LOCAL AND GLOBAL STABILITY IN PERFORMATIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In performative reinforcement learning the deployed policy shapes the environment that generates the learner's future data, and the natural solution concept is a *performatively stable* policy that is optimal in the environment it induces. Existing convergence guarantees rely on Lipschitz sensitivity assumptions on the environment map, which are hard to verify and fail in settings such as multi-agent best-response dynamics. We instead study stability for *mixtures* of policies, and show that the resulting picture is fundamentally different from performative prediction, where randomization removes the need for any sensitivity assumption. We distinguish *local* mixed stability, an occupancy-weighted first-order relaxation, from *global* mixed stability, which certifies against arbitrary deviating policies. The two notions genuinely differ: we exhibit an instance where local stability is achieved exactly but every mixture has global stability gap bounded away from zero. We show that a weighted per-state Hedge dynamic drives the local stability gap to zero at an $O(1/\sqrt{T})$ rate for an *arbitrary*, possibly discontinuous, environment map. For global stability we introduce a weaker bounded transition range (ϵ_P) assumption, under which unweighted per-state Hedge converges up to a floor of $O(\gamma\epsilon_P/(1-\gamma)^3)$. This floor is unavoidable as we prove a matching-in-ϵ_P lower bound under trajectory feedback. Finally, we extend both notions to n-player Markov games, obtaining local stability with no assumption on the environment map or game structure, and global stability for Markov potential games.

1 INTRODUCTION

The classic framework of reinforcement learning (Sutton & Barto, 1998) assumes that the environment is *static*, i.e. the reward and the transition probabilities do not change while the policy is being learned. However, in many applications the deployed policy shapes the very environment that generates the learner's future data. For example, a recommendation policy can change population behaviour through interactions (Jiang et al., 2019), and traffic-routing policy changes congestion patterns (Wu et al., 2021). Moreover, in multi-agent settings (Zhang et al., 2021), as the learner adapts its policy, the opponents and/or followers adjust their policies, often leading to arbitrary shift in the environment. Mandal et al. (2023) formalised this phenomenon as *performative reinforcement learning* and introduced *performative stability* as the natural fixed-point solution concept — a policy that remains optimal in the MDP it induces. Their result generalizes the framework of *performative prediction* from supervised learning (Perdomo et al., 2020), and shows that under sensitivity assumption on both the *reward* and *transition kernel* (i.e. the map $\pi \mapsto (P_\pi, r_\pi)$), a regularized variation of repeated optimization converges to a performatively stable policy.

Although repeated optimization can converge to a performatively stable policy, it requires strong sensitivity assumptions on the reward and transition kernel (Mandal et al., 2023; Mandal & Radanovic, 2025). These assumptions are hard to verify, and may not hold in certain applications. For example, in multi-agent settings, best response dynamics are often discontinuous and chaotic (Cheung & Piliouras, 2019), and if the learner changes its policy, the opponents may change their policies adversarially (Mohammadi et al., 2023). In the absence of sensitivity assumptions, a *deterministic* performatively stable policy may not exist, and a natural question is the choice of the right solution concept in this case. Fortunately, recent work in performative prediction (Farina & Perdomo, 2026) shows that if the learner is allowed to randomize over a mixture of policies, then stability can be achieved in expectation with no sensitivity assumption on the environment map.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This motivates us to study *mixed* stability in performative reinforcement learning through a mixture of policies in the absence of sensitivity assumptions. However, we discover that the notion of stability in performative reinforcement learning is fundamentally different than in supervised learning (Perdomo et al., 2020). In reinforcement learning, the environment consists of two components: a *reward* map ($\pi \mapsto r_\pi$) and a *transition* map ($\pi \mapsto P_\pi$). When both maps are performative, even a mixture of policies can fail to achieve the standard notion of *global stability*. The primary reason is that achieving sublinear regret in online reinforcement learning with *changing transition kernel* is both computationally hard (Abbasi-Yadkori et al., 2013) and information-theoretically hard with bandit feedback (Tian et al., 2021). Therefore, we introduce the notion of *local stability*, a first-order relaxation of global stability that is also equivalent to the notion of *stationarity* (Li & Wai, 2024), and study two questions:

(i) How much of the sensitivity assumption in Mandal et al. (2023) is actually needed to drive an online learning algorithm toward a performatively stable mixture?

(ii) Does the answer depend on which notion of stability we target: *local* stability against first-order deviations, or *global* stability against arbitrary deviations?

We find that the mixture of policies generated by an online algorithm can converge to performative local stability *unconditionally*. However, in contrast to performative prediction (Farina & Perdomo, 2026), unconditional convergence to global stability is not possible even with mixtures. We then show that under the *bounded transition range*, an assumption weaker than sensitivity of transition map, mixtures generated by an online algorithm can converge to performative global stability. Finally, we extend both notions of stability to n-player performative Markov games, and show that local stability can be achieved unconditionally, while global stability can be achieved in performative Markov potential games. We make no assumption about the sensitivity of the reward map.

1.1 Overview of the main contributions

(a) **Local stability needs no assumption.** We propose a weighted per-state Hedge algorithm that achieves arbitrary local stability at an $O(1/\sqrt{T})$ rate without any assumption on the environment map. We then consider a setting where the learner only accesses the induced environment through sampled trajectories and give a finite-sample version of our algorithm with sample complexity $\tilde{O}\left(\frac{SA}{(1-\gamma)^4\epsilon^2}\right)$ to achieve ϵ-local stability.

(b) **A strict separation between local and global stability.** We exhibit a *survival-chain instance* on which local stability is exactly achievable, yet every mixture of policies has global stability gap bounded away from 0. This establishes a strict separation between local and global notion of stability in performative reinforcement learning.

(c) **Global stability under bounded transition range.** We introduce the assumption of *Bounded Transition Range*, an assumption strictly weaker than the sensitivity of transition map (Mandal et al., 2023). Under this assumption, we propose an unweighted per-state Hedge algorithm that attains global stability up to a floor of $O(\gamma\epsilon_P/(1-\gamma)^3)$, along with a finite-sample counterpart with similar sample complexity as part (a). We also show that under a local trajectory-feedback protocol, no algorithm can attain global stability gap better than $\Omega(\gamma\epsilon_P/(1-\gamma))$.

(d) **Extension to performative Markov games.** We generalise both notions of stability to n-player performative Markov games. An independent weighted per-state Hedge algorithm attains local stability with *no* assumption on the joint sensitivity of the environment map or on the game's structure, improving on multiplayer performative prediction results that require such conditions (Narang et al., 2023). Finally, we show that we can obtain global stability for performative Markov *potential* games using independent projected policy gradient ascent by the agents.

1.2 Related work

Performative prediction and its stability. Perdomo et al. (2020) first formalised the general phenomenon as *performative prediction*: the deployed model shifts the data-generating distribution, and the goal is to obtain either a performatively stable or optimal solution. Repeated optimization (Perdomo et al., 2020) and stochastic gradient based lazy-deployment methods (Mendler-Dünner et al., 2020; Drusvyatskiy & Xiao, 2023) converge to such a stable point with a rate governed by the sensitivity of the performative map. In parallel, Miller et al. (2021) study performative risk, and obtain

Setting	Assumptions	Speed of convergence	Convergence floor
Local stability	None	$O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{S\ln A}{T}}\right)$ (Theorem 1)	0
Global stability	Bounded transition range (Assn. 1)	$O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{\ln(SA)}{T}}\right)$ (Theorem 3)	$\Theta\left(\frac{\gamma\,\epsilon_P}{(1-\gamma)^3}\right)$
Multi-agent local stability	None	$O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{S\ln A_{\max}}{T}}\right)$ (Theorem 4)	0
Multi-agent global stability	Markov potential game (Assn. 2, 3)	$O\left(\kappa\sqrt{\frac{Sn}{T(1-\gamma)}}\right)$ (Theorem 5)	$\Theta\left(\kappa\sqrt{Sn}\,\epsilon_L\right)$

Table 1: **Summary of our stability guarantees**. Here ϵ_P is the bounded-transition-range parameter of Assumption 1, ϵ_L is the potential-sensitivity parameter of Assumption 3, $A_{\max} := \max_i A_i$, and κ is the distribution mismatch coefficient of Theorem 5. "Convergence floor" is the residual stability gap that persists as $T \to \infty$. "None" means the guarantee holds for an arbitrary environment map.

convergence to performatively optimal solutions under parametric assumptions on the distribution map. Brown et al. (2022) and Izzo et al. (2022) extend the framework to a *stateful* world where the population's response depends on the history of deployed models rather than the current one.

Our approach is closest to Farina & Perdomo (2026), who show that online learning over *mixtures* of models attains performative stability at an $O(1/\sqrt{T})$ rate with no sensitivity assumptions. Our local stability results (Section 3) transport this mixture-based, assumption-free approach to the reinforcement learning setting. Li & Wai (2024) and Hikima & Takeda (2025) study stationary points of nonconvex performative risk under a decision dependent, distributionally robust formulation. Lemma 2 shows that their stationary stability notion is equivalent to our local stability notion. Narang et al. (2023) and Piliouras & Yu (2023) extend performative prediction to multiplayer games, but their convergence guarantees require a joint sensitivity conditions together with convexity of the game. Our Theorem 4 shows that, once the target is local stability, none of these conditions are necessary.

Performative reinforcement learning. Mandal et al. (2023) introduced the framework of *performative reinforcement learning* and gave an algorithm that converges to a pure stable policy under an ϵ-sensitivity (Lipschitzness) assumption on the environment map. Assumption 1 is strictly weaker than this condition as it bounds only the *range*, and not the modulus of continuity of the induced kernels. Rank et al. (2024) relax the *instantaneous* nature of the environment map, allowing the environment to adjust gradually across deployments, and Pollatos et al. (2025) study robustness of performative stability to corrupted feedback. Both directions are complementary to ours as they consider more general feedback or deployment protocol. Chen & Huang (2025) and Basu et al. (2025) pursue policy-gradient-style algorithms for *performatively optimal* policy — one that is optimal in expectation over the induced MDP. Convergence to performatively optimal policy requires different regularity conditions e.g. regularizer dominance or softmax parametrization. Sahitaj et al. (2025) initiated the study of independent learning in performative Markov *potential* games, which is the setting of Section 5.3. We complement their analysis with an unconditional local stability guarantee that requires no potential-game structure at all (Theorem 4). Additional Related work including potential games, policy optimisation, non-stationary RL are discussed in Appendix A.

2 Setting and Notation

MDPs. Fix a finite state space $\mathcal{S}$ with $|\mathcal{S}| = S$, a finite action space $\mathcal{A}$ with $|\mathcal{A}| = A$, a discount factor $\gamma \in [0, 1)$ and an initial distribution $\rho \in \Delta(\mathcal{S})$. An MDP is a tuple $M = (P, \mathcal{S}, \mathcal{A}, r, \rho)$ with transition kernel $P(\cdot \mid s, a) \in \Delta(\mathcal{S})$ and reward $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Let $\Pi := \Delta(\mathcal{A})^{\mathcal{S}}$ denote the set of stationary (possibly stochastic) policies. For $\pi \in \Pi$ and a kernel P, the normalised discounted occupancy measure is

$$d^{\pi}_{P}(s, a) \;=\; (1-\gamma)\sum_{h\geq 0}\gamma^{h}\,\Pr\left[s_h = s,\, a_h = a \,\middle|\, s_0 \sim \rho,\, \pi,\, P\right],$$

with state marginal $d_P^\pi(s) = \sum_a d_P^\pi(s,a)$. It is easy to verify that $\sum_{s,a} d_P^\pi(s,a) = 1$. The value function of the policy π in the MDP M is defined as

$$V_M(\pi) = \mathbb{E}_{\pi,P}\Big[\sum_{h\geq 0} \gamma^h \cdot r(s_h,a_h) \,\Big|\, s_0 \sim \rho\Big] = \frac{1}{1-\gamma}\,\langle r, d_P^\pi\rangle\,.$$

The state-action function of policy π at a tuple (s,a) is defined as

$$Q_M(s,a) = \mathbb{E}_{\pi,P}\Big[\sum_{h\geq 0} \gamma^h \cdot r(s_h,a_h) \,\Big|\, s_0 = s, a_0 = a\Big].$$

Performative RL. We recall the framework of performative reinforcement learning from Mandal et al. (2023). There is no longer a single MDP, rather the MDPs are parametrized by the policies in a class Π. If a policy π is deployed, then the underlying environment changes to an MDP $M(\pi)$ with transition kernel P_π and reward function r_π. The learner then observes the data generated by π in $M(\pi)$. Compared to Mandal et al. (2023), we do not make any assumption about Lipschitzness of the environment map $\pi \mapsto M(\pi) = (P_\pi, r_\pi)$ unless stated explicitly.

Following Farina & Perdomo (2026), we work with mixtures. A mixture $\mu \in \Delta(\Pi)$ is deployed by drawing $\pi \sim \mu$ independently at the start of each episode and running π. The environment then responds to the *realised* π. Here we assume that the environment responds to the realised policy, and not to the mixture μ as an aggregate object. If instead the environment reacts to μ the environment map should be defined on $\Delta(\Pi)$ which we do not consider here.

Shorthand. We will primarily consider online learning algorithm which deploy a sequence of policies. For a sequence of deployed policies $\pi_1, \ldots, \pi_T$ we write

$$M_t := M(\pi_t) = (P_t, r_t), \qquad Q_t(s,a) := Q_{M_t}^{\pi_t}(s,a), \qquad w_t(s) := d_{P_t}^{\pi_t}(s),$$

and for a comparator policy $\pi' \in \Pi$,

$$\Delta_t(s,\pi') \;:=\; \big\langle \pi'(\cdot \mid s) - \pi_t(\cdot \mid s),\; Q_t(s,\cdot)\big\rangle. \tag{1}$$

More generally, for any $\pi \in \Pi$ we write $\Delta_\pi(s,\pi') := \langle \pi'(\cdot \mid s) - \pi(\cdot \mid s),\, Q_{M(\pi)}^\pi(s,\cdot)\rangle$.

The following classical result will be useful to motivate our solution concepts.

Lemma 1 (Performance difference; Kakade & Langford (2002))**.** *For any MDP $M = (P,r)$ and any $\pi, \pi' \in \Pi$,*

$$V_M(\pi') - V_M(\pi) \;=\; \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d_P^{\pi'}(s)\,\big\langle \pi'(\cdot \mid s) - \pi(\cdot \mid s),\; Q_M^\pi(s,\cdot)\big\rangle.$$

2.1 Two solution concepts

Mandal et al. (2023) first introduced the notion of *performative stability* in reinforcement learning. A policy π is performatively stable if $\pi \in \arg\max_{\pi'} V_{M(\pi)}(\pi')$. This means that the policy π is optimal in the induced MDP $M(\pi)$. We now extend this definition to a mixture of policies.

Definition 1 (Global mixed performative stability)**.** For $\mu \in \Delta(\Pi)$ define

$$\mathrm{Gap}^{\mathrm{glob}}(\mu) \;:=\; \max_{\pi'\in\Pi} \mathbb{E}_{\pi\sim\mu}\big[V_{M(\pi)}(\pi')\big] \;-\; \mathbb{E}_{\pi\sim\mu}\big[V_{M(\pi)}(\pi)\big].$$

We say μ is ϵ-performatively stable if $\mathrm{Gap}^{\mathrm{glob}}(\mu) \leq \epsilon$.

This is the direct analogue of the mixed stability notion of Farina & Perdomo (2026). The comparator π' is evaluated inside each induced MDP $M(\pi)$ but doesn't receive its own performative effect. When μ is a point mass, the definition reduces to the notion of ϵ-performatively stable policy (Mandal et al., 2023).

Definition 2 (Local mixed performative stability)**.** For $\mu \in \Delta(\Pi)$ define

$$\mathrm{Gap}^{\mathrm{loc}}(\mu) \;:=\; \frac{1}{1-\gamma}\,\max_{\pi'\in\Pi}\,\mathbb{E}_{\pi\sim\mu}\left[\sum_{s\in\mathcal{S}} d_{P_\pi}^\pi(s)\,\Delta_\pi(s,\pi')\right].$$

We say μ is ϵ-*locally* performatively stable if $\mathrm{Gap}^{\mathrm{loc}}(\mu) \leq \epsilon$.

The above definition is different than Definition 1. Using Lemma 1, we can rewrite Definition 1 as

$$\mathrm{Gap}^{\mathrm{glob}}(\mu) = \frac{1}{1-\gamma} \max_{\pi' \in \Pi} \mathbb{E}_{\pi \sim \mu} \left[\sum_{s \in \mathcal{S}} d^{\pi'}_{P_\pi}(s) \Delta_\pi(s, \pi') \right].$$

The only difference from Definition 1 is that the per-state advantage terms are weighted by the occupancy of the *deployed* policy $d^{\pi}_{P_\pi}$ rather than that of the comparator, $d^{\pi'}_{P_\pi}$. This comparison also highlights that we include the factor $\frac{1}{1-\gamma}$ in Definition 2 to ensure that $\mathrm{Gap}^{\mathrm{loc}}$ and $\mathrm{Gap}^{\mathrm{glob}}$ are of the same scale. Definition 2 is closely related to the notion of stationary performative stability introduced by Li & Wai (2024) in the context of non-convex loss functions. Adapted to our *constrained* setting ($\pi \in \Pi$), the definition reads as follows.

Definition 3 (Stationary performative stability)**.** A policy π is *stationary performatively stable* if $\max_{\pi' \in \Pi} \left\langle \nabla_{\pi'} V_{M(\pi)}(\pi')|_{\pi'=\pi}, \pi' - \pi \right\rangle \leq \delta$, i.e. if π is a δ-approximate first-order stationary point of the MDP it induces.

Lemma 2. *A policy π is δ-stationary performatively stable iff* $\mathrm{Gap}^{\mathrm{loc}}(\pi) \leq \delta$.

Lemma 2 establishes that for distribution with point mass, the notion of stationary performative stability and local performative stability. Therefore, Definition 2 can be seen as a generalization of local performative stability Li & Wai (2024) to mixtures over policies.

2.2 The Separating Instance

There is a strict separation between global and local mixed performative stability. A mixture of policies $\mu \in \Delta(\Pi)$ can be locally mixed performatively stable i.e. $\mathrm{Gap}^{\mathrm{loc}}(\mu) = 0$, however, it need not be globally mixed performatively stable i.e. $\mathrm{Gap}^{\mathrm{glob}}(\mu) > c$ for some universal constant $c > 0$. We consider the instance shown in Figure 1.

