\Delta^t \middle| \mathcal{F}_t\right] = 0$ but Δ^t is $\mathcal{F}_{t+\tau+1}$ measurable. Therefore, we invoke the following modified version of Bernstein's inequality.

Lemma H.9 (Modified Bernstein's Inequality)**.** *Assume $\{x_i\}_{i=1}^n$ is a sequence of random vectors in $\mathbb{R}^d$, such that $\mathbb{E}\left[x_t \middle| \mathcal{F}_t\right] = 0$ and x_t is $\mathcal{F}_{t+\tau}$ measurable. Assume that $\mathbb{E}\left[\|x_t\|^2 \middle| \mathcal{F}_t\right] \le \sigma^2$ and $\|x_t\| \le M$ a.s., then with probability at least $1-\delta$,*

$$\left\|\sum_{i=1}^{n} x^i\right\| \le 2\sigma\sqrt{n\tau \log\left(\frac{(d+1)\tau}{\delta}\right)} + 2M\tau \log\left(\frac{(d+1)\tau}{\delta}\right).$$

When the ℓ_2 norm is replaced by the ℓ_∞ norm, i.e., $\{x_i\}_{i=1}^n$ satisfies $\mathbb{E}\left[\|x_t\|_\infty^2 \middle| \mathcal{F}_t\right] \le \sigma^2$, we have

$$\left\|\sum_{i=1}^{n} x^i\right\|_\infty \le 2\sigma\sqrt{n\tau \log\left(\frac{2d\tau}{\delta}\right)} + 2M\tau \log\left(\frac{2d\tau}{\delta}\right)$$

with probability at least $1-\delta$.

We still decompose

$$\sum_{t=1}^{T-\tau}\Gamma_3^t = \underbrace{\sum_{t=1}^{T-\tau}\left(\left\langle\Delta_V^t, V' - V^1\right\rangle + \left\langle\Delta_\lambda^t, \lambda' - \lambda^1\right\rangle\right)}_{S_c}$$

$$+ \underbrace{\sum_{t=1}^{T-\tau}\left(\left\langle\Delta_V^t, V^1 - V^t\right\rangle + \left\langle\Delta_\lambda^t, \lambda^1 - \lambda^t\right\rangle + \left\langle-\Delta_x^t, x' - x^t\right\rangle\right)}_{S_m}.$$

Most of the following analysis is similar to the one in Appendix D.3.

Correlated part Rewrite

$$\begin{aligned} S_c &= \left\langle\sum_{t=1}^{T-\tau}\Delta_V^t, V' - V^1\right\rangle + \left\langle\sum_{t=1}^{T-\tau}\Delta_\lambda^t, \lambda' - \lambda^1\right\rangle \\ &\le \left\|V' - V^1\right\| \cdot \left\|\sum_{t=1}^{T-\tau}\Delta_V^t\right\| + \left\|\lambda' - \lambda^1\right\|_1 \cdot \left\|\sum_{t=1}^{T-\tau}\Delta_\lambda^t\right\|_\infty. \end{aligned}$$

For each t, by Proposition H.3 (or Proposition H.8), we have

$$\begin{aligned} &\mathbb{E}\left[\left\|\Delta_V^t\right\|^2\middle|\mathcal{F}_t\right] \le \mathbb{E}\left[\left\|\widehat{g}_V(Z^t;\zeta_{t+\tau})\right\|^2\middle|\mathcal{F}_t\right] \lesssim \frac{C(\tau)\psi}{(1-\gamma)^2}, \quad \left\|\Delta_V^t\right\| \lesssim \frac{\psi}{1-\gamma}, \\ &\mathbb{E}\left[\left\|\Delta_\lambda^t\right\|_\infty^2\middle|\mathcal{F}_t\right] \lesssim \mathbb{E}\left[\left\|\widehat{g}_\lambda(Z^t;\zeta_{t+\tau})\right\|_\infty^2\middle|\mathcal{F}_t\right] \lesssim \frac{C(\tau)\psi}{(1-\gamma)^2}, \quad \left\|\Delta_\lambda^t\right\|_\infty \lesssim \frac{\psi}{1-\gamma}. \end{aligned}$$

