Performative Reinforcement Learning with Linear Markov Decision Process

Debmalya Mandal
University of Warwick, UK

Goran Radanović
Max-Planck Institute for Software Systems, Germany

Abstract

We study the setting of *performative reinforcement learning* where the deployed policy affects both the reward, and the transition of the underlying Markov decision process. Prior work (Mandal et al., 2023) has addressed this problem under the tabular setting and established last-iterate convergence of repeated retraining with iteration complexity explicitly depending on the number of states. In this work, we generalize the results to *linear Markov decision processes* which is the primary theoretical model of large-scale MDPs. The main challenge with linear MDP is that the regularized objective is no longer strongly concave and we want a bound that scales with the dimension of the features, rather than states which can be infinite. Our first result shows that repeatedly optimizing a regularized objective converges to a *performatively stable policy*. In the absence of strong concavity, our analysis leverages a new recurrence relation that uses a specific linear combination of optimal dual solutions for proving convergence. We then tackle the finite sample setting where the learner has access to a set of trajectories drawn from the current policy. We consider a reparametrized version of the primal problem, and construct an empirical Lagrangian which is to be optimized from the samples. We show that, under a *bounded coverage* condition, repeatedly solving a saddle point of this empirical Lagrangian converges to a performatively stable solution, and also construct a primal-dual algorithm that solves the empirical Lagrangian efficiently. Finally, we show several applications of the general framework of performative RL including multi-agent systems.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

The success of reinforcement learning in challenging domains like Go (Silver et al., 2017), Poker (Brown and Sandholm, 2019), language modeling (Ouyang et al., 2022) has led to its adoption in various human-facing systems. Indeed we are increasingly observing the use of RL in social media recommendations (Afsar et al., 2022), healthcare (Yu et al., 2021), traffic modeling (Wei et al., 2021), and many other open-ended problems. However, the classical framework of reinforcement learning is not directly applicable in these reactive settings. For example, a news recommender system might affect people's preferences, and a traffic prediction system influences how drivers choose traffic routes (Dulac-Arnold et al., 2019). The common theme underlying these different settings is that there is no fixed Markov decision process (MDP), but rather the underlying MDP changes in response to the deployed policy.

Recently, Mandal et al. (2023) introduced the framework of *performative reinforcement learning* to encompass such reactive setting in reinforcement learning. In performative RL, the deployed policy affects both the reward and the probability transition function of the underlying MDP. In contrary to the classical RL, there is no fixed optimal policy, rather the new goal is to find a *performatively stable policy*, a policy which is optimal with respect to the updated MDP after deployment. The main result of Mandal et al. (2023) is that repeatedly optimizing a regularized value function converges to an approximate performative stable policy in a *tabular* setting.

However, most of the practical applications of RL involve a large number of states, and requires *function approximation* to deal with the complexity of the optimal policy or value function. Indeed, most of the success of RL (Mnih et al., 2015; Lillicrap et al., 2015; Silver et al., 2017) can be attributed to handling large scale state spaces through deep neural networks. On the theoretical side, a significant effort has also been invested to design provable RL algorithms with various complexities of function approximation (Jin et al., 2020, 2021a; Zhou et al., 2021; Yang and Wang, 2020). However, these works assume a *static* setting as the underlying MDP doesn't change in response to the policy, and as discussed earlier, cannot address the concerns of real-world applications. In this work, our goal is

the design of provable RL algorithms with function approximation in the presence of *performativity*. In particular, we ask the following questions.

> Can we design provably efficient RL algorithms that converge to *performatively stable policy* under linear function approximation? Moreover, when there is only access to a finite number of samples, can we obtain stable policies with statistical and computational complexities depending only on the dimension of the features?

The reader might ask whether existing works (Mandal et al., 2023; Perdomo et al., 2020a) can be naturally extended to our setting. Although Mandal et al. (2023) showed that repeated optimization converges to a performatively stable policy, their approach doesn't translate to linear function approximation as the regularized objective is no longer a strongly concave function, and we need to devise new strategies for ensuring convergence. Furthermore, when we can access the underlying MDP only through trajectories collected from the deployed policy, we need to ensure that sample complexity only grows polynomially with the dimension of the features D, and is independent of the size of the state space, which can be infinite. In particular, our contributions are the following.

1. We show that repeatedly optimizing a regularized objective converges to a *performatively stable policy* in a linear MDP. In the absence of strong concavity of the objective, we establish a new recurrence relation depending on a time-varying linear combination of the optimal dual solutions of the regularized objective.

2. Furthermore, we show how to tune the strength of the regularization to obtain an approximate *performatively* stable optimal policy based on the sensitivity of the environment. We then show that our method also converges to a *performatively optimal policy* by establishing a bound on the distance between performatively optimal and stable policies.

3. For the finite sample setting, we introduce a *reparametrized* version of the primal problem. Based on the new reparametrized problem, we construct an empirical Lagrangian which is to be optimized from the samples, and show that, under a *bounded coverage* condition, repeatedly solving a saddle point of this Lagrangian converges to a performatively stable policy, with a sample complexity polynomial in the dimension D of the features. We also design an efficient algorithm for solving the empirical Lagrangian.

4. Finally, we show several applications of the framework of performative RL involving stochastic Stackelberg games with one and multiple followers that further highlights the generality of the framework of performativity in RL.

1.1 Related Work

We recognize two important lines of related work: *performative prediction* and *reinforcement learning*.

Performative Prediction. Performative prediction was introduced by Perdomo et al. (2020b), who proposed two important solution concepts, performative stability and optimality, and established convergence guarantees of repeated retraining approaches. A series of papers extended these results (Mendler-Dünner et al., 2020; Miller et al., 2021; Izzo et al., 2021; Lu, 2023; Yan and Cao, 2024) and studied different variants of the canonical performative prediction setting. These variants include multi-agent settings (Narang et al., 2023; Li et al., 2022; Brown et al., 2020; Piliouras and Yu, 2023) or *stateful* settings (Brown et al., 2020; Li and Wai, 2022; Ray et al., 2022; Izzo et al., 2022), where the distribution shifts are gradual. We refer the reader to Hardt and Mendler-Dünner (2023) for an extensive overview of the results related to performative prediction. In contrast to this line or work, we do not focus on prediction, but sequential decision making.

Performative Reinforcement Learning. the results in this paper are most related to recent works on performative RL (Mandal et al., 2023; Rank et al., 2024; Pollatos et al., 2025). As we already mentioned in the introduction, we extend the framework of Mandal et al. (2023) by considering linear MDPs. The framework of Rank et al. (2024) is orthogonal to ours as it considers tabular MDPs but also a different model of pefromativity, similar to the stateful performative prediction setting. Additionally, Cai et al. (2025) recently extended the performative prediction setting to dynamical systems. Their algorithm and analysis exploit linearity similar to ours.

Stochastic Games. The performative RL framework is also closely related to stochastic games (Shapley, 1953) and multi-agent RL (Zhang et al., 2021). More specifically, the framework relates to settings that consider commitment policies (Von Stackelberg, 2010; Letchford et al., 2012; Vorobeychik and Singh, 2012; Dimitrakakis et al., 2017; Zhong et al., 2021), in which a leader agent commits a policy to which a follower (or followers) respond by optimizing its utility function. Performative RL provides a more general abstraction, removing the need to prescribe a specific utility function to the follower.

Reinforcement Learning Theory. Our work is related to the literature on function approximation in reinforcement learning (Jin et al., 2020; Yang and Wang, 2020; Zhou et al., 2021; Du et al., 2021). We adopt a common modeling assumption about the structural properties of the environment, formalizing it as the linear Markov decision process (e.g., see Jin et al. (2020)). For finite sample setting, our approach is based on offline RL (Levine et al., 2020; Jin et al., 2021b) and we utilize the results of Gabbianelli et al. (2024) to establish finite sample guarantees. However, it is

important to note that offline RL does not consider a type of distribution shift that we study here. Namely, in offline RL, a deployed policy does not affect the environment.

Non-Stationarity in RL. More broadly, our work falls under the scope of non-stationary reinforcement learning (Cheung et al., 2020; Wei and Luo, 2021). However, the main difference is that the environment shift in our setting is induced by the policy, whereas existing work in non-stationary RL typically assumes policy-independent but bounded amount of shifts (Besbes et al., 2014).

2 MODEL

We consider Markov Decision Processes (MDPs) with a state space S, action set A, discount factor γ, and starting state distribution ρ. The reward and the probability transition functions of the MDP depend on the adopted policy. We consider infinite-horizon setting where the learner's goal is to maximize the total sum of discounted rewards. When the learner adopts policy π, the underlying MDP has reward function r_π and probability transition function P_π. We will write $M(\pi)$ to denote the corresponding MDP, i.e., $M(\pi) = (S, A, P_\pi, r_\pi, \rho)$.

When the learner adopts policy π and the underlying MDP is $M(\pi') = (S, A, P_{\pi'}, r_{\pi'}, \rho)$, $V^\pi_{\pi'}(\rho)$ denotes the value function, i.e., the expected sum of discounted rewards given the starting state distribution ρ. In particular, we will refer to $V^\pi_\pi(\rho)$ as the *performative value function* which can be interpreted as follows. The learner adopts policy π, in response the MDP changes to $M(\pi)$ and then $V^\pi_\pi(\rho)$ is the learner's value function in the new MDP $M(\pi)$. We next define the *performatively optimal policy*, which maximizes performative value function.

Definition 1 (Performatively Optimal Policy)**.** *A policy π_P is performatively optimal if it maximizes performative value function, i.e., $\pi_P \in \arg\max_{\pi'} V^{\pi'}_{\pi'}(\rho)$.*

Although, π_P maximizes the performative value function, it need not be stable, i.e., it need not be optimal with respect to the changed environment $M(\pi_P)$. We next define the notion of performatively stable policy which captures this notion of stability.

Definition 2 (Performatively Stable Policy)**.** *A policy π_S is* performatively stable *if it satisfies the condition $\pi_S \in \arg\max_{\pi'} V^{\pi'}_{\pi_S}(\rho)$.*

The definition of performatively stable policy implies that if the underlying MDP is $M(\pi_S)$ then an optimal policy is π_S. It is not a priori clear if a performatively stable policy always exists as the value function is a non-convex function of policies for most standard parametric representations including softmax and direct parameterization. And, for non-convex functions, it is not always possible to guarantee the existence of a fixed point. However, Mandal et al. (2023) observed that a performatively stable occupancy measure exists. In particular, given a policy π, let us define its long-term discounted state-action occupancy measure in the MDP $M(\pi)$ as $d^\pi(s,a) = \mathbb{E}_{\tau\sim\mathbb{P}^\pi_\pi}\left[\sum_{k=0}^\infty \gamma^k 1\{s_k = s, a_k = a\} \mid \rho\right]$. Given such an occupancy measure d, one can consider the following policy π^d.

$$\pi^d(a|s) = \begin{cases} \frac{d(s,a)}{\sum_b d(s,b)} & \text{if } \sum_b d(s,b) > 0 \\ \frac{1}{A} & \text{otherwise} \end{cases} \tag{1}$$

With this definition, we can pose the problem of finding a performatively stable occupancy measure. An occupancy measure d_S is performatively stable if it is the optimal solution of the following problem.

$$\begin{aligned} d_S \in \arg\max_{d:d\geq 0} & \sum_{s,a} d(s,a) r_d(s,a) \\ \text{s.t.} & \sum_a d(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P_d(s',a,s)\ \forall s \end{aligned} \tag{2}$$

With slight abuse of notation we are writing $r_d := r_{\pi_d}$ and $P_d := P_{\pi^d}$ (as defined in eq. (1)). Note that the objective is linear and the feasible region is a convex set. Therefore, fixed-point theorems can be used to show that a performatively stable occupancy measure d_S always exists.

Now suppose the learner has found a stable occupancy measure d_S and the corresponding MDP is $M_S = M(d_S) = M(\pi^{d_S})$. Then after deploying the policy π^{d_S} the resulting occupancy measure is d_S and the learner doesn't want to re-optimize.

Linear Markov Decision Process. We now introduce the definition of the linear Markov decision process (Jin et al., 2020). We assume known features $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^D$ where each state, action pair (s,a) is represented by the feature $\phi(s,a)$. Although the classic paper of Jin et al. (2020) consider infinite state-space, we will assume finite state-space and write $\Phi \in \mathbb{R}^{SA\times D}$ to denote the feature matrix. However, the state-space can be arbitrarily large compared to the dimension D of the features. We assume finite state-space but we believe our argument can be generalized to infinite state-space as well. In particular, observe that, one can always approximate a continuous and bounded state space by a finite state-space up to any desired accuracy and consider the feature matrix for that finite state-space.

In our setting, the MDP's are parameterized by the occupancy measure d. Given an occupancy measure d, let r_d be the corresponding reward function and P_d be the corresponding probability transition function. Then there exist unknown parameters $\theta_d \in \mathbb{R}^D$ and D-dimensional measure $\boldsymbol{\mu}_d = (\mu^1_d, \ldots, \mu^D_d)$ such that the following holds.

$$r_d(s,a) = \langle \phi(s,a), \theta_d \rangle \qquad P_d(s' \mid s,a) = \langle \phi(s,a), \mu_d(s') \rangle \tag{3}$$

We will assume that the parameters are bounded, i.e., $\|\theta_d\|_2 \leq \sqrt{D}$ and $\boldsymbol{\mu}_d(\mathcal{S}) \leq \sqrt{D}$ for any occupancy measure d. In matrix notations, we can write reward $r_d = \Phi\theta_d$.

Moreover, let $P_d \in \mathbb{R}^{S\times SA}$ be the probability transition matrix with entries $P_d(s';s,a) = P_d(s' \mid s,a)$. Then we have $P_d = \boldsymbol{\mu}_d\Phi^\top$. Substituting the expression of r_d and P_d in the optimization problem eq. (2) we get the following problem.

$$\begin{aligned}\max_{d:d\geq 0}\ & d^\top\Phi\theta_d \\ \text{s.t. } & Bd = \rho + \gamma\cdot\boldsymbol{\mu}_d\Phi^\top d\end{aligned} \tag{4}$$

Here the matrix $B \in \mathbb{R}^{S\times SA}$ is defined as follows.

$$B(s;(s',a')) = \begin{cases} 1 & \text{if } s' = s \\ 0 & \text{o.w.}\end{cases}$$

3 REPEATED RETRAINING

In order to obtain a stable policy, we first design a repeated optimization scheme. Let r_t (resp. P_t) be the reward (resp. transition probability) at iteration t. Under the assumption of linear MDP, we can equivalently assume that the relevant parameters at iteration t are θ_t and $\boldsymbol{\mu}_t$. Then we solve the following regularized optimization problem to obtain the new occupancy measure d_{t+1}.

$$\begin{aligned}\max_{d:d\geq 0}\ & d^\top\Phi\theta_t - \frac{\lambda}{2}d^\top\Phi\Phi^\top d \\ \text{s.t. } & Bd = \rho + \gamma\cdot\boldsymbol{\mu}_t\Phi^\top d\end{aligned} \tag{5}$$

Given an occupancy measure d, we will write $\theta_d = \mathcal{F}_\theta(d)$ to denote the D-dimensional parameter for the reward, and $\boldsymbol{\mu}_d = \mathcal{F}_\mu(d)$ to denote the D-dimensional measure $\boldsymbol{\mu}_d$. Moreover, given an occupancy measure d we deploy the policy π_d as defined in eq. (1).

In response to the deployed policy π_d, the underlying environment changes and generates parameters θ_d and $\boldsymbol{\mu}_d$. We will assume that if two occupancy measures generate the same policy (according to eq. (1)) then the corresponding parameters are also the same.

Assumption 1. *For any two occupancy measures d_1 and d_2 such that $\pi_{d_1} = \pi_{d_2}$ (as defined in eq. (1)), it follows that $\theta_{d_1} = \theta_{d_2}$ and $\boldsymbol{\mu}_{d_1} = \boldsymbol{\mu}_{d_2}$.*

The above assumption essentially says that the environment responds in terms of the induced policy π_d for an occupancy measure d, and if $d_1 \neq d_2$ but $\pi_{d_1} = \pi_{d_2} = \pi$ then the environment changes identically for these two cases. Algorithm 1 details the repeated optimization method. In order to prove convergence of Algorithm 1 we will make the following assumptions.

Assumption 2. *The mappings $(\mathcal{F}_\theta(\cdot), \mathcal{F}_\mu(\cdot))$ are $(\varepsilon_\theta, \varepsilon_\mu)$-sensitive i.e. the following holds for any two occupancy measures d and d'*

$$\begin{aligned}\left\|\mathcal{F}_\theta(d) - \mathcal{F}_\theta(d')\right\|_2 &\leq \varepsilon_\theta\left\|d - d'\right\|_2 \text{ and} \\ \left\|\mathcal{F}_\mu(d) - \mathcal{F}_\mu(d')\right\|_2 &\leq \varepsilon_\mu\left\|d - d'\right\|_2\end{aligned}$$

Algorithm 1 Repeated Optimization

Initialize occupancy measure d_0.
for $t = 1, \ldots$ **do**
  Obtain parameters $\theta_t = \mathcal{F}_\theta(d_{t-1})$ and $\boldsymbol{\mu}_t = \mathcal{F}_\mu(d_{t-1})$.
  Solve optimization problem eq. (5) to obtain occupancy measure d_t.
  Deploy policy π_t given as

$$\pi_t(a|s) = \begin{cases} \frac{d_t(s,a)}{\sum_b d_t(s,b)} & \text{if } \sum_b d_t(s,b) > 0 \\ \frac{1}{A} & \text{otherwise}\end{cases} \tag{6}$$

Assumption 3. *The matrix Φ has rank D and satisfies $\lambda_{\max}(\Phi^\top\Phi) \leq \mathcal{M}$ for some constant $\mathcal{M}$. Moreover, for any two valid occupancy measures d, d' (i.e., $d, d' \in \{d \in \mathbb{R}^{SA} : d \geq 0$ and $Bd = \rho + \gamma\mu_d\Phi^\top d\}$), we have $(d-d')^\top\Phi\Phi^\top(d-d') \geq \kappa\|d-d'\|_2^2$ for some $\kappa > 0$.*

Assumption 2 is standard in the literature on performative prediction, and states that the environment doesn't change too much if the occupancy measure of the deployed policy doesn't change much. The second assumption 3 has two parts. First, the feature matrix has full column rank and $\lambda_{\max}(\Phi^\top\Phi) \leq \mathcal{M}$. Suppose the rank of the feature matrix is strictly less than d, then it is possible to reduce the dimension of the features through orthogonalization. We require the assumption of bounded eigenvalue only because we prove convergence of occupancy measures in L_2 norm which is much stronger than the convergence in L_1 norm for an object of arbitrarily large dimension. Since $\|\Phi\|_1 \leq 1$, without the bounded eigenvalue assumption, we can prove convergence of occupancy measures in L_1 norm.

The second part of the assumption concerns the row-space of the feature matrix Φ, and says that for two different occupancy measures d, d' the L_2-norm of $\Phi^\top(d-d')$ is at least $\sqrt{\kappa}$ times the L_2-norm of $d-d'$. Without this assumption, there can be many occupancy measures d such that $\Phi^\top d = \Phi^\top d_S$ where d_S is the stable occupancy measure, and Algorithm 1 can converge to one such measure, and not necessarily to d_S.

Theorem 1. *Suppose assumptions 1, 2, 3 hold and $\alpha = \frac{\sqrt{\mathcal{M}}}{\sqrt{A}(1-\gamma)}$ and $\varepsilon_\mu < \frac{2\sqrt{\kappa}}{25\gamma\alpha^2}$. If Algorithm 1 is run with regularization parameter $\lambda > \frac{25\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)}{8\sqrt{\kappa}}$ then we are guaranteed that*

$$\|d_t - d_S\|_2 \leq \delta \quad \forall t \geq \ln\left(\tfrac{2}{\delta(1-\gamma)}\right)/\ln(1/r),$$

where $r = \frac{5}{4}\sqrt{\frac{\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu}{\lambda\sqrt{\kappa}} + \frac{4\gamma\varepsilon_\mu\alpha^2}{\sqrt{\kappa}}} < 1.$

The full proof is provided in Appendix A. Here we discuss the main challenges and the differences with the approach taken in Mandal et al. (2023). We consider the dual formulation of the optimization problem defined in eq. (5). For

the tabular setting, prior approach showed that the sequence of dual optimal solutions converge to a stable solution, and used the convergence of the dual optimal solutions to establish convergence of the sequence of primal optimal solutions $\{d_t\}_{t\geq 1}$. However, for our setting, the dual problem need not be strongly-convex and we cannot use the same proof strategy. We first use assumption 3 to establish the following recurrence relation.

$$\|d_{t+1} - d_S\|_2 \leq \frac{1}{\lambda\sqrt{\kappa}}\left(\varepsilon_\theta \|d_t - d_S\|_2 + \|u_{t+1} - u_S\|_2\right) \quad (7)$$

Here, u_t is a linear combination of dual optimal solution (g_t, h_t), i.e., $u_t = M_t h_t + W g_t$ (similarly $u_S = M_s h_S + W g_S$). Then we focus on bounding the difference $\|u_{t+1} - u_S\|_2$. We use first-order optimality conditions of the dual problem at (g_t, h_t), and assumption 2 to establish the following bound.

$$\|u_{t+1} - u_S\|_2 \leq \left(\varepsilon_\theta + \gamma\varepsilon_\mu \|h_{t+1}\|_2 + 3\lambda\gamma\varepsilon_\mu \left\|M_t^\dagger\right\|_2^2\right)\|d_{t-1} - d_S\|_2 + \gamma\varepsilon_\mu \|h_{t+1}\|_2 \|d_t - d_S\|_2$$

We then use a bound on the dual optimal solution h_{t+1}(Lemma 4), and assumption 3 (i.e., bound on the norm of $M_t^\dagger$) to establish a bound of the form $\|u_{t+1} - u_S\|_2 \leq \alpha_1 \|d_{t-1} - d_S\|_2 + \alpha_2 \|d_t - d_S\|_2$. Substituting this bound in eq. (7), we can establish the following recurrence relation: $\|d_{t+1} - d_S\|_2 \leq \beta_1/\lambda \|d_t - d_S\|_2 + \beta_2/\lambda \|d_{t-1} - d_S\|_2$ for some constants β_1, β_2. Finally, by choosing an appropriate value of λ, we can ensure that the sequence $\{d_t\}_{t\geq 1}$ is a contraction.

We obtain a linear convergence rate similar to the result of Mandal et al. (2023), i.e., the number of iterations required to obtain a δ-approximate stable occupancy measure is $O(\log(1/\delta))$. Moreover, when instantiated to the tabular setting (i.e., $D = SA$) the required level of regularization is $\lambda = O\left(\varepsilon_\theta + \gamma\frac{S\sqrt{A}}{1-\gamma}\varepsilon_\mu\right)$ for $\kappa = O(1)$ and $\mathcal{M} = SA$. This is an improvement by a factor of $S^{3/2}/(1-\gamma)^3$ compared to prior work. In order to understand why the strength of regularization is important, we next present two results that show why the required value of λ controls the approximation with respect to the unregularized objective.

3.1 Approximating the Unregularized Objective

Theorem 1 proves that repeatedly optimizing the regularized objective of eq. (5) converges to a stable solution (say d_S^λ). We can show that this stable solution also approximates the performatively stable and optimal solution with respect to the original unregularized objective. Because of limited space, here we provide informal statements of the claims, and full details are provided in Appendix B and Appendix C. We will require the following definition of an approximately stable policy.

Definition 3. *An occupancy measure d_S is β-approximately stable if*

$$d_S^\top r_S \geq \max_{d\in C(P_S)} d^\top r_S - \beta.$$

Here $r_S = r_{d_S}$ (resp. $P_S = P_{d_S}$) is the reward (probability transition) induced by the policy π^{d_S}, and $C(P_S)$ is the set of valid occupancy measures with respect to P_S.

The next theorem states that the stable occupancy measure according to the regularized objective (d_S^λ) is approximately stable with respect to the unregularized objective.

Theorem 2. *Suppose the assumptions of Theorem 1 hold. Then there exists a choice of regularization parameter (λ) such that repeatedly optimizing objective define in eq. (5) converges to a stable solution d_S^λ that is $\frac{25\mathcal{M}\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)}{16\sqrt{\kappa}(1-\gamma)^2}$-approximately stable with respect to the unregularized objective.*

Note that as the performativity vanishes, i.e., $\varepsilon_\theta, \varepsilon_\mu \to 0$, the solution d_S^λ converges to the performatively stable solution with respect to the unregularized objective. We next turn to bounding the approximation with respect to the performatively optimal policy. Let d_{PO} be the performatively optimal occupancy measure. With a slight abuse of notation, we will write $V_{d_{\text{PO}}}^{d_{\text{PO}}}(\rho) = V_{\pi^{d_{\text{PO}}}}^{\pi^{d_{\text{PO}}}}(\rho)$ to denote the value of the policy $\pi^{d_{\text{PO}}}$ in the MDP induced by $\pi^{d_{\text{PO}}}$.

Theorem 3 (Informal Statement)**.** *Suppose the assumptions of Theorem 1 hold, and $\Delta = \frac{3\gamma\varepsilon_\mu\mathcal{M}\sqrt{D}}{(1-\gamma)^2} + \varepsilon_\theta\sqrt{\mathcal{M}}$, and $\lambda_0 = \frac{25}{8\sqrt{\kappa}}\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)$. Then there exists a choice of regularization parameter (λ) such that repeatedly optimizing objective (5) converges to a solution d_S^λ with the guarantee that*

$$V_{d_{\text{PO}}}^{d_{\text{PO}}}(\rho) - V_{d_S^\lambda}^{d_S^\lambda}(\rho) \leq O\left(\max\{1, \Delta\}\Delta + \lambda_0\right).$$

The full statement of the theorem requires introducing new definitions. Hence we provide the formal statement and other details in the appendix Appendix C. The main step of the proof is to show the distance between the performatively optimal solution and performatively stable solution can be bounded by $O(\Delta/\kappa\lambda)$ (lemma 5). Note that the above theorem bounds the gap in performative value function under d_{PO} and d_S^λ. As $\varepsilon_\theta, \varepsilon_\mu \to 0$ both Δ, λ_0 approach zero, the suboptimality gap approaches zero.

