Performative Reinforcement Learning

Debmalya Mandal [1] **Stelios Triantafyllou** [1] **Goran Radanovic** [1]

Abstract

We introduce the framework of performative reinforcement learning where the policy chosen by the learner affects the underlying reward and transition dynamics of the environment. Following the recent literature on performative prediction (Perdomo et al., 2020), we introduce the concept of performatively stable policy. We then consider a regularized version of the reinforcement learning problem and show that repeatedly optimizing this objective converges to a performatively stable policy under reasonable assumptions on the transition dynamics. Our proof utilizes the dual perspective of the reinforcement learning problem and may be of independent interest in analyzing the convergence of other algorithms with decision-dependent environments. We then extend our results for the setting where the learner just performs gradient ascent steps instead of fully optimizing the objective, and for the setting where the learner has access to a finite number of trajectories from the changed environment. For both the settings, we leverage the dual formulation of performative reinforcement learning, and establish convergence to a stable solution. Finally, through extensive experiments on a grid-world environment, we demonstrate the dependence of convergence on various parameters e.g. regularization, smoothness, and the number of samples.

1. Introduction

Over the last decade, advances in reinforcement learning techniques enabled several breakthroughs in AI. These milestones include AlphaGo (Silver et al., 2017), Pluribus (Brown & Sandholm, 2019), and AlphaStar (Vinyals et al., 2019). Such success stories of reinforcement learning in multi-agent game playing environments have naturally led to the adoption of RL in many real-world scenarios e.g. recommender systems (Aggarwal et al.), and healthcare (Esteva et al., 2019). However, these critical domains often pose new challenges including the mismatch between deployed policy and the target environment.

Existing frameworks of reinforcement learning ignore the fact that a deployed policy might change the underlying environment (i.e., reward, or probability transition function, or both). Such a mismatch between the deployed policy and the environment often arises in practice. For example, recommender systems often use contextual Markov decision process to model interaction with a user (Hansen et al., 2020). In such a contextual MDP, the initial context/user feature is drawn according a distribution, then the user interacts with the platform according to the context-specific MDP. However, it has been repeatedly observed that such recommender systems not only change the user demographics (i.e. distribution of contexts) but also how they interact with the platforms (Chaney et al., 2018; Mansoury et al., 2020). Our second example comes from autonomous vehicles (AV). Even if we ignore the multi-agent aspect of these learning algorithms, a deployed AV might change how the pedestrians, and other cars behave, and the resulting environment might be quite different from what the designers of the AV had in mind (Nikolaidis et al., 2017a).

Recently, Perdomo et al. (2020) introduced the notion of *performative prediction*, where the predictions made by a classifier changes the data distribution. However, in the context of reinforcement learning, the situation is different as the changing transition dynamics introduces additional complexities. If the underlying probability transition function changes, then the class of feasible policies and/or models changes with time. This implies that we need a framework that is more general than the framework of performative prediction, and can model policy-dependent outcomes in reinforcement learning.

Our Contributions: In this paper, we introduce the notion of *performative reinforcement learning* where the deployed policy not only changes the reward vector but also the underlying transition probability function. We introduce the notion of *performatively stable policy* and show under what conditions various repeated retraining methods (e.g., policy optimization, gradient ascent etc.) converges to such a

[1]Max Planck Institute for Software Systems, Saarbruecken, Germany. Correspondence to: Debmalya Mandal <dmandal@mpi-sws.org>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

stable solution. Our precise contributions are the following.

- We consider a regularized version of the reinforcement learning problem where the variables are long-term discounted state-action occupancy measures. We show that, when both the probability transition function and the reward function changes smoothly in response to the occupancy measure, repeated optimization of regularized reinforcement learning converges to a stable solution.
- We then show that if the learner performs repeated projected gradient ascent steps, then also convergence is guaranteed provided that the step-size is small enough. Compared to the supervised learning setting (Perdomo et al., 2020), the projection step is necessary as the probability transition function, and hence the space of occupancy measures change with time.
- Next we extend our result to the finite samples setting, where the learner has access to a collection of samples from the updated environment. For this setting, we use an empirical version of the Lagrangian of the regularized RL problem. We show that repeatedly solving a saddle point of this empirical Lagrangian (max player corresponds to the learner) also converges to a stable solution provided that the number of samples is large enough.
- Finally, we empirically evaluate the effect of various parameter choices (regularization, smoothness, number of samples etc.) through extensive experiments on a two-agent grid-world environment. In this environment, the first agent performs various types of repeated retraining, whereas the second agent responds according to a smooth response function.

Our Techniques: At a high level, our theoretical results might look similar to the results derived by Perdomo et al. (2020). However, there are many challenges.

- We repeatedly maximize a regularized objective whose feasible region is the space of feasible occupancy measures. As the probability transition function changes with time, the feasible region of the optimization problem also changes with time. So ideas from supervised classification setting (Mendler-Dünner et al., 2020) cannot be applied directly. Therefore, we look at the dual problem which is strongly-convex and mapping from occupancy measure to the corresponding dual optimal solution turns out to be a contraction. We believe that the dual perspective on performative prediction might be useful for analyzing the convergence of other algorithms with decision-dependent outcomes.
- For performative reinforcement learning, the finite sample setting is very different than the supervised learning setting. This is because we do not have independent sample access from the new environment. At time-step t, we can only access the new model through the policy π_t (or occupancy measure d_t). In that sense, the learner faces an offline reinforcement learning problem where the samples are collected from the behavior policy π_t. This is also the reason we need an additional overlap assumption, which is often standard in offline reinforcement learning (Munos & Szepesvári, 2008).

1.1. Related Work:

Perdomo et al. (2020) introduced the notion of *performative prediction*. Subsequent papers have considered several aspects of this framework including optimization (Mendler-Dünner et al., 2020; Miller et al., 2021), multi-agent systems (Narang et al., 2022), and population dynamics (Brown et al., 2020). However, to the best of our knowledge, performative prediction in sequential decision making is mostly unexplored. A possible exception is (Bell et al., 2021) who consider a setting where the transition and reward of the underlying MDP depend non-deterministically on the deployed policy. Since the mapping is non-deterministic, it doesn't lead to a notion of equilibrium, and the authors instead focus on the optimality and convergence of various RL algorithms.

Stochastic Stackelberg Games: Our work is also closely related to the literature on stochastic games (Shapley, 1953; Filar & Vrieze, 2012), and in particular, those that study Stackelberg (commitment) strategies (Von Stackelberg, 2010), where a leader commits a policy to which a follower (best) responds. While different algorithmic approaches have been proposed for computing Stackelberg equilibria (SE) in stochastic games or related frameworks (Vorobeychik & Singh, 2012; Letchford et al., 2012; Dimitrakakis et al., 2017), computing optimal commitment policies has shown to be a computationally intractable (NP-hard) problem (Letchford et al., 2012). More recent works have studied learning SE in stochastic games, both from practical perspective (Rajeswaran et al., 2020; Mishra et al., 2020; Huang et al., 2022) and theoretical perspective (Bai et al., 2021; Zhong et al., 2021). The results in this paper differ from this line of work in two ways. Firstly, our framework abstracts the response model of an agent's effective environment in that it does not model it through a rational agency with a utility function. Instead, it is more aligned with the approaches that learn the response function of the follower agent (Sinha et al., 2016; Kar et al., 2017; Sessa et al., 2020), out of which the closest to our work is (Sessa et al., 2020) that considers repeated games. Secondly, given that we consider solution concepts from performative prediction rather than SE, we focus on repeated retraining as the algorithm of interest, rather than bi-level optimization approaches.

Other related work: We also draw a connection to other RL frameworks. Naturally, this work relates to RL settings that study non-stationary environments. These include recent learning-theoretic results, such as (Gajane et al., 2018;

Fei et al., 2020; Domingues et al., 2021; Cheung et al., 2020; Wei & Luo, 2021) that allow non-stationary rewards and transitions provided a bounded number or amount of changes (under no-regret regime), the extensive literature on learning under adversarial reward functions (e.g., (Even-Dar et al., 2004; Neu et al., 2012; Dekel & Hazan, 2013; Rosenberg & Mansour, 2019)), or the recent works on learning under corrupted feedback (Lykouris et al., 2021). However, the setting of this paper is more structured, i.e., the environment responds to the deployed policy and does not arbitrarily change. Our work is also broadly related to the extensive literature on multi-agent RL literature – we refer the reader to (Zhang et al., 2021) for a selective overview. A canonical example of a multi-agent setting that closely relates to the setting of this paper is human-AI cooperation, where the AI's policy influences the human behavior (Dimitrakakis et al., 2017; Nikolaidis et al., 2017b; Crandall et al., 2018; Radanovic et al., 2019; Carroll et al., 2019). In fact, our experiments are inspired by human-AI cooperative interaction. While the algorithmic framework of repeated retraining has been discussed in some of the works on cooperative AI (e.g., see (Carroll et al., 2019)), these works do not provide a formal treatment of the problem at hand. Finally, this paper also relates to the extensive literature on offline RL (Levine et al., 2020) as the learner faces an offline RL problem when performing repeated retraining with finite samples.We utilize the analysis of (Zhan et al., 2022) to establish finite sample guarantees, under a standard assumption on sample generation (Munos & Szepesvári, 2008; Farahmand et al., 2010; Xie & Jiang, 2021), and overlap in occupancy measure (Munos & Szepesvári, 2008; Zhan et al., 2022). Note that offline RL has primarily focused on static RL settings in which the policy of a learner does not affect the model of the underlying environment.

2. Model

We are primarily concerned with Markov Decision Processes (MDPs) with a fixed state space S, action set A, discount factor γ, and starting state distribution ρ. The reward and the probability transition functions of the MDP will be functions of the adopted policy. We consider infinite-horizon setting where the learner's goal is to minimize the total sum of discounted rewards. We will write s_k to denote the state visited at time-step k and a_k to denote the action taken at time-step k. 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., $M(\pi) = (S, A, P_\pi, r_\pi, \rho)$. Note that only the reward and the transition probability function change according to the policy π.

When an agent adopts policy π and the underlying MDP is $M(\pi') = (S, A, P_{\pi'}, r_{\pi'}, \rho)$ the probability of a trajectory $\tau = (s_k, a_k)_{k=0}^{\infty}$ is given as $\mathbb{P}(\tau) = \rho(s_0)\prod_{k=1}^{\infty} P_{\pi'}(s_{k+1}|s_k, \pi(s_k))$. We will write $\tau \sim \mathbb{P}^{\pi}_{\pi'}$ to denote such a trajectory τ. Given a policy π and an underlying MDP $M(\pi')$ we write $V^{\pi}_{\pi'}(\rho)$ to define the value function w.r.t. the starting state distribution ρ. This is defined as follows.

$$V^{\pi}_{\pi'}(\rho) = \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{\pi'}}\left[\sum_{k=0}^{\infty}\gamma^k r_{\pi'}(s_k, a_k)|\rho\right] \tag{1}$$

Solution Concepts: Given the definition of the value function eq. (1), we can now define the solution concepts for our setting. First we define the performative value function of a policy which is the expected total return of the policy given the environment changes in response to the policy.

Definition 1 (Performative Value Function)**.** *Given a policy π, and a starting state distribution $\rho \in \Delta(S)$, the performative value function $V^{\pi}_{\pi}(\rho)$ is defined as $V^{\pi}_{\pi}(\rho) = \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{\pi}}[\sum_{t=0}^{\infty}\gamma^t r_{\pi}(s_t, a_t)|\rho]$.*

We now define the performatively optimal policy, which maximizes performative value function.

Definition 2 (Performatively Optimal Policy)**.** *A policy π is performatively optimal if it maximizes performative value function, i.e., $\pi \in \arg\max_{\pi'} V^{\pi'}_{\pi'}(\rho)$.*

We will write π_P to denote the performatively optimal policy. Although, π_P maximizes the performative value function, it need not be stable, i.e., the policy 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 3 (Performatively Stable Policy)**.** *A policy π is performatively stable if it satisfies the condition $\pi \in \arg\max_{\pi'} V^{\pi'}_{\pi}(\rho)$.*

We will write π_S to denote the performatively stable policy. The definition of performatively stable policy implies that if the underlying MDP is $M(\pi_S)$ then an optimal policy is π_S. This means after deploying the policy π_S in the MDP $M(\pi_S)$ the environment doesn't change and the learner is also optimizing her reward in this stable environment.We next show that even for an MDP with a single state, these two solution concepts can be very different.

Example: Consider an MDP with single state s and two actions a and b. If a policy plays arm a with probability θ and b with probability $1-\theta$ then we have $r(s, a) = \frac{1}{2} - \epsilon\theta$ and $r(s, b) = \frac{1}{2} + \epsilon\theta$ for some $\epsilon < \frac{1}{4}$. Note that if $\theta_S = 0$ then both the actions give same rewards, and the learner can just play action b. Therefore, a policy that always plays action b is a stable policy and achieves a reward of $\frac{1}{2(1-\gamma)}$. On the other hand, a policy that always plays action a with

probability $\theta = 1/4$ has the performative value function of

$$\frac{\theta(1/2 - \epsilon\theta)}{1-\gamma} + \frac{(1-\theta)(1/2 + \epsilon\theta)}{1-\gamma} = \frac{1/2 + \epsilon/8}{1-\gamma}$$

So, for $\epsilon > 0$, a performatively optimal policy can achieve higher value function than a stable policy.

We will mainly consider with the tabular MDP setting where the number of states and actions are finite. Even though solving tabular MDP in classic reinforcement learning problem is relatively straight-forward, we will see that even the tabular setting raises many interesting challenges for the setting of performative reinforcement learning.

Discounted State, Action Occupancy Measure: Note that it is not a priori clear if there always exists a performatively stable policy (as defined in (2)). This is because such existence guarantee is usually established through a fixed-point argument, but the set of optimal solutions need not be convex. If both π_1 and π_2 optimizes (2), then their convex combination might not be optimal. So, in order to find a stable policy, we instead consider the linear programming formulation of reinforcement learning. Given a policy π, its long-term discounted state-action occupancy measure in the MDP $M(\pi)$ is defined 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 an occupancy measure d, one can consider the following policy π^d which has occupancy measure d.

$$\pi^d(a|s) = \begin{cases} \frac{d(s,a)}{\sum_b d(s,b)} & \text{if } \sum_a d(s,a) > 0 \\ \frac{1}{A} & \text{otherwise} \end{cases} \tag{2}$$

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

$$d_S \in \arg\max_{d\geq 0} \sum_{s,a} d(s,a) r_d(s,a) \tag{3}$$
$$\text{s.t.} \sum_a d(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P_d(s',a,s)\ \forall s$$

With slight abuse of notation we are writing P_d as P_{π^d} (as defined in equation (2)). If either the probability transition function or the reward function changes drastically in response to the occupancy measure then the optimization problem 3 need not even have a stable point. Therefore, as is standard in performative prediction, we make the following sensitivity assumption regarding the underlying environment.

Assumption 1. *The reward and probability transition mappings are $(\varepsilon_r, \varepsilon_p)$-sensitive i.e. the following holds for any two occupancy measures d and d'*

$$\|r_d - r_{d'}\|_2 \leq \varepsilon_r \|d - d'\|_2, \ \|P_d - P_{d'}\|_2 \leq \varepsilon_p \|d - d'\|_2$$

The motivation behind this assumption comes from multi-agent systems, particularly the Stackelberg game. Suppose the leader agent changes her policy and in response, the follower agents change their policies according to a smooth response function. Then from the leader's perspective, the environmental change is smooth. Furthermore, it is also possible to state the assumption in terms of policies i.e. change in the environment is bounded by the change in policies. This is because similar policies imply similar state visitations (e.g. see lemma 14.1 in (Agarwal et al., 2021)). The converse is not always true and our assumption is weaker.

Suppose reward (r_d) and the transition (P_d) are fixed in eq. (3), then the objective function of eq. (3) is convex (in fact linear), and the set of optimal solutions is convex, a simple application of Kakutani fixed point theorem (Glicksberg, 1952) shows that there always exists a performative stable solution.[1]

Proposition 1. *Suppose assumption (1) holds for some constants $(\varepsilon_r, \varepsilon_p)$, then the optimization problem 3 always has a fixed point.*

3. Convergence of Repeated Retraining

Even though the optimization problem (3) is guaranteed to have a stable solution, it is not clear that repeatedly optimizing this objective converges to such a point. We now consider a regularized version of the optimization problem (3), and attempt to obtain a stable solution of the regularized problem. In subsection (3.3) we will show that such a stable solution guarantees approximate stability with respect to the original unregularized objective (3).

$$\max_{d\geq 0} \sum_{s,a} d(s,a) r_d(s,a) - \frac{\lambda}{2}\|d\|_2^2 \tag{4}$$
$$\text{s.t.} \sum_a d(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P_d(s',a,s)\ \forall s$$

Here $\lambda > 0$ is a constant that determines the strong-concavity of the objective. Before analyzing the behavior of repeatedly optimizing the new objective (4) we discuss two important issues. First, we consider quadratic regularization for simplicity, and our results can be easily extended to any strongly-convex regularizer. In particular, we use the strong convexity of L_2 norm to show that the solution of the optimization problem in (3) forms a contraction mapping. If we use L_1 norm then this mapping might not be a contraction. But note that, the regularized objective in (3) still has a fixed point since the objective is still concave. So we still believe that repeated optimization converges to a stable point but we

[1] The proof of this result and all other results are provided in the appendix.

might only have convergence in the limit. Second, we apply regularization in the occupancy measure space, whereas regularization in policy space is commonly used (Mnih et al., 2016). However, value function is generally a non-convex function of policy and it is not immediately clear whether the solution of the optimization problem (eq. (3)) in the policy space also gives a contraction mapping. Since the performatively stable occupancy measure d_S is not known, a common strategy adopted is repeated policy optimization. At time t, the learner obtains the occupancy measure d_t, and deploys the policy π_t (as defined in eq. (2)). In response, the environment changes to $P_t = P_{d_t}$ and $r_t = r_{d_t}$, and the learning agent solves the following optimization problem to obtain the next occupancy measure d_{t+1}.

$$\max_{d \geq 0} \sum_{s,a} d(s,a) r_t(s,a) - \frac{\lambda}{2} \|d\|_2^2 \tag{5}$$
$$\text{s.t.} \sum_a d(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P_t(s',a,s) \; \forall s$$

We next show that repeatedly solving the problem (5) converges to a stable point.

Theorem 1. *Suppose assumption 1 holds with* $\lambda > \frac{12S^{3/2}(2\epsilon_r + 5S\epsilon_p)}{(1-\gamma)^4}$. *Let* $\mu = \frac{12S^{3/2}(2\epsilon_r + 5S\epsilon_p)}{\lambda(1-\gamma)^4}$. *Then for any* $\delta > 0$ *we have*

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

Here we discuss some of the main challenges behind the proof of this theorem.

- The primal objective function (5) is strongly concave but the feasible region of the optimization problem changes with each iteration. So we cannot apply the results from performative prediction (Perdomo et al., 2020), and instead, look at the dual objective which is $A(1-\gamma)^2/\lambda$-strongly convex.
- Although the dual problem is strongly convex, it does not satisfy Lipschitz continuity w.r.t. P. However, we show that the norm of the optimal solution of the dual problem is bounded by $O\left(S/(1-\gamma)^2\right)$ and this is sufficient to show that the dual objective is Lipschitz-continuous with respect to P at the dual optimal solution. We show that the proof argument used in Perdomo et al. (2020) works if we replace global Lipschitz-continuity by such local Lipschitz-continuity.
- Finally, we translate back the bound about the dual solution to a guarantee about the primal solution ($\|d_t - d_S\|_2$) using the strong duality of the optimization problem (5). This step crucially uses the quadratic regularization in the primal.

Here we make several observations regarding the assumptions required by Theorem 1. First, Theorem 1 suggests that for a given sensitivity (ϵ_r, ϵ_p) and discount factor γ, one can choose the parameter λ so that the convergence to a stable point is guaranteed. Second, for a given value of λ and γ if the sensitivity is small enough, then repeatedly optimizing 5 converges to a stable point.

3.1. Gradient Ascent Based Algorithm

We now extend our result for the setting where the learner does not fully solve the optimization problem 5 every iteration. Rather, the learner takes a gradient step with respect to the changed environment every iteration. Let $\mathcal{C}_t$ denote the set of occupancy measures that are compatible with probability transition function P_t.

$$\mathcal{C}_t = \Bigg\{ d : \sum_a d(s,a) = \rho(s) + \gamma \sum_{s',a} d(s',a) P_t(s',a,s) \; \forall s \\ \text{and } d(s,a) \geq 0 \; \forall s,a \Bigg\} \tag{6}$$

Then the gradient ascent algorithm first takes a gradient step according to the objective function $r_t^\top d - \frac{\lambda}{2}\|d\|_2^2$ and then projects the resulting occupancy measure onto $\mathcal{C}_t$.

$$\begin{aligned} d_{t+1} &= \text{Proj}_{\mathcal{C}_t}\left(d_t + \eta \cdot (r_t - \lambda d_t)\right) \\ &= \text{Proj}_{\mathcal{C}_t}\left((1-\eta\lambda) d_t + \eta r_t\right) \end{aligned} \tag{7}$$

Here $\text{Proj}_{\mathcal{C}}(v)$ finds a point in $\mathcal{C}$ that is closest to v in L_2-norm. We next show that repeatedly taking projected gradient ascent steps with appropriate step size η converges to a stable point.