Thus, we can apply Lemma H.9 to derive that, with probability at least $1 - \delta/20$,

$$\begin{aligned} \left\|\sum_{t=1}^{T-\tau}\Delta_V^t\right\| &\lesssim \frac{1}{1-\gamma}\sqrt{T\tau C(\tau)\psi\log(1/\delta)} + \frac{\psi}{1-\gamma}\cdot\tau\log(1/\delta), \\ \left\|\sum_{t=1}^{T-\tau}\Delta_\lambda^t\right\|_\infty &\lesssim \frac{1}{1-\gamma}\sqrt{T\tau C(\tau)\psi\log(I/\delta)} + \frac{\psi}{1-\gamma}\cdot\tau\log(I/\delta). \end{aligned}$$

Therefore, it holds that with probability at least $1 - \delta/20$,

$$S_c \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)|\mathcal{S}|\psi\iota} + \frac{\tau\psi\iota}{1-\gamma} \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)|\mathcal{S}|\psi\iota}. \tag{56}$$

Martingale part In order to bound S_m, we have to consider $\overline{\Delta}_V^t := \left\langle\Delta_V^t, V^1 - V^t\right\rangle$,$\overline{\Delta}_\lambda^t := \left\langle\Delta_\lambda^t, \lambda^1 - \lambda^t\right\rangle$, $\overline{\Delta}_x^t := \langle\Delta_x^t, x^t - x'\rangle$. By Proposition H.3, it holds that

$$\begin{aligned} &\left|\overline{\Delta}_V^t\right| \lesssim \frac{\psi}{\varphi(1-\gamma)^2}, \quad \mathbb{E}\left[\left(\overline{\Delta}_V^t\right)^2\middle|\mathcal{F}_t\right] \le D_V^2\mathbb{E}\left[\left\|\Delta_V^t\right\|^2\middle|\mathcal{F}_t\right] \lesssim \frac{C(\tau)\psi}{\varphi^2(1-\gamma)^4}, \\ &\left|\overline{\Delta}_\lambda^t\right| \le \frac{\psi}{\varphi(1-\gamma)}, \quad \mathbb{E}\left[\left(\overline{\Delta}_\lambda^t\right)^2\middle|\mathcal{F}_t\right] \le D_{\lambda,1}^2\mathbb{E}\left[\left\|\Delta_\lambda^t\right\|_\infty^2\middle|\mathcal{F}_t\right] \lesssim \frac{C(\tau)\psi}{\varphi^2(1-\gamma)^2}, \\ &\left|\overline{\Delta}_x^t\right| \le \frac{\psi}{\varphi(1-\gamma)^2}, \quad \mathbb{E}\left[\left(\overline{\Delta}_x^t\right)^2\middle|\mathcal{F}_t\right] \le \mathbb{E}\left[\left\|\frac{x' - x^t}{\sqrt{x' + x^t}}\right\|^2\left\|\Delta_x^t\right\|_{x'+x^t}^2\middle|\mathcal{F}_t\right] \lesssim \frac{C(\tau)\mathcal{N}\psi}{\varphi^2(1-\gamma)^4}. \end{aligned}$$

Thus, by applying Lemma H.9, the following three estimations hold with probability at least $1 - \delta/20$

$$\begin{aligned} \sum_{t=1}^{T-\tau}\overline{\Delta}_V^t &\lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)\psi\log(1/\delta)} + \frac{\psi}{\varphi(1-\gamma)^2}\cdot\tau\log(1/\delta), \\ \sum_{t=1}^{T-\tau}\overline{\Delta}_\lambda^t &\lesssim \frac{1}{\varphi(1-\gamma)}\sqrt{T\tau C(\tau)\psi\log(1/\delta)} + \frac{\psi}{\varphi(1-\gamma)}\cdot\tau\log(1/\delta), \end{aligned}$$

$$\sum_{t=1}^{T-\tau} \overline{\Delta}_x^t \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)\mathcal{N}\psi\log(1/\delta)} + \frac{\psi}{\varphi(1-\gamma)^2}\cdot\tau\log(1/\delta).$$

Therefore,

$$S_m \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)\mathcal{N}\psi\iota} + \frac{\tau\psi\iota}{\varphi(1-\gamma)^2} \lesssim \frac{1}{\varphi(1-\gamma)^2}\sqrt{T\tau C(\tau)\mathcal{N}\psi\iota}. \tag{57}$$

Combining (57) with (56) completes the proof.