4 FINITE SAMPLE SETTING

In the previous section, we assumed the learner has full access to the new model (P_t, r_t) at every iteration t. In this section, we relax this assumption and consider a setting where the learner deploys policy π_t in MDP M_t and only accesses samples through this deployed policy. In particular, we assume the following data generation process at iteration t.

Data: Given the occupancy measure d_t, let $\widetilde{d}_t$ be the normalized occupancy measure defined as $\widetilde{d}_t(s,a) = (1-\gamma)d_t(s,a)$ for any state, action pair (s,a). For each $j \in [m_t]$, we first sample a starting state $s_j^0 \sim \rho(\cdot)$, then sample $(s_j, a_j) \sim \widetilde{d}_t(\cdot)$. Finally, the next state $s_j' \sim P_t(\cdot \mid s_j, a_j)$ and reward $r_j \sim r_t(s_j, a_j)$. Therefore, for each $j \in [m_t]$, we have the tuple $(s_j^0, s_j, a_j, r_j, s_j')$ and the collection of m_t such tuples constitute the dataset $\mathcal{D}_t$ at iteration t.

We will follow an approach similar to Mandal et al. (2023) by constructing the Lagrangian corresponding to the optimization problem Equation (5), and then solving for a saddle point of the Lagrangian. In order for this approach to work, we need to show that the empirical version of the Lagrangian is close to the true Lagrangian through an ε-net construction. However, the standard Lagrangian of Equation (5) has infinite dimensional variables, and the size of any ε-net is unbounded, and hence the previous argument cannot be applied. Therefore, we introduce a reparametrization of the primal variable d by introducing a finite-dimensional variable. Let us rewrite the optimization problem Equation (5) by introducing the variable $\nu = \Phi^\top d$. Note that $\nu \in \mathbb{R}^D$ whereas d can be infinite-dimensional.

$$\begin{aligned} \max_{\nu, d \geq 0} \; & \nu^\top \theta_t - \frac{\lambda}{2}\nu^\top \nu \\ \text{s.t. } & Bd = \rho + \gamma \cdot \boldsymbol{\mu}_t \nu \\ & \nu = \Phi^\top d \end{aligned} \tag{8}$$

The Lagrangian of the optimization problem 8 is

$$\begin{aligned} \mathcal{L}(d, \nu; g, \omega) = \nu^\top \theta_t - \frac{\lambda}{2}\nu^\top \nu + \langle g, \rho + \gamma \cdot \boldsymbol{\mu}_t \nu - Bd \rangle \\ + \left\langle \omega, \Phi^\top d - \nu \right\rangle \end{aligned}$$

We now introduce an empirical version of the Lagrangian which can be estimated from the data. Let us write $\Sigma_t = \mathbb{E}_{(s,a)\sim\pi_t}[\phi(s,a)\phi(s,a)^\top]$ as the expected covariance matrix of the features observed under policy π_t in the model M_t. Then it is possible to derive the following equivalent expression of the Lagrangian $\mathcal{L}(\cdot)$ (the proof is in Appendix E.1).

$$\begin{aligned} \mathcal{L}(d, \nu; g, \omega) &= \nu^\top \left(\theta_t + \gamma \cdot \boldsymbol{\mu}_t^\top g - \omega\right) - \frac{\lambda}{2}\nu^\top \nu \\ &\quad + \langle g, \rho \rangle + \left\langle d, \Phi\omega - B^\top g \right\rangle \\ &= \nu^\top \Sigma_t^{-1} \mathop{\mathbb{E}}_{\substack{(s,a)\sim\pi_t \\ s'\sim P_t(\cdot|s,a)}} \big[\phi(s,a) r_t(s,a) + \gamma \cdot g(s') \\ &\quad - \phi(s,a)\phi(s,a)^\top \omega\big] - \frac{\lambda}{2}\nu^\top \nu \\ &\quad + \mathop{\mathbb{E}}_{s^0\sim\rho}\left[g(s^0)\right] + \left\langle d, \Phi\omega - B^\top g \right\rangle \end{aligned} \tag{9}$$

This motivates the following choice of the empirical Lagrangian.

$$\begin{aligned} \widehat{\mathcal{L}}_t(d, \nu; g, \omega) = \nu^\top \Sigma_t^{-1} \cdot \frac{1}{m_t} \sum_{j=1}^{m_t} \phi(s_j, a_j) \Big(r_t(s_j, a_j) + \gamma \cdot g(s_j') \\ - \phi(s_j, a_j)^\top \omega \Big) - \frac{\lambda}{2}\nu^\top \nu + \frac{1}{m_t} \sum_{j=1}^{m_t} g(s_j^0) + \left\langle d, \Phi\omega - B^\top g \right\rangle \end{aligned} \tag{10}$$

Algorithm 2 depicts our approach for the finite sample setting, which repeatedly solves a saddle point of the empirical Lagrangian defined in Equation (10). Since this is an infinite-dimensional optimization problem, it is not obvious that the problem can be solved. We first assume that we can exactly solve the optimization problem, and provide convergence guarantees for Algorithm 2. In the next subsection, we will show how to efficiently solve the saddle point optimization problem. We will make the following assumption regarding the deployed policy.

Algorithm 2 Repeated Optimization from Finite Samples

Initialize initial policy π_0.
for $t = 1, \ldots$ **do**
  Deploy policy π_t and collect dataset $\mathcal{D}_t$ of size m_t.
  Solve the min-max problem.

$$(\widehat{d}_t, \widehat{\nu}_t; \widehat{g}_t, \widehat{\omega}_t) \leftarrow \max_{d,\nu\in\mathbb{R}^D} \min_{g,\omega\in\mathbb{R}^D} \widehat{\mathcal{L}}_t(d, \nu; g, \omega) \tag{11}$$

  Set policy π_{t+1} as

$$\pi_{t+1}(a|s) = \begin{cases} \frac{\widehat{d}_t(s,a)}{\sum_b \widehat{d}_t(s,b)} & \text{if } \sum_b \widehat{d}_t(s,b) > 0 \\ \frac{1}{A} & \text{otherwise} \end{cases} \tag{12}$$

Assumption 4 (Bounded Coverage Ratio)**.** *For any policy π, let d_π be the occupancy measure of the policy π in the MDP M_π, and let $\Sigma_\pi = \mathbb{E}_{(s,a)\sim d_\pi}[\phi(s,a)\phi(s,a)^\top]$. Moreover, let $d_\star$ be the occupancy measure of the optimal policy in M_π[1]. Then, there exists a constant $B > 0$ such that*

$$\mathbb{E}_{(s,a)\sim d_\star}[\phi(s,a)^\top]\, \Sigma_\pi^{-2}\, \mathbb{E}_{(s,a)\sim d_\star}[\phi(s,a)] \leq B.$$

Assumption 4 generalizes single-policy concentrability assumption popular in the literature on offline reinforcement learning Jin et al. (2021b). In offline RL, it is assumed that there is an overlap between the data generating policy and the optimal policy. Without this assumption, even with an infinite amount of data, no information about the optimal policy can be obtained. Consequently, the best policy that can be learned from the dataset will always be suboptimal. Note that, our algorithm proceeds in iterations, and at iteration t, the learner deploys a policy π_t, collects a dataset

[1]This is the standard optimal policy in the MDP M_π and not the performatively optimal policy.

D_t from the new MDP $M(\pi_t)$ and learns an optimal policy in the new environment. In order to approximate the optimal policy in the new environment there must be an overlap between the deployed policy π_t and $\pi_t^\star$ the policy that is optimal in the new MDP $M(\pi_t)$. This is precisely what assumption 4 states.

Theorem 4. *Suppose assumptions 2, 3, and 4 hold and* $\alpha = \frac{\sqrt{M}}{\sqrt{A}(1-\gamma)}$, $c_1 = \min\left\{1, \gamma^2 \cdot \sigma^\star_{\min}(\mu_d \mu_d^\top)\right\}$[2], *and* $\varepsilon_\mu < \frac{\sqrt{\kappa}}{24\gamma\alpha^2}$. *If Algorithm 2 is run with* $\lambda > \frac{6\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)}{\sqrt{\kappa B}}$ *and the number of samples* $m_t = O\left(\frac{D^5 B^2 \lambda^4}{(1-\gamma)^2 c_1^4 \delta^4 \kappa} \log \frac{DB\lambda t}{c_1 \delta \kappa p}\right)$, *then with probability at least* $1-p$,

$$\|\widehat{d}_t - d_S\|_2 \le \delta \quad \forall t \ge \ln\left(\tfrac{2}{\delta(1-\gamma)}\right) / \ln(1/r),$$

with r as defined in Theorem 1.

Suppose $B, \kappa, c_1 = O(1)$. Then Theorem 4 shows that the required number of samples $m_t = \widetilde{O}\left(\frac{D^5}{(1-\gamma)^2\delta^4}\right)$ which is independent of the number of states and depends polynomially on the dimension of the features. We can apply our algorithm to the tabular setting by instantiating $D = |\mathcal{S}|\,|\mathcal{A}|$. This gives a sample complexity of $\widetilde{O}\left(\frac{S^5A^5}{(1-\gamma)^2\delta^4}\right)$ for obtaining a δ-approximate stable solution. On the other hand, Mandal et al. (2023) has a sample complexity of $\widetilde{O}\left(\frac{S^6A^2}{(1-\gamma)^4\delta^4}\right)$. Both approaches solve the empirical Lagrangian for the finite sample setting. Below we provide a brief sketch of the proof.

We first show that the saddle point of the Lagrangian has bounded norm. In particular, if $(d^\star, \nu^\star, g^\star, \omega^\star) \leftarrow \max_{d,\nu} \min_{g,\omega} \mathcal{L}_t(d, \nu; g, \omega)$ then $\left\|\Sigma_t^{-1}\nu^\star\right\|_2 \le \sqrt{B}$, $\left\|g^\star\right\|_2 + \left\|\omega^\star\right\|_2 \le \frac{\lambda+2D}{c_1}$, and $\left\|d^\star\right\|_2 \le \frac{1}{1-\gamma}$. This allows us to consider an equivalent max-min optimization problem where the norms of the variables are bounded.

The next step is to show that the empirical Lagrangian is close to the true Lagrangian so that a saddle point of the latter is also an approximate saddle point of the true Lagrangian. Since the variables ν and ω are D-dimensional and bounded norms, we can construct an ε-net and use union bound to show that the approximate Lagrangian $\widehat{\mathcal{L}}_t(d, \nu, g, \omega)$ is close to the true Lagrangian $\mathcal{L}_t(d, \nu, g, \omega)$ for any ν and ω. However, the same argument does not apply to the variable g as it is infinite-dimensional. In fact, we first prove

$$\left|\widehat{\mathcal{L}}_t(d, \nu, \widetilde{g}, \omega) - \mathcal{L}_t(d, \nu, \widetilde{g}, \omega)\right| \le \varepsilon$$

holds for any d, ν, ω, and $\widetilde{g} \in \arg\min_g \widehat{\mathcal{L}}_t(d, \nu, g, \omega)$. This allows us to use ε-net construction only for the set of possible minimizers of $\widehat{\mathcal{L}}_t(d, \nu, \cdot, \omega)$. Using this result, Lemma 8 shows that a saddle point $(\widehat{d}, \widehat{\nu}, \widehat{g}, \widehat{\omega})$ of the empirical Lagrangian is an approximate saddle point of the true Lagrangian in the sense that $\mathcal{L}(d_t^\star, \nu_t^\star; g_t^\star, \omega_t^\star) - \mathcal{L}(\widehat{d}_t, \widehat{\nu}_t; g_t^\star, \omega_t^\star) \le 2\varepsilon$. This allows us to show that the occupancy measures $d_t^\star$ and $\widehat{d}_t$ are close i.e. $\|d_t^\star - \widehat{d}_t\|_2 \le O\left(\sqrt{\frac{\varepsilon}{\lambda\kappa}} + \frac{\varepsilon}{\sqrt{\kappa}}\right)$. The rest of the proof uses results of Theorem 1 which shows that the sequence $d_t^\star$ converges to d_S with the right value of λ, and hence the sequence $\widehat{d}_t$ also converges to d_S.

4.1 Solving the Empirical Lagrangian

Algorithm 2 finds an approximate stable policy in the finite samples setting, but it assumes that the empirical Lagrangian problem (as defined in Equation (10)) is solvable. In this subsection, we construct an efficient algorithm to obtain a saddle point of the empirical Lagrangian. Our algorithm is motivated by the recent work of Gabbianelli et al. (2024) who designed a primal-dual algorithm for solving offline RL problem with linear MDPs. In our setting, the Lagrangian is strongly concave in ν, but linear in other variables, and we update their method accordingly.

Algorithm 3 describes a method to obtain an approximate saddle point of the empirical Lagrangian Equation (10). A standard strategy to solve such an optimization problem would be to perform alternate gradient descent in variables g, ω and ascent in variables d, ν. However, note that the variables d and g are infinite-dimensional, and their gradients cannot be explicitly computed. Hence we take gradient descent steps with respect to ω and ascent steps with respect to ν, and represent d and g using ν and ω. Furthermore, the objective is strongly concave in ν and we can replace the gradient ascent step in ν by a single update step. The algorithm runs for T iterations and at each iteration, it takes K gradient steps for ω, and one update step for ν. The gradient with respect to ω is given as follows.

$$\nabla_\omega \widehat{\mathcal{L}}_t(d, \nu; g, \omega) = \Phi^\top d - \frac{1}{m_t}\sum_{j=1}^{m_t} \phi(s_j, a_j)\phi(s_j, a_j)^\top \Sigma_t^{-1}\nu$$

The gradient with respect to ν is given as follows.

$$\nabla_\nu \widehat{\mathcal{L}}_t(d, \nu; g, \omega) = \Sigma_t^{-1} \cdot \frac{1}{m_t}\sum_{j=1}^{m_t} \phi(s_j, a_j)\left(r_t(s_j, a_j) + \gamma \cdot g(s_j') - \phi(s_j, a_j)^\top\omega\right) - \lambda \cdot \nu$$

Setting $\nabla_\nu \widehat{\mathcal{L}}_t(d, \nu; g, \omega) = 0$, we get the update step for the variable ν (line 5, Algorithm 3). Note that, computing the gradient requires the knowledge of d and g. We maintain d and g as a function of the parameters ν and ω. In particular, at iteration t, define policy π_t as

$$\pi_t(a \mid s) = \sigma\left(\eta_\pi \cdot \sum_{j<t} \phi(s, a)^\top \omega_j\right)$$

[2] $\sigma^\star_{\min}(A)$ is the smallest positive eignevalue of a symmetric matrix A.

Algorithm 3 Offline Regularized Primal-Dual

Input: (a) Dataset $\mathcal{D}$, and and (b) Number of iterations T_{inner} and K.

Set $\mathcal{W} = \{\omega : \|\omega\|_2 \le \frac{2D}{1-\gamma}\}$, $\mathcal{V} = \{\nu : \|\nu\|_2 \le D\sqrt{B}\}$, $\eta_\omega = \frac{D\sqrt{B}}{\sqrt{K(B+(1-\gamma)^{-2})}}$, $\eta_\pi = \sqrt{\frac{\log A}{T}}\frac{1-\gamma}{D}$.

Initialize $\omega_0 \in \mathcal{W}$, $\nu_0 \in \mathcal{V}$ and π_0.

for $\ell = 1, \dots, T_{\text{inner}}$ **do**

/* Take a gradient step for ω_ℓ */

Initialize $\omega_{\ell,0} = \omega_{\ell-1}$.

for $k = 0, \dots, K-1$ **do**

Obtain sample $(s^0_{\ell,k}, s_{\ell,k}, a_{\ell,k}, r_{\ell,k}, s'_{\ell,k})$ uniformly at random from $\mathcal{D}$.

$d_{\ell,k}(s,a) = \pi_\ell(a|s) \cdot (\nu_{\ell-1}^\top \Sigma^{-1}\phi(s_{\ell,k}, a_{\ell,k})\mathbf{1}\{s = s'_{\ell,k}\} + \mathbf{1}\{s = s^0_{\ell,k}\})$.

$\widetilde{g}_{\omega_{\ell,k}} = \Phi^\top d_{\ell-1,k} - \phi(s_{\ell,k}, a_{\ell,k})\phi(s_{\ell,k}, a_{\ell,k})^\top \Sigma^{-1}\nu_{\ell-1}$.

$\omega_{\ell,k+1} \leftarrow \text{Proj}_{\mathcal{W}}(\omega_{\ell,k} - \eta_\omega \cdot \widetilde{g}_{w_{\ell,k}})$.

Set $\omega_\ell = \frac{1}{K}\sum_{k=1}^K \omega_{\ell,k}$.

/* Update variable ν_ℓ and π_ℓ */

$$\nu_\ell \leftarrow \frac{1}{\lambda}\Sigma^{-1} \cdot \frac{1}{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{D}|}\phi(s_j, a_j)\Big(r_j + \gamma \cdot g^{\pi_{\ell-1},\omega_{\ell-1}}(s'_j) - \phi(s_j, a_j)^\top \omega_{\ell-1}\Big).$$

$$\pi_\ell(a \mid s) = \sigma\Big(\eta_\pi \cdot \textstyle\sum_{j=1}^{\ell}\phi(s,a)^\top \omega_{\ell-1}\Big).$$

Return π_j where $j \sim \text{uniform}([T])$.

where $\sigma(\cdot)$ is the sigmoid function, and η_π is a constant. Then we choose g_t as g^{π_t,ω_t} with

$$g^{\pi_t,\omega_t}(s,a) = \sum_a \pi_t(a \mid s)\phi(s,a)^\top \omega_t.$$

And, we choose d_t as d^{π_t,ν_t} with

$$d^{\pi_t,\nu_t}(s,a) = \pi_t(a \mid s) \cdot \Big(\rho(s) + \gamma \cdot \boldsymbol{\mu}(s)^\top \nu_t\Big).$$

Since the measure $\boldsymbol{\mu}$ is unknown we cannot use this representation of the occupancy measure, and instead build an estimator of d^{π_t,ν_t}. Given a tuple (s^0, s, a, r, s') let

$$\widehat{d}^{\pi_t,\nu_t}(\widetilde{s},\widetilde{a}) = \pi_t(\widetilde{a} \mid \widetilde{s}) \cdot \Big(\mathbf{1}\{\widetilde{s} = s^0\} + \nu_t^\top \Sigma_t^{-1}\phi(s,a)\mathbf{1}\{\widetilde{s} = s'\}\Big).$$

It can be easily checked that $\mathbb{E}\Big[\widehat{d}^{\pi_t,\nu_t}(s,a)\Big] = d^{\pi_t,\nu_t}(s,a)$. The next theorem proves that the policy returned by Algorithm 3 is approximately optimal with respect to the regularized objective.

Theorem 5. *Suppose Algorithm 3 is run with* $K \ge \frac{144D^2B}{\varepsilon^2}\left(B + (1-\gamma)^{-2}\right)$ *and* $T \ge \frac{576D^2}{\varepsilon^2} \cdot \frac{\log A}{(1-\gamma)^2}$. *Let* $\widetilde{\nu}$ *be the average feature of a policy selected uniformly at random from* $\{\pi_t\}_{t\in[T]}$. *Then for a policy* $\pi^\star$ *with* $\nu^\star = \Phi^\top d^{\pi^\star}$ *we have,*

$$\langle \widetilde{\nu}, \theta_t\rangle - \tfrac{\lambda}{2}\|\widetilde{\nu}\|_2^2 \ge \nu^{\star\top}\theta_t - \tfrac{\lambda}{2}\left\|\nu^\star\right\|_2^2 - \varepsilon.$$

The proof follows an argument similar to the proof of theorem 5.3 from Gabbianelli et al. (2024). We upper bound the sub-optimality gap in the regularized objective by a dynamic regret, and then Lemma 12 in the appendix, first expresses the regret in terms of regret in ω, ν, and π and provides an upper bound on each term separately. As an immediate corollary of Theorem 5 we can show that the policy returned by Algorithm 3 is an approximate saddle point of the empirical Lagrangian, which is required by Algorithm 2.

Corollary 1. *Under the same setting as in Theorem 5, let the number of samples* $m_t \ge O\left(\frac{D^5B\lambda^4}{(1-\gamma)^2c_1^4\varepsilon^2}\log\frac{DB\lambda}{c_1\varepsilon p}\right)$ *where* $c_1 = \min\left\{1, \gamma^2 \cdot \sigma^\star_{\min}(\boldsymbol{\mu}_d\boldsymbol{\mu}_d^\top)\right\}$. *Then there exists* $\widetilde{g}, \widetilde{\omega}$ *so that with probability at least* $1-p$,

$$\max_{d,\nu}\widehat{\mathcal{L}}_t(d,\nu;\widetilde{g},\widetilde{\omega}) - 2\varepsilon \le \widehat{\mathcal{L}}_t(\widetilde{d},\widetilde{\nu};\widetilde{g},\widetilde{\omega}) \le \min_{g,\omega}\widehat{\mathcal{L}}_t(\widetilde{d},\widetilde{\nu};g,\omega) + 2\varepsilon.$$

The next result shows that we can use algorithm 3 as an approximate oracle to solve the saddle point optimization problem (Equation (11)) in Algorithm 2 and obtain convergence to performatively stable solution by choosing an appropriate value of the regularization parameter λ.

Corollary 2. *Under the same setting as in Theorem 4, suppose* $\lambda > \frac{18(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu}{\sqrt{\kappa B}}$. *If Algorithm 2 uses Algorithm 3 every iteration to approximately solve the saddle point optimization Equation (11), then with probability at least* $1-p$,

$$\|\widetilde{d}_t - d_S\|_2 \le \delta \quad \forall t \ge \ln\left(\tfrac{2}{\delta(1-\gamma)}\right)/\ln(1/r),$$

with r *as defined in Theorem 1.*

5 APPLICATIONS

We provide several applications of performative RL involving stochastic games.

Stochastic Stackelberg Game: Consider two RL agents (1 and 2) interacting in an MDP with a shared state-space S. The set of actions available to agent 1 (resp. 2) is A_1 (resp. A_2). The transitions and rewards are determined by the actions of both the players, i.e., there exist reward functions $r_1, r_2 : S \times A_1 \times A_2 \to \mathbb{R}$, and transition $P : S \times A_1 \times A_2 \to \Delta(S)$.

We adopt the framework of Stackelberg game (Von Stackelberg, 2010) where agent 1 is the *leader* and agent 2 is the *follower*. The first agent deploys a stationary policy (say π_1) to maximize her reward, and in response, the second agent adopts a policy π_2 that is obtained through Boltzmann softmax operator with temperature parameter β. To be precise, given policy π_1 of agent 1, let the modified MDP be $(S, A_2, \overline{r}_2, \overline{P}_2, \gamma)$ where $\overline{r}_2(s,a_2) = \sum_{a_1}\pi_1(a \mid s)r_2(s,a_1,a_2)$ and $\overline{P}_2(s' \mid s,a_2) = \sum_{a_1}\pi_1(a \mid s)P(s' \mid s,a_1,a_2)$. Now let $Q_2^\star : S \times A_2 \to \mathbb{R}$ be the optimal state,action Q-function

in this new MDP. Then the policy adopted by the second agent is $\pi_2(a \mid s) = \frac{\exp(\beta Q_2^\star(s,a))}{\sum_b \exp(\beta Q_2^\star(s,b))}$.

Therefore, for a policy π_1 by the agent 1, the second agent updates her policy, and from the perspective of the first agent the underlying MDP changes to $M(\pi_1)$. The next lemma bounds the sensitivity of the reward, and probability transition functions.

Lemma 1. *Suppose* $\|\pi_1(\cdot \mid s) - \widetilde{\pi}_1(\cdot \mid s)\|_1 \le \delta$, *and* $R = \max_i \max_{s,a} r_i(s,a)$. *Then,* $\forall s, a, s'$

$$\left|r_{\pi_1}(s,a) - r_{\widetilde{\pi}_1}(s,a)\right| \le \delta \cdot \frac{2\sqrt{2}\beta A_1 A_2^{3/2} R^2}{(1-\gamma)^2} \quad \textit{and}$$

$$\left|P_{\pi_1}(s'|s,a) - P_{\widetilde{\pi}_1}(s'|s,a)\right| \le \delta \cdot \frac{2\sqrt{2} A_1 A_2^{3/2} \beta R}{(1-\gamma)^2}.$$

If agent 1 assumes that the MDP is linear, then Lemma 1 implies that the sensitivity parameters are $\varepsilon_\theta = \frac{2\sqrt{2}\beta A_1 A_2^{3/2} r_{\max}^2}{(1-\gamma)^2}$ and $\varepsilon_\mu = \frac{2\sqrt{2}\beta A_1 A_2^{3/2} r_{\max}}{(1-\gamma)^2}$. Therefore, according to Theorem 1, if agent 1 repeatedly optimizes a regularized objective with appropriate λ, they would converge to a performatively stable policy.

Multiple Followers: We now generalize the previous two-player stochastic Stackelberg game to a setting where there are m followers. Now the reward of the j-th agent is $r_j : S \times \prod_{i=1}^{m+1} A_i \to \mathbb{R}$ and the probability transition function is given as $P : S \times \prod_{i=1}^{m+1} A_i \to \Delta(S)$. As before, agent 1 is the *leader* and adopts a stationary policy (say π_1). This induces a stochastic game $\overline{\mathcal{M}} = (S, \{A_i\}_{i=2}^{m+1}, \overline{P}, \{\overline{r}_i\}_{i=2}^{m+1}, \gamma)$ among the m follower agents where $\overline{r}_i(s, \boldsymbol{a}) = \sum_{a_1} \pi_1(a_1 \mid s) r_i(s, a_1, \boldsymbol{a})$ and $\overline{P}(s' \mid s, \boldsymbol{a}) = \sum_{a_1} \pi_1(a_1 \mid s) P(s' \mid s, a_1, \boldsymbol{a})$.

