On Global Convergence Rates for Federated Policy Gradient under Heterogeneous Environment

Safwan Labbi[1] **Paul Mangold**[1] **Daniil Tiapkin**[1,2] **Eric Moulines**[1,3]
[1]CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
[2]Université Paris-Saclay, CNRS, LMO, 91405 Orsay, France
[3]Mohamed bin Zayed University of Artificial Intelligence, UAE

Abstract

Ensuring convergence of policy gradient methods in federated reinforcement learning (FRL) under environment heterogeneity remains a major challenge. In this work, we first establish that heterogeneity, perhaps counter-intuitively, can necessitate optimal policies to be non-deterministic or even time-varying, even in tabular environments. Subsequently, we prove global convergence results for federated policy gradient (`FedPG`) algorithms employing local updates, under a Łojasiewicz condition that holds only for each individual agent, in both entropy-regularized and non-regularized scenarios. Crucially, our theoretical analysis shows that `FedPG` attains linear speed-up with respect to the number of agents, a property central to efficient federated learning. Leveraging insights from our theoretical findings, we introduce `b-RS-FedPG`, a novel policy gradient method that employs a carefully constructed softmax-inspired parameterization coupled with an appropriate regularization scheme. We further demonstrate explicit convergence rates for `b-RS-FedPG` toward near-optimal stationary policies. Finally, we demonstrate that empirically both `FedPG` and `b-RS-FedPG` consistently outperform federated Q-learning on heterogeneous settings.

1 Introduction

In Federated Reinforcement Learning (FRL), multiple agents collaboratively optimize a shared policy without directly exchanging their local actions or rewards (Qi et al., 2021; Zhuo et al., 2023). Instead, policy information is aggregated via a central server (Khodadadian et al., 2022). FRL improves traditional distributed methods by enhancing privacy and minimizing communication overhead. Despite these advantages, FRL implementations face key challenges, notably heterogeneity of the environment (Jin et al., 2022) and limited communication bandwidth (Zhu et al., 2022; Fan et al., 2023). Most existing studies focus on homogeneous settings, where all agents interact with identical Markov Decision Processes (MDPs). The heterogeneous scenario, where agents evolve in different MDPs, is still largely under-explored.

This paper specifically addresses federated policy gradient methods in heterogeneous settings. Previously, Wang et al. (2024a) established convergence only to first-order stationary points of average-value functions. Such guarantees are substantially weaker than those for single-agent policy gradient methods, which achieve global convergence under mild conditions (Mei et al., 2020). Our main contribution is to bridge this gap. We first analyze structural properties specific to FRL under heterogeneity. A key result is that optimal common policies, even in the tabular setting, can inherently be *non-deterministic* or even *non-stationary*, in stark contrast to classical reinforcement learning (RL). This finding underscores the novel challenges posed by environmental heterogeneity in FRL.

Motivated by this insight, we develop novel algorithmic strategies tailored for learning non-deterministic stationary policies in tabular environments. We analyze federated policy gradient

Preprint. Under review.

Table 1: Comparison with prior work in the setting of agents with heterogeneous dynamics. Our results are the first to prove global convergence of FedPG to a near-optimal policy.

Algorithm	Local Steps	Global convergence	Last iterate	Linear- Speedup
PAvg (Jin et al. 2022)	✓	✗	✗	✗
FEDSVRPG-M (Wang et al. 2024a)	✓	✗	✗	✓
FEDHAPG-M (Wang et al. 2024a)	✓	✗	✗	✓
FedPG (our work)	✓	✓	✓	✓

methods (FedPG) with softmax parameterization, demonstrating convergence to a neighborhood of the optimal policy even with heterogeneous agents. Crucially, our analysis quantifies how this neighborhood size depends on environmental heterogeneity. Additionally, we highlight substantial benefits from incorporating regularization techniques, particularly when combined with a novel and carefully designed parameterization of the policy.

Our theoretical results rely on local assumptions about agent-specific value functions, which do not extend to the global FRL objective. We establish convergence guarantees for FedPG under a non-uniform Łojasiewicz inequality, which generalizes gradient dominance in non-convex objectives. This analysis extends beyond FRL, offering novel convergence insights into Federated Averaging (FedAVG) (McMahan et al., 2017) under non-uniform Łojasiewicz conditions. Our key contributions can be summarized as follows:

- We show that, due to heterogeneity, the classical properties of RL do not hold anymore in FRL. Specifically, optimal policies may be stochastic or time-varying even in simple tabular settings.
- We establish the *first* global convergence theory for entropy-regularized policy gradients in heterogeneous FRL, proving that FedPG converges to near-optimal policies under local non-uniform Łojasiewicz conditions, and achieves linear speed-up in the number of agents.
- Based on our results, we introduce a novel softmax-inspired parameterization with tailored regularization, for which we derive a convergence rate *with explicit constants*.
- We conduct experiments on two FRL environments, confirming our theoretical predictions for FedPG's, and demonstrating its robust performance across varying levels of heterogeneity.

We summarize the main differences between our analysis and previous work in Table 1. We discuss related work in Section 2, and describe the particular properties of FRL in Section 3. We then present our main theoretical results in Section 4, and confirm them numerically in Section 5.

2 Related Work

Policy Gradient Methods. Policy gradient methods (Williams, 1992; Sutton et al., 1999) have been extensively studied in the single-agent discounted RL setting. In deterministic scenarios, global convergence results leverage the quasi-convexity of the value function of the state-action discounted occupancy measure to avoid convergence to suboptimal traps, utilizing a Łojasiewicz condition and ensuring a positive probability for optimal actions (Mei et al., 2020; Zhang et al., 2020; Xiao, 2022). Mei et al. (2020) (following Agarwal et al. 2020) analyze entropy-regularized policy gradient methods, establishing linear convergence rates and detailing how entropy regularization enhances optimization properties. Early analyses for stochastic policy gradient methods demonstrate convergence only to first-order stationary points (Zhang et al., 2021b,a; Yuan et al., 2022). Mei et al. (2021) explicitly examine the discrepancy between deterministic and stochastic policy gradient updates, identifying conditions necessary for global convergence under stochastic updates. Mei et al. (2024) addresses the special case of bandit settings, proving that policy-gradient algorithms converge globally almost surely for any constant step size.

Federated Reinforcement Learning. FRL has attracted much attention recently, with significant theoretical developments in both value-based and policy-based tabular methods; see Zhuo et al. (2023) and the references therein. In homogeneous environments, several studies have proposed federated Q-learning variants that achieve near-optimal sample and communication complexity (Salgia and Chi, 2024; Zheng et al., 2025). For heterogeneous environments – where agents face different MDPs – recent analyzes highlight inherent convergence trade-offs; federated algorithms achieve

linear speedup but suffer from unavoidable suboptimality proportional to the degree of heterogeneity (Wang et al., 2024b; Zhang et al., 2024; Labbi et al., 2025). Policy gradient (PG) algorithms in federated environments have been mainly explored under homogeneous assumptions, especially via natural policy gradient approaches, which provide strong global convergence guarantees and improved communication efficiency (Lan et al., 2023; Ganesh et al., 2024). In *federated multitask settings*—where agents share identical dynamics but differ in reward functions—several works have established either convergence to stationary points (Zhu et al., 2024; Chen et al., 2021) or global convergence guarantees (Yang et al., 2024). These studies highlight that reward heterogeneity alone does not introduce significant theoretical challenges, as the resulting problem retains the fundamental characteristics of standard RL settings; see Appendix B.1. In contrast, environments with heterogeneous dynamics present additional complexities since optimal policies may become dependent on the initial state distribution (Jin et al., 2022). Consequently, existing federated policy gradient methods in this scenario face intrinsic difficulties arising from non-convex optimization landscapes, restricting their convergence guarantees to stationary points (Jin et al., 2022; Wang et al., 2024a).

Federated Averaging. Federated learning (FL) has generated an extensive body of research, a comprehensive review of which is beyond the scope of this paragraph; see Kairouz et al. (2021); Wang et al. (2021). We specifically focus on FL in non-convex optimization settings involving finite sums of functions under the Polyak–Łojasiewicz (PL) condition. Existing literature predominantly assumes the PL condition on the global (average) objective; by contrast, our approach requires PL-like conditions only at the level of local client objectives. Deterministic local gradient descent (full-batch local updates) under PL conditions has been studied by Haddadpour and Mahdavi (2019), who demonstrated optimal convergence rates provided gradient diversity across clients is suitably bounded. Additionally, Haddadpour et al. (2019) analyzed stochastic local updates, establishing that local SGD achieves linear convergence speed-ups proportional to the number of participating devices. These results have been extended Demidovich et al. (2025) under (L_0, L_1)-smoothness: their results still require PL conditions both for the individual and the average functions.

3 Heterogeneous Federated Reinforcement Learning

Problem setting. In FRL, each of the M agents independently interacts with its own infinite-horizon discounted MDP, defined as $\mathcal{M}_c := (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_c, \mathsf{r}, \rho)$. Each agent-specific MDP shares the finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, discount factor γ, a common deterministic reward function r, and a distribution ρ over initial states. The differences among these MDPs lie solely in their transition kernels P_c. Consequently, an FRL instance is fully characterized by the set of agent-specific transition dynamics.

For an agent $c \in [M]$, we denote its *local history* until iteration t as $\mathcal{H}_c^t := (s_c^0, a_c^0, s_c^1, a_c^1, \ldots, s_c^t)$, with $\mathbb{H}_{\mathrm{loc}}^t := (\mathcal{S} \times \mathcal{A})^t$ the set of possible local histories. Similarly, we denote $\mathcal{H}^t := (\mathcal{H}_1^t, \ldots, \mathcal{H}_M^t) \in \mathbb{H}_M^t$ the *global history*. A *global decision rule* $\pi^t \colon \mathbb{H}_M^t \times [M] \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$, assigns to each global history and agent a distribution over actions. An *agent-aware history-dependent policy* is a sequence $\pi = (\pi^t)_{t\in\mathbb{N}}$. The class of all such policies is denoted by Π. A *local decision rule* $\pi_c^t \colon \mathbb{H}_{\mathrm{loc}}^t \to \mathcal{P}(\mathcal{A})$ maps local histories to action distributions. A *local history-dependent policy* is $((\pi_c^t)_{t\in\mathbb{N}})_c$, with the set of all such policies denoted Π_ℓ. A *local stationary stochastic policy* is $\pi \colon \mathcal{S} \to \mathcal{P}(\mathcal{A})$, mapping current states to distributions over actions; this class is denoted by Π_{sta}. A *local deterministic policy* maps states directly to actions, $\pi \colon \mathcal{S} \to \mathcal{A}$, forming class Π_{det}.

For a given policy class $\mathcal{X} \subset \Pi$, we define the FRL objective over $\mathcal{X}$ as

$$\max_{\pi\in\mathcal{X}} J(\pi) := \tfrac{1}{M}\textstyle\sum_{c=1}^M J_c(\pi) \quad , \quad J_c(\pi) := \mathbb{E}_\pi\left[\textstyle\sum_{t=0}^\infty \gamma^t \mathsf{r}(S_c^t, A_c^t)\right] \quad , \tag{1}$$

where $\mathbb{E}_\pi[\cdot]$ is the expectation over random trajectories generated by following a policy $\pi = (\pi^t)_{t\in\mathbb{N}}$: the initial state is sampled from a distribution $S_c^0 \sim \rho(\cdot)$ and $\forall t \geq 0 : A_c^t \sim \pi^t(\cdot|\mathcal{H}^t, c), S_c^{t+1} \sim \mathsf{P}_c(\cdot|S_c^t, A_c^t)$, for $\mathcal{H}^t = (\mathcal{H}_c^t)_{c\in[M]}$ where $\mathcal{H}_c^t = (S_c^0, A_c^0, \ldots, S_c^t)$ for all $c \in [M]$. In (1), $J_c(\pi)$ is the expected discounted return of the agent c, and we notice that the local objectives J_c may vary across agents since the corresponding MDPs $\mathcal{M}_c$ may differ.

By construction, $\Pi_{\mathrm{det}} \subset \Pi_{\mathrm{sta}} \subset \Pi_\ell \subset \Pi$: restricting to smaller policy classes can only reduce or maintain the supremum of J_{sm}. In single-agent RL, the optimal policy can always be found within the class of stationary deterministic policies; see e.g. (Agarwal et al., 2019, Theorem 1.7). This result does not extend to FRL with heterogeneous agents.

Algorithm 1 (`S, RS, b-RS)-FedPG`

Initialization: Learning rate $\eta > 0$, parameter θ^0, projection set $\mathcal{T}$
for $r = 0$ to $R-1$ **do**
  for $c = 1$ to M **do**
    Set $\theta_c^{r,0} = \theta^r$.
    for $h = 0$ to $H-1$ **do**
      Collect B trajectories of length T: $Z_c^{r,h+1} := (S_{c,b}^{r,h,1:T}, A_{c,b}^{r,h,1:T})_{b=1}^B$ using $\pi_{\theta_c^{r,h}}$
      Update $\theta_c^{r,h+1} = \theta_c^{r,h} + \eta \mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})$ where $\mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})$ is computed using (6) for `S-FedPG`, (9) for `RS-FedPG`, and (10) for `b-RS-FedPG`
  Server updates parameter: $\theta^{r+1} = \mathrm{proj}_{\mathcal{T}}(\bar{\theta}^{r+1})$ where $\bar{\theta}^{r+1} = \frac{1}{M}\sum_{c=1}^M \theta_c^{r,H}$

Theorem 3.1. *For each of the following properties, there exists an FRL instance with two infinite-horizon discounted MDPs that satisfy*

$$\max_{\pi\in\Pi_{\mathrm{det}}} J(\pi) < \max_{\pi\in\Pi_{\mathrm{sta}}} J(\pi) \ , \quad \max_{\pi\in\Pi_{\mathrm{sta}}} J(\pi) < \max_{\pi\in\Pi_{\ell}} J(\pi) \ , \quad \max_{\pi\in\Pi_{\ell}} J(\pi) < \max_{\pi\in\Pi} J(\pi) \ .$$

The proof is postponed to Appendix B. We emphasize that these additional challenges arise specifically from heterogeneity in transition kernels. When transition kernels are homogeneous, *even with heterogeneous rewards*, the federated setting simplifies to a standard RL problem with an averaged reward structure; see Appendix B.1.

Note that the structural distinctions between FRL and standard RL inform algorithm selection. Algorithms like Fed-Q-learning (Jin et al., 2022), which target deterministic policies, are suboptimal in FRL (see Theorem 3.1). Conversely, history-dependent policies require substantial resources, making them impractical. Hence, stationary policies emerge as an effective compromise between computational feasibility and decision quality.

4 Solving Federated Reinforcement Learning with Policy Gradient Methods

In what follows, we consider the optimization of the objective J over the class of stationary policies, aiming thus to solve the following optimization problem: $\sup_{\pi\in\Pi_{\mathrm{sta}}} J(\pi)$. We use a *softmax parameterization*, i.e. given a parameter $\theta \in \Theta = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, the corresponding policy is defined as

$$\pi_\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\theta(s,a'))} \ , \quad \theta = (\theta(s,a), (s,a) \in \mathcal{S}\times\mathcal{A}) \tag{2}$$

Although we write θ as a function for notational clarity, we equivalently view $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as a real-valued matrix indexed by $(s,a) \in \{0,\dots|\mathcal{S}|-1\}\times\{0,\dots|\mathcal{A}|-1\}$. That is, we identify the function $\theta : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ with its corresponding matrix representation. This slight abuse of notation allows us to use both functional and matrix-based views interchangeably, which simplifies exposition in the context of softmax policy parameterization. Note that, because for each $s \in \mathcal{S}$, $\pi_\theta(\cdot|s)$ is a function to the probability simplex of dimension $|\mathcal{A}|$, the Jacobian $\mathrm{Jac}_{\pi_\theta}(\cdot|s) = \mathrm{diag}(\pi_\theta(\cdot|s)) - \pi_\theta(\cdot|s)\pi_\theta^\top(\cdot|s)$ satisfies $\mathrm{Jac}_{\pi_\theta}(\cdot|s)\mathbf{1}_{|\mathcal{A}|} = \mathbf{0}_{|\mathcal{A}|}$, where $\mathbf{1}_{|\mathcal{A}|} = [1,\dots,1]^\top$ and $\mathbf{0}_{|\mathcal{A}|} = [0,\dots,0]^\top$. Alternative softmax-like parameterizations will also be considered as we discuss further in Section 4.4. Given a parameterized policy π_θ, the distribution of a trajectory truncated at $T-1$ step for agent $c \in [M]$ is given, for $z = (s^t, a^t)_{t=0}^{T-1} \in (\mathcal{S}\times\mathcal{A})^T$, by

$$\nu_c(\theta; z) = \rho(s^0)\pi_\theta(a^0|s^0)\cdot\textstyle\prod_{t=1}^{T-1}\mathsf{P}_c(s^t \mid s^{t-1}, a^{t-1})\pi_\theta(a^t|s_t) \ . \tag{3}$$

We introduce the following assumptions to underpin our subsequent analysis:

A-1. *For any $c, c' \in [M]$, it holds that* $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\mathsf{P}_c(\cdot|s,a) - \mathsf{P}_{c'}(\cdot|s,a)\|_1 \le \varepsilon_{\mathsf{P}}$.

A-2. *The initial state distribution ρ satisfies* $\min_s \rho(s) > 0$.

A-1 captures agent heterogeneity by bounding the total variation between transition kernels, following Zhang et al. (2024). **A**-2 ensures sufficient exploration and plays a key role in the analysis.

4.1 General FedPG framework

We introduce and analyze three novel algorithms—S-FedPG, RS-FedPG, and b-RS-FedPG—as federated extensions of the policy gradient method; see Mei et al. (2020); Agarwal et al. (2021) and the references therein. Each algorithm is a specific instance of the general FedPG framework, where multiple agents cooperatively optimize the global objective $F(\theta)$, defined as an average of agent-specific local objectives $f_c(\theta)$. The key differences among the algorithms lie in the choice of local objectives, gradient estimators, and policy parameterizations.

Each communication round involves the central server distributing global parameters θ^r to all agents. Subsequently, each agent performs H stochastic gradient ascent steps on its local objective $f_c(\theta)$:

$$\theta_c^{r,h+1} = \theta_c^{r,h} + \eta \cdot \mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h}), \quad \theta_c^{r,0} = \theta^r, \tag{4}$$

where $\eta > 0$ is a learning rate, $\mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})$ is a REINFORCE-like estimator (Williams, 1992) of $\nabla f_c(\theta_c^{r,h})$ that uses a batch of independent B trajectories of length T: $Z_c^{r,h+1} = (Z_{c,b}^{r,h+1})_{b=1}^B \sim [\nu_c(\theta_c^{r,h})]^{\otimes B}$(see (3) for a definition of $\nu_c(\theta)$). After H local steps, following the federated averaging procedure, we average the final parameters $\bar{\theta}^{r+1} = \frac{1}{M}\sum_{c=1}^M \theta_c^{r,H}$ followed (optionally) by a projection step $\theta^{r+1} = \mathrm{proj}_{\mathcal{T}}(\bar{\theta}^{r+1})$ onto a specified target set $\mathcal{T}$. The complete algorithm is summarized in Algorithm 1.

Our convergence results are based on a novel fine-grained analysis of federated averaging, which is of independent interest; see Appendix C. The proposed approach utilizes a second-order expansion of the local objective function and extends the proof techniques originally introduced in Glasgow et al. (2022) to achieve linear speed-up.

Lemma 4.1 (Ascent Lemma)**.** *Assume the following conditions:*

1. Smoothness of the objective function and expected gradient estimator: *For any $c \in [M]$, the function f_c is L_1-Lipschitz, functions $\mathrm{g}_c(\theta) := \mathbb{E}_{Z_c\sim\nu_c(\theta)}[\mathrm{g}_c^{Z_c}(\theta)]$, ∇f_c are L_2-Lipschitz, and $\langle \nabla f_c, v\rangle$ is L_3-smooth for any $\|v\|_2 = 1$;*
2. Bounded gradient heterogeneity: $\|\nabla F(\theta) - \nabla f_c(\theta)\|_2 \le \zeta$;
3. Bounded bias, variance, and fourth central moment of the gradient estimator: *For any parameter θ, $\|\nabla f_c(\theta) - \mathrm{g}_c(\theta)\|_2 \le \beta$, and for $p \in \{2,4\}$ it holds $\mathbb{E}_{Z_c\sim\nu_c(\theta)}\left[\|\mathrm{g}_c^{Z_c}(\theta) - \mathrm{g}_c(\theta)\|_2^p\right] \le \sigma_p^p$.*

Then, for any $\eta > 0$ such that $\eta H L_2 \le 1/6$ and $32\eta^2 H^2 L_3^2 L_1^2 \le L_2^2$, the iterates of FedPG (see (4)) satisfy

$$F(\theta^r) - \mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right] \le -\frac{\eta H}{4}\|\nabla F(\theta^r)\|_2^2 + \frac{3\eta^2 L_2 H \sigma_2^2}{2M} + 2\eta H\beta^2 + 8\eta^3 L_2^2 H^2 (H-1)\zeta^2 + 4\cdot 12^3 \eta^5 L_3^2 H^2 (H-1)\sigma_4^4,$$

where $\mathcal{F}^r$ is a global filtration induced by iterates of the algorithm.

For the proof and a more detailed discussion of this result, we refer the reader to Appendix C. It is worth noting that the effects of agent heterogeneity and second-order biases disappear when $H = 1$.

4.2 Analysis of S-FedPG

We consider S-FedPG, the vanilla federated Softmax Policy Gradient method. This algorithm is a specific instance of FedPG with a local objective defined as $J_{\mathrm{sm},c}(\theta) = J_c(\pi_\theta)$ and a global $J_{\mathrm{sm}}(\theta) = 1/M\sum_{c=1}^M J_{\mathrm{sm},c}(\theta)$. Also define $J_{\mathrm{sm},c}^\star = \sup_{\theta\in\Theta} J_{\mathrm{sm},c}(\theta)$ and $J_{\mathrm{sm}}^\star$ as an average of $J_{\mathrm{sm},c}^\star$.

Importantly, Mei et al. (2020) shows that each local function $J_{\mathrm{sm},c}$ is smooth with a Lipschitz gradient, implying the global objective J_{sm} is also smooth. They further prove (Lemmas 8 and 15) that *each local function* $J_{\mathrm{sm},c}$ satisfies a non-uniform Łojasiewicz inequality:

$$\left\|\nabla J_{\mathrm{sm},c}(\theta)\right\|_2^2 \ge 2\mu_{\mathrm{sm},c}(\theta)[J_{\mathrm{sm},c}^\star - J_{\mathrm{sm},c}(\theta)]^2, \tag{5}$$

where $\mu_{\mathrm{sm},c}(\theta)$ is a strictly positive function, under **A**-2 - expression is provided in Lemma D.9; see also as Mei et al. (2020). The non-uniform Łojasiewicz inequality (5) ensures that first-order

stationary points and global maxima of $J_{\mathrm{sm},c}$ coincide. However, note that the sum or average of functions individually satisfying (5) does not necessarily satisfy (5), as the non-uniform Łojasiewicz inequality is not stable under averaging. Consequently, existing analyses of federated averaging methods leveraging classical Łojasiewicz conditions (e.g., Demidovich et al. (2025)) cannot be directly applied in our context.