Proof of Lemma H.9. We reduce Lemma H.9 to the standard martingale Bernstein's inequality (Lemma B.1). The set $[n]$ can be decomposed into

$$[n] = \bigsqcup_{k=1}^{\tau}\mathcal{I}_k, \qquad \mathcal{I}_k := \{j \in [n] : j \equiv k \mod \tau\}.$$

For each k, the sequence $(X_j)_{j\in\mathcal{I}_k}$ is a martingale difference sequence w.r.t. the filtration $(\mathcal{F}_j)_{j\in\mathcal{I}_k}$. Hence by Lemma B.1, with probability at least $1 - \delta/\tau$, we have

$$\left\|\sum_{j\in\mathcal{I}_k} x^j\right\| \le 2\sigma\sqrt{|\mathcal{I}_k|\log\left(\frac{(d+1)\tau}{\delta}\right)} + 2M\log\left(\frac{(d+1)\tau}{\delta}\right).$$

Summing over $k = 1, \cdots, \tau$ yields that with probability at least $1-\delta$

$$\begin{aligned}\left\|\sum_{j=1}^{n} x^j\right\| &\le 2\sigma\sqrt{\log\left(\frac{(d+1)\tau}{\delta}\right)}\sum_{k=1}^{\tau}\sqrt{|\mathcal{I}_k|} + 2M\tau\log\left(\frac{(d+1)\tau}{\delta}\right)\\ &\le 2\sigma\sqrt{n\tau\log\left(\frac{(d+1)\tau}{\delta}\right)} + 2M\tau\log\left(\frac{(d+1)\tau}{\delta}\right),\end{aligned}$$

where the last inequality is due to the Cauchy inequality.

The analogous ℓ_∞ case can be done similarly. □

H.7.3 Derivation of inequality (54)

By (55), it holds that

$$\begin{aligned}\Gamma_4^t =& \left\langle \widehat{g}(Z^{t-\tau};\zeta_t), Z^{t-\tau} - Z' \right\rangle - \left\langle \widehat{g}(Z^t;\zeta_t), Z^t - Z' \right\rangle\\ =& \widehat{\mathcal{L}}_{\zeta_t}(V^{t-\tau},\lambda^{t-\tau},x') - \widehat{\mathcal{L}}_{\zeta_t}(V',\lambda',x^{t-\tau}) + \widehat{\mathcal{L}}_{\zeta_t}(V',\lambda',x^t) - \widehat{\mathcal{L}}_{\zeta_t}(V^t,\lambda^t,x').\end{aligned} \tag{58}$$

Then we have

$$\begin{aligned}&\left|\widehat{\mathcal{L}}_{\zeta_t}(V',\lambda',x^t) - \widehat{\mathcal{L}}_{\zeta_t}(V',\lambda',x^{t-\tau})\right|\\ =&\frac{|x^t - x^{t-\tau}|(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\left|r_t - V'(s_t) + \gamma V'(s_{t+1}) + \langle \lambda', \mathbf{u}_t^\kappa\rangle\right|\\ \le&\frac{|x^t - x^{t-\tau}|(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\left(1 + \frac{16}{1-\gamma}\left(1+\frac{2}{\varphi}\right) + \frac{8(1+\kappa)}{\varphi}\right)\\ \le&\frac{64}{\varphi(1-\gamma)}\frac{|x^t - x^{t-\tau}|(s_t,a_t)}{\hat{\mu}(s_t,a_t)}.\end{aligned}$$

Similarly,

$$\begin{aligned}&\left|\widehat{\mathcal{L}}_{\zeta_t}(V^t,\lambda^t,x') - \widehat{\mathcal{L}}_{\zeta_t}(V^{t-\tau},\lambda^{t-\tau},x')\right|\\ \le&\left|V^t(s_{0,t}) - V^{t-\tau}(s_{0,t})\right|\\ &+ \frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\left(\left|V^t(s_t) - V^{t-\tau}(s_t)\right| + \gamma\left|V^t(s_{t+1}) - V^{t-\tau}(s_{t+1})\right| + \left|\left\langle \lambda^t - \lambda^{t-\tau}, \mathbf{u}_t^\kappa\right\rangle\right|\right)\\ \le&\left\|V^t - V^{t-\tau}\right\|_\infty\left(1 + 2\frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}\right) + \left\|\lambda^t - \lambda^{t-\tau}\right\|_1\cdot 128\frac{x'(s_t,a_t)}{\hat{\mu}(s_t,a_t)}.\end{aligned}$$

The proof is completed by combining (58) with the estimations above.