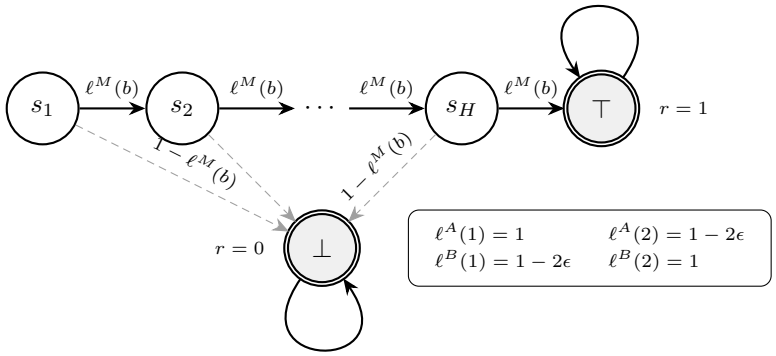

Figure 1: The survival chain of Definition 4 with H stages. From each s_h the process either advances one step, with the continuation probability $\ell^M(b)$ of the action b taken, or drops to the absorbing failure state $\perp$; the reward is 0 on the chain. The only reward is 1 in $\top$, so a policy's value is $\kappa = \gamma^H/(1-\gamma)$ times its probability of surviving all H stages. The environment map $M(\cdot)$ of Definition 5 assigns to each policy whichever model it survives less well in.

Definition 4 (Survival chain)**.** Fix $H \geq 1$ and $\epsilon \in (0, 1/2]$. The state space is $\{s_1, \ldots, s_H\} \cup \{\top, \perp\}$ with $\rho = \delta_{s_1}$, and $\mathcal{A} = \{1, 2\}$. The states $\top$ and $\perp$ are absorbing with $r(\top, \cdot) = 1$ and $r(\perp, \cdot) = 0$. The reward is 0 on the chain and is *the same in every model*. In MDP M, the process moves from s_h to s_{h+1} under action b (to $\top$ if $h = H$) with probability $\ell^M(b)$ and to $\perp$ otherwise, where

$$\ell^A(1) = 1, \quad \ell^A(2) = 1 - 2\epsilon; \qquad \ell^B(1) = 1 - 2\epsilon, \quad \ell^B(2) = 1.$$

Write M^A, M^B for the two resulting MDPs.

A policy is described by $y_h := \pi(1|s_h) \in [0, 1]$, and let us write $\kappa := \sum_{h \geq H} \gamma^h = \gamma^H/(1-\gamma)$. Then one can verify that

$$V_{M^A}(\pi) = \kappa\, F_A(y),\ F_A(y) := \textstyle\prod_{h=1}^{H} \big(1 - 2\epsilon(1 - y_h)\big),\ V_{M^B}(\pi) = \kappa\, F_B(y),\ F_B(y) = \textstyle\prod_{h=1}^{H} \big(1 - 2\epsilon\, y_h\big) \tag{2}$$

since the value is κ times the probability of surviving all H transitions.

Definition 5 (The environment map and the policy class)**.** The environment map $M(\pi) := M^A$ if $F_A(y) \leq F_B(y)$ and $M(\pi) := M^B$ otherwise. The environment selects the model in which the deployed policy fares worse.

Lemma 3. *For the instance described in Definitions 4 and 5, we have* $\inf_{\mu\in\Delta(\Pi)} \mathrm{Gap}^{\mathrm{loc}}(\mu) = 0$.

Lemma 4. *For any* $\epsilon \in [0, 1/2]$ *and* $H \geq 2$*, and for the instance described in Definitions 4 and 5,*

$$\inf_{\mu\in\Delta(\Pi)} \mathrm{Gap}^{\mathrm{glob}}(\mu) \geq \frac{1}{16}\frac{\gamma^H}{1-\gamma}\min\left\{1, \epsilon^2 H^2\right\}.$$

The above two results show a strict separation between local and global mixed performative stability. It is possible to achieve exact local mixed performative stability, however, no matter which mixture is chosen the environment defined in Definitions 4 and 5 do not admit a global mixed performatively stable mixture beyond a constant threshold. We note that, when constructing the chain instance, we need $H \geq 2$. This is because for $H = 1$, the value function is linear in the policy parameter and the notion of local and global performative stability coincide.

3 Local Stability without any Assumption

We now propose a simple algorithm that outputs a local performatively stable mixture of policies. The algorithm is based on per-state hedge algorithm. The policy for each state is updated using the Hedge algorithm with time-vayring state, action functions. In particular, at round t the learner deploys π_t, the environment changes to $M_t = M(\pi_t)$, and the learner observes the action-value vectors $Q_t(s, \cdot)$ for all s, and the occupancy measure of the deployed policy i.e. $w_t(s) = d_{P_t}^{\pi_t}(s)$ for all s. Initially, we assume that the vectors $Q_t(s, \cdot)$ and $w_t(\cdot)$ are directly available to the algorithm, and later we will see how to estimate them from samples.

Definition 6 (Weighted per-state Hedge)**.** Initialise $\pi_1(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$ for all s. For $t = 1, \ldots, T$: deploy π_t, observe Q_t and w_t, and set

$$\pi_{t+1}(a \mid s) \;\propto\; \pi_t(a \mid s)\,\exp\big(\eta\, w_t(s)\, Q_t(s, a)\big), \qquad s \in \mathcal{S},\ a \in \mathcal{A}. \tag{3}$$

The occupancy weight inside the exponent is important, as it aligns the algorithm's regret with the occupancy-weighted objective of Definition 2.

Theorem 1 (Unconditional local stability)**.** *Let* $M(\cdot)$ *be an* arbitrary *environment map. Let* $\pi_1, \ldots, \pi_T$ *be generated by equation 3 and let* $\hat{\mu} := \mathrm{Unif}\{\pi_1, \ldots, \pi_T\}$*. Then for every* $\eta > 0$*,*

$$\mathrm{Gap}^{\mathrm{loc}}(\hat{\mu}) \;\leq\; \frac{1}{1-\gamma}\left(\frac{S \ln A}{\eta T} + \frac{\eta}{8(1-\gamma)^2}\right),$$

and in particular, choosing $\eta = 2(1-\gamma)\sqrt{2S \ln A/T}$*, we have,* $\mathrm{Gap}^{\mathrm{loc}}(\hat{\mu}) \;\leq\; \frac{1}{(1-\gamma)^2}\sqrt{\frac{S \ln A}{2T}}$.

3.1 A finite-sample algorithm

Algorithm 6 assumes exact access to Q_t and w_t. We now replace both by estimates built from trajectories collected after the round-t deployment. In particular, we will assume that once π_t has been deployed and the environment $M_t = M(\pi_t)$ is realised, the learner may draw independent trajectories from M_t, starting from either the distribution ρ or at a chosen pair (s, a), and following π_t thereafter. We will assume access to an *occupancy sample* and a *value sample*.

Definition 7 (Occupancy and value samples)**.** Let $\tau \sim \mathrm{Geom}(1-\gamma)$ i.e. $\Pr[\tau = j] = (1-\gamma)\gamma^j$.

1. An *occupancy sample* draws τ, runs π_t from ρ in M_t for τ steps and returns the state s_τ.
2. A *value sample at* (s, a) draws τ, plays a at s and follows π_t thereafter, and returns $\frac{1}{1-\gamma} r_\tau$, where r_τ is the reward collected from the-τ step of that rollout.

Algorithm 1 estimates $w_t(s)$ and $Q_t(s, \cdot)$ from independent samples and uses them to execute the update step defined in eq. (3).

Theorem 2. *In order to reach* $\mathrm{Gap}^{\mathrm{loc}}(\widehat{\mu}) \leq \epsilon$ *with probability at least* $1-\delta$ *it suffices for Algorithm 1 to take* $T = O\left(\frac{S \ln(SA/\delta)}{(1-\gamma)^4\epsilon^2}\right)$ *iterations and* $O\left(\frac{SA \ln(SA/\delta)}{(1-\gamma)^4\epsilon^2}\right)$ *trajectories in total.*

Algorithm 1 Finite-sample weighted per-state Hedge

Input: Batch sizes $n, k \geq 1$ and a step size $\eta > 0$.

1: Initialize $\pi_1(\cdot|s) = \mathrm{Unif}(\mathcal{A})$ for each $s \in \mathcal{S}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Deploy π_t.
4: *Occupancy batch.* Draw n independent occupancy samples and let $\widehat{w}_t$ be their empirical distribution on $\mathcal{S}$.
5: *Value batch.* For each s with $\widehat{w}_t(s) > 0$ and each $a \in \mathcal{A}$, draw k independent value samples at (s, a) and let $\widehat{Q}_t(s, a)$ be their average. Set $\widehat{Q}_t(s, \cdot) := 0$ when $\widehat{w}_t(s) = 0$.
6: Update, with $\widehat{g}_t(s, a) := \widehat{w}_t(s)\widehat{Q}_t(s, a)$,

$$\pi_{t+1}(a \mid s) \;\propto\; \pi_t(a \mid s) \exp\big(\eta\, \widehat{g}_t(s, a)\big). \tag{4}$$

7: **end for**

Output: Unif $(\pi_1, \ldots, \pi_T)$.

4 Global Stability Under Bounded Transition Range

We now consider the problem of generating a global performatively stable mixture of policies. We will make the following assumption.

Assumption 1 (Bounded transition range)**.** There exists a transition kernel $\bar{P}$ and $\epsilon_P \geq 0$ e.g.

$$\max_{s,a} \big\|P_\pi(\cdot \mid s, a) - \bar{P}(\cdot \mid s, a)\big\|_1 \;\leq\; \epsilon_P \qquad \text{for all } \pi \in \Pi.$$

Assumption 1 constrains only the *range* of the induced kernels, and not the regularity of $\pi \mapsto P_\pi$: the map may be arbitrarily discontinuous within that range. It is implied by, but strictly weaker than, the ϵ-sensitivity condition used in the performative RL literature (Assumption 1, Mandal et al. (2023)). In particular, if $\sup_{\pi,\pi'} \max_{s,a} \|P_\pi(\cdot \mid s, a) - P_{\pi'}(\cdot \mid s, a)\|_1 \leq \tilde{\epsilon}\, \|\pi - \pi'\|_1$ then Assumption 1 holds with $\bar{P} := P_{\pi_1}$ and $\epsilon_P = 2\tilde{\epsilon}S$. Note that, we place no assumption on the reward map $\pi \mapsto r_\pi$.

Definition 8 (Unweighted per-state Hedge)**.** Initialise $\pi_1(\cdot|s) = \mathrm{Unif}(\mathcal{A})$. For $t = 1, \ldots, T$, deploy π_t, observe Q_t, and set

$$\pi_{t+1}(a \mid s) \;\propto\; \pi_t(a \mid s) \exp\big(\eta\, Q_t(s, a)\big). \tag{5}$$

Theorem 3 (Global stability up to the transition range)**.** *Suppose Assumption 1 holds. Let $\pi_1, \ldots, \pi_T$ be generated by equation 5 with $\eta = 2(1-\gamma)\sqrt{2 \ln A / T}$, and let $\hat{\mu} := \mathrm{Unif}\{\pi_1, \ldots, \pi_T\}$. Then*

$$\mathrm{Gap}^{\mathrm{glob}}(\hat{\mu}) \;\leq\; \frac{1}{(1-\gamma)^2}\sqrt{\frac{\ln A}{2T}} \;+\; \frac{\gamma\, \epsilon_P}{(1-\gamma)^3}.$$

Note that, in contrast to Theorem 1, Theorem 3 can guarantee convergence to a globally performatively stable mixture up to a floor of $\Theta(\gamma\epsilon_P/(1-\gamma)^3)$. In Appendix B we construct a collection of $K \geq 20$ instances such that for any algorithm $\mathcal{A}$ that samples at most $N \leq O(K/\epsilon_P^2)$ trajectories, there exist at least one instance $j \in [K]$ such that

$$\mathbb{E}\big[\mathrm{Gap}^{\mathrm{glob}}(\hat{\mu})\big] \geq \frac{\gamma\epsilon_P}{8(1-\gamma)}.$$

This implies that the result of Appendix 3 is optimal up to a factor of $(1-\gamma)^{-2}$. Moreover, similar to Algorithm 1, we can develop a finite-sample algorithm for global stability. The sample complexity is similar to Theorem 2 and the full details are provided in Appendix C.

5 Multi-Agent Setting

We now consider a multi-agent generalization of the framework of performative reinforcement learning (Sahitaj et al., 2025). There are n players holding policies $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^n)$ and the environment responds to the whole profile of policies. In this section, we aim to generalize the results of Theorem 1 (unconditional local stability) and Theorem 3 (conditional global stability) from single-agent to multi-agent settings. Let us first define the notion of performative Markov game.

5.1 Performative Markov games

Let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be finite action sets, $A_i := |\mathcal{A}_i|$, $\mathcal{A} := \prod_i \mathcal{A}_i$, and $\Pi^i := \Delta(\mathcal{A}_i)^{\mathcal{S}}$, $\mathbf{\Pi} := \prod_i \Pi^i$. A performative Markov game is an arbitrary map

$$M : \mathbf{\Pi} \longrightarrow \big\{(P, r^1, \dots, r^n)\big\}, \qquad \boldsymbol{\pi} \longmapsto M(\boldsymbol{\pi}) = \big(P_{\boldsymbol{\pi}},\, r^1_{\boldsymbol{\pi}}, \dots, r^n_{\boldsymbol{\pi}}\big),$$

with $P(\cdot \mid s, \boldsymbol{a}) \in \Delta(\mathcal{S})$ and $r^i : \mathcal{S} \times \mathcal{A} \to [0,1]$. Therefore, a policy profile $\boldsymbol{\pi}$ of the n players results in a Markov game with transition $P_{\boldsymbol{\pi}}$ and utility profile $(r^1_{\boldsymbol{\pi}}, \dots, r^n_{\boldsymbol{\pi}})$. The occupancy measure $d^{\boldsymbol{\pi}}_P$ is defined as before with respect to the joint action $\boldsymbol{a} \in \mathcal{A}$. Note that, there is a single state occupancy, shared by all players.

We will write $Q^i_M(s, \boldsymbol{a})$ for the action-value of player i in the Markov game M at state s and joint action $\boldsymbol{a}$. We can define the marginial action-value of player i against the others' current play as

$$Q^{i,\boldsymbol{\pi}}_M(s, a^i) \;:=\; \mathbb{E}_{a^{-i} \sim \boldsymbol{\pi}^{-i}(\cdot|s)} \big[Q^i_M\big(s, (a^i, a^{-i})\big) \big].$$

Additionally, we will write $V^i_M(\pi^i, \boldsymbol{\pi}^{-i})$ for the value of player i in the Markov game M when it plays π^i and the other players play $\boldsymbol{\pi}^{-i}$. We can now extend the definitions of global and local performative stability to the multi-agent setting. Note that, the mixtures $\mu \in \Delta(\mathbf{\Pi})$ are over *joint profiles* and are therefore correlated across players.

Definition 9 (Multi-Agent Mixed performative stability)**.** For $\mu \in \Delta(\mathbf{\Pi})$ set

$$\mathrm{Gap}^{\mathrm{glob}}_i(\mu) := \max_{\pi' \in \Pi^i} \mathbb{E}_{\boldsymbol{\pi} \sim \mu} \left[V^i_{M(\boldsymbol{\pi})}\big(\pi', \boldsymbol{\pi}^{-i}\big) \right] - \mathbb{E}_{\boldsymbol{\pi} \sim \mu} \left[V^i_{M(\boldsymbol{\pi})}(\boldsymbol{\pi}) \right],$$

$$\mathrm{Gap}^{\mathrm{loc}}_i(\mu) := \frac{1}{1-\gamma} \max_{\pi' \in \Pi^i} \mathbb{E}_{\boldsymbol{\pi} \sim \mu} \sum_s d^{\boldsymbol{\pi}}_{P_{\boldsymbol{\pi}}}(s) \left\langle \pi'(\cdot \mid s) - \pi^i(\cdot \mid s),\; Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s, \cdot) \right\rangle,$$

and $\mathrm{Gap}^{\mathrm{glob}}(\mu) := \max_i \mathrm{Gap}^{\mathrm{glob}}_i(\mu)$, $\mathrm{Gap}^{\mathrm{loc}}(\mu) := \max_i \mathrm{Gap}^{\mathrm{loc}}_i(\mu)$.

As in the single-agent case the deviating policy π' is scored inside the induced environment $M(\boldsymbol{\pi})$ and does not receive its own performative effect.

5.2 Multi-Agent Local Stability without any assumption

Definition 10 (Independent weighted per-state Hedge)**.** Player i initialises $\pi^i_1(\cdot|s) = \mathrm{Unif}(\mathcal{A}_i)$. At round t the profile $\boldsymbol{\pi}_t = (\pi^1_t, \dots, \pi^n_t)$ is deployed; each player observes its own $Q^i_t(s, \cdot) := Q^{i,\boldsymbol{\pi}_t}_{M_t}(s, \cdot)$ and the common occupancy measure $w_t(s) := d^{\boldsymbol{\pi}_t}_{P_t}(s)$, and updates

$$\pi^i_{t+1}(a|s) \;\propto\; \pi^i_t(a|s) \exp\big(\eta_i\, w_t(s)\, Q^i_t(s, a)\big). \tag{6}$$

Theorem 4 (Unconditional local stability, n players)**.** *Let $M(\cdot)$ be an arbitrary performative Markov game. Let $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_T$ be generated by equation 6 with $\eta_i = 2(1-\gamma)\sqrt{2S \ln A_i / T}$ and let $\hat{\mu} := \mathrm{Unif}\{\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_T\}$. Then for every player i,*

$$\mathrm{Gap}^{\mathrm{loc}}_i(\hat{\mu}) \;\le\; \frac{1}{(1-\gamma)^2} \sqrt{\frac{S \ln A_i}{2T}}, \;\textit{and}\; \mathrm{Gap}^{\mathrm{loc}}(\hat{\mu}) \;\le\; \frac{1}{(1-\gamma)^2} \sqrt{\frac{S \ln A_{\max}}{2T}}$$

with $A_{\max} := \max_i A_i$.