Theorem 2. *Let* $\lambda \geq \max\left\{4\epsilon_r, 2S, \frac{20\gamma^2 S^{1.5}(\epsilon_r + \epsilon_p)}{(1-\gamma)^2}\right\}$, *step-size* $\eta = \frac{1}{\lambda}$ *and* $\mu = \sqrt{\frac{64\gamma^2\epsilon_p^2}{(1-\gamma)^4}\left(1 + \frac{30\gamma^4 S^2}{(1-\gamma)^4}\right)}$. *Suppose assumption 1 holds with* $\epsilon_p < \min\left\{\frac{\gamma S}{3}, \frac{(1-\gamma)^4}{100\gamma^3 S}\right\}$. *Then for any* $\delta > 0$ *we have*

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

Proof Sketch: First, the projection step 7 can be computed through the following convex program.

$$\min_{d \geq 0} \frac{1}{2}\|d - (1-\eta\lambda) d_t - \eta r_t\|_2^2 \tag{8}$$
$$\text{s.t.} \sum_a d(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P_t(s',a,s) \; \forall s$$

Even though the objective above is convex, its feasible region changes with time. So we again look at the dual objective which is strongly concave and has a fixed feasible region. Given an occupancy measure d_t, let $\text{GD}_\eta(d_t)$ be the optimal solution of the problem (7). We show that if the step-size η is chosen small enough then the operator $\text{GD}_\eta(\cdot)$

is a contraction mapping by first proving the corresponding optimal dual solution forms a contraction, and then using strong duality to transfer the guarantee back to the primal optimal solutions. Finally, we show that the fixed point of the mapping $\mathrm{GD}_\eta(\cdot)$ indeed coincides with the performatively stable solution d_S.

In order to simplify the proof of Theorem 2, we substituted $\eta = 1/\lambda$ and then chose a suitable value of λ. This means that as $\epsilon_r, \epsilon_p \to 0$, λ needn't approach zero. However, it is possible to choose a value of λ that goes to zero as ϵ_r, ϵ_p approach zero. In particular, this can achieved by setting $\eta < 1/S$ and imposing additional constraints on ϵ_p. The proof of Theorem 2 attempts to show that the projected gradient step results in a contraction with a constant term of the form $\gamma^2\epsilon_p^2(1+2\eta S)^2$. Now if $\eta < 1/S$ then this term is smaller than 1 when ϵ_p is small.

3.2. Finite Sample Guarantees

So far we assumed that after deploying the policy corresponding to the occupancy measure d_t we observe the updated environment $M_t = (S, A, P_t, r_t, \gamma)$. However, in practice, we do not have access to the true model but only have access to samples from the updated environment. Our setting is more challenging than the finite samples setting considered by Perdomo et al. (2020). Unlike the supervised learning setting, we do not have access to independent samples from the new environment. At time t we can deploy policy π_t corresponding to the occupancy measure d_t, and can access trajectories from the new environment M_t only through the policy π_t. Therefore, at every step, the learner faces an offline reinforcement learning problem where the policy π_t is a behavioral policy.

A standard assumption in offline reinforcement learning is overlap in occupancy measure between the behavior policy and a class of target policies (Munos & Szepesvári, 2008). Without such overlap, one can get no information regarding the optimal policy from the trajectories visited by the behavioral policy. We make the following assumption regarding the overlap in occupancy measure between a deployed policy and the optimal policy in the changed environment.

Assumption 2. *Given an occupancy measure d, let $\rho_d^\star$ be the optimal occupancy measure maximizing eq. (5), and $\overline{d}$ be the occupancy measure of π^d in P_d. There exists $B > 0$ s.t.*

$$\max_{s,a} \left| \frac{\rho_d^\star(s,a)}{\overline{d}(s,a)} \right| \le B \quad \forall d$$

When there is no performativity, $\overline{d}$ equals d and the assumption states overlap between the occupancy measure of the deployed policy and the optimal policy. This is the standard assumption of single policy coverage in offline reinforcement learning. When there is performativity, $\overline{d}$ can be different than d since the deployed policy π^d might have occupancy measure different than d in the changed model P_d, and in that case we require overlap between $\overline{d}$ and the optimal occupancy measure. Assumption (2) is also significantly weaker than the uniform coverage assumption which requires $\max_d \max_{s,a} \overline{d}(s,a) > 0$ as it allows the possibility that $\overline{d}(s,a) = 0$ as long as the optimal policy doesn't visit state s or never takes action a in state s.

Data: We assume the following model of sample generation at time t. Given the occupancy measure d_t let the normalized occupancy measure be $\widetilde{d}_t(s,a) = (1-\gamma)d_t(s,a)$. First, sample a state, action pair (s_i, a_i) i.i.d as $(s_i, a_i) \sim \widetilde{d}_t$, then reward $r_i \sim r_t(s_i, a_i)$, and finally the next state $s_i' \sim P_t(\cdot|s_i, a_i)$. We have access to m_t such tuples at time t and the data collected at time is given as $\mathcal{D}_t = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{m_t}$. We would like to point out that this is a standard model of sample generation in offline reinforcement learning (see e.g. (Munos & Szepesvári, 2008; Farahmand et al., 2010; Xie & Jiang, 2021)).

With finite samples, the learner needs to optimize an empirical version of the optimization problem 5. We follow the recent literature on offline reinforcement learning (Zhan et al., 2022) and consider the Lagrangian of eq. (5).

$$\begin{aligned}
\mathcal{L}(d, h; M_t) &= d^\top r_t - \frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s) \times \\
&\left(-\sum_a d(s,a) + \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P_t(s',a,s) \right) \\
&= -\frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\rho(s) + \sum_{s,a} d_t(s,a)\frac{d(s,a)}{d_t(s,a)} \times \\
&\left(r_t(s,a) - h(s) + \gamma \sum_{s'} P_t(s,a,s')h(s') \right)
\end{aligned}$$

The above expression motivates the following empirical version of the Lagrangian.

$$\hat{\mathcal{L}}(d, h; M_t) = -\frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\rho(s) + \sum_{i=1}^{m_t} \frac{d(s_i,a_i)}{d_t(s_i,a_i)} \times \frac{r(s_i,a_i) - h(s_i) + \gamma h(s_i')}{m_t(1-\gamma)} \tag{9}$$

We repeatedly solve for a saddle point of the objective (9).

$$(d_{t+1}, h_{t+1}) = \arg\max_d \arg\min_h \hat{\mathcal{L}}(d, h; M_t) \tag{10}$$

The next theorem provides convergence guarantees of the repeated optimization procedure (10) provided that the number of samples is large enough.

Theorem 3 (Informal Statement)**.** *Suppose assumption 1 holds with $\lambda \ge \frac{24S^{3/2}(2\epsilon_r + 5S\epsilon_p)}{(1-\gamma)^4}$, and assumption 2 holds*

with parameter B. Let $\mu = \frac{24S^{3/2}(2\epsilon_r+5S\epsilon_p)}{(1-\gamma)^4}$. For any $\delta > 0$, and error probability p if we repeatedly solve the optimization problem (10) with number of samples $m_t \geq \widetilde{O}\left(\frac{A^2B^2}{\delta^4(2\epsilon_r+5S\epsilon_p)^2}\ln\left(\frac{t}{p}\right)\right)$[2] then with probability at least $1-p$ we have

$$\|d_t - d_S\|_2 \leq \delta \; \forall t \geq (1-\mu)^{-1}\ln\left(2/\delta(1-\gamma)\right)$$

Proof Sketch:

- We first show that the empirical version of the Lagrangian $\hat{\mathcal{L}}(d,h;M_t)$ is close to the true Lagrangian $\mathcal{L}(d,h;M_t)$ with high probability. This is shown using the Chernoff-Hoeffding inequality and an ϵ-net argument over the variables. Here we use the fact that for the dual variables we can just consider the space $\mathcal{H} = \left\{h : \|h\|_2 \leq 3S/(1-\gamma)^2\right\}$ as the optimal solution is guaranteed to exist in this space.
- We then show that an optimal saddle point of the empirical Lagrangian (9) is close to the optimal solution of the true Largrangian. In particular, if $\left|\mathcal{L}(d,h;M_t) - \hat{\mathcal{L}}(d,h;\hat{M}_t)\right| \leq \epsilon$ then we are guaranteed that $\|d_{t+1} - \mathrm{GD}(d_t)\|_2 \leq O(\epsilon)$. Here $\mathrm{GD}(d_t)$ denotes the solution obtained from optimizing the exact function.
- By choosing an appropriate number of samples, we can make the error term ϵ small enough, and establish the following recurrence relation: $\|d_{t+1} - d_S\|_2 \leq \beta\delta + \beta\|d_t - d_S\|_2$ for $\beta < 1/2$. The rest of the proof follows the main idea of the proof of Theorem 3.10 from (Mendler-Dünner et al., 2020). If $\|d_t - d_S\|_2 > \delta$ then we get a contraction mapping. On the other hand, if $\|d_t - d_S\|_2 \leq \delta$ then subsequent iterations cannot move d_t outside of the δ-neighborhood of d_S.

3.3. Approximating the Unregularized Objective

Theorem (1) shows that repeatedly optimizing objective (4) converges to a stable policy (say d_S^λ) with respect to the regularized objective (4). Here we show that the solution d_S^λ approximates the performatively stable and performatively optimal policy with respect to the unregularized objective (3).

Theorem 4. *There exists a choice of the regularization parameter (λ) such that repeatedly optimizing objective (5) converges to a policy (d_S^λ) with the following guarantee*[3]

$$\sum_{s,a} r_{d_S^\lambda}(s,a)d_S^\lambda(s,a) \geq \max_{d\in\mathcal{C}(d_S^\lambda)}\sum_{s,a} r_{d_S^\lambda}(s,a)d(s,a) - O\left(S^{3/2}(\epsilon_r + S\epsilon_p)/(1-\gamma)^6\right)$$

[2]Here we ignore terms that are logarithmic in S, A, and $1/\delta$.

[3]$C(\widetilde{d})$ denotes the set of occupancy measures that are feasible with respect to $\mathcal{P}(\widetilde{d}) = P_{\widetilde{d}}$.

Note that as $\epsilon = \max\{\epsilon_r, \epsilon_p\}$ converges to zero, the policy d_S^λ also approaches a performatively stable solution with respect to the original unregularized objective.

Theorem 5 (Informal Statement)**.** *Let d_{PO} be the performatively optimal solution with respect to the unregularized objective and let $\varepsilon = \max\{\epsilon_r, \epsilon_p\}$. Then there exists a value of λ such that repeatedly optimizing objective (5) converges to a policy (d_S^λ) with the following guarantee*

$$\sum_{s,a} r_{d_S^\lambda}(s,a)d_S^\lambda(s,a) \geq \sum_{s,a} r_{d_{PO}}(s,a)d_{PO}(s,a) - O\left(\max\left\{\frac{S^{5/3}A^{1/3}\epsilon^{2/3}}{(1-\gamma)^{14/3}}, \frac{\epsilon S}{(1-\gamma)^4}\right\}\right)$$

We again see that as ϵ converges to zero, d_S^λ approaches a performatively optimal solution with respect to the original objective. The proof of theorem (5) uses the following bound on the distance between the performatively stable solution and the optimal solution.$\left\|d_S^\lambda - d_{PO}^\lambda\right\|_2 \leq O\left(\frac{S^2\sqrt{A}}{\lambda(1-\gamma)^4}\left(\epsilon_r\left(1+\gamma\sqrt{S}\right) + \epsilon_p\frac{\gamma S}{(1-\gamma)^2}\right)\right)$

We believe that the bounds of theorems (5) and (4) can be improved with more careful analysis. However, the error bound should grow as γ decreases. This is because the diameter (or max L_2 norm) of occupancy measure is most $1/(1-\gamma)^2$ and even in performative prediction such an approximation bound grows with the diameter of the model.[4]

4. Experiments

In this section, we experimentally evaluate the performance of various repeated retraining methods, and determine the effects of various parameters on convergence. [5] All experiments are conducted on a grid-world environment proposed by (Triantafyllou et al., 2021).[6] We next describe how this environment is adapted for simulating performative reinforcement learning.

Gridworld: We consider the grid-world environment shown in Figure 3, in which $n+1$ agents share control over an actor. The agents' objective is to guide the actor from some initial state S to the terminal state G, while minimizing their total discounted cost. We will call the first agent the principal, and the other n agents the followers. In each state, the agents select their actions simultaneously. The principal agent, A_1, chooses the direction of the actor's next move by taking one of four actions (left, right, up, and down).

Any other agent, A_j decides to either intervene or not

[4]For example, see proposition E.1 (Perdomo et al., 2020), which is stated for diameter 1 and convex loss function.

[5]Code source: https://github.com/gradanovic/icml2023-performative-rl-paper-code

[6]The original single-agent version of this environment can be found in (Voloshin et al., 2019).

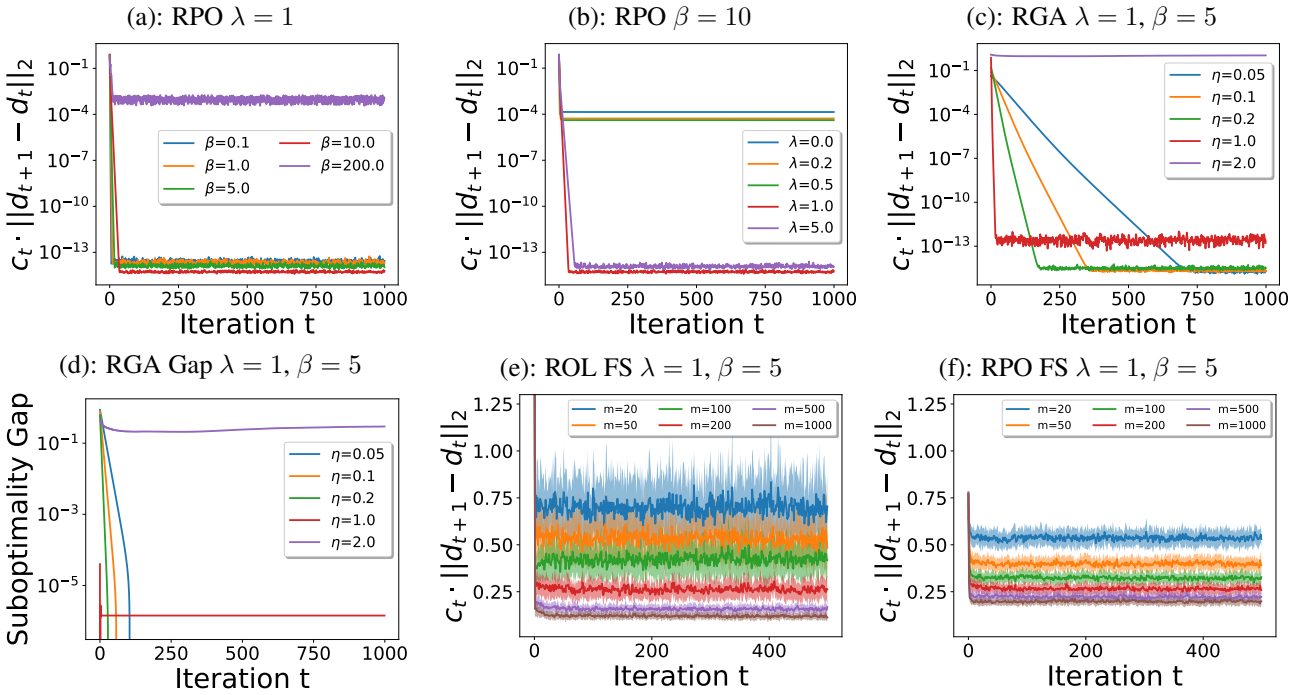

Figure 2: Experimental results for *Gridworld*. All plots were generated with $\gamma = 0.9$ and 1000 iterations. We normalized the distance between iterations t and $t+1$ with $c_t = \frac{1}{||d_t||_2}$. RPO stands for repeated policy optimization, RGA for repeated gradient ascent, ROL for repeatedly solving (empirical) Lagrangian and FS for finite samples. The parameters are λ (regularization), β (smoothness), η (step-size), and m (number of trajectories).

in A_1's action. If the majority of the n follower agents choose not to intervene, then the actor moves one cell towards the direction chosen by A_1, otherwise it moves one cell towards the new direction chosen by the majority of the followers. Note that the principal and the followers' policies determine a policy for the actor agent. The cost at each state is defined according to the grid-world shown in Figure 3. If the actor visits a blank or an S cell, then all the agents receive a small negative reward equal to -0.01. If an F cell is visited, then a slightly more increased cost equal to -0.02 is imposed and for H cells a severe penalty of -0.5 is inflicted. Additionally, whenever any A_j decides to intervene, it pays an additional cost of -0.05.

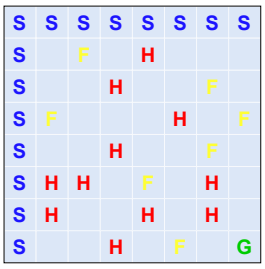

Figure 3: Gridworld

Response Model: We implement agent A_1 as a learner who performs repeated retraining. The initial policy of agent A_1 is an ϵ-optimal single-agent policy ($\epsilon = 0.1$) assuming that no other agent intervenes. Subsequently, agent A_1 performs one type of repeated retraining (e.g. gradient ascent).The follower agents, on the other hand, always respond to the policy of A_1 according to a response model. In particular, given a policy of A_1 (say π_1), we first compute an optimal Q-value function for agent A_j, $Q_j^{*|\pi_1}$. Note that $Q_j^{*|\pi_1}$ is computed w.r.t. a perturbed grid-world, and which was generated by performing a random cell perturbation on the grid-world of Figure 3. The perturbed grid-worlds are different for different agents. Then we compute an average Q-function defined as $\overline{Q}^{*|\pi_1} = \frac{1}{n}\sum_{j=2}^{n+1} \overline{Q}_j^{*|\pi_1}$. Then a policy π_2 adopted by the Boltzmann softmax operator with parameter β. Then a policy π_2 is determined by the Boltzmann softmax operator with temperature parameter β, $\pi_2(a_i|s) = \frac{e^{\beta \cdot \overline{Q}^{*|\pi_1}(s,a_i)}}{\sum_j e^{\beta \cdot \overline{Q}^{*|\pi_1}(s,a_j)}}$. Note that the new policy π_2 effectively plays the role of a changing environment by responding to the majority of the n follower agents. Additionally, parameter β controls the smoothness of the changing environment, as viewed by A_1.

Repeated Policy Optimization: We first consider the scenario where agent A_1 gets complete knowledge of the updated reward and probability transition function at time t. In our setting, this means that A_1 decides on π_1^t, all the other agents respond according to the softmax operator and jointly determines π_2^t, and then agent A_1 observes the new policy π_2^t. This lets A_1 to compute new probability transition function P_t, and reward function r_t, and solve optimization

problem (5). The solution is the new occupancy measure d_1^{t+1} for A_1, and A_1 computes new policy π_1^{t+1} for time $t+1$ by normalization using eq. (2). Figure 1(a) shows the convergence results of the repeated policy optimization for different values of β, with λ fixed to 1. We see that if the response function of the environment (i.e., the policy of agent A_2) is not smooth enough (e.g., for $\beta = 200$), the algorithm fails to converge to a stable solution. In figure 1(b) we fix β to 10 and vary the value of parameter λ (strength of regularization). We see that the algorithm converges only for large enough values of the constant λ. Furthermore, we observe that the convergence is faster as λ increases.

Repeated Gradient Ascent: We now see what happens if agent A_1 uses repeated gradient ascent instead of fully optimizing the objective each iteration. Here also A_1 fully observes π_2^t (hence P_t and r_t) at time t. However, instead of full optimization, A_1 performs a projected gradient ascent step (7) to compute the next occupancy measure d_1^{t+1}. Figure 1(c) shows the performance of repeated gradient ascent for different values of the step-size η. We observe that when η is small (e.g,. $\eta \leq O(1/\lambda)$), the learner converges to a stable solution. But the learner diverges for large η. Additionally, the convergence is faster for step-size closer to $1/\lambda$ (as suggested by Theorem 2). Since, repeated gradient ascent does not fully solve the optimization problem (5), we also plot the suboptimality gap of the current solution 1(d). This is measured as the difference between the objective value for the best feasible solution (w.r.t. M_t) and current solution (d_1^t), and then normalized by the former. As the step-size η is varied, we see a trend similar to figure 1(c).

Effect of Finite Samples: Finally, we investigate the scenario where A_1 does not know π_2^t at time t, and obtains samples from the new environment M_t by deploying π_1^t. In our experiments, instead of sampling from the occupancy measure, A_1 directly samples m trajectories of fixed length (up to 100) following policy π_1^t. We considered two approaches for obtaining the new policy π_1^{t+1}. First, A_1 solves the empirical Lagrangian (9) through an iterative method. In particular, we use an alternate optimization based method (algorithm (1)) where h_n is updated through follow the regularized leader (FTRL) algorithm and d_n is updated through best response. [7]

Second, A_1 computes estimates of probability transition function ($\widehat{P}_t$), and reward function ($\widehat{r}_t$), and solves eq. (5) with these estimates. For both versions, we ran our experiments with 20 different seeds, and figures 1(e) and 1(f) show the average errors along with the standard errors for the two approaches. For both settings, we observe that as m increases, the algorithms eventually converge to a smaller neighborhood around the stable solution. However, for large number of samples ($m = 500$ or 1000), directly solving the Lagrangian (figure 1(e)) gives improved result.