H.8 Proof of Proposition H.6

The proof of Proposition H.6 is separated into two steps.

Step 1. We derive bounds on $\sum p(x^t;\zeta_{t+\tau})^2$ and $\sum q(x^t;\zeta_{t+\tau})$ by directly applying Bernstein's inequality.

Step 2. We leverage the idea demonstrate in (13) again to bound $\sum p(x^t;\zeta_t)^2$ and $\sum q(x^t;\zeta_t)$, by bounding their difference with $\sum p(x^t;\zeta_{t+\tau})^2$ and $\sum q(x^t;\zeta_{t+\tau})$ respectively.

Then we finalize the proof by combining the results of Step 1 and Step 2.

H.8.1 Step 1. Bounding the asynchronous sums

First, let us present the following result for the ease of discussion.
Corollary. *Assume $\{x_i\}_{i=1}^n$ is a sequence of random variables, such that x_t is $\mathcal{F}_{t+\tau}$ measurable, and $\mathbb{E}\left[|x_t|\middle|\mathcal{F}_t\right] \le c$, $|x_t| \le M$ a.s. Then with probability at least $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n x^i\right| \le 2c\tau + 3M\tau\frac{\log(2\tau/\delta)}{n}.$$

By Proposition H.8, we have

$$q(x^t;\zeta_{t+\tau}) \le \frac{\psi}{(1-\gamma)\varsigma}, \quad \mathbb{E}\left[q(x^t;\zeta_{t+\tau})\middle|\mathcal{F}_t\right] \le C(\tau)\frac{\mathcal{N}\psi}{1-\gamma}.$$

Applying the above corollary yields that with probability at least $1-{}^{\delta}/_{20\tau_0}$,

$$\begin{aligned}\sum_{t=1}^{T-\tau} q(x^t;\zeta_{t+\tau}) &\lesssim TC(\tau)\frac{\mathcal{N}\psi}{1-\gamma} + \frac{\tau\psi}{(1-\gamma)\varsigma}\log\left(\frac{\tau_0}{\delta}\right)\\ &\lesssim TC(\tau)\frac{\mathcal{N}\psi}{1-\gamma} + \frac{\tau_0\mathcal{N}\psi^2\iota}{\varphi(1-\gamma)^3\epsilon_e}\\ &\lesssim TC(\tau)\frac{\mathcal{N}\psi}{1-\gamma},\end{aligned}$$

where the last inequality is due to $T \gtrsim \frac{\tau_0^2\mathcal{N}\psi\iota^3}{\varphi^2(1-\gamma)^4\epsilon_e^2} \ge \frac{\tau_0\psi\iota}{\varphi(1-\gamma)^2\epsilon_e}$.

Similarly, we have

$$p(x^t;\zeta_{t+\tau}) \le \frac{\psi}{1-\gamma}, \quad \mathbb{E}\left[p(x^t;\zeta_{t+\tau})^2\middle|\mathcal{F}_t\right] \le C(\tau)\frac{4\psi}{(1-\gamma)^2}.$$

Therefore, for each $1\le\tau\le\tau_0$, it holds with probability at least $1-{}^{\delta}/_{20\tau_0}$

$$\sum_{t=1}^{T-\tau} p(x^t;\zeta_{t+\tau})^2 \lesssim \frac{TC(\tau)\psi}{(1-\gamma)^2} + \frac{\tau\psi^2}{(1-\gamma)^2}\log\left(\frac{\tau_0}{\delta}\right) \lesssim \frac{TC(\tau)\psi}{(1-\gamma)^2}.$$

By taking the union bound for $1\le\tau\le\tau_0$, we conclude that with probability at least $1-{}^{\delta}/_{10}$,

$$\sum_{t=1}^{T-\tau} p(x^t;\zeta_{t+\tau})^2 \lesssim \frac{TC(\tau)\psi}{(1-\gamma)^2}, \qquad \sum_{t=1}^{T-\tau} q(x^t;\zeta_{t+\tau}) \lesssim \frac{TC(\tau)\mathcal{N}\psi}{1-\gamma}, \tag{59}$$

hold simultaneously and uniformly for $1\le\tau\le\tau_0$.