The bound in Theorem 4 is unconditional, and holds for any performative Markov game. Note that, convergence in multiplayer performative prediction Narang et al. (2023) requires conditions about the joint sensitivity of the environment map, and also about the structure of the game (e.g. convexity). In contrast, Theorem 4 avoids such requirements by considering a mixture over strategy profiles. Moreover, the bound depends on $A_{\max}$, not on the joint action space $|\mathcal{A}| = \prod_i A_i$. The reason is that player i's deviation and gains are already marginalised over $\boldsymbol{\pi}^{-i}_t$.

5.3 Global Stability for Markov Potential Games

The situation for the global notion of stability is qualitatively different than in single agent setting. Proof of Theorem 3 uses the argument that comparator's occupancy $d^{\pi'}_P$ is a *fixed* weight vector, so

that a uniform per-state regret bound could be integrated against it. However, in a Markov game, player i's deviation produces the profile $(\pi', \boldsymbol{\pi}_t^{-i})$, whose occupancy $d_{\hat{P}}^{(\pi', \boldsymbol{\pi}_t^{-i})}$ depends on t through the opponents' policies, however small ϵ_P is. Therefore, the weights are time-varying even at $\epsilon_P = 0$, and the proof of Theorem 3 cannot be generalized. In fact, in Appendix D we construct a simple two-player Markov game that admits distribution μ with

$$\mathrm{Gap}^{\mathrm{loc}}(\mu) \;=\; 0 \qquad \text{and} \qquad \mathrm{Gap}^{\mathrm{glob}}(\mu) \;\geq\; \tfrac{\gamma^H}{16(1-\gamma)} \min\left\{1, \epsilon^2 H^2\right\} \;\forall H \geq 2.$$

Assumption 2 (Performative Markov potential game)**.** For every profile $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ the game $M(\boldsymbol{\pi})$ admits a potential $\Phi_{M(\boldsymbol{\pi})} : \boldsymbol{\Pi} \to [0, \frac{1}{1-\gamma}]$, such that for all i, all $\sigma, \sigma' \in \Pi^i$ and all $\boldsymbol{\tau}^{-i}$,

$$V^i_{M(\boldsymbol{\pi})}(\sigma', \boldsymbol{\tau}^{-i}) - V^i_{M(\boldsymbol{\pi})}(\sigma, \boldsymbol{\tau}^{-i}) \;=\; \Phi_{M(\boldsymbol{\pi})}(\sigma', \boldsymbol{\tau}^{-i}) - \Phi_{M(\boldsymbol{\pi})}(\sigma, \boldsymbol{\tau}^{-i}).$$

We will make the following assumptions about the potential function $\Phi_{M(\boldsymbol{\pi})}$.

Assumption 3 (Regularity)**.** Let $\boldsymbol{\Pi} = \prod_i \Delta(\mathcal{A}_i)^{\mathcal{S}}$ be the joint policy space, and let $G := \sup_{\boldsymbol{\pi}, \boldsymbol{\sigma}} \left\|\nabla_{\sigma^i} \Phi_{M(\boldsymbol{\pi})}(\boldsymbol{\sigma})\right\|_2$ be the maximum norm of partial derivatives.

(P1) *Smoothness.* Each $\Phi_{M(\boldsymbol{\pi})}(\cdot)$ is L_Φ-smooth on $\boldsymbol{\Pi}$.

(P2) *Sensitivity.* $\left|\Phi_{M(\boldsymbol{\pi})}(\boldsymbol{\sigma}) - \Phi_{M(\boldsymbol{\pi}')}(\boldsymbol{\sigma})\right| \leq \epsilon_L \left\|\boldsymbol{\pi} - \boldsymbol{\pi}'\right\|_2$ for all $\boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\pi}'$.

The smoothness assumption (P1) is standard in the analysis of gradient-based methods, and the sensitivity assumption (P2) is a natural generalization of Assumption 1 to the multi-agent setting. We now consider an independent projected policy gradient ascent algorithm.

Definition 11 (Independent projected policy gradient ascent)**.** For each $t \in \{1, \ldots, T\}$ let $\Phi_t(\cdot) := \Phi_{M(\boldsymbol{\pi}_t)}(\cdot)$ and $\alpha \in (0, 1/L_\Phi]$. For each agent i, update

$$\pi^i_{t+1} \;:=\; \mathrm{proj}_{\boldsymbol{\Pi}^i}\left(\pi_t + \alpha \nabla_{\pi^i} \Phi_t(\boldsymbol{\pi}_t)\right). \tag{7}$$

Theorem 5. *Suppose Assumptions 2 and 3 hold. Let $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T$ be generated by Equation (7) with $\alpha \leq 1/L_\phi$, and $\widehat{\mu} := \mathit{Unif}\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T\}$. Then,*

$$\mathrm{Gap}^{\mathrm{glob}}(\widehat{\mu}) \leq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \max_{i, \sigma \in \boldsymbol{\Pi}} \left\| \frac{d^{(\sigma, \boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})}}{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}} \right\|_\infty \left(G + \frac{\sqrt{2S}}{\alpha}\right) \sqrt{n} \left(2\alpha\epsilon_L + \sqrt{\frac{2\alpha}{T(1-\gamma)}}\right).$$

Furthermore, choosing $\alpha \leq \min\left\{1/L_\phi, \sqrt{2S}/G\right\}$ and $T \geq \frac{1}{(1-\gamma)\alpha\epsilon_L^2}$ we obtain

$$\mathrm{Gap}^{\mathrm{glob}}(\widehat{\mu}) \leq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \max_{i, \sigma \in \boldsymbol{\Pi}} \left\| d^{(\sigma, \boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})} / d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})} \right\|_\infty 8\sqrt{2Sn}\epsilon_L.$$

Theorem 5 establishes that if the number of iterations $T = O(\epsilon_L^{-2})$ then it is possible to attain global stability gap of $O(\kappa\sqrt{Sn}\epsilon_L)$, where $\kappa := \|d^{(\sigma, \boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})} / d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}\|_\infty$ is the distribution mismatch coefficient. Compared to Theorem 4, the bound does not depend on the number of actions but grows as square root of the number of agents.

6 Conclusion and Future Work

We studied performative reinforcement learning through mixtures of policies, and considered *local* and *global* notions of stability. For converging to local stability, no assumption is needed, however, for global stability, the assumption of *Bounded Transition Range* (Assumption 1) is both necessary and sufficient. We also extend our results to performative Markov games, and show that imposing potential-game structure can guarantee multi-agent global stability.

Several questions remain open. First, as we replace deterministic stable policy with a mixture of policies, it is important the mixture is supported on a small number of policies to ensure easy deployment. Our current convergence rate of $O(1/\sqrt{T})$ implies that the mixture is supported on $O(1/\varepsilon^2)$ policies, and an interesting question to understand when it can be improved. Second, extending our framework to a large class of MDPs e.g. linear MDPs (Mandal & Radanovic, 2025) or general function approximation, is an interesting direction of future work. Finally, our current analysis of multi-agent stability relies on exact gradients of the induced potential function, and it would be interesting to understand whether a finite-sample counterpart can be established.

AI USE STATEMENT

In this work, we used generative AI to explore research ideas and create lower-bound examples. In particular, Fable 5 was used to generate the survival instance in Definition 4 and the counter-example used to derive linear lower bound for global stability (Theorem 6 in Appendix). Additionally, Opus 5 was used to search literature and generate text for some of the sections. All the AI-generated results have been carefully verified by the authors, and the authors take full responsibility for the final content of the work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies theoretical aspects in reinforcement learning. We do not anticipate any negative consequences of our work. Our study does not involve any human subjects, and does not involve any dataset release.

REPRODUCIBILITY STATEMENT

Note applicable, as this work is primarily theoretical.

REFERENCES

Yasin Abbasi-Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Alekh Agarwal, Kianté Brantley, Nan Jiang, Sham Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. 2026.

Mete Şeref Ahunbay. First-order (coarse) correlated equilibria in non-concave games. *arXiv preprint arXiv:2403.18174*, 2024. URL https://arxiv.org/abs/2403.18174.

Debabrota Basu, Udvas Das, Brahim Driss, and Uddalak Mukherjee. Performative policy gradient: Optimality in performative reinforcement learning. *arXiv preprint arXiv:2512.20576*, 2025.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Yang Cai, Constantinos Daskalakis, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. On tractable Φ-equilibria in non-concave games. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2403.08171.

Ziyi Chen and Heng Huang. Achieve performatively optimal policy for performative reinforcement learning. *arXiv preprint arXiv:2510.04430*, 2025.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1843–1854, 2020. URL https://proceedings.mlr.press/v119/cheung20a.html.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, pp. 807–834. PMLR, 2019.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023. doi: 10.1287/moor.2022.1287.

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

Gabriele Farina and Juan Carlos Perdomo. The stability of online algorithms in performative prediction. *arXiv preprint arXiv:2602.24207*, 2026.

Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1433–1441, 2017. URL https://proceedings.mlr.press/v70/hazan17a.html.

Yuya Hikima and Akiko Takeda. Zeroth-order methods for nonconvex stochastic problems with decision-dependent distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17195–17203, 2025.

Zachary Izzo, James Zou, and Lexing Ying. How to learn when data gradually reacts to your model. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3998–4035. PMLR, 28–30 Mar 2022.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390, 2019.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *arXiv preprint arXiv:2110.14555*, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Guanghui Lan. Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes: G. lan. *Mathematical programming*, 198(1):1059–1106, 2023.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*, 2022.

Qiang Li and Hoi-To Wai. Stochastic optimization schemes for performative prediction with nonconvex loss. *Advances in Neural Information Processing Systems*, 37:8673–8697, 2024.

Debmalya Mandal and Goran Radanovic. Performative reinforcement learning with linear markov decision process. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3232–3240, 2025.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *International Conference on Machine Learning*, pp. 23642–23680. PMLR, 2023.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7447–7458, 2021.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4929–4939, 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7710–7720. PMLR, 18–24 Jul 2021.

Mohammad Mohammadi, Jonathan Nöther, Debmalya Mandal, Adish Singla, and Goran Radanovic. Implicit poisoning attacks in two-agent reinforcement learning: Adversarial policies for training-time attacks. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pp. 1835–1844. International Foundation for Autonomous Agents and Multiagent Systems, 2023. ISBN 9781450394321.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202):1–56, 2023.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. EC '23, pp. 1047–1074, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597759. URL https://doi.org/10.1145/3580507.3597759.

Vasilis Pollatos, Debmalya Mandal, and Goran Radanovic. On corruption-robustness in performative reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19939–19947, 2025.

Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. In *Uncertainty in Artificial Intelligence*, pp. 3041–3075. PMLR, 2024.

Rilind Sahitaj, Paulius Sasnauskas, Yiğit Yalın, Debmalya Mandal, and Goran Radanovic. Independent learning in performative markov potential games. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312, 2025.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5668–5675, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10279–10288. PMLR, 18–24 Jul 2021.

Cathy Wu, Abdul Rahman Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M Bayen. Flow: A modular learning framework for mixed autonomy traffic. *IEEE Transactions on Robotics*, 38(2):1270–1286, 2021.

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

Appendix

Contents

A Additional Related Work **13**

B Lower Bound on Global Stability **14**

C A Finite-Sample Algorithm for Global Stability **14**

D Separating Local and Global Stability in Markov Games **17**

E Missing Proofs **18**

E.1 Equivalence Between Local and Stationary Stability 18

E.2 Local Stability of the Survival Chain Instance 18

E.3 Global Stability of the Survival Chain Instance 20

E.4 Local Stability without Any Assumption 21

E.5 Finite Sample Analysis of Local Stability 22

E.6 Global Stability Under Bounded Transition Range 26

E.7 Lower bound under Local Feedback 27

E.8 Multi-Agent Setting 29

E.9 Performative Markov Potential Games 29

A Additional Related Work

Policy optimisation and Markov potential games. Our finite-sample algorithms build on the classical policy-gradient (Sutton et al., 1999; Agarwal et al., 2021) and mirror-descent algorithms for MDPs (Shani et al., 2020; Lan, 2023). Leonardos et al. (2022) introduce Markov potential games and show that independent policy gradient converges to a stationary (Nash) policy in the non-performative setting; Ding et al. (2022) sharpen the rate and identify the distribution mismatch coefficient that also appears in Theorem 5. Our contribution is to adapt this line of analysis to the performative setting. Beyond potential games, Jin et al. (2021) gives a decentralised algorithm that converges to coarse correlated equilibrium (CCE) in general-sum Markov games without any structural assumption. CCE is structurally close to our unconditional multi-agent local stability guarantee (Theorem 4), though it targets the classical, non-performative equilibrium.

First-order correlated equilibria. Our local mixed stability notion is also connected to equilibrium relaxations for non-concave games. Hazan et al. (2017) introduce a local regret criterion based on windowed losses, while Ahunbay (2024) study first-order (coarse) correlated equilibria defined through expected gains from infinitesimal strategy modifications. Cai et al. (2024) study tractable Φ-equilibria, including local deviations that interpolate towards a comparator strategy, and reduce their computation in the first-order regime to regret minimisation on linearised losses. In a fixed Markov game, our local gap measures the expected first-order gain from moving each player's policy towards a fixed comparator, with the occupancy weights arising from the policy gradient theorem.

Online and non-stationary reinforcement learning. Even-Dar et al. (2009) study online MDPs with adversarial rewards and fixed dynamics using statewise expert-learning algorithms, providing an algorithmic antecedent to our per-state Hedge updates. With changing dynamics, Cheung et al.

(2020) and Mao et al. (2021) obtain dynamic-regret guarantees under variation budgets that control cumulative changes in rewards and transitions. Our bounded transition range assumption instead controls the distance of each induced kernel from a reference kernel: repeated switching between nearby kernels is permitted even when cumulative variation grows

B Lower Bound on Global Stability

In the previous section, we have shown that under the bounded transition range assumption, it is possible to obtain a globally performatively stable mixture of policies up to a floor of $\Theta(\gamma\epsilon_P/(1-\gamma)^3)$. Moreover, substituting $H = 2$ in Lemma 4 shows that there exists an instance where the global performative stability i.e. $\mathrm{Gap}^{\mathrm{glob}}(\mu)$is lower bounded by $\Omega(\gamma^2\epsilon_P^2/(1-\gamma))$ for any mixture μ. The reason the lower bound instance in Definition 4 generates a weaker bound is that the learning algorithm receives full feedback about the new environment. We will consider a feedback model that closely resembles the setup considered in Algorithms 1 and 2.

Definition 12 (Local feedback)**.** At round t the learner deploys π_t, the environment becomes $M_t = M(\pi_t)$, and the learner draws trajectories from M_t under π_t. Each trajectory is of the form $(s_t^0, a_t^0, r_t^0, s_t^1, a_t^1, r_t^1, \dots)$, and the learner never observes reward at unvisited pairs (s, a), nor $M(\pi)$ for any π it has not deployed. The process repeats for T round, and after N trajectories in total the learner outputs $\hat{\mu} = \mathrm{Unif}\{\pi_1, \dots, \pi_T\}$.

In order to prove the lower bound, we will consider a family of instances indexed by $j \in [K]$. Each instance is a simple two-action MDP with a single state, where one of the actions is hidden and contrarian. The environment rewards the hidden action exactly when the learner underweights it.

Definition 13 (Hidden-Action Instance)**.** Fix $\epsilon \in (0, 1/4]$ and a hidden action $j \in [K]$. The state space is $\{s_0, \top, \bot\}$ with $\rho = \delta_{s_0}$ and $\mathcal{A} = [K]$. The states $\top, \bot$ are absorbing with $r(\top, \cdot) = 1$, $r(\bot, \cdot) = 0$, and $r(s_0, \cdot) = 0$. From s_0, action a moves to $\top$ with probability p_a and to $\bot$ otherwise, where $p_a = \frac{1}{2}$ for $a \neq j$ and

$$p_j(\pi) \;=\; \begin{cases} \frac{1}{2} + \epsilon, & \pi(j|s_0) < \frac{1}{2}, \\ \frac{1}{2} - \epsilon, & \pi(j|s_0) \geq \frac{1}{2}. \end{cases}$$

Note that, writing $\epsilon_P := 2\epsilon$, the instance satisfies Assumption 1 with reference kernel $\bar{P}(\cdot|s_0, a) = \mathrm{Bern}(\frac{1}{2})$ for *every* a. The next theorem shows that if the algorithm is restricted to sample at most $O(K/\epsilon_P^2)$ trajectories, then there exists a hidden action j for which the expected global performative stability of the output mixture $\hat{\mu}$ is lower bounded by $\Omega(\gamma\epsilon_P/(1-\gamma))$. This shows that the upper bound in Theorem 3 is tight up to a factor of $O(1/(1-\gamma)^2)$.

Theorem 6. *Consider the instance of Definition 13 with $K \geq 20$ and $\epsilon = \epsilon_P/2 \leq 1/4$. Let any algorithm output a mixture of policies $\hat{\mu}$. If the total number of trajectories N satisfies $N \leq \frac{K}{24\,\epsilon_P^2}$, then there exists a hidden action $j \in [K]$ for which*

$$\mathbb{E}_{\mathbb{P}_j}\left[\mathrm{Gap}^{\mathrm{glob}}(\hat{\mu})\right] \geq \frac{\gamma\epsilon_P}{8(1-\gamma)},$$

where $\mathbb{P}_j$ is the law of the algorithm's trajectory of observations with hidden action j.

C A Finite-Sample Algorithm for Global Stability

Similar to Algorithm 1 we can develop an algorithm that obtains a globally stable mixture of policies using trajectories drawn from deployed policies.

Theorem 7 (Finite-sample global stability)**.** *Suppose Algorithm 2 is run with $\eta = 2(1-\gamma)\sqrt{2S\ln A/T}$ and $T \geq k\ln(SA/\delta)$ iterations, and let $\pi_1, \dots, \pi_T$ be generated by equation 8 and $\hat{\mu} := \mathrm{Unif}\{\pi_1, \dots, \pi_T\}$. Then with probability at least $1-\delta$,*

$$\mathrm{Gap}^{\mathrm{glob}}(\hat{\mu}) \;\leq\; \frac{1}{(1-\gamma)^2}\sqrt{\frac{\ln(SA)}{2T}} + \frac{3\gamma\,\epsilon_P}{(1-\gamma)^3} + \frac{4}{(1-\gamma)^2}\sqrt{\frac{\ln(SA/\delta)}{kT}}.$$

The total number of value and occupancy samples required by the algorithm is $TSAk$.