[7] Since the objective (9) is linear in h and concave in d, standard arguments (Freund & Schapire, 1996) show that algorithm (1) finds an ε-approximate saddle point in $O(SAB/(1-\gamma)^2\varepsilon^2)$ iterations.

Algorithm 1 Alternating Optimization for the Empirical Lagrangian

Set $H = \frac{3S}{(1-\gamma)^2}$, and $\mathcal{H} = \{h : \|h\|_2 \leq H\}$.
for $n = 1, 2, .., N$ **do**
$h_n = \arg\min_{h\in\mathcal{H}} \sum_{n'=1}^{n-1} \left\langle \nabla_h \hat{\mathcal{L}}(d_{n'}, h; M_t), h \right\rangle + \beta \|h\|_2^2$
$d_n = \arg\max_{d\geq 0,\ d(s,a)\leq B\hat{d}_t(s,a)\forall s,a} \hat{\mathcal{L}}(d, h_n; M_t)$
end for
Return $\overline{d} = \frac{1}{N}\sum_{n=1}^{N} d_n$.

5. Conclusion

In this work, we introduce the framework of performative reinforcement learning and study under what conditions repeated retraining methods (e.g., policy optimization, gradient ascent) converges to a stable policy. In the future, it would be interesting to extend our framework to handle high dimensional state-space, and general function approximation. The main challenge with general function approximation is that a stable policy might not exist, so the first step would be to characterize under what conditions there is a fixed point. Moreover, most RL algorithms with function approximation work in the policy space. So, another challenge lies in generalizing optimization problem 5 to handle representations of states and actions.

Another interesting question is to resolve the hardness of finding stable policy with respect to the unregularized objective. To the best of our knowledge, this question is unresolved even for performative prediction with just convex loss function. It could be interesting to explore connections between our repeated optimization procedure and standard reinforcement learning methods, e.g., policy gradient methods (Mnih et al., 2016; Neu et al., 2017). However, we note that policy gradient methods typically work in the policy space, and might not even converge to a stable point under changing environments. Finally, for the finite samples setting, it would be interesting to use offline reinforcement learning algorithms (Levine et al., 2020) for improving the speed of convergence to a stable policy.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. Stelios Triantafyllou and Goran Radanovic were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Agarwal, A., Jiang, N., Kakade, S., and Sun, W. Reinforcement learning: Theory and algorithms. 2021.

Aggarwal, C. C. et al. *Recommender systems*, volume 1. Springer.

Bai, Y., Jin, C., Wang, H., and Xiong, C. Sample-efficient learning of stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems*, 34, 2021.

Bell, J., Linsefors, L., Oesterheld, C., and Skalse, J. Reinforcement learning in newcomblike environments. *Advances in Neural Information Processing Systems*, 34, 2021.

Bertsekas, D. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

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

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

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Chaney, A. J., Stewart, B. M., and Engelhardt, B. E. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 224–232, 2018.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pp. 1843–1854. PMLR, 2020.

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., Rahwan, I., et al. Cooperating with machines. *Nature communications*, 9(1):1–12, 2018.

Dekel, O. and Hazan, E. Better rates for any adversarial deterministic mdp. In *International Conference on Machine Learning*, pp. 675–683. PMLR, 2013.

Dimitrakakis, C., Parkes, D. C., Radanovic, G., and Tylkin, P. Multi-view decision processes: the helper-ai problem. *Advances in Neural Information Processing Systems*, 30, 2017.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 3538–3546. PMLR, 2021.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Experts in a markov decision process. *Advances in neural information processing systems*, 17, 2004.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33: 6743–6754, 2020.

Filar, J. and Vrieze, K. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Freund, Y. and Schapire, R. E. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pp. 325–332, 1996.

Gajane, P., Ortner, R., and Auer, P. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Glicksberg, I. L. A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

Hansen, C., Hansen, C., Maystre, L., Mehrotra, R., Brost, B., Tomasi, F., and Lalmas, M. Contextual and sequential user embeddings for large-scale music recommendation. In *Fourteenth ACM conference on recommender systems*, pp. 53–62, 2020.

Huang, P., Xu, M., Fang, F., and Zhao, D. Robust reinforcement learning as a stackelberg game via adaptively-regularized adversarial training. *arXiv preprint arXiv:2202.09514*, 2022.

Kar, D., Ford, B., Gholami, S., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., et al. Cloudy with a chance of poaching: Adversary behavior modeling and forecasting with real-world poaching data. 2017.

Letchford, J., MacDermed, L., Conitzer, V., Parr, R., and Isbell, C. L. Computing optimal strategies to commit to in stochastic games. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

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

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., and Burke, R. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 2145–2148, 2020.

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

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

Mishra, R. K., Vasal, D., and Vishwanath, S. Model-free reinforcement learning for stochastic stackelberg security games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 348–353. IEEE, 2020.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. Multiplayer performative prediction: Learning in decision-dependent games. *arXiv preprint arXiv:2201.03398*, 2022.

Neu, G., Gyorgy, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pp. 805–813. PMLR, 2012.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Nikolaidis, S., Nath, S., Procaccia, A. D., and Srinivasa, S. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 323–331, 2017a.

Nikolaidis, S., Nath, S., Procaccia, A. D., and Srinivasa, S. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 323–331, 2017b.

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

Radanovic, G., Devidze, R., Parkes, D., and Singla, A. Learning to collaborate in markov decision processes. In *International Conference on Machine Learning*, pp. 5261–5270. PMLR, 2019.

Rajeswaran, A., Mordatch, I., and Kumar, V. A game theoretic framework for model based reinforcement learning. In *International conference on machine learning*, pp. 7953–7963. PMLR, 2020.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019.

Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. Learning to play sequential games versus unknown opponents. *Advances in Neural Information Processing Systems*, 33:8971–8981, 2020.

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

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

Sinha, A., Kar, D., and Tambe, M. Learning adversary behavior in security games: A pac model perspective. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 214–222, 2016.

Triantafyllou, S., Singla, A., and Radanovic, G. On blame attribution for accountable multi-agent sequential decision making. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

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

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

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

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

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

Appendix

Table of Contents

A. Additional Information on Experimental Setup and Results

In this section, we provide additional information on the experimental setup (Section A.1), as well as additional experimental results (Section A.2). We also provide information regarding the total amount of compute time and the type of resources that were used (Section A.3).

A.1. Additional Information on Experimental Setup

Gridworld: The initial state of the actor in the grid-world is selected uniformly at random from the cells denoted by S. Additionally, the actor remains at its current cell in case it attempts a move that would lead it outside the grid-world. Regarding the reward function, all the agents receive a positive reward equal to $+1$ whenever the actor reaches the terminal state G.

Response Model: The response model of a follower agent is based on a perturbed grid-world instead of the one in Fig. 3. In other words, each of the n follower agents sees different cell costs than the ones A_1 sees. As a result, they might respond to the policy of A_1, by adopting a policy that performs unnecessary or even harmful interventions w.r.t. the grid-world of Fig. 3. A perturbed grid-world is generated from the grid-world of Fig. 3 with the following procedure. First, G and S cells stay the same between the two grid-worlds. Then, any blank, F or H cell remains unchanged with probability 0.7, and with probability 0.3 we perturb its type to blank , F or H (the perturbation is done uniformly at random).

A.2. Additional Experimental Results

In this section, we provide additional insights on the interventional policies of the follower agents. The repeated retraining method we use in these experiments is the repeated policy optimization. More specifically, we present a visual representation of the limiting environment i.e. the majority of the agents' policy in the limit, i.e., after the method has converged to a stable solution. The configurations are set to $\lambda = 1$, $\beta = 5$, and we vary discount factor γ.

As mentioned in Section 4, the policy of the follower agents can be thought of as a changing environment that responds to the policy updates of A_1. To visualize how this environment looks like in the limit, we depict in Figure 6 several limiting policies of the follower agents. From the Figures 4, and 5 we observe that for smaller discount factor, the majority of the follower agents tend to intervene closer to the goal state.

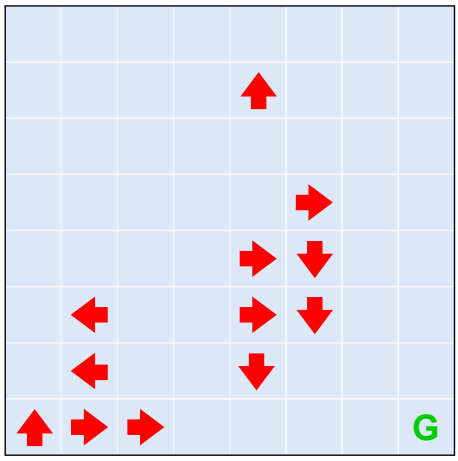

Figure 4: $\gamma = 0.5$

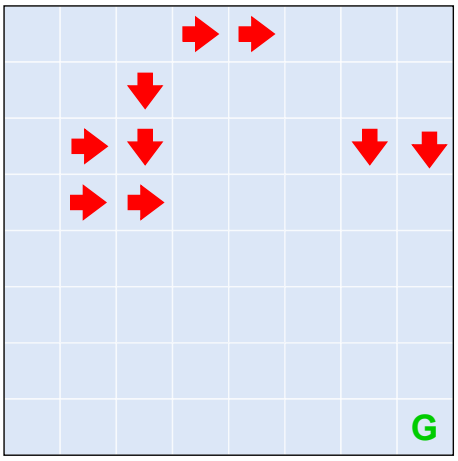

Figure 5: $\gamma = 0.9$

Figure 6: Figures 4, and 5 visualize two instances of the interventional policy of agent A_2 in the *Gridworld* environment. All figures correspond to the majority of the followers' policy at convergence for various values of the discount factor γ. Empty cells denote states where majority of the agents' most probable action is to not intervene. For cells with a red arrow the (highest probability) action of the majority of the follower agents is to intervene by forcing the actor to move one cell towards the direction of the arrow.

A.3. Total Amount of Compute and Type of Resources

All experiments were conducted on a computer cluster with machines equipped with 2 Intel Xeon E5-2667 v2 CPUs with 3.3GHz (16 cores) and 50 GB RAM. Table 1 reports the total computation times for our experiments (Section 4). Note that at each iteration of the repeated gradient ascent experiment, apart from the gradient step a full solution of the optimization problem 5 was also computed, in order to report the suboptimality gap.

Repeated Policy Optimization	767 sec
Repeated Gradient Ascent	964 sec
Repeated Policy Optimization with Finite Samples	33746 sec
Repeated Gradient Ascent with Finite Samples	35396 sec

Table 1: Total computation times for the different experiments described in Section 4.

B. Missing Proofs

B.1. Proof of Convergence of Repeated Maximization (Theorem 1)

Proof. We first compute the dual of the concave optimization problem 5. The Lagrangian is given as

$$\mathcal{L}(d,h) = d^\top r_t - \frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\left(-\sum_a d(s,a) + \rho(s) + \gamma \cdot \sum_{s',a} d(s',a)P_t(s',a,s)\right)$$

At an optimal solution we must have $\nabla_d \mathcal{L}(d,h) = 0$, which gives us the following expression for d.

$$d(s,a) = \frac{r_t(s,a)}{\lambda} - \frac{h(s)}{\lambda} + \frac{\gamma}{\lambda}\sum_{\widetilde{s}} h(\widetilde{s})P_t(s,a,\widetilde{s}) \tag{11}$$

Substituting the above value of d we get the following dual problem.

$$\begin{aligned}\min_{h\in\mathbb{R}^S} &-\frac{1}{\lambda}\sum_{s,a}h(s)r_t(s,a)+\frac{\gamma}{\lambda}\sum_{s}\sum_{s',a}h(s)r_t(s',a)P_t(s',a,s)+\sum_s h(s)\rho(s)\\ &+\frac{A}{2\lambda}\sum_s h(s)^2-\frac{\gamma}{\lambda}\sum_{s,a}h(s)\sum_{\widetilde{s}}h(\widetilde{s})P_t(s,a,\widetilde{s})+\frac{\gamma^2}{2\lambda}\sum_{s,a}\sum_{\widetilde{s},\hat{s}}h(\widetilde{s})h(\hat{s})P_t(s,a,\hat{s})P_t(s,a,\widetilde{s})\end{aligned} \tag{12}$$

Note that the dual objective is parameterized by reward function r_t and probability transition function P_t which are the parameters corresponding to the occupancy measure d_t. We will write $\mathcal{L}(\cdot; M_t)$ to denote this dual objective function.

For a given occupancy measure (i.e. $d_t = d$) we will write $\mathrm{GD}(d)$ to denote the optimal solution to the primal problem 5. We first aim to show that the operator $\mathrm{GD}(\cdot)$ is a contraction mapping. Consider two occupancy measures d and $\hat{d}$. Let r (resp. $\hat{r}$) be the reward functions in response to the occupancy measure d (resp. $\hat{d}$). Similarly, let P (resp. $\hat{P}$) be the probability transition function in response to the occupancy measure d (resp. $\hat{d}$).

Let h (resp. $\hat{h}$) be the optimal dual solutions corresponding to the occupancy measures d (resp. $\hat{d}$) i.e. $h \in \arg\max_{h'} \mathcal{L}(h'; M)$ and $\hat{h} \in \arg\max_{h'} \mathcal{L}(h'; \hat{M})$. Lemma 2 proves that the objective is $A(1-\gamma)^2/\lambda$ strongly convex. Therefore, we have the following two inequalities.

$$\mathcal{L}(h; M) - \mathcal{L}(\hat{h}; M) \geq \left(h - \hat{h}\right)^\top \nabla\mathcal{L}(\hat{h}; M) + \frac{A(1-\gamma)^2}{2\lambda}\left\|h - \hat{h}\right\|_2^2 \tag{13}$$

$$\mathcal{L}(\hat{h}; M) - \mathcal{L}(h; M) \geq \frac{A(1-\gamma)^2}{2\lambda}\left\|h - \hat{h}\right\|_2^2 \tag{14}$$

These two inequalities give us the following bound.

$$-\frac{A(1-\gamma)^2}{\lambda}\left\|h - \hat{h}\right\|_2^2 \geq (h - \hat{h})^\top \nabla\mathcal{L}(\hat{h}; M) \tag{15}$$

We now bound the Lipschitz constant of the term $(h - \hat{h})^\top \mathcal{L}_d(\hat{h}; M)$ with respect to the MDP M. Lemma 3 gives us the following bound.

$$\left\|\nabla\mathcal{L}(\hat{h}; M) - \nabla\mathcal{L}(\hat{h}; \hat{M})\right\|_2 \leq \frac{4S\sqrt{A}}{\lambda}\left\|r - \hat{r}\right\|_2 + \left(\frac{4\gamma\sqrt{SA}}{\lambda} + \frac{6\gamma\sqrt{A}S}{\lambda}\left\|\hat{h}\right\|_2\right)\left\|P - \hat{P}\right\|_2$$

Now notice that the dual variable $\hat{h}$ is actually an optimal solution and we can use lemma 4 to bound its norm by $\frac{3S}{(1-\gamma)^2}$. Furthermore, under assumption 1, we have $\|r - \hat{r}\|_2 \leq \epsilon_r \left\|d - \hat{d}\right\|_2$ and $\left\|P - \hat{P}\right\|_2 \leq \epsilon_p \left\|d - \hat{d}\right\|_2$. Substituting these bounds we get the following inequality.

$$\begin{aligned}&\left\|\nabla\mathcal{L}(\hat{h}; M) - \nabla\mathcal{L}(\hat{h}; \hat{M})\right\|_2 \leq \frac{4S\sqrt{A}}{\lambda}\epsilon_r\left\|d - \hat{d}\right\|_2 + \left(\frac{4\gamma\sqrt{SA}}{\lambda} + \frac{6\gamma\sqrt{A}S}{\lambda}\frac{3S}{(1-\gamma)^2}\right)\epsilon_p\left\|d - \hat{d}\right\|_2\\ &\leq \left(\frac{4S\sqrt{A}\epsilon_r}{\lambda} + \frac{10\gamma S^2\sqrt{A}\epsilon_p}{\lambda(1-\gamma)^2}\right)\left\|d - \hat{d}\right\|_2\end{aligned}$$

We now substitute the above bound in equation 15.

$$
\begin{aligned}
&-\frac{A(1-\gamma)^2}{\lambda}\left\|h-\hat{h}\right\|_2^2 \geq (h-\hat{h})^\top \nabla\mathcal{L}(\hat{h};M)\\
&= (h-\hat{h})^\top\left(\nabla\mathcal{L}(\hat{h};M)-\nabla\mathcal{L}(\hat{h};\hat{M})\right) \text{ [As } \hat{h} \text{ is optimal for } \mathcal{L}(\cdot;\hat{M})]\\
&\geq -\left\|h-\hat{h}\right\|_2\left\|\nabla\mathcal{L}(\hat{h};M)-\nabla\mathcal{L}(\hat{h};\hat{M})\right\|_2\\
&\geq -\left\|h-\hat{h}\right\|_2\left(\frac{4S\sqrt{A}\epsilon_r}{\lambda}+\frac{10\gamma S^2\sqrt{A}\epsilon_p}{\lambda(1-\gamma)^2}\right)\left\|d-\hat{d}\right\|_2
\end{aligned}
$$

Rearranging we get the following inequality.

$$
\left\|h-\hat{h}\right\|_2 \leq \frac{\lambda}{A(1-\gamma)^2}\left(\frac{4S\sqrt{A}\epsilon_r}{\lambda}+\frac{10\gamma S^2\sqrt{A}\epsilon_p}{\lambda(1-\gamma)^2}\right)\left\|d-\hat{d}\right\|_2
$$

Recall that $\mathrm{GD}(d)$ (resp. $\mathrm{GD}(\hat{d})$) are the optimal solution corresponding to the primal problem when the deployed occupancy measure is d (resp. $\hat{d}$). Therefore, we can apply lemma 1 to obtain the following bound.

$$
\begin{aligned}
&\left\|\mathrm{GD}(d)-\mathrm{GD}(\hat{d})\right\|_2 \leq \left(1+\frac{4\epsilon_r+6\epsilon_p\left\|\hat{h}\right\|_2}{\lambda}\right)\frac{3\sqrt{AS}}{\lambda}\left\|h-\hat{h}\right\|_2\\
&\leq \left(1+\frac{4\epsilon_r+6\epsilon_p\cdot 3S/(1-\gamma)^2}{\lambda}\right)\frac{3\sqrt{AS}}{\lambda}\frac{\lambda}{A(1-\gamma)^2}\left(\frac{4S\sqrt{A}\epsilon_r}{\lambda}+\frac{10\gamma S^2\sqrt{A}\epsilon_p}{\lambda(1-\gamma)^2}\right)\left\|d-\hat{d}\right\|_2\\
&\leq \underbrace{\left(1+\frac{4\epsilon_r+6\epsilon_p\cdot 3S/(1-\gamma)^2}{\lambda}\right)\frac{3\sqrt{S}}{\sqrt{A}(1-\gamma)^2}\left(\frac{4S\sqrt{A}\epsilon_r}{\lambda}+\frac{10\gamma S^2\sqrt{A}\epsilon_p}{\lambda(1-\gamma)^2}\right)}_{:=\beta}\left\|d-\hat{d}\right\|_2
\end{aligned}
$$

Now it can be easily verified that if $\lambda > 12S^{3/2}(1-\gamma)^{-4}(2\epsilon_r+5S\epsilon_p)$ then $\beta = \frac{12S^{3/2}(2\epsilon_r+5S\epsilon_p)}{\lambda(1-\gamma)^4} < 1$. This implies that the operator $\mathrm{GD}(\cdot)$ is a contraction mapping and the sequence of iterates $\{d_t\}_{t\geq 1}$ converges to a fixed point. In order to determine the speed of convergence let us substitute $d = d_t$ and $\hat{d} = d_S$. This gives us $\|\mathrm{GD}(d_t)-d_S\|_2 \leq \beta\|d_t-d_S\|_2$. As $\mathrm{GD}(d_t) = d_{t+1}$ we have $\|d_{t+1}-d_S\|_2 \leq \beta\|d_t-d_S\|_2$. After t iterations we have $\|d_t-d_S\|_2 \leq \beta^t\|d_0-d_S\|_2$. Therefore, if $t \geq \ln\left(\|d_0-d_S\|_2/\delta\right)/\ln(1/\beta)$ we are guaranteed that $\|d_t-d_S\|_2 \leq \delta$. Since $\|d_0-d_S\|_2 \leq \frac{2}{1-\gamma}$, the desired upper bound on the number of iterations becomes the following.

$$
\frac{\ln\left(\|d_0-d_S\|_2/\delta\right)}{\ln(1/\beta)} \leq 2(1-\beta)^{-1}\ln\left(\frac{2}{\delta(1-\gamma)}\right)
$$

□

Lemma 1. *Consider two state-action occupancy measures d and $\hat{d}$. Let $\lambda \geq 2\left(2\epsilon_r+3\epsilon_p\left\|\hat{h}\right\|_2\right)$. Then we have the following bound.*

$$
\left\|d-\hat{d}\right\|_2 \leq \left(1+\frac{4\epsilon_r+6\epsilon_p\left\|\hat{h}\right\|_2}{\lambda}\right)\frac{3\sqrt{AS}}{\lambda}\left\|h-\hat{h}\right\|_2
$$

Proof. Recall the relationship between the dual and the primal variables.