Regarding the gradient estimator, we consider the vanilla REINFORCE to estimate the gradient of the local function $J_{\mathrm{sm},c}$, and to do it, we need to sample truncated trajectories by following the current policy π_θ. The stochastic oracle for the gradient of the local objective function is

$$\mathrm{g}^{Z_c}_{\mathrm{sm},c}(\theta) := \tfrac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t\left(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(A^\ell_{c,b}\mid S^\ell_{c,b})\right)\mathsf{r}(S^t_{c,b},A^t_{c,b})\ . \tag{6}$$

where $Z_c=(Z_{c,b})_{b=0}^B$ and $Z_{c,b}=(S^t_{c,b},A^t_{c,b})_{t=0}^{T-1}$ are independent truncated trajectories sampled by following policy π_θ (i.e., from $\nu_c(\theta)$). Importantly, note that our analysis readily generalizes to more advanced gradient oracles, such as those employing variance-reduction techniques, importance sampling, and other related methods. Under **A**-1 and 2, we can verify that the properties needed for Lemma 4.1 holds (see Appendix D). Given this lemma and the non-uniform Łojasiewicz inequality (5), we establish the following convergence rate for `S-FedPG`.

Theorem 4.2 (Convergence rates of `S-FedPG`)**.** *Assume **A**-1 and 2 and no projection (i.e., $\mathcal{T}=\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$). Additionally, assume that there exists $\mu_{\mathrm{sm}}\in(0,1)$ such that, with probability 1, $\inf_{r\in\mathbb{N}}\mu_{\mathrm{sm}}(\theta^r)\geq\mu_{\mathrm{sm}}$. For any $\eta>0$ such that $\eta H\leq(1-\gamma)^3/592$, $T\geq 4(1-\gamma)^{-2}$, and $M\cdot B\geq(1-\gamma)^{-1}$, the iterates of `S-FedPG` satisfy*

$$J^\star_{\mathrm{sm}}-\mathbb{E}[J_{\mathrm{sm}}(\theta^R)]\lesssim\frac{J^\star_{\mathrm{sm}}-J_{\mathrm{sm}}(\theta^0)}{1+R\cdot(J^\star_{\mathrm{sm}}-J_{\mathrm{sm}}(\theta^0))\cdot\eta H\mu_{\mathrm{sm}}}+\frac{\eta^{1/2}}{\mu_{\mathrm{sm}}^{1/2}M^{1/2}B^{1/2}\cdot(1-\gamma)^{3.5}}$$
$$+\frac{\eta^2H^{1/2}(H-1)^{1/2}}{\mu_{\mathrm{sm}}^{1/2}(1-\gamma)^8B}+\frac{T\gamma^T}{\mu_{\mathrm{sm}}^{1/2}(1-\gamma)}+\frac{\varepsilon_{\mathsf{P}}}{\mu_{\mathrm{sm}}^{1/2}(1-\gamma)^3}\ .$$

We notice that in these convergence guarantees, the heterogeneity bias does not disappear even if $H=1$. This behavior is unavoidable since it is not guaranteed that the global objective function J_{sm} also satisfies (5) and thus it potentially may have several local minima and maxima. Based on this result, we also obtain the following communication and sample complexity results for `S-FedPG`:

Corollary 4.3 (Sample and Communication Complexity of `S-FedPG`)**.** *Under the assumptions of Theorem 4.2, let $\epsilon\gtrsim\varepsilon_{\mathsf{P}}\mu_{\mathrm{sm}}^{-1/2}(1-\gamma)^{-3}$. Then, for T such that $T\gtrsim(1-\gamma)^{-1}\max((1-\gamma),\log(\epsilon\mu_{\mathrm{sm}}^{1/2}(1-\gamma)))$, a properly chosen step size and number of local updates, `S-FedPG` learns an ϵ-approximation of the optimal objective with a number of communication rounds*

$$R\gtrsim\frac{[(J^\star_{\mathrm{sm}}-J_{\mathrm{sm}}(\theta^0))-\epsilon/5]}{(J^\star_{\mathrm{sm}}-J_{\mathrm{sm}}(\theta^0))\mu_{\mathrm{sm}}\epsilon(1-\gamma)^3}\ ,$$

for a total number of sampled trajectories per agent of

$$RHB\gtrsim\max\Big(\frac{B}{\mu_{\mathrm{sm}}(1-\gamma)^3},\frac{1}{\mu^2_{\mathrm{sm}}M(1-\gamma)^7\epsilon^2},\frac{1}{\mu^{3/2}_{\mathrm{sm}}\epsilon(1-\gamma)^5}\Big)\frac{[(J^\star_{\mathrm{sm}}-J_{\mathrm{sm}}(\theta^0))-\epsilon/5]}{(J^\star_{\mathrm{sm}}-J_{\mathrm{sm}}(\theta^0))\epsilon}\ .$$

This result shows that `S-FedPG` have linear speedup until $M\approx\min\Big(\frac{1}{\mu_{\mathrm{sm}}^{1/2}\epsilon(1-\gamma)^2},\frac{1}{\mu^2_{\mathrm{sm}}B(1-\gamma)^4\epsilon^2}\Big)$.

4.3 Analysis of `RS-FedPG`

Next, following Agarwal et al. (2020); Mei et al. (2020), we analyze the softmax policy gradient algorithm with entropy regularization. In particular, we are interested in the optimization of the following local objectives:

$$J_{\mathrm{r},c}(\theta)=:J_{\mathrm{sm},c}(\theta)+\lambda\mathcal{H}^\rho_c(\theta)\,,\quad\mathcal{H}^\rho_c(\theta):=-\mathbb{E}_{\pi_\theta}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t\log(\pi_\theta(A^t_c|S^t_c))\right]\ , \tag{7}$$

where $\lambda>0$ is an regularization coefficient, and the global objective $J_{\mathrm{r}}=1/M\sum_{c=1}^M J_{\mathrm{r},c}$.

The core idea of entropy regularization is to penalize deterministic policies and promote exploration (Williams and Peng, 1991; Mnih et al., 2016; Schulman et al., 2017; Ahmed et al., 2019). As

noted by (Mei et al., 2020, Theorem 6), the (deterministic) policy gradient algorithm applied to the entropy-regularized objective (7) enjoys a linear convergence rate toward the softmax optimal policy, in contrast to the polynomial convergence rates observed in non-regularized scenarios. This accelerated convergence results from the fulfillment of a stronger, non-uniform Polyak–Łojasiewicz condition

$$\|\nabla J_{\mathrm{r},c}(\theta)\|_2^2 \geq 2\mu_{\mathrm{r},c}^{\lambda}(\theta)\left[J_{\mathrm{r},c}^{\star} - J_{\mathrm{r},c}(\theta)\right] \quad . \tag{8}$$

The main difference to the previous version of the non-uniform Łojasiewicz inequality is that the sub-optimality gap is linear. For small sub-optimality gaps this means that the gradient must be larger. The constant $\mu_{\mathrm{r},c}^{\lambda}(\theta)$ is given in Lemma E.10; see also (Mei et al., 2020, Lemma 15). As above, we emphasize that the satisfaction of a non-uniform PL condition by each local function does not necessarily imply that their average inherits this condition; see Lemma E.9 where we provide a counter example.

To estimate the gradients of (7), we employ the REINFORCE objective using a log-policy augmented reward $\mathrm{g}_{\mathrm{r},c}^{Z_c}(\theta)$, identical in form to (6), with the reward r replaced by the augmented reward

$$\tilde{\mathsf{r}}_{\lambda}(S_{c,b}^t, A_{c,b}^t; \theta) := \mathsf{r}(S_{c,b}^t, A_{c,b}^t) - \lambda \log(\pi_{\theta}(A_{c,b}^t | S_{c,b}^t)) \, . \tag{9}$$

Here again, we do not apply the projection. This choice of the gradient estimator results in the algorithm `RS-FedPG` that achieves the following sample complexity result under an assumption of $\mu_{\mathrm{r},c}^{\lambda}(\theta_r)$ is bounded away from zero with probability 1:

Corollary 4.4 (Sample and Communication Complexity of `RS-FedPG`)**.** *Assume **A**-1 and 2 and no projection (i.e., $\mathcal{T} = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$). Moreover, assume that there exists $\mu_{\mathrm{r}}^{\lambda} \in (0,1)$ such that such that $\inf_{r\in[\mathbb{N}]} \mu_{\mathrm{r}}^{\lambda}(\theta^r) \geq \mu_{\mathrm{r}}^{\lambda} > 0$ with probability 1. Let $\epsilon \gtrsim (1+\lambda\log(|\mathcal{A}|))^2 \varepsilon_{\mathsf{P}}^2 (\mu_{\mathrm{r}}^{\lambda})^{-1}(1-\gamma)^{-6}$. Then, for a properly chosen truncation horizon, step size, and number of local updates, `RS-FedPG` learns an ϵ-approximation of the optimal objective $J_{\mathrm{r}}^{\star}$ with a number of communication rounds*

$$R \gtrsim \frac{(1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^3 \mu_{\mathrm{r}}^{\lambda}} \log\Big(\frac{(J_{\mathrm{r}}^{\star} - J_{\mathrm{r}}(\theta^0))}{\epsilon}\Big) \quad ,$$

for a total number of samples per agent of

$$RHB \gtrsim \max\Big(\frac{(1+\lambda\log(|\mathcal{A}|))B}{\mu_{\mathrm{r}}^{\lambda}(1-\gamma)^3}, \frac{(1+\lambda\log(|\mathcal{A}|))^3}{(\mu_{\mathrm{r}}^{\lambda})^2 \epsilon M (1-\gamma)^7}, \frac{(1+\lambda\log(|\mathcal{A}|))^2}{\epsilon^{1/2}(\mu_{\mathrm{r}}^{\lambda})^{3/2}(1-\gamma)^5}\Big) \log\Big(\frac{J_{\mathrm{r}}^{\star} - J_{\mathrm{r}}(\theta^0)}{\epsilon}\Big).$$

This result follows from the combination of Lemma 4.1 and PL-inequality (8). This result proves that `RS-FedPG` achieve a communication complexity that scales only logarithmically with the desired accuracy while guaranteeing linear speedup until $M \approx \min\Big(\frac{1+\lambda\log(|\mathcal{A}|)}{(\mu_{\mathrm{r}}^{\lambda})^{1/2}\epsilon^{1/2}(1-\gamma)^2}, \frac{(1+\lambda\log(|\mathcal{A}|))^2}{\mu_{\mathrm{r}}^{\lambda}\epsilon B(1-\gamma)^4}\Big)$.

4.4 Analysis of `b-RS-FedPG`

In Corollaries 4.3 and 4.4, we require the non-uniform Łojasiewicz constant to be bounded away from zero almost surely—a restrictive assumption that is difficult to verify in practice.

In Appendix E.3, we show that when $|\mathcal{A}| = 2$, the gradient field $\nabla J_{\mathrm{r},c}$ is *radial* outside a ball, for all $\|\theta\| \geq R$, i.e. $\langle \nabla J_{\mathrm{r},c}(\theta), \theta \rangle < 0$ so that $\nabla J_{\mathrm{r},c}$ always pushes θ toward the origin (see Lemma E.16 for a precise statement). This radiality allows us to identify a bounded region onto which projecting the global iterates of `RS-FedPG` *increases* the value of the local objective J_{r} (the easy argument is given in Lemma E.19). By enforcing this projection at every round, we ensure the policy parameters remain uniformly bounded away from the simplex boundary while improving J_{r}. Unlike convex problems—where projection on appropriate sets preserves or improves convex objectives—this behavior is highly non-trivial in the non-convex setting.

Exploiting Lemma E.10, we then derive an explicit lower bound on $\inf_{r\in[R]} \mu_{\mathrm{r}}^{\lambda}(\theta^r)$. Notably, as shown in Remark E.18, this radiality property fails for $|\mathcal{A}| \geq 3$, so analogous projection-based guarantees cannot be extended to larger action spaces.

We now present `b-RS-FedPG`, a novel federated policy gradient method that uses bit-level parameterization in combination with regularization. The idea of our approach is to reduce the general problem of policy optimization in MDP with $|\mathcal{A}| \geq 3$ to an equivalent problem in 2-action MDP. To do it, we frame the action selection process as a sequence of binary decisions over action-encoding

bits, thus reducing the original FRL problem to a simpler two-action problem. This reformulation allows to establish explicit lower bounds for the Łojasiewicz constant with the help of the projection.

Without loss of generality, let us consider an FRL instance $(\mathcal{M}^c)_{c\in[M]}$ with $|\mathcal{A}| = 2^{k}$[1]. Our goal is to build a *bit-level* FRL instance $(\bar{\mathcal{M}}_c := (\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{\gamma}, \bar{\mathsf{P}}_c, \bar{\mathsf{r}}))_{c\in[M]}$ with exactly two actions, such that solving the original FRL task with $(\mathcal{M}_c)_{c\in[M]}$ for stationary policies is reduced to solving this simpler instance. Let $\Sigma := \{0, 1\}$ be the binary alphabet. Consider $\Sigma^* = \bigcup_{k=0}^{\infty} \Sigma^k$, the set of finite words (including the empty word). The length of $w \in \Sigma^*$ is $|w|$. Concatenation of $w, w' \in \Sigma^*$ is denoted $w \circ w'$. For $w = w_1 \circ w_2 \circ \cdots \circ w_{|w|}$, define its prefix of length $k \leq |w|$ as $w_{:k} = w_1 \circ w_2 \circ \cdots \circ w_k$. Finally, let $\Sigma^{<k}$ denote all words shorter than k. Then, since $|\mathcal{A}| = 2^k$, we can associate the action space of the FRL instance $(\mathcal{M}_c)_{c\in[M]}$ with a set Σ^k of binary words of length exactly k, and define the corresponding action as a_w for $w \in \Sigma^k$. Conversely, we define $w(a)$ as the word associated with action a. Now, consider an FRL instance with a state space $\bar{\mathcal{S}}$ defined by $\bar{\mathcal{S}} := \mathcal{S} \times \Sigma^{<k}$, and with an action space $\bar{\mathcal{A}}$, given by the binary alphabet, i.e $\bar{\mathcal{A}} := \Sigma$.

For a given $c \in [M]$, the transition kernel of agent c in this bit-level FRL instance is defined as:

$$\bar{\mathsf{P}}_c((s', w')|(s, w), \bar{a}) := \begin{cases} \mathsf{P}_c(s'|s, a_{w\circ\bar{a}}) \cdot \mathbb{1}(w' = \emptyset) & \text{if } |w| = k - 1, \\ \mathbb{1}((s', w') = (s, w \circ \bar{a})) & \text{otherwise .} \end{cases}$$

In this bit-level FRL instance, the discount factor γ must be rescaled to reflect that the states in which the original FRL instance is embedded are k times further apart. We define the rescaled discount factor as $\bar{\gamma} := \gamma^{1/k}$ and the reward function as follows: $\bar{\mathsf{r}}((s, w), \bar{a}) := \bar{\gamma}^{-(k-1)}\mathsf{r}(s_i, a_{w\circ\bar{a}})$ if $|w| = k-1$ and 0 otherwise. For a given parameter $\theta \in \mathbb{R}^{|\bar{\mathcal{S}}|\times|\bar{\mathcal{A}}|}$, we define the following softmax policy in the extended environment as

$$\bar{\pi}_\theta(i|(s, w)) := \frac{\exp(\theta((s, w), i))}{\exp(\theta((s, w), 1)) + \exp(\theta((s, w), 0))}, \quad i \in \Sigma = \{0, 1\}$$

Drawing inspiration for auto-regressive sequence modeling, we can define the following corresponding policy in the original FRL instance $\pi_\theta(a_w|s) := \prod_{p=1}^{k} \bar{\pi}_\theta(w_p|(s, w_{:p}))$,

where $\theta = (\theta(i, w), i \in \Sigma, w \in \Sigma^{<k})$. Compared to a usual softmax parameterization, this bit-level softmax parameterization allows to execute a policy π_θ using only $k = \log_2(|\mathcal{A}|)$ operations instead of $|\mathcal{A}|$, that is very useful in the case of large action spaces that typically appear in the case of language modeling or recommendation systems. Define the bit-entropy regulariser as

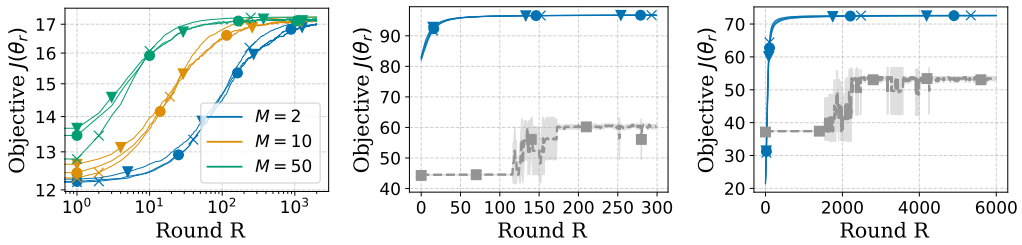
$$\mathcal{H}^{\rho}_{\mathrm{b},c}(\theta) := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t h^\theta_{\mathrm{b}}(S^t_c, A^t_c)\middle| S^0_c \sim \rho\right], \;\; h^\theta_{\mathrm{b}}(s, a) := -\sum_{p=0}^{k-1} \bar{\gamma}^p \log \bar{\pi}_\theta(w(a)_p|(s, w(a)_{:p})) \;\; .$$

Finally, denote by $\widetilde{V}^\theta_{\mathrm{b},c}(s) = V^{\pi_\theta}_c(s) + \lambda\mathcal{H}^\theta_{\mathrm{b},c}(s)$ and by $\bar{V}^\theta_c$ the entropy-regularized value function in this bit-level MDP associated to the c-the agent and to the policy $\bar{\pi}_\theta$. Proposition F.1 (stated and proved in the supplement) shows that $\bar{V}^\theta_c$ coincide with $\widetilde{V}^\theta_{\mathrm{b},c}$ demonstrating the consistency of our proposed formulation.

Building upon this FRL reduction method, we propose `b-RS-FedPG`, a special instance of `FedPG`, in which the local objective function is $J_{\mathrm{b},c} := J_{\mathrm{sm},c} + \lambda\mathcal{H}^\rho_{\mathrm{b},c}$ and global objective is defined as $J_{\mathrm{b}} := 1/M \sum_{c=1}^{M} J_{\mathrm{b},c}$. The key motivation for considering this specific local objective is that they match the local functions of `RS-FedPG` when dealing with a two-action FRL instance, allowing to provably establish the existence of a bounded set on which projecting increases the value of the objective J_{b}.

We now design the stochastic estimator of the gradient so that it matches the bit-entropy regularized stochastic estimator that would have been used in the bit-level MDP by `RS-FedPG`. For any parameter $\theta = \mathbb{R}^{|\bar{\mathcal{S}}|\times|\bar{\mathcal{A}}|}$, and given $Z_c \sim [\nu_c(\theta)]^{\otimes B}$ (defined at (3)), the biased estimator of the stochastic

[1]This property can be assumed without loss of generality by adding new artificial action identical to some fixed $a \in \mathcal{A}$.



(a) `Synthetic`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.3$ (b) `Synthetic`, $H = 5$, $\varepsilon_{\mathsf{P}} \gg 0.3$ (c) `GridWorld`, $H = 5$, $\varepsilon_{\mathsf{P}} \gg 0.3$

Figure 1: Comparison of `S-FedPG` (crosses), `RS-FedPG` (circles), `b-RS-FedPG` (triangles), and `Fed-Q-learning` (squares): (a) Value of the global objective $J(\theta^r)$ in the `Synthetic` environment, for the three `FedPG` variants and different numbers of agents $M \in \{2, 10, 50\}$, shown on a log-log scale; (b) Value of $J(\theta^r)$ in the `Synthetic` environment, comparing all four algorithms; (c) Value of $J(\theta^r)$ in the `GridWorld` environment, comparing all four algorithms.

gradient is chosen to be:

$$\mathrm{g}_{\mathrm{b},c}^{Z_c}(\theta) := \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{k(T-1)}\bar{\gamma}^t\left(\sum_{\ell=0}^{t}\nabla\log\bar{\pi}_\theta(w(A_{c,b}^{\mathrm{p}_t})_{\mathrm{q}_t}|(S_{c,b}^{\mathrm{p}_t}, w(A_{c,b}^{\mathrm{p}_t})_{:\mathrm{q}_t}))\right)R_{c,b}^t \ , \tag{10}$$

where $\mathrm{p}_t = \lfloor t/k \rfloor$, and $\mathrm{q}_t = t - k\lfloor t/k \rfloor$, and

$$R_{c,b}^t \triangleq \left[1_{\{\mathrm{q}_t = k-1\}}\mathtt{r}(S_{c,b}^{\mathrm{p}_t}, A_{c,b}^{\mathrm{p}_t}) - \lambda\log\bar{\pi}_\theta(w(A_{c,b}^{\mathrm{p}_t})_{\mathrm{q}_t}|(S_{c,b}^{\mathrm{p}_t}, w(A_{c,b}^{\mathrm{p}_t})_{:\mathrm{q}_t}))\right] \ .$$

Next, we establish the sample and communication complexity for `b-RS-FedPG` with explicit constants. Importantly, `b-RS-FedPG` achieves the linear speedup until a certain threshold (similar to that of `RS-FedPG`) with a logarithmic communication complexity with respect to the desired accuracy ϵ.

Corollary 4.5 (Sample and Communication Complexity of `b-RS-FedPG`). *Assume **A**-1 and **A**-2. Define $\bar{\gamma} = \gamma^{1/\log_2(|\mathcal{A}|)}$ and*

$$\mu_{\mathrm{b}}^{\lambda} \triangleq \frac{\gamma^3\lambda(1-\bar{\gamma})}{4^{\log(|\mathcal{A}|)}|\mathcal{S}|}\min_s \rho(s)^2\exp\left(-4\log(|\mathcal{A}|)\cdot\frac{1+\lambda\log(2)}{\lambda(1-\bar{\gamma})}\right) \ .$$

Set $\theta^0 = (0, \ldots, 0)^\top$ and let $\epsilon \gtrsim (1+\lambda)^2\varepsilon_{\mathsf{P}}^2(\mu_{\mathrm{b}}^{\lambda})^{-1}(1-\bar{\gamma})^{-6}$. Then, for a properly chosen truncation horizon, a properly chosen projection set $\mathcal{T}$, a properly chosen step size and number of local updates, `b-RS-FedPG` learns an ϵ-approximation of the optimal objective $J_{\mathrm{b}}^\star$ with a number of communication rounds

$$R \gtrsim \frac{(1+\lambda)}{(1-\bar{\gamma})^2\mu_{\mathrm{b}}^{\lambda}}\log\left(\frac{5(J_{\mathrm{b}}^\star - J_{\mathrm{b}}(\theta^0))}{\epsilon}\right) \ ,$$

for a total number of samples per agent of

$$RHB \gtrsim \max\left(\frac{(1+\lambda\log(2))B}{\mu_{\mathrm{b}}^{\lambda}(1-\bar{\gamma})^3}, \frac{(1+\lambda)^3}{(\mu_{\mathrm{b}}^{\lambda})^2\epsilon M(1-\bar{\gamma})^7}, \frac{(1+\lambda)^2}{\epsilon^{1/2}(\mu_{\mathrm{b}}^{\lambda})^{3/2}(1-\bar{\gamma})^5}\right)\log\left(\frac{J_{\mathrm{b}}^\star - J_{\mathrm{b}}(\theta^0)}{\epsilon}\right) \ .$$

5 Experiments

We study the empirical performance of the three proposed methods on two environments, that satisfy **A**-1 and **A**-2, and illustrate their advantage over `Fed-Q-learning` in heterogeneous settings. In the two problems, the transition kernel for agent c can be decomposed as a mixture of two kernels: are modeled as a mixture of two components: $\mathsf{P}_c = (1-\varepsilon_{\mathsf{P}})\mathsf{P}^{\mathsf{com}} + \varepsilon_{\mathsf{P}}\mathsf{P}_c^{\mathsf{ind}}$ where $\mathsf{P}^{\mathsf{com}}$ is a common kernel, and $\mathsf{P}_c^{\mathsf{ind}}$ is an individual kernel specific to each agent. Details are given in Appendix H.