Algorithm 2 Finite-sample unweighted per-state Hedge

Input: Batch sizes $k \geq 1$ and a step size $\eta > 0$.

1: Initialize $\pi_1(\cdot|s) = \mathrm{Unif}(\mathcal{A})$ for each $s \in \mathcal{S}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Deploy π_t.
4: *Value batch.* For each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, draw k independent value samples at (s, a) and let $\widehat{Q}_t(s, a)$ be their average.
5: Update,

$$\pi_{t+1}(a \mid s) \;\propto\; \pi_t(a \mid s) \exp\left(\eta\, \widehat{Q}_t(s, a)\right). \tag{8}$$

6: **end for**

Output: Unif $(\pi_1, \ldots, \pi_T)$.

By substituting $k = \Theta(1)$ we can get the following corollary.

Corollary 1. *In order to reach* $\mathrm{Gap}^{\mathrm{glob}}(\widehat{\mu}) \leq \epsilon$ *for any* $\epsilon \geq \frac{3\gamma\epsilon_P}{(1-\gamma)^3}$ *with probability at least* $1 - \delta$ *it suffices to take* $T = O\left(\frac{\ln(SA/\delta)}{(1-\gamma)^4\epsilon^2}\right)$ *iterations and* $O\left(\frac{SA\ln(SA/\delta)}{(1-\gamma)^4\epsilon^2}\right)$ *trajectories in total.*

Proof of Theorem 7. Let us write $\zeta_t(s, a) = Q_t(s, a) - \widehat{Q}_t(s, a)$. Then we can write

$$\begin{aligned}
\mathrm{Gap}^{\mathrm{glob}}(\hat{\mu}) &= \max_{\pi' \in \Pi} \frac{1}{T} \sum_{t=1}^{T} V_{M_t}(\pi') - V_{M_t}(\pi_t) \\
&= \max_{\pi' \in \Pi} \frac{1}{(1-\gamma)T} \sum_{t=1}^{T} \sum_{s} d_{P_t}^{\pi'}(s) \Delta_t(s, \pi') \\
&= \max_{\pi' \in \Pi} \frac{1}{(1-\gamma)T} \sum_{t=1}^{T} \sum_{s} d_{P_t}^{\pi'}(s) \left\langle \pi'(\cdot|s) - \pi_t(\cdot|s), Q_t(s, \cdot) \right\rangle \\
&= \max_{\pi' \in \Pi} \frac{1}{(1-\gamma)T} \underbrace{\sum_{t=1}^{T} \sum_{s} d_{P_t}^{\pi'}(s) \left\langle \pi'(\cdot|s) - \pi_t(\cdot|s), \widehat{Q}_t(s, \cdot) \right\rangle}_{:=(I)} \\
&\quad + \frac{1}{(1-\gamma)T} \underbrace{\sum_{t=1}^{T} \sum_{s} d_{P_t}^{\pi'}(s) \left\langle \pi'(\cdot|s) - \pi_t(\cdot|s), \zeta_t(s, \cdot) \right\rangle}_{:=(II)}.
\end{aligned}$$

We now proceed similar to the proof of Theorem 8, and bound (I) using Lemma 7 and (II) using Freedman's inequality.

Bounding (I)**:** Since $\widehat{Q}_t(s, a) \in [0, 1/(1-\gamma)]$, we can apply Lemma 7 to get that for any $\pi' \in \Pi$,

$$\sum_{t=1}^{T} \Delta_t(s, \pi') \leq \frac{\ln A}{\eta} + \frac{\eta T}{8(1-\gamma)^2} =: R_T.$$

Substituting $\eta = 2(1-\gamma)\sqrt{2\ln A/T}$, we get $R_T = \frac{1}{1-\gamma}\sqrt{T \ln A/2}$. Now we can proceed similar to the proof of Theorem 3 to get that

$$\max_{\pi' \in \Pi}(I) \leq \frac{1}{1-\gamma}\sqrt{T \ln A/2} + \frac{T\gamma\epsilon_P}{(1-\gamma)^2}. \tag{9}$$

Bounding (II)**:** Writing $d^{\pi'}_{P_t}(s) = d^{\pi'}_{\widehat{P}}(s) + e_t(s)$, we can write

$$\begin{aligned}\max_{\pi' \in \Pi}(II) &= \max_{\pi' \in \Pi} \sum_{t=1}^{T} \sum_{s} d^{\pi'}_{P_t}(s) \langle \pi'(\cdot|s) - \pi_t(\cdot|s), \zeta_t(s,\cdot)\rangle \\ &= \max_{\pi' \in \Pi} \sum_{t=1}^{T} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \langle \pi'(\cdot|s) - \pi(\cdot|s), \zeta_t(s,\cdot)\rangle - \sum_{t=1}^{T} \sum_{s} e_t(s) \langle \pi'(\cdot|s) - \pi_t(\cdot|s), \zeta_t(s,\cdot)\rangle\end{aligned}$$

Using Assumption 1 and Lemma 12, we can bound the second term as follows: $\frac{1}{1-\gamma}\sum_{t=1}^{T}\sum_s |e_t(s)|\,\|\pi'(\cdot|s) - \pi_t(\cdot|s)\|_1 \le \frac{T}{1-\gamma}\|e_t\|_1 \le \frac{T\gamma\epsilon_P}{(1-\gamma)^2}$. This gives us,

$$\begin{aligned}\max_{\pi \in \Pi}(II) &\le \max_{\pi' \in \Pi} \sum_{t=1}^{T} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \langle \pi'(\cdot|s) - \pi_t(\cdot|s), \zeta_t(s,\cdot)\rangle + \frac{T\gamma\epsilon_P}{(1-\gamma)^2} \\ &\le \max_{\pi' \in \Pi} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \sum_{t=1}^{T} \langle \pi'(\cdot|s), \zeta_t(s,\cdot)\rangle + \max_{\pi' \in \Pi} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \sum_{t=1}^{T} \langle \pi_t(\cdot|s), -\zeta_t(s,\cdot)\rangle + \frac{T\gamma\epsilon_P}{(1-\gamma)^2} \\ &\le \max_{\pi' \in \Pi} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \max_{a} \sum_{t=1}^{T} \zeta_t(s,a) + \max_{\pi' \in \Pi} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \sum_{t=1}^{T} \langle \pi_t(\cdot|s), -\zeta_t(s,\cdot)\rangle + \frac{T\gamma\epsilon_P}{(1-\gamma)^2}.\end{aligned}$$

Let us now write $Z_T(s,a) = \sum_{t=1}^{T} \zeta_t(s,a)$, and $y_{t,s} = \langle \pi_t(\cdot|s), \zeta_t(s,\cdot)\rangle$. Then we can write

$$\max_{\pi' \in \Pi}(II) \le \max_{\pi' \in \Pi} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \max_{a} Z_T(s,a) + \max_{\pi' \in \Pi} \sum_{s} d^{\pi'}_{\widehat{P}}(s) \sum_{t=1}^{T} y_{t,s} + \frac{T\gamma\epsilon_P}{(1-\gamma)^2} \tag{10}$$

We now bound the remaining two terms. First, note that $\mathbb{E}[\zeta_t(s,a)|\mathcal{F}_{t-1}] = 0$ as $\widehat{Q}_t(s,a)$ is an unbiased estimate of $Q_t(s,a)$. Moreover, $\widehat{Q}_t(s,a)$ is the average of k independent random variables in $[0, 1/(1-\gamma)]$, and hence we have

$$\mathbb{E}[\zeta_t(s,a)^2|\mathcal{F}_{t-1}] = \mathrm{Var}(\widehat{Q}_t(s,a)) \le \frac{1}{k(1-\gamma)^2}.$$

Therefore, by applying Freedman's inequality Lemma 9, we get that with probability at least $1-\delta$, for any $\lambda \in [0, (1-\gamma)]$

$$\forall s, a,\ Z_T(s,a) = \sum_{t=1}^{T} \zeta_t(s,a) \le (e-2)\lambda \cdot \frac{T}{k(1-\gamma)^2} + \frac{\ln(SA/\delta)}{\lambda}.$$

Substituting $\lambda = \sqrt{k(1-\gamma)^2 \ln(SA/\delta)/(T(e-2))}$, we get that with probability at least $1-\delta$, for all s, a,

$$Z_T(s,a) \le \frac{2}{(1-\gamma)}\sqrt{\frac{T}{k}\ln(SA/\delta)}.$$

Note that as long as $T \ge k\ln(SA/\delta)$, we can guarantee that the requried $\lambda \in [0, (1-\gamma)]$.

Now we upper bound the second term in Equation (10). Note that $y_{t,s}$ is a martingale difference sequence with respect to $\mathcal{F}_t$,as $\mathbb{E}[y_{t,s}|\mathcal{F}_{t-1}] = \mathbb{E}_{\pi_t}[\mathbb{E}[y_{t,s}|\mathcal{F}_{t-1}, \pi_t]] = 0$. Furthermore, $|y_{t,s}| \le$

$1/(1-\gamma)$.

$$\begin{aligned}
\mathbb{E}[y_{t,s}^2|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\left(\sum_a \pi_t(a|s)\zeta_t(s,a)\right)^2 \middle| \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_a \pi_t(a|s)^2\zeta_t(s,a)^2 \middle| \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\sum_{a,b} \pi_t(a|s)\pi_t(b|s)\zeta_t(s,a)\zeta_t(s,b) \middle| \mathcal{F}_{t-1}\right] \\
&\leq \sum_a \pi_t(a|s)\mathbb{E}[\zeta_t(s,a)^2|\mathcal{F}_{t-1}] \\
&\leq \frac{1}{k(1-\gamma)^2}
\end{aligned}$$

Therefore, we can apply Freedman's inequality again to get that with probability at least $1-\delta$, for any $\lambda \in [0,(1-\gamma)]$,

$$\sum_{t=1}^{T} y_{t,s} \leq (e-2)\lambda \cdot \frac{T}{k(1-\gamma)^2} + \frac{\ln(S/\delta)}{\lambda}.$$

Substituting $\lambda = \sqrt{k(1-\gamma)^2\ln(S/\delta)/(T(e-2))}$, we get that with probability at least $1-\delta$, for all s,

$$\sum_{t=1}^{T} y_{t,s} \leq \frac{2}{(1-\gamma)}\sqrt{\frac{T}{k}\ln(S/\delta)}.$$

Substituting the above two bounds in Equation (10), we get that with probability at least $1-2\delta$,

$$\begin{aligned}
\max_{\pi' \in \Pi}(II) &\leq \frac{2}{(1-\gamma)}\sqrt{\frac{T\ln(SA/\delta)}{k}} + \frac{2}{(1-\gamma)}\sqrt{\frac{T}{k}\ln(S/\delta)} + \frac{T\gamma\epsilon_P}{(1-\gamma)^2} \\
&\leq \frac{4}{(1-\gamma)}\sqrt{\frac{T\ln(SA/\delta)}{k}} + \frac{T\gamma\epsilon_P}{(1-\gamma)^2},
\end{aligned}$$

Now adding the upper bound for (I) given in Equation (9), and dividing by $T(1-\gamma)$ we get the desired bound. □

D Separating Local and Global Stability in Markov Games

We construct a simple two-player Markov game that separates local and global stability, even in the absence of performative effects ($\epsilon_P = 0$).

Definition 14 (Two-player survival chain)**.** Fix $H \geq 2$ and $\epsilon \in (0, 1/2)$. States $\{s_1, \ldots, s_H\} \cup \{\top, \bot\}$ with $\rho = \delta_{s_1}$; $\mathcal{A}_1 = \{1,2\}$ and $\mathcal{A}_2 = \{A, B\}$. From s_h under the joint action (j, b) the process advances to s_{h+1} (to $\top$ if $h = H$) with probability $\ell^b(j)$ and falls to $\bot$ otherwise, where $\ell^A = (1,\ 1-2\epsilon)$ and $\ell^B = (1-2\epsilon,\ 1)$. The rewards are $r^1(\top,\cdot) = 1$, $r^1(\bot,\cdot) = 0$ and $r^2 = 1 - r^1$, zero on the chain. The game is *fixed*, so $\epsilon_P = 0$.

This example generalizes Definition 4 and player 2 chooses which of player 1's actions is safe — precisely the choice that the environment map made in Definition 5.

Proposition 1 (Separation)**.** *Let $\bar{\pi}^1$ play $\mathrm{Unif}(\mathcal{A}_1)$ at every s_h, let π_A^2 play A at every state and π_B^2 play B at every state, and set $\mu^\star := \frac{1}{2}\delta_{(\bar{\pi}^1, \pi_A^2)} + \frac{1}{2}\delta_{(\bar{\pi}^1, \pi_B^2)}$. Then,*

$$\mathrm{Gap}^{\mathrm{loc}}(\mu^\star) = 0 \qquad \textit{and} \qquad \mathrm{Gap}^{\mathrm{glob}}(\mu^\star) \geq \frac{\gamma^H}{16(1-\gamma)}\min\left\{1, \epsilon^2 H^2\right\} \ \forall H \geq 2.$$

Proof of Proposition 1. The two components share their occupancies. Player 1's per-step survival probability is $\langle \bar{\pi}^1, \ell^A\rangle = \langle \bar{\pi}^1, \ell^B\rangle = 1-\epsilon$, so the state occupancy d and player 1's value function V along the chain are identical under the two policies π_A^2 and π_B^2 of player 2. We will write $d(s_h)$ to denote the state occupancy and $V^1(s_{h+1})$ for the common value function of player 1.

Player 1 *is locally indifferent.* By Proposition 3, $\bar{q}^1_{\mu^\star}(s_h, j) = d(s_h)\,\gamma V^1(s_{h+1}) \cdot \frac{1}{2}\big(\ell^A(j) + \ell^B(j)\big)$, and $\ell^A(1)+\ell^B(1) = \ell^A(2)+\ell^B(2) = 2-2\epsilon$, so this is constant in j and equals $d(s_h)\gamma V^1(s_{h+1})(1-\epsilon)$. The deployed term $\bar{v}^1_{\mu^\star}(s_h)$ can be computed as

$$\begin{aligned}\bar{v}^1_{\mu^\star}(s_h) &= d(s_h)\tfrac{1}{2}\left(Q^1(s_h,1) + Q^1(s_h,2)\right)\\ &= d(s_h)\,\gamma V^1(s_{h+1})(1-\epsilon).\end{aligned}$$

Therefore, $\bar{v}^1_{\mu^\star}(s_h)$ takes the same value as $\bar{q}^1_{\mu^\star}(s_h, j)$, so every per-state gap vanishes and according to Proposition 3, $\mathrm{Gap}_1^{\mathrm{loc}}(\mu^\star) = 0$.

Player 2 *is locally indifferent.* Against $\bar{\pi}^1$, player 2's choice gives the same survival probability $1-\epsilon$ to player 1, so it gives the same probability ϵ of reaching state $\perp$ for player 2. Therefore, by a very similar calculation to the one above, we can show that $\bar{q}^2_{\mu^\star}(s_h, b)$ is constant in b and equals $\bar{v}^2_{\mu^\star}(s_h)$. Therefore, $\mathrm{Gap}_2^{\mathrm{loc}}(\mu^\star) = 0$, and $\mathrm{Gap}^{\mathrm{loc}}(\mu^\star) = \max_{i\in\{1,2\}} \mathrm{Gap}_i^{\mathrm{loc}}(\mu^\star) = 0$.

Player 1 *gains globally.* Let us write σ to denote a comparator policy of player 1.With $F_A(y) := \prod_h \left(1-2\epsilon(1-y_h)\right)$ and $F_B(y) := \prod_h \left(1-2\epsilon y_h\right)$ as in equation 2, $\mathbb{E}_{\mu^\star}[V^1(\sigma,\pi^2)] = \kappa\cdot\frac{1}{2}\big(F_A(\sigma)+F_B(\sigma)\big)$, where $\kappa = \gamma^H/(1-\gamma)$. In particular, when $\sigma = \pi^1_1$ (i.e. always play 1), we have $\mathbb{E}_{\mu^\star}[V^1(\sigma,\pi^2)] = \kappa\cdot\frac{1}{2}\big(1+(1-2\epsilon)^H\big)$. The deployed value i.e. $\mathbb{E}_{\mu^\star}[V^1(\boldsymbol{\pi})]$ equals $\kappa(1-\epsilon)^H$, as player 1's probability of survival is $1-\epsilon$ at every step. Therefore, $\mathrm{Gap}_1^{\mathrm{glob}}(\mu^\star) \geq \kappa f(\epsilon, H)$, where $f(\epsilon,H) = \frac{1}{2}(1+(1-2\epsilon)^H) - (1-\epsilon)^H$. The final bound now follows by using the lower bound on $f(\epsilon,H)$ from Lemma 5. □

E Missing Proofs

E.1 Equivalence Between Local and Stationary Stability

Proof of Lemma 2. From the policy gradient functional form Sutton et al. (1999) and Agarwal et al. (2026) (Theorem 9.4), we have the following expression of the partial derivative.

$$\frac{\partial V_{M(\pi)}(\pi')}{\partial \pi'(a|s)} = \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}\sum_a d^{\pi'}_{M(\pi)}(s) Q^{\pi'}_{M(\pi)}(s,a)$$

Taking inner product with $\pi' - \pi$ we obtain the following.

$$\begin{aligned}\left\langle \nabla_{\pi'} V^{\pi'}_{M(\pi)}\big|_{\pi'=\pi}, \pi' - \pi\right\rangle &= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d^{\pi}_{P_\pi}(s)\left\langle \pi'(\cdot|s) - \pi(\cdot|s), Q^{\pi}_{M(\pi)}(s,\cdot)\right\rangle\\ &= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d^{\pi}_{P_\pi}(s)\Delta_\pi(s,\pi')\end{aligned}$$

Therefore, the policy π is δ-stationary performatively stable if and only if $\frac{1}{1-\gamma}\max_{\pi'\in\Pi}\sum_{s\in\mathcal{S}} d^{\pi}_{P_\pi}(s)\Delta_\pi(s,\pi') = \mathrm{Gap}^{\mathrm{loc}}(\pi) \leq \delta$. □

E.2 Local Stability of the Survival Chain Instance

Proof of Lemma 3. In order to prove the first bound, we will consider the following policy class. $\Pi_0 = \left\{\pi^+_\eta, \pi^-_\eta : \eta \in [0, 1/2]\right\}$, where π^+_η is a stationary policy with $y_h = \pi(1|s_h) = \frac{1}{2} + \eta$. And, similarly π^-_η is a stationary policy with $y_h = \pi(1|s_h) = \frac{1}{2} - \eta$. Let us also define the mixture μ_η to be supported with equal probability over π^+_η and π^-_η i.e.

$$\mu_\eta = \frac{1}{2}\cdot\delta_{\pi^+_\eta} + \frac{1}{2}\cdot\delta_{\pi^-_\eta}.$$

First note that, $V_{M^A}(\pi^+_\eta) = \kappa\left(1-2\epsilon(1-(1/2+\eta))\right)^H = \kappa(1-\epsilon+2\epsilon\eta)^H$, and $V_{M^B}(\pi^+_\eta) = \kappa(1-2\epsilon(1/2+\eta))^H = \kappa(1-\epsilon-2\epsilon\eta)^H$. Since $F_A(\pi^+_\eta) \geq F_B(\pi^+_\eta)$ we have $M(\pi^+_\eta) = M^B$. Similarly, we can show that $M(\pi^-_\eta) = M^A$.