$$
d(s,a) = \frac{r_t(s,a)}{\lambda}-\frac{h(s)}{\lambda}+\frac{\gamma}{\lambda}\sum_{\widetilde{s}}h(\widetilde{s})P(s,a,\widetilde{s})
$$

This gives us the following bound on the difference $(d(s,a) - \hat{d}(s,a))^2$.

$$\begin{aligned}
\left(d(s,a) - \hat{d}(s,a)\right)^2 &\leq \frac{3}{\lambda^2}\left(r(s,a) - \hat{r}(s,a)\right)^2 + \frac{1}{\lambda^2}\left(h(s) - \hat{h}(s)\right)^2 \\
&+ \frac{3\gamma^2}{\lambda^2}\left(\sum_{s'} h(s')P(s,a,s') - \sum_{s'} \hat{h}(s')\hat{P}(s,a,s')\right)^2 \text{ [By Jensen's inequality]} \\
&= \frac{3}{\lambda^2}\left(r(s,a) - \hat{r}(s,a)\right)^2 + \frac{1}{\lambda^2}\left(h(s) - \hat{h}(s)\right)^2 \\
&+ \frac{3}{\lambda^2}\left(\sum_{s'}\left(h(s') - \hat{h}(s')\right)P(s,a,s') + \hat{h}(s')\left(P(s,a,s') - \hat{P}(s,a,s')\right)\right)^2 \\
&\leq \frac{3}{\lambda^2}\left(r(s,a) - \hat{r}(s,a)\right)^2 + \frac{1}{\lambda^2}\left(h(s) - \hat{h}(s)\right)^2 \\
&+ \frac{6}{\lambda^2}\left(\sum_{s'}\left(h(s') - \hat{h}(s')\right)P(s,a,s')\right)^2 \\
&+ \frac{6}{\lambda^2}\left(\sum_{s'}\hat{h}(s')\left(P(s,a,s') - \hat{P}(s,a,s')\right)\right)^2 \text{ [By Jensen's inequality]} \\
&\leq \frac{3}{\lambda^2}\left(r(s,a) - \hat{r}(s,a)\right)^2 + \frac{1}{\lambda^2}\left(h(s) - \hat{h}(s)\right)^2 \\
&+ \frac{6}{\lambda^2}\left\|h - \hat{h}\right\|_2^2 + \frac{6}{\lambda^2}\left\|\hat{h}\right\|_2^2 \sum_{s'}\left(P(s,a,s') - \hat{P}(s,a,s')\right)^2 \text{ [By Cauchy-Schwarz inequality]}
\end{aligned}$$

Now summing over s and a we get the following bound.

$$\left\|d - \hat{d}\right\|_2^2 \leq \frac{3}{\lambda^2}\left\|r - \hat{r}\right\|_2^2 + \frac{7AS}{\lambda^2}\left\|h - \hat{h}\right\|_2^2 + \frac{6}{\lambda^2}\left\|\hat{h}\right\|_2^2\left\|P - \hat{P}\right\|_2^2$$

We now use the assumptions $\left\|r - \hat{r}\right\|_2 \leq \epsilon_r \left\|d - \hat{d}\right\|_2$ and $\left\|P - \hat{P}\right\|_2 \leq \epsilon_p \left\|d - \hat{d}\right\|_2$.

$$\left\|d - \hat{d}\right\|_2 \leq \frac{2\epsilon_r}{\lambda}\left\|d - \hat{d}\right\|_2 + \frac{3\sqrt{AS}}{\lambda}\left\|h - \hat{h}\right\|_2 + \frac{3\epsilon_p}{\lambda}\left\|\hat{h}\right\|_2\left\|d - \hat{d}\right\|_2$$

Rearranging we get the following bound.

$$\left\|d - \hat{d}\right\|_2 \leq \left(1 - \frac{2\epsilon_r + 3\epsilon_p\left\|\hat{h}\right\|_2}{\lambda}\right)^{-1}\frac{3\sqrt{AS}}{\lambda}\left\|h - \hat{h}\right\|_2 \leq \left(1 + \frac{4\epsilon_r + 6\epsilon_p\left\|\hat{h}\right\|_2}{\lambda}\right)\frac{3\sqrt{AS}}{\lambda}\left\|h - \hat{h}\right\|_2$$

The last inequality uses the fact that $\lambda \geq 2(2\epsilon_r + 3\epsilon_p\left\|\hat{h}\right\|_2)$. □

Lemma 2. *The dual objective $\mathcal{L}_d$ (as defined in 12) is $\frac{A(1-\gamma)^2}{\lambda}$-strongly convex.*

Proof. The derivative of the dual objective $\mathcal{L}_d$ with respect to $h(s)$ is given as follows.

$$\begin{aligned}
\frac{\partial \mathcal{L}_d(h)}{\partial h(s)} &= -\frac{1}{\lambda}\sum_a r_t(s,a) + \frac{\gamma}{\lambda}\sum_{s',a} r_t(s',a)P_t(s',a,s) + \rho(s) + \frac{A}{\lambda}h(s) \\
&- \frac{\gamma}{\lambda}\sum_{\widetilde{s},a} h(\widetilde{s})\left(P_t(s,a,\widetilde{s}) + P_t(\widetilde{s},a,s)\right) + \frac{\gamma^2}{\lambda}\sum_{s',a,\widetilde{s}} h(\widetilde{s})P_t(s',a,\widetilde{s})P_t(s',a,s)
\end{aligned} \tag{16}$$

This gives us the following identity.

$$\begin{aligned}\left(\nabla\mathcal{L}_d(h) - \nabla\mathcal{L}_d(\widetilde{h})\right)^\top (h - \widetilde{h}) = \frac{A}{\lambda}\left\|h - \widetilde{h}\right\|_2^2 \\ - \frac{\gamma}{\lambda}\sum_{s,\widetilde{s},a}(h(s) - \widetilde{h}(s))\left(P_t(s,a,\widetilde{s}) + P_t(\widetilde{s},a,s)\right)(h(\widetilde{s}) - \widetilde{h}(\widetilde{s})) \\ + \frac{\gamma^2}{\lambda}\sum_{s',a}\sum_{s,\widetilde{s}}(h(s) - \widetilde{h}(\widetilde{s}))P_t(s',a,\widetilde{s})P_t(s',a,s)(h(s) - \widetilde{h}(s))\end{aligned}$$

Let us now define the matrix $M_a \in \mathbb{R}^{S\times S}$ with entries $M_a(s,s') = P_t(s,a,s')$.

$$\begin{aligned}&\left(\nabla\mathcal{L}_d(h) - \nabla\mathcal{L}_d(\widetilde{h})\right)^\top (h - \widetilde{h}) = \frac{A}{\lambda}\left\|h - \widetilde{h}\right\|_2^2 \\ &- \frac{\gamma}{\lambda}\sum_a (h - \widetilde{h})^\top\left(M_a + M_a^\top\right)(h - \widetilde{h}) + \frac{\gamma^2}{\lambda}\sum_a (h - \widetilde{h})^\top M_a^\top M_a (h - \widetilde{h}) \\ &= \frac{1}{\lambda}\sum_a (h - \widetilde{h})^\top\left(\mathrm{Id} - \gamma M_a - \gamma M_a^\top + \gamma^2 M_a^\top M_a\right)(h - \widetilde{h}) \\ &\geq \frac{A(1-\gamma)^2}{\lambda}\left\|h - \widetilde{h}\right\|_2^2\end{aligned}$$

The last inequality uses lemma 5. □

Lemma 3. *The dual function $\mathcal{L}_d$ (as defined in eq. (12)) satisfies the following bound for any h and MDP $M, \widehat{M}$.*

$$\left\|\nabla\mathcal{L}_d(h,M) - \nabla\mathcal{L}_d(h,\widehat{M})\right\|_2 \leq \frac{4S\sqrt{A}}{\lambda}\left\|r - \hat{r}\right\|_2 + \left(\frac{4\gamma\sqrt{SA}}{\lambda} + \frac{6\gamma S\sqrt{A}}{\lambda}\left\|h\right\|_2\right)\left\|P - \hat{P}\right\|_2$$

Proof. From the expression of the derivative of $\mathcal{L}_d$ with respect to h (eq. (16)) we get the following bound.

$$\begin{aligned}&\left\|\nabla\mathcal{L}_d(h,M) - \nabla\mathcal{L}_d(h,\widehat{M})\right\|_2^2 = \sum_s\Bigg\{-\frac{1}{\lambda}\sum_a(r(s,a) - \hat{r}(s,a)) \\ &+ \frac{\gamma}{\lambda}\sum_{s',a}\left(r(s',a)P(s',a,s) - \hat{r}(s',a)\hat{P}(s',a,s)\right) - \frac{\gamma}{\lambda}\sum_{\widetilde{s},a}h(\widetilde{s})(P(\widetilde{s},a,s) - \hat{P}(\widetilde{s},a,s)) \\ &- \frac{\gamma}{\lambda}\sum_{\widetilde{s},a}h(\widetilde{s})(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s})) + \frac{\gamma^2}{\lambda}\sum_{s',a,\widetilde{s}}h(\widetilde{s})\left(P(s',a,\widetilde{s})P(s',a,s) - \hat{P}(s',a,\widetilde{s})\hat{P}(s',a,s)\right)\Bigg\} \\ &\leq \frac{5A}{\lambda^2}\left\|r - \hat{r}\right\|_2^2 + \frac{5\gamma^2}{\lambda^2}\sum_s\left(\sum_{s',a}\left(r(s',a)P(s',a,s) - \hat{r}(s',a)\hat{P}(s',a,s)\right)\right)^2 \\ &+ \frac{5\gamma^2}{\lambda^2}\sum_s\left(\sum_{\widetilde{s},a}h(s)(P(\widetilde{s},a,s) - \hat{P}(\widetilde{s},a,s))\right)^2 + \frac{5\gamma^2}{\lambda^2}\sum_s\left(\sum_{\widetilde{s},a}h(\widetilde{s})(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s}))\right)^2 \\ &+ \frac{5\gamma^4}{\lambda^2}\sum_s\left(\sum_{s',a,\widetilde{s}}h(\widetilde{s})\left(P(s',a,\widetilde{s})P(s',a,s) - \hat{P}(s',a,\widetilde{s})\hat{P}(s',a,s)\right)\right)^2\end{aligned}$$

We now use four bounds to complete the proof. The following bounds use Jensen's inequality and Cauchy-Schwarz

inequality.

$$\begin{aligned}
&\text{Bound 1}: \sum_{s',a}\Big(r(s',a)P(s',a,s)-\hat{r}(s',a)\hat{P}(s',a,s)\Big)\\
&\leq \sum_{s',a}|r(s',a)-\hat{r}(s',a)|\,P(s',a,s)+\hat{r}(s',a)\left|P(s',a,s)-\hat{P}(s',a,s)\right|\\
&\leq \|r-\hat{r}\|_1+\sum_{s',a}\left|P(s',a,s)-\hat{P}(s',a,s)\right|
\end{aligned}$$

$$\begin{aligned}
&\text{Bound 2}: \sum_s\left(\sum_{\widetilde{s},a}h(\widetilde{s})(P(\widetilde{s},a,s)-\hat{P}(\widetilde{s},a,s))\right)^2\leq A\sum_s\sum_{\widetilde{s}}h(\widetilde{s})^2\sum_{\widetilde{s},a}\Big(P(\widetilde{s},a,s)-\hat{P}(\widetilde{s},a,s)\Big)^2\\
&\leq A\|h\|_2^2\left\|P-\hat{P}\right\|_2^2\\
&\text{Bound 3}: \sum_s\left(\sum_{\widetilde{s},a}h(\widetilde{s})(P(s,a,\widetilde{s})-\hat{P}(s,a,\widetilde{s}))\right)^2\leq A\sum_s\sum_{\widetilde{s}}h(\widetilde{s})^2\sum_{\widetilde{s},a}\Big(P(s,a,\widetilde{s})-\hat{P}(s,a,\widetilde{s})\Big)^2\\
&\leq A\|h\|_2^2\left\|P-\hat{P}\right\|_2^2
\end{aligned}$$

$$\begin{aligned}
&\text{Bound 4}: \sum_s\left(\sum_{s',a,\widetilde{s}}h(\widetilde{s})\Big(P(s',a,\widetilde{s})P(s',a,s)-\hat{P}(s',a,\widetilde{s})\hat{P}(s',a,s)\Big)\right)^2\\
&\leq \sum_s\sum_{\widetilde{s}}h(\widetilde{s})^2\sum_{\widetilde{s}}\left(\sum_{s',a}\Big(P(s',a,\widetilde{s})P(s',a,s)-\hat{P}(s',a,\widetilde{s})\hat{P}(s',a,s)\Big)\right)^2\\
&\leq SA\|h\|_2^2\sum_{s,\widetilde{s}}\sum_{s',a}\Big(P(s',a,\widetilde{s})P(s',a,s)-\hat{P}(s',a,\widetilde{s})\hat{P}(s',a,s)\Big)^2\\
&\leq SA\|h\|_2^2\sum_{s,\widetilde{s}}\sum_{s',a}\Big(P(s',a,\widetilde{s})(P(s',a,s)-\hat{P}(s',a,s))+\hat{P}(s',a,s)(P(s',a,\widetilde{s})-\hat{P}(s',a,\widetilde{s}))\Big)\\
&\leq 2SA\|h\|_2^2\sum_{s,\widetilde{s}}\sum_{s',a}\left|P(s',a,s)-\hat{P}(s',a,s)\right|^2+\left|P(s',a,\widetilde{s})-\hat{P}(s',a,\widetilde{s})\right|^2\\
&\leq 4S^2A\|h\|_2^2\left\|P-\hat{P}\right\|_2^2
\end{aligned}$$

Using the four upper bounds shown above, we can complete the proof.

$$\begin{aligned}
&\left\|\nabla\mathcal{L}_d(h,M)-\nabla\mathcal{L}_d(h,\widehat{M})\right\|_2^2\leq\frac{5A}{\lambda^2}\|r-\hat{r}\|_2^2+\frac{5\gamma^2}{\lambda^2}\sum_s\Big(\|r-\hat{r}\|_1+\left\|P(\cdot,\cdot,s)-\hat{P}(\cdot,\cdot,s)\right\|_1\Big)^2\\
&+\frac{10A\gamma^2}{\lambda^2}\|h\|_2^2\left\|P-\hat{P}\right\|_2^2+\frac{20S^2A\gamma^4}{\lambda^2}\|h\|_2^2\left\|P-\hat{P}\right\|_2^2\\
&\leq\left(\frac{5A}{\lambda^2}+\frac{10S^2A\gamma^2}{\lambda^2}\right)\|r-\hat{r}\|_2^2+\left(\frac{10\gamma^2SA}{\lambda^2}+\frac{10A\gamma^2}{\lambda^2}\|h\|_2^2+\frac{20S^2A\gamma^4}{\lambda^2}\|h\|_2^2\right)\left\|P-\hat{P}\right\|_2^2
\end{aligned}$$

□

Lemma 4. *The norm of the optimal solution to the dual problem (defined in 12) is bounded by $\frac{3S}{(1-\gamma)^2}$ for any choice of MDP M.*

Proof. The dual objective $\mathcal{L}_d$ is strongly-convex and has a unique solution. The optimal solution can be obtained by setting the derivative with respect to h to zero. Rearranging the derivative of the dual objective (16) we get the following systems of equations.

$$\begin{aligned}
&h(s)\left(\frac{A}{\lambda} - \frac{2\gamma}{\lambda}\sum_a P_t(s,a,s) + \frac{\gamma^2}{\lambda}\sum_{s',a} P_t(s',a,s)^2\right)\\
&+\sum_{\hat{s}\neq s} h(\hat{s})\left(-\frac{\gamma}{\lambda}\sum_a P_t(s,a,\hat{s}) - \frac{\gamma}{\lambda}\sum_a P_t(\hat{s},a,s) + \frac{\gamma^2}{\lambda}\sum_{s',a} P_t(s',a,\hat{s})P_t(s',a,s)\right)\\
&= \frac{1}{\lambda}\sum_a r_t(s,a) - \rho(s) - \frac{\gamma}{\lambda}\sum_{s',a} r_t(s',a)P_t(s',a,s)
\end{aligned}$$

Therefore let us define a matrix $B \in \mathbb{R}^{S\times S}$ and a vector $b \in \mathbb{R}^S$ with the following entries.

$$B(s,\hat{s}) = \begin{cases} \frac{A}{\lambda} - \frac{2\gamma}{\lambda}\sum_a P_t(s,a,s) + \frac{\gamma^2}{\lambda}\sum_{s',a} P_t(s',a,s)^2 & \text{if } s = \hat{s}\\ -\frac{\gamma}{\lambda}\sum_a P_t(s,a,\hat{s}) - \frac{\gamma}{\lambda}\sum_a P_t(\hat{s},a,s) + \frac{\gamma^2}{\lambda}\sum_{s',a} P_t(s',a,\hat{s})P_t(s',a,s) & \text{o.w.}\end{cases}$$

$$b(s) = \frac{1}{\lambda}\sum_a r_t(s,a) - \rho(s) - \frac{\gamma}{\lambda}\sum_{s',a} r_t(s',a)P_t(s',a,s)$$

Then the optimal solution is the solution of the system of equations $Bh = b$. We now provide a bound on the L_2-norm of such a solution. For each a, we define matrix $M_a \in R^{S\times S}$ with entries $M_a(s,\hat{s}) = P_t(s,a,\hat{s})$. Then the matrix B can be expressed as follows.

$$B = \frac{A}{\lambda}\mathrm{Id} - \frac{\gamma}{\lambda}\sum_a \left(M_a + M_a^\top\right) + \frac{\gamma^2}{\lambda}\sum_a M_a^\top M_a \succcurlyeq \frac{A(1-\gamma)^2}{\lambda}\mathrm{Id}$$

The last inequality uses lemma 5. Notice that for $\gamma < 1$ this also shows that the matrix is invertible. We can also bound the norm of the vector b.

$$\begin{aligned}
\|b\|_2^2 &\leq \sum_s \left(\frac{A}{\lambda} + \rho(s) + \frac{\gamma}{\lambda}\sum_{s',a} P_t(s',a,s)\right)^2\\
&\leq 3\frac{A^2S}{\lambda^2} + 3\|\rho\|_2^2 + 3\frac{\gamma^2}{\lambda^2}\sum_s\left(\sum_{s',a} P_t(s',a,s)\right)^2\\
&\leq 3\frac{A^2S}{\lambda^2} + 3 + \frac{3SA\gamma^2}{\lambda^2}\sum_s\sum_{s',a} P_t(s',a,s) \leq \frac{9S^2A^2}{\lambda^2}
\end{aligned}$$

Therefore, we have the following bound on the optimal value.

$$\left\|A^{-1}b\right\|_2 \leq \frac{\|b\|_2}{\lambda_{\min}(A)} \leq \frac{3S}{(1-\gamma)^2}$$

□

Lemma 5. *For each a, let the matrix $M_a \in \mathbb{R}^{S\times S}$ be defined so that $M_a(s,s') = P(s,a,s')$.*

$$\lambda_{\min}\left(\sum_a \mathrm{Id} - \gamma(M_a + M_a^\top) + \gamma^2 M_a^\top M_a\right) \geq A(1-\gamma)^2$$

Proof. Let $M_a = U_a \Sigma_a U_a^\top$ be the Eigen-decomposition of the matrix M_a. Then we have

$$\begin{aligned}
&\sum_a \mathrm{Id} - \gamma(M_a + M_a^\top) + \gamma^2 M_a^\top M_a = \sum_a (\mathrm{Id} - \gamma M_a)^\top (\mathrm{Id} - \gamma M_a) \\
&= \sum_a \left(U_a(\mathrm{Id} - \gamma\Sigma_a)U_a^\top\right)^\top \left(U_a(\mathrm{Id} - \gamma\Sigma_a)U_a^\top\right) = \sum_a U_a (\mathrm{Id} - \gamma\Sigma_a)^2 U_a^\top \\
&\succcurlyeq \sum_a \mathrm{Id}(1-\gamma)^2 = A(1-\gamma)^2\mathrm{Id}
\end{aligned}$$

The last line follows the largest eigenvalue of M_a is 1, and therefore the smallest diagonal entry of the matrix $(\mathrm{Id} - \gamma\Sigma_a)^2$ is at least $(1-\gamma)^2$. □

B.2. Proof of Convergence of Repeated Gradient Ascent (Theorem 2)

Proof. The dual of the optimization problem 8 is given as follows.

$$\begin{aligned}
\max_{h \in \mathbb{R}^S} \; & \sum_{s,a} h(s)\left((1-\eta\lambda)d_t(s,a) + \eta r_t(s,a)\right) - \sum_s h(s)\rho(s) \\
& - \gamma \cdot \sum_s h(s) \sum_{s',a} P_t(s',a,s)\left((1-\eta\lambda)d_t(s',a) + \eta r_t(s',a)\right) - \frac{A}{2}\sum_s h(s)^2 \\
& + \gamma \cdot \sum_s h(s) \sum_{s',a} h(s') P_t(s',a,s) - \frac{\gamma^2}{2} \sum_{s',s''} h(s')h(s'') \sum_{s,a} P_t(s,a,s')P_t(s,a,s'')
\end{aligned} \tag{17}$$

We will consider the equivalent minimization problem.

$$\begin{aligned}
\min_{h \in \mathbb{R}^S} \; & -\sum_{s,a} h(s)\left((1-\eta\lambda)d_t(s,a) + \eta r_t(s,a)\right) + \sum_s h(s)\rho(s) \\
& + \gamma \cdot \sum_s h(s) \sum_{s',a} P_t(s',a,s)\left((1-\eta\lambda)d_t(s',a) + \eta r_t(s',a)\right) + \frac{A}{2}\sum_s h(s)^2 \\
& - \gamma \cdot \sum_s h(s) \sum_{s',a} h(s') P_t(s',a,s) + \frac{\gamma^2}{2} \sum_{s',s''} h(s')h(s'') \sum_{s,a} P_t(s,a,s')P_t(s,a,s'')
\end{aligned} \tag{18}$$

Let us call the above objective function $\mathcal{P}(\cdot; M)$ for a given MDP M. Consider two occupancy measures d and $\hat{d}$. Let r (resp. $\hat{r}$) be the reward functions in response to the occupancy measure d (resp. $\hat{d}$) i.e. $r = \mathcal{R}(d)$ and $\hat{r} = \mathcal{R}(\hat{d})$. Similarly let P (resp. $\hat{P}$) be the probability transition functions in response to the occupancy measures d (resp. $\hat{d}$).