In Figure 1a, we illustrate the *linear speedup* by evaluating the three variants of `FedPG` in a heterogeneous environment. Specifically, we report the global objective J during the learning process for various numbers of agents, using the theoretically motivated step size. Empirically, all three algorithms achieve the speedup, thereby highlighting the benefits of collaboration even among heterogeneous agents. In Figures 1b and 1c, we compare the performance of `Fed-Q-learning` and the three variants on two highly heterogeneous FRL problems. The three algorithms *learn better policies*, demonstrating, as suggested by Theorem 3.1, the advantage of learning a stochastic policy.

6 Conclusion

This work extends the theoretical foundations of FRL in heterogeneous environments. It identifies structural differences that challenge classical RL properties and shows that deterministic, stationary strategies can be suboptimal. Our main contribution is the first global convergence guarantee for both non-regularized and entropy-regularized policy gradient methods in heterogeneous FRL. We also introduce a new algorithm, `b-RS-FedPG`, which combines a softmax-inspired parameterization with tailored regularization, and for which we derive explicit convergence rates. Experiments on two FRL benchmarks support our theoretical findings and demonstrate the effectiveness of `FedPG` across varying levels of heterogeneity. A promising direction for future work is to establish exact convergence to the optimal policy by developing new methods that correct for heterogeneity.

Acknowledgements

The work of S. Labbi, and P. Mangold has been supported by Technology Innovation Institute (TII), project Fed2Learn. The work of D.Tiapkin has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF. The work of E. Moulines has been partly funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2020). Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR.

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

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR.

Chen, T., Zhang, K., Giannakis, G. B., and Başar, T. (2021). Communication-efficient policy gradient methods for distributed reinforcement learning. *IEEE Transactions on Control of Network Systems*, 9(2):917–929.

Demidovich, Y., Ostroukhov, P., Malinovsky, G., Horváth, S., Takáč, M., Richtárik, P., and Gorbunov, E. (2025). Methods with local steps and random reshuffling for generally smooth nonconvex federated optimization. In *The Thirteenth International Conference on Learning Representations*.

Ding, Y., Zhang, J., Lee, H., and Lavaei, J. (2025). Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *IEEE Transactions on Automatic Control*.

Domingues, O. D., Flet-Berliac, Y., Leurent, E., Ménard, P., Shang, X., and Valko, M. (2021). rlberry - A Reinforcement Learning Library for Research and Education.

Fan, X., Liu, H., and Yang, Q. (2023). Federated learning with heterogeneous data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):297–317.

Ganesh, S., Chen, J., Thoppe, G., and Aggarwal, V. (2024). Global convergence guarantees for federated policy gradient methods with adversaries. *Transactions on Machine Learning Research*.

Geist, M., Scherrer, B., and Pietquin, O. (2019). A theory of regularized Markov decision processes. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2160–2169. PMLR.

Glasgow, M. R., Yuan, H., and Ma, T. (2022). Sharp bounds for federated averaging (local SGD) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. (2019). Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32.

Haddadpour, F. and Mahdavi, M. (2019). On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. (2022). Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 18–37. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Khodadadian, S., Vahdat, A., Mohri, M., and Talwalkar, A. (2022). Federated reinforcement learning via adaptive averaging and quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 11692–11704.

Labbi, S., Tiapkin, D., Mancini, L., Mangold, P., and Moulines, E. (2025). Federated ucbvi: Communication-efficient federated regret minimization with heterogeneous agents. In Li, Y., Mandt, S., Agrawal, S., and Khan, E., editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1315–1323. PMLR.

Lan, G., Wang, H., Anderson, J., Brinton, C., and Aggarwal, V. (2023). Improved communication efficiency in federated natural policy gradient via admm-based gradient updates. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 59873–59885. Curran Associates, Inc.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Mei, J., Dai, B., Agarwal, A., Vaswani, S., Raj, A., Szepesvari, C., and Schuurmans, D. (2024). Small steps no more: Global convergence of stochastic gradient bandits for arbitrary learning rates. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Mei, J., Dai, B., Xiao, C., Szepesvari, C., and Schuurmans, D. (2021). Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA. PMLR.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30.

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition.

Osękowski, A. (2012). *Sharp martingale and semimartingale inequalities*, volume 72. Springer Science & Business Media.

Puterman, M. L. (1994). *Discounted Markov Decision Problems*, chapter 6, pages 142–276. John Wiley and Sons, Ltd.

Qi, J., Zhou, Q., Lei, L., and Zheng, K. (2021). Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence and Robotics*, 1(1).

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32.

Salgia, S. and Chi, Y. (2024). The sample-communication complexity trade-off in federated Q-learning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 39694–39747. Curran Associates, Inc.

Schulman, J., Chen, X., and Abbeel, P. (2017). Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Wang, H., He, S., Zhang, Z., Miao, F., and Anderson, J. (2024a). Momentum for the win: Collaborative federated reinforcement learning across heterogeneous environments. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50530–50560. PMLR.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. (2021). A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*.

Wang, M., Yang, P., and Su, L. (2024b). On the convergence rates of federated Q-learning across heterogeneous environments. *arXiv preprint arXiv:2409.03897*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

Xiao, L. (2022). On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36.

Yang, T., Cen, S., Wei, Y., Chen, Y., and Chi, Y. (2024). Federated natural policy gradient and actor critic methods for multi-task reinforcement learning. *Advances in Neural Information Processing Systems*, 37:121304–121375.

Yuan, R., Gower, R. M., and Lazaric, A. (2022). A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR.

Zhang, C., Wang, H., Mitra, A., and Anderson, J. (2024). Finite-time analysis of on-policy heterogeneous federated reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. (2021a). Sample efficient reinforcement learning with reinforce. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10887–10895.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. (2021b). On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240.

Zheng, Z., Gao, F., Xue, L., and Yang, J. (2023). Federated Q-learning: Linear regret speedup with low communication cost. In *The Twelfth International Conference on Learning Representations*.

Zheng, Z., Zhang, H., and Xue, L. (2025). Federated Q-learning with reference-advantage decomposition: Almost optimal regret and logarithmic communication cost. In *The Thirteenth International Conference on Learning Representations*.

Zhu, F., Heath, R. W., and Mitra, A. (2024). Towards fast rates for federated and multi-task reinforcement learning. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 2658–2663. IEEE.

Zhu, L., Liu, X., Han, Y., Hu, S., Chen, Y., and Li, W. (2022). Federated learning: Challenges, methods, and future directions. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2684–2703.

Zhuo, H., Liu, Y., Huang, J., Zhang, T., Chen, X., Li, P., and Zhou, J. (2023). Federated reinforcement learning: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15745–15753.

Contents

A Notations

For clarity, we summarize here the notations that we use

Symbols	Meaning	Definition
$\mathcal{S}$	State space	Section 3
$\mathcal{A}$	Action space	Section 3
M	Number of agents	Section 3
P_c	Transition kernel of agent c	Section 3
r	Reward function	Section 3
γ	Discount factor	Section 3
ρ	Initial distribution over the state space	Section 3
J	Global FRL objective	Equation (1)
J_c	Local objective of agent c	Equation (1)
ε_{P}	Heterogeneity on the transition kernels	**A**-1
R	Number of Communication rounds of `FedPG`	Section 4.1
H	Number of local steps of `FedPG`	Section 4.1
T	Length of the sampled trajectories	Section 4.1
B	Number of trajectories collected per iteration	Section 4.1
$\mathcal{T}$	Projection set of `FedPG`	Section 4.1
π_θ	Stationary policy parametrized by $\theta \in \Theta$	Section 4.1
$\nu_c(\theta)$	Distribution of sampled trajectory by agent c	Equation (3)
F	Global objective optimised by `FedPG`	Section 4.1
f_c	Local function of agent c in `FedPG`	Section 4.1
V_c^π	Value function of agent c under policy π	Equation (38)
Q_c^π	Q-function of agent c under policy π	Equation (39)
$d_c^{\rho,\pi}$	Occupancy measure of agent c under π	Equation (40)
A_c^π	Advantage function of agent c under π	Equation (41)
λ	Regularization temperature	Section 4.3
$\widetilde{V}_c^\pi$	Regularized value function of agent c	Equation (49)
$\widetilde{Q}_c^\pi$	Regularized Q-function of agent c	Equation (50)
$\widetilde{A}_c^\pi$	Regularized Advantage function of agent c	Equation (51)
$J_{\mathrm{sm},c}$, $J_{\mathrm{r},c}$, and $J_{\mathrm{b},c}$	Local functions in `S-FedPG`, `RS-FedPG`, and `b-RS-FedPG`	Sections 4.2 to 4.4
J_{sm}, J_{r}, and J_{b}	Respective global functions	Sections 4.2 to 4.4
$\mathrm{g}^Z_{\mathrm{sm},c}(\theta)$, $\mathrm{g}^Z_{\mathrm{r},c}(\theta)$, and $\mathrm{g}^Z_{\mathrm{b},c}(\theta)$	Stochastic estimators of the gradient at θ in `S-FedPG`, `RS-FedPG`, and `b-RS-FedPG`	Equations (6), (9) and (10)
$L_{1,\mathrm{sm}}$, $L_{1,\mathrm{r}}$, and $L_{1,\mathrm{b}}$	Bound on the gradients of J_{sm}, J_{r}, and J_{b}	Lemmas D.4 and E.4 and Appendix F.2
$L_{2,\mathrm{sm}}$, $L_{2,\mathrm{r}}$, and $L_{2,\mathrm{b}}$	Smoothness constants of J_{sm}, J_{r}, and J_{b}	Lemmas D.3 and E.3 and Appendix F.2
$L_{3,\mathrm{sm}}$, $L_{3,\mathrm{r}}$, and $L_{3,\mathrm{b}}$	Bounds on the third-order derivative tensors of J_{sm}, J_{r}, and J_{b}	Lemmas D.5 and E.5 and Appendix F.2
$\sigma^p_{\mathrm{sm},p}$, $\sigma^p_{\mathrm{r},p}$, and $\sigma^p_{\mathrm{b},p}$	Bounds on the p-th central moments of $\mathrm{g}^Z_{\mathrm{sm},c}(\theta)$, $\mathrm{g}^Z_{\mathrm{r},c}(\theta)$, and $\mathrm{g}^Z_{\mathrm{b},c}(\theta)$ for $p \in \{2,4\}$	Lemmas D.7 and E.7 and Appendix F.2
β_{sm}, β_{r}, and β_{b}	Bounds on bias of $\mathrm{g}^Z_{\mathrm{sm},c}(\theta)$, $\mathrm{g}^Z_{\mathrm{r},c}(\theta)$, and $\mathrm{g}^Z_{\mathrm{b},c}(\theta)$.	Lemmas D.7 and E.7 and Appendix F.2
ζ_{sm}, ζ_{r}, and ζ_{b}	Bound on the gradient heterogeneity of the objectives J_{sm}, J_{r}, and J_{b}	Lemmas D.6 and E.6 and Appendix F.2
μ_{sm}, μ^λ_{r}, and μ^λ_{b}	Minimal Łojasiewicz coefficient over the agents of of $J_{\mathrm{sm},c}$, $J_{\mathrm{r},c}$, and $J_{\mathrm{b},c}$	Lemmas D.10 and E.11 and Appendix F.2

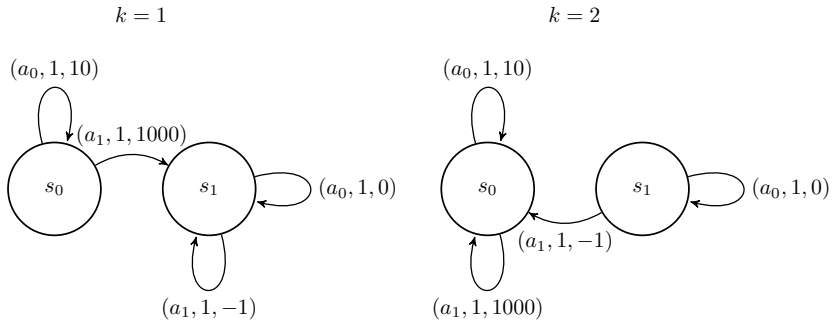


Figure 2: FRL task with no optimal local history-dependant policy. The triplet means (action, probability, reward) and $\gamma = 0.9$. Note that these two environments share the same action space, same state space, and same reward function.

The cardinality (the number of elements) of a set Y is denoted $|Y|$. We define the indicator function of an element $y \in Y$ as

$$\begin{aligned} \mathbf{1}_y(\cdot)\colon Y &\longrightarrow \{0,1\} \\ w &\longmapsto \begin{cases} 1 & \text{if } w = y \\ 0, & \text{otherwise} \end{cases} . \end{aligned}$$

$\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product. For a three-times differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we denote $\nabla f \in \mathbb{R}^d$ its gradient, $\nabla^2 f \in \mathbb{R}^{d\times d}$ its Hessian and $\nabla^3 f \in \mathbb{R}^{d\times d\times d}$ its third-order derivative tensor. and $X^{\otimes k}$ the k-th tensor power of a tensor X. For two real-valued sequences $(a_r)_{r=0}^{\infty}$ and $(b_r)_{r=0}^{\infty}$, we write $a_r \lesssim b_r$ if there exists a constant $C > 0$ such that $a_r \leq C b_r$ for any $r \geq 0$.

B On the different classes of policies

The goal of this section is to prove Theorem 3.1. For clarity and readability, we prove each statement of the theorem in a separate lemma. First, we define the value function of an agent $c \in [M]$, of a policy $\pi \in \Pi$, and for an initial distribution ρ as

$$V_c^{\pi}(\rho) \triangleq \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathsf{r}(S_c^t, A_c^t)\right] , \tag{11}$$

where $\mathbb{E}_{\pi}[\cdot]$ is the expectation over random trajectories generated by following a policy $\pi = (\pi^t)_{t\in\mathbb{N}}$: the initial state is sampled from a distribution $S_c^0 \sim \rho(\cdot)$ and $\forall t \geq 0 : A_c^t \sim \pi^t(\cdot|\mathcal{H}^t, c), S_c^{t+1} \sim \mathsf{P}_c(\cdot|S_c^t, A_c^t)$, for $\mathcal{H}^t = (\mathcal{H}_c^t)_{c\in[M]}$ where $\mathcal{H}_c^t = (S_c^0, A_c^0, \ldots, S_c^t)$ for all $c \in [M]$.

Lemma B.1. *There exists an FRL instance such that any local history-dependent policy is suboptimal with respect to some agent-aware history-dependent policy.*

Proof. We consider the same FRL instance as in Theorem 1 of Jin et al. (2022) that is represented in Figure 2 with $\rho = (1/2, 1/2)$.

We show here that it holds

$$\max_{\pi\in\Pi_\ell} \frac{1}{2}\left(V_1^{\pi}(\rho) + V_2^{\pi}(\rho)\right) < \max_{\pi\in\Pi} \frac{1}{2}\left(V_1^{\pi}(\rho) + V_2^{\pi}(\rho)\right) .$$

Let $\pi_\ell^\star$ be an optimal local history-dependant policy. Define $\pi_\ell^{(0,0)}$, $\pi_\ell^{(1,0)}$, $\pi_\ell^{(0,1)}$ and $\pi_\ell^{(1,1)}$ as local history-dependant policies that maximises respectively $V_1^{\pi}(s_0) + V_2^{\pi}(s_0)$, $V_1^{\pi}(s_1) + V_2^{\pi}(s_0)$, $V_1^{\pi}(s_0) + V_2^{\pi}(s_1)$, and $V_1^{\pi}(s_1) + V_2^{\pi}(s_1)$ on the set of local history-dependant policies Π_ℓ. Now, define the following agent-aware history-dependant policy

$$\pi = \mathbf{1}_{(s_0,s_0)}(s_1^0, s_2^0)\pi_\ell^{(0,0)} + \mathbf{1}_{(s_0,s_1)}(s_1^0, s_2^0)\pi_\ell^{(0,1)} + \mathbf{1}_{(s_1,s_0)}(s_1^0, s_2^0)\pi_\ell^{(1,0)} + \mathbf{1}_{(s_1,s_1)}(s_1^0, s_2^0)\pi_\ell^{(1,1)} .$$

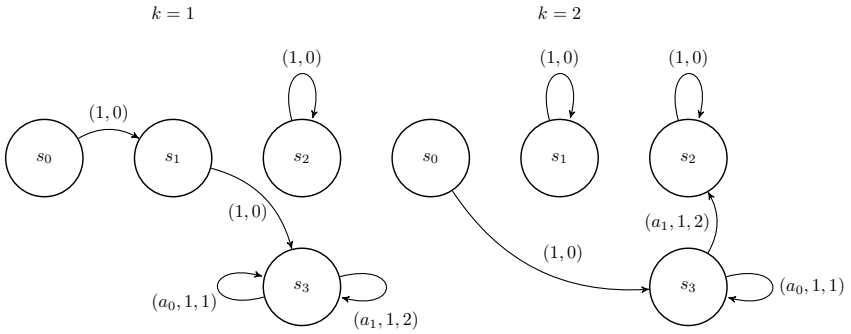


Figure 3: FRL task with no optimal stationary policy. The triplet means (action, probability, reward) and $\gamma = 0.9$. If the action is not specified, it means that all the actions give the same reward and lead to the same state

Denote by $p = \pi_\ell^\star(a_0|s_0)$ and $q = \pi_\ell^\star(a_0|s_1)$. We distinguish the two following cases:

Case $q = 1$: In this case we have

$$\frac{1}{2}\left(V_1^{\pi_\ell^\star}(s_1) + V_2^{\pi_\ell^\star}(s_1)\right) < \frac{1}{2}\left(V_1^{\pi_\ell^{(1,1)}}(s_1) + V_2^{\pi_\ell^{(1,1)}}(s_1)\right) \quad , \tag{12}$$

as $q = 1$ means that the second agent will not select a_1 at s_1 at $t = 0$ and thus will not reach s_0 at $t = 1$ in the second environment which is suboptimal. Thus we have

$$\begin{aligned}
\frac{1}{2}\left(V_1^{\pi_\ell^\star}(\rho) + V_2^{\pi_\ell^\star}(\rho)\right) &= \frac{1}{4}\left[V_1^{\pi_\ell^\star}(s_0) + V_2^{\pi_\ell^\star}(s_0) + V_1^{\pi_\ell^\star}(s_1) + V_2^{\pi_\ell^\star}(s_1)\right] \\
&< \frac{1}{4}\left[V_1^{\pi_l^{(0,0)}}(s_0) + V_2^{\pi_\ell^{(0,0)}}(s_0) + V_1^{\pi_\ell^{(1,1)}}(s_1) + V_2^{\pi_\ell^{(1,1)}}(s_1)\right] \\
&\leq \frac{1}{2}\left(V_1^{\pi}(\rho) + V_2^{\pi}(\rho)\right) \quad ,
\end{aligned}$$

where the before last inequality is a consequence of (12) and the optimality of the policy $\pi_\ell^{(1,1)}$ on the set Π_ℓ when the starting point is s_1 for both agents.