We assume that the m follower agents play a policy that is an approximation of the optimal coarse correlated equilibrium (CCE) obtained through the Boltzmann softmax operator with temperature parameter β. In particular, let $\pi_f^\star : S \to \Delta\left(\prod_{i=2}^{m+1} A_i\right)$ be the CCE in the game $\overline{\mathcal{M}}$ that maximizes the welfare, i.e., the sum of expected rewards of the m followers. Additionally, let $Q_f^\star(s, \boldsymbol{a}) = \sum_{i=2}^{m+1} \mathbb{E}_{\pi_f^\star}\left[\sum_{t=0}^{\infty} \gamma^t \overline{r}_i(s_t, \boldsymbol{a}_t) \mid s, \boldsymbol{a}\right]$ be the state-action welfare function under the policy $\pi_f^\star$. Then the m followers adopt a policy π_f that is given as $\pi_f(\boldsymbol{a} \mid s) = \frac{\exp(\beta Q_f^\star(s,\boldsymbol{a}))}{\sum_{\boldsymbol{b}} \exp(\beta Q_f^\star(s,\boldsymbol{b}))}$. Then the following bound holds.

Lemma 2. *Let* $\|\pi_1(\cdot|s) - \widetilde{\pi}_1(\cdot|s)\|_1 \le \delta$, $A = \max_i A_i$, *and* $R = \max_{i,s,\boldsymbol{a}} r_i(s, \boldsymbol{a})$. *Then,* $\forall s, a$

$$\left|r_{\pi_1}(s,a) - r_{\widetilde{\pi}_1}(s,a)\right| \le \delta \cdot \frac{3\sqrt{2}\beta m A^{3m/2+1} R^2}{(1-\gamma)^2} \quad \textit{and}$$

$$\left|P_{\pi_1}(s'|s,a) - P_{\widetilde{\pi}_1}(s'|s,a)\right| \le \delta \cdot \frac{3\sqrt{2}\beta m A^{3m/2+1} R}{(1-\gamma)^2}$$

Now, the sensitivity parameters grow exponentially with the number of followers. This is unavoidable for a general stochastic game, but can be significantly reduced for succinct games (e.g., (Kearns et al., 2013)).

6 CONCLUSION

In this work, we provide computationally and statistically efficient algorithms for performative RL with large-scale MDPs. Our work is centered around the linear MDP model, and it would be interesting to generalize our approach to nonlinear function approximation (e.g., (Wang et al., 2020)). However, there are two challenges with general function approximation. First, a performatively stable policy might not exist since the value function is non-convex in both policy and occupancy measures. Therefore, we might have to resort to a locally stable solution (Li and Wai, 2024). Second, obtaining *last-iterate* convergence in non-convex performative prediction is difficult, and to the best of our knowledge, the existing result provides *best-iterate* convergence. In this work, we have provided *last-iterate* convergence for linear MDPs, and obtaining similar convergence for general MDPs is significantly harder.

It could also be exciting to consider the effects of multiple learners in performative RL by imposing a specific structure on the underlying game so that independent repeated optimization still converges (Piliouras and Yu, 2023). Finally, we briefly study the approximation of performatively optimal policy, and exploring the design of such policies is another avenue of future research.

Acknowledgements

The work of Goran Radanovic was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Afsar, M. M., Crump, T., and Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

Brown, G., Hod, S., and Kalemaj, I. (2020). Performative prediction in a stateful world. *arXiv preprint arXiv:2011.03885*.

Brown, N. and Sandholm, T. (2019). Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890.

Cai, S., Han, F., and Cao, X. (2025). Performative control for linear dynamical systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2020). Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *Interna-*

tional Conference on Machine Learning, pages 1843–1854. PMLR.

Dimitrakakis, C., Parkes, D. C., Radanovic, G., and Tylkin, P. (2017). Multi-view decision processes: the helper-ai problem. *Advances in neural information processing systems*, 30.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Gabbianelli, G., Neu, G., Papini, M., and Okolo, N. M. (2024). Offline primal-dual reinforcement learning for linear mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 3169–3177. PMLR.

Hardt, M. and Mendler-Dünner, C. (2023). Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Izzo, Z., Ying, L., and Zou, J. (2021). How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR.

Izzo, Z., Zou, J., and Ying, L. (2022). How to learn when data gradually reacts to your model. In *International Conference on Artificial Intelligence and Statistics*, pages 3998–4035. PMLR.

Jin, C., Liu, Q., and Miryoosefi, S. (2021a). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021b). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

Kearns, M., Littman, M. L., and Singh, S. (2013). Graphical models for game theory. *arXiv preprint arXiv:1301.2281*.

Letchford, J., MacDermed, L., Conitzer, V., Parr, R., and Isbell, C. (2012). Computing optimal strategies to commit to in stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1380–1386.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, Q. and Wai, H.-T. (2022). State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR.

Li, Q. and Wai, H.-T. (2024). Stochastic optimization schemes for performative prediction with nonconvex loss. *arXiv preprint arXiv:2405.17922*.

Li, Q., Yau, C.-Y., and Wai, H.-T. (2022). Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems*, 35:38449–38460.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lu, S. (2023). Bilevel optimization with coupled decision-dependent distributions. In *International Conference on Machine Learning*, pages 22758–22789. PMLR.

Mandal, D., Triantafyllou, S., and Radanovic, G. (2023). Performative reinforcement learning. In *International Conference on Machine Learning*, pages 23642–23680. PMLR.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939.

Miller, J. P., Perdomo, J. C., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. (2023). Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202):1–56.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020a). Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020b). Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

Piliouras, G. and Yu, F.-Y. (2023). Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1047–1074.

Pollatos, V., Mandal, D., and Radanovic, G. (2025). On corruption-robustness in performative reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Rank, B., Triantafyllou, S., Mandal, D., and Radanovic, G. (2024). Performative reinforcement learning in gradually shifting environments. *arXiv preprint arXiv:2402.09838*.

Ray, M., Ratliff, L. J., Drusvyatskiy, D., and Fazel, M. (2022). Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8081–8088.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Von Stackelberg, H. (2010). *Market structure and equilibrium*. Springer Science & Business Media.

Vorobeychik, Y. and Singh, S. (2012). Computing stackelberg equilibria in discounted stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1478–1484.

Wang, R., Salakhutdinov, R. R., and Yang, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135.

Wedin, P.-Å. (1973). Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13:217–232.

Wei, C.-Y. and Luo, H. (2021). Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pages 4300–4354. PMLR.

Wei, H., Zheng, G., Gayah, V., and Li, Z. (2021). Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(2):12–18.

Yan, W. and Cao, X. (2024). Zero-regret performative prediction under inequality constraints. *Advances in Neural Information Processing Systems*, 36.

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384.

Zhong, H., Yang, Z., Wang, Z., and Jordan, M. I. (2021). Can reinforcement learning find stackelberg-nash equilibria in general-sum markov games with myopic followers? *arXiv preprint arXiv:2112.13521*.

Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator if your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Table of Contents

A PROOF OF THEOREM 1

Proof. The Lagrangian of the optimization problem eq. (5) is given as follows.

$$\mathcal{L}_t(d, h, g) = d^\top \Phi\theta_t - \frac{\lambda}{2} d^\top \Phi\Phi^\top d + h^\top \left(Bd - \rho - \gamma \cdot \boldsymbol{\mu}_t \Phi^\top d\right) + d^\top g$$

Let $\mathcal{F}_t$ be the function defined as $\mathcal{F}_t(h, g) = \max_d \mathcal{L}_t(d, h, g)$. Then the dual of the optimization problem defined in eq. (5) is given as follows.

$$\min_{h,g\geq 0} \mathcal{F}_t(h, g)$$

For a fixed h and g, the gradient of $\mathcal{L}_t(d, h)$ with respect to d is given as follows.

$$\nabla_d \mathcal{L}_t(d, h) = \Phi\theta_t - \lambda\Phi\Phi^\top d + B^\top h - \gamma\Phi\boldsymbol{\mu}_t^\top h + g$$

At an optimal solution $d^\star$, $\nabla_d \mathcal{L}_t(d^\star, h, g) = 0$, and we have

$$\begin{aligned} g &= -\Phi\theta_t + \lambda\Phi\Phi^\top d^\star - B^\top h + \gamma\Phi\boldsymbol{\mu}_t^\top h \\ \Rightarrow g^\top d^\star &= -(d^\star)^\top \Phi\theta_t + \lambda(d^\star)^\top \Phi\Phi^\top d^\star - (d^\star)^\top B^\top h + \gamma(d^\star)^\top \Phi\boldsymbol{\mu}_t^\top h \end{aligned}$$

Substituting the above expression of $g^\top d^\star$ we get the following form of the Lagrangian.

$$\mathcal{F}_t(h, g) = \mathcal{L}_t(d^\star, h, g) = \frac{\lambda}{2}(d^\star)^\top \Phi\Phi^\top d^\star - h^\top \rho \tag{13}$$

Moreover, from the equation $\nabla_d \mathcal{L}_t(d^\star, h, g) = 0$ we also have the following expression of $\Phi^\top d^\star$.

$$\begin{aligned} \Phi\Phi^\top d^\star &= \frac{1}{\lambda}\left(\Phi\theta_t + B^\top h - \gamma\Phi\boldsymbol{\mu}_t^\top h + g\right) \\ \Rightarrow \Phi^\top d^\star &= \frac{1}{\lambda}\left(\theta_t + (\Phi^\dagger B^\top - \gamma\boldsymbol{\mu}_t^\top)h + \Phi^\dagger g\right) \end{aligned} \tag{14}$$

Here we use the fact that the matrix Φ has full row rank and its left pseudoinverse is $\Phi^\dagger = (\Phi^\top\Phi)^{-1}\Phi^\top$. Finally substituting the above expression of $\Phi^\top d^\star$ in eq. (13) we get the following expression for the dual problem.

$$\min_{h,g\geq 0} \mathcal{F}_t(h, g) = \frac{1}{2\lambda}\left\|(\Phi^\dagger B^\top - \gamma\boldsymbol{\mu}_t^\top)h + \Phi^\dagger g + \theta_t\right\|_2^2 - h^\top \rho \tag{15}$$

Let (h_{t+1}, g_{t+1}) be the optimal dual solutions corresponding to the occupancy measure d_{t+1} i.e. $(h_{t+1}, g_{t+1}) \in \arg\min_{h,g\geq 0} \mathcal{F}_t(h, g)$. Additionally, let (h_S, g_S) be the optimal dual solutions corresponding to the stable occupancy measure d_S i.e. $(h_S, g_S) \in \arg\min_{h,g\geq 0} \mathcal{F}_S(h, g)$, where the objective $\mathcal{F}_S(h, g)$ is given as follows.

$$\mathcal{F}_S(h, g) = \frac{1}{2\lambda} \left\|(\Phi^\dagger B^\top - \gamma \boldsymbol{\mu}_S^\top)h + \Phi^\dagger g + \theta_S\right\|_2^2 - h^\top \rho$$

Let $u_t = (\Phi^\dagger B^\top - \gamma \boldsymbol{\mu}_t^\top)h_t + \Phi^\dagger g_t$ and $u_S = (\Phi^\dagger B^\top - \gamma \boldsymbol{\mu}_S^\top)h_S + \Phi^\dagger g_S$. Then using eq. (14) we get the following bound.

$$\begin{aligned}\sqrt{\kappa}\,\|d_{t+1} - d_S\|_2 \leq \left\|\Phi^\top d_{t+1} - \Phi^\top d_S\right\|_2 &\leq \frac{1}{\lambda}\left(\|\theta_{t+1} - \theta_S\|_2 + \|u_{t+1} - u_S\|_2\right) \\ &\leq \frac{1}{\lambda}\left(\varepsilon_\theta \|d_t - d_S\|_2 + \|u_{t+1} - u_S\|_2\right) \end{aligned} \tag{16}$$

The first inequality uses assumption 3 and $d_{t+1} \neq d_S$. If $d_{t+1} = d_S$ we are already done. Now we need to bound the term $\|u_{t+1} - u_S\|_2$. Note that the dual optimization problem in variable h is unconstrained. Therefore, given g_{t+1} we must have $\nabla_h \mathcal{F}_t(h_{t+1}, g_{t+1}) = 0$. We will denote $M_t = \Phi^\dagger B^\top - \gamma \boldsymbol{\mu}_t^\top$. Moreover, the proof of lemma 3 shows that the matrix M_t has full column rank, and hence its right pseudoinverse exists and is given as $M_t^\dagger = M_t^\top (M_t M_t^\top)^{-1}$.

$$\begin{aligned}&\frac{1}{\lambda} M_t^\top M_t h_{t+1} + \frac{1}{\lambda} M_t^\top (\Phi^\dagger g_{t+1} + \theta_t) - \rho = 0 \\ \Rightarrow \quad & M_t h_{t+1} = -\Phi^\dagger g_{t+1} - \theta_t + \lambda \left(M_t^\dagger\right)^\top \rho = -u_{t+1} + M_{t+1} h_{t+1} - \theta_t + \lambda \left(M_t^\dagger\right)^\top \rho\end{aligned}$$

Rearranging we get the following expression for u_{t+1}.

$$u_{t+1} = (M_{t+1} - M_t)h_{t+1} - \theta_t + \lambda \left(M_t^\dagger\right)^\top \rho$$

Similarly, we can establish the following relation.

$$u_S = -\theta_S + \lambda \left(M_S^\dagger\right)^\top \rho$$

Using the expressions of u_{t+1} and u_S we get the following upper bound.

$$\begin{aligned}\|u_{t+1} - u_S\|_2 &\leq \|\theta_t - \theta_S\|_2 + \lambda \left\|\left((M_t^\dagger)^\top - (M_S^\dagger)^\top\right)\rho\right\|_2 + \|M_{t+1} - M_t\|_2 \|h_{t+1}\|_2 \\ &\leq \varepsilon_\theta \|d_{t-1} - d_S\|_2 + \lambda \left\|(M_t^\dagger)^\top - (M_S^\dagger)^\top\right\|_2 + \gamma \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t+1}\|_2 \|h_{t+1}\|_2 \quad \text{[By assumption 2]} \\ &\leq \varepsilon_\theta \|d_{t-1} - d_S\|_2 + \lambda \left\|(M_t^\dagger)^\top - (M_S^\dagger)^\top\right\|_2 + \gamma \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_S\|_2 \|h_{t+1}\|_2 + \gamma \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_S\|_2 \|h_{t+1}\|_2 \\ &\leq (\varepsilon_\theta + \gamma\varepsilon_\mu \|h_{t+1}\|_2) \|d_{t-1} - d_S\|_2 + \gamma\varepsilon_\mu \|h_{t+1}\|_2 \|d_t - d_S\|_2 \\ &+ 3\lambda \left\|M_t^\top - M_S^\top\right\|_2 \left\|(M_t^\dagger)^\top\right\|_2 \left\|(M_S^\dagger)^\top\right\|_2 \quad \text{[By matrix perturbation bound[3] and assumption 2]} \\ &\leq (\varepsilon_\theta + \gamma\varepsilon_\mu \|h_{t+1}\|_2) \|d_{t-1} - d_S\|_2 + \gamma\varepsilon_\mu \|h_{t+1}\|_2 \|d_t - d_S\|_2 \\ &+ 3\lambda\gamma \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_S\|_2 \left\|(M_t M_t^\top)^{-1} M_t\right\|_2 \left\|(M_S M_S^\top)^{-1} M_S\right\|_2 \\ &\leq (\varepsilon_\theta + \gamma\varepsilon_\mu \|h_{t+1}\|_2) \|d_{t-1} - d_S\|_2 + \gamma\varepsilon_\mu \|h_{t+1}\|_2 \|d_t - d_S\|_2 \\ &+ 3\lambda\gamma\varepsilon_\mu \|d_{t-1} - d_S\|_2 \left\|(M_t M_t^\top)^{-1} M_t\right\|_2 \left\|(M_S M_S^\top)^{-1} M_S\right\|_2\end{aligned}$$

Using Lemma 3 we get the following bound, $\left\|(M_t M_t^\top)^{-1} M_t\right\|_2 \leq \frac{\sqrt{M}}{\sqrt{A}(1-\gamma)}$. Similarly, we can bound $\left\|(M_S M_S^\top)^{-1} M_S\right\|_2 \leq \frac{\sqrt{M}}{\sqrt{A}(1-\gamma)}$. Finally, observe that h_{t+1} is an optimal dual solution, and we can use Lemma 4 to bound norm of the solution h_{t+1}. Let $\alpha = \frac{\sqrt{M}}{\sqrt{A}(1-\gamma)}$. Then we have,

$$\|u_{t+1} - u_S\|_2 \leq \left(\varepsilon_\theta + 3\lambda\gamma\varepsilon_\mu\alpha^2 + \alpha(\lambda\alpha + \sqrt{D})\gamma\varepsilon_\mu\right)\|d_{t-1} - d_S\|_2 + \gamma\varepsilon_\mu\alpha(\lambda\alpha + \sqrt{D})\|d_t - d_S\|_2$$

[3] For example, see Theorem 4.1 of Wedin (1973).

Substituting the above bound in eq. (16) we get the following recurrence relation.

$$\|d_{t+1} - d_S\|_2 \leq \underbrace{\frac{1}{\lambda\sqrt{\kappa}}\left(\varepsilon_\theta + 3\lambda\gamma\varepsilon_\mu\alpha^2 + \alpha(\lambda\alpha + \sqrt{D})\gamma\varepsilon_\mu\right)}_{:=\beta}\|d_{t-1} - d_S\|_2$$
$$+ \underbrace{\frac{1}{\lambda\sqrt{\kappa}}\left(\varepsilon_\theta + \gamma\varepsilon_\mu\alpha(\lambda\alpha + \sqrt{D})\right)}_{:=\beta_1}\|d_t - d_S\|_2 \tag{17}$$

Notice that $\beta \geq \beta_1$, which gives us the following recurrence relation $\|d_{t+1} - d_S\|_2 \leq \beta(\|d_t - d_S\|_2 + \|d_{t-1} - d_S\|_2)$. We claim that $\|d_{t+1} - d_S\|_2 \leq \frac{2}{1-\gamma}r^t$ for $r = \frac{\beta}{2}\left(1 + \frac{1}{2}\sqrt{1 + \frac{4}{\beta}}\right)$. The proof of this claim is through induction. Indeed, $\|d_0 - d_S\|_2 \leq \frac{1}{1-\gamma}$ as $\sum_{s,a} d(s,a) = \frac{1}{1-\gamma}$ for any occupancy measure d. Furthermore, assuming the claim holds for any index less than or equal to t, we get the following bound.

$$\|d_{t+1} - d_S\|_2 \leq \frac{2\beta}{1-\gamma}\left(r^{t-1} + r^{t-2}\right) = \frac{2r^{t-2}}{1-\gamma}(\beta r + \beta)$$

It can be checked that $r = \frac{\beta}{2}\left(1 + \frac{1}{2}\sqrt{1 + \frac{4}{\beta}}\right)$ is one of the roots of $x^2 - \beta x - \beta = 0$ and $\beta r + \beta = r^2$. This proves that $\|d_{t+1} - d_S\|_2 \leq \frac{2}{1-\gamma}r^t$. If $r < 1$, then as long as $t \geq \ln\left(\frac{2}{\delta(1-\gamma)}\right)/\ln(1/r)$ we are guaranteed that $\|d_t - d_S\|_2 \leq \delta$.

Now we determine the condition that guarantees $r < 1$. Moreover, we can express the parameter β as follows. Note that $r < \frac{\beta}{2}\left(1 + \frac{1}{2} + \frac{1}{\sqrt{\beta}}\right) < \frac{3}{4}\beta + \frac{\sqrt{\beta}}{2}$. If we ensure that $\beta < \frac{16}{25}$ then we get $r < \left(\frac{3}{4} + \frac{1}{2}\right)\sqrt{\beta} < \frac{5}{4}\sqrt{\beta} < 1$. From the definition of β in eq. (17), we can express β as follows.

$$\beta = \frac{\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu}{\lambda\sqrt{\kappa}} + \frac{4\gamma\varepsilon_\mu\alpha^2}{\sqrt{\kappa}}$$

If $\lambda > \frac{25(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu)}{8\sqrt{\kappa}}$ then the first term in β is less than $8/25$. In order for $\beta < 16/25$ we need $\varepsilon_\mu < \frac{2\sqrt{\kappa}}{25\gamma\alpha^2}$. □

Lemma 3. *Suppose assumption 3 holds, then* $\left\|M_t^\dagger\right\|_2 \leq \frac{\sqrt{\mathcal{M}}}{\sqrt{A}(1-\gamma)}$.

Proof. Since $M_t = \Phi^\dagger B^\top - \gamma\boldsymbol{\mu}_t^\top$ and $\Phi^\top\Phi$ is invertible (by assumption 3), we obtain the following expression for M_t.

$$\begin{aligned} M_t &= \Phi^\dagger B^\top - \gamma\boldsymbol{\mu}_t^\top \\ &= (\Phi^\top\Phi)^{-1}\left[\Phi^\top B^\top - \gamma\Phi^\top\Phi\boldsymbol{\mu}_t^\top\right] \\ &= (\Phi^\top\Phi)^{-1}\Phi^\top[B^\top - \gamma P^\top] \\ &= \Phi^\dagger[B^\top - \gamma P_t^\top] \end{aligned}$$

where $P_t = \boldsymbol{\mu}_t\Phi^\top$ is the probability transition matrix. The singular values of $\Phi^\dagger$ are obtained by inverting all the non-zero singular values of Φ and leaving zero singular values as they are. By assumption 3 Φ has rank d and hence only non-zero singular values. Moreover, since $\lambda_{\max}(\Phi^\top\Phi) \leq \mathcal{M}$, we have $\sigma_{\max}(\Phi) \leq \sqrt{\mathcal{M}}$ and $\sigma_{\min}(\Phi^\dagger) \geq \frac{1}{\sqrt{\mathcal{M}}}$. By using an argument very similar to lemma 5 of Mandal et al. (2023) we can show that $\sigma_{\min}(B^\top - \gamma P_t^\top) \geq \sqrt{A}(1-\gamma)$. Therefore,

$$\sigma_{\min}(M_t) \geq \sigma_{\min}(\Phi^\dagger)\sigma_{\min}(B^\top - \gamma P_t^\top) \geq \frac{\sqrt{A}(1-\gamma)}{\sqrt{\mathcal{M}}}$$

Since the singular values of $M_t^\dagger$ are formed by inverting the non-zero singular values of M_t and leaving the zero singular values as they are, we obtain $\left\|M_t^\dagger\right\|_2 \leq \frac{\sqrt{\mathcal{M}}}{\sqrt{A}(1-\gamma)}$. □

Lemma 4. *Let* $(h^\star, g^\star)$ *be an optimal solution of the optimization problem equation 15 of the minimum norm. Then* $\left\|h^\star\right\|_2 \leq \alpha\left(\lambda\alpha + \sqrt{D}\right)$, *where* $\alpha = \frac{\sqrt{\mathcal{M}}}{\sqrt{A}(1-\gamma)}$.

Proof. At an optimal solution $(h^\star, g^\star)$ we must have $\nabla_h \mathcal{F}_t(h^\star, g^\star) = 0$ and $\left\langle \nabla_g \mathcal{F}_t(h^\star, g^\star), g - g^\star \right\rangle \geq 0$ for any $g \geq 0$. Let $M_t = \Phi^\dagger B^\top - \gamma \boldsymbol{\mu}_t^\top$. Then we have,

$$\begin{aligned} \nabla_h \mathcal{F}_t(h^\star, g^\star) = \frac{1}{\lambda} M_t^\top M_t h^\star + \frac{1}{\lambda} M_t^\top \left(\Phi^\dagger g^\star + \theta_t\right) - \rho = 0 \\ \Rightarrow M_t h^\star + \theta_t = \left(M_t^\dagger\right)^\top \lambda\rho - \Phi^\dagger g^\star. \end{aligned} \tag{18}$$

On the other hand,

$$\nabla_g \mathcal{F}_t(h^\star, g^\star) = \frac{1}{\lambda}(\Phi^\dagger)^\top \Phi^\dagger g^\star + \frac{1}{\lambda}(\Phi^\dagger)^\top \left(M_t h^\star + \theta_t\right) = (\Phi^\dagger)^\top \left(M_t^\dagger\right)^\top \rho$$

Let the j-th coordinate of the vector $(\Phi^\dagger)^\top \left(M_t^\dagger\right)^\top \rho$ is non-zero. This implies that the j-th coordinate of $g^\star$ is zero, as $\left\langle \nabla_g \mathcal{F}_t(h^\star, g^\star), g - g^\star \right\rangle \geq 0$ for any $g \geq 0$ and one can take $g = g^\star \pm \frac{g_j^\star}{2} \cdot \boldsymbol{e}_j$ to conclude that $g_j^\star = 0$. This is also equivalent to the condition ${g^\star}^\top (\Phi^\dagger)^\top \left(M_t^\dagger\right)^\top \rho = 0$.