We will write $GD(\cdot)$ to denote the projected gradient ascent step defined in eq. (7). In particular, if we write $\mathcal{C}$ to define the set of occupancy measures feasible with respect to P, then we have

$$GD(d) = \mathrm{Proj}_{\mathcal{C}}\left((1-\eta\lambda)d + \eta r\right) \tag{19}$$

Note that $GD_\eta(d)$ is the optimal solution to the primal problem 8 with $d_t = d$. Let h be the corresponding dual optimal solution. Similarly let $\hat{h}$ be the optimal dual solution corresponding to the occupancy measure $\hat{d}$. Since h is the unique minimizer of $\mathcal{P}(\cdot; M)$ and $\mathcal{P}(\cdot; M)$ is $A(1-2\gamma)$-strongly convex for any M (lemma 7) we have the following set of inequalities.

$$\begin{aligned}
&\mathcal{P}(h; M, d) - \mathcal{P}(\hat{h}; M, d) \geq (h - \hat{h})^\top \nabla\mathcal{P}(\hat{h}; M, d) + A(1-\gamma)^2/2 \left\|h - \hat{h}\right\|_2^2 \\
&\mathcal{P}(\hat{h}; M, d) - \mathcal{P}(h; M, d) \geq A(1-\gamma)^2/2 \left\|h - \hat{h}\right\|_2^2
\end{aligned}$$

These two inequalities give us the following bound.

$$-A(1-\gamma)^2 \left\|h - \hat{h}\right\|_2^2 \geq (h - \hat{h})^\top \nabla\mathcal{P}(\hat{h}; M, d) \tag{20}$$

We now apply lemma 8 to bound the Lipschitz constant of the term $(h-\hat{h})^\top \nabla\mathcal{P}(\hat{h};M)$.

$$\begin{aligned}
&\left\|\nabla\mathcal{P}(\hat{h};M,d)-\nabla\mathcal{P}(\hat{h};\hat{M},\hat{d})\right\|_2^2 \le 5A(1-\eta\lambda)^2(1+2\gamma^2S^2)\left\|d-\hat{d}\right\|_2^2 + 5\eta^2A\left((1-\eta\lambda)^2+2\gamma^2S^2\right)\|r-\hat{r}\|_2^2 \\
&+ 5\gamma^2SA\left(\frac{2(1-\eta\lambda)^2}{(1-\gamma)^2}+2\eta^2+6\gamma^2\left\|\hat{h}\right\|_2^2\right)\left\|P-\hat{P}\right\|_2^2 \\
&\le \left(5A(1-\eta\lambda)^2(1+\eta^2\epsilon_r^2+2\gamma^2S^2)+10\eta^2\gamma^2AS^2\epsilon_r^2\right)\left\|d-\hat{d}\right\|_2^2 \\
&+ 5\gamma^2SA\left(\frac{2(1-\eta\lambda)^2}{(1-\gamma)^2}+2\eta^2+12\gamma^2\left(\frac{(1+2\eta S)^2}{(1-\gamma)^4}+4(1-\eta\lambda)^2\frac{S^2}{A(1-\gamma)^6}\right)\right)\epsilon_p^2\left\|d-\hat{d}\right\|_2^2 \\
&\le \left(5A(1-\eta\lambda)^2\left(1+\eta^2\epsilon_r^2+2\gamma^2S^2+\frac{2\gamma^2S}{(1-\gamma)^2}\epsilon_p^2+\frac{48\gamma^4S^3}{A(1-\gamma)^6}\epsilon_p^2\right)\right. \\
&\left.+\ 10\eta^2\gamma^2AS^2\epsilon_r^2+10\eta^2\gamma^2SA\epsilon_p^2+60\gamma^4SA\epsilon_p^2\frac{(1+2\eta S)^2}{(1-\gamma)^4}\right)\left\|d-\hat{d}\right\|_2^2 \\
&\le \underbrace{\left(5A(1-\eta\lambda)^2\left(1+\eta^2\epsilon_r^2+2\gamma^2S^2+\frac{50\gamma^2S^3}{(1-\gamma)^6}\epsilon_p^2\right)+10\eta^2\gamma^2S^2A(\epsilon_p^2+\epsilon_r^2)+60\gamma^4SA\epsilon_p^2\frac{(1+2\eta S)^2}{(1-\gamma)^4}\right)}_{:=\Delta^2}\left\|d-\hat{d}\right\|_2^2
\end{aligned}$$

The last inequality uses lemma 9 and assumption 1. Substituting this bound in equation 20 we get the following inequality.

$$\begin{aligned}
&-A(1-\gamma)^2\left\|h-\hat{h}\right\|_2^2 \ge (h-\hat{h})^\top\nabla\mathcal{P}(\hat{h};M,d) \\
&= (h-\hat{h})^\top\nabla\mathcal{P}(\hat{h};M,d)-(h-\hat{h})^\top\nabla\mathcal{P}(\hat{h};\hat{M},\hat{d}) \\
&\ge -\left\|h-\hat{h}\right\|_2\left\|\nabla\mathcal{P}(\hat{h};M,d)-\nabla\mathcal{P}(\hat{h};\hat{M},\hat{d})\right\|_2 \ge -\Delta\left\|h-\hat{h}\right\|_2\left\|d-\hat{d}\right\|_2
\end{aligned}$$

Rearranging we get the following inequality.

$$\begin{aligned}
&\left\|h-\hat{h}\right\|_2 \le \frac{\Delta}{A(1-\gamma)^2}\left\|d-\hat{d}\right\|_2 \\
\Rightarrow &\left\|h-\hat{h}\right\|_2^2 \le \frac{\Delta^2}{A^2(1-\gamma)^4}\left\|d-\hat{d}\right\|_2^2 \\
\Rightarrow &\frac{\left\|\mathrm{GD}_\eta(d)-\mathrm{GD}_\eta(\hat{d})\right\|_2^2}{8\gamma^2SA}-\frac{4(1-\eta\lambda)^2+4\eta^2\epsilon_r^2+8\gamma^2\left\|\hat{h}\right\|_2^2\epsilon_p^2}{8\gamma^2SA}\left\|d-\hat{d}\right\|_2^2 \le \frac{\Delta^2}{A^2(1-\gamma)^4}\left\|d-\hat{d}\right\|_2^2
\end{aligned}$$

The last line uses lemma 6. After rearranging we get the following inequality.

$$\begin{aligned}
&\left\|\mathrm{GD}_\eta(d)-\mathrm{GD}_\eta(\hat{d})\right\|_2^2 \le \left(4(1-\eta\lambda)^2+4\eta^2\epsilon_r^2+8\gamma^2\left\|\hat{h}\right\|_2^2\epsilon_p^2+\frac{8\gamma^2\Delta^2S}{A(1-\gamma)^4}\right)\left\|d-\hat{d}\right\|_2^2 \\
&\le \left(4(1-\eta\lambda)^2+4\eta^2\epsilon_r^2+\frac{16\gamma^2\epsilon_p^2(1+2\eta S)^2}{(1-\gamma)^4}+32(1-\eta\lambda)^2\frac{\gamma^2\epsilon_p^2S^2}{A(1-\gamma)^6}+\frac{8\gamma^2\Delta^2S}{A(1-\gamma)^4}\right)\left\|d-\hat{d}\right\|_2^2
\end{aligned}$$

For $\eta = 1/\lambda$ we get the following bound.

$$\left\|\mathrm{GD}_\eta(d)-\mathrm{GD}_\eta(\hat{d})\right\|_2^2 \le \left(\frac{4\epsilon_r^2}{\lambda^2}+\frac{16\gamma^2\epsilon_p^2(1+2S/\lambda)^2}{(1-\gamma)^4}+\frac{80\gamma^4S^3(\epsilon_r^2+\epsilon_p^2)}{\lambda^2(1-\gamma)^4}+\frac{480\gamma^6S^2\epsilon_p^2(1+2S/\lambda)^2}{(1-\gamma)^8}\right)\left\|d-\hat{d}\right\|_2^2$$

If we choose $\lambda \ge \max\left\{4\epsilon_r, 2S, \frac{20\gamma^2S^{1.5}(\epsilon_r+\epsilon_p)}{(1-\gamma)^2}\right\}$ we get the following condition.

$$\left\|\mathrm{GD}_\eta(d)-\mathrm{GD}_\eta(\hat{d})\right\|_2^2 \le \left(\frac{1}{2}+\frac{64\gamma^2\epsilon_p^2}{(1-\gamma)^4}+\frac{1920\gamma^6S^2\epsilon_p^2}{(1-\gamma)^8}\right)\left\|d-\hat{d}\right\|_2^2$$

For a contraction mapping we need the following condition.

$$\frac{64\gamma^2\epsilon_p^2}{(1-\gamma)^4}\left(1+\frac{30\gamma^4S^2}{(1-\gamma)^4}\right)<\frac{1}{2}$$

We consider two cases. First, if $\frac{30\gamma^4S^2}{(1-\gamma)^4}<1$ then one can show that a sufficient condition is $\epsilon_p<\gamma S/3$. On the other hand, if $\frac{30\gamma^4S^2}{(1-\gamma)^4}\geq 1$ then we need $\epsilon_p<\frac{(1-\gamma)^4}{100\gamma^3S}$. Combining the two conditions above a sufficient condition for contraction is the following.

$$\epsilon_p<\min\left\{\frac{\gamma S}{3},\frac{(1-\gamma)^4}{100\gamma^3S}\right\}$$

Now if we set $\mu=\sqrt{\frac{1}{2}+\frac{64\gamma^2\epsilon_p^2}{(1-\gamma)^4}+\frac{1920\gamma^6S^2\epsilon_p^2}{(1-\gamma)^8}}$ we get the contraction mapping: $\left\|\mathrm{GD}_\eta(d)-\mathrm{GD}_\eta(\hat{d})\right\|_2\leq\mu\left\|d-\hat{d}\right\|_2$. Let $d^\star$ be the fixed point of this contraction mapping. Using $d=d_t$ and $\hat{d}=d^\star$ we get the following sequence of inequalities.

$$\|d_{t+1}-d^\star\|_2\leq\mu\|d_t-d^\star\|_2\leq\ldots\leq\mu^t\|d_1-d^\star\|_2\leq\mu^t\frac{2}{1-\gamma}$$

The last inequality uses the fact that for any occupancy measure d we have $\|d\|_2\leq\|d\|_1\leq\frac{1}{1-\gamma}$. Rearranging we get that as long as $t\geq\ln\left(\frac{2}{\delta(1-\gamma)}\right)/\ln(1/\mu)$ we have $\|d_t-d^\star\|_2\leq\delta$.

We now show that the fixed point $d^\star$ is a stable point. In response to $d^\star$, let the probability transition function (resp. reward function) be $P^\star$ (resp. $d^\star$). Let $\mathcal{C}^\star$ be the set of occupancy measures corresponding to $d^\star$. Note that $\mathcal{C}^\star$ is a convex set. We consider two cases. First, $(1-\eta\lambda)d^\star+\eta r^\star\in\mathcal{C}^\star$. Then $d^\star=\mathrm{GD}_\eta(d^\star)=(1-\eta\lambda)d^\star+\eta r^\star$ and $r^\star-\lambda d^\star=\nabla\mathcal{P}(d^\star;P^\star,r^\star)=0$. Since $\mathcal{P}(\cdot;P^\star,r^\star)$ is a concave function the occupancy measure $d^\star$ is the optimal point and is a stable point.

Second, we consider the case when $(1-\eta\lambda)d^\star+\eta r^\star\notin\mathcal{C}^\star$. Since $d^\star=\mathrm{Proj}_{\mathcal{C}^\star}\left((1-\eta\lambda)d^\star+\eta r^\star\right)$. Since $\mathcal{C}^\star$ is a convex set, by the projection theorem (see e.g. (Bertsekas, 2009)) we have the following inequality for any $d\in\mathcal{C}^\star$.

$$\begin{aligned}&\left((1-\eta\lambda)d^\star+\eta r^\star-d^\star\right)^\top(d-d^\star)\leq 0\\ \Rightarrow\ &\eta(\lambda d^\star-r^\star)^\top(d-d^\star)\leq 0\\ \Rightarrow\ &\nabla\mathcal{P}(d^\star;P^\star,r^\star)^\top(d-d^\star)\geq 0\end{aligned}$$

This implies that $d^\star$ maximizes the function $\mathcal{P}(\cdot;P^\star,r^\star)$ over the set $\mathcal{C}^\star$ and is a stable point. □

Lemma 6. *Consider two state-action occupancy measures d and $\hat{d}$. Let h (resp. $\hat{h}$) be the dual optimal solutions to the projection (eq. (18)) corresponding to occupancy measure $d_t=d$ (resp. $\hat{d}$). Then we have the following inequality.*

$$\left\|\mathrm{GD}_\eta(d)-\mathrm{GD}_\eta(\hat{d})\right\|_2^2\leq\left(4(1-\eta\lambda)^2+4\eta^2\epsilon_r^2+8\gamma^2\left\|\hat{h}\right\|_2^2\epsilon_p^2\right)\left\|d-\hat{d}\right\|_2^2+8\gamma^2SA\left\|h-\hat{h}\right\|_2^2$$

Proof. Recall the relationship between the dual and the primal variables.

$$\mathrm{GD}_\eta(d)(s,a)=(1-\eta\lambda)d(s,a)+\eta r(s,a)-h(s)+\gamma\sum_{\widetilde{s}}h(\widetilde{s})P(s,a,\widetilde{s})$$

This gives us the following bound on the difference $(\mathrm{GD}_\eta(d)(s,a) - \mathrm{GD}_\eta(\hat{d})(s,a))^2$.

$$\begin{aligned}
&\left(\mathrm{GD}_\eta(d)(s,a) - \mathrm{GD}_\eta(\hat{d})(s,a)\right)^2 \leq 4(1-\eta\lambda)^2\left(d(s,a) - \hat{d}(s,a)\right)^2 + 4\eta^2\left(r(s,a) - \hat{r}(s,a)\right)^2 \\
&+ 4\left(h(s) - \hat{h}(s)\right)^2 + 4\gamma^2\left(\sum_{s'} h(s')P(s,a,s') - \sum_{s'}\hat{h}(s')\hat{P}(s,a,s')\right)^2 \\
&\leq 4(1-\eta\lambda)^2\left(d(s,a) - \hat{d}(s,a)\right)^2 + 4\eta^2\left(r(s,a) - \hat{r}(s,a)\right)^2 \\
&+ 4\left(h(s) - \hat{h}(s)\right)^2 + 4\gamma^2\left(\sum_{s'}\left(h(s') - \hat{h}(s')\right)P(s,a,s') + \hat{h}(s')\left(P(s,a,s') - \hat{P}(s,a,s')\right)\right)^2 \\
&\leq 4(1-\eta\lambda)^2\left(d(s,a) - \hat{d}(s,a)\right)^2 + 4\eta^2\left(r(s,a) - \hat{r}(s,a)\right)^2 \\
&+ 8\gamma^2\left(\sum_{s'}\left(h(s') - \hat{h}(s')\right)P(s,a,s')\right)^2 + 8\gamma^2\left(\sum_{s'}\hat{h}(s')\left(P(s,a,s') - \hat{P}(s,a,s')\right)\right)^2 \\
&\leq 4(1-\eta\lambda)^2\left(d(s,a) - \hat{d}(s,a)\right)^2 + 4\eta^2\left(r(s,a) - \hat{r}(s,a)\right)^2 \\
&+ 8\gamma^2\left\|h - \hat{h}\right\|_2^2 + 8\gamma^2\left\|\hat{h}\right\|_2^2\sum_{s'}\left(P(s,a,s') - \hat{P}(s,a,s')\right)^2
\end{aligned}$$

Now summing over s and a we get the following bound.

$$\left\|\mathrm{GD}_\eta(d) - \mathrm{GD}_\eta(\hat{d})\right\|_2^2 \leq 4(1-\eta\lambda)^2\left\|d - \hat{d}\right\|_2^2 + 4\eta^2\left\|r - \hat{r}\right\|_2^2 + 8\gamma^2 SA\left\|h - \hat{h}\right\|_2^2 + 8\gamma^2\left\|\hat{h}\right\|_2^2\left\|P - \hat{P}\right\|_2^2$$

We now use the assumptions $\left\|r - \hat{r}\right\|_2 \leq \epsilon_r\left\|d - \hat{d}\right\|_2$ and $\left\|P - \hat{P}\right\|_2 \leq \epsilon_p\left\|d - \hat{d}\right\|_2$.

$$\left\|\mathrm{GD}_\eta(d) - \mathrm{GD}_\eta(\hat{d})\right\|_2^2 \leq \left(4(1-\eta\lambda)^2 + 4\eta^2\epsilon_r^2 + 8\gamma^2\left\|\hat{h}\right\|_2^2\epsilon_p^2\right)\left\|d - \hat{d}\right\|_2^2 + 8\gamma^2 SA\left\|h - \hat{h}\right\|_2^2$$

□

Lemma 7. *The objective function $\mathcal{P}(\cdot; M)$ (as defined in 18) is $A(1-\gamma)^2$-strongly convex.*

Proof. The derivative of the objective function $\mathcal{P}(\cdot; M)$ with respect to $h(s)$ is given as follows.

$$\begin{aligned}
\frac{\partial\mathcal{P}(h; M_t)}{\partial h(s)} = &-\sum_a\left((1-\eta\lambda)d_t(s,a) + \eta r_t(s,a)\right) + \rho(s) + \gamma\cdot\sum_{s',a}P_t(s',a,s)\left((1-\eta\lambda)d_t(s,a) + \eta r_t(s,a)\right) \\
&+ Ah(s) - \gamma\cdot\sum_{\widetilde{s},a}h(\widetilde{s})\left(P_t(\widetilde{s},a,s) + P_t(s,a,\widetilde{s})\right) + \gamma^2\sum_{s'}h(s')\sum_{\widetilde{s},a}P_t(\widetilde{s},a,s')P_t(\widetilde{s},a,s)
\end{aligned} \tag{21}$$

This gives us the following identity.

$$\begin{aligned}
\left(\nabla\mathcal{P}(h; M_t) - \nabla\mathcal{P}(\widetilde{h}; M_t)\right)^\top(h - \widetilde{h}) = &A\left\|h - \widetilde{h}\right\|_2^2 \\
&- \gamma\cdot\sum_{s,s',a}(h(s') - \widetilde{h}(s'))(P_t(s',a,s) + P_t(s,a,s'))(h(s) - \widetilde{h}(s)) \\
&+ \gamma^2\sum_{s',s}(h(s') - \widetilde{h}(s'))\sum_{\widetilde{s},a}P_t(\widetilde{s},a,s')P_t(\widetilde{s},a,s)(h(s) - \widetilde{h}(s))
\end{aligned}$$

Now for each action a, we define the following matrix $M_a \in R^{S\times S}$ with entries $M_a(s,s') = P_t(s,a,s')$. Note that matrix M_a is row-stochastic and has eigenvalues bounded between -1 and 1.

$$\begin{aligned}\left(\nabla\mathcal{P}(h;M_t) - \nabla\mathcal{P}(\widetilde{h};M_t)\right)^\top (h-\widetilde{h}) &= A\left\|h-\widetilde{h}\right\|_2^2 - \gamma\cdot(h-\widetilde{h})^\top\left(\sum_a M_a + M_a^\top\right)(h-\widetilde{h}) \\ &\quad + \gamma^2(h-\widetilde{h})^\top\left(\sum_a M_a^\top M_a\right)(h-\widetilde{h}) \\ &= (h-\widetilde{h})^\top\sum_a\left(\mathrm{Id} - \gamma(M_a+M_a^\top) + \gamma^2 M_a^\top M_a\right)(h-\widetilde{h}) \\ &\geq A(1-\gamma)^2\left\|h-\widetilde{h}\right\|_2^2\end{aligned}$$

The last line uses lemma 5. □

Lemma 8. *The dual function (as defined in 18) satisfies the following guarantee for any h, occupancy measures $d, \hat{d}$, and MDP $M, \hat{M}$.*

$$\begin{aligned}\left\|\nabla\mathcal{P}(h;M,d) - \nabla\mathcal{P}(h;\widehat{M},\hat{d})\right\|_2 &\leq 5A(1-\eta\lambda)^2(1+2\gamma^2S^2)\left\|d-\hat{d}\right\|_2^2 \\ &\quad + 5\eta^2A\left((1-\eta\lambda)^2+2\gamma^2S^2\right)\|r-\hat{r}\|_2^2 \\ &\quad + 5\gamma^2SA\left(\frac{2(1-\eta\lambda)^2}{(1-\gamma)^2} + 2\eta^2 + 6\gamma^2\|h\|_2^2\right)\left\|P-\hat{P}\right\|_2^2 t\end{aligned}$$