Case $q < 1$: If $(s_1^0, s_2^0) = (s_1, s_0)$ then $q < 1$ is strictly suboptimal as the first agent selects with a non-zero probability the action a_1 that leads to a -1 reward while the action a_0 leads to a better reward. Thus we have

$$\begin{aligned}
\frac{1}{2}\left(V_1^{\pi_\ell^\star}(\rho) + V_2^{\pi_\ell^\star}(\rho)\right) &= \frac{1}{4}\left[V_1^{\pi_\ell^\star}(s_1) + V_2^{\pi_\ell^\star}(s_0) + V_1^{\pi_\ell^\star}(s_0) + V_2^{\pi_\ell^\star}(s_1)\right] \\
&< \frac{1}{4}\left[V_1^{\pi_\ell^{(1,0)}}(s_0) + V_2^{\pi_\ell^{(1,0)}}(s_0) + V_1^{\pi_\ell^{(0,1)}}(s_1) + V_2^{\pi_\ell^{(0,1)}}(s_1)\right] \\
&\leq \frac{1}{2}\left(V_1^{\pi}(\rho) + V_2^{\pi}(\rho)\right) \quad ,
\end{aligned}$$

where the before last inequality is a consequence of the suboptimality of $\pi_\ell^\star$ as $q < 1$ and the optimality of the policy $\pi_\ell^{(0,1)}$ on the set Π_l when the starting point is s_1 for both agents. □

Lemma B.2. *There exists an FRL instance such that any local stochastic stationary policy is sub-optimal with respect to some local history-dependent policy.*

Proof. We consider the FRL task described in Figure 3 with $\rho = (1, 0, 0, 0)$. We show here that it holds

$$\max_{\pi \in \Pi_{\text{sta}}} \frac{1}{2}\left(V_1^{\pi}(s_0) + V_2^{\pi}(s_0)\right) < \max_{\pi \in \Pi_\ell} \frac{1}{2}\left(V_1^{\pi}(s_0) + V_2^{\pi}(s_0)\right) \quad .$$

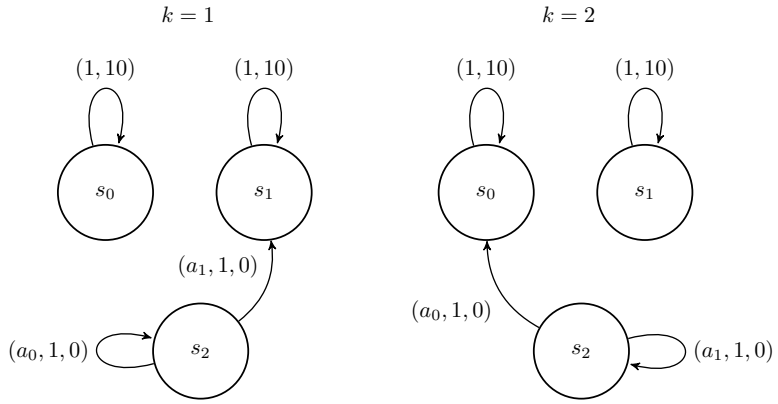


Figure 4: FRL task with no optimal local deterministic policy. The triplet means (action, probability, reward) and $\gamma = 0.9$. If the action is not specified, it means that all the actions give the same reward and lead to the same state

Define the following local history-dependent policy $\pi_\ell = (\pi_\ell^t)_{t\in\mathbb{N}}$ that satisfies

$$\pi_\ell^1(a_0|s_3) = 1 \cdot 1_{t=1} + 0 \cdot 1_{t\geq 2} \ ,$$

which intuitively describes the policy that takes action a_0 at the instant where the second agent reaches the state s_3 and then takes action a_1 when the first agent reaches the state s_3. The (double of the) FRL objective of this policy is equal

$$V_1^{\pi_\ell}(s_0) + V_2^{\pi_\ell}(s_0) = \frac{2\gamma^2}{1-\gamma} + \gamma + 2\gamma^2 \ .$$

Let $\pi_{\text{sta}}^\star$ be a local stationary policy that maximizes $V_1^\pi(s_0) + V_2^\pi(s_0)$ on the set of the local stationary policies Π_{sta}. We define $p = \pi_{\text{sta}}^\star(a_0|s_3)$. The (double of the) federated objective for this policy is

$$V_1^{\pi_{\text{sta}}^\star}(s_0) + V_2^{\pi_{\text{sta}}^\star}(s_0) = \sum_{k=2}^{\infty} \gamma^k (1 \cdot p + 2 \cdot (1-p)) + V_2^{\pi_{\text{sta}}^\star}(s_0) \ .$$

The first instant at which the second agent takes actions a_1 follows a geometric distribution of parameter $1-p$. Thus, we have

$$\begin{aligned}
V_2^{\pi_{\text{sta}}^\star}(s_0) &= \gamma \sum_{k=0}^{\infty} \left((1-p)^k p \cdot \left(\sum_{i=0}^{k-1} \gamma^i \cdot 1 + 2\gamma^k \right) \right) \\
&= \gamma \sum_{k=0}^{\infty} \left((1-p)^k p \cdot \left(\frac{1-\gamma^k}{1-\gamma} + 2\gamma^k \right) \right) \\
&= \frac{\gamma}{1-\gamma} \sum_{k=0}^{\infty} \left((1-p)^k p \cdot (1 - \gamma^k + 2\gamma^k - 2\gamma^{k+1}) \right) \\
&= \frac{\gamma}{1-\gamma} \sum_{k=0}^{\infty} \left((1-p)^k p \cdot (1 + \gamma^k - 2\gamma^{k+1}) \right) \\
&= \frac{\gamma p}{1-\gamma} \sum_{k=0}^{\infty} \left((1-p)^k + ((1-p)\gamma)^k - 2\gamma((1-p)\gamma)^k \right) \\
&= \frac{\gamma p}{1-\gamma} \left(\frac{1}{p} + \frac{1}{1-\gamma+p\gamma} - \frac{2\gamma}{1-\gamma+p\gamma} \right) \ .
\end{aligned}$$

By gathering the two precedent expressions, we get

$$\begin{aligned} V_1^{\pi^\star_{\rm sta}}(s_0) + V_2^{\pi^\star_{\rm sta}}(s_0) &= \frac{(2-p)\gamma^2}{1-\gamma} + \frac{\gamma p}{1-\gamma}\left(\frac{1}{p} + \frac{1}{1-\gamma+p\gamma} - \frac{2\gamma}{1-\gamma+p\gamma}\right) \\ &= \frac{1}{1-\gamma}\left[(2-p)\gamma^2 + \gamma + (1-2\gamma)\frac{\gamma p}{1-\gamma+p\gamma}\right] \\ &= \frac{1}{1-\gamma}\left[(2-p)\gamma^2 + \gamma + (1-2\gamma) - (1-2\gamma)\frac{1-\gamma}{1-\gamma+p\gamma}\right] \\ &\leq \frac{1}{1-\gamma}\left[2\gamma^2 + (1-\gamma) - (1-2\gamma)\frac{1-\gamma}{1-\gamma+p\gamma}\right] \leq \frac{2\gamma^2}{1-\gamma} + 2\gamma \ , \end{aligned}$$

where the last inequality holds as $\gamma > 1/2$. As for any $\gamma > 1/2$, we have $2\gamma < \gamma + 2\gamma^2$ then this proves the suboptimality of the stationary policy $\pi^\star_{\rm sta}$ with respect to the local history-dependent policy π_ℓ. □

Lemma B.3. *There exists an FRL instance such that any local deterministic policy is suboptimal with respect to some local stationary stochastic policy.*

Proof. We consider the FRL task of Figure 4, and we consider the setting where the two agents start from the state s_2, i.e, $\rho = (0, 0, 1)$. We show here that it holds

$$\max_{\pi \in \Pi_{\rm det}} \frac{1}{2}\left(V_1^\pi(s_2) + V_2^\pi(s_2)\right) < \max_{\pi \in \Pi_{\rm sta}} \frac{1}{2}\left(V_1^\pi(s_2) + V_2^\pi(s_2)\right) \ .$$

We define the stationary policy $\pi_{\rm sta}$ that satisfies $\pi_{\rm sta}(a_0|s_2) = 1/2$ and $\pi_{\rm sta}(a_1|s_2) = 1/2$. First, note that the probability of each agent being in state s_2 at time t, while following $\pi_{\rm sta}$, is $1/2^t$. Thus, the FRL objective of this policy is equal to

$$\frac{1}{2}\left(V_1^{\pi_{\rm sta}}(s_2) + V_2^{\pi_{\rm sta}}(s_2)\right) = \frac{10\gamma}{1-\gamma} - \frac{10\gamma}{1-\gamma/2} = \frac{6\gamma}{1-\gamma} + \frac{2\gamma^2 - 6\gamma(1-\gamma)}{(1-\gamma)(1-\gamma/2)} \geq \frac{6\gamma}{1-\gamma} \ ,$$

where the last inequality follows from the fact that $\gamma = 0.9$. Let $\pi^\star_{\rm det}$ be an optimal deterministic policy. We distinguish two cases

Case $\pi^\star_{\rm det}(s_2) = a_0$: In this case, the second agent will reach state s_0 at first iteration, but the first agent will be stuck at s_2 where he will get no reward. Thus, the FRL objective for this policy is equal to

$$\frac{1}{2}\left(V_1^{\pi^\star_{\rm det}}(s_2) + V_2^{\pi^\star_{\rm det}}(s_2)\right) = \frac{1}{2}\sum_{t=1}^{\infty} 10\gamma^t = \frac{5\gamma}{1-\gamma} \ ,$$

proving that $\pi_{\rm sta}$ achieves a higher value than $\pi^\star_{\rm det}$.

Case $\pi^\star_{\rm det}(s_1) = a_1$: This case is similar to the previous one. □

Combining the three previous lemmas concludes the proof of Theorem 3.1.

B.1 Heterogeneous rewards

To further clarify the novelty of our setting, we contrast it with a commonly studied setup in the literature, often referred to as the federated multi-task RL setting, where agents share identical dynamics but differ in their reward functions. This setting has been explored in prior work Zhu et al. (2024); Chen et al. (2021); Yang et al. (2024). This setup does not introduce additional structural challenges and, thus, more closely aligns with the standard single-agent setting. In particular, when agents differ only in rewards, the optimal FRL objective over the space of history-dependent policies is achieved by a deterministic policy. The following lemma formalizes this observation:

Lemma B.4. *Let $\{\mathcal{M}_c\}_{c=1}^M$ be an FRL instance consisting of M MDPs that share the same transition kernel P and initial distribution ρ, but have distinct reward functions r_c. Denote by J the corresponding FRL objective. Then,*

$$\max_{\pi \in \Pi_{\rm det}} J(\pi) = \max_{\pi \in \Pi_\ell} J(\pi) \ ,$$

and furthermore, the FRL objective is equivalent to the RL objective of a single MDP with reward function equal to the average of the individual rewards.

Algorithm 2 `proj-FedAVG`

Initialization: Learning rate $\eta > 0$, parameter θ^0, projection set $\mathcal{T}$
for $r = 0$ to $R - 1$ **do**
  for $c = 1$ to M **do**
    Set $\theta_c^{r,0} = \theta^r$.
    for $h = 0$ to $H - 1$ **do**
      Receive random state $Z_c^{r,h+1}$
      Update $\theta_c^{r,h+1} = \theta_c^{r,h} + \eta \mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})$
  Server updates parameter: $\theta^{r+1} = \mathrm{proj}_{\mathcal{T}}(\bar{\theta}^{r+1})$ where $\bar{\theta}^{r+1} = \frac{1}{M}\sum_{c=1}^{M} \theta_c^{r,H}$

Proof. Consider an FRL instance where each agent's MDP is defined as $\mathcal{M}_c := (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}, \mathsf{r}_c, \rho)$. Let $\pi = (\pi^t)_{t\in\mathbb{N}} \in \Pi$ be an arbitrary history-dependent policy. Since all agents share the same transition kernel, their trajectories under π follow identical distributions. Precisely, for any $t \geq 0$ and $c \in [M]$, it holds that $(S_c^t, A_c^t) \sim (S_1^t, A_1^t)$. Thus, the FRL objective simplifies as:

$$\frac{1}{M}\sum_{c=1}^{M} J_c(\pi) = \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_\pi\left[\frac{1}{M}\sum_{c=1}^{M}\mathsf{r}_c(S_c^t, A_c^t)\right] = \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_\pi\left[\bar{\mathsf{r}}(S_1^t, A_1^t)\right] \ ,$$

where $\bar{\mathsf{r}} := \frac{1}{M}\sum_{c=1}^{M}\mathsf{r}_c$ denotes the average reward function. This expression corresponds to the standard RL objective of the MDP $(\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}, \bar{\mathsf{r}}, \rho)$. By (Agarwal et al., 2019, Theorem 1.7), the optimal value of this objective is attained by a deterministic policy, which concludes the proof. □

C Ascent lemma

Problem setting. In this section, we provide an ascent lemma that can be applied to a general class of distributed non-convex optimization problems of the form

$$\max_{\theta\in\mathbb{R}^d} F(\theta) = \frac{1}{M}\sum_{c=1}^{M} f_c(\theta) \ , \quad \text{where } f_c(\theta) := \mathbb{E}_{Z_c\sim\xi_c(\theta)}[f_c^{Z_c}(\theta)] \ , \tag{13}$$

where each Z_c is a random variable which takes its value from a distribution $\xi_c(\theta)$, which may depend on θ, and takes values in a measurable set $(\mathsf{Z}, \mathcal{Z})$, and where the function $(z, \theta) \mapsto f_c^z(\theta)$ are measurable functions. Each function f_c is only available to the client c through *biased* stochastic gradients $\mathrm{g}_c^z(\theta)$, whose expected value is

$$\mathrm{g}_c(\theta) := \mathbb{E}_{Z_c\sim\xi_c(\theta)}[\mathrm{g}_c^{Z_c}(\theta)] \ , \tag{14}$$

but is typically different from the gradient of f_c.

To solve (13), we use `proj-FedAVG`, an extension of projected gradient ascent to the federated setting, which performs local stochastic gradient updates at the client level with step size η, aggregates the locally updated model, and projects the resulting model on a closed convex set $\mathcal{T} \subseteq \mathbb{R}^d$. For completeness, we give the pseudo-code for this algorithm in Algorithm 2.

Assumptions. To derive our new convergence result for this algorithm under the following assumptions, which slightly differ from the classical setting, but are typical in reinforcement learning. First, we assume that both the true gradient and its biased estimator are Lispchitz-continuous, that the true gradient is bounded, and that the objective functions' third derivates are uniformly bounded.

FL-1. *For any $c \in [M]$, the functions ∇f_c and the biased gradients g_c are L_2-Lipschitz, that is*

$$\|\nabla f_c(\theta) - \nabla f_c(\theta')\| \leq L_2\|\theta - \theta'\| \ , \quad \textit{for all } \theta, \theta' \in \mathbb{R}^d \ , \tag{15}$$

$$\|\mathrm{g}_c(\theta) - \mathrm{g}_c(\theta')\| \leq L_2\|\theta - \theta'\| \ , \quad \textit{for all } \theta, \theta' \in \mathbb{R}^d \ . \tag{16}$$

FL-2. *There exists $L_1 > 0$, such that for all $c \in [M]$ and $\theta \in \mathbb{R}^d$,*

$$\|\nabla f_c(\theta)\| \leq L_1 \ , \quad \textit{for all } c \in [M] \ , \ \theta \in \mathbb{R}^d \ . \tag{17}$$

FL-3. *For any $c \in [M]$, the function f_c is three times differentiable and has bounded third derivative tensor, that is, there exists $L_3 < \infty$ such that*

$$\|\nabla^3 f_c(\theta) u^{\otimes 2}\| \leq L_3 \|u\|^2 \ , \quad \text{for all } \theta \in \mathbb{R}^d \ , \ u \in \mathbb{R}^d \ . \tag{18}$$

FL-4. *For any $c \in [M]$, the gradient gradient heterogeneity is uniformly bounded, that is, there exists $\zeta \geq 0$ such that*

$$\|\nabla f_c(\theta) - \nabla F(\theta)\| \leq \zeta \ , \quad \text{for all } c \in [M] \ , \ \theta \in \mathbb{R}^d \ . \tag{19}$$

FL-5. *For $p \in \{2, 4\}$, $c \in [M]$, there exists $\sigma_p^p \geq 0$ such that*

$$\mathbb{E}_{Z_c \sim \xi_c(\theta)}[\|\mathrm{g}_c^{Z_c}(\theta) - \mathrm{g}_c(\theta)\|^p] \leq \sigma_p^p \ , \quad \text{for all } c \in [M] \ , \ \theta \in \mathbb{R}^d \ . \tag{20}$$

FL-6. *For any $c \in [M]$, there exists $\beta \geq 0$ such that*

$$\|\mathrm{g}_c(\theta) - \nabla f_c(\theta)\| \leq \beta \ , \quad \text{for all } c \in [M] \ , \ \theta \in \mathbb{R}^d \ . \tag{21}$$

Proof of ascent lemma. To establish an ascent lemma for Algorithm 2, we first provide two lemmas: in Lemma C.1, we give a bound on the expected drift, and in Lemma C.2, we provide a bound on the variance of the global averaged parameters. We then use these two lemmas to prove Lemma C.3, which is our main result.

In the following, we define the filtration adapted to the global and local iterates of Algorithm 2 as

$$\mathcal{F}^r \triangleq \sigma\Big(Z_c^{r',h'} : r' < r, h' \in \{0, \dots, H\}, c' \in \{1, \dots, M\} \Big) \ .$$

We now prove our first lemma on the expected drift of Algorithm 2.

Lemma C.1 (Bound on Expected Drift)**.** *Assume **FL**-1 to **FL**-6. Let $\eta > 0$ such that $\eta H L_2 \leq 1/6$ and $32\eta^2 H^2 L_3^2 L_1^2 \leq L_2^2$, where L_2 and L_1 are defined in **FL**-1 and **FL**-2. respectively. Then the iterates of `proj-FedAVG` satisfy*

$$\begin{aligned}
&\frac{1}{MH} \sum_{c=1}^{M} \sum_{h=1}^{H-1} \|\mathbb{E}\left[\nabla f_c(\theta^r) - \nabla f_c(\theta_c^{r,h}) \middle| \mathcal{F}^r\right]\|_2^2 \\
&\quad \leq \frac{8\eta^2 L_2^2 H(H-1)}{M} \sum_{c=1}^{M} \|\nabla f_c(\theta^r)\|_2^2 + 8\eta^2 L_2^2 H(H-1)\beta^2 + 4 \cdot 12^3 \eta^4 L_3^2 H(H-1)\sigma_4^4 \ .
\end{aligned}$$

Proof. (Definition of Drift Error Terms.) To prove this lemma, we will bound each term of the sum

$$\mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} \triangleq \|\mathbb{E}\left[\nabla f_c(\theta^r) - \nabla f_c(\theta_c^{r,h}) \middle| \mathcal{F}^r\right]\|_2^2 \ .$$

(Bound on Drift Error Terms.) First, we use Taylor expansion to expand

$$\mathbb{E}\left[\nabla f_c(\theta_c^{r,h}) \middle| \mathcal{F}^r\right] - \nabla f_c(\theta^r) = \nabla^2 f_c(\theta^r) \mathbb{E}\left[\theta_c^{r,h} - \theta^r \middle| \mathcal{F}^r\right] + \mathbb{E}\left[\mathrm{D}_{3,c}^r(\theta_c^{r,h})(\theta_c^{r,h} - \theta^r)^{\otimes 2} \middle| \mathcal{F}^r\right] \ ,$$

where we defined the integral remainder as

$$\mathrm{D}_{3,c}^r(\theta_c^{r,h}) = \int_0^1 (1-t) \nabla^3 f_c(\theta^r + t(\theta_c^{r,h} - \theta^r)) \mathrm{d}t \ . \tag{22}$$

We thus obtain the following bound, using Jensen's inequality and the bound on the third derivatives tensor of f_c,

$$\begin{aligned}
\mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} &\leq 2\|\nabla^2 f_c(\theta^r) \mathbb{E}\left[\theta_c^{r,h} - \theta^r \middle| \mathcal{F}^r\right]\|_2^2 + 2\|\mathbb{E}\left[\mathrm{D}_{3,c}^r(\theta_c^{r,h})(\theta_c^{r,h} - \theta^r)^{\otimes 2} \middle| \mathcal{F}^r\right]\|_2^2 \\
&\leq 2L_2^2 \|\mathbb{E}\left[\theta_c^{r,h} - \theta^r \middle| \mathcal{F}^r\right]\|_2^2 + 2L_3^2 \mathbb{E}\left[\|\theta_c^{r,h} - \theta^r\|_2^4 \middle| \mathcal{F}^r\right] \ ,
\end{aligned} \tag{23}$$

We now use the fact that $\theta_c^{r,h} = \theta^r - \eta \sum_{\ell=0}^{h-1} \mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell})$ and (21) to write

$$\begin{aligned}
&2L_2^2 \|\mathbb{E}\left[\theta_c^{r,h} - \theta^r \middle| \mathcal{F}^r\right]\|_2^2 \\
&\quad = 2\eta^2 L_2^2 \Big\| \mathbb{E}\Big[\sum_{\ell=0}^{h-1} \nabla f_c(\theta^r) + \nabla f_c(\theta_c^{r,\ell}) - \nabla f_c(\theta^r) + \mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell}) - \nabla f_c(\theta_c^{r,\ell}) \Big| \mathcal{F}^r \Big] \Big\|_2^2
\end{aligned}$$

$$\leq 6\eta^2 L_2^2 h^2 \|\nabla f_c(\theta^r)\|_2^2 + 6\eta^2 L_2^2 h \sum_{\ell=0}^{h-1} \left\| \mathbb{E}\left[\nabla f_c(\theta^r) - \nabla f_c(\theta_c^{r,\ell}) \middle| \mathcal{F}^r\right] \right\|_2^2 + 6\eta^2 L_2^2 h^2 \beta^2 \ ,$$

Completing the sum until $\ell = H-1$ and plugging this inequality in (23), we obtain

$$\begin{aligned}\mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq 6\eta^2 L_2^2 h^2 \|\nabla f_c(\theta^r)\|_2^2 + 6\eta^2 L_2^2 h \sum_{\ell=0}^{H-1} \left\| \mathbb{E}\left[\nabla f_c(\theta^r) - \nabla f_c(\theta_c^{r,\ell}) \middle| \mathcal{F}^r\right] \right\|_2^2 \\ + 6\eta^2 L_2^2 h^2 \beta^2 + 2L_3^2 \mathbb{E}\left[\|\theta_c^{r,h} - \theta^r\|_2^4 \middle| \mathcal{F}^r\right] \ .\end{aligned}$$

Now, we average the above inequality for $h = 0$ to $H-1$ and $c = 1$ to M, which gives

$$\begin{aligned}\frac{1}{MH}\sum_{c=1}^{M}\sum_{h=1}^{H-1} \mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq \frac{3\eta^2 L_2^2 H(H-1)}{M} \sum_{c=1}^{M} \|\nabla f_c(\theta^r)\|_2^2 + \frac{3\eta^2 L_2^2 H(H-1)}{MH} \sum_{c=1}^{M}\sum_{h=0}^{H-1} \mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} \\ + 3\eta^2 L_2^2 H(H-1)\beta^2 + \frac{2L_3^2}{MH} \sum_{c=1}^{M}\sum_{h=0}^{H-1} \mathbb{E}\left[\|\theta_c^{r,h} - \theta^r\|_2^4 \middle| \mathcal{F}^r\right] \ ,\end{aligned}$$

where we used $\sum_{h=0}^{H-1} h^2 \leq H \sum_{h=0}^{H-1} h = \frac{H^2(H-1)}{2}$. Using that $3\eta^2 L_2^2 H(H-1) \leq 1/2$, reorganizing the terms, and multiplying the resulting inequality by 2, we obtain

$$\begin{aligned}\frac{1}{MH}\sum_{c=1}^{M}\sum_{h=1}^{H-1} \mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq \frac{6\eta^2 L_2^2 H(H-1)}{M} \sum_{c=1}^{M} \|\nabla f_c(\theta^r)\|_2^2 \\ + \frac{4L_3^2}{MH} \sum_{c=1}^{M}\sum_{h=0}^{H-1} \mathbb{E}\left[\|\theta_c^{r,h} - \theta^r\|_2^4 \middle| \mathcal{F}^r\right] + 6\eta^2 L_2^2 H(H-1)\beta^2 \ . \end{aligned} \tag{24}$$

(Fourth Order Drift Terms.) We now bound the fourth moment of the drift. To this end, we define

$$\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \triangleq \mathbb{E}\left[\|\theta_c^{r,h} - \theta^r\|^4 \middle| \mathcal{F}^r\right] \ ,$$

and we write $\theta_c^{r,h} = \theta^r + \eta \sum_{\ell=0}^{h-1} \mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell})$, and we decompose each update as

$$\mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell}) = \mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell}) - \mathrm{g}_c(\theta_c^{r,\ell}) + \mathrm{g}_c(\theta_c^{r,\ell}) - \nabla f_c(\theta_c^{r,\ell}) + \nabla f_c(\theta_c^{r,\ell}) - \nabla f_c(\theta^r) + \nabla f_c(\theta^r) \ .$$

This gives the bound

$$\begin{aligned}\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq 4^3\eta^4 \underbrace{\mathbb{E}\left[\left\| \sum_{\ell=0}^{h-1} \mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell}) - \mathrm{g}_c(\theta_c^{r,\ell}) \right\|^4 \middle| \mathcal{F}^r\right]}_{T_1} + 4^3\eta^4 \underbrace{\mathbb{E}\left[\left\| \sum_{\ell=0}^{h-1} \mathrm{g}_c(\theta_c^{r,\ell}) - \nabla f_c(\theta_c^{r,\ell}) \right\|^4 \middle| \mathcal{F}^r\right]}_{T_2} \\ + 4^3\eta^4 \underbrace{\mathbb{E}\left[\left\| \sum_{\ell=0}^{h-1} \nabla f_c(\theta_c^{r,\ell}) - \nabla f_c(\theta^r) \right\|^4 \middle| \mathcal{F}^r\right]}_{T_3} + 4^3\eta^4 \underbrace{h^4 \|\nabla f_c(\theta^r)\|^4}_{T_4} \ .\end{aligned} \tag{25}$$

We bound T_1 using Burkholder's inequality (Theorem 8.6, Osękowski, 2012), which gives

$$T_1 \leq 3^4 \Big\{ \sum_{\ell=0}^{h-1} \mathbb{E}^{1/2}\left[\|\mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell}) - \mathrm{g}_c(\theta_c^{r,\ell})\|^4 \middle| \mathcal{F}^r\right] \Big\}^2 \leq 3^4 h^2 \sigma_4^4 \ . \tag{26}$$