Local Gap. Now we compute $\mathrm{Gap}^{\mathrm{loc}}(\mu_\eta)$ using Proposition 2. Let us evaluate the following two quantities.

$$\bar{q}_\mu(s,a) := \mathbb{E}_{\pi\sim\mu}\big[d^\pi_{P_\pi}(s)\, Q^\pi_{M(\pi)}(s,a)\big], \qquad \bar{v}_\mu(s) := \mathbb{E}_{\pi\sim\mu}\big[d^\pi_{P_\pi}(s)\, \langle \pi(\cdot \mid s), Q^\pi_{M(\pi)}(s,\cdot)\rangle\big].$$

Let us write $\bar{\ell}(\eta) = 1 - \epsilon - 2\epsilon\eta$. Then we have,

$$Q^{\pi^+_\eta}_{M^B}(s_h,a) = \ell^B(a)\cdot V^{\pi^+_\eta}_{M^B}(s_{h+1}) = \ell^B(a)\cdot\kappa\cdot\bar{\ell}(\eta)^{H-h} \text{ and } d^{\pi^+_\eta}_{P^B}(s_h) = \bar{\ell}(\eta)^h.$$

Similarly,

$$Q^{\pi^-_\eta}_{M^A}(s_h,a) = \ell^A(a)\cdot V^{\pi^+_\eta}_{M^A}(s_{h+1}) = \ell^A(a)\cdot\kappa\cdot\bar{\ell}(\eta)^{H-h} \text{ and } d^{\pi^-_\eta}_{P^A}(s_h) = \bar{\ell}(\eta)^h.$$

As $\ell^A(a) + \ell^B(a) = 2 - 2\epsilon$ for any action $a \in \mathcal{A}$, we obtain

$$\bar{q}_{\mu_\eta}(s,a) = \frac{1}{2}Q^{\pi^+_\eta}_{M^B}(s,a)\cdot d^{\pi^+_\eta}_{P^B}(s) + \frac{1}{2}Q^{\pi^-_\eta}_{M^A}(s,a)\cdot d^{\pi^-_\eta}_{P^A}(s) = \kappa\cdot(1-\epsilon)\bar{\ell}(\eta)^H.$$

Therefore, $\bar{q}_{\mu_\eta}(s,a)$ is independent of action a. In order to evaluate $\bar{v}_\mu(s)$, let us first evaluate $\left\langle \pi^+_\eta(\cdot|s_h), Q^{\pi^+_\eta}_{M^B}(s_h,\cdot)\right\rangle$.

$$\begin{aligned}\left\langle \pi^+_\eta(\cdot|s_h), Q^{\pi^+_\eta}_{M^B}(s_h,\cdot)\right\rangle &= \left(\frac{1}{2}+\eta\right)\ell^B(1)\cdot\kappa\cdot\bar{\ell}(\eta)^{H-h} + \left(\frac{1}{2}-\eta\right)\ell^B(2)\cdot\kappa\cdot\bar{\ell}(\eta)^{H-h}\\ &= \kappa\cdot\bar{\ell}(\eta)^{H-h+1}\end{aligned}$$

Therefore,

$$\begin{aligned}\bar{v}_\mu(s_h) &= \frac{1}{2}d^{\pi^+_\eta}_{P^B}(s)\left\langle \pi^+_\eta(\cdot|s_h), Q^{\pi^+_\eta}_{M^B}(s_h,\cdot)\right\rangle + \frac{1}{2}d^{\pi^-_\eta}_{P^A}(s)\left\langle \pi^-_\eta(\cdot|s_h), Q^{\pi^-_\eta}_{M^A}(s_h,\cdot)\right\rangle\\ &= \kappa\cdot\bar{\ell}(\eta)^{H+1}.\end{aligned}$$

We now use Proposition 2 to bound $\mathrm{Gap}^{\mathrm{loc}}(\mu_\eta)$.

$$\begin{aligned}\mathrm{Gap}^{\mathrm{loc}}(\mu_\eta) &= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}\Big(\max_{a\in\mathcal{A}}\bar{q}_\mu(s,a) \;-\; \bar{v}_\mu(s)\Big)\\ &= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}\big(\kappa\cdot(1-\epsilon)\bar{\ell}(\eta)^H - \kappa\cdot\bar{\ell}(\eta)^{H+1}\big)\\ &= \frac{2\epsilon\eta\kappa(H+2)}{1-\gamma}\bar{\ell}(\eta)^H\end{aligned}$$

Now, $\inf_{\mu\in\Delta(\Pi)}\mathrm{Gap}^{\mathrm{loc}}(\mu) \le \inf_{\mu\in\Delta(\Pi_0)}\mathrm{Gap}^{\mathrm{loc}}(\mu) \le \inf_{\eta\to 0}\mathrm{Gap}^{\mathrm{loc}}(\mu_\eta) = 0$.

□

Proposition 2. *Given a mixture μ of policies, let us define, for each $s \in \mathcal{S}$,*

$$\bar{q}_\mu(s,a) := \mathbb{E}_{\pi\sim\mu}\big[d^\pi_{P_\pi}(s)\, Q^\pi_{M(\pi)}(s,a)\big], \qquad \bar{v}_\mu(s) := \mathbb{E}_{\pi\sim\mu}\big[d^\pi_{P_\pi}(s)\, \langle \pi(\cdot \mid s), Q^\pi_{M(\pi)}(s,\cdot)\rangle\big].$$

Then

$$\mathrm{Gap}^{\mathrm{loc}}(\mu) \;=\; \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}\Big(\max_{a\in\mathcal{A}}\bar{q}_\mu(s,a) \;-\; \bar{v}_\mu(s)\Big).$$

Proof. Expanding the inner product in $\Delta_\pi(s,\pi')$ and exchanging the (finite) sum over s with the expectation,

$$\begin{aligned}\mathbb{E}_{\pi\sim\mu}\sum_s d^\pi_{P_\pi}(s)\Delta_\pi(s,\pi') &= \mathbb{E}_{\pi\sim\mu}\sum_s d^\pi_{P_\pi}(s)\left\langle \pi'(\cdot|s) - \pi(\cdot|s), Q^\pi_{M(\pi)}(s,\cdot)\right\rangle\\ &= \sum_s\Big(\big\langle \pi'(\cdot \mid s),\, \bar{q}_\mu(s,\cdot)\big\rangle - \bar{v}_\mu(s)\Big).\end{aligned}$$

The summand depends on π' only through $\pi'(\cdot \mid s)$, and the constraint set $\Pi = \Delta(\mathcal{A})^{\mathcal{S}}$ is a product over s. Hence the maximisation decouples across states, and for each s the maximum of the linear functional $q \mapsto \langle q, \bar{q}_\mu(s,\cdot)\rangle$ over $q \in \Delta(\mathcal{A})$ is attained at a vertex, giving $\max_a \bar{q}_\mu(s,a)$. □

E.3 Global Stability of the Survival Chain Instance

Proof of Lemma 4. We will use Lemma 6 to provide a lower bound on the approximation of global mixed performative stability $\mathrm{Gap}^{\mathrm{glob}}(\mu)$. Let π^A (resp. π^B) plays action 1 (resp. 2) at every s_h. Then we define $\nu = \frac{1}{2}\delta_{\pi^A} + \frac{1}{2}\delta_{\pi^B}$. Now our aim is to establish a lower bound on the following quantity.

$$\inf_{\pi\in\Pi}\left(\mathbb{E}_{\pi'\sim\nu}[V_{M(\pi)}(\pi')] - V_{M(\pi)}(\pi)\right).$$

First suppose, we choose π (with $y_h = \pi(1|s_h)$) such that $M(\pi) = M^A$ i.e. $F_A(y) \le F_B(y)$. Then,

$$\mathbb{E}_{\pi'\sim\nu}\left[V_{M^A}(\pi')\right] = \frac{\kappa}{2}\left(1 + (1-2\epsilon)^H\right).$$

Similarly, it can be verified that

$$\mathbb{E}_{\pi'\sim\nu}\left[V_{M^B}(\pi')\right] = \frac{\kappa}{2}\left(1 + (1-2\epsilon)^H\right).$$

On the other hand, for any policy $\pi \in \Pi$,

$$\begin{aligned} V_{M(\pi)}(\pi) &= \kappa \cdot \min\left\{F_A(y), F_B(y)\right\} \le \kappa\sqrt{F_A(y)F_B(y)} \\ &= \kappa\sqrt{\prod_{h=1}^{H}(1-2\epsilon(1-y_h))(1-2\epsilon y_h)} \\ &\le \kappa(1-\epsilon)^H \end{aligned}$$

This is because each term $g(y) = (1-2\epsilon(1-y))(1-2\epsilon y)$ is maximized at $1/2$. Therefore, for any policy $\pi \in \Pi$

$$\begin{aligned} \mathbb{E}_{\pi'\sim\nu}\left[V_{M(\pi)}(\pi')\right] - V_{M(\pi)}(\pi) &\ge \kappa\left(\frac{1}{2}\left(1 + (1-2\epsilon)^H\right) - (1-\epsilon)^H\right) \\ &\ge \kappa\frac{1}{16}\min\left\{1, \epsilon^2H^2\right\}. \end{aligned}$$

□

Lemma 5. *Let us define* $f(\epsilon, H) = \frac{1+(1-2\epsilon)^H}{2} - (1-\epsilon)^H$. *Then,*

$$f(\epsilon, H) \ge \frac{1}{16}\min\left\{1, \epsilon^2H^2\right\} \quad \textit{for } \epsilon \in [0, 1/2], H \ge 2.$$

Proof. Let $X_1, \ldots, X_H$ be i.i.d. Bernoulli trials with $\Pr[X_i = 1] = \epsilon$. First note that,

$$1 = (1-\epsilon+\epsilon)^H = \sum_{k=0}^{H}\binom{H}{k}\epsilon^k(1-\epsilon)^k$$

and

$$(1-2\epsilon)^H = \sum_{k=0}^{H}\binom{H}{k}(-1)^k\epsilon^k(1-\epsilon)^{H-k}.$$

Summing the above two identities and observing that the terms corresponding to k odd cancel out, we obtain the following identity.

$$\frac{1+(1-2\epsilon)^H}{2} = \sum_{k:k \text{ even}}^{H}\binom{H}{k}\epsilon^k(1-\epsilon)^{H-k}$$

Therefore,

$$f(\epsilon, H) = \frac{1+(1-2\epsilon)^H}{2} - (1-\epsilon)^H = \sum_{\substack{k:k \text{ even}\\ k\ge 2}}^{H}\binom{H}{k}\epsilon^k(1-\epsilon)^{H-k} \ge \binom{H}{2}\epsilon^2(1-\epsilon)^{H-2}.$$

Now we consider two cases. First, if $(1-\epsilon)^H \geq 1/4$, we have

$$f(\epsilon, H) \geq \frac{H^2}{4}\epsilon^2 \frac{(1-\epsilon)^H}{(1-\epsilon)^2} \geq \frac{1}{16}\epsilon^2 H^2.$$

Second, if $(1-\epsilon)^H \leq 1/4$, $f(\epsilon, H) \geq 1/4 + (1-2\epsilon)^H \geq 1/4$. Combining the two cases, we have the desired result. □

Lemma 6 (Certificate lemma). *Let $\nu \in \Delta(\Pi)$ and $c \in \mathbb{R}$ satisfy*

$$\mathbb{E}_{\pi' \sim \nu}\big[V_{M(\pi)}(\pi')\big] - V_{M(\pi)}(\pi) \;\geq\; c \qquad \textit{for every } \pi \in \Pi. \tag{11}$$

Then $\mathrm{Gap}^{\mathrm{glob}}(\mu) \geq c$ for every $\mu \in \Delta(\Pi)$.

Proof. Fix μ. A maximum dominates an average, so

$$\max_{\pi' \in \Pi} \mathbb{E}_{\pi \sim \mu}\big[V_{M(\pi)}(\pi')\big] \;\geq\; \mathbb{E}_{\pi' \sim \nu}\, \mathbb{E}_{\pi \sim \mu}\big[V_{M(\pi)}(\pi')\big] \;=\; \mathbb{E}_{\pi \sim \mu}\, \mathbb{E}_{\pi' \sim \nu}\big[V_{M(\pi)}(\pi')\big],$$

by Fubini (all quantities are bounded by $1/(1-\gamma)$). Subtracting $\mathbb{E}_{\pi \sim \mu}[V_{M(\pi)}(\pi)]$ and applying equation 11 pointwise in π gives $\mathrm{Gap}^{\mathrm{glob}}(\mu) \geq c$. □

E.4 Local Stability without Any Assumption

Proof of Theorem 1. Fix an arbitrary comparator $\pi' \in \Pi$ and a state $s \in \mathcal{S}$. Apply Lemma 7 at state s with $p_t = \pi_t(\cdot \mid s)$, gain vectors $g_t := w_t(s)\, Q_t(s, \cdot)$ and comparator $q := \pi'(\cdot \mid s)$. Since $Q_t(s,a) \in [0, \frac{1}{1-\gamma}]$ and $w_t(s) \geq 0$, the gains lie in $[0, G_t]$ with $G_t = w_t(s)/(1-\gamma)$. The update equation 3 is exactly the Hedge update on these gains, so

$$\sum_{t=1}^{T} w_t(s)\, \Delta_t(s, \pi') \;=\; \sum_{t=1}^{T} \big\langle \pi'(\cdot \mid s) - \pi_t(\cdot \mid s),\, g_t(s, \cdot) \big\rangle \;\leq\; \frac{\ln A}{\eta} + \frac{\eta}{8(1-\gamma)^2} \sum_{t=1}^{T} w_t(s)^2. \tag{12}$$

Crucially, equation 12 holds for every s simultaneously and deterministically on the realised trajectory, whatever be the sequence (w_t, Q_t).

Summing equation 12 over $s \in \mathcal{S}$,

$$\sum_{t=1}^{T} \sum_{s} w_t(s) \Delta_t(s, \pi') \;\leq\; \frac{S \ln A}{\eta} + \frac{\eta}{8(1-\gamma)^2} \sum_{t=1}^{T} \sum_{s} w_t(s)^2.$$

Since $w_t(\cdot)$ is a probability distribution over $\mathcal{S}$, $\sum_s w_t(s)^2 \leq \big(\max_s w_t(s)\big) \sum_s w_t(s) \leq 1$, so the second sum is at most T, giving the cumulative bound

$$\sum_{t=1}^{T} \sum_{s} w_t(s) \Delta_t(s, \pi') \;\leq\; \frac{S \ln A}{\eta} + \frac{\eta\, T}{8(1-\gamma)^2}. \tag{13}$$

Dividing equation 13 by T,

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{s} w_t(s) \Delta_t(s, \pi') \;\leq\; \frac{S \ln A}{\eta T} + \frac{\eta}{8(1-\gamma)^2}.$$

By construction of $\hat{\mu}$ as the uniform distribution over the realised iterates, the left-hand side equals $\mathbb{E}_{\pi \sim \hat{\mu}}\big[\sum_s d^{\pi}_{P_\pi}(s) \Delta_\pi(s, \pi')\big]$. The bound holds for every $\pi' \in \Pi$ and its right-hand side does not depend on π', so it holds for the maximum over π'. Multiplying by $\frac{1}{1-\gamma}$ gives the first result.