Proof. From the expression of the derivative of the function $\mathcal{P}(\cdot;M,d)$ (21) we have the following bound.

$$\begin{aligned}&\left\|\nabla\mathcal{P}(h;M,d) - \nabla\mathcal{P}(h;\widehat{M},\hat{d})\right\|_2^2 = \sum_s\Bigg\{-(1-\eta\lambda)\sum_a(d(s,a)-\hat{d}(s,a)) \\ &+\eta(1-\eta\lambda)\sum_a(r(s,a)-\hat{r}(s,a)) + \gamma\eta\sum_{s',a}\left(P(s',a,s)r(s,a) - \hat{P}(s',a,s)\hat{r}(s,a)\right) \\ &+\gamma(1-\eta\lambda)\sum_{s',a}\left(P(s',a,s)d(s,a) - \hat{P}(s',a,s)\hat{d}(s,a)\right) \\ &-\gamma\cdot\sum_{\widetilde{s},a}h(\widetilde{s})\left(P(\widetilde{s},a,s) + P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s) - \hat{P}(s,a,\widetilde{s})\right) \\ &+\gamma^2\cdot\sum_{s'}h(s')\sum_{\widetilde{s},a}\left(P(\widetilde{s},a,s')P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s')\hat{P}(s,a,\widetilde{s})\right)\Bigg\}^2\end{aligned}$$

$$
\begin{aligned}
&\leq 5(1-\eta\lambda)^2 \sum_s \left(\sum_a (d(s,a) - \hat{d}(s,a)) \right)^2 + 5\eta^2(1-\eta\lambda)^2 \sum_s \left(\sum_a (r(s,a) - \hat{r}(s,a)) \right)^2 \\
&+ 5\gamma^2\eta^2 \sum_s \left(\sum_{s',a} \left(P(s',a,s) r(s,a) - \hat{P}(s',a,s)\hat{r}(s,a) \right) \right)^2 \\
&+ 5\gamma^2(1-\eta\lambda)^2 \sum_s \left(\sum_{s',a} \left(P(s',a,s) d(s,a) - \hat{P}(s',a,s)\hat{d}(s,a) \right) \right)^2 \\
&+ 5\gamma^2 \sum_s \left(\sum_{\widetilde{s},a} h(\widetilde{s}) \left(P(\widetilde{s},a,s) + P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s) - \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&+ 5\gamma^4 \sum_s \left(\sum_{s'} h(s') \sum_{\widetilde{s},a} \left(P(\widetilde{s},a,s') P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s') \hat{P}(s,a,\widetilde{s}) \right) \right)^2
\end{aligned}
$$

We now establish several bounds to complete the proof. The bounds mainly use the Cauchy-Schwarz inequality and the Jensen's inequality.

$$
\begin{aligned}
\text{Bound 1} : &\sum_s \left(\sum_{s',a} \left(P(s',a,s) d(s,a) - \hat{P}(s',a,s)\hat{d}(s,a) \right) \right)^2 \\
&\leq \sum_s \left(\sum_{s',a} P(s',a,s) \left(d(s,a) - \hat{d}(s,a) \right) + \hat{d}(s,a) \left(P(s',a,s) - \hat{P}(s',a,s) \right) \right)^2 \\
&\leq 2 \sum_s \left(\sum_{s',a} P(s',a,s) \left(d(s,a) - \hat{d}(s,a) \right) \right)^2 + 2 \sum_s \left(\sum_{s',a} \hat{d}(s,a) \left(P(s',a,s) - \hat{P}(s',a,s) \right) \right)^2 \\
&\leq 2 \sum_s \sum_{s',a} \left(d(s,a) - \hat{d}(s,a) \right)^2 \sum_{s',a} \left(P(s',a,s) \right)^2 \\
&+ 2 \sum_s \sum_a (\hat{d}(s,a))^2 \sum_a \left(\sum_{s'} P(s',a,s) - \hat{P}(s',a,s) \right)^2 \\
&\leq 2S \left\| d - \hat{d} \right\|_2^2 \sum_{s,a,s'} P(s',a,s) + \frac{2AS}{(1-\gamma)^2} \sum_{s,a,s'} \left| P(s',a,s) - \hat{P}(s',a,s) \right|^2 \\
&\leq 2S^2 A \left\| d - \hat{d} \right\|_2^2 + \frac{2AS}{(1-\gamma)^2} \left\| P - \hat{P} \right\|_2^2
\end{aligned}
$$

Similarly one can establish the following bound.

$$
\text{Bound 2} : \sum_s \left(\sum_{s',a} \left(P(s',a,s) r(s,a) - \hat{P}(s',a,s)\hat{r}(s,a) \right) \right)^2 \leq 2S^2 A \left\| r - \hat{r} \right\|_2^2 + 2AS \left\| P - \hat{P} \right\|_2^2
$$

$$\begin{aligned}
\text{Bound 3} : &\sum_s \left(\sum_{\widetilde{s},a} h(\widetilde{s}) \left(P(\widetilde{s},a,s) + P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s) - \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&\leq \sum_s \sum_{\widetilde{s}} h(\widetilde{s})^2 \sum_{\widetilde{s}} \left(\sum_a \left(P(\widetilde{s},a,s) + P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s) - \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&\leq A \left\| h \right\|_2^2 \sum_{s,\widetilde{s},a} \left(P(\widetilde{s},a,s) + P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s) - \hat{P}(s,a,\widetilde{s}) \right)^2 \\
&\leq 2A \left\| h \right\|_2^2 \left\| P - \hat{P} \right\|_2^2
\end{aligned}$$

$$\begin{aligned}
\text{Bound 4} : &\sum_s \left(\sum_{s'} h(s') \sum_{\widetilde{s},a} \left(P(\widetilde{s},a,s') P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s') \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&\leq \sum_s \sum_{s'} h(s')^2 \sum_{s'} \left(\sum_{\widetilde{s},a} \left(P(\widetilde{s},a,s') P(s,a,\widetilde{s}) - \hat{P}(\widetilde{s},a,s') \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&= \left\| h \right\|_2^2 \sum_{s,s'} \left(\sum_{\widetilde{s},a} P(s,a,\widetilde{s}) \left(P(\widetilde{s},a,s') - \hat{P}(\widetilde{s},a,s') \right) + \hat{P}(\widetilde{s},a,s') \left(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&\leq 2 \left\| h \right\|_2^2 \sum_{s,s'} \left(\sum_{\widetilde{s},a} P(s,a,\widetilde{s}) \left(P(\widetilde{s},a,s') - \hat{P}(\widetilde{s},a,s') \right) \right)^2 \\
&+ 2 \left\| h \right\|_2^2 \sum_{s,s'} \left(\sum_{\widetilde{s},a} \hat{P}(\widetilde{s},a,s') \left(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s}) \right) \right)^2 \\
&\leq 2 \left\| h \right\|_2^2 \sum_{s,s'} \sum_{\widetilde{s},a} P(s,a,\widetilde{s}) \sum_{\widetilde{s},a} \left(P(\widetilde{s},a,s') - \hat{P}(\widetilde{s},a,s') \right)^2 \\
&+ 2 \left\| h \right\|_2^2 \sum_{s,s'} \sum_{\widetilde{s},a} \hat{P}(\widetilde{s},a,s') \left(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s}) \right)^2 \\
&\leq 2 \left\| h \right\|_2^2 \left\| P - \hat{P} \right\|_2^2 \sum_{s,\widetilde{s},a} P(s,a,\widetilde{s}) + 2 \left\| h \right\|_2^2 \sum_{s,\widetilde{s},a} \left(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s}) \right)^2 \\
&\leq 4SA \left\| h \right\|_2^2 \left\| P - \hat{P} \right\|_2^2
\end{aligned}$$

We now substitute the four bounds above to complete the proof.

$$\begin{aligned}
&\left\|\nabla\mathcal{P}(h;M,d)-\nabla\mathcal{P}(h;\widehat{M},\hat{d})\right\|_2^2 \le 5(1-\eta\lambda)^2A\left\|d-\hat{d}\right\|_2^2+5\eta^2(1-\eta\lambda)^2A\left\|r-\hat{r}\right\|_2^2\\
&+5\gamma^2(1-\eta\lambda)^2\left(2S^2A\left\|d-\hat{d}\right\|_2^2+\frac{2AS}{(1-\gamma)^2}\left\|P-\hat{P}\right\|_2^2\right)\\
&+5\gamma^2\eta^2\left(2S^2A\left\|r-\hat{r}\right\|_2^2+2AS\left\|P-\hat{P}\right\|_2^2\right)\\
&+10\gamma^2\left\|h\right\|_2^2\left\|P-\hat{P}\right\|_2^2+20\gamma^4SA\left\|h\right\|_2^2\left\|P-\hat{P}\right\|_2^2\\
&\le 5A(1-\eta\lambda)^2(1+2\gamma^2S^2)\left\|d-\hat{d}\right\|_2^2+5\eta^2A\left((1-\eta\lambda)^2+2\gamma^2S^2\right)\left\|r-\hat{r}\right\|_2^2\\
&+5\gamma^2SA\left(\frac{2(1-\eta\lambda)^2}{(1-\gamma)^2}+2\eta^2+6\gamma^2\left\|h\right\|_2^2\right)\left\|P-\hat{P}\right\|_2^2
\end{aligned}$$

□

Lemma 9. *For any choice of MDP M and occupancy measure d, the L_2-norm of the optimal solution to the dual objective (as defined in 18) is bounded by*

$$\frac{1+2\eta S}{(1-\gamma)^2}+2\left|1-\eta\lambda\right|\frac{S/\sqrt{A}}{(1-\gamma)^3}.$$

Proof. The objective function $\mathcal{P}$ is strongly-convex and has a unique solution. We set the derivative with respect to h to zero and get the following system of equations.

$$\begin{aligned}
&h(s)\left(A-2\gamma\sum_a P(s,a,s)+\gamma^2\sum_{\widetilde{s},a}P(\widetilde{s},a,s)^2\right)\\
&+\sum_{s'\neq s}h(s')\left(-\gamma\sum_a(P(s',a,s)+P(s,a,s'))+\gamma^2\sum_{\widetilde{s},a}P(\widetilde{s},a,s')P(\widetilde{s},a,s)\right)\\
&=\sum_a\left((1-\eta\lambda)d(s,a)+\eta r(s,a)\right)-\rho(s)-\gamma\cdot\sum_{s',a}P(s',a,s)\left((1-\eta\lambda)d(s,a)+\eta r(s,a)\right)
\end{aligned}$$

Let us now define matrix $B\in\mathbb{R}^{S\times S}$ and a vector $b\in\mathbb{R}^S$ with the following entries.

$$B(s,s')=\begin{cases}A-2\gamma\sum_a P(s,a,s)+\gamma^2\sum_{\widetilde{s},a}P(\widetilde{s},a,s)^2 & \text{if } s'=s\\ -\gamma\sum_a(P(s',a,s)+P(s,a,s'))+\gamma^2\sum_{\widetilde{s},a}P(\widetilde{s},a,s')P(\widetilde{s},a,s) & \text{o.w.}\end{cases}$$

$$b(s)=\sum_a\left((1-\eta\lambda)d(s,a)+\eta r(s,a)\right)-\rho(s)-\gamma\cdot\sum_{s',a}P(s',a,s)\left((1-\eta\lambda)d(s,a)+\eta r(s,a)\right)$$

Then the optimal solution is the solution to the system of equations $Bx=b$. We now provide a bound on the L_2-norm of such a solution. For each action a, we define a matrix $M_a\in\mathbb{R}^{S\times S}$ with entries $M_a(s,s')=P(s,a,s')$. Then the matrix B can be expressed as follows.

$$B=A\cdot\mathrm{Id}-\gamma\sum_a(M_a+M_a^\top)+\gamma^2\sum_a M_a^\top M_a\succcurlyeq A(1-\gamma)^2\mathrm{Id}$$

The last inequality uses lemma 5. This also implies that for $\gamma<1$ the matrix B is invertible.

We now bound the norm of the vector b. We will use the fact that for any state $\sum_{s,a} d(s,a) = 1/(1-\gamma)$.

$$\begin{aligned}
\|b\|_2 \leq \|b\|_1 &\leq |1-\eta\lambda| \sum_{s,a} d(s,a) + \eta \sum_{s,a} r(s,a) + \sum_{s,a} \rho(s) \\
&+ \gamma\,|1-\eta\lambda| \cdot \sum_{s,s',a} P(s',a,s) d(s,a) + \eta\gamma \sum_{s,s',a} P(s',a,s)\,|r(s,a)| \\
&\leq \frac{|1-\eta\lambda|}{1-\gamma} + \eta SA + A + \gamma\,|1-\eta\lambda| \sum_{s'} \sqrt{\sum_{s,a} P(s',a,s)^2} \sqrt{\sum_{s,a} d(s,a)^2} + \eta\gamma SA \\
&\leq \frac{|1-\eta\lambda|}{1-\gamma} + \eta SA + A + \gamma\,|1-\eta\lambda|\,\|d\|_2 \sum_{s'} \sqrt{\sum_{s,a} P(s',a,s)} + \eta\gamma SA \\
&\leq \frac{|1-\eta\lambda|}{1-\gamma} + \eta SA + A + \gamma\,|1-\eta\lambda| \frac{S\sqrt{A}}{1-\gamma} + \eta\gamma SA \leq A(1+2\eta S) + 2\,|1-\eta\lambda| \frac{S\sqrt{A}}{1-\gamma}
\end{aligned}$$

The optimal solution to the dual objective is bounded by

$$\left\|A^{-1}b\right\|_2 \leq \frac{\|b\|_2}{\lambda_{\min}(A)} \leq \frac{1+2\eta S}{(1-\gamma)^2} + 2\,|1-\eta\lambda| \frac{S/\sqrt{A}}{(1-\gamma)^3}$$

□

B.3. Formal Statement and Proof of Convergence with Finite Samples (Theorem 3)

Theorem 6. *Suppose assumption 1 holds with $\lambda \geq \frac{24S^{3/2}(2\epsilon_r + 5S\epsilon_p)}{(1-\gamma)^4}$, and assumption 2 holds with parameter B. For a given δ, and error probability p, if we repeatedly solve the optimization problem 10 with number of samples*

$$m_t \geq \frac{64A(B+\sqrt{A})^2}{\beta^4\delta^4(2\epsilon_r + 5S\epsilon_p)^2} \left(\ln\left(\frac{t}{p}\right) + \ln\left(\frac{4S(B+\sqrt{A})}{\beta\delta(2\epsilon_r + 5S\epsilon_p)}\right)\right)$$

then we have

$$\|d_t - d_S\|_2 \leq \delta\ \forall t \geq (1-\mu)^{-1} \ln\left(\frac{2}{\delta(1-\gamma)}\right) \quad \textit{with probability at least } 1-p$$

where $\mu = \frac{24S^{3/2}(2\epsilon_r + 5S\epsilon_p)}{(1-\gamma)^4}$.

Proof. We will write $\widehat{\mathrm{GD}}(d_t)$ to denote the occupancy measure d_{t+1}.

$$(\widehat{\mathrm{GD}}(d_t), \hat{h}_{t+1}) = \arg\max_d \arg\min_h \hat{\mathcal{L}}(d,h;M_t)$$

Let us also write $\widehat{\mathrm{GD}}(d_S)$ to denote the primal solution corresponding to the stable solution d_S i.e.

$$(\widehat{\mathrm{GD}}(d_S), \hat{h}_S) = \arg\max_d \arg\min_h \hat{\mathcal{L}}(d,h;M_S)$$

Let us also recall the definition of the operator $\mathrm{GD}(\cdot)$. Given occupancy measure d_t, $\mathrm{GD}(d_t)$ is the optimal solution to the optimization problem 5 when we use the exact model M_t. Because of strong-duality this implies there exists h_{t+1} so that

$$(\mathrm{GD}(d_t), h_{t+1}) = \arg\max_d \arg\min_h \mathcal{L}(d,h;M_t)$$

Since $\mathrm{GD}(d_S) = d_S$ there also exists h_S so that

$$(d_S, h_S) = \arg\max_d \arg\min_h \mathcal{L}(d,h;M_S)$$

Because of lemma 4 we can assume the L_2-norms of the dual solutions h_{t+1}, $\hat{h}_S$, and $\hat{h}_S$ are bounded by $\frac{3S}{(1-\gamma)^2}$. Since there exists a saddle point with bounded norm, we can just consider the restricted set $\mathcal{H} = \left\{h : \|h\|_2 \leq \frac{3S}{(1-\gamma)^2}\right\}$.[8] Moreover, by assumption 2 we know that $\mathrm{GD}(d_t)(s,a)/d_t(s,a) \leq B$ for any (s,a). Therefore, we can apply lemma 10 with $\delta_1 = p/2t^2$ and $H = 3S/(1-\gamma)^2$ to get the following bound,

$$\left|\hat{\mathcal{L}}(d, h; M_t) - \mathcal{L}(d, h; M_t)\right| \leq \frac{18S^{1.5}(B+\sqrt{A})\epsilon}{(1-\gamma)^3} \tag{22}$$

as long as

$$m_t \geq \frac{4A}{\epsilon^2}(\ln(t/p) + \ln(S/(1-\gamma)\epsilon)) \tag{23}$$

$h \in \mathcal{H}$ and $\max_{s,a} d(s,a)/d_t(s,a) \leq B$. Since the event (22) holds at time t with probability at least $1 - \frac{p}{2t^2}$, by a union bound over all time steps, this event holds with probability at least $1-p$.

Note that the objective $\mathcal{L}(\cdot, h_{t+1}; M_t)$ is λ-strongly concave. Therefore, we have

$$\mathcal{L}(\widehat{\mathrm{GD}}(d_t), h_{t+1}; M_t) - \mathcal{L}(\mathrm{GD}(d_t), h_{t+1}; M_t) \leq -\frac{\lambda}{2}\left\|\mathrm{GD}(d_t) - \widehat{\mathrm{GD}}(d_t)\right\|_2^2.$$

Rearranging and using lemma 12 we get the following bound.

$$\begin{aligned}\left\|\mathrm{GD}(d_t) - \widehat{\mathrm{GD}}(d_t)\right\|_2 &\leq \sqrt{\frac{2\left(\mathcal{L}(\mathrm{GD}(d_t), h_{t+1}; M_t) - \mathcal{L}(\widehat{\mathrm{GD}}(d_t), h_{t+1}; M_t)\right)}{\lambda}} \\ &\leq \frac{6\sqrt{S^{1.5}(B+\sqrt{A})\epsilon}}{(1-\gamma)^{1.5}}\frac{1}{\sqrt{\lambda}}\end{aligned}$$

The proof of theorem 1 establishes that the operator $\mathrm{GD}(\cdot)$ is a contraction. In particular, it shows that

$$\|\mathrm{GD}(d_t) - d_S\|_2 \leq \beta \|d_t - d_S\|_2 \quad \text{for } \beta = \frac{12S^{3/2}(2\epsilon_r + 5\epsilon_p)}{\lambda(1-\gamma)^4} \text{ and } \lambda > 12S^{3/2}(1-\gamma)^{-4}(2\epsilon_r + 5S\epsilon_p)$$

This gives us the following recurrence relation on the iterates of the algorithm.

$$\begin{aligned}\|d_{t+1} - d_S\|_2 = \left\|\widehat{\mathrm{GD}}(d_t) - d_S\right\|_2 &\leq \left\|\widehat{\mathrm{GD}}(d_t) - \mathrm{GD}(d_t)\right\|_2 + \|\mathrm{GD}(d_t) - d_S\|_2 \\ &\leq \frac{6\sqrt{S^{1.5}(B+\sqrt{A})\epsilon}}{(1-\gamma)^{1.5}}\frac{1}{\sqrt{\lambda}} + \beta\|d_t - d_S\|_2\end{aligned}$$

We choose $\lambda = 24S^{3/2}(1-\gamma)^{-4}(2\epsilon_r + 5S\epsilon_p)$ which ensures $\beta < 1/2$ and gives the following recurrence.

$$\|d_{t+1} - d_S\|_2 \leq 2\sqrt{1-\gamma}\sqrt{\frac{(B+\sqrt{A})\epsilon}{2\epsilon_r + 5S\epsilon_p}} + \beta\|d_t - d_S\|_2 \leq \beta\delta + \beta\|d_t - d_S\|_2 \tag{24}$$

The last line requires the following bound on ϵ.

$$\epsilon \leq \frac{\beta^2\delta^2(2\epsilon_r + 5S\epsilon_p)}{4(1-\gamma)(B+\sqrt{A})} \tag{25}$$

Substituting the bound on ϵ in equation (23) the required number of samples at time-step t is given as follows.

$$m_t \geq \frac{64A(B+\sqrt{A})^2}{\beta^4\delta^4(2\epsilon_r + 5S\epsilon_p)^2}\left(\ln\left(\frac{t}{p}\right) + \ln\left(\frac{4S(B+\sqrt{A})}{\beta\delta(2\epsilon_r + 5S\epsilon_p)}\right)\right)$$

[8] See lemma 3 of (Zhan et al., 2022) for a proof of this statement.