The term T_2 is a bias term, which we bound using (21),

$$T_2 \leq h^3 \sum_{\ell=0}^{h-1} \mathbb{E}\left[\|\mathrm{g}_c(\theta_c^{r,\ell}) - \nabla f_c(\theta_c^{r,\ell})\|^4 \middle| \mathcal{F}^r\right] \leq h^4 \beta^4 \ . \tag{27}$$

Then, we bound T_3 using (15)

$$T_3 \leq h^3 \sum_{\ell=0}^{h-1} \mathbb{E}\left[\|\nabla f_c(\theta_c^{r,\ell}) - \nabla f_c(\theta^r)\|^4 \middle| \mathcal{F}^r\right] \leq L_2^4 h^3 \sum_{\ell=0}^{h-1} \mathbb{E}\left[\|\theta_c^{r,\ell} - \theta^r\|^4 \middle| \mathcal{F}^r\right] \ . \tag{28}$$

Finally, we bound T_4 using gradient's boundedness (17),

$$T_4 \leq L_1^2 h^4 \|\nabla f_c(\theta^r)\|^2 \ . \tag{29}$$

Plugging (26), (27), (28), (29) in (25), we obtain

$$\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq 4^3\eta^4 h^4 L_1^2 \|\nabla f_c(\theta^r)\|^2 + 4^3\eta^4 h^4 \beta^4 + 4^3\eta^4 L_2^4 h^3 \sum_{\ell=0}^{h-1} \mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{\ell}} + 3\cdot 12^3 \eta^4 h^2 \sigma_4^4 \ .$$

Like for the terms $\mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}}$, we complete the sum and average over $h = 0$ to $H-1$, which gives

$$\begin{aligned}\frac{1}{H}\sum_{h=0}^{H-1} \mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq \frac{4^3\eta^4 L_2^4 H^3 (H-1)}{5H} \sum_{h=0}^{H-1} \mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} + \frac{3\cdot 12^3 \eta^4 H(H-1)}{3}\sigma_4^4 \\ + \frac{4^3\eta^4 H^2 (H-1)^2}{5}\Big(L_1^2\|\nabla f_c(\theta^r)\|^2 + \beta^4\Big) \ .\end{aligned}$$

Using $\eta H L_2 \leq 1/6$, averaging over $c = 1$ to M, collecting the terms in $\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}}$ on the left hand side, and multiplying by 2, we obtain

$$\frac{1}{MH}\sum_{c=1}^{M}\sum_{h=0}^{H-1} \mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq \frac{2\cdot 4^3\eta^4 H^2(H-1)^2}{5}\Big\{\beta^4 + L_1^2\|\nabla f_c(\theta^r)\|^2\Big\} + \frac{6\cdot 12^3\eta^4 H(H-1)}{3}\sigma_4^4 \ . \tag{30}$$

(Final Result.) Plugging (30) back in (24) and using $L_3^2\eta^2 H^2 L_1^2 \leq L_2^2/32$ and $\beta \leq L_1$ gives

$$\begin{aligned}\frac{1}{MH}\sum_{c=1}^{M}\sum_{h=1}^{H-1} \mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} &\leq \Big(6\eta^2 L_2^2 H(H-1) + \frac{4^4\eta^4 L_3^2 L_1^2 H^2 (H-1)^2}{5}\Big)\frac{1}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|_2^2 \\ &\quad + \frac{12^4\eta^4 L_3^2 H(H-1)}{3}\sigma_4^4 + 6\eta^2 L_2^2 H(H-1)\beta^2 + \frac{4^4\eta^4 L_3^2 H^2(H-1)^2}{5}\beta^4 \\ &\leq \Big(6\eta^2 L_2^2 H(H-1) + 2\eta^2 L_2^2 (H-1)^2\Big)\frac{1}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|_2^2 \\ &\quad + 4\cdot 12^3 L_3^2\eta^4 H(H-1)\sigma_4^4 + \Big(6\eta^2 L_2^2 H(H-1) + 2\eta^2 L_2^2 (H-1)^2\Big)\beta^2 \ ,\end{aligned}$$

and the result follows. □

Lemma C.2 (Bound on global iterates variance)**.** *Assume **FL**-1 to **FL**-6. Assume that $\eta H L_2 \leq 1/6$ Then the iterates of `proj-FedAVG` satisfy*

$$\mathbb{E}\Big[\|\bar\theta^{r+1} - \mathbb{E}\left[\bar\theta^{r+1}\middle|\mathcal{F}^r\right]\|^2\Big] \leq \frac{3\eta^2 H\sigma_2^2}{M} \ .$$

Proof. Since $\bar\theta^{r+1} = 1/M\sum_{c=1}^M \theta_c^{r,H}$ and $\{\theta_c^{r,H}\}_{c=1}^M$ are independent conditional to $\mathcal{F}^r$,

$$\mathbb{E}\Big[\Big\|\bar\theta^{r+1} - \mathbb{E}\left[\bar\theta^{r+1}\middle|\mathcal{F}^r\right]\Big\|^2\Big] = \frac{1}{M^2}\sum_{c=1}^{M}\mathbb{E}\Big[\|\theta_c^{r,H} - \mathbb{E}\left[\theta_c^{r,H}\middle|\mathcal{F}^r\right]\|^2\Big] \ .$$

Then, we have, for $h \in \{0,\dots,H-1\}$, using that $\mathbb{E}\left[\mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})\middle|\mathcal{F}^r\right] = \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]$,

$$\begin{aligned}\mathbf{A}_{\boldsymbol{c}}^{\boldsymbol{r,h+1}} &:= \mathbb{E}\Big[\|\theta_c^{r,h+1} - \mathbb{E}\left[\theta_c^{r,h+1}\middle|\mathcal{F}^r\right]\|^2\Big] \\ &= \mathbb{E}\Big[\Big\|\theta_c^{r,h} - \mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right] + \eta\big(\mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h}) - \mathrm{g}_c(\theta_c^{r,h}) + \mathrm{g}_c(\theta_c^{r,h}) - \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\big)\Big\|^2\Big] \ .\end{aligned}$$

Since $\{Z_c^{r,h}\}_{h=1}^H$ are independent conditional to $\mathcal{F}^r$, we have, using (20),

$$\mathbf{A}_{\boldsymbol{c}}^{\boldsymbol{r,h+1}} = \mathbb{E}\Big[\Big\|\theta_c^{r,h} - \mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right] + \eta\big(\mathrm{g}_c(\theta_c^{r,h}) - \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\big)\Big\|^2\Big] + \eta^2\sigma_2^2. \tag{31}$$

Then, by Young's inequality, we have

$$\mathbb{E}\Big[\Big\|\theta_c^{r,h} - \mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right] + \eta\Big(\mathrm{g}_c(\theta_c^{r,h}) - \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\Big)\Big\|^2\Big]$$
$$\leq (1+\eta L_2)\mathbb{E}\Big[\Big\|\theta_c^{r,h} - \mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right]\Big\|^2\Big] + (\eta^2 + \eta/L_2)\mathbb{E}\Big[\Big\|\mathrm{g}_c(\theta_c^{r,h}) - \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\Big\|^2\Big]$$

Finally, we have, by Young's inequality and (15),

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\mathrm{g}_c(\theta_c^{r,h}) - \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\Big\|^2\Big] \\
&\leq 2\mathbb{E}\Big[\Big\|\mathrm{g}_c(\theta_c^{r,h}) - \mathrm{g}_c(\mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right])\Big\|^2\Big] + 2\mathbb{E}\Big[\Big\|\mathrm{g}_c(\mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right]) - \mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\Big\|^2\Big] \\
&\leq 2L_2^2\mathbb{E}\Big[\Big\|\theta_c^{r,h} - \mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right]\Big\|^2\Big] + 2L_2^2\mathbb{E}\Big[\Big\|\mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right] - \theta_c^{r,h}\Big\|^2\Big] \\
&\leq 4L_2^2\mathbb{E}\Big[\Big\|\theta_c^{r,h} - \mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right]\Big\|^2\Big] \ ,
\end{aligned}$$

where we used Jensen's inequality in the last inequality. Then, notice that $4(\eta^2 + \eta/L_2)L_2^2 = 4(\eta^2 L_2^2 + \eta L_2) \leq 5\eta L_2$ since $\eta L_2 \leq 1/4$. Plugging this in (31), we obtain

$$\mathbf{A_c^{r,h+1}} \leq (1+6\eta L_2)\mathbf{A_c^{r,h}} + \eta^2\sigma_2^2 \ .$$

And unrolling this inequality gives

$$\mathbb{E}\Big[\|\theta_c^{r,H} - \mathbb{E}\left[\theta_c^{r,H}\middle|\mathcal{F}^r\right]\|^2\Big] \leq \eta^2\sum_{h=0}^{H}(1+6\eta L_2)^h\sigma_2^2 \leq 3\eta^2 H\sigma_2^2 \ ,$$

where the second inequality comes from $\eta H L_2 \leq 1/6$, which gives $(1+6\eta L_2)^h \leq (1+1/H)^H \leq 3$, and the lemma follows. □

Lemma C.3 (Ascent Lemma)**.** *Assume **FL-1 to FL-6**. For any $\eta > 0$ such that $\eta H L_2 \leq 1/6$ and $32\eta^2 H^2 L_3^2 L_1^2 \leq L_2^2$, the iterates of `proj-FedAVG` satisfy*

$$\begin{aligned}
-\mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right] \leq -F(\theta^r) - \frac{\eta H}{4}\|\nabla F(\theta^r)\|_2^2 + \frac{3\eta^2 L_2 H\sigma_2^2}{2M} \\
+ 2\eta H\beta^2 + 8\eta^3 L_2^2 H^2(H-1)\zeta^2 + 4\cdot 12^3\eta^5 L_3^2 H^2(H-1)\sigma_4^4 \ .
\end{aligned}$$

Proof. Smoothness of f_c gives $|F(\bar{\theta}^{r+1}) - F(\theta^r) - \langle\nabla F(\theta^r), \bar{\theta}^{r+1} - \theta^r\rangle| \leq (L_2/2)\|\bar{\theta}^{r+1} - \theta^r\|^2$, which implies that

$$-F(\bar{\theta}^{r+1}) \leq -F(\theta^r) - \langle\nabla F(\theta^r), \bar{\theta}^{r+1} - \theta^r\rangle + \frac{L_2}{2}\|\bar{\theta}^{r+1} - \theta^r\|_2^2 \ .$$

Let $\kappa = \frac{1}{\sqrt{\eta H}}$. Taking the expectation conditionally on $\mathcal{F}^r$ and using the polarization identity $2\langle a, b\rangle = \|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2$ for $a, b \in \mathbb{R}^d$, we get

$$\begin{aligned}
-\mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right] + F(\theta^r) &\leq -\langle\kappa^{-1}\nabla F(\theta^r), \kappa\mathbb{E}\left[\bar{\theta}^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\rangle + \frac{L_2}{2}\mathbb{E}\left[\|\bar{\theta}^{r+1} - \theta^r\|_2^2\middle|\mathcal{F}^r\right] \\
&= -\frac{1}{2\kappa^2}\|\nabla F(\theta^r)\|_2^2 + \underbrace{\frac{1}{2\kappa^2}\|\nabla F(\theta^r) + \kappa^2\mathbb{E}\left[\theta^r - \bar{\theta}^{r+1}\middle|\mathcal{F}^r\right]\|_2^2}_{(\mathbf{A})} \\
&\quad + \underbrace{\frac{L_2}{2}\mathbb{E}\left[\|\bar{\theta}^{r+1} - \theta^r\|_2^2\middle|\mathcal{F}^r\right] - \frac{\kappa^2}{2}\|\mathbb{E}\left[\bar{\theta}^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\|_2^2}_{(\mathbf{B})} \ . \qquad (32)
\end{aligned}$$

The term $(\mathbf{A})$ is a drift term, that is due to local updates, and is due to heterogeneity, while the term $(\mathbf{B})$ is a second order term error term and a variance term. We now bound each of these two terms.

Bounding $(\mathbf{A})$. Using the fact that $F = \frac{1}{M}\sum_{c=1}^M f_c$, the definition $\kappa^2 = 1/\eta H$, the definition of $\bar{\theta}^{r+1}$ and Jensen's inequality, we have

$$\Big\|\nabla F(\theta^r) + \kappa^2\mathbb{E}\left[\theta^r - \bar{\theta}^{r+1}\middle|\mathcal{F}^r\right]\Big\|_2^2 = \Big\|\mathbb{E}\Big[\frac{1}{M}\sum_{c=1}^M\Big(\nabla f_c(\theta^r) - \frac{1}{H}\sum_{h=0}^{H-1}\mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})\Big)\Big|\mathcal{F}^r\Big]\Big\|_2^2$$

$$\leq \frac{1}{HM}\sum_{c=1}^{M}\sum_{h=0}^{H-1}\left\|\mathbb{E}\left[\nabla f_c(\theta^r) - \mathrm{g}_c^{Z_c^{r,h}}(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\right\|_2^2$$
$$= \frac{1}{HM}\sum_{c=1}^{M}\sum_{h=0}^{H-1}\left\|\mathbb{E}\left[\nabla f_c(\theta^r) - \mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\right\|_2^2 ,$$

where the last equality holds by independence of $Z_c^{r,h+1}$ and $\mathcal{F}_{r,h}^c$. By decomposing

$$\nabla f_c(\theta^r) - \mathrm{g}_c(\theta_c^{r,h}) = \nabla f_c(\theta^r) - \nabla f_c(\theta_c^{r,h}) + \nabla f_c(\theta_c^{r,h}) - \mathrm{g}_c(\theta_c^{r,h}) \ .$$

Using Young's inequality and bounding the bias using (21), we obtain

$$\left\|\nabla F(\theta^r) + \kappa^2\mathbb{E}\left[\theta^r - \bar{\theta}^{r+1}\middle|\mathcal{F}^r\right]\right\|_2^2 \leq \frac{2}{HM}\sum_{c=1}^{M}\sum_{h=0}^{H-1}\|\mathbb{E}\left[\nabla f_c(\theta^r) - \nabla f_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\|_2^2 + 2\beta^2 \ .$$

Using Lemma C.1 to bound the first term, and multiplying by $1/(2\kappa^2) = \eta H/2$, we obtain

$$\begin{aligned}(\mathbf{A}) \leq \frac{8\eta^3 L_2^2 H^2(H-1)}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|_2^2 \\ + 4\cdot 12^3\cdot 2L_3^2\eta^5 H^2(H-1)\sigma_4^4 + (1+8\eta^2L_2^2H(H-1))\eta H\beta^2 \ .\end{aligned} \tag{33}$$

Bounding $(\mathbf{B})$**.** We decompose $(\mathbf{B})$ by writing $\bar{\theta}^{r+1} = \mathbb{E}\left[\bar{\theta}^{r+1}\middle|\mathcal{F}^r\right] + \bar{\theta}^{r+1} - \mathbb{E}\left[\bar{\theta}^{r+1}\middle|\mathcal{F}^r\right]$, which gives

$$\begin{aligned}(\mathbf{B}) &= \frac{L_2}{2}\mathbb{E}\left[\|\mathbb{E}\left[\bar{\theta}^{r+1}\middle|\mathcal{F}^r\right] - \bar{\theta}^{r+1}\|^2\middle|\mathcal{F}^r\right] + \frac{L_2}{2}\|\mathbb{E}\left[\bar{\theta}^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\|_2^2 - \frac{\kappa^2}{2}\|\mathbb{E}\left[\bar{\theta}^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\|_2^2 \\ &= \frac{L_2}{2}\mathbb{E}\left[\|\mathbb{E}\left[\bar{\theta}^{r+1}\middle|\mathcal{F}^r\right] - \bar{\theta}^{r+1}\|^2\middle|\mathcal{F}^r\right] + \left(\frac{L_2}{2} - \frac{\kappa^2}{2}\right)\|\mathbb{E}\left[\bar{\theta}^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\|_2^2 \ .\end{aligned}$$

Since $\eta H L_2 \leq 1$, we have $\frac{L_2}{2} - \frac{\kappa^2}{2} \leq \frac{L_2}{2} - \frac{1}{2\eta H} \leq 0$, and the second term is negative. The second term is a variance term, that we bound using Lemma C.2, which gives

$$(\mathbf{B}) \leq \frac{3\eta^2 L_2 H\sigma_2^2}{2M} \ . \tag{34}$$

Bound on (32)**.** Plugging in the bounds (33) and (34) on $(\mathbf{A})$ and $(\mathbf{B})$ in (32) yields

$$\begin{aligned}-\mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right] + F(\theta^r) \leq -\frac{\eta H}{2}\|\nabla F(\theta^r)\|_2^2 + \frac{8\eta^3L_2^2H^2(H-1)}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|_2^2 \\ + 4\cdot 12^3\eta^5L_3^2H^2(H-1)\sigma_4^4 + 2\eta H\beta^2 + \frac{3\eta^2L_2H\sigma_2^2}{2M} \ ,\end{aligned} \tag{35}$$

where we used $\eta H L_2 \leq 1/6$ to bound $(1+8\eta^2L_2^2H(H-1))\eta H\beta^2 \leq 2\eta H\beta^2$. Moreover, we have $8\eta^2L_2^2H^2 \leq 1/4$ and $\frac{1}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|^2 \leq \|\nabla F(\theta^r)\|^2 + \zeta^2$, which gives the bound

$$\frac{8\eta^3L_2^2H^2(H-1)}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|_2^2 \leq \frac{\eta H}{4}\|\nabla F(\theta^r)\|_2^2 + 8\eta^3L_2^2H^2(H-1)\zeta^2 \ ,$$

and the result of the lemma follows from plugging this inequality in (35). □

D Analysis of `S-FedPG`

`S-FedPG` can be interpreted as a specific instance of `proj-FedAVG`, where the projection set is chosen as $\mathcal{T} = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, the local objective is defined as $f_c = J_{\mathrm{sm},c}$, and the agent data distribution $\xi_c(\theta)$ corresponds to $\nu_c(\theta)$—the distribution induced by sampling B truncated trajectories from the policy π_θ, as defined in (3).

Given a parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and an observation $Z_c \sim \nu_c(\theta)$, we recall the form of the biased estimator (defined in (6)) for the stochastic gradient:

$$\mathrm{g}_{\mathrm{sm},c}^{Z_c}(\theta) := \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t\left(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(A_{c,b}^{\ell}\mid S_{c,b}^{\ell})\right)\mathsf{r}(S_{c,b}^{t},A_{c,b}^{t}) \ . \tag{36}$$

Define also

$$\mathrm{g}_{\mathrm{sm},c}(\theta) = \mathbb{E}_{Z_c\sim\nu_c(\theta)}[\mathrm{g}_{\mathrm{sm},c}^{Z_c}(\theta)]. \tag{37}$$

To apply the ascent lemma (Lemma C.3), it remains to verify that Assumptions **FL**-1 through **FL**-6 are satisfied. We establish these conditions in the following.

D.1 Checking the assumptions and establishing a local Łojasiewicz structure

For a given policy π and agent $c \in [M]$, the value function $V_c^\pi : \mathcal{S} \to \mathbb{R}$, is defined as:

$$V_c^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t\mathsf{r}(S_c^t,A_c^t)\middle| S_c^0 = s\right], \tag{38}$$

where for all $t \geq 0$, $A_c^t \sim \pi(\cdot|S_c^t)$ is chosen using the shared policy, and $S_c^{t+1} \sim \mathsf{P}_c(.|S_c^t,A_c^t)$ follows the local dynamics of agent c's environment. We define $V_c^\pi(\rho)$ as the value function when the initial distribution is ρ. Similarly, the Q-function of a policy π for agent c is

$$Q_c^\pi(s,a) := \mathsf{r}(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_c(s'|s,a)V_c^\pi(s') \ . \tag{39}$$

This allows to define the advantage function $A_c^\pi(s,a) = Q_c^\pi(s,a) - V_c^\pi(s)$. Define $J_{\mathrm{sm},c}(\theta) := J_c(\pi_\theta)$. Given a policy π and initial distribution ρ, we define the occupancy measure as

$$d_c^{\rho,\pi}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\rho\mathsf{P}_{c,\pi}^t(s) \ , \quad \mathsf{P}_{c,\pi}(s'|s) = \sum\pi(a|s)\mathsf{P}_c(s'|s,a) \ . \tag{40}$$

We define the advantage function of a policy π_θ as

$$A_c^{\pi_\theta}(s,a) := Q_c^{\pi_\theta}(s,a) - V_c^{\pi_\theta}(s) \ , \quad \text{for all } (s,a)\in\mathcal{S}\times\mathcal{A} \ . \tag{41}$$

Following Mei et al. (2020), we will use the following expression of the gradient.

Lemma D.1 (Lemma 10 from Mei et al. (2020))**.** *We have*

$$\frac{\partial J_{\mathrm{sm},c}(\theta)}{\partial\theta(s,a)} = \frac{1}{1-\gamma}\cdot d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s)A_c^{\pi_\theta}(s,a) \ , \tag{42}$$

where $A_c^{\pi_\theta}$ is defined in (41)*.*

First, we establish the smoothness of $\mathrm{g}_{\mathrm{sm},c}(\theta)$.

Lemma D.2. *For any $c \in [M]$, the function $\mathrm{g}_{\mathrm{sm},c}$ is $L_{2,\mathrm{sm}} \triangleq 8/(1-\gamma)^3$-smooth, that is for all $\theta,\theta' \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\|\mathrm{g}_{\mathrm{sm},c}(\theta') - \mathrm{g}_{\mathrm{sm},c}(\theta)\| \leq L_{2,\mathrm{sm}}\|\theta'-\theta\|_2 \ .$$

Proof. The result follows from setting $\lambda = 0$ in the bound of Lemma E.2.

□

Lemma D.3. *For $c \in [M]$, the function $J_{\mathrm{sm},c}$ is $L_{2,\mathrm{sm}} = 8/(1-\gamma)^3$-smooth and J_{sm} is also $L_{2,\mathrm{sm}}$-smooth.*

Proof. The result follows from Lemma 7 of Mei et al. (2020) and the fact that a mean of smooth functions is a smooth functions with the same smoothness coefficient. □

Lemma D.4. *For all $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds*

$$\|\nabla J_{\mathrm{sm},c}(\theta)\|_2 \le L_{1,\mathrm{sm}} \ , \quad \textit{where } L_{1,\mathrm{sm}} \triangleq \frac{1}{(1-\gamma)^2} \ .$$

Proof. By norm comparisons, and Lemma D.1, it holds that

$$\|J_{\mathrm{sm},c}(\theta)\|_2 \le \|J_{\mathrm{sm},c}(\theta)\|_1 \le \frac{1}{1-\gamma}\sum_{s,a} d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s)|A_c^{\pi_\theta}(s,a)| \ .$$

Finally, using that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $|A_c^{\pi_\theta}(s,a)| \le 1/(1-\gamma)$ concludes the proof. □

Lemma D.5. *The spectral norm of the third derivative tensor is bounded by $L_{3,\mathrm{sm}} := 480\cdot(1-\gamma)^{-4}$, i.e., for any $u, v, w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ it holds*

$$|\mathrm{d}^3 J_{\mathrm{sm},c}(\theta)[u,v,w]| = |\nabla^3 J_{\mathrm{sm},c}(\theta)u\otimes v\otimes w| \le \frac{480}{(1-\gamma)^4}\|u\|_2\|v\|_2\|w\|_2 \, .$$

Proof. By Lemma G.9 with $\lambda = 0$ we have for any $u, v, w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

$$\|\mathrm{d}^3 V_c^{\pi_\theta}[u,v,w]\|_\infty \le \frac{480}{(1-\gamma)^4}\|u\|_2\|v\|_2\|w\|_2 \ .$$

Next, we notice that

$$\mathrm{d}^3 J_{\mathrm{sm},c}(\theta)[u,v,w] = \rho^\top \mathrm{d}^3 V_c^{\pi_\theta}[u,v,w] \ ,$$

and the result follows from the fact that ρ is a probability distribution. □

Lemma D.6. *Assume **A**-1. Let $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. It holds that*

$$\|\nabla J_{\mathrm{sm}}(\theta) - \nabla J_{\mathrm{sm},c}(\theta)\|_2^2 \le \zeta_{\mathrm{sm}}^2 \ , \quad \textit{where } \zeta_{\mathrm{sm}}^2 := \frac{38\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} \ .$$

Proof. The result follows from setting $\lambda = 0$ in the bound of Lemma E.6.