We now show that without loss of generality, one can choose $g^\star = 0$. Given a solution $(h^\star, g^\star)$ let us choose $h = h^\star + \Delta$ where $\Delta = M_t^\dagger \Phi^\dagger g^\star$. Then the objective at the solution $(h, 0)$ is the following.

$$\begin{aligned} \mathcal{F}_t(h, 0) &= \frac{1}{2\lambda} \|M_t h + \theta_t\|_2^2 - h^\top \rho \\ &= \frac{1}{2\lambda} \left\|M_t(h^\star + \Delta) + \theta_t\right\|_2^2 - (h^\star + \Delta)\rho \\ &= \frac{1}{2\lambda} \left\|M_t h^\star + M_t \Delta + \theta_t\right\|_2^2 - (h^\star)^\top \rho - \Delta^\top \rho \\ &= \frac{1}{2\lambda} \left\|M_t h^\star + \Phi^\dagger g^\star + \theta_t\right\|_2^2 - (h^\star)^\top \rho - {g^\star}^\top (\Phi^\dagger)^\top \left(M_t^\dagger\right)^\top \rho \\ &= \frac{1}{2\lambda} \left\|M_t h^\star + \Phi^\dagger g^\star + \theta_t\right\|_2^2 - (h^\star)^\top \rho = \mathcal{F}_t(h^\star, g^\star) \end{aligned}$$

Substituting $g^\star = 0$ in eq. (18) we get the following equation: $M_t h^\star + \theta_t = \lambda \left(M_t^\dagger\right)^\top \rho$. The solution of this equation is $h^\star = M_t^\dagger \left(\lambda \left(M_t^\dagger\right)^\top \rho - \theta_t\right)$ and we can bound its norm as follows.

$$\begin{aligned} \left\|h^\star\right\|_2 &\leq \left\|M_t^\dagger\right\|_2 \left(\lambda \left\|\left(M_t^\dagger\right)^\top\right\|_2 \|\rho\|_2 + \|\theta_t\|_2\right) \\ &\leq \frac{\sqrt{\mathcal{M}}}{\sqrt{A}(1-\gamma)} \left(\lambda \frac{\sqrt{\mathcal{M}}}{\sqrt{A}(1-\gamma)} + \sqrt{D}\right) \end{aligned}$$

The last inequality uses Lemma 3. □

B PROOF OF THEOREM 2

Proof. Let $C(d_S^\lambda)$ be the set of occupancy measures that are feasible with respect to the measure $\boldsymbol{\mu}_\lambda = \mathcal{F}_\mu(d_S^\lambda)$ i.e. $C(d_S^\lambda) = \{d : Bd = \rho + \gamma \cdot \boldsymbol{\mu}_\lambda \Phi^\top d,\ d \geq 0\}$. As d_S^λ maximizes the objective equation 5, we have the following bound.

$$d_S^{\lambda\top} \Phi \theta_S - \frac{\lambda}{2} d_S^{\lambda\top} \Phi\Phi^\top d_S^\lambda \geq \max_{d \in C(d_S^\lambda)} d^\top \Phi \theta_S - \frac{\lambda}{2} d^\top \Phi\Phi^\top d \tag{19}$$

After rearranging and using assumption 3 we get the following lower bound.

$$\begin{aligned} d_S^{\lambda\top} \Phi \theta_S &\geq \max_{d \in C(d_S^\lambda)} d^\top \Phi \theta_S - \frac{\lambda}{2} d^\top \Phi\Phi^\top d + \frac{\lambda}{2} d_S^{\lambda\top} \Phi\Phi^\top d_S^\lambda \\ &\geq \max_{d \in C(d_S^\lambda)} d^\top \Phi \theta_S - \frac{\lambda}{2} d^\top \Phi\Phi^\top d \\ &\geq \max_{d \in C(d_S^\lambda)} d^\top \Phi \theta_S - \frac{\lambda}{2} \mathcal{M} \sum_a d(\cdot, a)^\top d(\cdot, a) \\ &\geq \max_{d \in C(d_S^\lambda)} d^\top \Phi \theta_S - \frac{\lambda \mathcal{M}}{2(1-\gamma)^2} \end{aligned}$$

The last line uses $\|d\|_2^2 = \sum_{s,a} d^2(s,a) = (1-\gamma)^{-2} \sum_{s,a}((1-\gamma)d(s,a))^2 \leq (1-\gamma)^{-2} \sum_{s,a}(1-\gamma)d(s,a) = (1-\gamma)^{-2}$. Now we substitute $\lambda = \frac{25\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)}{8\sqrt{\kappa}}$ and obtain the following inequality.

$$d_S^{\lambda\top}\Phi\theta_S \geq \max_{d \in C(d_S^\lambda)} d^\top \Phi\theta_S - \frac{25\mathcal{M}\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)}{16\sqrt{\kappa}(1-\gamma)^2}$$

□

C PROOF OF THEOREM 3

In order to formally state, and prove Theorem 3 we will need to introduce some definitions. Let d_{PO}^λ be the performatively optimal occupancy measure, and $\pi_{\mathrm{PO}}^\lambda$ be the performatively optimal policy which is defined as follows.

$$\pi_{\mathrm{PO}}^\lambda(a \mid s) = \begin{cases} \frac{d_{\mathrm{PO}}^\lambda(s,a)}{\sum_b d_{\mathrm{PO}}^\lambda(s,b)} & \text{if } \sum_b d_{\mathrm{PO}}^\lambda(s,b) > 0 \\ \frac{1}{A} & \text{o.w.} \end{cases}$$

Let us also define $d_{\mathrm{PO}\to\mathrm{PO}}^\lambda$ (resp. $d_{\mathrm{PO}\to S}^\lambda$) to be the occupancy measure of the policy $\pi_{\mathrm{PO}}^\lambda$ in the MDP M_{PO}^λ (resp. M_S^λ). Note that, $d_{\mathrm{PO}\to\mathrm{PO}}^\lambda$ need not be equal to d_{PO}^λ.

Theorem 6 (Formal statement of Theorem 3)**.** *Suppose the assumptions of Theorem 1 hold, and* $\Delta = \frac{3\gamma\varepsilon_\mu\mathcal{M}\sqrt{D}}{(1-\gamma)^2} + \varepsilon_\theta\sqrt{\mathcal{M}}$*, and* $\lambda_0 = \frac{25}{8\sqrt{\kappa}}\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)$ *(required lower bound from Theorem 1). Then there exists a choice of regularization parameter* (λ) *such that repeatedly optimizing objective (5) converges to a solution* d_S^λ *with the following guarantee.*

$$d_{\mathrm{PO}\to\mathrm{PO}}^\top \Phi\theta_{\mathrm{PO}} - d_S^{\lambda\top}\Phi\theta_S^\lambda \leq 4\sqrt{\frac{(1+\Delta)\Delta}{\kappa} \cdot \frac{\mathcal{M}}{(1-\gamma)^2}} + \lambda_0 \cdot \frac{\mathcal{M}}{(1-\gamma)^2}$$

The suboptimality gap established in Theorem 3 asymptotically scales as $O(\max\{1,\Delta\}\,\Delta + \lambda_0)$. As $\varepsilon_\theta, \varepsilon_\mu \to 0$, both λ_0 and Δ converge to zero, and d_S^λ approaches a performatively optimal solution with respect to the unregularized objective.

The proof of Theorem 3 first upper bounds the suboptimality gap in terms of $\left\|d_{\mathrm{PO}\to S}^\lambda - d_S^\lambda\right\|_2$ the distance between the occupancy measures resulting from deploying (regularized) optimal policy $(\pi_{\mathrm{PO}}^\lambda)$ and stable policy (π_S^λ) in the stable environment. Then Lemma 5 provides an upper bound on the distance between these two measures that scales as $O\left(\frac{\Delta}{\lambda}\right)$ under certain conditions. Substituting this upper bound and then choosing an appropriate value of the regularizer λ gives the desired bound.

Proof. Given a policy π, we define the matrix $\Pi \in \mathbb{R}^{SA\times S}$ as follows.

$$\Pi = \begin{bmatrix} \pi(\cdot \mid s_1) & 0 & \dots \\ 0 & \pi(\cdot \mid s_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Given a measure $\boldsymbol{\mu}$ (i.e. probability transition $P = \boldsymbol{\mu}\Phi^\top$ the occupancy measure d of the deployed policy is given as

$$d = \Pi\rho + \gamma\Pi\boldsymbol{\mu}\Phi^\top\Pi\rho + \gamma^2\left(\Pi\boldsymbol{\mu}\Phi^\top\right)^2\Pi\rho + \ldots = \left(\mathrm{Id} - \gamma\Pi\boldsymbol{\mu}\Phi^\top\right)^{-1}\Pi\rho \tag{20}$$

Furthermore, following objective eq. (5), given an MDP $M = (\theta, \boldsymbol{\mu})$ we will write the regularized reward of an occupancy measure d as

$$RR(d; M) = d^\top\Phi\theta - \frac{\lambda}{2}d^\top\Phi\Phi^\top d.$$

Then, using the definition of performative optimality and stability, we get the following sequence of inequalities.

$$RR(d_{\mathrm{PO}\to\mathrm{PO}}^\lambda; M_{\mathrm{PO}}^\lambda) \geq RR(d_S^\lambda; M_S^\lambda) \geq RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda) \tag{21}$$

$$
\begin{aligned}
& d_{\mathrm{PO}\to\mathrm{PO}}^{\top}\Phi\theta_{\mathrm{PO}} - d_S^{\lambda^\top}\Phi\theta_S^{\lambda} \\
&= \left(d_{\mathrm{PO}\to\mathrm{PO}}^{\top}\Phi\theta_{\mathrm{PO}} - \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}}\right) + \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}} \\
&- \left(d_S^{\lambda^\top}\Phi\theta_S^{\lambda} - \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda^\top}\right) - \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda^\top} \\
&\leq \left(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda^\top}\Phi\theta_{\mathrm{PO}}^{\lambda} - \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda^\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda}\right) + \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}} \\
&- \left(d_S^{\lambda^\top}\Phi\theta_S^{\lambda} - \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda^\top}\right) \\
&= \underbrace{RR(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda}; M_{\mathrm{PO}}^{\lambda}) - RR(d_S^{\lambda}; M_S^{\lambda})}_{:=T_1} + \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}}
\end{aligned} \tag{22}
$$

Using eq. (21) we can upper bound the term T_1 as follows.

$$
\begin{aligned}
T_1 &\leq RR(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda}; M_{\mathrm{PO}}^{\lambda}) - RR(d_S^{\lambda}; M_S^{\lambda}) \\
&\leq \left(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda^\top}\Phi\theta_{\mathrm{PO}}^{\lambda} - \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda^\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda}\right) - \left(d_S^{\lambda^\top}\Phi\theta_S^{\lambda} - \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda}\right) \\
&= \left(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda^\top}\Phi\theta_{\mathrm{PO}}^{\lambda} - d_S^{\lambda^\top}\Phi\theta_{\mathrm{PO}}^{\lambda}\right) + \left(d_S^{\lambda^\top}\Phi\theta_{\mathrm{PO}}^{\lambda} - d_S^{\lambda^\top}\Phi\theta_S^{\lambda}\right) \\
&- \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda^\top}\Phi\Phi^{\top}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda} + \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda} \\
&\leq \left\|d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda} - d_S^{\lambda}\right\|_2 \left\|\Phi\theta_{\mathrm{PO}}^{\lambda}\right\|_2 + \left\|d_S^{\lambda^\top}\Phi\right\|_2 \left\|\theta_{\mathrm{PO}}^{\lambda} - \theta_S^{\lambda}\right\|_2 + \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda} \\
&\leq \left(\left\|d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda} - d_{\mathrm{PO}\to S}^{\lambda}\right\|_2 + \left\|d_{\mathrm{PO}\to S}^{\lambda} - d_S^{\lambda}\right\|_2\right)\left\|\Phi\theta_{\mathrm{PO}}^{\lambda}\right\|_2 + \left\|d_S^{\lambda^\top}\Phi\right\|_2 \left\|\theta_{\mathrm{PO}}^{\lambda} - \theta_S^{\lambda}\right\|_2 + \frac{\lambda}{2}d_S^{\lambda^\top}\Phi\Phi^{\top}d_S^{\lambda}
\end{aligned} \tag{23}
$$

Using eq. (20) we get,

$$
\begin{aligned}
\left\|d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda} - d_{\mathrm{PO}\to S}^{\lambda}\right\|_2 &= \left\|\left\{\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1} - \left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_S\Phi^{\top}\right)^{-1}\right\}\Pi_{\mathrm{PO}}\rho\right\|_2 \\
&\leq \left\|\left\{\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1} - \left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_S\Phi^{\top}\right)^{-1}\right\}\right\|_2 \|\Pi_{\mathrm{PO}}\rho\|_2 \\
&\leq 3\gamma\left\|\Pi_{\mathrm{PO}}\left(\boldsymbol{\mu}_{\mathrm{PO}} - \boldsymbol{\mu}_S\right)\Phi^{\top}\right\|_2 \left\|\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1}\right\|_2 \left\|\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1}\right\|_2 \|\Pi_{\mathrm{PO}}\rho\|_2
\end{aligned}
$$

The last inequality uses perturbation bound for the inverse of a matrix, in particular $\left\|A^{-1}\right\|_2 - \left\|B^{-1}\right\|_2 \leq 3\|A - B\|_2\left\|A^{-1}\right\|_2\left\|B^{-1}\right\|_2$. We will now use the following set of observations.

1. $\|\Pi_{\mathrm{PO}}\|_2 \leq 1$ as $\|\Pi_{\mathrm{PO}}v\|_2^2 = \sum_{s,a}\pi^2(a \mid s)v_{s,a}^2 \leq \|v\|_2^2$.
2. Since $\boldsymbol{\mu}_S\Phi^{\top}$ is a probability transition function we have $\left\|\Pi_{\mathrm{PO}}\boldsymbol{\mu}_S\Phi^{\top}\right\|_2 \leq \|\Pi_{\mathrm{PO}}\|_2\left\|\boldsymbol{\mu}_S\Phi^{\top}\right\|_2 \leq 1$. This also implies that $\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_S\Phi^{\top} \succcurlyeq (1-\gamma)\cdot\mathrm{Id}$.
3. By assumption 3, we have $\left\|\Phi^{\top}\right\|_2 \leq \sqrt{M}$.

$$
\left\|d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda} - d_{\mathrm{PO}\to S}^{\lambda}\right\|_2 \leq \frac{3\gamma\sqrt{M}}{(1-\gamma)^2}\|\boldsymbol{\mu}_{\mathrm{PO}} - \boldsymbol{\mu}_S\|_2 \leq \frac{3\gamma\varepsilon_{\mu}\sqrt{M}}{(1-\gamma)^2}\left\|d_{\mathrm{PO}\to S}^{\lambda} - d_S^{\lambda}\right\|_2
$$

The last inequality uses assumption equation 2 and the fact that the measure $d_{\mathrm{PO}\to S}$ induces policy π_{PO} i.e. $\pi_{d_{\mathrm{PO}\to S}}$, as defined in equation 1 equals π_{PO}. Substituting the above bound in eq. (23) we get the following upper bound on T_1.

$$
\begin{aligned}
T_1 &\leq \left(1 + \frac{3\gamma\varepsilon_{\mu}\sqrt{M}}{(1-\gamma)^2}\right)\left\|d_{\mathrm{PO}\to S}^{\lambda} - d_S^{\lambda}\right\|_2 \|\Phi\|_2\left\|\theta_{\mathrm{PO}}^{\lambda}\right\|_2 + \left\|d_S^{\lambda}\right\|_2\|\Phi\|_2\,\varepsilon_{\theta}\left\|d_{\mathrm{PO}\to S}^{\lambda} - d_S^{\lambda}\right\|_2 + \frac{\lambda}{2}\left\|d_S^{\lambda}\right\|_{\Phi\Phi^{\top}} \\
&\leq \left(1 + \underbrace{\frac{3\gamma\varepsilon_{\mu}M\sqrt{D}}{(1-\gamma)^2} + \varepsilon_{\theta}\sqrt{M}}_{:=\Delta}\right)\left\|d_{\mathrm{PO}\to S}^{\lambda} - d_S^{\lambda}\right\|_2 + \frac{\lambda}{2}\left\|d_S^{\lambda}\right\|_{\Phi\Phi^{\top}}
\end{aligned}
$$

Substituting the above bound in eq. (22) and using the observation that for any occupancy measure d we have $d^\top \Phi\Phi^\top d \le \frac{\mathcal{M}}{(1-\gamma)^2}$ we get the following bound.

$$\begin{aligned} d_{\mathrm{PO}\to\mathrm{PO}}^\top \Phi\theta_{\mathrm{PO}} - d_S^{\lambda\top}\Phi\theta_S^\lambda &\le (1+\Delta)\left\|d_{\mathrm{PO}\to S}^\lambda - d_S^\lambda\right\|_2 + \frac{\lambda\mathcal{M}}{(1-\gamma)^2} \\ &\le (1+\Delta)\frac{4\Delta}{\kappa\lambda} + \frac{\lambda\mathcal{M}}{(1-\gamma)^2} \text{ [By Lemma 5]} \end{aligned} \tag{24}$$

Note that the above expression can be written as $\frac{S_1}{\lambda} + \lambda \cdot S_2$ where

$$S_1 = (1+\Delta)\frac{4\Delta}{\kappa} \text{ and } S_2 = \frac{\mathcal{M}}{(1-\gamma)^2}$$

Let $\lambda_0 = \frac{25}{8\sqrt{\kappa}}\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)$. There are two cases to consider. If $\left(\frac{T_1}{T_2}\right)^{1/2} \ge \lambda_0$, we can substitute $\lambda = \left(\frac{T_1}{T_2}\right)^{1/2}$ in eq. (24) to obtain

$$d_{\mathrm{PO}\to\mathrm{PO}}^\top \Phi\theta_{\mathrm{PO}} - d_S^{\lambda\top}\Phi\theta_S^\lambda \le 2\sqrt{S_1 S_2}.$$

On the other hand, if $\left(\frac{S_1}{S_2}\right)^{1/2} < \lambda_0$, we can substitute $\lambda = \lambda_0$ in eq. (24) to obtain the following bound.

$$d_{\mathrm{PO}\to\mathrm{PO}}^\top \Phi\theta_{\mathrm{PO}} - d_S^{\lambda\top}\Phi\theta_S^\lambda \le \frac{S_1}{\sqrt{\lambda_0}} + \lambda_0 \cdot S_2 \le \sqrt{S_1 S_2} + \lambda_0 \cdot S_2$$

Combining the two bounds above, we are always guaranteed an upper bound of $2\sqrt{S_1 S_2} + \lambda_0 \cdot S_2$ on the suboptimality gap. □

C.1 Distance Between Performatively Optimal and Stable Solution

Lemma 5. *Let* $\Delta = \left(\frac{3\gamma\varepsilon_\mu\mathcal{M}\sqrt{D}}{(1-\gamma)^2} + \varepsilon_\theta\sqrt{\mathcal{M}}\right)$ *and* $c\cdot\Delta \ge \lambda \ge \frac{25}{8\sqrt{\kappa}}\left(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu\right)$ *for a constant* $c \ge \frac{4}{\sqrt{\kappa\mathcal{M}}}$. *Then*

$$\left\|d_{\mathrm{PO}\to S}^\lambda - d_S^\lambda\right\|_2 \le \frac{4\Delta}{\kappa\lambda}.$$

Proof. We first provide a lower bound on the difference $RR(d_S^\lambda; M_S^\lambda) - RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda)$. Since the occupancy measure d_S^λ maximizes $RR(\cdot; M_S^\lambda)$ we have $\left\langle \nabla_d RR(d_S^\lambda; M_S^\lambda), d_S^\lambda - d_{\mathrm{PO}\to S}^\lambda \right\rangle \ge 0$. This implies the following inequality.

$$\begin{aligned} &\left(d_S^\lambda - d_{\mathrm{PO}\to S}^\lambda\right)^\top\left(\Phi\theta_S^\lambda - \lambda\Phi\Phi^\top d_S^\lambda\right) \ge 0 \\ \Rightarrow\ & d_S^{\lambda\top}\Phi\theta_S^\lambda - \lambda d_S^{\lambda\top}\Phi\Phi^\top d_S^\lambda \ge d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\theta_S^\lambda - \lambda d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\Phi^\top d_S^\lambda \\ \Rightarrow\ & RR(d_S^\lambda; M_S^\lambda) - \frac{\lambda}{2}d_S^{\lambda\top}\Phi\Phi^\top d_S^\lambda \ge RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda) - \lambda d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\Phi^\top d_S^\lambda + \frac{\lambda}{2}d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\Phi^\top d_{\mathrm{PO}\to S}^\lambda \\ \Rightarrow\ & RR(d_S^\lambda; M_S^\lambda) - RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda) \ge \frac{\lambda}{2}\left(d_S^\lambda - d_{\mathrm{PO}\to S}^\lambda\right)^\top\Phi\Phi^\top\left(d_S^\lambda - d_{\mathrm{PO}\to S}^\lambda\right) \ge \frac{\kappa\lambda}{2}\left\|d_S^\lambda - d_{\mathrm{PO}\to S}^\lambda\right\|_2^2 \end{aligned} \tag{25}$$

The last inequality uses assumption equation 3. We now provide an upper bound on the term $RR(d_S^\lambda; M_S^\lambda) - RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda)$. Note that $RR(d_S^\lambda; M_S^\lambda) - RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda) \le RR(d_{\mathrm{PO}\to\mathrm{PO}}^\lambda; M_{\mathrm{PO}}^\lambda) - RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda)$, and we upper bound the latter.

$$\begin{aligned} &RR(d_{\mathrm{PO}\to\mathrm{PO}}^\lambda; M_{\mathrm{PO}}^\lambda) - RR(d_{\mathrm{PO}\to S}^\lambda; M_S^\lambda) \\ &\le \left(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda\top}\Phi\theta_{\mathrm{PO}}^\lambda - \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda\top}\Phi\Phi^\top d_{\mathrm{PO}\to\mathrm{PO}}^\lambda\right) - \left(d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\theta_S^\lambda - \frac{\lambda}{2}d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\Phi^\top d_{\mathrm{PO}\to S}^\lambda\right) \\ &= \left(d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda\top}\Phi\theta_{\mathrm{PO}}^\lambda - d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\theta_{\mathrm{PO}}^\lambda\right) + \left(d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\theta_{\mathrm{PO}}^\lambda - d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\theta_S^\lambda\right) \\ &- \frac{\lambda}{2}d_{\mathrm{PO}\to\mathrm{PO}}^{\lambda\top}\Phi\Phi^\top d_{\mathrm{PO}\to\mathrm{PO}}^\lambda + \frac{\lambda}{2}d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\Phi^\top d_{\mathrm{PO}\to S}^\lambda \\ &\le \left\|d_{\mathrm{PO}\to\mathrm{PO}}^\lambda - d_{\mathrm{PO}\to S}^\lambda\right\|_2\left\|\Phi\theta_{\mathrm{PO}}^\lambda\right\|_2 + \left\|d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\right\|_2\left\|\theta_{\mathrm{PO}}^\lambda - \theta_S^\lambda\right\|_2 + \frac{\lambda}{2}d_{\mathrm{PO}\to S}^{\lambda\top}\Phi\Phi^\top d_{\mathrm{PO}\to S}^\lambda \end{aligned} \tag{26}$$

Using eq. (20) we get,

$$\begin{aligned}
\left\|d^{\lambda}_{\mathrm{PO}\to\mathrm{PO}} - d^{\lambda}_{\mathrm{PO}\to S}\right\|_2 &= \left\|\left\{\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1} - \left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{S}\Phi^{\top}\right)^{-1}\right\}\Pi_{\mathrm{PO}}\rho\right\|_2 \\
&\leq \left\|\left\{\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1} - \left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{S}\Phi^{\top}\right)^{-1}\right\}\right\|_2 \|\Pi_{\mathrm{PO}}\rho\|_2 \\
&\leq 3\gamma\left\|\Pi_{\mathrm{PO}}\left(\boldsymbol{\mu}_{\mathrm{PO}} - \boldsymbol{\mu}_{S}\right)\Phi^{\top}\right\|_2 \left\|\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1}\right\|_2 \left\|\left(\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_{\mathrm{PO}}\Phi^{\top}\right)^{-1}\right\|_2 \|\Pi_{\mathrm{PO}}\rho\|_2
\end{aligned}$$

The last inequality uses perturbation bound for the inverse of a matrix, in particular $\left\|A^{-1}\right\|_2 - \left\|B^{-1}\right\|_2 \leq 3\|A - B\|_2 \left\|A^{-1}\right\|_2 \left\|B^{-1}\right\|_2$. We will now use the following set of observations.

1. $\|\Pi_{\mathrm{PO}}\|_2 \leq 1$ as $\|\Pi_{\mathrm{PO}}v\|_2^2 = \sum_{s,a}\pi^2(a \mid s)v_{s,a}^2 \leq \|v\|_2^2$.
2. Since $\boldsymbol{\mu}_S\Phi^{\top}$ is a probability transition function we have $\left\|\Pi_{\mathrm{PO}}\boldsymbol{\mu}_S\Phi^{\top}\right\|_2 \leq \|\Pi_{\mathrm{PO}}\|_2 \left\|\boldsymbol{\mu}_S\Phi^{\top}\right\|_2 \leq 1$. This also implies that $\mathrm{Id} - \gamma\Pi_{\mathrm{PO}}\boldsymbol{\mu}_S\Phi^{\top} \succcurlyeq (1-\gamma)\cdot\mathrm{Id}$.
3. By assumption 3 we have, $\left\|\Phi^{\top}\right\|_2 \leq \sqrt{M}$.

$$\left\|d^{\lambda}_{\mathrm{PO}\to\mathrm{PO}} - d^{\lambda}_{\mathrm{PO}\to S}\right\|_2 \leq \frac{3\gamma\sqrt{M}}{(1-\gamma)^2}\|\boldsymbol{\mu}_{\mathrm{PO}} - \boldsymbol{\mu}_S\|_2 \leq \frac{3\gamma\varepsilon_{\mu}\sqrt{M}}{(1-\gamma)^2}\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2$$

The last inequality uses assumption 2 and the fact that the measure $d_{\mathrm{PO}\to S}$ induces policy π_{PO} i.e. $\pi_{d_{\mathrm{PO}\to S}}$, as defined in equation 1 equals π_{PO}. Substituting the above bound in eq. (26) we get the following upper bound.

$$\begin{aligned}
&RR(d^{\lambda}_{\mathrm{PO}\to\mathrm{PO}}; M^{\lambda}_{\mathrm{PO}}) - RR(d^{\lambda}_{\mathrm{PO}\to S}; M^{\lambda}_{S}) \\
&\leq \frac{3\gamma\varepsilon_{\mu}\sqrt{M}}{(1-\gamma)^2}\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2 \|\Phi\|_2 \left\|\theta^{\lambda}_{\mathrm{PO}}\right\|_2 + \left\|d^{\lambda}_{\mathrm{PO}\to S}\right\|_2 \|\Phi\|_2 \varepsilon_{\theta}\left\|d^{\lambda}_{\mathrm{PO}} - d^{\lambda}_{S}\right\|_2 + \frac{\lambda}{2}\left\|d^{\lambda}_{\mathrm{PO}\to S}\right\|_{\Phi\Phi^{\top}} \\
&\leq \left(\frac{3\gamma\varepsilon_{\mu}M\sqrt{D}}{(1-\gamma)^2} + \varepsilon_{\theta}\sqrt{M}\right)\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2 + \frac{\lambda}{2}\left\|d^{\lambda}_{\mathrm{PO}\to S}\right\|_{\Phi\Phi^{\top}} \\
&\leq \underbrace{\left(\frac{3\gamma\varepsilon_{\mu}M\sqrt{D}}{(1-\gamma)^2} + \varepsilon_{\theta}\sqrt{M}\right)}_{:=\Delta}\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2 + \frac{\lambda M}{2(1-\gamma)^2}
\end{aligned}$$

Using the lower bound established in eq. (25) we obtain the following inequality.

$$\frac{\kappa\lambda}{2}\left\|d^{\lambda}_{S} - d^{\lambda}_{\mathrm{PO}\to S}\right\|_2^2 \leq \Delta\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2 + \frac{\lambda M}{2(1-\gamma)^2}$$

Now there are two cases to consider. First, $\Delta\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2 \leq \frac{\lambda M}{2(1-\gamma)^2}$. Then the upper bound on $\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2$ is $\frac{\lambda M}{2\Delta(1-\gamma)^2}$. Second, $\Delta\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2 > \frac{\lambda M}{2(1-\gamma)^2}$. Then we have $\frac{\kappa\lambda}{2}\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2^2 \leq 2\Delta\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2$ and the upper bound on $\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2$ is $\frac{4\Delta}{\kappa\lambda}$.