H.8.2 Step 2. Bounding the difference

Utilizing the closeness between Z^t and $Z^{t+\tau}$, we bound the difference $q(x^t;\zeta_t) - q(x^t;\zeta_{t+\tau})$ as

$$\begin{aligned}\sum_{t=1}^{T} q(x^t;\zeta_t) - \sum_{t=1}^{T-\tau} q(x^t;\zeta_{t+\tau}) &\le \sum_{t=1}^{\tau} q(x^t;\zeta_t) + \sum_{t=1}^{T-\tau} q(\left|x^t - x^{t+\tau}\right|;\zeta_{t+\tau})\\ &\le \frac{\tau\psi}{(1-\gamma)\varsigma} + \frac{1}{\varsigma}\sum_{t=1}^{T-\tau} p(\left|x^t - x^{t+\tau}\right|;\zeta_{t+\tau}).\end{aligned} \tag{60}$$

We next deal with the quantity $p(|x^t - x^{t+\tau}\big|\,; \zeta_{t+\tau})$ carefully. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, it holds that

$$\begin{aligned}
p(\big|x^t - x^{t+\tau}\big|\,; s,a) &= \frac{|x^t(s,a) - x^{t+\tau}(s,a)|}{\hat{\mu}(s,a)} \\
&\le \frac{1}{\hat{\mu}(s,a)} \sum_{t'=t}^{t+\tau-1} \left| x^{t'}(s,a) - x^{t'+1}(s,a) \right| \\
&\le \frac{1}{\hat{\mu}(s,a)} \sum_{t'=t}^{t+\tau-1} \sqrt{x^{t'}(s,a) + x^{t'+1}(s,a)} \left\| \frac{x^{t'} - x^{t'+1}}{\sqrt{x^{t'} + x^{t'+1}}} \right\| \\
&\overset{(a)}{\lesssim} \frac{\eta}{\alpha_x} \frac{1}{\hat{\mu}(s,a)} \sum_{t'=t}^{t+\tau-1} \sqrt{x^{t'}(s,a) + x^{t'+1}(s,a)} \left\| \widehat{g}_x(Z^{t'}; \zeta_{t'}) \right\|_{x^{t'}} \\
&\overset{(b)}{\lesssim} \frac{\eta}{\alpha_x} \frac{1}{\hat{\mu}(s,a)} \sum_{t'=t}^{t+\tau-1} \sqrt{x^{t'}(s,a) + x^{t'+1}(s,a)} \cdot \frac{1}{\varphi(1-\gamma)} \sqrt{q(x^t; \zeta_t)} \\
&= \frac{\eta}{\alpha_x} \cdot \frac{1}{\varphi(1-\gamma)} \sum_{t'=t}^{t+\tau-1} \sqrt{q(x^{t'}; s,a) + q(x^{t'+1}; s,a)} \sqrt{q(x^t; \zeta_t)} \\
&\overset{(c)}{\le} \frac{\eta}{\alpha_x} \cdot \frac{1}{\varphi(1-\gamma)} \sqrt{\sum_{t'=t}^{t+\tau-1} q(x^t; \zeta_t)} \sqrt{\sum_{t'=t}^{t+\tau} q(x^{t'}; s,a)}.
\end{aligned}$$