□

The following lemma bounds the bias and the variance of the estimator of this stochastic gradient.

Lemma D.7 (Lemmas 6 and 7 from Ding et al. (2025))**.** *Consider the stochastic gradient defined in* (36)*. For any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\|\nabla J_{\mathrm{sm},c}(\theta) - \mathrm{g}_{\mathrm{sm},c}(\theta)\|_2 \le \beta_{\mathrm{sm}} := \frac{2\gamma^T}{1-\gamma}\left(T + \frac{1}{1-\gamma}\right) \ ,$$
$$\mathrm{Var}(\mathrm{g}_{\mathrm{sm},c}^{Z_c}(\theta)) \le \sigma_{\mathrm{sm},2}^2 := \frac{12}{B(1-\gamma)^4} \ .$$

Finally, we show that the fourth-order moment of our biased estimator is bounded.

Lemma D.8. *For any $c \in [M]$, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, the fourth central moment of $\mathrm{g}_{\mathrm{sm},c}^{Z_c}$ is bounded, that is*

$$\mathbb{E}_{Z_c\sim\nu_c(\theta)}\left[\|\mathrm{g}_{\mathrm{sm},c}^{Z_c}(\theta) - \mathrm{g}_{\mathrm{sm},c}(\theta)\|_2^4\right] \le \sigma_{\mathrm{sm},4}^4 := \frac{1120}{B^2(1-\gamma)^8} \ . \tag{43}$$

Proof. The result follows from setting $\lambda = 0$ in the bound of Lemma E.8. □

Lemma D.9 (Lemma 8 of Mei et al. (2020))**.** *For all $c \in [M]$, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds*

$$\|\nabla J_{\mathrm{sm},c}(\theta)\|_2^2 \ge 2\mu_{\mathrm{sm},c}(\theta)\cdot\left[J_{\mathrm{sm},c}^\star - J_{\mathrm{sm},c}(\theta)\right]^2 \ ,$$

where

$$\mu_{\mathrm{sm},c}(\theta) \triangleq \frac{1}{2|\mathcal{S}|}\cdot\min_s \pi_\theta(a^\star(s)|s)^2 \cdot \left\|\frac{d_c^{\rho,\pi_c^\star}}{d_c^{\rho,\theta}}\right\|_\infty^{-2} \ ,$$

and where $\pi_c^\star$ is an optimal deterministic policy of agent c, and $a^\star(s)$ is the action picked by this policy when the agent is in state s.

In case of small heterogeneity between the agents, i.e **A**-1, this precedent Lemma, combined with Lemma D.6 allows us to establish that the global objective satisfies a Relaxed Łojasiewicz-type inequality.

Lemma D.10. *Assume **A**-1. For any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\zeta_{\rm sm}^2 + \|\nabla J_{\rm sm}(\theta)\|_2^2 \geq \mu_{\rm sm}(\theta)(J_{\rm sm}^\star - J_{\rm sm}(\theta))^2 \ , \quad \textit{with } \mu_{\rm sm}(\theta) = \min_{c\in[M]} \mu_{{\rm sm},c}(\theta) \ .$$

Proof. Let $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Using Lemma D.9 and the triangle inequality, we have for any $c \in [M]$

$$\sqrt{\min_{c\in[M]} 2\mu_{{\rm sm},c}(\theta)}\left[J_{{\rm sm},c}^\star - J_{{\rm sm},c}(\theta)\right] \leq \sqrt{2\mu_{{\rm sm},c}(\theta)}\left[J_{{\rm sm},c}^\star - J_{{\rm sm},c}(\theta)\right] \leq \|\nabla J_{{\rm sm},c}(\theta)\|_2 \ .$$

We then decompose $\nabla J_{{\rm sm},c}(\theta) = \nabla J_{{\rm sm},c}(\theta) - \nabla J_{\rm sm}(\theta) + \nabla J_{\rm sm}(\theta)$ and use triangle inequality and Lemma D.6 to bound

$$\|\nabla J_{{\rm sm},c}(\theta)\|_2 \leq \|\nabla J_{{\rm sm},c}(\theta) - \nabla J_{\rm sm}(\theta)\|_2 + \|\nabla J_{\rm sm}(\theta)\|_2 \leq \zeta_{\rm sm} + \|\nabla J_{\rm sm}(\theta)\|_2 \ .$$

Averaging the resulting inequality over all the agents, taking the square, using that $J_{\rm sm}^\star \leq 1/M \sum_{c=1}^M J_{{\rm sm},c}^\star$, and applying Young's inequality concludes the proof. □

D.2 Convergence rates, sample, and communication complexities

We preface the proof by an elementary Lemma.

Lemma D.11. *Let $(w_r)_{r=0}^\infty$ be a sequence of positive real numbers, and let $\kappa > 0$, $B > 0$. Assume that for all $r \geq 0$,*

$$w_{r+1} \leq w_r - \kappa\, w_r^2 + B.$$

Then for every integer $r \geq 0$ one has

$$w_r \leq \sqrt{\frac{B}{\kappa}} + B + \frac{w_0}{1 + \kappa\, r\, w_0}.$$

Proof. Set $M = \sqrt{B/\kappa}$ and fix $r \in \mathbb{N}$. We split into two cases:

Case 1: $w_k > M$ for all $k \in \{0, \dots r\}$. Define $v_k := w_k - M$ which is positive as $w_k > M$. Then for any $k \in \{0, \dots r\}$, it holds that

$$v_{k+1} = w_{k+1} - M \leq w_k - M - \kappa(w_k - M + M)^2 + B \leq v_k - \kappa v_k^2 \ ,$$

where in the last inequality, we used that for any $a, b \geq 0$, we have $(a+b)^2 \geq a^2 + b^2$. Dividing the preceding inequality by v_k^2 yields

$$\frac{v_{k+1} - v_k}{v_k^2} \leq -\kappa \ . \tag{44}$$

For $x > 0$, define $g(x) = x^{-1}$. By convexity of g on $\mathbb{R}_+^\star$, we have $g(v_{k+1}) \geq g(v_k) + (v_{k+1} - v_k)g'(v_k)$ which can be rewritten as

$$v_{k+1}^{-1} \geq v_k^{-1} - (v_{k+1} - v_k)\,\frac{1}{v_k^2} \ ,$$

and which implies, after using (44)

$$v_k^{-1} - v_{k+1}^{-1} \leq \frac{v_{k+1} - v_k}{v_k^2} \leq -\kappa \ .$$

Summing up both sides over $k = 0 \dots r$ and rearranging the terms yields

$$(w_r - M)^{-1} \geq \kappa r + w_0^{-1} \ .$$

Finally, we get

$$w_r \leq M + \frac{w_0}{1 + \kappa r w_0} \ .$$

Case 2: There exists some $0 \leq r_0 \leq r$ with $w_{r_0} \leq M$. Let us prove that for any $0 \leq x \leq M + B$, it holds that $0 \leq x - \kappa x^2 + B \leq M + B$. We distinguish two sub-cases. First, if $x \leq M$ then it holds that $x - \kappa x^2 + B \leq x + B \leq M + B$. Alternatively, if $M \leq x \leq M + B$ then $x - \kappa x^2 + B \leq x - \kappa M^2 + B = x \leq M + B$. Finally, using the preceding inequality combined with an immediate recursion proves that for all $k \geq r_0$, we have $w_k \leq B + M$. □

Theorem D.12 (Convergence rates of S-FedPG)**.** *Assume **A**-1 and **A**-2 and set $\mathcal{T} = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Additionally, assume that there exists $\mu_{\mathrm{sm}} > 0$ such that such that $\inf_{r\in[\mathbb{N}]} \mu_{\mathrm{sm}}(\theta^r) \geq \mu_{\mathrm{sm}} > 0$. For any $\eta > 0$ such that $\eta H L_{2,\mathrm{sm}} \leq 1/74$ the iterates of S-FedPG satisfy*

$$J^\star - \mathbb{E}[J(\pi_{\theta^R})] \leq \frac{J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0)}{1 + R\cdot(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\cdot(\eta H \mu_{\mathrm{sm}}/4)} + \left(\frac{6\eta L_{2,\mathrm{sm}}\sigma^2_{\mathrm{sm},2}}{M\mu_{\mathrm{sm}}}\right)^{1/2} + \left(\frac{16\cdot 12^3\eta^4 L^2_{3,\mathrm{sm}} H(H-1)\sigma^4_{\mathrm{sm},4}}{\mu_{\mathrm{sm}}}\right)^{1/2} + \left(\frac{2\zeta^2_{\mathrm{sm}}}{\mu_{\mathrm{sm}}}\right)^{1/2} + \left(\frac{8\beta^2_{\mathrm{sm}}}{\mu_{\mathrm{sm}}}\right)^{1/2} + \frac{\zeta^2_{\mathrm{sm}}}{37 L_{2,\mathrm{sm}}} + \frac{\eta\sigma^2_{\mathrm{sm},2}}{12M} + \frac{\beta^2_{\mathrm{sm}}}{9L_{2,\mathrm{sm}}} + \frac{3\cdot 12^2\eta^4 L^2_{3,\mathrm{sm}} H(H-1)\sigma^4_{\mathrm{sm},4}}{L_{2,\mathrm{sm}}} \ .$$

Proof. First note that no projection is used, which implies that $\bar{\theta}^{r+1} = \theta^{r+1}$. Under **A**-1 and **A**-2, all the results of Appendix D.1 hold, satisfying thereby the assumptions of Lemma C.3. Importantly note that if $\eta H L_{2,\mathrm{sm}} \leq 1/74$ then it holds that $32\eta^2 H^2 L^2_{3,\mathrm{sm}} L^2_{1,\mathrm{sm}} \leq L^2_{2,\mathrm{sm}}$ (as $32 L^2_{3,\mathrm{sm}} L^2_{1,\mathrm{sm}} \leq 74^2 L^4_{2,\mathrm{sm}}$ by Lemma D.3, Lemma D.4, and Lemma D.5). Applying Lemma C.3 yields

$$-\mathbb{E}\left[J_{\mathrm{sm}}(\theta^{r+1})\middle|\mathcal{F}^r\right] \leq -J_{\mathrm{sm}}(\theta^r) - \frac{\eta H}{4}\|\nabla J_{\mathrm{sm}}(\theta^r)\|_2^2 + \frac{3\eta^2 L_{2,\mathrm{sm}} H\sigma^2_{\mathrm{sm},2}}{2M} + 2\eta H\beta^2_{\mathrm{sm}} + 8\eta^3 L^2_{2,\mathrm{sm}} H^2(H-1)\zeta^2_{\mathrm{sm}} + 4\cdot 12^3\eta^5 L^2_{3,\mathrm{sm}} H^2(H-1)\sigma^4_{\mathrm{sm},4} \ .$$

Adding $J^\star_{\mathrm{sm}}$, and using Lemma D.10 yields

$$\begin{aligned} &J^\star_{\mathrm{sm}} - \mathbb{E}\left[J_{\mathrm{sm}}(\theta^{r+1})\middle|\theta_r\right] \\ &\leq J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^r) - \frac{\eta H\mu_{\mathrm{sm}}}{4}(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^r))^2 + \frac{\eta H}{4}\zeta^2_{\mathrm{sm}} + \frac{3\eta^2 L_{2,\mathrm{sm}} H\sigma^2_{\mathrm{sm},2}}{2M} + 2\eta H\beta^2_{\mathrm{sm}} \\ &+ 8\eta^3 L^2_{2,\mathrm{sm}} H^2(H-1)\zeta^2_{\mathrm{sm}} + 4\cdot 12^3\eta^5 L^2_{3,\mathrm{sm}} H^2(H-1)\sigma^4_{\mathrm{sm},4} \ . \end{aligned} \tag{45}$$

Taking the expectation with respect to all the stochasticity, applying Jensen's inequality, and using that $\eta H L_{2,\mathrm{sm}} \leq 1/74$ to simplify the heterogeneity terms gives

$$\delta^{r+1} \leq \delta^r - \kappa(\delta^r)^2 + B \ ,$$

where we defined $\delta^r = J^\star_{\mathrm{sm}} - \mathbb{E}[J_{\mathrm{sm}}(\theta^r)]$, $\kappa = \frac{\eta H\mu_{\mathrm{sm}}}{4}$, and

$$B = \frac{\eta H}{2}\zeta^2_{\mathrm{sm}} + \frac{3\eta^2 L_{2,\mathrm{sm}} H\sigma^2_{\mathrm{sm},2}}{2M} + \frac{4\beta^2_{\mathrm{sm}}}{3L_{2,\mathrm{sm}}} + 4\cdot 12^3\eta^5 L^2_{3,\mathrm{sm}} H^2(H-1)\sigma^4_{\mathrm{sm},4} \ .$$

Finally, applying Lemma D.11 on the sequence δ^r concludes the proof. □

Recall that

$$L_{1,\mathrm{sm}} = \frac{1}{(1-\gamma)^2} \ , \ \ L_{2,\mathrm{sm}} = \frac{8}{(1-\gamma)^3} \ , \ \ L_{3,\mathrm{sm}} = \frac{480}{(1-\gamma)^4} \ , \ \ \zeta^2_{\mathrm{sm}} = \frac{38\varepsilon^2_{\mathsf{P}}}{(1-\gamma)^6} \ ,$$
$$\beta_{\mathrm{sm}} = \frac{2\gamma^T T}{1-\gamma} + \frac{2\gamma^T}{(1-\gamma)^2} \ , \ \ \sigma^2_{\mathrm{sm},2} = \frac{12}{(1-\gamma)^4 B} \ , \ \ \sigma^4_{\mathrm{sm},4} = \frac{1120}{(1-\gamma)^8 B^2} \ ,$$

which are defined respectively in Lemmas D.2 and D.4 to D.8. We obtain the following simplified result.

Corollary D.13 (Simplified convergence rates of S-FedPG)**.** *Assume **A**-1 and **A**-2, and no projection (i.e., set $\mathcal{T} = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$). Additionally, assume that there exists $\mu_{\mathrm{sm}} \in (0,1)$ such that, with probability 1, $\inf_{r\in[\mathbb{N}]} \mu_{\mathrm{sm}}(\theta^r) \geq \mu_{\mathrm{sm}}$. For any $\eta > 0$ such that $\eta H \leq (1-\gamma)^3/592$, $T \geq 4(1-\gamma)^{-2}$, and $M\cdot B \geq (1-\gamma)^{-1}$, the iterates of S-FedPG satisfy*

$$J^\star_{\mathrm{sm}} - \mathbb{E}[J_{\mathrm{sm}}(\theta^R)] \leq \frac{J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0)}{1 + R\cdot(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\cdot\eta H\mu_{\mathrm{sm}}/4} + \frac{25\eta^{1/2}}{\mu^{1/2}_{\mathrm{sm}} M^{1/2} B^{1/2}\cdot(1-\gamma)^{3.5}} + \frac{2\cdot 7^8\eta^2 H^{1/2}(H-1)^{1/2}}{\mu^{1/2}_{\mathrm{sm}}(1-\gamma)^8 B} + \frac{13T\gamma^T}{\mu^{1/2}_{\mathrm{sm}}(1-\gamma)} + \frac{10\varepsilon_{\mathsf{P}}}{\mu^{1/2}_{\mathrm{sm}}(1-\gamma)^3} \ .$$

Corollary D.14 (Sample and Communication Complexity of S-FedPG)**.** *Assume **A**-1 and **A**-2 and no projection (i.e., set $\mathcal{T} = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$). Additionally, assume that there exists $\mu_{\mathrm{sm}} \in (0,1)$ such that such that $\inf_{r\in[\mathbb{N}]} \mu_{\mathrm{sm}}(\theta^r) \geq \mu_{\mathrm{sm}} > 0$. Let $1 \geq \epsilon \geq 50\varepsilon_{\mathtt{P}}\mu_{\mathrm{sm}}^{-1/2}(1-\gamma)^{-3}$. Then, for a properly chosen truncation horizon, step size, and number of local updates, S-FedPG learns an ϵ-approximation of the optimal objective with a number of communication rounds*

$$R \geq \frac{11^4 \cdot [(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0)) - \epsilon/9]}{(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\mu_{\mathrm{sm}}\epsilon(1-\gamma)^3} \ ,$$

for a total number of sampled trajectories per agent of

$$RHB \geq \max\Big(\frac{8}{\mu_{\mathrm{sm}}(1-\gamma)^3}, \frac{5^6}{\mu^2_{\mathrm{sm}}MB(1-\gamma)^7\epsilon^2}, \frac{7^7}{B\mu^{3/2}_{\mathrm{sm}}\epsilon(1-\gamma)^5}\Big)\frac{36B[(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0)) - \epsilon/9]}{(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\epsilon} \ .$$

Proof. First, we set the truncation horizon to

$$T \geq \frac{2}{1-\gamma}\max\left(\frac{2}{1-\gamma}, \log\left(\frac{65}{\epsilon\mu^{1/2}_{\mathrm{sm}}(1-\gamma)}\right)\right) \ .$$

Secondly, we require (i) that $\eta \leq 1/L_{2,\mathrm{sm}}$, and that (ii) each variance terms to be smaller than $\epsilon/5$, which gives the condition on the step size

$$\eta \leq \min\Big((1-\gamma)^3/8, \frac{\mu_{\mathrm{sm}}MB(1-\gamma)^7\epsilon^2}{5^6}, \frac{B\mu^{1/2}_{\mathrm{sm},}\epsilon(1-\gamma)^5}{7^7}\Big) \ , \tag{46}$$

where the second element of the $\min$ comes from

$$\frac{25\eta^{1/2}}{\mu^{1/2}_{\mathrm{sm}}M^{1/2}B^{1/2}\cdot(1-\gamma)^{3.5}} \leq \frac{\epsilon}{5} \ ,$$

and the third comes from

$$\frac{2\cdot 7^8\eta^2H^{1/2}(H-1)^{1/2}}{\mu^{1/2}_{\mathrm{sm}}(1-\gamma)^8B} \leq \frac{2\cdot 7^6\eta}{6\mu^{1/2}_{\mathrm{sm}}(1-\gamma)^5B} \leq \frac{\epsilon}{5} \ .$$

Then, H has to satisfy $\eta H \leq (1-\gamma)^3/592$ This requires

$$H \leq \frac{(1-\gamma)^3}{592\eta} \ .$$

Finally, we require that the number of communications is at least

$$R \geq \frac{J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0) - \epsilon/5}{(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\eta H\mu_{\mathrm{sm}}\epsilon/20} \geq \frac{11^4L_{2,\mathrm{sm}}[(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0)) - \epsilon/9]}{(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\mu_{\mathrm{sm}}\epsilon(1-\gamma)^3} \ .$$

The sample complexity follows from $RHB \geq \frac{20[(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0)) - \epsilon/5]}{(J^\star_{\mathrm{sm}} - J_{\mathrm{sm}}(\theta^0))\epsilon}\frac{1}{\mu_{\mathrm{sm}}\eta}B$ and (46). □

E Analysis of RS-FedPG

RS-FedPG is a special instance of proj-FedAVG in which, the local objective function is $f_c = J_{\mathrm{r},c} =: J_{\mathrm{sm},c} + \lambda\mathcal{H}^\rho_c$, where for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ we have

$$\mathcal{H}^\rho_c(\theta) \triangleq -\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\log(\pi_\theta(A^t_c|S^t_c)) \mid S^0_c \sim \rho\right] \ . \tag{47}$$

We additionally define the global objective of the algorithm as $J_{\mathrm{r}} := \frac{1}{M}\sum_{c=1}^M J_{\mathrm{r},c}$. The client-specific data distribution $\xi_c(\theta)$ corresponds to $\nu_c(\theta)$, as defined in Eq. (3). For a given parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and an observation $Z_c \sim \nu_c(\theta)$, we define the biased stochastic estimator of the gradient of the local objective $J_{\mathrm{r},c}$ as:

$$\mathrm{g}^{Z_c}_{\mathrm{r},c}(\theta) := \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t\left(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(a^\ell_{c,b} \mid S^\ell_{c,b})\right)\left[\mathrm{r}(S^t_{c,b}, A^t_{c,b}) - \lambda\log(\pi_\theta(A^t_{c,b}, S^t_{c,b}))\right] \ . \tag{48}$$

We also define

$$\mathrm{g}_{\mathrm{r},c}(\theta) = \mathbb{E}_{Z_c\sim\nu_c(\theta)}[\mathrm{g}^{Z_c}_{\mathrm{r},c}(\theta)] \ .$$

The applicability of the ascent lemma relies on the validity of Assumptions **FL**-1–**FL**-6, which we proceed to verify in the subsequent analysis.

E.1 Checking the assumptions and establishing a local Łojasiewicz structure

For convenience, we recall the definitions of the regularised value function, the regularised Q-function, and the regularised advantage function defined in Geist et al. (2019):

$$\widetilde{V}_c^{\pi_\theta}(s) := V_c^{\pi_\theta}(s) + \lambda \mathcal{H}_c^s(\theta) \ , \ \text{where } \mathcal{H}_c^s(\theta) = -\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \log(\pi_\theta(A_c^t|S_c^t)) \mid S_c^0 = s\right] \tag{49}$$

$$\widetilde{Q}_c^{\pi_\theta}(s,a) := \mathsf{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathsf{P}_c(s'|s,a) \widetilde{V}_c^{\pi_\theta}(s') \ , \tag{50}$$

$$\widetilde{A}_c^{\pi_\theta}(s,a) := \widetilde{Q}_c^{\pi_\theta}(s,a) - \lambda \log(\pi_\theta(a \mid s)) - \widetilde{V}_c^{\pi_\theta}(s) \ , \quad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A} \ . \tag{51}$$

Following Mei et al. (2020), we will use the following expression of the gradient.

Lemma E.1 (Lemma 10 from Mei et al. (2020))**.** *We have*

$$\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_c^{\rho,\pi_\theta}(s) \pi_\theta(a|s) \widetilde{A}_c^{\pi_\theta}(s,a) \ , \tag{52}$$

where $\widetilde{A}_c^{\pi_\theta}$ is defined in (51)*.*

First, we establish the smoothness of $\mathrm{g}_{\mathrm{r},c}(\theta)$.