Let $\lambda_0 = \frac{25}{8\sqrt{\kappa}}\left(\varepsilon_{\theta} + \gamma\sqrt{D}\varepsilon_{\mu}\frac{\sqrt{M}}{\sqrt{A}(1-\gamma)}\right)$ be the required lower bound on λ. From the definition of Δ we have, $\frac{\Delta}{\sqrt{M}} \geq \varepsilon_{\theta} + \gamma\sqrt{D}\varepsilon_{\mu}\frac{\sqrt{M}}{\sqrt{A}(1-\gamma)} = \lambda_0\cdot\frac{8\sqrt{\kappa}}{25}$. Therefore, the inequality $c\cdot\Delta \geq \lambda \geq \lambda_0$ is feasible as long as $c \geq \frac{25}{8\sqrt{\kappa M}}$. Now note that, if $\lambda < \frac{2\sqrt{2}\Delta(1-\gamma)}{\sqrt{\kappa M}}$ then $\frac{4\Delta}{\kappa\lambda} > \frac{\lambda M}{2\Delta(1-\gamma)^2}$ and the upper bound on $\left\|d^{\lambda}_{\mathrm{PO}\to S} - d^{\lambda}_{S}\right\|_2$ is $\frac{4\Delta}{\kappa\lambda}$. Therefore, we need the constant c to satisfy the following inequality.

$$c \geq \max\left\{\frac{25}{8\sqrt{\kappa M}}, \frac{2\sqrt{2}(1-\gamma)}{\sqrt{\kappa M}}\right\}$$

Therefore, it is sufficient to take $c \geq 4/\sqrt{\kappa M}$. □

D PROOF OF THEOREM 4

Proof. We first construct the dual problem of the optimization problem eq. (8). Let $\mathcal{F}_t$ be the function defined as $\mathcal{F}_t(g,\omega) = \max_{d\geq 0,\nu} \mathcal{L}_t(d,\nu;g,\omega)$. Then the dual optimization problem defined in eq. (8) is given as

$$\min_{g,\omega} \mathcal{F}_t(g,\omega)$$

Fix a choice of g and ω. An optimal solution ν satisfies

$$\nabla_\nu \mathcal{F}_t(d,\nu;g,\omega) = \theta_t - \lambda\nu + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega = 0 \;\text{ and }\; \nu = \frac{1}{\lambda}\left(\theta_t + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega\right)$$

On the other hand, the derivative with respect to d is given as follows.

$$\nabla_d \mathcal{F}_t(d,\nu;g,\omega) = -B^\top g + \Phi\omega$$

If any entry of $-B^\top g + \Phi\omega$ is positive, we can choose the corresponding entry of d to be arbitrarily large, and the value $\mathcal{F}_t(g,\omega)$ would be unbounded. Therefore, we must have $-B^\top g + \Phi\omega \leq 0$. Now substituting the choice of ν derived above we get the following dual optimization problem.

$$\begin{aligned}\min_{g,\omega}\ & \frac{1}{2\lambda}\left\|\theta_t + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega\right\|_2^2 + \langle g,\rho\rangle\\ \text{s.t.}\ & -B^\top g + \Phi\omega \leq 0\end{aligned}$$

We now apply Lemma 10 to bound the norm of the optimal solution to the above optimization problem. Note that the objective can be written in the form $x^\top A_t x + b_t^\top v$ where

$$A_t = \frac{1}{2\lambda}\cdot\begin{bmatrix}\mathrm{Id}_D & -\gamma\cdot\boldsymbol{\mu}_t^\top\\ -\gamma\cdot\boldsymbol{\mu}_t & \gamma^2\cdot\boldsymbol{\mu}_t\boldsymbol{\mu}_t^\top\end{bmatrix}$$

and

$$b_t = \frac{1}{\lambda}\cdot[-\theta_t;\ \lambda\rho + \boldsymbol{\mu}_t\theta_t].$$

Suppose the eigenvalues of $\boldsymbol{\mu}_t\boldsymbol{\mu}_t^\top$ are $\sigma_1,\ldots,\sigma_D$. Then we claim that the eigenvalues of A_t are $\frac{1}{2\lambda},\ldots,\frac{1}{2\lambda},\frac{\gamma^2}{2\lambda}\sigma_1,\ldots,\frac{\gamma^2}{2\lambda}\sigma_D$. Indeed, let v_i be the i-th eigenvector of $\boldsymbol{\mu}_t\boldsymbol{\mu}_t^\top$ and let $u_i = [0_D, v_i]$. Then $u_i^\top A_t u_i = \frac{\gamma^2}{2\lambda}u_i^\top\boldsymbol{\mu}_t\boldsymbol{\mu}_t^\top u_i = \frac{\gamma^2}{2\lambda}\sigma_i\|u_i\|_2^2$. Therefore, the smallest positive eigenvalue of the matrix A_t, denoted as $\sigma^\star_{\min}(A_t)$ is bounded below as

$$\sigma_t = \sigma^\star_{\min}(A_t) \geq \frac{\min\left\{1,\gamma^2\cdot\sigma^\star_{\min}(\boldsymbol{\mu}_t\boldsymbol{\mu}_t^\top)\right\}}{2\lambda} \geq \frac{\min\left\{1,\gamma^2\underline{\sigma}\right\}}{2\lambda}$$

where $\underline{\sigma} = \min_d \sigma^\star_{\min}\boldsymbol{\mu}_d\boldsymbol{\mu}_d^\top$. Furthermore, $\|b\|_2 \leq \frac{1}{\lambda}\left(\|\theta_t\|_2 + \lambda\|\rho\|_2 + \|\boldsymbol{\mu}_t\theta_t\|_2\right) \leq \frac{1}{\lambda}\left(\sqrt{D} + \lambda + D\right)$. Therefore, we can apply Lemma 10 to obtain the following bound on the optimal dual solution.

$$\left\|\omega^\star\right\|_2 + \left\|g^\star\right\|_2 \leq \frac{\lambda + 2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}$$

We will write c_1 as $\min\left\{1,\gamma^2\underline{\sigma}\right\}$.

Let $(d_t^\star,\nu_t^\star;g_t^\star,\omega_t^\star) \in \arg\max_{d,\nu}\min_{g,\omega}\mathcal{L}_t(d,\nu;g,\omega)$. In the previous paragraph, we showed that it is sufficient to consider $\left\|\omega_t^\star\right\|_2, \left\|g_t^\star\right\|_2 \leq \frac{\lambda+2D}{\gamma^2\cdot\underline{\sigma}}$. Since $d_t^\star$ is an occupancy measure and $\sum_{s,a} d_t^\star(s,a) = \frac{1}{1-\gamma}$ we have

$$\sqrt{\sum_{s,a}\left(d_t^\star(s,a)\right)^2} = \frac{1}{1-\gamma}\sqrt{\sum_{s,a}\left((1-\gamma)\cdot d_t^\star(s,a)\right)^2} \leq \frac{1}{1-\gamma}\sqrt{\sum_{s,a}(1-\gamma)\cdot d_t^\star(s,a)} = \frac{1}{1-\gamma}.$$

On the other hand, note that $\nu_t^\star = \Phi^\top d_t^\star$ and by assumption 4 we obtain the following bound.

$$\left\|\Sigma_t^{-1}\nu_t^\star\right\|_2^2 = \left\|\Sigma_t^{-1}\Phi^\top d_t^\star\right\|_2^2 = {d_t^\star}^\top\Phi\Sigma_t^{-2}\Phi^\top d_t^\star = \mathbb{E}_{(s,a)\sim d_t^\star}[\phi(s,a)^\top]\Sigma_t^{-2}\mathbb{E}_{(s,a)\sim d_t^\star}[\phi(s,a)] \leq B$$

Therefore, $\left\|\Sigma_t^{-1}\nu_t^\star\right\|_2 \le \sqrt{B}$ and for solving the empirical Lagrangian we can restrict the parameters so that $\left\|\Sigma_t^{-1}\nu_t^\star\right\|_2 \le \sqrt{B}$. Therefore, we can apply Lemma 6 with $m_t = O\left(\frac{D^5B\lambda^4}{(1-\gamma)^2c_1^4\varepsilon^2}\log\frac{DB\lambda t}{c_1\varepsilon p}\right)$, and obtain that the following bound holds with probability at least $1-\frac{p}{t^2}\cdot\frac{6}{2\pi^2}$.

$$\mathcal{L}_t(d_t^\star,\nu_t^\star;g_t^\star,\omega_t^\star) - \mathcal{L}_t(\widehat{d}_t,\widehat{\nu}_t;g_t^\star,\omega_t^\star) \le 2\varepsilon \tag{27}$$

Therefore, by a union bound the bound in eq. holds for any t with probability at least $1-\sum_t\frac{p}{t^2}\cdot\frac{6}{2\pi^2} = 1-p/2$. Now observe that the objective $\mathcal{L}_t(d,\nu;g,\omega)$ is strongly concave in ν as $\nabla_\nu^2\mathcal{L}_t(d,\nu;g,\omega) = -\lambda\cdot\mathrm{Id}_D$. Since given $g_t^\star,\omega_t^\star$, $(d_t^\star,\nu_t^\star)$ is an optimal solution of $\mathcal{L}_t(\cdot,g_t^\star,\omega_t^\star)$ we have,

$$\left\|\nu_t^\star-\widehat{\nu}_t\right\|_2 \le \sqrt{\frac{\mathcal{L}_t(d_t^\star,\nu_t^\star;g_t^\star,\omega_t^\star)-\mathcal{L}_t(\widehat{d}_t,\widehat{\nu}_t;g_t^\star,\omega_t^\star)}{2\lambda}} \le \sqrt{\frac{\varepsilon}{\lambda}} \tag{28}$$

Now note that $\nu_t^\star = \Phi^\top d_t^\star$, but $\widehat{\nu}_t \neq \Phi^\top\widehat{d}_t$. However, observe that given $\widehat{d},\widehat{\nu},\widehat{g},\widehat{\omega}$ is an optimal solution to the optimization problem $\min_\omega \widehat{\mathcal{L}}_t(\widehat{d},\widehat{\nu},\widehat{g},\omega)$. Therefore, $\nabla_\omega\widehat{\mathcal{L}}_t(\widehat{d},\widehat{\nu},\widehat{g},\widehat{\omega}) = 0$ and we obtain the following equality.

$$\frac{1}{m_t}\sum_{j=1}^{m_t}\phi(s_j,a_j)\phi(s_j,a_j)^\top\Sigma_t^{-1}\widehat{\nu}_t = \Phi^\top\widehat{d}_t$$

This gives us the following bound with probability at least $1-\frac{p}{t^2}\cdot\frac{6}{2\pi^2}$

$$\begin{aligned}
\left\|\widehat{\nu}_t-\Phi^\top\widehat{d}_t\right\|_2 &= \left\|\left(\frac{1}{m_t}\sum_{j=1}^{m_t}\phi(s_j,a_j)\phi(s_j,a_j)^\top-\mathrm{Id}_d\right)\Sigma_t^{-1}\widehat{\nu}\right\|_2\\
&\le \left\|\left(\frac{1}{m_t}\sum_{j=1}^{m_t}\phi(s_j,a_j)\phi(s_j,a_j)^\top-\mathrm{Id}_d\right)\right\|\left\|\Sigma_t^{-1}\widehat{\nu}\right\|_2\\
&\le 8\sqrt{B}\max\left\{\sqrt{\frac{D+\log(t/p)}{m_t}},\frac{D+\log(t/p)}{m_t}\right\}
\end{aligned}$$

The last inequality uses standard concentration inequality for sample covariance matrix (e.g. see theorem 1.6.2 of Tropp et al. (2015)). Now substituting $m_t \ge O\left(\frac{D^5B\lambda^4}{(1-\gamma)^2c_1^4\varepsilon^2}\log\frac{DB\lambda t}{c_1\varepsilon p}\right)$ we obtain $\left\|\widehat{\nu}_t-\Phi^\top\widehat{d}_t\right\|_2 \le 8\sqrt{B}\varepsilon$ with probability at least $1-\frac{p}{t^2}\cdot\frac{6}{2\pi^2}$. Therefore, by a union bound, we obtain that with probability at least $1-p/2$, for any t we have $\left\|\widehat{\nu}_t-\Phi^\top\widehat{d}_t\right\|_2 \le 8\sqrt{B}\varepsilon$. This result, along with eq. (28) gives us the following inequality.

$$\sqrt{\kappa}\left\|d_t^\star-\widehat{d}_t\right\| \le \left\|\Phi^\top d_t^\star-\Phi^\top\widehat{d}_t\right\| \le \left\|\nu_t^\star-\widehat{\nu}_t\right\|_2+\left\|\widehat{\nu}_t-\Phi^\top\widehat{d}_t\right\|_2 \le \sqrt{\frac{\varepsilon}{\lambda}}+8\sqrt{B}\varepsilon$$

After rearranging we obtain,

$$\left\|d_t^\star-\widehat{d}_t\right\| \le \sqrt{\frac{\varepsilon}{\lambda\kappa}}+8\sqrt{\frac{B}{\kappa}}\varepsilon. \tag{29}$$

The proof of theorem equation 1 establishes the following recurrence relation.

$$\left\|d_{t+1}^\star-d_S\right\|_2 \le \beta\left(\left\|d_t^\star-d_S\right\|_2+\left\|d_{t-1}^\star-d_S\right\|_2\right) \tag{30}$$

for a constant β. Using the two inequalities above, we can establish a recurrence relation on the norm of the difference between $\widehat{d}_t$ and d_S.

$$\begin{aligned}
\left\|\widehat{d}_{t+1}-d_S\right\|_2 &\le \left\|d_{t+1}^\star-\widehat{d}_{t+1}\right\|+\left\|d_{t+1}^\star-d_S\right\|_2\\
&\le \beta\left(\left\|d_t^\star-d_S\right\|_2+\left\|d_{t-1}^\star-d_S\right\|_2\right)+\sqrt{\frac{\varepsilon}{\lambda\nu}}+8\sqrt{\frac{B}{\nu}}\varepsilon\\
&\le \beta\left(\left\|\widehat{d}_t-d_S\right\|_2+\left\|\widehat{d}_{t-1}-d_S\right\|_2\right)+3\sqrt{\frac{\varepsilon}{\lambda\kappa}}+24\sqrt{\frac{B}{\kappa}}\varepsilon
\end{aligned}$$

Now if $\lambda > \frac{1}{\sqrt{\kappa B}}$ and $\varepsilon < \frac{\beta^2\delta^2}{48}\sqrt{\frac{\kappa}{B}}$ then it can be easily checked that the recurrence relation is the following.

$$\left\|\widehat{d}_{t+1} - d_S\right\|_2 \leq \beta\left(\left\|\widehat{d}_t - d_S\right\|_2 + \left\|\widehat{d}_{t-1} - d_S\right\|_2\right) + \beta\delta$$

Now suppose $\beta < 1/3$. Then there are two cases to consider. First, if $\max\left\{\left\|\widehat{d}_t - d_S\right\|_2, \left\|\widehat{d}_{t-1} - d_S\right\|_2\right\} < \delta$ then $\left\|\widehat{d}_{t+1} - d_S\right\|_2 < \delta$ and for all subsequent $t' > t+1$ we also have $\left\|\widehat{d}_{t'} - d_S\right\|_2 < \delta$. Otherwise, we have $\left\|\widehat{d}_{t+1} - d_S\right\|_2 \leq 2\beta\left(\left\|\widehat{d}_t - d_S\right\|_2 + \left\|\widehat{d}_{t-1} - d_S\right\|_2\right)$. Now following an argument very similar to the proof of Theorem equation 1 we can establish that $\left\|\widehat{d}_{t+1} - d_S\right\|_2 \leq \frac{2}{1-\gamma}r^t$ for $r = \beta\left(1 + \frac{1}{2}\sqrt{1+\frac{2}{\beta}}\right)$. As $\beta < 1/3$ we have $r < \beta\left(1 + \frac{1}{2} + \frac{1}{\sqrt{2\beta}}\right) = \frac{3\beta}{2} + \sqrt{\frac{\beta}{2}} < 1$. Therefore, as long as $t \geq \ln\left(\frac{2}{\delta(1-\gamma)}\right)/\ln(1/r)$, we are guaranteed that $\left\|\widehat{d}_t - d_S\right\|_2 \leq \delta$.

Now we determine the sufficient conditions for ensuring $\beta < \frac{1}{3}$ and $\varepsilon < \frac{\beta^2\delta^2}{48}\sqrt{\frac{\kappa}{B}}$. The proof of theorem equation 1 establishes

$$\beta = \frac{\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu}{\lambda\sqrt{\kappa}} + \frac{4\gamma\varepsilon_\mu\alpha^2}{\sqrt{\kappa}}$$

Therefore, if $\lambda > \frac{6(\varepsilon_\theta + \alpha\gamma\sqrt{D}\varepsilon_\mu)}{\sqrt{\kappa B}}$ and $\varepsilon_\mu < \frac{\sqrt{\kappa}}{24\gamma\alpha^2}$ then $\beta < 1/3$. Finally for the upper bound on ε we need the number of samples $m_t = O\left(\frac{D^5B\lambda^4}{(1-\gamma)^2c_1^4\varepsilon^2}\log\frac{DB\lambda t}{c_1\varepsilon p}\right) = O\left(\frac{D^5B^2\lambda^4}{(1-\gamma)^2c_1^4\delta^4\kappa}\log\frac{DB\lambda t}{c_1\delta\kappa p}\right)$. □

Lemma 6. *Let us define the saddle points* $(d^\star, \nu^\star, g^\star, \omega^\star)$ *and* $(\widehat{d}, \widehat{\nu}, \widehat{g}, \widehat{\omega})$ *as*

$$(d^\star, \nu^\star; g^\star, \omega^\star) \in \arg\max_{d,\nu}\min_{g,\omega}\mathcal{L}_t(d,\nu;g,\omega), \text{ and } (\widehat{d},\widehat{\nu};\widehat{g},\widehat{\omega}) \in \arg\max_{d,\nu}\min_{g,\omega}\widehat{\mathcal{L}}_t(d,\nu;g,\omega).$$

Suppose $\max\left\{\left\|\Sigma_t^{-1}\nu^\star\right\|_2, \left\|\Sigma_t^{-1}\widehat{\nu}\right\|_2\right\} \leq \sqrt{B}$, *and* $\max\left\{\left\|g^\star\right\|_2, \left\|\widehat{g}\right\|_2, \left\|\omega^\star\right\|_2, \left\|\widehat{\omega}\right\|_2\right\} \leq \frac{\lambda+2D}{\min\{1,\gamma^2\cdot\underline{\sigma}\}}$, *and the number of samples* $m_t \geq O\left(\frac{D^5B\lambda^4}{(1-\gamma)^2c_1^4\varepsilon^2}\log\frac{DB\lambda}{c_1\varepsilon\delta_0}\right)$ *where* $c_1 = \min\left\{1, \gamma^2\cdot\sigma^\star_{\min}(\boldsymbol{\mu}_d\boldsymbol{\mu}_d^\top)\right\}$. *Then with probability at least* $1-\delta_0$ *we have,*

$$\mathcal{L}_t(d^\star, \nu^\star; g^\star, \omega^\star) - \mathcal{L}_t(\widehat{d}, \widehat{\nu}; g^\star, \omega^\star) \leq 2\varepsilon.$$

Proof. First, note that the expected value of the empirical Lagrangian equals $\mathcal{L}(d, v; g, \omega)$.

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_{\pi_t}}\left[\widehat{\mathcal{L}}_t(d, v; g, \omega)\right] \\
&= v^\top \Sigma_t^{-1} \cdot \frac{1}{m_t}\sum_{j=1}^{m_t} \mathbb{E}\left[\phi(s_j, a_j)\left(r_t(s_j, a_j) + \gamma \cdot g(s_j') - \phi(s_j, a_j)^\top \omega\right)\right] \\
&- \frac{\lambda}{2} v^\top v + \frac{1}{m_t}\sum_{j=1}^{m_t} \mathbb{E}\left[g(s_j^0)\right] + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top \Sigma_t^{-1} \cdot \frac{1}{m_t}\sum_{j=1}^{m_t} \mathbb{E}\left[\phi(s_j, a_j)\left(\phi(s_j, a_j)^\top \theta_t + \sum_{s'} P(s' \mid s_j, a_j) g(s') - \phi(s_j, a_j)^\top \omega\right)\right] \\
&- \frac{\lambda}{2} v^\top v + \frac{1}{m_t}\sum_{j=1}^{m_t} \sum_{s} \rho(s) g(s) + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top \Sigma_t^{-1} \cdot \frac{1}{m_t}\sum_{j=1}^{m_t} \mathbb{E}\left[\phi(s_j, a_j)\left(\phi(s_j, a_j)^\top \theta_t + \sum_{s'} \phi(s_j, a_j)^\top \boldsymbol{\mu}_t(s') g(s') - \phi(s_j, a_j)^\top \omega\right)\right] \\
&- \frac{\lambda}{2} v^\top v + \rho^\top g + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top \Sigma_t^{-1} \cdot \frac{1}{m_t}\sum_{j=1}^{m_t} \mathbb{E}\left[\phi(s_j, a_j)\phi(s_j, a_j)^\top\right](\theta_t + \gamma \cdot \boldsymbol{\mu}_t g - \omega) \\
&- \frac{\lambda}{2} v^\top v + \rho^\top g + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top (\theta_t + \gamma \cdot \boldsymbol{\mu}_t g - \omega) - \frac{\lambda}{2} v^\top v + \rho^\top g + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= \mathcal{L}_t(d, v; g, \omega)
\end{aligned}
$$

Since $\|g\|_2 \le \frac{\lambda + 2D}{\min\{1, \gamma^2 \underline{\sigma}\}}$, we can apply the Chernoff-Hoeffding inequality and obtain the following bound.

$$
\Pr\left(\left|\frac{1}{m_t}\sum_{j=1}^{m_t} g(s_j^0) - \rho^\top g\right| \ge \frac{\lambda + 2D}{\min\left\{1, \gamma^2 \underline{\sigma}\right\}} \sqrt{\frac{\log(4/\delta_1)}{m_t}}\right) \le \frac{\delta_1}{2} \tag{31}
$$

Moreover, for any j we have,

$$
\begin{aligned}
&v^\top \Sigma_t^{-1} \phi(s_j, a_j)\left(r_t(s_j, a_j) + \gamma \cdot g(s_j') - \phi(s_j, a_j)^\top \omega\right) \\
&\le \left\|v^\top \Sigma_t^{-1}\right\|_2 \left\|\phi(s_j, a_j)\right\|_2 \left|r_t(s_j, a_j) + \gamma \cdot g(s_j') - \phi(s_j, a_j)^\top \omega\right| \\
&\le \sqrt{BD}\left(\left|r_t(s_j, a_j)\right| + \gamma \cdot \left|g(s_j')\right| + \left\|\phi(s_j, a_j)\right\|_2 \|\omega\|_2\right) \\
&\le \sqrt{BD}\left(1 + \gamma \cdot \frac{\lambda + 2D}{\min\left\{1, \gamma^2 \underline{\sigma}\right\}} + \sqrt{D} \cdot \frac{\lambda + 2D}{\min\left\{1, \gamma^2 \underline{\sigma}\right\}}\right) = \underbrace{\sqrt{BD}\left(1 + \frac{(\gamma + \sqrt{D})(\lambda + 2D)}{\min\left\{1, \gamma^2 \underline{\sigma}\right\}}\right)}_{:=H}
\end{aligned}
$$

Then we can apply Chernoff-Hoeffding inequality and obtain the following bound.

$$
\begin{aligned}
&\Pr\left(\left|\frac{1}{m_t}\sum_{j=1}^{m_t} v^\top \Sigma_t^{-1} \phi(s_j, a_j)\left(r_t(s_j, a_j) + \gamma \cdot g(s_j') - \phi(s_j, a_j)^\top \omega\right) - v^\top\left(\theta_t + \gamma \cdot \boldsymbol{\mu}_t^\top g - \omega\right)\right|\right. \\
&\left. \ge H \sqrt{\frac{\log(4/\delta_1)}{m_t}}\right) \le \frac{\delta_1}{2}
\end{aligned} \tag{32}
$$

Using the bounds derived in equations equation 31 and equation 32 we get the following bound.

$$
\Pr\left(\left|\widehat{\mathcal{L}}_t(d, v; g, \omega) - \mathcal{L}_t(d, v; g, \omega)\right| \ge \left(H + \frac{\lambda + 2D}{\min\left\{1, \gamma^2 \underline{\sigma}\right\}}\right) \sqrt{\frac{\log(4/\delta_1)}{m_t}}\right) \le \delta_1 \tag{33}
$$

Note that the difference term $\left|\widehat{\mathcal{L}}_t(d,\nu;g,\omega)-\mathcal{L}_t(d,\nu;g,\omega)\right|$ is independent of d and the bound derived in eq. (33) holds for any d. We now extend the bound for any ν and ω.

We can assume that $\omega\in\Omega=\left\{\omega:\|\omega\|_2\le\frac{\lambda+2D}{\min\{1,\gamma^2\underline{\sigma}\}}\right\}$. By lemma 5.2 of Vershynin (2010) there is an ε-net Ω_ε of size at most $\left(1+\frac{2(\lambda+2D)}{\min\{1,\gamma^2\underline{\sigma}\}\cdot\varepsilon}\right)^D$, of the set Ω so that for any ω there exists ω' so that $\|\omega-\omega'\|_2\le\varepsilon$.