Here the inequality (a) is due to Corollary D.5, the inequality (b) is due to Proposition H.8, and the inequality (c) comes from Cauchy inequality. Hence, we have

$$\begin{aligned}
\sum_{t=1}^{T-\tau} p(\big|x^t - x^{t+\tau}\big|\,; \zeta_{t+\tau}) \lesssim& \frac{\eta}{\alpha_x} \cdot \frac{1}{\varphi(1-\gamma)} \sum_{t=1}^{T-\tau} \sqrt{\sum_{t'=t}^{t+\tau-1} q(x^t; \zeta_t)} \sqrt{\sum_{t'=t}^{t+\tau} q(x^{t'}; \zeta_{t+\tau})} \\
\le& \frac{\eta}{\alpha_x} \cdot \frac{1}{\varphi(1-\gamma)} \sqrt{\sum_{t=1}^{T-\tau} \sum_{t'=t}^{t+\tau-1} q(x^t; \zeta_t)} \sqrt{\sum_{t=1}^{T-\tau} \sum_{t'=t}^{t+\tau} q(x^{t'}; \zeta_{t+\tau})} \\
\le& \frac{\eta}{\alpha_x} \cdot \frac{1}{\varphi(1-\gamma)} \sqrt{\tau \sum_{t=1}^{T} q(x^t; \zeta_t)} \sqrt{\sum_{j=1}^{\tau} \sum_{t=1}^{T-j} q(x^t; \zeta_{t+j})}. \qquad (61)
\end{aligned}$$

Combining (61) with (60) yields

$$\begin{aligned}
&\sum_{t=1}^{T-\tau} q(x^t; \zeta_t) - \sum_{t=1}^{T-\tau} q(x^t; \zeta_{t+\tau}) \\
\lesssim\; & \frac{\tau\psi}{(1-\gamma)\varsigma} + \frac{\eta}{\alpha_x} \frac{1}{\varphi(1-\gamma)\varsigma} \sqrt{\tau \sum_{t=1}^{T} q(x^t; \zeta_t)} \sqrt{\sum_{t=1}^{T-\tau} \sum_{t'=t}^{t+\tau} q(x^{t'}; \zeta_{t+\tau})}
\end{aligned} \qquad (62)$$

Similarly, it holds that for $0 \le j \le \tau$,

$$\begin{aligned}
&\sum_{t=1}^{T-j} q(x^t; \zeta_{t+j}) - \sum_{t=1}^{T-\tau} q(x^t; \zeta_{t+\tau}) \\
\lesssim\; & \frac{\tau\psi}{(1-\gamma)\varsigma} + \frac{\eta}{\alpha_x} \frac{1}{\varphi(1-\gamma)\varsigma} \sqrt{\tau \sum_{t=1}^{T} q(x^t; \zeta_t)} \sqrt{\sum_{t=1}^{T-\tau} \sum_{t'=t}^{t+\tau} q(x^{t'}; \zeta_{t+\tau})}.
\end{aligned} \qquad (63)$$

H.8.3 Combining Step 1 and Step 2

Actually, (63) is already enough to bound $\sum_{t=1}^{T} q(x^t; \zeta_t)$. For simplicity, we denote

$$Q_1 := \frac{c_0 \tau \psi}{(1-\gamma)\varsigma} + \sum_{t=1}^{T-\tau} q(x^t; \zeta_{t+\tau}), \qquad Q_2 := \sum_{t=1}^{T} q(x^t; \zeta_t), \qquad c := c_0 \frac{\eta}{\alpha_x} \frac{1}{\varphi(1-\gamma)\varsigma},$$

$$Q_3 := \frac{1}{\tau}\sum_{t=1}^{T-\tau}\sum_{t'=t}^{t+\tau} q(x^{t'};\zeta_{t+\tau}) = \frac{1}{\tau}\sum_{j=1}^{\tau}\sum_{t=1}^{T-\tau+j} q(x^t;\zeta_{t+\tau-j}),$$

where c_0 is a universal constant hidden by the $\lesssim$ in (63). Now, (63) implies

$$\begin{aligned} &Q_2 \le Q_1 + c\tau\sqrt{Q_2Q_3}, \quad Q_3 \le Q_1 + c\tau\sqrt{Q_2Q_3}, \\ \Rightarrow &Q_2 + Q_3 \le Q_1 + c\tau(Q_2+Q_3). \end{aligned} \tag{64}$$