For the second, substitute $\eta = 2(1-\gamma)\sqrt{2S \ln A / T}$:

$$\frac{S \ln A}{\eta T} = \frac{1}{2\sqrt{2}} \cdot \frac{\sqrt{S \ln A}}{(1-\gamma)\sqrt{T}}, \qquad \frac{\eta}{8(1-\gamma)^2} = \frac{\sqrt{2}}{4} \cdot \frac{\sqrt{S \ln A}}{(1-\gamma)\sqrt{T}},$$

and $\frac{1}{2\sqrt{2}} + \frac{\sqrt{2}}{4} = \frac{1}{\sqrt{2}}$, giving $\mathrm{Gap}^{\mathrm{loc}}(\hat{\mu}) \leq \frac{1}{(1-\gamma)^2}\sqrt{S \ln A / (2T)}$. □

Lemma 7 (Hedge with time-varying gain ranges). *Fix a state s. Let $p_1 = \mathrm{Unif}(\mathcal{A})$ and, for gain vectors $g_t \in [0, G_t]^{\mathcal{A}}$, let $p_{t+1}(a) \propto p_t(a)\exp(\eta\, g_t(a))$ with $\eta > 0$. Then for every $q \in \Delta(\mathcal{A})$,*

$$\sum_{t=1}^{T} \langle q - p_t,\ g_t \rangle \ \le\ \frac{\ln A}{\eta} \ +\ \frac{\eta}{8}\sum_{t=1}^{T} G_t^2.$$

Proof. Let us define $Z_t := \sum_a p_1(a)\exp\left(\eta \sum_{\tau<t} g_\tau(a)\right)$, so $Z_1 = 1$ and $p_t(a) = p_1(a)\exp(\eta\sum_{\tau<t} g_\tau(a))/Z_t$. Then

$$\ln \frac{Z_{t+1}}{Z_t} = \ln \mathbb{E}_{a\sim p_t}\left[e^{\eta g_t(a)}\right].$$

The random variable $\eta g_t(a)$, $a \sim p_t$, takes values in an interval of length at most ηG_t, so Hoeffding's lemma gives $\ln \mathbb{E}[e^{\eta g_t}] \le \eta\langle p_t, g_t\rangle + \eta^2 G_t^2/8$. Summing over $t = 1, \dots, T$ and telescoping,

$$\ln Z_{T+1} = \sum_{t=1}^{T} \ln \frac{Z_{t+1}}{Z_t} = \sum_{t=1}^{T} \ln \mathbb{E}_{a\sim p_t}\left[e^{\eta g_t(a)}\right] \ \le\ \eta \sum_{t=1}^{T}\langle p_t, g_t\rangle + \frac{\eta^2}{8}\sum_{t=1}^{T} G_t^2.$$

For the lower bound, for any $q \in \Delta(\mathcal{A})$, we can apply Jensen's inequality to the convex function $-\ln$.

$$\ln Z_{T+1} = \ln \sum_a p_1(a) e^{\eta\sum_t g_t(a)} \ \ge\ \sum_a q(a) \ln \frac{p_1(a) e^{\eta\sum_t g_t(a)}}{q(a)} = \eta \sum_t \langle q, g_t\rangle - \mathrm{KL}(q\,\|\,p_1),$$

and $\mathrm{KL}(q\|p_1) \le \ln A$ since p_1 is uniform. Combining the above two displayed equations and dividing by η gives the claim. □

E.5 Finite Sample Analysis of Local Stability

Proof of Theorem 2. The proof immediately follows from Theorem 8 if we substitute $n = k = \Theta(1)$. □

Theorem 8 (Finite-sample local stability). *Suppose Algorithm 1 is run for $T \ge \frac{S\ln(SA/\delta)}{2/k+1/n}$ iterations, and let $\pi_1, \dots, \pi_T$ be generated by equation 4 and $\hat{\mu} := \mathrm{Unif}\{\pi_1, \dots, \pi_T\}$. Then for any $\eta > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Gap}^{\mathrm{loc}}(\hat{\mu}) \ \le\ \frac{S\ln A}{\eta T} + \frac{\eta}{8(1-\gamma)^2} + \frac{4}{(1-\gamma)^2}\sqrt{\frac{(2/k+1/n)S\ln(SA/\delta)}{T}},$$

and with $\eta = 2(1-\gamma)\sqrt{2S\ln A/T}$, $\mathrm{Gap}^{\mathrm{loc}}(\hat{\mu}) \le \frac{6}{(1-\gamma)^2}\sqrt{\frac{(2/k+1/n)SA\ln(SA/\delta)}{T}}$. The total number of value and occupancy samples required by the algorithm is $Tn(1+kA)$.

Proof of Theorem 8. Fix $\pi' \in \Pi$ and let us write $\xi_t(s,a) := g_t(s,a) - \hat{g}_t(s,a)$. Then we can establish the following decomposition

$$\begin{aligned}\sum_t\sum_s w_t(s)\Delta_t(s,\pi') &= \sum_t\sum_s \langle \pi'(\cdot|s) - \pi_t(\cdot|s), \hat{g}_t(s,\cdot)\rangle \\ &= \underbrace{\sum_t\sum_s \left\langle \pi'(\cdot|s) - \pi_t(\cdot|s),\ \hat{g}_t(s,\cdot)\right\rangle}_{\text{(I)}} + \underbrace{\sum_t\sum_s \left\langle \pi'(\cdot|s) - \pi_t(\cdot|s),\ \xi_t(s,\cdot)\right\rangle}_{\text{(II)}}.\end{aligned} \tag{14}$$

We now bound the two terms separately. Let us also write $\mathcal{F}_t$ to denote the σ-algebra generated by M_t, π_t or equivalently, the trajectories drawn at iteration t.

Term (I). The empirical distribution satisfies $\sum_s \widehat{w}_t(s)^2 \le 1$. Moreover, the estimate of $Q_t(s,a)$ is bounded between 0 and $1/(1-\gamma)$ as it is $1/(1-\gamma)$ times the reward at the τ-th step. Therefore, the proof of Theorem 1 applies verbatim with w_t replaced by $\widehat{w}_t$ giving

$$\text{(I)} \le \frac{S\ln A}{\eta} + \frac{\eta T}{8(1-\gamma)^2}$$

as in equation 13.

Term (II). Let us write $\Xi(s,a) := \sum_t \xi_t(s,a)$ and $Y_T := \sum_t \sum_s \langle \pi_t(\cdot \mid s), \xi_t(s,\cdot)\rangle$. Since the constraint set is the product $\Delta(\mathcal{A})^{\mathcal{S}}$,

$$\begin{aligned}\max_{\pi' \in \Pi} (\text{II}) &= \max_{\pi' \in \Pi} \sum_s \sum_t \langle \pi'(\cdot|s) - \pi_t(\cdot|s),\ \xi_t(s,\cdot)\rangle \\ &= \max_{\pi' \in \Pi} \sum_s \sum_a \pi'(a|s)\Xi(s,a) - Y_T \\ &= \sum_s \max_a \Xi(s,a) - Y_T.\end{aligned}$$

We now bound the two terms separately.

Bounding $\sum_s \max_a \Xi(s,a)$: Using Lemma 10 we have the following inequality for any $\lambda \in [0, (1-\gamma)]$.

$$\Pr\left(\forall s,a, \Xi(s,a) \le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t w_t(s) + \frac{\ln(SA/\delta)}{\lambda}\right) \ge 1-\delta.$$

This gives us,

$$\begin{aligned}\sum_s \max_a \Xi(s,a) &\le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t \sum_s w_t(s) + \frac{S\ln(SA/\delta)}{\lambda} \\ &\le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)T + \frac{S\ln(SA/\delta)}{\lambda}\end{aligned}$$

Substituting $\lambda = (1-\gamma)\sqrt{\frac{S\ln(SA/\delta)}{(e-2)T(2/k+1/n)}}$ we obtain the following upper bound.

$$\sum_s \max_a \Xi(s,a) \le \frac{2}{1-\gamma}\sqrt{(e-2)\left(\frac{2}{k}+\frac{1}{n}\right)TS\ln(SA/\delta)} \tag{15}$$

Note that, as long as $T \ge \frac{S\ln(SA/\delta)}{(e-2)(2/k+1/n)}$ we are guaranteed that $\lambda \in [0,(1-\gamma)]$ which is required to apply Freedman's inequality.

Bounding $-Y_T$: Let us define $y_{t,s} = -\langle \pi_t(\cdot|s), \xi_t(s,\cdot)\rangle$. Then $-Y_t = \sum_s \sum_t y_{t,s}$. We can apply Lemma 11 to obtain the following inequality for any $\lambda \in [0,(1-\gamma)]$.

$$\Pr\left(\forall s, \sum_t y_{t,s} \le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t w_t(s) + \frac{\ln(S/\delta)}{\lambda}\right) \ge 1-\delta.$$

This gives us,

$$\begin{aligned}-Y_T = \sum_s \sum_t y_{t,s} &\le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t \sum_s w_t(s) + \frac{S\ln(S/\delta)}{\lambda} \\ &\le \frac{(e-2)\lambda}{(1-\gamma)^2}\left(\frac{2}{k}+\frac{1}{n}\right)T + \frac{S\ln(S/\delta)}{\lambda}\end{aligned}$$

Substituting $\lambda = (1-\gamma)\sqrt{\frac{S\ln(SA/\delta)}{(e-2)T(2/k+1/n)}}$ we obtain the following upper bound.

$$-Y_T \le \frac{2}{1-\gamma}\sqrt{(e-2)\left(\frac{2}{k}+\frac{1}{n}\right)TS\ln(S/\delta)} \tag{16}$$

Now using Equation (15) and Equation (16) we obtain the following upper bound on part (II).

$$\max_{\pi'}(II) \le \frac{4}{1-\gamma}\sqrt{(e-2)\left(\frac{2}{k}+\frac{1}{n}\right)TS\ln(SA/\delta)}$$

Recall the definition of $\mathrm{Gap}^{\mathrm{loc}}(\widehat{\mu})$.

$$\begin{aligned}\mathrm{Gap}^{\mathrm{loc}}(\widehat{\mu}) &= \frac{1}{(1-\gamma)T}\sum_t\sum_s w_t(s)\Delta_t(s,\pi') \\ &\le \frac{S\ln A}{\eta T} + \frac{\eta}{8(1-\gamma)^2} + \frac{4}{(1-\gamma)^2}\sqrt{\frac{(2/k+1/n)S\ln(SA/\delta)}{T}}\end{aligned}$$

The bound on the sample complexity follows from the observation that iteration t collects n occupancy samples and $kA\,|\mathrm{supp}(\widehat{w}_t)| \le nkA$ value samples. □

Lemma 8 (Estimator properties)**.** *Conditioned on π_t and M_t, an occupancy sample is distributed as $d_{P_t}^{\pi_t}$, and a value sample at (s,a) is an unbiased estimate of $Q_t(s,a)$ taking values in $[0, \frac{1}{1-\gamma}]$.*

Proof. For the first, $\Pr[s_\tau = s] = \sum_j (1-\gamma)\gamma^j \Pr[s_j = s] = d_{P_t}^{\pi_t}(s)$ by definition of the normalised occupancy.

For the second, $\mathbb{E}[\frac{r_\tau}{1-\gamma}] = \frac{1}{1-\gamma}\sum_j (1-\gamma)\gamma^j\,\mathbb{E}[r_j \mid s_0 = s, a_0 = a] = Q_{M_t}^{\pi_t}(s,a)$, and $r_\tau \in [0,1]$ bounds the range. □

Lemma 9 (Freedman's inequality, Beygelzimer et al. (2011))**.** *Let $\{X_t\}_{t=1}^T$ be a martingale difference sequence measurable with respect to the sequence of σ-algebras satisfying $X_t \le R$ a.s. for all $t \in [T]$. Then for any $\lambda \in (0, 1/R]$ and $\delta \in (0,1)$*

$$\Pr\left(\sum_{t=1}^T X_t \le (e-2)\lambda\sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] + \frac{\ln(1/\delta)}{\lambda}\right) \ge 1-\delta.$$

Lemma 10. *As in the proof of Theorem 8, let $\xi_t(s,a) = w_t(s)Q_t(s,a) - \widehat{w}_t(s)\widehat{Q}_t(s,a)$. Then $\xi_t(s,a)$ is a martingale difference sequence with respect to $\mathcal{F}_{t-1}$ and for any $\lambda \in [0,(1-\gamma)]$*

$$\Pr\left(\forall s,a,\ \sum_t \xi_t(s,a) \le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t w_t(s) + \frac{\ln(SA/\delta)}{\lambda}\right) \ge 1-\delta.$$

Proof. First, we show that $\widehat{g}_t$ is an unbiased estimator for g_t. Let us write b_t to denote the occupancy batch collected during iteration t. Conditioned on the batch b_t, the estimator $\widehat{w}_t$ and $\mathrm{supp}(\widehat{w}_t)$ are determined. The value batch uses independent samples, and therefore by Lemma 8, $\mathbb{E}[\widehat{Q}_t(s,a) \mid \mathcal{F}_{t-1}, b_t] = Q_t(s,a)$ for every $s \in \mathrm{supp}(\hat{w}_t)$. Since conditioned on the occupancy batch b_t, the estimators $\widehat{w}_t$ and $\widehat{Q}_t$ are independent, we have

$$\begin{aligned}\mathbb{E}[\widehat{w}_t(s)\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\mathbb{E}_{b_t}[\widehat{w}_t(s)\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}, b_t]\right] \\ &= \mathbb{E}\left[\mathbb{E}_{b_t}[\widehat{w}_t(s)|\mathcal{F}_{t-1}, b_t]\cdot\mathbb{E}_{b_t}[\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}, b_t]\right] \\ &= \mathbb{E}[\widehat{w}_t(s)|\mathcal{F}_{t-1}]Q_t(s,a)\end{aligned} \tag{17}$$

for all $s \in \mathrm{supp}(\widehat{w}_t)$. The claim also holds trivially for $s \notin \mathrm{supp}(\widehat{w}_t)$. Now taking expectation over b_t and using the observation $\mathbb{E}[\widehat{w}_t(s)|\mathcal{F}_{t-1}] = w_t(s)$ we obtain

$$\mathbb{E}[\widehat{g}_t(s,a)|\mathcal{F}_{t-1}] = \mathbb{E}[\widehat{w}_t(s)\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}] = w_t(s)Q_t(s,a) = g_t(s,a)$$

Therefore, $\mathbb{E}[\xi_t(s,a)|\mathcal{F}_{t-1}] = \mathbb{E}[\widehat{g}_t(s,a) - g_t(s,a)|\mathcal{F}_{t-1}] = 0$ and $\{\xi_t(s,a)\}_{t=1}^T$ is a Martingale difference sequence. We now bound $\sum_t \xi_t(s,a)$ using Freedman's inequality Lemma 9.

$$\begin{aligned}&\mathbb{E}[\xi_t^2(s,a)|\mathcal{F}_{t-1}] = \mathrm{Var}(\widehat{g}_t(s,a)|\mathcal{F}_{t-1}) = \mathrm{Var}(\widehat{w}_t(s,a)\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}) \\ =&\mathbb{E}\left[\widehat{w}_t(s,a)^2|\mathcal{F}_{t-1}\right]\mathbb{E}\left[\widehat{Q}_t(s,a)^2|\mathcal{F}_{t-1}\right] - \left(\mathbb{E}\left[\widehat{w}_t(s,a)|\mathcal{F}_{t-1}\right]\right)^2\left(\mathbb{E}\left[\widehat{Q}_t(s,a)^2|\mathcal{F}_{t-1}\right]\right)^2\end{aligned}$$

The last line follows the same argument as displayed in Equation (17). Now observe, that $\widehat{w}_t(s)$ is the average of n Bernoulli trials with success probability $w_t(s)$. Therefore, $\mathbb{E}[\widehat{w}_t(s)^2] = w_t(s)(1 -$

$w_t(s))/n + w_t^2(s)$. This gives us the following upper bound.

$$\begin{aligned}\mathbb{E}[\xi_t^2(s,a)|\mathcal{F}_{t-1}] &= \left(\frac{w_t(s)(1-w_t(s))}{n} + w_t(s)^2\right)\left(Q_t(s,a)^2 + \text{Var}\left(\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}\right)\right) - w_t(s)^2 Q_t(s,a)^2 \\ &= \left(\frac{w_t(s)(1-w_t(s))}{n} + w_t(s)^2\right)\cdot \text{Var}\left(\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}\right) + \frac{w_t(s)(1-w_t(s))}{n}\cdot Q_t(s,a)^2\end{aligned}$$

We now use $w_t(s) \in [0,1]$, and $Q_t(s,a) \in [0, 1/(1-\gamma)]$. Moreover, $\widehat{Q}_t(s,a)$ is the sample average of k independent random variables bounded between 0 and $1/(1-\gamma)$, giving us $\text{Var}\left(\widehat{Q}_t(s,a)|\mathcal{F}_{t-1}\right) \le 1/k \cdot 1/(1-\gamma)^2$.

$$\begin{aligned}\mathbb{E}[\xi_t^2(s,a)|\mathcal{F}_{t-1}] &\le \left(\frac{w_t(s)}{n} + w_t(s)^2\right)\cdot\frac{1}{k(1-\gamma)^2} + \frac{w_t(s)}{n}\cdot\frac{1}{(1-\gamma)^2} \\ &\le \frac{2w_t(s)}{k}\cdot\frac{1}{(1-\gamma)^2} + \frac{w_t(s)}{n}\cdot\frac{1}{(1-\gamma)^2}.\end{aligned} \tag{18}$$

We can apply Lemma 9 and union bound over the SA tuples to obtain the following inequality for any $\lambda \in [0, (1-\gamma)]$.

$$\Pr\left(\forall s,a \sum_t \xi_t(s,a) \le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t w_t(s) + \frac{\ln(SA/\delta)}{\lambda}\right) \ge 1-\delta. \tag{19}$$

□

Lemma 11. *As in the proof of Theorem 8, let* $y_{t,s} = -\langle \pi_t(\cdot|s), \xi_t(s,\cdot)\rangle$ *with* $\xi_t(s,a) = w_t(s)Q_t(s,a) - \widehat{w}_t(s)\widehat{Q}_t(s,a)$. *Then* $y_{t,s}$ *is a martingale difference sequence with respect to* $\mathcal{F}_{t-1}$ *and for any* $\lambda \in [0,(1-\gamma)]$

$$\Pr\left(\forall s, \sum_t y_{t,s} \le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t w_t(s) + \frac{\ln(S/\delta)}{\lambda}\right) \ge 1-\delta.$$

Proof. First note that,

$$\begin{aligned}\mathbb{E}[y_{t,s}|\mathcal{F}_{t-1}] &= -\sum_a \mathbb{E}\left[\pi_t(a|s)\xi_t(s,a)|\mathcal{F}_{t-1}\right] \\ &= -\sum_a \mathbb{E}_{\pi_t}\left[\pi_t(a|s)\mathbb{E}[\xi_t(s,a)|\mathcal{F}_{t-1},\pi_t]\right] = 0.\end{aligned}$$

The last line follows since conditioned on the policy π_t, the mean of $\widehat{g}_t(s,a) = \widehat{w}_t(s)\widehat{Q}_t(s,a)$ equals $g_t(s,a) = w_t(s)Q_t(s,a)$ and $\mathbb{E}[\xi_t(s,a)|\mathcal{F}_{t-1},\pi_t] = 0$. Therefore, $y_{t,s}$ is a martingale difference sequence. Moreover, from the definition of $\xi_t(s,a)$ we have, $|y_{t,s}| \le \|\pi_t(\cdot|s)\|_1 \max_a \xi_t(s,a) \le 1/(1-\gamma)$.

$$\begin{aligned}\mathbb{E}[y_{t,s}^2|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\left(\sum_a \pi_t(a|s)\xi_t(s,a)\right)^2\middle|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\sum_a \pi_t(a|s)^2\xi_t(s,a)^2\middle|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\sum_{a,b}\pi_t(a|s)\pi_t(b|s)\xi_t(s,a)\xi_t(s,b)\middle|\mathcal{F}_{t-1}\right] \\ &\le \sum_a \pi_t(a|s)\mathbb{E}[\xi_t(s,a)^2|\mathcal{F}_{t-1}]\end{aligned}$$

The last line follows from the observation that conditioned on $\mathcal{F}_{t-1}$ and occupancy batch samples, $\xi_t(s,a)$ and $\xi_t(s,b)$ as they are estimated from different sets of trajectories. Therefore, $\mathbb{E}[\xi_t(s,a)\xi_t(s,b)|\mathcal{F}_{t-1},b_t] = \mathbb{E}[\xi_t(s,a)|\mathcal{F}_{t-1},b_t]\mathbb{E}[\xi_t(s,b)|\mathcal{F}_{t-1},b_t] = 0$. Now using Equation (18) we obtain,

$$\mathbb{E}[y_{t,s}^2|\mathcal{F}_{t-1}] \le \frac{2w_t(s)}{k}\cdot\frac{1}{(1-\gamma)^2} + \frac{w_t(s)}{n}\cdot\frac{1}{(1-\gamma)^2}$$

We can now apply Lemma 9 and union bound over S states to obtain the following inequality for any $\lambda \in [0, (1-\gamma)]$.

$$\Pr\left(\forall s,\ \sum_t y_{t,s} \le (e-2)\lambda\left(\frac{2}{k(1-\gamma)^2} + \frac{1}{n(1-\gamma)^2}\right)\sum_t w_t(s) + \frac{\ln(S/\delta)}{\lambda}\right) \ge 1-\delta.$$

□

E.6 Global Stability Under Bounded Transition Range

Proof of Theorem 3. Fix $\pi' \in \Pi$. Applying Lemma 7 at each state s with $g_t = Q_t(s,\cdot) \in [0, \frac{1}{1-\gamma}]^A$, i.e. $G_t \equiv \frac{1}{1-\gamma}$, gives the *uniform* per-state regret bound for every $s \in \mathcal{S}$.

$$\sum_{t=1}^{T} \Delta_t(s,\pi') \ \le\ \frac{\ln A}{\eta} + \frac{\eta T}{8(1-\gamma)^2} \ =:\ R_T \tag{20}$$

and with $\eta = 2(1-\gamma)\sqrt{2\ln A/T}$ one gets, exactly as at the end of the proof of Theorem 1, $R_T = \frac{1}{1-\gamma}\sqrt{T\ln A/2}$.