On the other hand, $\nu\in\mathcal{V}=\left\{\nu:\left\|\Sigma_t^{-1}\nu\right\|_2\le\sqrt{B}\right\}$. In order to construct an ε-net of the set $\mathcal{V}$, consider the set $\left\{y:\|y\|\le\sqrt{B}\right\}$. There is an ε-net C_ε of this set of cardinality $\left(1+\frac{2\sqrt{B}}{\varepsilon}\right)^D$. Now consider the following set $\mathcal{V}_\varepsilon=\{\Sigma_t y:y\in C_\varepsilon\}$. Given any $\nu\in\mathcal{V}$ we know there exists $\widetilde{y}\in C_\varepsilon$ so that $\left\|\Sigma_t^{-1}\nu-\widetilde{y}\right\|_2\le\varepsilon$. Equivalently, there exists $\widetilde{\nu}$ so that $\left\|\Sigma_t^{-1}(\nu-\widetilde{\nu})\right\|_2\le\varepsilon$.

Moreover, for a pair of $(\omega,\nu)\in\Omega_\varepsilon\times\mathcal{V}_\varepsilon$, we will fix a choice of g which is given as the maximizer of the following optimization problem.

$$g=g(\omega,\nu)\in\operatorname*{arg\,min}_{g':\|g'\|_2\le G}\widehat{\mathcal{L}}(d,\nu;g',\omega)\tag{34}$$

Note that, from the definition of the empirical Lagrangian equation 10, we can choose $g(\omega,\nu)$ to be G times unit vector with support in the set $\left\{s_j^0:j\in[m_t]\right\}$ or $\cup\left\{s_j':j\in[m_t]\right\}$. In particular let $\mathcal{G}=\left\{G\cdot\mathbf{1}_s:s\in\left\{s_j^0,s_j'\right\},\ j\in[m_t]\right\}$. Now, by union bound over $|\Omega_\varepsilon|\times|\mathcal{V}_\varepsilon|$ tuples, we can extend eq. (33) i.e. the following event holds with probability at least $1-\delta_0$ for any $\omega\in\Omega_\varepsilon$, $\nu\in\mathcal{V}_\varepsilon$, and $g\in\mathcal{G}$ defined above.

$$\left|\widehat{\mathcal{L}}_t(d,\nu;g,\omega)-\mathcal{L}_t(d,\nu;g,\omega)\right|\le\underbrace{\left(H+\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}\right)\sqrt{\frac{D}{m_t}}\sqrt{\log\left(\frac{32DB(\lambda+2D)m_t}{\min\left\{1,\gamma^2\underline{\sigma}\right\}\cdot\varepsilon\cdot\delta_0}\right)}}_{:=f(\varepsilon)}\tag{35}$$

Now given any ν,ω let us pick $\widetilde{\omega}\in\Omega_\varepsilon$ and $\widetilde{\nu}\in\mathcal{V}_\varepsilon$ so that $\|\omega-\widetilde{\omega}\|_2\le\varepsilon$ and $\left\|\Sigma_t^{-1}(\nu-\widetilde{\nu})\right\|_2\le\varepsilon$. Moreover, we define g as $g=g(\omega,\nu)$ (as defined in eq. (34)). Then we have,

$$\begin{aligned}\mathcal{L}_t(d,\nu;g,\omega)-\widehat{\mathcal{L}}_t(d,\nu;g,\omega)\le{}&|\mathcal{L}_t(d,\nu;g,\omega)-\mathcal{L}_t(d,\widetilde{\nu};g,\widetilde{\omega})|+\left|\mathcal{L}_t(d,\widetilde{\nu};g,\widetilde{\omega})-\widehat{\mathcal{L}}_t(d,\widetilde{\nu};g,\widetilde{\omega})\right|\\&+\left|\widehat{\mathcal{L}}_t(d,\nu;g,\omega)-\widehat{\mathcal{L}}_t(d,\widetilde{\nu};g,\widetilde{\omega})\right|\end{aligned}$$

Since $g\in\mathcal{G}$, the second term is bounded by $f(\varepsilon)$, by eq. (35). Moreover, we can apply Lemma 7 to bound the first and the third term as follows.

$$|\mathcal{L}_t(d,\nu;g,\omega)-\mathcal{L}_t(d,\widetilde{\nu};g,\widetilde{\omega})|\le\varepsilon D\left(\sqrt{D}+\left(1+\gamma\sqrt{D}\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}\right)\cdot\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}+\lambda D\sqrt{B}\right)+\varepsilon\left(D\sqrt{B}+\frac{1}{1-\gamma}\right)$$

and

$$\left|\widehat{\mathcal{L}}_t(d,\nu;g,\omega)-\widehat{\mathcal{L}}_t(d,\widetilde{\nu};g,\widetilde{\omega})\right|\le\varepsilon\left(\sqrt{D}+\left(1+\gamma\sqrt{D}\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}\right)\cdot\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}+\lambda\cdot D^2\sqrt{B}\right)+\varepsilon\left(\sqrt{B}+\frac{1}{1-\gamma}\right)$$

Suppose m_t is chosen so that

$$\frac{D}{m_t}\log\frac{64D^2Bm_t}{\min\left\{1,\gamma^2\underline{\sigma}\right\}\varepsilon\delta_0}\le\varepsilon_1^2\tag{36}$$

then it can be easily seen that $f(\varepsilon_1)\le\left(H+\frac{\lambda+2D}{\min\{1,\gamma^2\underline{\sigma}\}}\right)\varepsilon_1\sqrt{2\log\lambda}$. Then the following bound holds for any $\nu\in\mathcal{V}$, $\omega\in\Omega$

and $g = g(\omega, \nu)$.

$$\begin{aligned}
\left|\mathcal{L}_t(d,\nu;g,\omega) - \widehat{\mathcal{L}}_t(d,\nu;g,\omega)\right| &\leq 2\varepsilon_1\left(H + \frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}\right)\left(1+\gamma\sqrt{D}\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}} + \sqrt{2\log\lambda}\right) \\
&\quad + 2\varepsilon_1\left(\lambda D^2\sqrt{B} + D^{3/2}\sqrt{B} + \frac{1}{1-\gamma}\right) \\
&\leq 8\varepsilon_1\sqrt{B}D^{3/2}\left(\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}\right)^2 + \frac{4\varepsilon\cdot\lambda D^2\sqrt{B}}{1-\gamma} \\
&\leq \frac{16\varepsilon_1 D^2\sqrt{B}}{1-\gamma}\left(\frac{\lambda+2D}{\min\left\{1,\gamma^2\underline{\sigma}\right\}}\right)^2
\end{aligned}$$

Now we can substitute $\varepsilon = \frac{16\varepsilon_1 D^2\sqrt{B}}{1-\gamma}\left(\frac{\lambda+2D}{\min\{1,\gamma^2\underline{\sigma}\}}\right)^2$ in eq. (36), and apply lemma 8 to obtain the required bound. □

Lemma 7. *Suppose, we are given $\nu, \widetilde{\nu}$ such that $\left\|\Sigma^{-1}(\nu-\widetilde{\nu})\right\|_2 \leq \varepsilon_1$ and $\max\left\{\left\|\Sigma^{-1}\nu\right\|_2, \left\|\Sigma^{-1}\widetilde{\nu}\right\|_2\right\} \leq V$, and $\omega, \widetilde{\omega}$ such that $\|\omega-\widetilde{\omega}\|_2 \leq \varepsilon_2$, and $\max\{\|\omega\|_2, \|\widetilde{\omega}\|_2\} \leq W$. Then for any g with $\|g\|_2 \leq G$ we have,*

$$|\mathcal{L}(d,\nu;g,\omega) - \mathcal{L}(d,\widetilde{\nu};g,\widetilde{\omega})| \leq \varepsilon_1 D\left(\sqrt{D} + \gamma\sqrt{D}G + W + \lambda DV\right) + \varepsilon_2\left(DV + \frac{1}{1-\gamma}\right)$$

and

$$\left|\widehat{\mathcal{L}}(d,\nu;g,\omega) - \widehat{\mathcal{L}}(d,\widetilde{\nu};g,\widetilde{\omega})\right| \leq \varepsilon_1\left(\sqrt{D} + \gamma G + W + \lambda\cdot D^2 V\right) + \varepsilon_2\left(V + \frac{1}{1-\gamma}\right)$$

Proof.

$$|\mathcal{L}(d,\nu;g,\omega) - \mathcal{L}(d,\widetilde{\nu};g,\widetilde{\omega})| \leq |\mathcal{L}(d,\nu;g,\omega) - \mathcal{L}(d,\widetilde{\nu};g,\omega)| + |\mathcal{L}(d,\widetilde{\nu};g,\omega) - \mathcal{L}(d,\widetilde{\nu};g,\widetilde{\omega})|$$

Using the definition of $\mathcal{L}(\cdot)$ eq. (9) we obtain the following bound.

$$\begin{aligned}
&|\mathcal{L}(d,\nu;g,\omega) - \mathcal{L}(d,\widetilde{\nu};g,\widetilde{\omega})| \leq \left|(\nu-\widetilde{\nu})^\top\left(\theta_t + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega\right) - \frac{\lambda}{2}\left(\|\nu\|_2^2 - \|\widetilde{\nu}\|_2^2\right)\right| \\
&+ \left|(\widetilde{\omega}-\omega)^\top\widetilde{\nu} + d^\top\Phi(\omega-\widetilde{\omega})\right| \\
&\leq \|\nu-\widetilde{\nu}\|_2\left\|\theta_t + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega - \frac{\lambda}{2}(\nu+\widetilde{\nu})\right\|_2 + \|\omega-\widetilde{\omega}\|_2\left\|-\widetilde{\nu} + \Phi^\top d\right\|_2 \\
&\leq D\varepsilon_1\left(\sqrt{D} + \gamma\sqrt{D}G + W + \lambda DV\right) + \varepsilon_2\left(DV + \frac{1}{1-\gamma}\right)
\end{aligned}$$

The last inequality uses (a) $\|\nu-\widetilde{\nu}\|_2 = \left\|\Sigma\Sigma^{-1}(\nu-\widetilde{\nu})\right\|_2 \leq \varepsilon_1\|\Sigma\|_2 \leq D\varepsilon_1$, (b) for a linear MDP, $\|\theta_t\|_2 \leq \sqrt{D}$ and $\|\boldsymbol{\mu}_t\|_2 \leq \sqrt{D}$. Second, $\left\|\Phi^\top d\right\|_2 = \left\|\sum_{s,a} d(s,a)\phi(s,a)\right\|_2 \leq \sum_{s,a} d(s,a)\|\phi(s,a)\|_2 \leq \frac{1}{1-\gamma}$.

Now let us write $\phi_j = \phi(s_j,a_j)$, and $r_j = r(s_j,a_j)$. Then from the definition of empirical Lagrangian equation 10 we obtain

the following bound.

$$\begin{aligned}
&\left|\widehat{\mathcal{L}}(d,\nu;g,\omega)-\widehat{\mathcal{L}}(d,\widetilde{\nu};g,\widetilde{\omega})\right| \le \left|\widehat{\mathcal{L}}(d,\nu;g,\omega)-\widehat{\mathcal{L}}(d,\widetilde{\nu};g,\omega)\right| + \left|\widehat{\mathcal{L}}(d,\widetilde{\nu};g,\omega)-\widehat{\mathcal{L}}(d,\widetilde{\nu};g,\widetilde{\omega})\right| \\
&\le \left|(\nu-\widetilde{\nu})^\top\Sigma^{-1}\cdot\frac{1}{m}\sum_{j=1}^m\phi_j\left(r_j+\gamma\cdot g(s_j')-\phi_j^\top\omega\right)-\frac{\lambda}{2}\left(\|\nu\|_2^2-\|\widetilde{\nu}\|_2^2\right)\right| \\
&+\left|\widetilde{\nu}^\top\Sigma^{-1}\cdot\frac{1}{m}\sum_{j=1}^m\phi_j\phi_j^\top\left(\widetilde{\omega}-\omega\right)\right| + |\langle d,\Phi(\omega-\widetilde{\omega})\rangle| \\
&\le \left\|(\nu-\widetilde{\nu})^\top\Sigma^{-1}\right\|_2\left(\left\|\frac{1}{m}\sum_{j=1}^m\phi_j\left(r_j+\gamma\cdot g(s_j')-\phi_j^\top\omega\right)\right\|_2+\frac{\lambda}{2}\|\Sigma(\nu+\widetilde{\nu})\|_2\right) \\
&+\left(\left\|\widetilde{\nu}^\top\Sigma^{-1}\right\|_2\left\|\frac{1}{m}\sum_{j=1}^m\phi_j\phi_j^\top\right\|_2+\left\|\Phi^\top d\right\|_2\right)\|\omega-\widetilde{\omega}\|_2 \\
&\le \varepsilon_1\left(\sqrt{D}+\gamma G+W+\lambda\cdot D^2V\right)+\varepsilon_2\left(V+\frac{1}{1-\gamma}\right)
\end{aligned}$$

The last inequality uses the following observations – (a) $\left|r_j\right| = \left|\phi_j^\top\theta_t\right| \le \left\|\phi_j\right\|_2\|\theta_t\|_2 \le \sqrt{D}$, (b) $\left\|\Phi^\top d\right\|_2 = \left\|\sum_{s,a}d(s,a)\phi(s,a)\right\|_2 \le \sum_{s,a}d(s,a)\|\phi(s,a)\|_2 \le \frac{1}{1-\gamma}$, and (c)$\|\nu\|_2 \le \|\Sigma\|_2\left\|\Sigma^{-1}\nu\right\|_2 \le DV$. □

Lemma 8. *Suppose*

$$(d^\star,\nu^\star;g^\star,\omega^\star)\in\arg\max_{d,\nu}\min_{g,\omega}\mathcal{L}(d,\nu;g,\omega)$$

and

$$(\widehat{d},\widehat{\nu};\widehat{g},\widehat{\omega})\in\arg\max_{d,\nu}\min_{g,\omega}\widehat{\mathcal{L}}(d,\nu;g,\omega).$$

Moreover, given any d,ν,ω suppose the following inequality holds.

$$\left|\mathcal{L}(d,\nu;\widehat{g},\omega)-\widehat{\mathcal{L}}(d,\nu;\widehat{g},\omega)\right|\le\varepsilon \quad \textit{where } \widehat{g}\in\arg\min_g\widehat{\mathcal{L}}(d,\nu;g,\omega)$$

Then we have,

$$\mathcal{L}(d^\star,\nu^\star;g^\star,\omega^\star)-\mathcal{L}(\widehat{d},\widehat{\nu};g^\star,\omega^\star)\le 2\varepsilon.$$

Proof. Given d,ν let us define $\widehat{g}(d,\nu)$ and $\widehat{\omega}(d,\nu)$ as follows.

$$(\widehat{g}(d,\nu),\widehat{\omega}(d,\nu))\in\arg\min_{g,\omega}\widehat{\mathcal{L}}(d,\nu;g,\omega)$$

Moreover, let $\widetilde{g}=\arg\min_g\mathcal{L}(\hat{d},\hat{\nu};g,\omega^\star)$.

$$\begin{aligned}
&\mathcal{L}(d^\star,\nu^\star;g^\star,\omega^\star)-\mathcal{L}(\widehat{d},\widehat{\nu};g^\star,\omega^\star) \\
&=\underbrace{\mathcal{L}(d^\star,\nu^\star;g^\star,\omega^\star)-\mathcal{L}(d^\star,\nu^\star;\widehat{g}(d^\star,\nu^\star),\widehat{\omega}(d^\star,\nu^\star))}_{:=T_1} \\
&+\underbrace{\mathcal{L}(d^\star,\nu^\star;\widehat{g}(d^\star,\nu^\star),\widehat{\omega}(d^\star,\nu^\star))-\widehat{\mathcal{L}}(d^\star,\nu^\star;\widehat{g}(d^\star,\nu^\star),\widehat{\omega}(d^\star,\nu^\star))}_{:=T_2} \\
&+\underbrace{\widehat{\mathcal{L}}(d^\star,\nu^\star;\widehat{g}(d^\star,\nu^\star),\widehat{\omega}(d^\star,\nu^\star))-\widehat{\mathcal{L}}(\widehat{d},\widehat{\nu};\widehat{g},\widehat{\omega})}_{:=T_3} \\
&+\underbrace{\widehat{\mathcal{L}}(\widehat{d},\widehat{\nu};\widehat{g},\widehat{\omega})-\widehat{\mathcal{L}}(\widehat{d},\widehat{\nu};\widetilde{g},\omega^\star)}_{:=T_4}+\underbrace{\widehat{\mathcal{L}}(\widehat{d},\widehat{\nu};\widetilde{g},\omega^\star)-\mathcal{L}(\widehat{d},\widehat{\nu};\widetilde{g},\omega^\star)}_{:=T_5}+\underbrace{\mathcal{L}(\widehat{d},\widehat{\nu};\widetilde{g},\omega^\star)-\mathcal{L}(\widehat{d},\widehat{\nu};g^\star,\omega^\star)}_{:=T_6}
\end{aligned}$$

Given $(d^\star,\nu^\star)$, $(g^\star,\omega^\star)$ is the optimal minimizer, and $T_1\le 0$. Given $d^\star,\nu^\star,\widehat{\omega}_\star$, $\widehat{g}(d^\star,\nu^\star)$ minimizes the function $\widehat{\mathcal{L}}(d^\star,\nu^\star;\cdot,\widehat{\omega}_\star)$, and $T_2\le\varepsilon$. Since $(\widehat{d},\widehat{\nu};\widehat{g},\widehat{\omega})$ is a saddle point, $(\widehat{d},\widehat{\nu})$ is the maximizer of the following optimization problem $\max_{d,\nu}\widehat{\mathcal{L}}(d,\nu;\widehat{g}(d,\nu),\widehat{\omega}(d,\nu))$. This implies that $T_3\le 0$.

Since $(\widehat{g}, \widehat{\omega})$ minimizes $\widehat{\mathcal{L}}(\widehat{d}, \widehat{v}; \cdot)$ the term $T_4 \leq 0$. By a similar argument as T_2, we have $T_5 \leq 0$. Finally, given $\widehat{d}, \widehat{v}, \omega^\star, \widetilde{g}$ is the minimizer of $\mathcal{L}(\widehat{d}, \widehat{v}; \cdot, \omega^\star)$ and $T_6 \leq 0$. Combining the six inequalities we get the desired result. □

Lemma 9. *Let $g \in \arg\min_{g'} \|g'\|_2^2$ s.t. $v^\top g' = c$ and $\widetilde{g} \in \arg\min_{g'} \|g'\|_2^2$ s.t. $\widetilde{v}^\top g' = \widetilde{c}$. Then $\|g - \widetilde{g}\|_2 \leq |c - \widetilde{c}| + 2\,\widehat{|c|}\,\frac{\|\widehat{v} - v\|_2}{\|\widehat{v}\|_2}$.*

Proof. The solution to the optimization problem $\min_{g'} \|g'\|_2^2$ s.t. $v^\top g' = c$ is $\frac{c}{\|v\|_2} \cdot v$. Therefore,

$$\begin{aligned}
\|g - \widetilde{g}\|_2 &= \left\| \frac{c}{\|v\|_2} \cdot v - \frac{\widetilde{c}}{\|\widehat{v}\|_2} \cdot \widetilde{v} \right\|_2 \leq |c - \widetilde{c}|\, \|v/\|v\|_2\|_2 + \widehat{|c|} \left\| \frac{v}{\|v\|_2} - \frac{\widetilde{v}}{\|\widehat{v}\|_2} \right\|_2 \\
&\leq |c - \widetilde{c}| + \widehat{|c|}\, \frac{\|v \cdot \|\widehat{v}\|_2 - \widetilde{v} \cdot \|v\|_2\|_2}{\|v\|_2\, \|\widehat{v}\|_2} \\
&\leq |c - \widetilde{c}| + \widehat{|c|}\, \frac{\|v\,(\|\widehat{v}\|_2 - \|v\|_2) + (v - \widetilde{v}) \cdot \|v\|_2\|_2}{\|v\|_2\, \|\widehat{v}\|_2} \\
&\leq |c - \widetilde{c}| + \widehat{|c|} \left(\frac{|\|\widehat{v}\|_2 - \|v\|_2|}{\|\widehat{v}\|_2} + \frac{\|\widehat{v} - v\|_2}{\|\widehat{v}\|_2} \right) \\
&\leq |c - \widetilde{c}| + 2\,\widehat{|c|}\, \frac{\|\widehat{v} - v\|_2}{\|\widehat{v}\|_2}
\end{aligned}$$

□

Lemma 10. *Let A be a positive semidefinite matrix, Q be a matrix with full row-rank and $x^\star$ be an optimal solution of the following optimization problem.*

$$\begin{aligned}
&\min_{x \in \mathbb{R}^n} x^\top A x + b^\top x \\
&\text{s.t.} \quad Qx \geq d
\end{aligned}$$

Then we have

$$\left\|x^\star\right\|_2 \leq \frac{\|b\|_2}{2\sigma^\star_{\min}(A)} + \frac{\|A\|_2\, \|d\|_2}{\sigma^\star_{\min}(A) \lambda_{\min}(QQ^\top)},$$

where $\sigma^\star_{\min}(A)$ is the smallest positive eigenvalue of the matrix A.

Proof. Let us write q_i to denote the rows of the matrix Q. We will say that constraint i is active for $x^\star$ if $q_i^\top x^\star = d_i$. Let C be the set of rows that correspond to active constraints for $x^\star$. If $C = \emptyset$ then $x^\star$ coincides with the optimal solution of the unconstrained problem i.e. $2Ax^\star + b = 0$. In that case, $-\frac{1}{2}A^\dagger b$ is the solution of the minimum norm. If $A = U\Sigma U^\top$ is the singular value decomposition of A then $A^\dagger = U\Sigma^+ U^\top$ where Σ^+ is obtained by inverting all non-zero eigenvalues of A and leaving the zero eignevalues as they are. This gives us $\left\|x^\star\right\|_2 \leq \frac{1}{2}\left\|A^\dagger\right\|_2 \|b\|_2 \leq \frac{\|b\|_2}{2\sigma^\star_{\min}(A)}$ where $\sigma^\star_{\min}(A)$ is the smallest positive eigenvalue of A.

Now we consider the case where $|C| = r \leq n$. Then the optimal solution lies in the subspace $\Pi = \{x : Q_C x = d_C\}$ where Q_C (resp. d_C) is the submatrix with the rows of the matrix Q (resp. vector d) indexed by the set C. We can also assume that Q_C has full row rank, otherwise, we can eliminate some of the rows, and yet the subspace Π will remain the same. Any element of the subspace Π is given as $x = Q_C^\dagger d_C + [\mathrm{Id} - Q_C^\dagger Q_C]w$ for arbitrary w. Substituting this value of x we get the following unconstrained problem.

$$\min_{w \in \mathbb{R}^n} w^\top (\mathrm{Id} - Q_C^\dagger Q_C)^\top A (\mathrm{Id} - Q_C^\dagger Q_C) w + \left(b^\top + 2 d_C^\top Q_C^{\dagger\top} A \right) (I - Q_C^\dagger Q_C) w$$

At an optimal solution $w^\star$ we have the following equality.

$$2A(\mathrm{Id} - Q_C^\dagger Q_C) w^\star + (b + 2A^\top Q_C^\dagger d_C) = 0$$

The minimum norm solution to this problem is the following.

$$w^\star = -\frac{1}{2} \left(A(\mathrm{Id} - Q_C^\dagger Q_C) \right)^\dagger (b + 2A^\top Q_C^\dagger d_C)$$

and

$$\left\|w^\star\right\|_2 \le \frac{1}{2}\left\|\left(A(\mathrm{Id} - Q_C^\dagger Q_C)\right)^\dagger\right\|_2 \left(\|b\|_2 + 2\|A\|_2 \left\|Q_C^\dagger\right\|_2 \|d_C\|_2\right) \tag{37}$$

We first claim that $\left\|\left(A(\mathrm{Id} - Q_C^\dagger Q_C)\right)^\dagger\right\|_2 \le \frac{1}{\sigma^\star_{\min}(A)}$. Let $Q_C = U^1\Sigma^1 V^{1\top}$. Then $\mathrm{Id} - Q_C^\dagger Q_C = \mathrm{Id} - V^1 V^{1\top}$. Since the columns of V^1 are orthonormal the eigenvalues of $\mathrm{Id} - Q_C^\dagger Q_C$ are either 0 and 1. Moreover, we can assume that not all eigenvalues are zero, otherwise, the claim already holds. Now given two matrices A and B the smallest non-zero eigenvalue of AB i.e. $\sigma^\star_{\min}(AB)$ is bounded from below by $\sigma^\star_{\min}(A) \times \sigma^\star_{\min}(B)$ since

$$\sigma^\star_{\min}(AB) = \min_{x:ABx\neq 0} \frac{\|ABx\|_2}{\|x\|_2} = \min_{x:ABx\neq 0} \frac{\|ABx\|_2}{\|Bx\|_2}\frac{\|Bx\|_2}{\|x\|_2} \ge \sigma^\star_{\min}(A)\sigma^\star_{\min}(B).$$

Therefore,

$$\left\|\left(A(\mathrm{Id} - Q_C^\dagger Q_C)\right)^\dagger\right\|_2 \le \frac{1}{\sigma^\star_{\min}\left(A(\mathrm{Id} - Q_C^\dagger Q_C)\right)} \le \frac{1}{\sigma^\star_{\min}(A)\sigma^\star_{\min}(\mathrm{Id} - Q_C^\dagger Q_C)} = \frac{1}{\sigma^\star_{\min}(A)}$$

Substituting the bound above in eq. (37) we obtain the following bound.

$$\left\|w^\star\right\|_2 \le \frac{1}{2\sigma^\star_{\min}(A)}\left(\|b\|_2 + 2\|A\|_2 \left\|Q_C^\dagger\right\|_2 \|d\|_2\right)$$