Thus, as long as $c\tau_0 \le \frac{1}{2}$, we have $Q_2 + Q_3 \le 2Q_1$. The condition $c\tau_0 \le \frac{1}{2}$ is equivalent to

$$\frac{1}{2c_0} \ge \tau_0\frac{\eta}{\alpha_x}\frac{1}{\varphi(1-\gamma)\varsigma} = \tau_0\cdot\sqrt{\frac{1}{T}}\cdot\left(\frac{1}{\varphi(1-\gamma)}\sqrt{\frac{\mathcal{N}\psi}{\log\psi}}\right)^{-1}\frac{1}{\varphi(1-\gamma)\varsigma} = \sqrt{\frac{4\tau_0^2\mathcal{N}\psi\log\psi}{\varphi^2(1-\gamma)^4\epsilon_e^2}\cdot\frac{1}{T}}.$$

Thus, $T \ge 16c_0^2\frac{\tau_0^2\mathcal{N}\psi\log\psi}{\varphi^2(1-\gamma)^4\epsilon_e^2}$ is enough to ensure $Q_2 \le 2Q_1, Q_3 \le 2Q_1$ for any $\tau \le \tau_0$. Here, according to (59) we have

$$Q_1 = \frac{c_0\tau\psi}{(1-\gamma)\varsigma} + \sum_{t=1}^{T-\tau} q(x^t;\zeta_{t+\tau}) \lesssim \frac{c_0\tau_0\psi}{(1-\gamma)\varsigma} + \frac{TC(\tau)\mathcal{N}\psi}{1-\gamma} \lesssim \frac{TC(\tau)\mathcal{N}\psi}{1-\gamma}.$$

Consequently, we obtain

$$\begin{aligned} &\sum_{t=1}^{T} q(x^t;\zeta_t) \lesssim \frac{TC(\tau)\mathcal{N}\psi}{1-\gamma}, \\ &\sum_{t=1}^{T-\tau}\sum_{t'=t}^{t+\tau} q(x^{t'};\zeta_{t+\tau}) \lesssim \tau\frac{TC(\tau)\mathcal{N}\psi}{1-\gamma}. \end{aligned} \tag{65}$$

Hence, by (61),

$$\sum_{t=\tau+1}^{T} p(\left|x^t - x^{t-\tau}\right|;\zeta_t) \lesssim \frac{\tau C(\tau)}{1-\gamma}\sqrt{T\mathcal{N}\psi\log\psi}.$$

We can further establish the bound for $\sum_{t=1}^{T} p(x^t;\zeta_t)^2$ as

$$\begin{aligned} \sum_{t=1}^{T} p(x^t;\zeta_t)^2 &\lesssim \sum_{t=1}^{\tau} p(x^t;\zeta_t)^2 + \sum_{t=1}^{T-\tau}\left[p(x^t;\zeta_{t+\tau})^2 + p(\left|x^t - x^{t+\tau}\right|;\zeta_{t+\tau})^2\right] \\ &\overset{(a)}{\lesssim} \tau\frac{\psi^2}{(1-\gamma)^2} + \sum_{t=1}^{T-\tau} p(x^t;\zeta_{t+\tau})^2 + \frac{\psi}{1-\gamma}\sum_{t=1}^{T-\tau} p(\left|x^t - x^{t+\tau}\right|;\zeta_{t+\tau}) \\ &\lesssim \frac{\tau\psi^2}{(1-\gamma)^2} + \frac{TC(\tau)\psi}{(1-\gamma)^2} + \frac{\tau C(\tau)}{1-\gamma}\sqrt{T\mathcal{N}\psi\log\psi} \\ &\overset{(b)}{\lesssim} \frac{TC(\tau)\psi}{(1-\gamma)^2}, \end{aligned} \tag{66}$$

where the inequality (a) is due to $p(x^t;\zeta_t) \le \frac{\psi}{1-\gamma}$, $p(\left|x^t - x^{t+\tau}\right|;\zeta_{t+\tau}) \le \frac{2\psi}{1-\gamma}$, and the inequality (b) is due to our requirement $T \gtrsim \frac{\tau_0^2\mathcal{N}\psi\iota}{\varphi^2(1-\gamma)^4\epsilon_e^2}$.

The proof is completed by taking $\tau = \tau_0$ in (66) and (65).