Lemma E.2. *For any $c \in [M]$, the function $\mathrm{g}_{\mathrm{r},c}$ is $L_{2,\mathrm{r}} \triangleq (8 + \lambda(4 + 8\log(|\mathcal{A}|))/(1-\gamma)^3$-smooth, that is for all $\theta, \theta' \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\|\mathrm{g}_{\mathrm{r},c}(\theta') - \mathrm{g}_{\mathrm{r},c}(\theta)\| \leq L_{2,\mathrm{r}} \|\theta' - \theta\|_2 \ .$$

Proof. Fix any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $c \in [M]$. Let $\mathfrak{T} := (S^0, A^0, \ldots, S^{T-1}, A^{T-1})$ be a random variable distributed according to $\nu_c(\theta)$, as defined in (3). Then, $\mathrm{g}_{\mathrm{r},c}(\theta)$ can be equivalently expressed as

$$\mathrm{g}_{\mathrm{r},c}(\theta) = \sum_{t=0}^{T-1} \gamma^t \sum_{\ell=0}^{t} \underbrace{\mathbb{E}_{\mathfrak{T}\sim\nu_c(\theta)}\left[\nabla \log \pi_\theta(A^\ell \mid S^\ell) \left(\mathsf{r}(S^t, A^t) - \lambda \log(\pi_\theta(A^t \mid S^t))\right)\right]}_{\mathrm{E}_\ell^t(\theta)} \ .$$

Denote by $\mathrm{E}_\ell^t(\theta, s, a)$ the coefficient at coordinate (s,a) of $\mathrm{E}_\ell^t(\theta)$. Using the REINFORCE formula (Lemma G.2), for any $(\bar{s}, \bar{a})$, we can express the partial derivative of $\mathrm{E}_\ell^t(\theta, s, a)$ with respect to $\theta(\bar{s}, \bar{a})$ as

$$\begin{aligned}
\frac{\partial \mathrm{E}_\ell^t(\theta,s,a)}{\partial \theta(\bar{s},\bar{a})} &= \frac{\partial}{\partial \theta(\bar{s},\bar{a})}\left[\mathbb{E}_{\mathfrak{T}\sim\nu_c(\theta)}\left[\frac{\partial \log \pi_\theta(A^\ell \mid S^\ell)}{\partial \theta(s,a)}\left(\mathsf{r}(S^t,A^t) - \lambda \log(\pi_\theta(A^t \mid S^t))\right)\right]\right] \\
&= \underbrace{\mathbb{E}_{\mathfrak{T}\sim\nu_c(\theta)}\left[\frac{\partial \log(\nu_c(\theta;\mathfrak{T}))}{\partial \theta(\bar{s},\bar{a})} \cdot \frac{\partial \log \pi_\theta(A^\ell \mid S^\ell)}{\partial \theta(s,a)}\left(\mathsf{r}(S^t,A^t) - \lambda \log(\pi_\theta(A^t \mid S^t))\right)\right]}_{\mathrm{F}_\ell^t(s,a,\bar{s},\bar{a})} \\
&\quad + \underbrace{\mathbb{E}_{\mathfrak{T}\sim\nu_c(\theta)}\left[\frac{\partial^2 \log \pi_\theta(A^\ell \mid S^\ell)}{\partial \theta(s,a)\partial\theta(\bar{s},\bar{a})}\left(\mathsf{r}(S^t,A^t) - \lambda \log(\pi_\theta(A^t \mid S^t))\right)\right]}_{\mathrm{G}_\ell^t(s,a,\bar{s},\bar{a})} \\
&\quad - \lambda \underbrace{\left[\mathbb{E}_{\mathfrak{T}\sim\nu_c(\theta)}\left[\frac{\partial \log \pi_\theta(A^\ell \mid S^\ell)}{\partial \theta(s,a)} \cdot \frac{\partial \log(\pi_\theta(A^t \mid S^t))}{\partial \theta(\bar{s},\bar{a})}\right]\right]}_{\mathrm{H}_\ell^t(s,a,\bar{s},\bar{a})} \ .
\end{aligned}$$

We now bound each of these three terms separately. Beforehand, recall that for any $(s,a,\bar{s},\bar{a})$, we have

$$\frac{\partial \pi_\theta(a \mid s)}{\partial \theta(\bar{s},\bar{a})} = 1_{\bar{s}}(s)(1_{\bar{a}}(a)\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(\bar{a}|s)) \ . \tag{53}$$

Bounding $\mathrm{F}_\ell^t(s,a,\bar{s},\bar{a})$. Using (53), note that, for any (s,a,s^ℓ,a^ℓ), we have

$$\frac{\partial \log(\pi_\theta(a^\ell \mid s^\ell))}{\partial \theta(s,a)} = \mathsf{1}_s(s^\ell)\left(\mathsf{1}_a(a^\ell) - \pi_\theta(a|s)\right) \enspace .$$

Now, consider a trajectory $z = (s^0, a^0, \dots s^{T-1}, a^{T-1})$. It holds that

$$\frac{\partial \log(\nu_c(\theta; z))}{\partial \theta(\bar{s},\bar{a})} = \sum_{k=0}^{T-1} \mathsf{1}_{\bar{s}}(s^k)\left(\mathsf{1}_{\bar{a}}(a^k) - \pi_\theta(\bar{a}|\bar{s})\right) \enspace .$$

Additionally, note that for $k \geq \max(t,\ell)$, we have

$$\mathbb{E}_{\mathfrak{T}}\left[\mathsf{1}_{\bar{s}}(S^k)\left(\mathsf{1}_{\bar{a}}(A^k) - \pi_\theta(\bar{a}|\bar{s})\right) \cdot \frac{\partial \log(\pi_\theta(A^\ell \mid S^\ell))}{\partial \theta(s,a)}\left(\mathsf{r}(S^t,A^t) - \lambda \log(\pi_\theta(A^t \mid S^t))\right)\right] = 0 \enspace .$$

Combining the three previous identities, the triangle inequality and the fact that the reward is bounded by 1 yields

$$\begin{aligned}
\left|\mathrm{F}_\ell^t(s,a,\bar{s},\bar{a})\right| \leq & \sum_{k=0}^{t} \mathbb{E}\left[\mathsf{1}_{\bar{s}}(S^k)\mathsf{1}_s(S^\ell)\left(\mathsf{1}_{\bar{a}}(A^k)\mathsf{1}_a(A^\ell) + \pi_\theta(\bar{a}|\bar{s})\mathsf{1}_a(A^\ell)\right)\right] \\
& + \sum_{k=0}^{t} \mathbb{E}\left[\mathsf{1}_{\bar{s}}(S^k)\mathsf{1}_s(S^\ell)\left(\mathsf{1}_{\bar{a}}(A^k)\pi_\theta(a \mid s) + \pi_\theta(a \mid s)\pi_\theta(\bar{a}|\bar{s})\right)\right] \\
& - \lambda \sum_{k=0}^{t} \mathbb{E}\left[\mathsf{1}_{\bar{s}}(S^k)\mathsf{1}_s(S^\ell)\left(\mathsf{1}_{\bar{a}}(A^k)\mathsf{1}_a(A^\ell) + \pi_\theta(\bar{a}|\bar{s})\mathsf{1}_a(A^\ell)\right)\log \pi_\theta(A^t \mid S^t)\right] \\
& - \lambda \sum_{k=0}^{t} \mathbb{E}\left[\mathsf{1}_{\bar{s}}(S^k)\mathsf{1}_s(S^\ell)\left(\mathsf{1}_{\bar{a}}(A^k)\pi_\theta(a \mid s) + \pi_\theta(\bar{a}|\bar{s})\pi_\theta(a \mid s)\right)\log \pi_\theta(A^t \mid S^t)\right] \enspace .
\end{aligned}$$

Bounding $\mathrm{G}_\ell^t(s,a,\bar{s},\bar{a})$. Consider a trajectory $z = (s^0, a^0, \dots s^{T-1}, a^{T-1})$. It holds that

$$\frac{\partial \log(\pi_\theta(a^\ell \mid s^\ell))}{\partial \theta(s,a)} = \mathsf{1}_s(s^\ell)\left(\mathsf{1}_a(a^\ell) - \pi_\theta(a|s)\right) \enspace .$$

Next, deriving with respect to $\theta(\bar{s},\bar{a})$ yields

$$\frac{\partial^2 \log \pi_\theta(a^\ell \mid s^\ell)}{\partial \theta(s,a)\partial \theta(\bar{s},\bar{a})} = -\mathsf{1}_{\bar{s}}(s^\ell)\mathsf{1}_{\bar{s}}(s)\left[\mathsf{1}_a(\bar{a})\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(\bar{a}|\bar{s})\right] \enspace .$$

Combining the previous equality, the triangle inequality, and using that the reward is bounded by 1 yields

$$\begin{aligned}
\left|\mathrm{G}_\ell^t(s,a,\bar{s},\bar{a})\right| \leq & \, \mathsf{1}_{\bar{s}}(s)\mathsf{1}_{\bar{a}}(a)\mathbb{E}[\mathsf{1}_{\bar{s}}(S^\ell)]\pi_\theta(a \mid s) + \mathsf{1}_{\bar{s}}(s)\mathbb{E}[\mathsf{1}_{\bar{s}}(S^\ell)]\pi_\theta(a \mid s)\pi_\theta(\bar{a} \mid \bar{s}) \\
& - \lambda \mathsf{1}_{\bar{s}}(s)\mathsf{1}_{\bar{a}}(a)\mathbb{E}[\mathsf{1}_{\bar{s}}(S^\ell)\log \pi_\theta(A^t \mid S^t)]\pi_\theta(a \mid s) \\
& - \lambda \mathsf{1}_{\bar{s}}(s)\mathbb{E}[\mathsf{1}_{\bar{s}}(S^\ell)\log \pi_\theta(A^t \mid S^t)]\pi_\theta(a \mid s)\pi_\theta(\bar{a} \mid \bar{s}) \enspace .
\end{aligned}$$

Bounding $\mathrm{H}_\ell^t(s,a,\bar{s},\bar{a})$. Applying the triangle inequality yields

$$\begin{aligned}
|\mathrm{H}_\ell^t(s,a,\bar{s},\bar{a})| &= \lambda\left|\mathbb{E}_{\mathfrak{T}}\left[\mathsf{1}_s(S^\ell)\left(\mathsf{1}_a(A^\ell) - \pi_\theta(a \mid s)\right)\mathsf{1}_{\bar{s}}(S^t)\left(\mathsf{1}_{\bar{a}}(A^t) - \pi_\theta(\bar{a} \mid \bar{s})\right)\right]\right| \\
&\leq \lambda \mathbb{E}_{\mathfrak{T}}\left[\mathsf{1}_s(S^\ell)\mathsf{1}_{\bar{s}}(S^t)\left(\mathsf{1}_a(A^\ell)\mathsf{1}_{\bar{a}}(A^t) + \mathsf{1}_a(A^\ell)\pi_\theta(\bar{a} \mid \bar{s})\right)\right] \\
&\quad + \lambda \mathbb{E}_{\mathfrak{T}}\left[\mathsf{1}_s(S^\ell)\mathsf{1}_{\bar{s}}(S^t)\left(\pi_\theta(a \mid s)\mathsf{1}_{\bar{a}}(A^t) + \pi_\theta(a \mid s)\pi_\theta(\bar{a} \mid \bar{s})\right)\right] \enspace .
\end{aligned}$$

Denote by $\mathrm{g}_{\mathrm{r},c}(\theta,s,a)$ the coefficient at coordinate (s,a) of $\mathrm{g}_{\mathrm{r},c}(\theta)$. Applying the triangle inequality yields

$$\left|\frac{\partial \mathrm{g}_{\mathrm{sm},c}(\theta,s,a)}{\partial \theta(\bar{s},\bar{a})}\right| \leq \sum_{t=0}^{T-1} \gamma^t \sum_{\ell=0}^{t}\left[\left|\mathrm{F}_\ell^t(s,a,\bar{s},\bar{a})\right| + \left|\mathrm{G}_\ell^t(s,a,\bar{s},\bar{a})\right| + \left|\mathrm{H}_\ell^t(s,a,\bar{s},\bar{a})\right|\right]$$

Using that for any $s' \in \mathcal{S}$, $\sum_{s\in\mathcal{S}} 1_s(s') = 1$, and that for any $a' \in \mathcal{A}$, $\sum_{a\in\mathcal{A}} 1_s(a') = 1$ gives

$$\sum_{s,a,\bar{s},\bar{a}} \left| \frac{\partial g_{\mathrm{sm},c}(\theta, s, a)}{\partial \theta(\bar{s}, \bar{a})} \right| \leq \sum_{t=0}^{T-1} \gamma^t \sum_{\ell=0}^{t} \left[2t + 2 + 4\lambda - (2\lambda + 2t\lambda) \mathbb{E}_{\mathfrak{T}} \left[\log \pi_\theta(A^t \mid S^t) \right] \right]$$

Now using that for any $x \in [0, 1[$, $\sum_{k=0}^{\infty} k^2 x^k \leq 2/(1-x)^3$, $\sum_{k=0}^{\infty} k x^k \leq 1/(1-x)^2$, and that $-\mathbb{E}_{\mathfrak{T}} \left[\log \pi_\theta(A^t \mid S^t) \right] \leq \log(|\mathcal{A}|)$ yields

$$\left\| \frac{\partial g_{\mathrm{sm},c}}{\partial \theta} \right\|_2 \leq \left\| \frac{\partial g_{\mathrm{sm},c}}{\partial \theta} \right\|_1 \leq \frac{8 + \lambda(4 + 8\log(|\mathcal{A}|))}{(1-\gamma)^3} \ ,$$

which concludes the proof. □

Lemma E.3. *For any $c \in [M]$, $J_{\mathrm{r},c}$ and J_{r} are $L_{2,\mathrm{r}} := (8 + \lambda(4 + 8\log(|\mathcal{A}|))/(1-\gamma)^3$-smooth.*

Proof. Follows from (Mei et al., 2020, Lemma 14) and the properties of averaging of smooth functions. □

Lemma E.4. *For all $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds*

$$\|\nabla J_{\mathrm{r},c}(\theta)\|_2 \leq L_{1,\mathrm{r}} \ , \quad \textit{where } L_{1,\mathrm{r}} := \frac{1 + \lambda \log(|\mathcal{A}|)}{(1-\gamma)^2} \ .$$

Proof. By norm comparisons, and Lemma E.1, it holds that

$$\|J_{\mathrm{r},c}(\theta)\|_2 \leq \|J_{\mathrm{r},c}(\theta)\|_1 = \frac{1}{1-\gamma} \sum_{s,a} d_c^{\rho,\pi_\theta}(s) \pi_\theta(a|s) |\widetilde{A}_c^{\pi_\theta}(s,a)| \ .$$

Now, using that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $|\widetilde{A}_c^{\pi_\theta}(s,a)| \leq (1 + \lambda \log(|\mathcal{A}|))/(1-\gamma)$ yields

$$\|J_{\mathrm{r},c}(\theta)\|_2 \leq \|J_{\mathrm{r},c}(\theta)\|_1 \leq \frac{1 + \lambda \log(|\mathcal{A}|)}{(1-\gamma)^2} \sum_{s,a} d_c^{\rho,\pi_\theta}(s) \pi_\theta(a|s) = \frac{1 + \lambda \log(|\mathcal{A}|)}{(1-\gamma)^2} \ .$$

which concludes the proof. □

Lemma E.5. *The spectral norm of the third derivative tensor of $J_{\mathrm{r},c}$ is bounded by $L_{3,\mathrm{r}} := (480 + 832\lambda \log|\mathcal{A}|) \cdot (1-\gamma)^{-4}$, i.e., for any $u, v, w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ it holds*

$$|\mathrm{d}^3 J_{\mathrm{r},c}(\theta)[u,v,w]| = |\nabla^3 J_{\mathrm{r},c}(\theta) u \otimes v \otimes w| \leq \frac{480 + 832\lambda \log|\mathcal{A}|}{(1-\gamma)^4} \|u\|_2 \|v\|_2 \|w\|_2 \, .$$

Proof. By Lemma G.9 we have for any $u, v, w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

$$\|\mathrm{d}^3 \widetilde{V}_c^{\pi_\theta}[u,v,w]\|_\infty \leq \frac{480 + 832\lambda \log|\mathcal{A}|}{(1-\gamma)^4} \|u\|_2 \|v\|_2 \|w\|_2 \, .$$

Next, we notice that

$$\mathrm{d}^3 J_{\mathrm{r},c}(\theta)[u,v,w] = \rho^\top \mathrm{d}^3 V_c^{\pi_\theta}[u,v,w] \, ,$$

thus

$$|\mathrm{d}^3 J_{\mathrm{r},c}(\theta)[u,v,w]| \leq \frac{480 + 832\lambda \log|\mathcal{A}|}{(1-\gamma)^4} \|u\|_2 \|v\|_2 \|w\|_2 \, .$$

□

Lemma E.6. *Let $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. It holds that*

$$\|\nabla J_{\mathrm{r}}(\theta) - \nabla J_{\mathrm{r},c}(\theta)\|_2^2 \leq \zeta_{\mathrm{r}}^2 \ , \quad \textit{where } \zeta_{\mathrm{r}}^2 := \frac{38(1 + \lambda \log(|\mathcal{A}|))^2 \varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} \ .$$

Proof. Fix $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Using Lemma E.1, we have

$$\left|\frac{\partial J_{\mathrm{r}}(\theta)}{\partial\theta(s,a)} - \frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,a)}\right| \leq \frac{1}{M}\frac{1}{1-\gamma}\sum_{k=1}^{M}\pi_\theta(a|s)\left|d_k^{\rho,\pi_\theta}(s)\widetilde{A}_k^{\pi_\theta}(s,a) - d_c^{\rho,\pi_\theta}(s)\widetilde{A}_c^{\pi_\theta}(s,a)\right|$$

$$\leq \frac{1}{M}\frac{\pi_\theta(a|s)}{1-\gamma}\sum_{k=1}^{M}|d_k^{\rho,\pi_\theta}(s) - d_c^{\rho,\pi_\theta}(s)|\underbrace{\left|\widetilde{A}_k^{\pi_\theta}(s,a)\right|}_{(\mathbf{X})} + \underbrace{\left|\widetilde{A}_k^{\pi_\theta}(s,a) - \widetilde{A}_c^{\pi_\theta}(s,a)\right|}_{(\mathbf{Y})} d_c^{\rho,\pi_\theta}(s) \ .$$

We bound each of $(\mathbf{X})$ and $(\mathbf{Y})$ separately.

Bounding $(\mathbf{X})$. First note that we have

$$0 \leq \widetilde{V}_c^{\pi_\theta}(s) \leq \frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma} \ , \tag{54}$$

where $\widetilde{V}_c^{\pi_\theta}(s)$ is defined in (49). Combining the previous inequality and applying the triangle inequality yields

$$\begin{aligned}(\mathbf{X}) &= \left|\mathsf{r}(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_k(s'|s,a)\widetilde{V}_{k,\pi_\theta}(s') - \lambda\log(\pi_\theta(a|s)) - \widetilde{V}_{k,\pi_\theta}(s)\right| \\ &\leq \frac{2+2\lambda\log(|\mathcal{A}|)}{1-\gamma} + \lambda|\log(\pi_\theta(a|s))| \ .\end{aligned}$$

Bounding $(\mathbf{Y})$. Using the triangle inequality, we get

$$\begin{aligned}(\mathbf{Y}) &= \left|\gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_k(s'|s,a)\widetilde{V}_k^{\pi_\theta}(s') - \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_c(s'|s,a)\widetilde{V}_c^{\pi_\theta}(s') + \widetilde{V}_c^{\pi_\theta}(s') - \widetilde{V}_k^{\pi_\theta}(s')\right| \\ &\leq \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_k(s'|s,a)\left|\widetilde{V}_k^{\pi_\theta}(s') - \widetilde{V}_c^{\pi_\theta}(s')\right| + \gamma\sum_{s'\in\mathcal{S}}|\mathsf{P}_k(s'|s,a) - \mathsf{P}_c(s'|s,a)|\,\widetilde{V}_c^{\pi_\theta}(s') \\ &+ \left|\widetilde{V}_k^{\pi_\theta}(s) - \widetilde{V}_c^{\pi_\theta}(s)\right| \ .\end{aligned}$$

Using **A**-1 and (54), we obtain

$$(\mathbf{Y}) \leq \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_k(s'|s,a)\left|\widetilde{V}_k^{\pi_\theta}(s') - \widetilde{V}_c^{\pi_\theta}(s')\right| + \frac{(1+\lambda\log(|\mathcal{A}|))\,\varepsilon_{\mathsf{P}}}{1-\gamma} + \left|\widetilde{V}_k^{\pi_\theta}(s) - \widetilde{V}_c^{\pi_\theta}(s)\right|$$

Using (49), note that we have

$$\left|\widetilde{V}_k^{\pi_\theta}(s) - \widetilde{V}_c^{\pi_\theta}(s)\right| \leq |V_k^{\pi_\theta}(s) - V_c^{\pi_\theta}(s)| + \lambda\,|\mathcal{H}_k^s(\theta) - \mathcal{H}_c^s(\theta)| \ .$$

The bound on the first term of the previous bound is provided by Lemma G.5. For the second term, we have

$$\begin{aligned}\lambda\,|\mathcal{H}_k^s(\theta) - \mathcal{H}_c^s(\theta)| &\leq \frac{\lambda}{1-\gamma}\sum_{s_0\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left|d_k^{s,\theta}(s_0) - d_c^{s,\theta}(s_0)\right|\,|\pi_\theta(a|s_0)\log(\pi_\theta(a|s_0))| \\ &\leq \frac{\lambda\log(|\mathcal{A}|)}{1-\gamma}\sum_{s_0\in\mathcal{S}}\left|d_k^{s,\theta}(s_0) - d_c^{s,\theta}(s_0)\right| \ ,\end{aligned}$$

where in the last inequality we used $-\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\log(\pi_\theta(a|s)) \leq \log(|\mathcal{A}|)$. Finally plugging in the bound of Lemma G.6 yields

$$\lambda\,|\mathcal{H}_k^s(\theta) - \mathcal{H}_c^s(\theta)| \leq \frac{\lambda\log(|\mathcal{A}|)\varepsilon_{\mathsf{P}}}{(1-\gamma)^2} \ .$$

Thus, we get the following bound on $(\mathbf{Y})$

$$(\mathbf{Y}) \leq 3\cdot\frac{(1+\lambda\log(|\mathcal{A}|))\varepsilon_{\mathsf{P}}}{(1-\gamma)^2} \ .$$

Combining the bounds on $(\mathbf{X})$ and $(\mathbf{Y})$ yields

$$\begin{aligned}
&\left|\frac{\partial J_{\mathrm{r}}(\theta)}{\partial\theta(s,a)}-\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,a)}\right| \\
&\quad\leq\frac{1}{M}\sum_{k=1}^{M}\left[\frac{2(1+\lambda\log(|\mathcal{A}|))}{1-\gamma}+\lambda|\log(\pi(a|s))|\right]\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|\frac{\pi_\theta(a|s)}{1-\gamma} \\
&\quad+\frac{3(1+\lambda\log(|\mathcal{A}|))\varepsilon_{\mathsf{P}}}{(1-\gamma)^2}\cdot\frac{d_c^{\rho,\theta}(s)\pi_\theta(a|s)}{1-\gamma}
\end{aligned}$$