In order to analyze the recurrence relation (24) we consider two cases. First, if $\|d_t - d_S\|_2 \geq \delta$ we have

$$\|d_{t+1} - d_S\|_2 \leq 2\beta \|d_t - d_S\|_2$$

Since $\beta < 1/2$ this is a contraction, and after $\ln(\|d_0 - d_S\|_2)/\ln(1/2\beta)$ iterations we are guaranteed that $\|d_t - d_S\|_2 \leq \delta$. Since $\|d_0 - d_S\|_2 \leq \frac{2}{1-\gamma}$, the required number of iterations for this event to occur is given by the following upper bound.

$$\frac{\ln(\|d_0 - d_S\|_2)}{\ln(1/2\beta)} \leq 2(1-2\beta)^{-1} \ln\left(\frac{2}{\delta(1-\gamma)}\right)$$

On the other hand, if $\|d_t - d_S\| \leq \delta$, equation (24) gives us

$$\|d_{t+1} - d_S\|_2 \leq 2\beta\delta < \delta \quad \text{[Since } \beta < 1/2\text{]}$$

Therefore, once $\|d_t - d_S\|_2 \leq \delta$ we are guaranteed that $\|d_{t'} - d_S\|_2 \leq \delta$ for any $t' \geq t$. □

Lemma 10. *Suppose $m \geq \frac{1}{\epsilon^2}\left(A\ln(2/\delta_1) + \ln(4H/\epsilon) + 2A\ln(\ln(SABH/\epsilon)/\epsilon)\right)$. Then for any occupancy measure d satisfying $\max_{s,a} d(s,a)/d_t(s,a) \leq B$ and any h with $\|h\|_2 \leq H$ the following bound holds with probability at least $1-\delta_1$.*

$$\left|\hat{\mathcal{L}}(d,h;M_t) - \mathcal{L}(d,h;M_t)\right| \leq \frac{6H\sqrt{S}(B+\sqrt{A})\epsilon}{1-\gamma}$$

Proof. Note that the expected value of the objective above equals $\mathcal{L}(d,h;M_t)$.

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathcal{L}}(d,h;M_t)\right] &= -\frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\rho(s) + \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\frac{d(s_i,a_i)}{d_t(s_i,a_i)}\left(r(s_i,a_i) - h(s_i) + \gamma h(s_i')\right)\right] \\
&= -\frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\rho(s) + \sum_{s,a} d_t(s,a)\frac{d(s,a)}{d_t(s,a)}\left(r(s,a) - h(s) + \gamma\sum_{s'} P_t(s'|s,a)h(s')\right) \\
&= \mathcal{L}(d,h;M_t)
\end{aligned}$$

By the overlap assumption 2 and the assumption $\|h\|_2 \leq H$ the following bound holds for each i,

$$\frac{1}{1-\gamma}\frac{d(s_i,a_i)}{d_t(s_i,a_i)}\left(r(s_i,a_i) - h(s_i) + \gamma h(s_i')\right) \leq \frac{2BH}{1-\gamma}.$$

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

$$P\left(\left|\hat{\mathcal{L}}(d,h;M_t) - \mathcal{L}(d,h;M_t)\right| \geq \frac{2BH}{1-\gamma}\sqrt{\frac{\ln(2/\delta_1)}{m}}\right) \leq \delta_1$$

We now extend this bound to any occupancy measure $d \in \mathcal{D}$ and h in the set $\mathcal{H} = \{h : \|h\|_2 \leq H\}$. By lemma 5.2 of (Vershynin, 2010) we can choose an ϵ-net, $\mathcal{H}_\epsilon$ of the set $\mathcal{H}$ of size at most $\left(1 + \frac{2H}{\epsilon}\right)^S$ and for any point $h \in \mathcal{H}$ we are guaranteed to find $h' \in \mathcal{H}_\epsilon$ so that $\|h - h'\|_2$.

However, such an additive error bound is not sufficient for the set of occupancy measures because of the overlap assumption 2. So instead we choose a multiplicative ϵ-net as follows. For any (s,a) we consider the grid points $d_t(s,a), (1+\epsilon)d_t(s,a), \ldots, (1+\epsilon)^p d_t(s,a)$ for $p = \log(B/d_t(s,a))/\log(1+\epsilon)$. Although $d_t(s,a)$ can be arbitrarily small, without loss of generality we can assume that $d_t(s,a) \geq \frac{\epsilon}{4SABH}$. This is because from the expression of $\mathcal{L}(d,h;M_t)$ (9), it is clear that ignoring such small terms introduces an error of at most $\epsilon/4$. Therefore we can choose $p = 2\log(SABH/\epsilon)/\log(1+\epsilon)$. Taking a Cartesian product over the set of all state, action pairs we see that we can choose an ϵ-net $\mathcal{D}_\epsilon$ so that $|\mathcal{D}_\epsilon| \leq \left(\frac{2\log(SABH/\epsilon)}{\log(1+\epsilon)}\right)^{SA} \leq \left(\frac{2\log(SABH/\epsilon)}{\epsilon}\right)^{SA}$. Notice that we are guaranteed that for any $d \in \mathcal{D}$ there exists a $d' \in \mathcal{D}_\epsilon$ such that $d(s,a)/d'(s,a) \leq 1+\epsilon$.

By a union bound over the elements in $\mathcal{H}_\epsilon$ and $\mathcal{D}_\epsilon$ the following bound holds for any $d \in \mathcal{D}_\epsilon$ and $h \in \mathcal{H}_\epsilon$ with probability at least $1 - \delta_1$.

$$\left|\hat{\mathcal{L}}(d, h; M_t) - \mathcal{L}(d, h; M_t)\right| \leq \frac{2BH}{1-\gamma} \underbrace{\sqrt{\frac{SA \ln\left(\frac{2}{\delta_1}\right) + S \ln\left(\frac{4H}{\epsilon}\right) + 2SA \ln(\ln(SABH/\epsilon)/\epsilon)}{m}}}_{:=T_m}$$

We now extend the bound above for any $d \in \mathcal{D}$ and $h \in \mathcal{H}$ using lemma 11. There exists $\widetilde{d} \in \mathcal{D}_\epsilon$ so that $\max_{s,a} d(s,a)/\widetilde{d}(s,a) \leq \epsilon$. Similarly there exists $h_\epsilon \in \mathcal{H}_\epsilon$ so that $\left\|h - \widetilde{h}\right\|_2 \leq \epsilon$. Let $\mathcal{L}^0(d, h; M_t) = \mathcal{L}(d, h; M_t) + \frac{\lambda}{2}\|d\|_2^2 - \sum_s h(s)\rho(s)$.

$$\begin{aligned}
&\left|\hat{\mathcal{L}}(d, h; M_t) - \mathcal{L}(d, h; M_t)\right| \leq \left|\hat{\mathcal{L}}^0(d, h; M_t) - \hat{\mathcal{L}}^0(\widetilde{d}, \widetilde{h}; M_t)\right| \\
&+ \left|\hat{\mathcal{L}}(\widetilde{d}, \widetilde{h}; M_t) - \mathcal{L}(\widetilde{d}, \widetilde{h}; M_t)\right| + \left|\mathcal{L}^0(\widetilde{d}, \widetilde{h}; M_t) - \mathcal{L}^0(d, h; M_t)\right| \\
&\leq \frac{2BHT_m}{1-\gamma} + \frac{6\sqrt{SA}H\epsilon}{1-\gamma} + \frac{4BH\sqrt{S}\epsilon}{1-\gamma}
\end{aligned}$$

Therefore, if $m \geq \frac{1}{\epsilon^2}\left(A \ln(2/\delta_1) + \ln(4H/\epsilon) + 2A \ln(\ln(SABH/\epsilon)/\epsilon)\right)$ then $T_m \leq \sqrt{S}\epsilon$ and we have the following bound.

$$\left|\hat{\mathcal{L}}(d, h; M_t) - \mathcal{L}(d, h; M_t)\right| \leq \frac{6H\sqrt{S}(B + \sqrt{A})\epsilon}{1-\gamma}$$

□

Lemma 11. *Suppose we are guaranteed that $\frac{d(s,a)}{\widetilde{d}(s,a)} \leq 1 + \epsilon$ and $\left\|h - \widetilde{h}\right\|_2 \leq \epsilon$, and $\|h\|_2, \left\|\widetilde{h}\right\|_2 \leq H$. Let $\mathcal{L}^0(d, h; M_t) = \mathcal{L}(d, h; M_t) + \frac{\lambda}{2}\|d\|_2^2 - \sum_s h(s)\rho(s)$ and define $\hat{\mathcal{L}}^0(d, h; M_t)$ analogously. Then the following inequalities hold.*

$$\left|\mathcal{L}^0(d, h; M_t) - \mathcal{L}^0(\widetilde{d}, \widetilde{h}; M_t)\right| \leq \frac{6\sqrt{SA}H\epsilon}{1-\gamma} \quad \textit{and}$$

$$\left|\hat{\mathcal{L}}^0(d, h; M_t) - \hat{\mathcal{L}}^0(\widetilde{d}, \widetilde{h}; M_t)\right| \leq \frac{4BH\sqrt{S}\epsilon}{1-\gamma}$$

Proof. First note that $\left\|d - \widetilde{d}\right\|_2^2 = \sum_{s,a}\left(d(s,a) - \widetilde{d}(s,a)\right)^2 \leq \sum_{s,a} d(s,a)^2\epsilon^2 \leq \frac{\epsilon^2}{(1-\gamma)^2}$.

$$\begin{aligned}
&\left|\mathcal{L}^0(d, h; M_t) - \mathcal{L}^0(\widetilde{d}, \widetilde{h}; M_t)\right| \leq \sum_{s,a}\left|d(s,a) - \widetilde{d}(s,a)\right| \\
&+ \underbrace{\left|\sum_{s,a} d(s,a)h(s) - \widetilde{d}(s,a)\widetilde{h}(s)\right|}_{:=T_1} + \gamma \underbrace{\sum_{s,a}\left|d(s,a)\sum_{s'} P_t(s,a,s')h(s') - \widetilde{d}(s,a)\sum_{s'} P_t(s,a,s')\widetilde{h}(s')\right|}_{:=T_2}
\end{aligned}$$

We now bound the terms T_1 and T_2.

$$\begin{aligned}
T_1 &= \left|\sum_{s,a} d(s,a)\left(h(s) - \widetilde{h}(s)\right) + \widetilde{h}(s)\left(d(s,a) - \widetilde{d}(s,a)\right)\right| \\
&\leq \left\|h - \widetilde{h}\right\|_1 \sum_{s,a} d(s,a) + \sqrt{\sum_s (\widetilde{h}(s))^2}\sqrt{\sum_s\left(\sum_a (d(s,a) - \widetilde{d}(s,a)\right)^2} \\
&\leq \frac{\sqrt{S}\epsilon}{1-\gamma} + \frac{H\sqrt{A}\epsilon}{1-\gamma}
\end{aligned}$$

$$
\begin{aligned}
T_2 &= \sum_{s,a} \left| d(s,a) \sum_{s'} P_t(s,a,s') h(s') - \widetilde{d}(s,a) \sum_{s'} P_t(s,a,s') \widetilde{h}(s') \right| \\
&= \sum_{s,a} \left| d(s,a) - \widetilde{d}(s,a) \right| \sum_{s'} P_t(s,a,s') \left| h(s') \right| + \sum_{s,a} \widetilde{d}(s,a) \sum_{s'} P_t(s,a,s') \left| h(s') - \widetilde{h}(s') \right| \\
&\leq \|h\|_2 \left\| d - \widetilde{d} \right\|_1 + \left\| h - \widetilde{h} \right\|_1 \sum_{s,a} \widetilde{d}(s,a) \\
&\leq \frac{\sqrt{SA} H \epsilon}{1-\gamma} + \frac{\sqrt{S}\epsilon}{1-\gamma}
\end{aligned}
$$

Substituting the bounds on T_1 and T_2 we get the following bound on $\left| \mathcal{L}(d,h;M_t) - \mathcal{L}(\widetilde{d},\widetilde{h};M_t) \right|$.

$$
\left| \mathcal{L}^0(d,h;M_t) - \mathcal{L}^0(\widetilde{d},\widetilde{h};M_t) \right| \leq \sqrt{SA}\epsilon + \frac{2\sqrt{S}\epsilon}{1-\gamma} + \frac{2\sqrt{SA}H\epsilon}{1-\gamma}
$$

We now consider bounding the difference $\left| \hat{\mathcal{L}}^0(d,h;M_t) - \hat{\mathcal{L}}^0(\widetilde{d},\widetilde{h};M_t) \right|$.

$$
\begin{aligned}
&\left| \hat{\mathcal{L}}^0(d,h;M_t) - \hat{\mathcal{L}}^0(\widetilde{d},\widetilde{h};M_t) \right| \leq \\
&+ \underbrace{\frac{1}{m(1-\gamma)} \sum_{i=1}^{m} \left| \frac{d(s_i,a_i)}{d_t(s_i,a_i)} (r(s_i,a_i) - h(s_i) + \gamma h(s_i')) - \frac{\widetilde{d}(s_i,a_i)}{d_t(s_i,a_i)} (r(s_i,a_i) - \widetilde{h}(s_i) + \gamma \widetilde{h}(s_i')) \right|}_{:=T_3}
\end{aligned}
$$

We now bound the term T_3.

$$
\begin{aligned}
T_3 &= \frac{1}{m(1-\gamma)} \sum_{i=1}^{m} \left| \frac{d(s_i,a_i)}{d_t(s_i,a_i)} (r(s_i,a_i) - h(s_i) + \gamma h(s_i')) - \frac{\widetilde{d}(s_i,a_i)}{d_t(s_i,a_i)} (r(s_i,a_i) - \widetilde{h}(s_i) + \gamma \widetilde{h}(s_i')) \right| \\
&\leq \frac{1}{m(1-\gamma)} \sum_{i=1}^{m} B \left(\left| h(s_i) - \widetilde{h}(s_i) \right| + \gamma \left| h(s_i') - \widetilde{h}(s_i') \right| \right) \\
&+ \frac{1}{m(1-\gamma)} \sum_{i=1}^{m} \left| \frac{d(s_i,a_i) - \widetilde{d}(s_i,a_i)}{d_t(s_i,a_i)} \right| \left| r(s_i,a_i) - \widetilde{h}(s_i) + \gamma \widetilde{h}(s_i') \right| \\
&\leq \frac{B}{1-\gamma} \left\| h - \widetilde{h} \right\|_1 + \frac{\epsilon B}{1-\gamma} \left(1 + 2 \|h\|_1 \right) \\
&\leq \frac{B\sqrt{S}\epsilon}{1-\gamma} + \frac{3BH\sqrt{S}\epsilon}{1-\gamma}
\end{aligned}
$$

□

Lemma 12. *Let* $(d^\star, h^\star) \in \arg\max_d \arg\min_h \mathcal{L}(d,h;M)$ *and* $(\hat{d},\hat{h}) \in \arg\max_d \arg\min_h \mathcal{L}(d,h;\widehat{M})$. *Moreover, suppose* $\left| \mathcal{L}(d,h;M) - \mathcal{L}(d,h;\widehat{M}) \right| \leq \epsilon$ *for all* d *and* h *with* $\|h\|_2 \leq 3S/(1-\gamma)^2$. *Then we have*

$$
\mathcal{L}(d^\star,h^\star) - \mathcal{L}(\hat{d},h^\star) \leq 2\epsilon
$$

Proof. We will drop the dependence on the underlying model and write $\mathcal{L}(d,h;M)$ as $\mathcal{L}(d,h)$ and $\mathcal{L}(d,h;\widehat{M})$ as $\hat{\mathcal{L}}(d,h)$.

$$
\begin{aligned}
\mathcal{L}(d^\star,h^\star) - \mathcal{L}(\hat{d},h^\star) = &\underbrace{\mathcal{L}(d^\star,h^\star) - \mathcal{L}(d^\star,\hat{h}(d^\star))}_{:=T_1} + \underbrace{\mathcal{L}(d^\star,\hat{h}(d^\star)) - \hat{\mathcal{L}}(d^\star,\hat{h}(d^\star))}_{:=T_2} \\
&+ \underbrace{\hat{\mathcal{L}}(d^\star,\hat{h}(d^\star)) - \hat{\mathcal{L}}(\hat{d},\hat{h})}_{:=T_3} + \underbrace{\hat{\mathcal{L}}(\hat{d},\hat{h}) - \hat{\mathcal{L}}(\hat{d},h^\star)}_{:=T_4} + \underbrace{\hat{\mathcal{L}}(\hat{d},h^\star) - \mathcal{L}(\hat{d},h^\star)}_{:=T_5}
\end{aligned}
$$

Here we write $\hat{h}(\widetilde{d}) = \arg\min_h \hat{\mathcal{L}}(\widetilde{d}, h)$ i.e. the dual solution that minimizes the objective $\hat{\mathcal{L}}(\widetilde{d}, \cdot)$. Since $h^\star$ minimizes $\mathcal{L}(d^\star, \cdot)$, by lemma 4 we have $\|h^\star\|_2 \leq 3S/(1-\gamma)^2$. By a similar argument $\left\|\hat{h}\right\|_2 \leq 3S/(1-\gamma)^2$. Therefore, both T_2 and T_5 are at most ϵ.

Given $d^\star$, $h^\star$ minimizes $\mathcal{L}(d^\star, \cdot)$. Therefore, the term T_1 is at most zero. By a similar argument the term T_4 is at most zero. Now for the term T_3, notice that $\hat{h} = \hat{h}(\hat{d})$ and it also minimizes the objective $\hat{\mathcal{L}}(\hat{d}, \hat{h}(d))$. Therefore, the term T_3 is also at most zero. □

B.4. Proof of Proposition 1

Proof. Let $\mathcal{D}$ be the following set $\mathcal{D} = \left\{ d \in \mathbb{R}^{S\times A} : d(s,a) \geq 0 \; \forall s, a \text{ and } \sum_{s,a} d(s,a) = \frac{1}{1-\gamma} \right\}$. We define a set-valued function $\phi : \mathcal{D} \to 2^{\mathcal{D}}$ as follows.

$$\begin{aligned} \phi(d) = \arg\max_{\widetilde{d}\geq 0} & \sum_{s,a} \widetilde{d}(s,a) r_d(s,a) \\ \text{s.t.} & \sum_a \widetilde{d}(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} \widetilde{d}(s',a) P_d(s',a,s) \; \forall s \end{aligned} \tag{26}$$

Any fixed point of $\phi(\cdot)$ corresponds to a stable point. First, note that $\phi(d)$ is non-empty as one can always choose $\widetilde{d}$ to be the occupancy measure associated with any arbitrary policy π in an MDP with probability transition function P_d. Now suppose $d_1, d_2 \in S(d)$. Then for any $\rho \in [0,1]$ it is easy to show that $\rho d_1 + (1-\rho) d_2 \in \phi(d)$. This is because all the constraints are linear, so $\rho d_1 + (1-\rho) d_2$ is feasible. Moreover, the objective is linear, so $\rho d_1 + (1-\rho) d_2$ also attains the same objective value.

We now show that the function ϕ is upper hemicontinuous. Let L be the Lagrangian of the optimization problem (26).

$$L(\widetilde{d}, h; M_d) = \sum_{s,a} \widetilde{d}(s,a) r_d(s,a) + \sum_s h(s) \left(\sum_a \widetilde{d}(s,a) - \rho(s) - \gamma \cdot \sum_{s',a} \widetilde{d}(s',a) P_d(s',a,s) \right)$$

Note that the Lagrangian is continuous (in fact linear) in $\widetilde{d}$, and continuous in d (from the assumption of (ϵ_r, ϵ_p)-sensitivity). Finally, observe the alternative definition of the function ϕ.

$$\phi(d) = \arg\max_{\widetilde{d}\in\mathcal{D}} \min_h L(\widetilde{d}, h; M_d)$$

Since the minimum of continuous functions is also continuous, and the set $\mathcal{D}$ is compact, we can apply Berge's maximum theorem to conclude that the function $\phi(\cdot)$ is upper hemicontinuous. Now an application of Kakutani fixed point theorem (Glicksberg, 1952) shows that ϕ has a fixed point.

□

B.5. Assumptions Regarding Quadratic Regularizer

Throughout we performed repeated optimization with quadratic regularization. Our proof techniques can be easily generalized if we consider a strongly convex regularizer $\mathrm{R}(d)$. Suppose, at time t we solve the following optimization problem.

$$\begin{aligned} \max_{\widetilde{d}\geq 0} & \sum_{s,a} \widetilde{d}(s,a) r_t(s,a) - \mathrm{R}(\widetilde{d}) \\ \text{s.t.} & \sum_a \widetilde{d}(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} \widetilde{d}(s',a) P_t(s',a,s) \; \forall s \end{aligned} \tag{27}$$

Since R is strongly convex, $(\mathrm{R}')^{-1}$ exists and we can use this result to show that the dual of the (27) is strongly convex. In fact, as in the proof of theorem 1 we can write down the lagrangian $\mathcal{L}(d, h)$ and at an optimal solution we must have

$\nabla_d \mathcal{L}(d,h) = 0$. This gives the following expression.

$$d(s,a) = (\mathrm{R}')^{-1}\left(r_t(s,a) - h(s) + \gamma \cdot \sum_{\widetilde{s}} h(\widetilde{s}) P_t(s,a,\widetilde{s})\right) \tag{28}$$

We can use the result above to show that the dual is strongly convex and the optimal dual solutions form a contraction. Then we can translate this guarantee back to the primal using (28).