By Lemma 1 applied in $M_t = (P_t, r_t)$,

$$\frac{1}{T}\sum_{t=1}^{T}\left[V_{M_t}(\pi') - V_{M_t}(\pi_t)\right] = \frac{1}{(1-\gamma)T}\sum_{t=1}^{T}\sum_s d^{\pi'}_{P_t}(s)\,\Delta_t(s,\pi'). \tag{21}$$

Write $d^{\pi'}_{P_t}(s) = d^{\pi'}_{\bar P}(s) + e_t(s)$. By Assumption 1 and Lemma 12,

$$\|e_t\|_1 \le \frac{\gamma\epsilon_P}{1-\gamma} \qquad \text{for every } t. \tag{22}$$

We now replace $d^{\pi'}_{P_t}(s)$ in Equation (21) by $d^{\pi'}_{\bar P}(s) + e_t(s)$ and analyze the two terms separately.

Main term. The weights $d^{\pi'}_{\bar P}(s)$ are nonnegative, and sum to one. Moreover, $\bar P$ is a single fixed kernel. Therefore we may exchange the order of summation and apply equation 20 state-wise.

$$\sum_{t=1}^{T}\sum_s d^{\pi'}_{\bar P}(s)\Delta_t(s,\pi') = \sum_s d^{\pi'}_{\bar P}(s)\sum_{t=1}^{T}\Delta_t(s,\pi') \le \sum_s d^{\pi'}_{\bar P}(s)\,R_T = R_T.$$

Perturbation term. Since $Q_t(s,a) \in [0, 1/(1-\gamma)]$, $\Delta_t(s,\pi') = \langle \pi'(\cdot|s) - \pi(\cdot|s), Q_t(s,\cdot)\rangle \in [0, 1/(1-\gamma)]$ for any state s. By Hölder's inequality, and equation 22,

$$\Big|\sum_{t=1}^{T}\sum_s e_t(s)\Delta_t(s,\pi')\Big| \le \sum_{t=1}^{T}\|e_t\|_1 \max_s|\Delta_t(s,\pi')| \le T\cdot\frac{\gamma\epsilon_P}{1-\gamma}\cdot\frac{1}{1-\gamma}.$$

Combining. Dividing by $(1-\gamma)T$,

$$\frac{1}{T}\sum_{t=1}^{T}\left[V_{M_t}(\pi') - V_{M_t}(\pi_t)\right] \ \le\ \frac{R_T}{(1-\gamma)T} + \frac{\gamma\epsilon_P}{(1-\gamma)^3} \ =\ \frac{1}{(1-\gamma)^2}\sqrt{\frac{\ln A}{2T}} + \frac{\gamma\epsilon_P}{(1-\gamma)^3}.$$

The left-hand side equals $\mathbb{E}_{\pi\sim\hat\mu}[V_{M(\pi)}(\pi')] - \mathbb{E}_{\pi\sim\hat\mu}[V_{M(\pi)}(\pi)]$, the right-hand side is independent of π', and π' was arbitrary; taking the maximum over π' gives the claim. □

Lemma 12 (Occupancy simulation lemma)**.** *Let $P, \bar P$ be transition kernels with $\max_{s,a}\|P(\cdot|s,a) - \bar P(\cdot|s,a)\|_1 \le \epsilon$. Then for every $\pi' \in \Pi$,*

$$\left\|d^{\pi'}_P - d^{\pi'}_{\bar P}\right\|_1 \ \le\ \frac{\gamma\,\epsilon}{1-\gamma},$$

where the norm is over state–action pairs; the same bound holds for the state marginals.

Proof. For a kernel P define the affine operator T_P on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ by

$$(T_P d)(s,a) \;=\; \pi'(a \mid s)\Big[(1-\gamma)\rho(s) + \gamma \sum_{s',a'} P(s \mid s',a')\, d(s',a')\Big].$$

The Bellman flow constraints state exactly that $d_P^{\pi'}$ is the unique fixed point of T_P in the simplex.

Contraction. For d, d' in the simplex,

$$\begin{aligned}\|T_P d - T_P d'\|_1 &= \gamma \sum_{s,a} \pi'(a \mid s) \Big| \sum_{s',a'} P(s \mid s',a')\,(d-d')(s',a') \Big| \\ &\le \gamma \sum_{s} \sum_{s',a'} P(s \mid s',a') \,\big|(d-d')(s',a')\big| \;=\; \gamma \|d-d'\|_1,\end{aligned}$$

using $\sum_a \pi'(a \mid s) = 1$ and then exchanging the order of summation.

Perturbation. For any d in the simplex,

$$\begin{aligned}\|T_P d - T_{\bar P} d\|_1 &= \gamma \sum_{s} \Big| \sum_{s',a'} \big(P - \bar P\big)(s \mid s',a')\, d(s',a') \Big| \\ &\le \gamma \sum_{s',a'} d(s',a') \,\big\|P(\cdot \mid s',a') - \bar P(\cdot \mid s',a')\big\|_1 \;\le\; \gamma\epsilon.\end{aligned}$$

Conclusion. Writing $d := d_P^{\pi'}$ and $\bar d := d_{\bar P}^{\pi'}$,

$$\|d - \bar d\|_1 = \|T_P d - T_{\bar P}\bar d\|_1 \le \|T_P d - T_P \bar d\|_1 + \|T_P \bar d - T_{\bar P}\bar d\|_1 \le \gamma \|d - \bar d\|_1 + \gamma\epsilon,$$

so $\|d-\bar d\|_1 \le \gamma\epsilon/(1-\gamma)$. The marginal bound follows since summing over actions is a contraction in ℓ_1. □

E.7 Lower bound under Local Feedback

Proof of Theorem 6. We will consider a slightly stronger interaction model than specified in Definition 12. During round t, the algorithm draws $m_t \ge 0$ trajectories. Trajectory (t,i) selects an action $A_{t,i} \in [K]$, and returns $X_{t,i} \sim \mathrm{Bern}\big(p_{A_{t,i}}(\pi_t)\big)$, the indicator of absorption in $\top$. The total budget satisfies $\sum_t m_t \le N$ almost surely, and the algorithm outputs $\hat\mu := \mathrm{Unif}\{\pi_1, \dots, \pi_T\}$. Note that, letting the algorithm choose $A_{t,i}$, rather than sampling it from π_t, only strengthens the oracle and hence the lower bound

For ease of notation, let us write $X := \gamma\epsilon/(1-\gamma)$ and $G_j(\mu) := \mathrm{Gap}_j^{\mathrm{glob}}(\mu)$ the global stability gap on the j-th hidden action instance. We will need three definitions: (a) $q_j(\mu) := \mu\big(\{\pi : y_\pi^j < \frac12\}\big)$, the probability that π underweights the hidden action j; (b) $\sigma_\pi^j := +1$ if $y_\pi^j < \frac12$ and -1 otherwise; and (c) $y_\pi^j := \pi(j \mid s_0)$, the probability of playing the hidden action j under policy π.

By Lemma 13, $G_j(\mu) = X\big[(2q_j(\mu)-1)_+ - \mathbb{E}_\mu[y_\pi^j \sigma_\pi^j]\big]$, so $G_j(\mu)/X \in [-1,2]$ and in particular $|G_j| \le 2X$ for every μ and every j. For the empirical mixture, $q_j(\hat\mu) = \frac1T \big\{t : y_{\pi_t}^j < \frac12\big\}$ and $\mathbb{E}_{\hat\mu}[y^j\sigma^j] = \frac1T \sum_t y_{\pi_t}^j \sigma_{\pi_t}^j$ are measurable functions of $(\pi_1, \dots, \pi_T)$, and $G_j(\hat\mu)$ is a bounded random variable. We will also write $\mathcal{D}$ to denote the dataset of all the trajectories drawn by the algorithm, $\mathbb{P}_j$ to denote the law of $\mathcal{D}$ under the j-th hidden action instance, and $\mathbb{P}_0$ to denote the law of $\mathcal{D}$ under the null instance in which $p_a \equiv \frac12$ for every a.

The null instance. Lemma 14 holds for *every* $\mu \in \Delta(\Pi)$, hence pathwise for the random $\hat\mu$. Taking expectation under $\mathbb{P}_0$ and exchanging the finite sum with the expectation,

$$\frac1K \sum_{j=1}^K \mathbb{E}_{\mathcal{D}\sim\mathbb{P}_0}\big[G_j(\hat\mu(\mathcal{D}))\big] \;\ge\; X\Big(1 - \frac5K\Big). \tag{23}$$

Transferring to $\mathbb{P}_j$. For a random variable Z with $|Z| \le M$ and two laws P, Q on the same space, $|\mathbb{E}_P Z - \mathbb{E}_Q Z| \le 2M\|P-Q\|_{\mathrm{TV}}$. Applying this with $Z = G_j(\hat\mu)$ and $M = 2X$,

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}_j}\big[G_j(\hat\mu(\mathcal{D}))\big] \;\ge\; \mathbb{E}_{\mathcal{D}\sim\mathbb{P}_0}\big[G_j(\hat\mu(\mathcal{D}))\big] - 4X\,\big\|\mathbb{P}_j - \mathbb{P}_0\big\|_{\mathrm{TV}}.$$

Bounding the total variation on average. By Pinsker's inequality and Lemma 15, $\|\mathbb{P}_j - \mathbb{P}_0\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_j)} \le \sqrt{\frac{3}{2}\epsilon^2\,\mathbb{E}_0[N_j]}$. Averaging over j and using the concavity of the square root (Jensen) together with $\sum_j \mathbb{E}_0[N_j] \le N$,

$$\frac{1}{K}\sum_{j=1}^{K}\|\mathbb{P}_j - \mathbb{P}_0\|_{\mathrm{TV}} \;\le\; \sqrt{\frac{3\epsilon^2}{2}\cdot\frac{1}{K}\sum_j \mathbb{E}_0[N_j]} \;\le\; \sqrt{\frac{3\,\epsilon^2\,N}{2K}}. \tag{24}$$

From equation 23 and equation 24,

$$\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}_{\mathcal{D}\sim\mathbb{P}_j}\big[G_j(\hat{\mu}(\mathcal{D}))\big] \;\ge\; X\left(1 - \frac{5}{K} - 4\sqrt{\frac{3\epsilon^2 N}{2K}}\right).$$

Since $\epsilon = \epsilon_P/2$, the hypothesis $N \le K/(24\epsilon_P^2)$ reads $N \le K/(96\epsilon^2)$, which implies $\frac{3\epsilon^2 N}{2K} \le \frac{1}{64}$ and $4\sqrt{\frac{3\epsilon^2 N}{2K}} \le \frac{1}{2}$; and $K \ge 20$ gives $5/K \le \frac{1}{4}$. This gives a lower bound of at least $\frac{1}{4}$, and $X/4 = \gamma\epsilon_P/(8(1-\gamma))$. A maximum dominates an average, giving the final claim. □

Lemma 13. *Write $y^j_\pi := \pi(j \mid s_0)$, $\sigma^j_\pi := +1$ if $y^j_\pi < \frac{1}{2}$ and -1 otherwise, and $q_j(\mu) := \mu\big(\{\pi : y^j_\pi < \frac{1}{2}\}\big)$. Then for every $\mu \in \Delta(\Pi)$,*

$$\mathrm{Gap}^{\mathrm{glob}}(\mu) \;=\; \frac{\gamma\,\epsilon}{1-\gamma}\Big[\big(2q_j(\mu) - 1\big)_+ \;-\; \mathbb{E}_{\pi\sim\mu}\big[y^j_\pi\sigma^j_\pi\big]\Big].$$

Proof. For any π' and any realised model M, $V_M(\pi') = \frac{\gamma}{1-\gamma}\sum_a \pi'(a \mid s_0)p_a$, since reaching $\top$ at step 1 yields the discounted unit stream. As $p_a = \frac{1}{2}$ for $a \ne j$, this equals

$$V_M(\pi') = \frac{\gamma}{1-\gamma}\big[(1 - y^j_{\pi'})\tfrac{1}{2} + p_j \cdot y^j_{\pi'}\big] = \frac{\gamma}{1-\gamma}\big[\tfrac{1}{2} + y^j_{\pi'}(\epsilon - \tfrac{1}{2})\big].$$

Taking $\pi' = \pi$ and averaging gives $\mathbb{E}_\mu[V_{M(\pi)}(\pi)] = \frac{\gamma}{1-\gamma}\big[\frac{1}{2} + \epsilon\,\mathbb{E}_\mu[y^j_\pi\sigma^j_\pi]\big]$. For the comparator policy, $\mathbb{E}_\mu[p_j(\pi)] - \frac{1}{2} = \epsilon(2q_j(\mu) - 1)$, so $\mathbb{E}_\mu[V_{M(\pi)}(\pi')] = \frac{\gamma}{1-\gamma}\big[\frac{1}{2} + y^j_{\pi'}\epsilon(2q_j(\mu) - 1)\big]$. This is is linear in $y^j_{\pi'} \in [0,1]$ and is maximum value is exactly $\frac{1}{2} + \epsilon(2q_j(\mu) - 1)_+$. Subtracting gives the claim. □

Lemma 14 (Counting)**.** *For every $\mu \in \Delta(\Pi)$,*

$$\frac{1}{K}\sum_{j=1}^{K}\mathrm{Gap}^{\mathrm{glob}}_j(\mu) \;\ge\; \frac{\gamma\,\epsilon}{1-\gamma}\Big(1 - \frac{5}{K}\Big),$$

where $\mathrm{Gap}^{\mathrm{glob}}_j$ is the global stability gap in the instance with hidden action j.

Proof. Lemma 13 establishes the following bound

$$\mathrm{Gap}^{\mathrm{glob}}_j(\mu) = \tfrac{\gamma\,\epsilon}{1-\gamma}\big[(2q_j - 1)_+ - \mathbb{E}_\mu[y^j\sigma^j]\big].$$

Since $y\sigma \le y$ always, $\mathbb{E}_\mu[y^j_\pi\sigma^j_\pi] \le \mathbb{E}_\mu[y^j_\pi]$, and $\frac{1}{K}\sum_j \mathbb{E}_\mu[y^j_\pi\sigma^j_\pi] = \mathbb{E}_\mu\big[\frac{1}{K}\sum_j \pi(j \mid s_0)\big] = \frac{1}{K}$. For the first term, each policy can have $y^j_\pi \ge \frac{1}{2}$ for at most two indices j, so

$$\sum_j (1 - q_j(\mu)) = \sum_j \mu\big(\{\pi : y^j_\pi \ge \tfrac{1}{2}\}\big) \le 2.$$

This gives us $\frac{1}{K}\sum_j(2q_j - 1) \ge 1 - \frac{4}{K}$. Combining with $(x)_+ \ge x$ and Lemma 13 gives the average bound. □

Lemma 15 (Divergence decomposition)**.** *Let $\mathbb{P}_j$ be the law of the algorithm's trajectory of observations under the instance of Definition 13 with hidden action j, and $\mathbb{P}_0$ the law under the null instance in which $p_a \equiv \frac{1}{2}$ for every a. If $\epsilon \le 1/4$ then for every $j \in [K]$,*

$$\mathrm{KL}\big(\mathbb{P}_0 \,\big\|\, \mathbb{P}_j\big) \;\le\; 3\epsilon^2\,\mathbb{E}_0\big[N_j\big].$$

Proof. As in the proof of Theorem 6, let us define $\sigma^j_{\pi_t} := +1$ if $\pi_t(j \mid s_0) < \frac{1}{2}$ and -1 otherwise. Now, conditioned on the history preceding the draw (t, i), the deployed policy π_t, the selected action $A_{t,i}$, and the branch indicator $\sigma^j_{\pi_t}$ are measurable.