As Q_C has full row rank we can write $Q_C = UDV^\top$ for some invertible matrix $D \in \mathbb{R}^{r\times r}$. Then $Q_C^\dagger = VD^{-1}U^\top$ and $\left\|Q_C^\dagger\right\|_2$ is inverse of the smallest singular value of Q_C i.e. $1/\lambda_{\min}(Q_C Q_C^\top)$. Now recall that the matrix Q_C was formed by removing rows of the matrix Q. Therefore, given any $x \in \mathbb{R}^r$ we can choose $y = [x\ 0_{n-r}]$ so that $QQ^\top y = Q_C Q_C^\top x$. This means that $\min_{x\neq 0} \frac{\|Q_C Q_C^\top x\|_2}{\|x\|_2} \ge \min_{y\neq 0} \frac{\|QQ^\top y\|_2}{\|y\|_2} = \lambda_{\min}(QQ^\top)$. Therefore, we can upper bound $\left\|Q_C^\dagger\right\|_2$ by $1/\lambda_{\min}(QQ^\top)$ and get the desired bound. □

E MISSING PROOFS FROM SECTION 4

E.1 Equivalent Expression of the Lagrangian

$$\begin{aligned}
\mathcal{L}(d, v; g, \omega) &= v^\top\left(\theta_t + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega\right) - \frac{\lambda}{2}v^\top v + \langle g, \rho\rangle + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top\Sigma_t^{-1}\Sigma_t\left(\theta_t + \gamma\cdot\boldsymbol{\mu}_t^\top g - \omega\right) - \frac{\lambda}{2}v^\top v + \langle g, \rho\rangle + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top\Sigma_t^{-1}\mathbb{E}_{(s,a)\sim\pi_t}\left[\phi(s,a)\phi(s,a)^\top\left(\theta_t + \gamma\cdot\boldsymbol{\mu}_t v - \omega\right)\right] - \frac{\lambda}{2}v^\top v + \mathbb{E}_{s^0\sim\rho}\left[g(s^0)\right] + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top\Sigma_t^{-1}\mathbb{E}_{(s,a)\sim\pi_t}\left[\phi(s,a)r_t(s,a) + \gamma\cdot g^\top P_t(\cdot\mid s,a) - \phi(s,a)\phi(s,a)^\top\omega\right] \\
&- \frac{\lambda}{2}v^\top v + \mathbb{E}_{s^0\sim\rho}\left[g(s^0)\right] + \left\langle d, \Phi w - B^\top g\right\rangle \\
&= v^\top\Sigma_t^{-1}\mathbb{E}_{(s,a)\sim\pi_t, s'\sim P_t(\cdot|s,a)}\left[\phi(s,a)r_t(s,a) + \gamma\cdot g(s') - \phi(s,a)\phi(s,a)^\top\omega\right] \\
&- \frac{\lambda}{2}v^\top v + \mathbb{E}_{s^0\sim\rho}\left[g(s^0)\right] + \left\langle d, \Phi w - B^\top g\right\rangle
\end{aligned}$$

E.2 Proof of Theorem 5

Definition of Regret. Given ω let us define the policy π as,

$$\pi(a\mid s) = \frac{\exp(\phi(s,a)^\top\omega)}{\sum_b \exp(\phi(s,b)^\top\omega)}$$

Moreover, given policy π we can define $d^{\pi,v}$ and $g^{\pi,\omega}$ as follows.

$$d^{\pi,v}(s,a) = \pi(a\mid s)\cdot\left(\rho(s) + \gamma\cdot\boldsymbol{\mu}_t(s)^\top v\right) \tag{38}$$

$$g^{\pi,\omega}(s) = \sum_a \pi(a\mid s)\phi(s,a)^\top\omega$$

Let us define the function $f(\pi, \nu, \omega) = \mathcal{L}_t(d^{\pi,\nu}, \nu, g^{\pi,\omega}, \omega)$. We now define the following notion of regret.

$$\mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu^\star, \pi^\star, \omega_{\ell-1}) - f(\nu_\ell, \pi_{\ell-1}, \omega^\star_{\ell-1}) \tag{39}$$

Proof. Given policies $\{\pi_{\ell-1}\}_{\ell=1}^{T_{\text{inner}}}$ let us define $d^{\widetilde{\pi}} = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} d^{\pi_{\ell-1}}$ and $\widetilde{\nu} = \Phi^\top d^{\widetilde{\pi}}$. Then $\langle \widetilde{\nu}, \theta \rangle = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \langle d^{\pi_{\ell-1}}, \Phi\theta \rangle = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \langle V^{\pi_{\ell-1}}, \rho \rangle$. Now using Lemma 11 we obtain the following bound.

$$\langle \widetilde{\nu}, \theta \rangle - \frac{\lambda}{2} \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \nu_\ell^\top \nu_\ell = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \langle V^{\pi_{\ell-1}}, \rho \rangle - \frac{\lambda}{2} \nu_\ell^\top \nu_\ell = {\nu^\star}^\top \theta - \frac{\lambda}{2} \left\| \nu^\star \right\|_2^2 - \mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) \tag{40}$$

Let us also define the policy $\pi^\dagger$ as follows $\pi^\dagger(a \mid s) = \frac{d^{\widetilde{\pi}}(s,a)}{\sum_b d^{\widetilde{\pi}}(s,b)}$. Then $\phi^\top d^{\pi^\dagger} = \Phi^\top d^{\widetilde{\pi}} = \widetilde{\nu}$. We now apply Lemma 11 with policy $\pi^\dagger$.

$$\begin{aligned}
\mathcal{R}(\widetilde{\nu}, \pi^\dagger, \widetilde{\omega}_{1:T_{\text{inner}}}) &= \widetilde{\nu}^\top \theta - \frac{\lambda}{2} \left\| \widetilde{\nu} \right\|_2^2 - \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \left(\langle V^{\pi_{\ell-1}}, \rho \rangle - \frac{\lambda}{2} \left\| \nu_\ell \right\|_2^2 \right) \\
&= -\frac{\lambda}{2} \left\| \widetilde{\nu} \right\|_2^2 + \frac{\lambda}{2} \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \left\| \nu_\ell \right\|_2^2
\end{aligned}$$

Substituting the value of $\frac{\lambda}{2} \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \left\| \nu_\ell \right\|_2^2$ in eq. (40) we obtain the following identity.

$$\langle \widetilde{\nu}, \theta \rangle - \frac{\lambda}{2} \left\| \widetilde{\nu} \right\|_2^2 = {\nu^\star}^\top \theta - \frac{\lambda}{2} \left\| \nu^\star \right\|_2^2 - \mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) + \mathcal{R}(\widetilde{\nu}, \pi^\dagger, \widetilde{\omega}_{1:T_{\text{inner}}}) \tag{41}$$

It can be easily verified that with the choices of η_ω and η_π in algorithm equation 3, lemma 12 provides the following upper bound on the regret.

$$\mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) \le 2 \sqrt{\frac{D^2 B}{K} \left(B + \frac{1}{(1-\gamma)^2} \right)} + \frac{4D}{1-\gamma} \sqrt{\frac{\log A}{T_{\text{inner}}}}$$

If $K \ge \frac{144 D^2 B}{\varepsilon^2} \left(B + (1-\gamma)^{-2} \right)$ and $T \ge \frac{576 D^2}{\varepsilon^2} \cdot \frac{\log A}{(1-\gamma)^2}$, then each term above is bounded by $\varepsilon/6$ and the regret is bounded by $\varepsilon/3$. Moreover, the result of Lemma 12 applies for any target policy, and in particular for $\pi^\dagger$, and so $\mathcal{R}(\widetilde{\nu}, \pi^\dagger, \widetilde{\omega}_{1:T_{\text{inner}}}) \le \varepsilon/3$. Then by eq. (41) we have the following inequality.

$$\langle \widetilde{\nu}, \theta \rangle - \frac{\lambda}{2} \left\| \widetilde{\nu} \right\|_2^2 \ge {\nu^\star}^\top \theta - \frac{\lambda}{2} \left\| \nu^\star \right\|_2^2 - \varepsilon$$

□

Lemma 11. *Given a policy $\pi^\star$ let us define $\nu^\star = \Phi^\top d^{\pi^\star}$, and let $\omega^\star_\ell = \theta_t + \gamma \cdot \boldsymbol{\mu}_t^\top V^{\pi_\ell}$. Then we have,*

$$\mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \left({\nu^\star}^\top \theta - \frac{\lambda}{2} \left\| \nu^\star \right\|_2^2 - \langle V^{\pi_{\ell-1}}, \rho \rangle + \frac{\lambda}{2} \left\| \nu_\ell \right\|_2^2 \right).$$

Proof.

$$\begin{aligned}
f(\nu^\star, \pi^\star, \omega_\ell) &= \mathcal{L}(d^{\pi^\star,\nu^\star}, \nu^\star, g^{\pi^\star,\omega_\ell}, \omega_\ell) \\
&= {\nu^\star}^\top \theta_t - \frac{\lambda}{2} \left\| \nu^\star \right\|_2^2 + \left\langle g^{\pi^\star,\omega_\ell}, \rho + \gamma \cdot \boldsymbol{\mu}_t \nu^\star - B d^{\pi^\star,\nu^\star} \right\rangle + \left\langle \omega_\ell, \Phi^\top d^{\pi^\star,\nu^\star} - \nu^\star \right\rangle \\
&= {\nu^\star}^\top \theta_t - \frac{\lambda}{2} \left\| \nu^\star \right\|_2^2
\end{aligned}$$

The last equality uses two observations. First, from the definition of $d^{\pi^\star,\nu^\star}$ in eq. (38), we have $B d^{\pi^\star,\nu^\star} = \rho + \gamma \cdot \boldsymbol{\mu}_t \nu^\star$, and the third term vanishes. Moreover, from the definition of $\nu^\star$ we can show that $d^{\pi^\star,\nu^\star}(s,a) = d^{\pi^\star}(s,a)$, and the fourth term

vanishes.

$$\begin{aligned} d^{\pi^\star,\nu^\star}(s,a) &= \pi(a \mid s) \cdot \left(\rho(s) + \gamma \cdot \boldsymbol{\mu}_t(s)^\top \sum_{s',b} \phi(s',b) d^{\pi^\star}(s',b) \right) \\ &= \pi(a \mid s) \cdot \left(\rho(s) + \gamma \cdot \sum_{s',b} P(s \mid s',b) d^{\pi^\star}(s',b) \right) \\ &= \pi(a \mid s) \cdot d^{\pi^\star}(s) = d^{\pi^\star}(s,a) \end{aligned}$$

$$\begin{aligned} f(\nu_\ell, \pi_{\ell-1}, \omega^\star_{\ell-1}) &= \mathcal{L}(d^{\pi_{\ell-1},\nu_\ell}, \nu_\ell, g^{\pi_{\ell-1},\omega^\star_{\ell-1}}, \omega^\star_{\ell-1}) \\ &= \nu_\ell^\top \left(\theta_t + \gamma \cdot \boldsymbol{\mu}_t^\top g^{\pi_{\ell-1},\omega^\star_{\ell-1}} - \omega^\star_{\ell-1} \right) - \frac{\lambda}{2} \nu_\ell^\top \nu_\ell + \left\langle g^{\pi_{\ell-1},\omega^\star_{\ell-1}}, \rho \right\rangle + \left\langle d^{\pi_{\ell-1},\nu_\ell}, \Phi \omega^\star_{\ell-1} - B^\top g^{\pi_{\ell-1},\omega^\star_{\ell-1}} \right\rangle \end{aligned} \tag{42}$$

From the definition of $\omega^\star_{\ell-1}$ we have,

$$\begin{aligned} \phi(s,a)^\top \omega^\star_{\ell-1} &= r_t(s,a) + \gamma \phi(s,a)^\top \boldsymbol{\mu}_t^\top V^{\pi_{\ell-1}} \\ &= r_t(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^{\pi_{\ell-1}}(s') = Q^{\pi_{\ell-1}}(s,a). \end{aligned}$$

This also implies $g^{\pi_{\ell-1},\omega^\star_{\ell-1}}(s) = \sum_a \pi_{\ell-1}(a \mid s) \phi(s,a)^\top \omega^\star_\ell = V^{\pi_{\ell-1}}(s)$. Therefore, the first and the fourth term in eq. (42) vanish and we obtain the following identity.

$$f(\nu_\ell, \pi_{\ell-1}, \omega^\star_{\ell-1}) = \langle V^{\pi_{\ell-1}}, \rho \rangle - \frac{\lambda}{2} \nu_\ell^\top \nu_\ell$$

Now the result follows from the definition of the regret $\mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}})$.

□

Lemma 12. *Given a policy $\pi^\star$ let us define $\nu^\star = \Phi^\top d^{\pi^\star}$, and let $\omega^\star_\ell = \theta_t + \gamma \cdot \boldsymbol{\mu}_t^\top V^{\pi_\ell}$. Then we have,*

$$\mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) \le \frac{D^2 B}{\eta_\omega K} + \eta_\omega \left(B + \frac{1}{(1-\gamma)^2} \right) + \frac{\log A}{T_{\text{inner}}\, \eta_\pi} + \frac{2\eta_\pi D^2}{(1-\gamma)^2}$$

Proof. As we set $\omega^\star_\ell = \theta_t + \gamma \cdot \boldsymbol{\mu}_t^\top V^{\pi_\ell}$, its norm is bounded by $\|\theta_t\|_2 + \gamma \cdot \|\boldsymbol{\mu}_t\|_2 |V^{\pi_\ell}| \le \sqrt{D} + \gamma \cdot \sqrt{D} \frac{\sqrt{D}}{1-\gamma} = \frac{2D}{1-\gamma}$. On the other hand, $\left\| \Sigma^{-1} \nu_\ell \right\|_2 \le \sqrt{B}$. If we write $y = \Sigma^{-1} \nu_\ell$ then $\|y\|_2 \le \sqrt{B}$. Moreover, $\|\nu_\ell\|_2 = \|\Sigma y\|_2 \le \mathrm{Tr}(\Sigma) \|y\|_2 \le D\sqrt{B}$. Therefore, we can choose the norm of each ν_ℓ to be bounded by at most $D\sqrt{B}$.

We will use the following decomposition of regret.

$$\begin{aligned} \mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T_{\text{inner}}}) &= \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu^\star, \pi^\star, \omega_{\ell-1}) - f(\nu_\ell, \pi_{\ell-1}, \omega^\star_{\ell-1}) \\ &= \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu^\star, \pi_{\ell-1}, \omega_{\ell-1}) - f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1}) \\ &+ \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1}) - f(\nu_\ell, \pi_{\ell-1}, \omega^\star_{\ell-1}) \\ &+ \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu^\star, \pi^\star, \omega_{\ell-1}) - f(\nu^\star, \pi_{\ell-1}, \omega_{\ell-1}) \end{aligned}$$

Regret in ν: Define $\mathcal{R}_\nu = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu^\star, \pi_{\ell-1}, \omega_{\ell-1}) - f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1})$. The function $f(\cdot, \pi_{\ell-1}, \omega_{\ell-1})$ is λ-strongly concave, and algorithm 3 sets ν_ℓ to be a maximize of this function. Therefore, $\mathcal{R}_\nu \le 0$.

Regret in ω: Define $\mathcal{R}_\omega = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1}) - f(\nu_\ell, \pi_{\ell-1}, \omega^\star_{\ell-1})$. Since $f(\nu_t, \pi_\ell, \cdot)$ is a linear function we have,

$$\mathcal{R}_\omega = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \left\langle \omega_{\ell-1} - \omega^\star_{\ell-1}, \nabla_{\omega_{\ell-1}} f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1}) \right\rangle = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} \frac{1}{K} \sum_{k=1}^{K} \left\langle \omega_{\ell-1,k} - \omega^\star_{\ell-1}, \nabla_{\omega_{\ell-1}} f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1}) \right\rangle$$

Since the k-th sample is drawn uniformly at random from the dataset, we have,

$$\mathbb{E}\left[\widetilde{g}_{\omega_{\ell,k}}\right] = \mathbb{E}\left[\Phi^\top d^{\pi_{\ell-1},\nu_{\ell-1}} - \frac{1}{m_t} \sum_{j=1}^{m_t} \phi(s_j, a_j)\phi(s_j, a_j)^\top \Sigma_\ell^{-1} \nu_{\ell-1}\right] = \Phi^\top d^{\pi_{\ell-1},\nu_{\ell-1}} - \nu_{\ell-1}$$

We now bound the norm of $\widetilde{g}_{\omega_{\ell,k}}$.

$$\begin{aligned}
\left\|\widetilde{g}_{\omega_{\ell,k}}\right\|_2 &\le \left\|\Phi^\top d_\ell\right\|_2 + \left\|\phi(s_j, a_j)\phi(s_j, a_j)^\top \Sigma^{-1} \nu_{\ell-1}\right\|_2 \\
&\le \sum_{s,a} d_\ell(s, a) \left\|\phi(s, a)\right\|_2 + \left\|\Sigma^{-1} \nu_{\ell-1}\right\|_2 \\
&\le \frac{1}{1-\gamma} + \sqrt{B}
\end{aligned}$$

Therefore, we can apply Lemma 13 to obtain the following bound.

$$\frac{1}{K} \sum_{k=1}^{K} \left\langle \omega_{\ell-1,k} - \omega^\star_{\ell-1}, \nabla_{\omega_{\ell-1}} f(\nu_\ell, \pi_{\ell-1}, \omega_{\ell-1}) \right\rangle \le \frac{\left\|\omega_{\ell-1,1} - \omega^\star_{\ell-1}\right\|_2^2}{2\eta_\omega K} + \eta_\omega (B + (1-\gamma)^{-2})$$

Summing over all ℓ and using the fact $\left\|\omega_{\ell,k}\right\|_2 \le D\sqrt{B}$ we obtain the following upper bound on the regret $\mathcal{R}_\omega$.

$$\mathcal{R}_\omega \le \frac{D^2 B}{\eta_\omega K} + \eta_\omega \left(B + \frac{1}{(1-\gamma)^2}\right)$$

Regret in π: Let us define $\mathcal{R}_\pi = \frac{1}{T_{\text{inner}}} \sum_{t=1}^{T_{\text{inner}}} f(\nu^\star, \pi^\star, \omega_{\ell-1}) - f(\nu^\star, \pi_{\ell-1}, \omega_{\ell-1})$. Using the definition of $f(\nu, \pi, \omega)$ we can rewrite the regret as follows.

$$\begin{aligned}
&f(\nu^\star, \pi^\star, \omega_\ell) - f(\nu^\star, \pi_\ell, \omega_\ell) = \mathcal{L}(d^{\pi^\star,\nu^\star}, \nu^\star, g^{\pi^\star,\omega_\ell}, \omega_\ell) - \mathcal{L}(d^{\pi_\ell,\nu^\star}, \nu^\star, g^{\pi_\ell,\omega_\ell}, \omega_\ell) \\
&= \underbrace{\gamma \cdot \nu^{\star\top} \mu^\top \left(g^{\pi^\star,\omega_\ell} - g^{\pi_\ell,\omega_\ell}\right) + \rho^\top \left(g^{\pi^\star,\omega_\ell} - g^{\pi_\ell,\omega_\ell}\right)}_{:=T_1} \\
&+ \underbrace{\left\langle d^{\pi^\star,\nu^\star}, \Phi\omega_\ell - B^\top g^{\pi^\star,\omega_\ell} \right\rangle - \left\langle d^{\pi_\ell,\nu^\star}, \Phi\omega_\ell - B^\top g^{\pi_\ell,\omega_\ell} \right\rangle}_{:=T_2}
\end{aligned}$$

Now from the definition of $g^{\pi,\omega}$ we have $g^{\pi^\star,\omega_\ell} - g^{\pi_\ell,\omega_\ell} = \sum_a (\pi^\star(a \mid s) - \pi_\ell(a \mid s))\phi(s, a)^\top \omega_\ell$. Moreover, with $\nu^\star = \Phi^\top d^{\pi^\star}$, $\rho + \gamma\mu\nu^\star = d^{\pi^\star}$. This implies that the term T_1 equals $\sum_s d^{\pi^\star}(s) \sum_a (\pi^\star(a \mid s) - \pi_\ell(a \mid s))\phi(s, a)^\top \omega_\ell$. In order to bound the term T_2 note that

$$\begin{aligned}
&\left\langle d^{\pi^\star,\nu^\star}, \Phi\omega_\ell - B^\top g^{\pi^\star,\omega_\ell} \right\rangle - \left\langle d^{\pi_\ell,\nu^\star}, \Phi\omega_\ell - B^\top g^{\pi_\ell,\omega_\ell} \right\rangle \\
&= \sum_s d^{\pi^\star}(s) \sum_a \pi^\star(a \mid s) \left(\phi(s, a)^\top \omega_\ell - \sum_b \pi^\star(b \mid s)\phi(s, b)^\top \omega_\ell\right) = 0
\end{aligned}$$

Similarly the second term in T_2 evaluates to zero. Therefore, we have the following expression of regret.

$$\mathcal{R}_\pi = \frac{1}{T_{\text{inner}}} \sum_{\ell=1}^{T_{\text{inner}}} f(\nu^\star, \pi^\star, \omega_{\ell-1}) - f(\nu^\star, \pi_{\ell-1}, \omega_{\ell-1}) \tag{43}$$

$$= \frac{1}{T_{\text{inner}}} \sum_{t=1}^{T_{\text{inner}}} \sum_s d^{\pi^\star}(s) \sum_a (\pi^\star(a \mid s) - \pi_{\ell-1}(a \mid s))\phi(s, a)^\top \omega_{\ell-1} \tag{44}$$

Now we can apply Lemma 15 to obtain the following bound on regret.

$$\mathcal{R}_\pi \le \frac{\mathcal{H}(\pi^\star \| \pi_1)}{T_{\text{inner}}\eta_\pi} + \frac{2\eta_\pi D^2}{(1-\gamma)^2} \le \frac{\log A}{T_{\text{inner}}\,\eta_\pi} + \frac{2\eta_\pi D^2}{(1-\gamma)^2}$$

Therefore, the total regret can be bounded as follows.

$$\begin{aligned}\mathcal{R}(\nu^\star, \pi^\star, \omega^\star_{1:T}) &= \mathcal{R}_\nu + \mathcal{R}_\omega + \mathcal{R}_\pi \\ &\le \frac{D^2 B}{\eta_\omega K} + \eta_\omega\left(B + \frac{1}{(1-\gamma)^2}\right) + \frac{\log A}{T_{\text{inner}} \cdot \eta_\pi} + \frac{2\eta_\pi D^2}{(1-\gamma)^2}\end{aligned}$$

□

Lemma 13 (Online Stochastic Gradient Descent)**.** *Let $y_1 \in W$, and $\eta > 0$. Define the sequence $y_2, \dots, y_{n+1}$ and $h_1, \dots, h_n$ such that for $k = 1, \dots, n$*

$$y_{k+1} = Proj_W\left(y_k + \eta \widehat{h}_k\right)$$

and $\widehat{h}_k$ satisfies $\mathbb{E}\left[\widehat{h}_k \mid \mathcal{F}_{k-1}\right] = h_k$ and $\mathbb{E}\left[\left\|\widehat{h}_k\right\|_2^2 \mid \mathcal{F}_{k-1}\right] \le G^2$. Then for any $y^\star \in W$,

$$\mathbb{E}\left[\sum_{k=1}^n \left\langle y^\star - y_k, h_k\right\rangle\right] \le \frac{\left\|y_1 - y^\star\right\|_2^2}{2\eta} + \frac{\eta n G^2}{2}.$$

Lemma 14 (Online Stochastic Gradient Descent for Strongly Convex Loss Function)**.** *Let $y_1 \in W$, and $\eta > 0$. Define the sequence $y_2, \dots, y_{n+1}$ and $h_1, \dots, h_n$ such that for $k = 1, \dots, n$*

$$y_{k+1} = Proj_W\left(y_k + \eta_k \widehat{h}_k\right)$$

and $\widehat{h}_k$ satisfies $\mathbb{E}\left[\widehat{h}_k \mid \mathcal{F}_{k-1}\right] = \nabla f_k(y_k)$ and $\mathbb{E}\left[\left\|\widehat{h}_k\right\|_2^2 \mid \mathcal{F}_{k-1}\right] \le G^2$. Moreover, suppose f_k is α-strongly convex for each k and stepsize $\eta_k = \frac{1}{\alpha k}$. Then for any $y^\star \in W$,

$$\mathbb{E}\left[\sum_{k=1}^n f_k(y^\star) - f_k(y_k)\right] \le \frac{G^2}{\alpha}\log n.$$

Proof. See Hazan (2016) (chapter 3) for a proof. □

Lemma 15 (Mirror Descent, Lemma D.2 of Gabbianelli et al. (2024))**.** *Let $q_1, q_2, \dots, q_T$ be a sequence of functions from $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ so that $\|q_t\|_\infty \le D$. Given an initial policy π_1, and a learning rate $\alpha > 0$, define a sequence of policies*

$$\pi_{t+1}(a \mid s) \propto \pi_t(a \mid s) e^{\alpha q_t(s,a)}$$

for $t = 1, 2, \dots, T-1$. Then for any comparator policy $\pi^\star$,

$$\frac{1}{T}\sum_{t=1}^T \sum_{s\in\mathcal{S}} q^{\pi^\star}(s)\left\langle \pi^\star(\cdot \mid s) - \pi_t(\cdot \mid s), q_t(s,\cdot)\right\rangle \le \frac{\mathcal{H}(\pi^\star \| \pi_1)}{T\alpha} + \frac{\alpha D^2}{2}$$

E.3 Proof of Corollary 1

Proof. The proof of Lemma 6 shows that as long as the number of samples $m_t \ge O\left(\frac{D^5 B\lambda^4}{(1-\gamma)^2 c_1^4 \varepsilon^2}\log\frac{DB\lambda}{c_1\varepsilon\delta_0}\right)$

$$\left|\mathcal{L}_t(d, \nu; g, \omega) - \widehat{\mathcal{L}}_t(d, \nu; g, \omega)\right| \le \varepsilon$$

for any $d, \nu \in \mathcal{V}, \omega \in \mathcal{W}$, and $g = \arg\min_{g'} \widehat{\mathcal{L}}_t(d, \nu; g', \omega)$.