B.6. Omitted Proofs from Subsection 3.3

We will write d^λ_{PO} to write the performatively optimal solution when using the regularization parameter λ i.e.

$$\begin{aligned} d^\lambda_{PO} \in \arg\max_{d \geq 0} & \sum_{s,a} d(s,a) r^\lambda_{PO}(s,a) - \frac{\lambda}{2}\|d\|_2^2 \\ \text{s.t.} & \sum_a d(s,a) = \rho(s) + \gamma \cdot \sum_{s',a} d(s',a) P^\lambda_{PO}(s',a,s)\ \forall s \end{aligned} \tag{29}$$

Here we write $r^\lambda_{PO} = \mathcal{R}(d^\lambda_{PO})$ and $P^\lambda_{PO} = \mathcal{P}(d^\lambda_{PO})$ to denote the reward function and probability transition function in response to the optimal occupancy measure d^λ_{PO}. We will also write d^λ_S to denote the performatively stable solution and r^λ_S (resp. P^λ_S) to denote the corresponding reward (resp. probability transition) function. The next lemma bounds the distance between d^λ_S and d^λ_{PO}.

B.6.1. Proof of Theorem 4

Proof. Suppose repeatedly maximizing the regularized objective converges to a stable solution d^λ_S i.e.

$$\sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) - \frac{\lambda}{2}\left\|d^\lambda_S\right\|_2^2 \geq \max_{d \in \mathcal{C}(d^\lambda_S)} \sum_{s,a} r_{d^\lambda_S}(s,a) d(s,a) - \frac{\lambda}{2}\|d\|_2^2$$

Therefore,

$$\begin{aligned} \sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) &\geq \max_{d \in \mathcal{C}(d^\lambda_S)} \sum_{s,a} r_{d^\lambda_S}(s,a) d(s,a) - \frac{\lambda}{2}\|d\|_2^2 \\ &\geq \max_{d \in \mathcal{C}(d^\lambda_S)} \sum_{s,a} r_{d^\lambda_S}(s,a) d(s,a) - \frac{\lambda}{2(1-\gamma)^2} \end{aligned}$$

The last inequality uses $\|d\|_2^2 = \sum_{s,a} d(s,a)^2 = (1-\gamma)^{-2} \sum_{s,a} \left((1-\gamma) d(s,a)\right)^2 \leq (1-\gamma)^{-2} \sum_{s,a} (1-\gamma) d(s,a) = (1-\gamma)^{-2}$. Now we substitute $\lambda = \frac{12 S^{3/2}(2\epsilon_r + 5 S \epsilon_p)}{(1-\gamma)^4}$ from theorem 1 and get the following bound.

$$\sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) \geq \max_{d \in \mathcal{C}(d^\lambda_S)} \sum_{s,a} r_{d^\lambda_S}(s,a) d(s,a) - \frac{6 S^{3/2}(2\epsilon_r + 5 S \epsilon_p)}{(1-\gamma)^6}$$

□

B.6.2. Formal Statement and Proof of Theorem 5

Theorem 7. *Let d_{PO} be the performatively optimal solution with respect to the original (unregularized) objective. Then there exists a choice of regularization parameter (λ) such that repeatedly optimizing objective (12) converges to a policy (d^λ_S) with the following guarantee*

$$\sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) \geq \sum_{s,a} r_{d_{PO}}(s,a) d_{PO}(s,a) - \Delta$$

where

$$\Delta = O\left(\max\left\{\frac{S A^{1/3}}{(1-\gamma)^{10/3}}\left((1+\gamma\sqrt{S})\epsilon_r + \frac{\gamma S \epsilon_p}{(1-\gamma)^2}\right)^{2/3}, \frac{\epsilon_r}{(1-\gamma)^2} + \frac{\epsilon_p S}{(1-\gamma)^4}\right\}\right)$$

Proof. Let us write h^λ_{PO} to denote the dual optimal solution i.e.

$$h^\lambda_{PO} \in \arg\min_h \mathcal{L}_d(h; M^\lambda_{PO})$$

Moreover, let h^λ_S be the dual optimal solution corresponding to the stable solution d^λ_S.

$$\begin{aligned}
&\sum_{s,a} r_{d_{PO}}(s,a) d_{PO}(s,a) - \sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) \\
=&\left(\sum_{s,a} r_{d_{PO}}(s,a) d_{PO}(s,a) - \frac{\lambda}{2}\|d_{PO}\|_2^2\right) + \frac{\lambda}{2}\|d_{PO}\|_2^2 \\
&- \left(\sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) - \frac{\lambda}{2}\left\|d^\lambda_S\right\|_2^2\right) - \frac{\lambda}{2}\left\|d^\lambda_S\right\|_2^2 \\
\leq&\left(\sum_{s,a} r_{d^\lambda_{PO}}(s,a) d^\lambda_{PO}(s,a) - \frac{\lambda}{2}\left\|d^\lambda_{PO}\right\|_2^2\right) + \frac{\lambda}{2}\|d_{PO}\|_2^2 \\
&- \left(\sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) - \frac{\lambda}{2}\left\|d^\lambda_S\right\|_2^2\right) - \frac{\lambda}{2}\left\|d^\lambda_S\right\|_2^2 \\
\leq&\mathcal{L}_d(h^\lambda_{PO}; M^\lambda_{PO}) - \mathcal{L}_d(h^\lambda_S; M^\lambda_S) + \frac{\lambda}{2}\|d_{PO}\|_2^2
\end{aligned}$$

The first inequality uses the fact that d^λ_{PO} is the performatively optimal solution with regularization parameter λ. The second inequality uses strong duality and expresses the objective in terms of optimal dual variables. We now bound the difference $\mathcal{L}_d(h^\lambda_{PO}; M^\lambda_{PO}) - \mathcal{L}_d(h^\lambda_S; M^\lambda_S)$.

$$\begin{aligned}
&\mathcal{L}_d(h^\lambda_{PO}; M^\lambda_{PO}) - \mathcal{L}_d(h^\lambda_S; M^\lambda_S) \\
=&\ \mathcal{L}_d(h^\lambda_{PO}; M^\lambda_{PO}) - \mathcal{L}_d(h^\lambda_S; M^\lambda_{PO}) + \mathcal{L}_d(h^\lambda_S; M^\lambda_{PO}) - \mathcal{L}_d(h^\lambda_S; M^\lambda_S) \\
\leq&\ \mathcal{L}_d(h^\lambda_S; M^\lambda_{PO}) - \mathcal{L}_d(h^\lambda_S; M^\lambda_S) \quad \text{[Since } h^\lambda_{PO} \text{ minimizes } \mathcal{L}_d(\cdot; M^\lambda_{PO})\text{]} \\
\leq&\ \frac{\left\|h^\lambda_S\right\|_2 \sqrt{A}}{\lambda}\left((1+\gamma\sqrt{S})\epsilon_r + \gamma(2\sqrt{S} + \left\|h^\lambda_S\right\|_2)\epsilon_p\right)\left\|d^\lambda_S - d^\lambda_{PO}\right\|_2 \text{ [By inequality 32]} \\
\leq&\ \frac{S^3 A}{\lambda^2(1-\gamma)^6}\left((1+\gamma\sqrt{S})\epsilon_r + \gamma\left(2\sqrt{S} + \frac{3S}{(1-\gamma)^2}\right)\epsilon_p\right)^2 \text{ [By lemma 13]}
\end{aligned}$$

The term $\|d_{PO}\|_2^2$ can be bounded as $\sum_{s,a} d^2_{PO}(s,a) = (1-\gamma)^{-2}\sum_{s,a}(d_{PO}(s,a)(1-\gamma))^2 \leq (1-\gamma)^{-2}\sum_{s,a} d_{PO}(s,a)(1-\gamma) = (1-\gamma)^{-2}$. This gives us the following bound.

$$\begin{aligned}
&\sum_{s,a} r_{d_{PO}}(s,a) d_{PO}(s,a) - \sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) \\
\leq&\ \frac{S^3 A}{\lambda^2(1-\gamma)^6}\left((1+\gamma\sqrt{S})\epsilon_r + \gamma\left(2\sqrt{S} + \frac{3S}{(1-\gamma)^2}\right)\epsilon_p\right)^2 + \frac{\lambda}{2(1-\gamma)^2} \\
=&\ \frac{1}{\lambda^2}\underbrace{\frac{S^3 A}{(1-\gamma)^6}\left((1+\gamma\sqrt{S})\epsilon_r + \gamma\left(2\sqrt{S} + \frac{3S}{(1-\gamma)^2}\right)\epsilon_p\right)^2}_{:=T_1} + \lambda\underbrace{\frac{1}{2(1-\gamma)^2}}_{:=T_2}
\end{aligned}$$

Note that in order to apply lemma 13 we need $\lambda \geq \lambda_0 = 2\left(2\epsilon_r + 9\epsilon_p S(1-\gamma)^{-2}\right)$. So we consider two cases. First if $(2T_1/T_2)^{-1/3} > \lambda_0$. In that case, we can use $\lambda = (2T_1/T_2)^{-1/3}$ and get the following upper bound.

$$\begin{aligned}
&\sum_{s,a} r_{d_{PO}}(s,a) d_{PO}(s,a) - \sum_{s,a} r_{d^\lambda_S}(s,a) d^\lambda_S(s,a) \leq O\left(T_1^{1/3} T_2^{2/3}\right) \\
=&\ O\left(\frac{S A^{1/3}}{(1-\gamma)^{10/3}}\left((1+\gamma\sqrt{S})\epsilon_r + \gamma\left(2\sqrt{S} + \frac{3S}{(1-\gamma)^2}\right)\epsilon_p\right)^{2/3}\right)
\end{aligned}$$

On the other hand, if $(2T_1/T_2)^{-1/3} \leq \lambda_0$ then we can substitute $\lambda = \lambda_0$ and get the following bound.

$$\begin{aligned}
&\sum_{s,a} r_{d_{PO}}(s,a) d_{PO}(s,a) - \sum_{s,a} r_{d_S^\lambda}(s,a) d_S^\lambda(s,a) \leq \frac{1}{\lambda_0^2} T_1 + \lambda_0 T_2 \\
=&\frac{1}{(T_1^{1/3} T_2^{-1/3})^2} + \lambda_0 T_2 = T_1^{1/3} T_2^{2/3} + \lambda_0 T_2 \leq 2\lambda_0 T_2 \\
=& O\left(\frac{\epsilon_r}{(1-\gamma)^2} + \frac{\epsilon_p S}{(1-\gamma)^4}\right)
\end{aligned}$$

□

Lemma 13. *Suppose* $\lambda \geq 2\left(2\epsilon_r + \frac{9\epsilon_p S}{(1-\gamma)^2}\right)$. *Then we have*

$$\left\|d_S^\lambda - d_{PO}^\lambda\right\|_2 \leq O\left(\frac{S^2\sqrt{A}}{\lambda(1-\gamma)^4}\left(\epsilon_r\left(1+\gamma\sqrt{S}\right) + \epsilon_p \frac{\gamma S}{(1-\gamma)^2}\right)\right)$$

Proof. Let us write h_{PO}^λ to denote the dual optimal solution i.e.

$$h_{PO}^\lambda \in \arg\min_h \mathcal{L}_d(h; M_{PO}^\lambda)$$

Moreover, let h_S^λ be the dual optimal solution corresponding to the stable solution d_S^λ.

Since the dual $\mathcal{L}_d(\cdot; M_{PO}^\lambda)$ objective is strongly convex (lemma 2) and h_{PO}^λ is the corresponding optimal solution we have,

$$\mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) - \mathcal{L}_d(h_{PO}^\lambda; M_{PO}^\lambda) \geq \frac{A(1-\gamma)^2}{2\lambda}\left\|h_S^\lambda - h_{PO}^\lambda\right\|_2^2 \tag{30}$$

From lemma (1) we get the following bound.

$$\left(1 - \frac{2\epsilon_r + 3\epsilon_p \left\|h_S^\lambda\right\|_2}{\lambda}\right)\left\|d_S^\lambda - d_{PO}^\lambda\right\|_2 \leq \frac{3\sqrt{AS}}{\lambda}\left\|h_S^\lambda - h_{PO}^\lambda\right\|_2$$

Substituting the above bound in eq. (30) and using lemma 4 we get the following inequality for any $\lambda > 2\epsilon_r + 9\epsilon_r S/(1-\gamma)^2$.

$$\mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) - \mathcal{L}_d(h_{PO}^\lambda; M_{PO}^\lambda) \geq \frac{(1-\gamma)^2}{18S}\left(1 - \frac{2\epsilon_r + 3\epsilon_p \left\|h_S^\lambda\right\|_2}{\lambda}\right)^2 \left\|d_S^\lambda - d_{PO}^\lambda\right\|_2^2 \tag{31}$$

We now upper bound $\mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) - \mathcal{L}_d(h_S^\lambda; M_S^\lambda)$. Using lemma 14 we get the following bound.

$$\begin{aligned}
\mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) - \mathcal{L}_d(h_S^\lambda; M_S^\lambda) &\leq \frac{\left\|h_S^\lambda\right\|_2 \sqrt{A}}{\lambda}\left(\left(1+\gamma\sqrt{S}\right)\left\|r_S^\lambda - r_{PO}^\lambda\right\|_2 + \gamma(2\sqrt{S} + \left\|h_S^\lambda\right\|_2)\left\|P_S^\lambda - P_{PO}^\lambda\right\|_2\right) \\
&\leq \frac{\left\|h_S^\lambda\right\|_2 \sqrt{A}}{\lambda}\left(\left(1+\gamma\sqrt{S}\right)\epsilon_r + \gamma(2\sqrt{S} + \left\|h_S^\lambda\right\|_2)\epsilon_p\right)\left\|d_S^\lambda - d_{PO}^\lambda\right\|_2
\end{aligned} \tag{32}$$

Note that the following sequence of inequalities hold.

$$\mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) \geq \mathcal{L}_d(h_{PO}^\lambda; M_{PO}^\lambda) \geq \mathcal{L}_d(h_S^\lambda; M_S^\lambda)$$

The first inequality is true because h_{PO}^λ minimizes $\mathcal{L}_d(\cdot; M_{PO}^\lambda)$. The second inequality holds because the primal objective at performative optimal solution (d_{PO}^λ) upper bound the primal objective at performatively stable solution (d_S^λ) and by strong duality the primal objectives are equal to the corresponding dual objectives. Therefore we must have $\mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) -$

$\mathcal{L}_d(h_S^\lambda; M_S^\lambda) \geq \mathcal{L}_d(h_S^\lambda; M_{PO}^\lambda) - \mathcal{L}_d(h_{PO}^\lambda; M_{PO}^\lambda)$, and by using equations (32) and (31) we get the following bound on $\left\|d_S^\lambda - d_{PO}^\lambda\right\|_2$.

$$\begin{aligned}\left\|d_S^\lambda - d_{PO}^\lambda\right\|_2 &\leq \frac{\left\|h_S^\lambda\right\|_2 \sqrt{A}}{\lambda}\left((1+\gamma\sqrt{S})\epsilon_r + \gamma(2\sqrt{S} + \left\|h_S^\lambda\right\|_2)\epsilon_p\right)\frac{18S}{(1-\gamma)^2}\left(1 - \frac{2\epsilon_r + 3\epsilon_p\left\|h_S^\lambda\right\|_2}{\lambda}\right)^{-2} \\ &\leq \frac{108S^2\sqrt{A}\lambda}{(1-\gamma)^4}\left(\epsilon_r(1+\gamma\sqrt{S}) + \gamma\sqrt{S}\epsilon_p\left(2 + \frac{3\sqrt{S}}{(1-\gamma)^2}\right)\epsilon_p\right)\end{aligned}$$

The last line uses lemma 4 and $\lambda \geq 2(2\epsilon_r + 9\epsilon_p S/(1-\gamma)^2)$

□

Lemma 14. $\left|\mathcal{L}_d(h; M) - \mathcal{L}_d(h; \widehat{M})\right| \leq \frac{\|h\|_2\sqrt{A}}{\lambda}\left((1+\gamma\sqrt{S})\left\|r - \hat{r}\right\|_2 + \gamma(2\sqrt{S} + \|h\|_2)\left\|P - \widehat{P}\right\|_2\right)$

Proof. From the definition of the dual objective 12 we have,

$$\begin{aligned}\left|\mathcal{L}_d(h; M) - \mathcal{L}_d(h; \widehat{M})\right| \leq \frac{1}{\lambda}\underbrace{\left|\sum_{s,a} h(s)(r(s,a) - \hat{r}(s,a))\right|}_{:=T_1} \\ + \frac{\gamma}{\lambda}\underbrace{\left|\sum_{s,s',a} h(s)\left(r(s',a)P(s',a,s) - \hat{r}(s',a)\hat{P}(s',a,s)\right)\right|}_{:=T_2} \\ + \frac{\gamma}{\lambda}\underbrace{\left|\sum_{s,a} h(s)\sum_{\widetilde{s}} h(\widetilde{s})\left(P(s,a,\widetilde{s}) - \hat{P}(s,a,\widetilde{s})\right)\right|}_{:=T_3} \\ + \frac{\gamma^2}{2\lambda}\underbrace{\left|\sum_{s,a}\sum_{\widetilde{s},\hat{s}} h(\widetilde{s})h(\hat{s})\left(P(s,a,\hat{s})P(s,a,\widetilde{s}) - \widehat{P}(s,a,\hat{s})\widehat{P}(s,a,\widetilde{s})\right)\right|}_{:=T_4}\end{aligned}$$

$$T_1 \leq \sqrt{\sum_s h(s)^2}\sqrt{\sum_s\left(\sum_a (r(s,a) - \hat{r}(s,a))\right)^2} \leq \|h\|_2\sqrt{A}\left\|r - \hat{r}\right\|_2$$

$$\begin{aligned}T_2 &\leq \left|\sum_{s,s',a} h(s)P(s',a,s)\left(r(s',a) - \hat{r}(s',a)\right)\right| + \left|\sum_{s,s',a} h(s)\hat{r}(s',a)\left(P(s',a,s) - \hat{P}(s',a,s)\right)\right| \\ &\leq \sqrt{\sum_s h(s)^2}\sqrt{\sum_s\left(\sum_{s',a} P(s',a,s)\left(r(s',a) - \hat{r}(s',a)\right)\right)^2} \\ &+ \sqrt{\sum_s h(s)^2}\sqrt{\sum_s\left(\sum_{s',a}\left(P(s',a,s) - \hat{P}(s',a,s)\right)\right)^2} \\ &\leq \|h\|_2\sqrt{SA}\sqrt{\sum_s\sum_{s',a} P(s',a,s)\left(r(s',a) - \hat{r}(s',a)\right)^2} + \|h\|_2\sqrt{SA}\sqrt{\sum_s\sum_{s',a}\left(P(s',a,s) - \hat{P}(s',a,s)\right)^2} \\ &\leq \|h\|_2\sqrt{SA}\left\|r - \hat{r}\right\|_2 + \|h\|_2\sqrt{SA}\left\|P - \hat{P}\right\|_2\end{aligned}$$

$$T_3 \le \|h\|_2 \sqrt{\sum_s \left(\sum_{\widetilde{s},a} h(\widetilde{s}) \left(P(s,a,\widetilde{s}) - \widehat{P}(s,a,\widetilde{s}) \right) \right)^2}$$
$$\le \|h\|_2 \sqrt{\sum_s \|h\|_2^2 \sum_{\widetilde{s}} \left(\sum_a \left(P(s,a,\widetilde{s}) - \widehat{P}(s,a,\widetilde{s}) \right) \right)^2} \le \|h\|_2^2 \sqrt{A} \left\| P - \widehat{P} \right\|_2$$

$$T_4 \le \left| \sum_{s,a} \sum_{\widetilde{s},\hat{s}} h(\widetilde{s}) h(\hat{s}) P(s,a,\hat{s}) \left(P(s,a,\widetilde{s}) - \widehat{P}(s,a,\widetilde{s}) \right) \right|$$
$$+ \left| \sum_{s,a} \sum_{\widetilde{s},\hat{s}} h(\widetilde{s}) h(\hat{s}) \widehat{P}(s,a,\hat{s}) \left(P(s,a,\hat{s}) - \widehat{P}(s,a,\hat{s}) \right) \right|$$
$$\le 2 \|h\|_2 \left| \sum_{s,a} \sum_{\widetilde{s}} h(\widetilde{s}) \left| P(s,a,\widetilde{s}) - \widehat{P}(s,a,\widetilde{s}) \right| \right| \quad [\text{By} \sum_{\hat{s}} h(\hat{s}) P(s,a,\hat{s}) \le \|h\|_2 \sqrt{\sum_{\hat{s}} P(s,a,\hat{s})} = \|h\|_2]$$
$$\le 2 \|h\|_2^2 \sqrt{\sum_{\widetilde{s}} \left(\sum_{s,a} \left| P(s,a,\widetilde{s}) - \widehat{P}(s,a,\widetilde{s}) \right| \right)^2}$$
$$\le 2 \|h\|_2^2 \sqrt{SA} \left\| P - \widehat{P} \right\|_2$$

Substituting the upper bounds on T_1, T_2, T_3, and T_4 into the upper bound on $\mathcal{L}_d(h; M) - \mathcal{L}_d(h; \widehat{M})$ gives the desired bound. □