If $A_{t,i} \neq j$ the two models both return $\mathrm{Bern}(\frac{1}{2})$ and the conditional divergence is 0. If $A_{t,i} = j$ the conditional laws are $\mathrm{Bern}(\frac{1}{2})$ under $\mathbb{P}_0$ and $\mathrm{Bern}\big(\frac{1}{2} + \epsilon\sigma^j_{\pi_t}\big)$ under $\mathbb{P}_j$, so the conditional divergence is

$$\mathrm{kl}\Big(\tfrac{1}{2} \,\Big\|\, \tfrac{1}{2} \pm \epsilon\Big) = \tfrac{1}{2} \ln \frac{1/4}{1/4 - \epsilon^2} = -\tfrac{1}{2} \ln\big(1 - 4\epsilon^2\big),$$

The value is identical for both branches. For $\epsilon \leq 1/4$ we have $4\epsilon^2 \leq \frac{1}{4}$ and $-\ln(1-x) \leq \frac{x}{1-x} \leq \frac{4}{3}x$ on $[0, \frac{1}{4}]$, so this is at most $\frac{8}{3}\epsilon^2 \leq 3\epsilon^2$. The chain rule for Kullback–Leibler divergence over the (finitely many) draws, together with the fact that the algorithm's internal randomisation has the same law under both models, gives $\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_j) \leq 3\epsilon^2\, \mathbb{E}_0[N_j]$. □

E.8 Multi-Agent Setting

Proof of Theorem 4. Fix a player i, a comparator $\pi' \in \Pi^i$ and a state s. Player i's realised gains are $g^i_t(s, \cdot) = w_t(s) Q^i_t(s, \cdot) \in [0, w_t(s)/(1-\gamma)]^{\mathcal{A}_i}$, and equation 6 is performing Hedge on them, so Lemma 7 applies verbatim. Summing over s and using $\sum_s w_t(s)^2 \leq 1$ gives us

$$\begin{aligned}\sum_{t=1}^{T} \sum_s w_t(s) \left\langle \pi'(\cdot \mid s) - \pi^i_t(\cdot \mid s),\, Q^i_t(s, \cdot) \right\rangle &\leq \frac{S \ln A_i}{\eta_i} + \frac{\eta_i}{8(1-\gamma)^2} \sum_{t=1}^{T} \sum_s w_t(s)^2 \\ &\leq \frac{S \ln A_i}{\eta_i} + \frac{\eta_i T}{8(1-\gamma)^2}.\end{aligned}$$

Dividing by $T(1-\gamma)$, and substituting $\eta_i = 2(1-\gamma)\sqrt{2S \ln A_i / T}$ and maximising over π' gives us the per-player upper bound. Note that the per-player bound is deterministic, and taking the maximum over the n players gives us the upper bound on $\mathrm{Gap}^{\mathrm{loc}}(\hat{\mu})$. □

Proposition 3 (State-wise CCE decomposition). *With* $\bar{q}^i_\mu(s, a) := \mathbb{E}_\mu\big[d^{\boldsymbol{\pi}}_{P_{\boldsymbol{\pi}}}(s) Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s, a)\big]$ *and* $\bar{v}^i_\mu(s) := \mathbb{E}_\mu\big[d^{\boldsymbol{\pi}}_{P_{\boldsymbol{\pi}}}(s) \langle \pi^i(\cdot \mid s), Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s, \cdot)\rangle\big]$,

$$\mathrm{Gap}^{\mathrm{loc}}_i(\mu) \;=\; \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} \Big(\max_{a \in \mathcal{A}_i} \bar{q}^i_\mu(s, a) - \bar{v}^i_\mu(s) \Big).$$

Proof. Identical to Proposition 2: the objective depends on π' only through $\pi'(\cdot \mid s)$, the constraint set $\Delta(\mathcal{A}_i)^{\mathcal{S}}$ is a product over s, and the inner maximum of a linear functional over $\Delta(\mathcal{A}_i)$ is attained at a vertex. □

E.9 Performative Markov Potential Games

Proof of Theorem 5. Recall the definition of global stability gap for an agent i.

$$\mathrm{Gap}^{\mathrm{glob}}_i(\widehat{\mu}) = \max_{\sigma \in \boldsymbol{\Pi}_i} \mathbb{E}_{\boldsymbol{\pi} \sim \widehat{\mu}} \left[V^i_{M(\boldsymbol{\pi})}(\sigma, \boldsymbol{\pi}^{-i}) - V^i_{M(\boldsymbol{\pi})}(\pi^i, \boldsymbol{\pi}^{-i}) \right]$$

We first bound $\mathbb{E}_{\boldsymbol{\pi} \sim \widehat{\mu}} \left[V^i_{M(\boldsymbol{\pi})}(\sigma, \boldsymbol{\pi}^{-i}) - V^i_{M(\boldsymbol{\pi})}(\pi^i, \boldsymbol{\pi}^{-i}) \right]$ for a fixed policy profile $\boldsymbol{\pi}$. Holding $\boldsymbol{\pi}^{-i}$ fixed, player i faces a stationary single-agent MDP on $\mathcal{S}$ with action set $\mathcal{A}_i$, transition kernel $\mathbb{E}_{a^{-i} \sim \boldsymbol{\pi}^{-i}(\cdot|s)}[P(\cdot|s, (a^i, a^{-i}))]$ and the correspondingly marginalised action-value function is $Q^{i,\boldsymbol{\pi}}_M$. Applying Lemma 1 in that MDP, we obtain

$$\begin{aligned} V^i_{M(\boldsymbol{\pi})}(\sigma, \boldsymbol{\pi}^{-i}) - V^i_{M(\boldsymbol{\pi})}(\boldsymbol{\pi}) &= \frac{1}{1-\gamma} \sum_s d^{(\sigma, \boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})}(s) \left\langle \sigma(\cdot \mid s) - \pi^i(\cdot \mid s),\, Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s, \cdot) \right\rangle \\ &= \frac{1}{1-\gamma} \sum_s \frac{d^{(\sigma, \boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})}(s)}{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s)}\, d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s) \max_{\bar{\sigma} \in \Delta(\mathcal{A})} \left\langle \bar{\sigma} - \pi^i(\cdot \mid s),\, Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s, \cdot) \right\rangle. \end{aligned}$$

Now observe that the term $\max_{\bar{\sigma}\in\Delta(\mathcal{A})} \left\langle \bar{\sigma} - \pi^i(\cdot \mid s),\ Q_M^{i,\boldsymbol{\pi}}(s,\cdot)\right\rangle$ is non-negative as one can just take $\bar{\sigma}$ to be the action that maximizes $Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s,\cdot)$. Therefore,

$$\begin{aligned}
V^i_{M(\boldsymbol{\pi})}(\sigma,\boldsymbol{\pi}^{-i}) - V^i_{M(\boldsymbol{\pi})}(\boldsymbol{\pi}) &\le \left\| \frac{d^{(\sigma,\boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})}}{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}} \right\|_\infty \sum_s \frac{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s)}{1-\gamma} \max_{\bar{\sigma}\in\Delta(\mathcal{A})} \left\langle \bar{\sigma} - \pi^i(\cdot \mid s),\ Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s,\cdot)\right\rangle \\
&\le \left\| \frac{d^{(\sigma,\boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})}}{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}} \right\|_\infty \max_{\bar{\sigma}\in\Delta(\mathcal{A})^S} \sum_s \frac{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s)}{1-\gamma} \left\langle \bar{\sigma} - \pi^i(\cdot \mid s),\ Q^{i,\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}(s,\cdot)\right\rangle \\
&= \left\| \frac{d^{(\sigma,\boldsymbol{\pi}^{-i})}_{M(\boldsymbol{\pi})}}{d^{\boldsymbol{\pi}}_{M(\boldsymbol{\pi})}} \right\|_\infty \max_{\bar{\sigma}\in\Delta(\mathcal{A})^S} \left\langle \bar{\sigma} - \pi^i, \nabla_{\bar{\pi}^i} V^i_{M(\boldsymbol{\pi})}(\bar{\pi}^i,\boldsymbol{\pi}^{-i})\big|_{\bar{\pi}^i=\pi^i}\right\rangle
\end{aligned}$$

The last line uses the policy gradient theorem Sutton et al. (1999) for a single-agent i.e. $\frac{\partial V_M(\pi)}{\partial \pi(a|s)} = d^\pi_M(s)Q^\pi_M(s,a)/(1-\gamma)$. Instantiating the above inequality for each $\boldsymbol{\pi} = \boldsymbol{\pi}_t$ and averaging over T iterations we obtain,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T V^i_{M(\boldsymbol{\pi}_t)}(\sigma,\boldsymbol{\pi}_t^{-i}) - V^i_{M(\boldsymbol{\pi}_t)}(\boldsymbol{\pi}_t) \\
&\le \max_{t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \frac{1}{T}\sum_{t=1}^T \underbrace{\max_{\bar{\sigma}\in\Delta(\mathcal{A})^S} \left\langle \bar{\sigma} - \pi^i_t, \nabla_{\bar{\pi}^i} V^i_{M(\boldsymbol{\pi}_t)}(\bar{\pi}^i,\boldsymbol{\pi}_t^{-i})\big|_{\bar{\pi}^i=\pi^i_t}\right\rangle}_{:=\mathrm{FW}_i(\boldsymbol{\pi}_t)}.
\end{aligned}$$

Taking maximum over $\sigma \in \boldsymbol{\Pi}^i$ and using the definition of $\mathrm{Gap}^{\mathrm{glob}}_i(\widehat{\mu})$ we obtain the following upper bound.

$$\mathrm{Gap}^{\mathrm{glob}}_i(\widehat{\mu}) \le \max_{\sigma\in\boldsymbol{\Pi}^i}\max_{t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \frac{1}{T}\sum_{t=1}^T \mathrm{FW}_i(\boldsymbol{\pi}_t).$$

We now bound the Frank-Wolfe gap $\mathrm{FW}_i(\boldsymbol{\pi}_t)$. Let us define $g^i = \nabla_{\pi^i}\Phi_t(\boldsymbol{\pi}_t)$. Then for any $z \in \boldsymbol{\Pi}_i$, the variational inequality gives

$$\left\langle \pi^i_t + \alpha g^i - \pi^i_{t+1}, z - \pi^i_{t+1}\right\rangle \le 0,$$

and after rearranging we obtain,

$$\alpha\left\langle g^i, z - \pi^i_{t+1}\right\rangle \le \left\langle \pi^i_{t+1} - \pi^i_t, z - \pi^i_{t+1}\right\rangle \le \left\|\pi^i_{t+1} - \pi^i_t\right\|_2 \left\|z - \pi^i_{t+1}\right\|_2 \le \sqrt{2S}\left\|\pi^i_{t+1} - \pi^i_t\right\|_2.$$

The last inequality uses the observation $\left\|z - \pi^i_{t+1}\right\|_2 = \sqrt{\sum_{s,a}(z(a|s) - \pi^i_{t+1}(a|s))^2} \le \sqrt{\sum_{s,a} z(a|s)^2 + \pi^i_{t+1}(a|s)^2} \le \sqrt{2S}$. Therefore,

$$\left\langle g^i, z - \pi^i_t\right\rangle = \left\langle g^i, z - \pi^i_{t+1}\right\rangle + \left\langle g^i, \pi^i_{t+1} - \pi^i_t\right\rangle \le \left(G + \frac{\sqrt{2S}}{\alpha}\right)\left\|\pi^i_{t+1} - \pi^i_t\right\|_2$$

by Assumption 3. Taking maximum over $z \in \boldsymbol{\Pi}^i$ and averaging over T iterations we obtain,

$$\frac{1}{T}\sum_{t=1}^T \mathrm{FW}_i(\boldsymbol{\pi}_t) \le \left(G + \frac{\sqrt{2S}}{\alpha}\right)\frac{1}{T}\sum_{t=1}^T \left\|\pi^i_{t+1} - \pi^i_t\right\|_2.$$

This gives us

$$\begin{aligned}\mathrm{Gap}^{\mathrm{glob}}(\widehat{\mu}) \le \max_i \mathrm{Gap}_i^{\mathrm{glob}}(\widehat{\mu}) &\le \max_{i,\sigma\in\mathbf{\Pi}^i,t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \left(G + \frac{\sqrt{2S}}{\alpha}\right) \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \left\|\pi_{t+1}^i - \pi_t^i\right\|_2 \\ &\le \max_{i,\sigma\in\mathbf{\Pi}^i,t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \left(G + \frac{\sqrt{2S}}{\alpha}\right) \frac{\sqrt{n}}{T}\sum_{t=1}^{T} \left\|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\right\|_2 \\ &\le \max_{i,\sigma\in\mathbf{\Pi}^i,t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \left(G + \frac{\sqrt{2S}}{\alpha}\right) \sqrt{n}\left(2\alpha\epsilon_L + \sqrt{\frac{2\alpha}{T(1-\gamma)}}\right).\end{aligned}$$

For the second result, note that choosing $T \ge \frac{1}{(1-\gamma)\alpha\epsilon_L^2}$ gives us the following upper bound.

$$\begin{aligned}\mathrm{Gap}^{\mathrm{glob}}(\widehat{\mu}) &\le 4\alpha\sqrt{n}\epsilon_L \left(G + \sqrt{2S}/\alpha\right) \max_{i,\sigma\in\mathbf{\Pi}^i,t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \\ &\le 4\epsilon_L\sqrt{n}(G\alpha + \sqrt{2S}) \max_{i,\sigma\in\mathbf{\Pi}^i,t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty \\ &\le 8\sqrt{2Sn}\epsilon_L \max_{i,\sigma\in\mathbf{\Pi}^i,t\in[T]} \left\| \frac{d^{(\sigma,\boldsymbol{\pi}_t^{-i})}_{M(\boldsymbol{\pi}_t)}}{d^{\boldsymbol{\pi}_t}_{M(\boldsymbol{\pi}_t)}} \right\|_\infty\end{aligned}$$

as long as $\alpha \le \sqrt{2S}/G$. □

Lemma 16. *Let $\Delta_t := \|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\|_2$ and . Under Assumption 3 and $\alpha \le 1/L_\Phi$,*

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t \le 2\alpha\epsilon_L + \sqrt{\frac{2\alpha}{T(1-\gamma)}}.$$

Proof. We will write $\psi(\boldsymbol{\pi}) := \Phi_{M(\boldsymbol{\pi})}(\boldsymbol{\pi})$ and $g := \nabla\Phi_t(\boldsymbol{\pi}_t)$. The projection $\boldsymbol{\pi}_{t+1}$ of $\boldsymbol{\pi}_t + \alpha g$ onto $\mathbf{\Pi}$ satisfies the variational inequality $\langle \boldsymbol{\pi}_t + \alpha g - \boldsymbol{\pi}_{t+1},\ \boldsymbol{z} - \boldsymbol{\pi}_{t+1}\rangle \le 0$ for every $\boldsymbol{z} \in \mathbf{\Pi}$. Taking $\boldsymbol{z} = \boldsymbol{\pi}_t$ and rearranging,

$$\Delta_t^2 \ \le\ \alpha\,\langle g,\ \boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\rangle. \tag{25}$$

Since $\Phi_t(\cdot)$ is L_Φ-smooth.

$$\begin{aligned}\Phi_t(\boldsymbol{\pi}_{t+1}) &\ge\ \Phi_t(\boldsymbol{\pi}_t) + \langle g,\ \boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\rangle - \frac{L_\Phi}{2}\Delta_t^2 \\ &\ge\ \Phi_t(\boldsymbol{\pi}_t) + \frac{\Delta_t^2}{\alpha} - \frac{L_\Phi}{2}\Delta_t^2 \\ &\ge\ \Phi_t(\boldsymbol{\pi}_t) + \frac{\Delta_t^2}{2\alpha},\end{aligned}$$

using equation 25 and then $\alpha \le 1/L_\Phi$. Finally, we obtain the following inequality.

$$\begin{aligned}\psi(\boldsymbol{\pi}_{t+1}) - \psi(\boldsymbol{\pi}_t) &= \left[\Phi_{t+1}(\boldsymbol{\pi}_{t+1}) - \Phi_t(\boldsymbol{\pi}_{t+1})\right] + \left[\Phi_t(\boldsymbol{\pi}_{t+1}) - \Phi_t(\boldsymbol{\pi}_t)\right] \\ &\ge \Phi_{M(\boldsymbol{\pi}_{t+1})}(\boldsymbol{\pi}_{t+1}) - \Phi_{M(\boldsymbol{\pi}_t)}(\boldsymbol{\pi}_{t+1}) + \frac{\Delta_t^2}{2\alpha} \\ &\ge -\epsilon_L\left\|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\right\|_2 + \frac{\Delta_t^2}{2\alpha}.\end{aligned}$$

The last line uses (P2) of Assumption 3. This gives us the following recurrence relation.

$$\psi(\boldsymbol{\pi}_{t+1}) - \psi(\boldsymbol{\pi}_t) \ge -\epsilon_L\Delta_t + \frac{\Delta_t^2}{2\alpha} \tag{26}$$

Summing Equation (26) for $t = 1, \ldots, T$, we obtain

$$\psi(\boldsymbol{\pi}_{T+1}) - \psi(\boldsymbol{\pi}_1) \geq -\epsilon_L \sum_{t=1}^{T} \Delta_t + \frac{1}{2\alpha} \sum_{t=1}^{T} \Delta_t^2.$$

We now use two observations. First, rewards are bounded between 0 and 1 and $\psi(\boldsymbol{\pi}_{T+1}) - \psi(\boldsymbol{\pi}_1) = \Phi_{M(\boldsymbol{\pi}_{T+1})}(\boldsymbol{\pi}_{T+1}) - \Phi_{M(\boldsymbol{\pi}_1)}(\boldsymbol{\pi}_1) \leq 1/(1-\gamma)$. Second, Cauchy-Schwarz inequality gives $\sum_{t=1}^{T} \Delta_t \leq \sqrt{T \sum_{t=1}^{T} \Delta_t^2}$. Writing $S_2^2 = \sum_{t=1}^{T} \Delta_t^2$ we obtain,

$$\frac{1}{1-\gamma} \geq -\epsilon_L \sqrt{T} \cdot S_2 + \frac{1}{2\alpha} S_2^2.$$

This gives us $S_2 \leq \alpha \left(\epsilon_L \sqrt{T} + \sqrt{\epsilon_L^2 T + \frac{2}{\alpha(1-\gamma)}} \right) \leq 2\alpha\epsilon_L \sqrt{T} + \sqrt{\frac{2\alpha}{1-\gamma}}$. Now we use $\frac{1}{T} \sum_{t=1}^{T} \Delta_t \leq \frac{S_2}{\sqrt{T}}$ to obtain the desired bound. □