Given $\widetilde{d}, \widetilde{\nu}$ let us pick $(\widetilde{g}, \widetilde{\omega}) \in \arg\min_{g,\omega} \widehat{\mathcal{L}}_t(\widetilde{d}, \widetilde{\nu}; g, \omega)$. Then, we have $\widehat{\mathcal{L}}_t(\widetilde{d}, \widetilde{\nu}; \widetilde{g}, \widetilde{\omega}) \le \min_{g,\omega} \widehat{\mathcal{L}}_t(\widetilde{d}, \widetilde{\nu}; g, \omega)$ and

$$\begin{aligned}\widehat{\mathcal{L}}_t(\widetilde{d}, \widetilde{\nu}; \widetilde{g}, \widetilde{\omega}) &\ge \mathcal{L}_t(\widetilde{d}, \widetilde{\nu}; \widetilde{g}, \widetilde{\omega}) - \varepsilon = \widetilde{\nu}^\top \theta_t - \frac{\lambda}{2}\|\widetilde{\nu}\|_2^2 - \varepsilon \ge \nu^{\star\top}\theta_t - \frac{\lambda}{2}\left\|\nu^\star\right\|_2^2 - 2\varepsilon \\ &= \mathcal{L}_t(d^{\pi^\star}, \nu^\star, \widetilde{g}, \widetilde{\nu}) - 2\varepsilon \ge \max_{d,\nu} \mathcal{L}_t(d, \nu, \widetilde{g}, \widetilde{\nu}) - 2\varepsilon\end{aligned}$$

□

E.4 Proof of Corollary 2

Proof. We first bound the distance between $\widetilde{\nu}_t$ and $\widehat{\nu}_t$. Since $(\widehat{d}_t, \widehat{\nu}_t)$ is the optimal solution to $\widehat{\mathcal{L}}_t(\cdot, \widehat{g}_t, \widehat{\omega}_t)$ we obtain the following inequality.

$$\begin{aligned}
\left\|\widehat{\nu}_t - \widetilde{\nu}_t\right\|_2 &\leq \sqrt{\frac{\widehat{\mathcal{L}}_t(\widehat{d}_t, \widehat{\nu}_t, \widehat{g}_t, \widehat{\omega}_t) - \widehat{\mathcal{L}}_t(\widetilde{d}_t, \widetilde{\nu}_t, \widehat{g}_t, \widehat{\omega}_t)}{2\lambda}} \\
&\leq \sqrt{\frac{\widehat{\mathcal{L}}_t(\widehat{d}_t, \widehat{\nu}_t, \widetilde{g}_t, \widetilde{\omega}_t) - \widehat{\mathcal{L}}_t(\widetilde{d}_t, \widetilde{\nu}_t, \widehat{g}_t, \widehat{\omega}_t)}{2\lambda}} \\
&\leq \sqrt{\frac{\widehat{\mathcal{L}}_t(\widetilde{d}_t, \widetilde{\nu}_t, \widetilde{g}_t, \widetilde{\omega}_t) - \widehat{\mathcal{L}}_t(\widetilde{d}_t, \widetilde{\nu}_t, \widehat{g}_t, \widehat{\omega}_t) + 2\varepsilon}{2\lambda}} \\
&\leq \sqrt{\frac{4\varepsilon}{2\lambda}}
\end{aligned}$$

The first inequality follows from the fact that $(\widehat{d}_t, \widehat{\nu}_t, \widehat{g}_t, \widehat{\omega}_t)$ is an exact saddle point of the objective $\widehat{\mathcal{L}}_t(\cdot, \cdot)$. The last two inequalities follow because $(\widetilde{d}_t, \widetilde{\nu}_t, \widetilde{g}_t, \widetilde{\omega}_t)$ is a 2ε-approximate saddle point of the objective $\widehat{\mathcal{L}}_t(\cdot, \cdot)$.

Now we bound the distance between $\widehat{d}_t$ and $\widetilde{d}_t$. First note that, the proof of theorem 4 shows that if $m_t \geq O\left(\frac{D^5 B\lambda^4}{(1-\gamma)^2 c_1^4 \varepsilon^2} \log \frac{DB\lambda t}{c_1 \varepsilon p}\right)$ we obtain $\left\|\widehat{\nu}_t - \Phi^\top \widehat{d}_t\right\|_2 \leq 8\sqrt{B\varepsilon}$ with probability at least $1 - \frac{p}{t^2} \cdot \frac{6}{2\pi^2}$. Furthermore, $\widetilde{\nu}_t = \Phi^\top \widetilde{d}_t$, and we obtain the following inequality.

$$\sqrt{\kappa}\left\|\widehat{d}_t - \widetilde{d}_t\right\|_2 \leq \left\|\Phi^\top \widehat{d}_t - \Phi^\top \widetilde{d}_t\right\|_2 \leq \left\|\widetilde{\nu}_t - \widehat{\nu}_t\right\|_2 + \left\|\widehat{\nu}_t - \Phi^\top \widehat{d}_t\right\|_2 \leq \sqrt{\frac{4\varepsilon}{2\lambda}} + 8\sqrt{B\varepsilon}$$

After rearranging we obtain,

$$\left\|\widehat{d}_t - \widetilde{d}_t\right\|_2 \leq \sqrt{\frac{2\varepsilon}{\lambda\kappa}} + 8\sqrt{\frac{B}{\kappa}}\varepsilon \tag{45}$$

The proof of theorem 4 also establishes the following inequality.

$$\left\|d_t^\star - \widehat{d}_t\right\|_2 \leq \sqrt{\frac{\varepsilon}{\lambda\kappa}} + 8\sqrt{\frac{B}{\kappa}}\varepsilon$$

The above two inequalities imply the following bound on the distance between $d_t^\star$ and $\widetilde{d}_t$.

$$\left\|d_t^\star - \widetilde{d}_t\right\|_2 \leq 3\sqrt{\frac{\varepsilon}{\lambda\kappa}} + 16\sqrt{\frac{B}{\kappa}}\varepsilon$$

Now we can proceed very similarly to the proof of theorem 4 and establish that the sequence $\{\widetilde{d}_t\}_{t\geq 1}$ converges to d_S if the regularization constant λ is chosen slightly larger than required in theorem 4. □

F MISSING PROOFS FROM SECTION 5

F.1 Proof of Lemma 1

Proof. Suppose $Q_2^\star$ (resp. $\widetilde{Q}_2^\star$) be the optimal Q-function when the first agent deploys policy π_1 (resp. $\widetilde{\pi}_1$). Then we have,

$$|\pi_2(a \mid s) - \widetilde{\pi}_2(a \mid s)| = \left|\frac{\exp(\beta Q_2^\star(s,a))}{\sum_b \exp(\beta Q_2^\star(s,b))} - \frac{\exp(\beta \widetilde{Q}_2^\star(s,a))}{\sum_b \exp(\beta \widetilde{Q}_2^\star(s,b))}\right| \leq \sqrt{2}\beta \left\|Q_2^\star(s,\cdot) - \widetilde{Q}_2^\star(s,\cdot)\right\|_2 \tag{46}$$

The last inequality follows from the observation that for any L-Lipschitz function f we have $|f(x) - f(y)| \leq L\|x - y\|_2$, and for the function $f_j(x) = \frac{\exp(\beta x_j)}{\sum_i \exp(\beta x_i)}$ we have

$$\nabla_{x_k} f_j(x) = \begin{cases} \beta f_j(x)\left(1 - f_j(x)\right) & \text{if } k = j \\ -\beta f_j(x) f_k(x) & \text{if } k \neq j \end{cases}$$

which implies $\left\|\nabla_x f_j(x)\right\|_2 \leq \sqrt{2}\beta$.

Now we provide a bound on the norm $\left\|Q_2^\star(s,\cdot) - \widetilde{Q}_2^\star(s,\cdot)\right\|_2$. Let $\bar{r}_2$ (resp. $\widetilde{r}_2$) be the reward function of agent 2 in response to policy π_1 (resp. $\widetilde{\pi}_1$).

$$\begin{aligned}
|\bar{r}_2(s,a) - \widetilde{r}_2(s,a)| &\leq \sum_{a_1} |\pi_1(a_1 \mid s) - \widetilde{\pi}_1(a_1 \mid s)|\, |r_2(s,a_1,a_2)| \\
&\leq \|\pi_1(\cdot \mid s) - \widetilde{\pi}_1(\cdot \mid s)\|_1 \max_{a_1,a_2} |r_2(s,a_1,a_2)| \\
&\leq \delta \cdot r_{\max}
\end{aligned}$$

Similarly, let $\overline{P}_2$ (resp. $\widetilde{P}_2$) be the probability transition function of agent 2 in response to policy π_1 (resp. $\widetilde{\pi}_1$). Then for any state s, and action a_2 we have,

$$\begin{aligned}
\sum_{s'} \left|\overline{P}_2(s' \mid s,a_2) - \widetilde{P}_2(s' \mid s,a_2)\right| &\leq \sum_{s'}\sum_{a_1} |\pi_1(a_1 \mid s) - \widetilde{\pi}_1(a_1 \mid s)|\, P(s' \mid s,a_1,a_2) \\
&\leq \|\pi_1(\cdot \mid s) - \widetilde{\pi}_1(\cdot \mid s)\|_1 \sum_{a_1}\sum_{s'} P(s' \mid s,a_1,a_2) \\
&\leq \delta \cdot A_1
\end{aligned}$$

Now consider the problem of computing the optimal policies of agent 2 starting from the state, action pair (s,a). Let $\overline{\pi}_2$ (resp. $\widetilde{\pi}_2$) be the optimal policy with reward $\bar{r}_2$, and transition $\overline{P}_2$ (resp. reward $\widetilde{r}_2$, and transition $\widetilde{P}_2$). Moreover, let $\overline{d}_2$ (resp. $\overline{q}_2$) be the state, action occupancy measure of the policy $\overline{\pi}_2$ under probability transition $\overline{P}_2$ (resp. $\widetilde{P}_2$). Similarly, let $\widetilde{d}_2$ (resp. $\widetilde{q}_2$) be the state, action occupancy measure of the policy $\widetilde{\pi}_2$ under probability transition $\widetilde{P}_2$ (resp. $\overline{P}_2$).

Let $\overline{d}_2^\star$ (resp. $\widetilde{d}_2^\star$) be the optimal state, action occupancy measure under reward function $\bar{r}_2$ (resp. $\widetilde{r}_2$). Then we have,

$$\begin{aligned}
Q_2^\star(s,a) &= \sum_{s',b'} \bar{r}_2(s',b')\overline{d}_2(s',b') \geq \sum_{s',b'} \bar{r}_2(s',b')\widetilde{q}_2(s',b') \\
&= \sum_{s',b'} \bar{r}_2(s',b')\widetilde{d}_2(s',b') + \sum_{s',b'} \bar{r}_2(s',b')\left(\widetilde{q}_2(s',b') - \widetilde{d}_2(s',b')\right) \\
&\geq \sum_{s',b'} \widetilde{r}_2(s',b')\widetilde{d}_2(s',b') + \sum_{s',b'} \left(\bar{r}_2(s',b') - \widetilde{r}_2(s',b')\right)\widetilde{d}_2(s',b') - r_{\max}\cdot\left\|\widetilde{q}_2 - \widetilde{d}_2\right\|_1 \\
&\geq \widetilde{Q}_2^\star(s,a) - \delta\cdot r_{\max}\sum_{s',b'}\widetilde{d}_2(s',b') - r_{\max}\cdot\frac{\delta A_1\gamma}{(1-\gamma)^2} \\
&\geq \widetilde{Q}_2^\star(s,a) - \frac{\delta\cdot r_{\max}}{1-\gamma} - r_{\max}\cdot\frac{\delta A_1\gamma}{(1-\gamma)^2}
\end{aligned}$$

The first inequality follows because the policy $\widetilde{\pi}_2$ is sub-optimal under the reward $\bar{r}_2$ and transition $\overline{P}_2$. The third inequality uses Lemma 16 and $\widetilde{Q}_2^\star(s,a) = \sum_{s',b'}\widetilde{r}_2(s',b')\widetilde{d}_2(s',b')$. The final inequality uses that $\sum_{s',a'} d(s',a') = 1/(1-\gamma)$ for any occupancy measure d. Similarly we can show that $\widetilde{Q}_2^\star(s,a) \geq Q_2^\star(s,a) - \frac{\delta\cdot r_{\max}}{1-\gamma} - r_{\max}\cdot\frac{\delta A_1\gamma}{(1-\gamma)^2}$. Therefore,

$$\left\|Q_2^\star(s,\cdot) - \widetilde{Q}_2^\star(s,\cdot)\right\|_2 \leq \sqrt{A_2} r_{\max}\frac{2\delta A_1}{(1-\gamma)^2}.$$

Therefore, using Equation (46) we obtain the following bound.

$$|\pi_2(a \mid s) - \widetilde{\pi}_2(a \mid s)| \leq 2\sqrt{2} r_{\max}\frac{A_1\sqrt{A_2}\beta\delta}{(1-\gamma)^2}$$

Now, let r_1 (resp. $\widetilde{r}_1$) be the reward function of agent 1 when agent 2 adopts policy π_2 (resp. $\widetilde{\pi}_2$).

$$\begin{aligned}
|r_1(s,a) - \widetilde{r}_1(s,a)| &\leq \sum_{a_2} |\pi_2(a_2 \mid s) - \widetilde{\pi}_2(a_2 \mid s)|\, |r_1(s,a_1,a_2)| \\
&\leq \delta\cdot\frac{2\sqrt{2}\beta A_1 A_2^{3/2} r_{\max}^2}{(1-\gamma)^2}
\end{aligned}$$

Similarly, let P_1 (resp. $\widetilde{P}_1$) be the probability transition function of agent 1 when agent 2 adopts policy π_2 (resp. $\widetilde{\pi}_2$).

$$\begin{aligned}\left|P_1(s' \mid s,a) - \widetilde{P}_1(s' \mid s,a)\right| &\leq \sum_{a_2} |\pi_2(a_2 \mid s) - \widetilde{\pi}_2(a_2 \mid s)|\, P(s,a_1,a_2) \\ &\leq \delta \cdot \frac{2\sqrt{2}\beta A_1 A_2^{3/2} r_{\max}}{(1-\gamma)^2}\end{aligned}$$

□

Lemma 16. *Suppose* $\left\|P(\cdot \mid s,a) - \widetilde{P}(\cdot \mid s,a)\right\|_1 \leq \beta$ *for any state,action pair* (s,a)*. Then for any policy* π *and any starting state distribution* ρ *we have* $\left\|d_\rho^\pi - \widetilde{d}_\rho^\pi\right\|_1 \leq \frac{\beta\gamma}{(1-\gamma)^2}$.

Proof. Let P_h^π (resp. $\widetilde{P}^\pi$) be the state distribution at time-step h resulting from the starting distribution ρ under the probability transition function P (resp. $\widetilde{P}$).

$$\begin{aligned}P_h^\pi(s') - \widetilde{P}_h^\pi(s') &= \sum_{s,a}\left(P_{h-1}^\pi(s)P(s' \mid s,a) - \widetilde{P}_{h-1}^\pi(s)\widetilde{P}(s' \mid s,a)\right)\pi(a \mid s) \\ &= \sum_s P_{h-1}^\pi(s)\sum_a\left(P(s' \mid s,a) - \widetilde{P}(s' \mid s,a)\right)\pi(a \mid s) \\ &+ \sum_s\left(P_{h-1}^\pi(s) - \widetilde{P}_{h-1}^\pi(s)\right)\sum_a \widetilde{P}(s' \mid s,a)\pi(a \mid s)\end{aligned}$$

Taking absolute values and summing over all the states we obtain the following inequality.

$$\begin{aligned}\sum_{s'}\left|P_h^\pi(s') - \widetilde{P}_h^\pi(s')\right| &\leq \sum_{s,a}\left\|P(\cdot \mid s,a) - \widetilde{P}(\cdot \mid s,a)\right\|_1 P_{h-1}^\pi(s)\pi(a \mid s) \\ &+ \sum_{s'}\sum_s\left|P_{h-1}^\pi(s) - \widetilde{P}_{h-1}^\pi(s)\right|\sum_a \widetilde{P}(s' \mid s,a)\pi(a \mid s) \\ &\leq \beta \cdot \sum_{s,a} P_{h-1}^\pi(s)\pi(a \mid s) + \sum_s\left|P_{h-1}^\pi(s) - \widetilde{P}_{h-1}^\pi(s)\right|\sum_a \pi(a \mid s) \\ &\leq \beta + \sum_s\left|P_{h-1}^\pi(s) - \widetilde{P}_{h-1}^\pi(s)\right|\end{aligned}$$

Since $P^p i_0(s) = \rho(s)$ we have $\left\|P_0^\pi - \widetilde{P}_0^\pi\right\|_1 = 0$, and $\left\|P_h^\pi - \widetilde{P}_h^\pi\right\|_1 = \beta \cdot h$. Now using the definition of state, action occupancy measure we get,

$$d^\pi(s,b) - \widetilde{d}^\pi(s,b) = \sum_h \gamma^h\left(P_h^\pi(s) - \widetilde{P}_h^\pi(s)\right)\pi(b \mid s) \leq \sum_h \gamma^h\left|P_h^\pi(s) - \widetilde{P}_h^\pi(s)\right|\pi(b \mid s).$$

Therefore,

$$\left\|d^\pi - \widetilde{d}^\pi\right\|_1 \leq \sum_h\sum_{s'}\gamma^h\left|P_h^\pi(s) - \widetilde{P}_h^\pi(s)\right|\sum_b \pi(b \mid s) = \sum_h \beta\gamma^h h = \frac{\beta\gamma}{(1-\gamma)^2}$$

□

F.2 Proof of Lemma 2

Proof. Suppose $Q_f^\star$ (resp. $\widetilde{Q}_f^\star$) be the optimal state,action welfare-function when the first agent deploys policy π_1 (resp. $\widetilde{\pi}_1$). Similar to the proof of Lemma 1 we can establish the following inequality.

$$\left|\pi_f(\boldsymbol{a} \mid s) - \widetilde{\pi}_f(\boldsymbol{a} \mid s)\right| \leq \sqrt{2}\beta\left\|Q_f^\star(s,\cdot) - \widetilde{Q}_f^\star(s,\cdot)\right\|_2$$

Let us also write $\overline{r}_j(s,\boldsymbol{a})$ (resp. $\widetilde{r}_j(s,\boldsymbol{a})$ be the reward function of agent $j \in \{2,\ldots,m+1\}$ when the first agent plays policy π_1 (resp. $\widetilde{\pi}_1$). Then we have,

$$\begin{aligned}\left|\overline{r}_j(s,\boldsymbol{a}) - \widetilde{r}_j(s,\boldsymbol{a})\right| &\leq \sum_{a_1}|\pi_1(a_1 \mid s) - \widetilde{\pi}_1(a_1 \mid s)|\left|r_j(s,a_1,\boldsymbol{a}\right| \\ &\leq \|\pi_1(\cdot \mid s) - \widetilde{\pi}_1(\cdot \mid s)\|\, r_{\max} \\ &\leq \delta \cdot r_{\max}.\end{aligned}$$

Let us also write $\overline{P}$ (resp. $\widetilde{P}$) to be the probability transition function when the first agent plays policy π_1 (resp. $\widetilde{\pi}_1$). Then we have,

$$\begin{aligned}\sum_{s'}\left|\overline{P}_2(s' \mid s, \boldsymbol{a}) - \widetilde{P}_2(s' \mid s, \boldsymbol{a})\right| &\leq \sum_{s'}\sum_{a_1}\left|\pi_1(a_1 \mid s) - \widetilde{\pi}_1(a_1 \mid s)\right| P(s' \mid s, a_1, \boldsymbol{a}) \\ &\leq \|\pi_1(\cdot \mid s) - \widetilde{\pi}_1(\cdot \mid s)\|_1 \sum_{a_1}\sum_{s'} P(s' \mid s, a_1, \boldsymbol{a}) \\ &\leq \delta \cdot A_1.\end{aligned}$$

Now consider the problem of computing the optimal CCE of agents $\{2, \ldots, m+1\}$ starting from the state, action pair $(s, \boldsymbol{a})$. Let $\overline{\pi}_f$ (resp. $\widetilde{\pi}_f$) be the optimal policy with reward $\overline{r}_j$, and transition $\overline{P}$ (resp. reward $\widetilde{r}_j$, and transition $\widetilde{P}_j$). There are two cases to consider. First, the policy $\widetilde{\pi}_f$ is a CCE under reward $\overline{r}_j$ and transition $\overline{P}$. In that case, we can proceed similar to the proof Lemma 1 and establish the following inequality.

$$Q_f^\star(s, \boldsymbol{a}) \geq \widetilde{Q}_f^\star(s, \boldsymbol{a}) - \frac{\delta \cdot r_{\max}}{1-\gamma} - \frac{\delta A_1 \gamma \cdot r_{\max}}{(1-\gamma)^2} \tag{47}$$

The second case occurs when $\widetilde{\pi}_f$ is no longer a CCE. In that case, let $P \subseteq \{2, \ldots, m+1\}$ be the set of agents that can improve under $\widetilde{\pi}_f$ i.e. for each $j \in P$ there exists a strategy $\pi'_j : S \to \Delta(A_j)$ so that $Q_j^{\pi'_j, \widetilde{\pi}_{f,-j}}(s, \boldsymbol{a}) > Q_j^{\widetilde{\pi}_f}(s, \boldsymbol{a})$. However, $\widetilde{\pi}_f$ is a CCE under reward $\widetilde{r}_j$ and transition $\widetilde{P}$, which implies $\widetilde{Q}_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) \geq \widetilde{Q}_j^{\pi'_j, \widetilde{\pi}_{f,-j}}(s, \boldsymbol{a})$. Now using Lemma 16 we get and $Q_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) \geq \widetilde{Q}_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) - r_{\max} \cdot \frac{\delta A_1 \gamma}{(1-\gamma)^2}$. Therefore, we obtain the following inequality.

$$\begin{aligned}Q_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) + 2r_{\max} \cdot \frac{\delta A_1 \gamma}{(1-\gamma)^2} &\geq \widetilde{Q}_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) + r_{\max} \cdot \frac{\delta A_1 \gamma}{(1-\gamma)^2} \\ &\geq \widetilde{Q}_j^{\pi'_j, \widetilde{\pi}_{f,-j}}(s, \boldsymbol{a}) + r_{\max} \cdot \frac{\delta A_1 \gamma}{(1-\gamma)^2} \\ &\geq Q_j^{\pi'_j, \widetilde{\pi}_{f,-j}}(s, \boldsymbol{a}) > Q_j^{\widetilde{\pi}_f}(s, \boldsymbol{a})\end{aligned}$$

This inequality shows that $\widetilde{\pi}_f$ is ε-CCE under reward $\overline{r}_j$ and transition $\overline{P}$ for $\varepsilon = 2r_{\max} \cdot \frac{\delta A_1 \gamma}{(1-\gamma)^2}$. Since the optimal welfare over the set of ε-CCE is at most εm plus the optimal welfare over the set of exact CCE, we have the following inequality.

$$\begin{aligned}Q_f^\star(s, \boldsymbol{a}) &\geq Q^{\widetilde{\pi}_f}(s, \boldsymbol{a}) - 2r_{\max} \cdot \frac{m\delta A_1 \gamma}{(1-\gamma)^2} \\ &= \sum_j Q_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) - 2r_{\max} \cdot \frac{m\delta A_1 \gamma}{(1-\gamma)^2} \\ &\geq \sum_j \widetilde{Q}_j^{\widetilde{\pi}_f}(s, \boldsymbol{a}) - 3r_{\max} \cdot \frac{m\delta A_1 \gamma}{(1-\gamma)^2} \\ &= \widetilde{Q}_f^\star(s, \boldsymbol{a}) - 3r_{\max} \cdot \frac{m\delta A_1 \gamma}{(1-\gamma)^2}\end{aligned}$$

The above inequality and eq. (47) imply $Q_f^\star(s, \boldsymbol{a}) \geq \widetilde{Q}_f^\star(s, \boldsymbol{a}) - 3r_{\max} \cdot \frac{m\delta A_1}{(1-\gamma)^2}$. Similarly we can show $\widetilde{Q}_f^\star(s, \boldsymbol{a}) \geq Q_f^\star(s, \boldsymbol{a}) - 3r_{\max} \cdot \frac{m\delta A_1}{(1-\gamma)^2}$, and we get the following bound.

$$\left|\pi_f(\boldsymbol{a} \mid s) - \widetilde{\pi}_f(\boldsymbol{a} \mid s)\right| \leq \sqrt{2}\beta \left\|Q_f^\star(s, \cdot) - \widetilde{Q}_f^\star(s, \cdot)\right\|_2 \leq \beta\delta \cdot \frac{3\sqrt{2}mA^{m/2+1} \cdot r_{\max}}{(1-\gamma)^2}$$

Now, let r_1 (resp. $\widetilde{r}_1$) be the reward function of agent 1 when the follower agents adopt policy π_f (resp. $\widetilde{\pi}_f$).

$$\begin{aligned}|r_1(s, a_1) - \widetilde{r}_1(s, a_1)| &\leq \sum_{\boldsymbol{a}} \left|\pi_f(\boldsymbol{a} \mid s) - \widetilde{\pi}_f(\boldsymbol{a} \mid s)\right| |r_1(s, a_1, \boldsymbol{a})| \\ &\leq A^m \cdot \beta\delta \cdot \frac{3\sqrt{2}mA^{m/2+1} \cdot r_{\max}}{(1-\gamma)^2} \cdot r_{\max} \\ &\leq \delta \cdot \frac{3\sqrt{2}\beta mA^{3m/2+1} r_{\max}^2}{(1-\gamma)^2}\end{aligned}$$

Similarly, let P_1 (resp. $\widetilde{P}_1$) be the probability transition function of agent 1 when the follower agents adopt policy π_f (resp. $\widetilde{\pi}_f$).

$$\begin{aligned}
\left|P_1(s' \mid s, a_1) - \widetilde{P}_1(s' \mid s, a_1)\right| &\leq \sum_{\boldsymbol{a}} \left|\pi_f(\boldsymbol{a} \mid s) - \widetilde{\pi}_f(\boldsymbol{a} \mid s)\right| P(s' \mid s, a_1, \boldsymbol{a}) \\
&\leq A^m \cdot \beta\delta \cdot \frac{3\sqrt{2}mA^{m/2+1} \cdot r_{\max}}{(1-\gamma)^2} \cdot 1 \\
&\leq \delta \cdot \frac{3\sqrt{2}\beta mA^{3m/2+1} r_{\max}}{(1-\gamma)^2}
\end{aligned}$$

□