Thus, we get by Young's inequality

$$\begin{aligned}
&\|\nabla J_{\mathrm{r}}(\theta)-\nabla J_{\mathrm{r},c}(\theta)\|_2^2 \\
&\quad=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\frac{\partial J_{\mathrm{r}}(\theta)}{\partial\theta(s,a)}-\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,a)}\right|^2 \\
&\quad\leq\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}2\cdot\left(\frac{1}{M}\sum_{k=1}^{M}\left[\frac{2(1+\lambda\log(|\mathcal{A}|))}{1-\gamma}+\lambda|\log(\pi(a|s))|\right]\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|\frac{\pi_\theta(a|s)}{1-\gamma}\right)^2 \\
&\quad+\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}2\cdot\left(\frac{3(1+\lambda\log(|\mathcal{A}|))\varepsilon_{\mathsf{P}}}{(1-\gamma)^2}\cdot\frac{d_c^{\rho,\theta}(s)\pi_\theta(a|s)}{1-\gamma}\right)^2 \quad .
\end{aligned}$$

Now applying Jensen's inequality yields

$$\begin{aligned}
&\|\nabla J_{\mathrm{r}}(\theta)-\nabla J_{\mathrm{r},c}(\theta)\|_2^2 \\
&\quad\leq\frac{2}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^{M}\left[\frac{2(1+\lambda\log(|\mathcal{A}|))}{1-\gamma}+\lambda|\log(\pi(a|s))|\right]^2\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|^2\frac{\pi_\theta(a|s)^2}{(1-\gamma)^2} \\
&\quad+\frac{2}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{M}\frac{9(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}\cdot d_c^{\rho,\theta}(s)^2\pi_\theta(a|s)^2 \\
&\quad\leq\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^{M}\frac{16(1+\lambda\log(|\mathcal{A}|))^2}{(1-\gamma)^2}\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|^2\frac{\pi_\theta(a|s)^2}{(1-\gamma)^2} \\
&\quad+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^{M}4\lambda^2|\log(\pi(a|s))|^2\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|^2\frac{\pi_\theta(a|s)^2}{(1-\gamma)^2} \\
&\quad+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{M}\frac{18(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}\cdot d_c^{\rho,\theta}(s)^2\pi_\theta(a|s)^2 \quad ,
\end{aligned}$$

For the first term, using that $\pi_\theta(a|s)\leq 1$, $|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)|\leq\varepsilon_{\mathsf{P}}/(1-\gamma)$, for the second term using that $|\log(\pi_\theta(a|s))|\pi_\theta(a|s)\leq 1$, $|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)|\leq\varepsilon_{\mathsf{P}}/(1-\gamma)$, and for the third term applying that $\pi_\theta(a|s)d_c^{\rho,\theta}(s)\leq 1$ gives

$$\begin{aligned}
\|\nabla J_{\mathrm{r}}(\theta)-\nabla J_{\mathrm{r},c}(\theta)\|_2^2 \quad&\leq\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^{M}\frac{16(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}}{(1-\gamma)^3}\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|\frac{\pi_\theta(a|s)}{(1-\gamma)^2} \\
&+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^{M}4\lambda^2|\log(\pi(a|s))|\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|\frac{\pi_\theta(a|s)\varepsilon_{\mathsf{P}}}{(1-\gamma)^3} \\
&+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{M}\frac{18(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}\cdot d_c^{\rho,\theta}(s)\pi_\theta(a|s) \quad ,
\end{aligned}$$

Finally, for the first term using that $\sum_{a\in\mathcal{A}} |\log(\pi_\theta(a|s))|\pi_\theta(a|s) \le \log(|\mathcal{A}|)$, and using Lemma G.6 for both the first and second term yields

$$\|\nabla J_{\mathrm{r}}(\theta) - \nabla J_{\mathrm{r},c}(\theta)\|_2^2 \le \frac{38(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} \ ,$$

which concludes the proof. □

The following lemma bounds the bias and the variance of this stochastic gradient.

Lemma E.7 (Lemma 6 from Ding et al. (2025))**.** *Consider the stochastic gradient defined in* (48)*. We have*

$$\|\mathrm{g}_{\mathrm{r},c}(\theta) - \nabla J_{\mathrm{r},c}(\theta)\|_2 \le \beta_{\mathrm{r}} := \frac{2(1+\lambda\log(|\mathcal{A}|))\gamma^T}{1-\gamma}\left(T + \frac{1}{1-\gamma}\right) \ ,$$
$$\mathrm{Var}(\mathrm{g}_{\mathrm{r},c}^{Z_c}) \le \sigma_{\mathrm{r},2}^2 := \frac{12 + 24\lambda^2(\log(|\mathcal{A}|))^2}{B(1-\gamma)^4} \ .$$

Finally, we show that the fourth-order moment of our biased estimator is bounded.

Lemma E.8. *For any $c \in [M]$, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, the fourth central moment of $\mathrm{g}_{\mathrm{r},c}^{Z_c}$ is bounded, that is*

$$\mathbb{E}_{Z_c\sim\nu_c(\theta)}\left[\|\mathrm{g}_{\mathrm{r},c}^{Z_c}(\theta) - \mathrm{g}_{\mathrm{r},c}(\theta)\|_2^4\right] \le \sigma_{\mathrm{r},4}^4 := \frac{1120 + 4480\lambda^4\log(|\mathcal{A}|)^4}{B^2(1-\gamma)^8} \ . \tag{55}$$

Proof. Fix $c \in [M]$, $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and an observation $Z_c = (Z_{c,1}, \dots Z_{c,B}) \sim \nu_c(\theta)^{\otimes B}$. For more readability of the proof, we define for any $z = (s^t, a^t)_{t=0}^{T-1} \in (\mathcal{S}\times\mathcal{A})^T$:

$$\mathrm{u}(z) \triangleq \sum_{t=0}^{T-1}\gamma^t\left(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(a^t \mid s^t)\right)\left[\mathsf{r}(s^t,a^t) - \lambda\log(\pi_\theta(a^t \mid s^t)\right] \ .$$

Importantly, note that

$$\mathrm{g}_{\mathrm{r},c}^{Z_c}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\mathrm{u}(Z_{c,b}) \ , \quad \text{and } \mathrm{g}_{\mathrm{r},c}(\theta) = \bar{\mathrm{u}} \ ,$$

where we define $\bar{\mathrm{u}} = \mathbb{E}_{Z_{c,b}\sim\nu_c(\theta)}[\mathrm{u}(Z_{c,b})]$. Using this decomposition, we can bound the fourth order of $\mathrm{g}_{\mathrm{r},c}^{Z_c}(\theta)$ by the fourth central moment of $\mathrm{u}(Z_{c,b})$. Indeed, expanding the norm to the fourth power yields

$$\mathbb{E}_{Z_c\sim\nu_c(\theta)^{\otimes B}}\left[\|\mathrm{g}_{\mathrm{r},c}^{Z_c}(\theta) - \mathrm{g}_{\mathrm{r},c}(\theta)\|_2^4\right] = \mathbb{E}_{Z_c}\left[\left\|\frac{1}{B}\sum_{b=1}^{B}[\mathrm{u}(Z_{c,b}) - \bar{\mathrm{u}}]\right\|_2^4\right]$$
$$= \frac{1}{B^4}\sum_{b_1=1}^{B}\sum_{b_2=1}^{B}\sum_{b_3=1}^{B}\sum_{b_4=1}^{B}\underbrace{\mathbb{E}_{Z_c}\left[\langle\mathrm{u}(Z_{c,b_1}) - \bar{\mathrm{u}}, \mathrm{u}(Z_{c,b_2}) - \bar{\mathrm{u}}\rangle\langle\mathrm{u}(Z_{c,b_3}) - \bar{\mathrm{u}}, \mathrm{u}(Z_{c,b_4}) - \bar{\mathrm{u}}\rangle\right]}_{U(b_1,b_2,b_3,b_4)} \ .$$

Note that by independence of the trajectories, $U(b_1,b_2,b_3,b_4)$ is non-zero if and only if all of the indices are equal or there are two pairs of equal indices. In this case, as the trajectories are identically distributed, $U(b_1,b_2,b_3,b_4)$ is respectively equal to $\mathbb{E}\left[\|\mathrm{u}(Z_{c,1}) - \bar{\mathrm{u}}\|_2^4\right]$ and $\mathbb{E}\left[\|\mathrm{u}(Z_{c,1}) - \bar{\mathrm{u}}\|^2\right]^2$. There are exactly B combinations where all indices are equal, and

$$\frac{B(B-1)}{2}\cdot\frac{4\cdot 3}{2}$$

combinations corresponding to the two distinct pairs of equal indices case. Combining these, we arrive at the following identity:

$$\mathbb{E}_{Z_c}\left[\|\mathrm{g}_{\mathrm{r},c}^{Z_c}(\theta) - \mathrm{g}_{\mathrm{r},c}(\theta)\|_2^4\right] = \frac{1}{B^3}\left[\underbrace{\mathbb{E}\left[\|\mathrm{u}(Z_{c,1}) - \bar{\mathrm{u}}\|_2^4\right]}_{(\mathbf{M})} + 3(B-1)\mathbb{E}\left[\|\mathrm{u}(Z_{c,1}) - \bar{\mathrm{u}}\|^2\right]^2\right] \ . \tag{56}$$

We now decompose $\mathrm{u}(Z_{c,1})$ into two components, one that comes from the rewards of the MDP and a second associated with the regularization. Precisely, we define

$$\mathrm{u_r}(Z_{c,1}) \triangleq \sum_{t=0}^{T-1} \gamma^t \left(\sum_{\ell=0}^{t} \nabla \log \pi_\theta(A_{c,1}^\ell \mid S_{c,1}^\ell) \right) \mathsf{r}(S_{c,1}^t, A_{c,1}^t) \ ,$$
$$\mathrm{u}_\lambda(Z_{c,1}) \triangleq -\lambda \sum_{t=0}^{T-1} \gamma^t \left(\sum_{\ell=0}^{t} \nabla \log \pi_\theta(A_{c,1}^\ell \mid S_{c,1}^\ell) \right) \log(\pi_\theta(A_{c,1}^t, S_{c,1}^t)) \ .$$

Additionally, define $\mathrm{u_r}$ and u_λ respectively as the expectations of $\mathrm{u_r}(Z_{c,1})$ and $\mathrm{u}_\lambda(Z_{c,1})$. Importantly, note that

$$\mathrm{u}(Z_{c,1}) = \mathrm{u_r}(Z_{c,1}) + \mathrm{u}_\lambda(Z_{c,1}) \ , \quad \text{and } \bar{\mathrm{u}} = \mathrm{u_r} + \mathrm{u}_\lambda \ .$$

Thus, using the triangle inequality, we have

$$\begin{aligned}
(\mathbf{M}) &\leq \mathbb{E}_{Z_c \sim \nu_c(\theta)} \left[\left(\|\mathrm{u_r}(Z_{c,1}) - \mathrm{u_r}\|_2 + \|\mathrm{u}_\lambda(Z_{c,1}) - \mathrm{u}_\lambda\|_2 \right)^4 \right] \\
&\leq \mathbb{E}_{Z_c \sim \nu_c(\theta)} \left[\left(2 \max(\|\mathrm{u_r}(Z_{c,1}) - \mathrm{u_r}\|_2, \|\mathrm{u}_\lambda(Z_{c,1}) - \mathrm{u}_\lambda\|_2) \right)^4 \right] \\
&\leq 16 \underbrace{\mathbb{E}_{Z_c} \left[\|\mathrm{u_r}(Z_{c,1}) - \mathrm{u_r}\|_2^4 \right]}_{(\mathbf{M_1})} + 16 \underbrace{\mathbb{E}_{Z_c} \left[\|\mathrm{u}_\lambda(Z_{c,1}) - \mathrm{u}_\lambda\|_2^4 \right]}_{(\mathbf{M_2})}
\end{aligned}$$

Subsequently, we bound each of these two terms separately.

Bounding $(\mathbf{M_1})$. Applying the triangle inequality, combined with Jensen's inequality, and using the fact that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\nabla \log(\pi_\theta(a \mid s))\|_2 \leq 2$ (see, e.g., proof of Lemma 7 in Ding et al. (2025)), gives

$$(\mathbf{M_1}) = \mathbb{E}_{Z_c} \left[\|\mathrm{u_r}(Z_{c,1}) - \mathrm{u_r}\|_2^4 \right] \leq \mathbb{E}_{Z_c} \left[\|\mathrm{u_r}(Z_{c,1})\|_2^4 \right] \leq \left(\sum_{t=0}^{T-1} 2t\gamma^t \right)^4 \leq \frac{16}{(1-\gamma)^8} \ ,$$

where in the last inequality, we used that for any $x \in [0, 1[$, $\sum_{k=0}^{\infty} k x^k \leq 1/(1-x)^2$.

Bounding $(\mathbf{M_2})$. Applying the triangle inequality combined with Jensen's inequality yields

$$\begin{aligned}
(\mathbf{M_2}) &\leq \mathbb{E}_{Z_c} \left[\|\mathrm{u}_\lambda(Z_{c,1})\|_2^4 \right] \\
&= \lambda^4 \mathbb{E}_{Z_c} \left[\| \sum_{t=0}^{T-1} \gamma^t \left(\sum_{\ell=0}^{t} \nabla \log \pi_\theta(A_{c,1}^\ell \mid S_{c,1}^\ell) \right) \log(\pi_\theta(A_{c,1}^t, S_{c,1}^t)) \|_2^4 \right] \\
&\leq \lambda^4 \mathbb{E}_{Z_c} \left[\left(\sum_{t=0}^{T-1} \gamma^t \sum_{\ell=0}^{t} \|\nabla \log \pi_\theta(A_{c,1}^\ell \mid S_{c,1}^\ell)\|_2 |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))| \right)^4 \right] \ ,
\end{aligned}$$

where in the last inequality, we used the triangle inequality. Now, using that $\|\nabla \log(\pi_\theta(a \mid s))\|_2 \leq 2$, we obtain

$$(\mathbf{M_2}) \leq \lambda^4 \mathbb{E}_{Z_c} \left[\left(\sum_{t=0}^{T-1} 2t\gamma^t |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))| \right)^4 \right] \ .$$

Next, applying Cauchy-Schwarz inequality gives

$$\begin{aligned}
(\mathbf{M_2}) &\leq \lambda^4 \mathbb{E}_{Z_c} \left[\left(\sum_{t=0}^{T-1} 2t\gamma^{t/2}\gamma^{t/2} |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))| \right)^4 \right] \\
&\leq \lambda^4 \left(\sum_{t=0}^{T-1} 4t^2\gamma^t \right)^2 \mathbb{E}_{Z_c} \left[\left(\sum_{t=0}^{T-1} \gamma^t |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))|^2 \right)^2 \right]
\end{aligned}$$

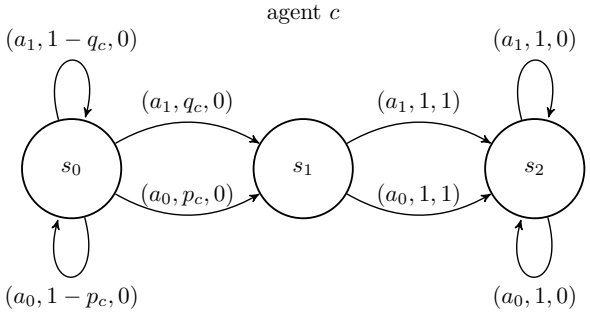


Figure 5: FRL task with no optimal local deterministic policy. The triplet means (action, probability, reward) , $\gamma = 0.999$, and $\lambda = 1$. If the action is not specified, it means that all the actions give the same reward and lead to the same state.

$$\leq \lambda^4 \left(\sum_{t=0}^{T-1} 4t^2 \gamma^t \right)^2 \mathbb{E}_{Z_c} \left[\frac{(1-\gamma^T)^2}{(1-\gamma)^2} \left(\frac{1-\gamma}{1-\gamma^T} \sum_{t=0}^{T-1} \gamma^t |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))|^2 \right)^2 \right] \ .$$

For the first sum, using that for any $x \in [0,1[$, $\sum_{k=0}^{\infty} k^2 x^k \leq 2/(1-x)^3$, and for the second sum using Jensen's inequality gives

$$\begin{aligned}(\mathbf{M_2}) &\leq \lambda^4 \frac{64}{(1-\gamma)^6} \mathbb{E}_{Z_c} \left[\frac{1-\gamma^T}{1-\gamma} \sum_{t=0}^{T-1} \gamma^t |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))|^4 \right] \\ &\leq \lambda^4 \frac{64}{(1-\gamma)^7} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{Z_c} \left[|\log(\pi_\theta(A_{c,1}^t \mid S_{c,1}^t))|^4 \right] \ . \end{aligned} \tag{57}$$

Denote by $\mathcal{P}(\mathcal{A})$ the set of probability distributions on $\mathcal{A}$. Note that for any policy. Note, that, we have

$$\max_{\pi \in \mathcal{P}(\mathcal{A})} - \sum_{a \in \mathcal{A}} \pi(a \mid s) \log(\pi(a \mid s))^4 = (\log(|\mathcal{A}|))^4 \ . \tag{58}$$

Plugging in the previous bound in (57) yields

$$(\mathbf{M_2}) \leq \frac{64\lambda^4}{(1-\gamma)^8} \log(|\mathcal{A}|)^4 \ .$$

Combining the bounds on $(\mathbf{M_1})$ and $(\mathbf{M_2})$ gives the following bound on $(\mathbf{M})$.

$$(\mathbf{M}) \leq 16(\mathbf{M_1}) + 16(\mathbf{M_2}) \leq \frac{256}{(1-\gamma)^8} + \frac{1024\lambda^4}{(1-\gamma)^8} \log(|\mathcal{A}|)^4 \ .$$

Plugging in the previous bound in (56) concludes the proof. □

We first show that, in general, the objective J_{r} does not have Łojasiewicz structure.

Lemma E.9. *There exists an FRL instance where the objective J_{r} admits a strict local minima.*

Proof. Consider the FRL task defined in Figure 5. Define $x := \pi_\theta(a_1|s_0)$. From the flow conservation constraints for occupancy measures for any agent $c \in [M]$, it holds that

$$\begin{aligned} d_c^{\rho,\theta}(s_2) &= \gamma d_c^{\rho,\theta}(s_1) \ , \quad d_c^{\rho,\theta}(s_1) = \gamma \left(q_c x + (1-x) p_c \right) d_c^{\rho,\theta}(s_1) \\ d_k^{\rho,\theta}(s_0) &= (1-\gamma) + \gamma \left((1-q_c)x + (1-p_c)(1-x) \right) d_c^{\rho,\theta}(s_0) \ . \end{aligned}$$

Rearranging the precedent terms yields

$$d_c^{\rho,\theta}(s_0) := \frac{1-\gamma}{1 - \gamma \left((1-q_c)x + (1-p_c)(1-x) \right)} \ ,$$

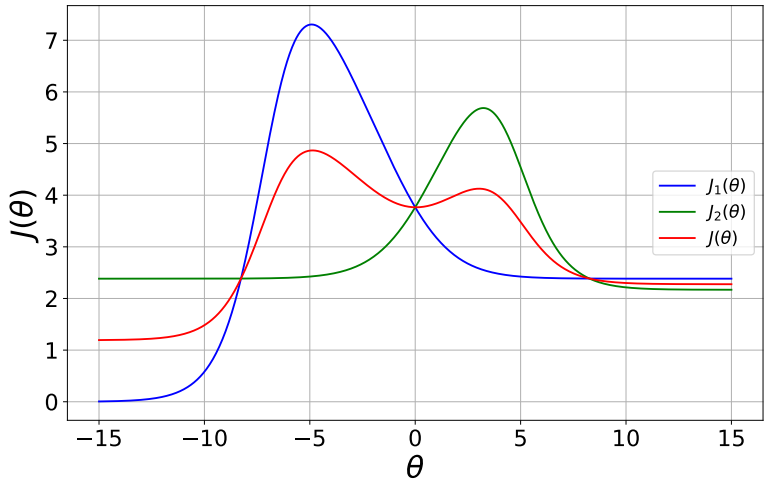


Figure 6: An example that shows that FRL objective J_{r} does not necessarily have a Łojasiewicz structure.

which implies

$$d_c^{\rho,\theta}(s_1) := \frac{\gamma(1-\gamma)\left(q_c x + (1-x)p_c\right)}{1-\gamma\left((1-q_c)x + (1-p_c)(1-x)\right)} = \frac{\gamma(1-\gamma)\left(p_c + (q_c-p_c)x\right)}{1-\gamma\left(1-p_c+x(p_c-q_c)\right)} \ .$$

The value of the objective function is thus

$$J_{\mathrm{r},c}(\theta) = \frac{\lambda}{1-\gamma}\left[d_c^{\rho,\theta}(s_0)H(x) + \frac{d_c^{\rho,\theta}(s_1)}{\lambda} + d_c^{\rho,\theta}(s_1)H(\pi_\theta(a_1|s_1)) + d_c^{\rho,\theta}(s_2)H(\pi_\theta(a_1|s_2))\right] \ ,$$

where for any $y \in (0,1), H(y) \overset{\Delta}{=} -y\log y - (1-y)\log(1-y)$. Now, let us assume that the policy for states s_1 and s_2 is uniform since it is an optimal solution given any values of p_c and q_c, and in this case we have $H(\pi_\theta(a_1|s_1)) = H(\pi_\theta(a_1|s_2)) = \log 2$. Then, let us define a value $f(x;p_c,q_c) = p_c + (q_c - p_c)x$, where $x = \sigma(\theta)$ for $\sigma(\theta) = \frac{1}{1+\exp(-\theta)}$ is a sigmoid parametrization.

Thus, after plugging in a value of our occupancy measures in our MDP, we have

$$J_{\mathrm{r},c}(\theta) = \frac{\tau H(\sigma(\theta)) + \gamma \cdot f(\sigma(\theta);p_c,q_c)\cdot(1+\tau\log 2 + \gamma\tau\log 2)}{1-\gamma(1-f(\sigma(\theta);p_c,q_c))}.$$

The plot of J_{r} (for $M=2$, $p_1=0$, $q_1=1$, $p_2=0.99$, $q_2=0.01$, $\gamma=0.999$, and $\lambda=1$) in Figure 6 shows that this problem does not have a Łojasiewicz structure. □

However, each agent locally satisfies a Łojasiewicz-type property

Lemma E.10 ((Equation 539) from Lemma 15 of Mei et al. (2020))**.** *For any agent $c \in [M]$, denote by $\pi_\lambda^{\star,c}$ the unique optimal regularized policy (see e.g Nachum et al. (2017) for the proof of existence and unicity) of this agent. It holds*

$$\|\nabla J_{\mathrm{r},c}(\theta)\|_2^2 \geq 2\mu_{\mathrm{r},c}^{\lambda}(\theta)\left[J_{\mathrm{r},c}^{\star} - J_{\mathrm{r},c}(\theta)\right] \ ,$$

where $\mu_{\mathrm{r},c}^{\lambda}(\theta)$ is defined as

$$\mu_{\mathrm{r},c}^{\lambda}(\theta) = \frac{\lambda \min_s d_c^{\rho,\pi_\theta}(s)\min_{s,a}\pi_\theta(a|s)^2}{|\mathcal{S}|(1-\gamma)}\left\|\frac{d_c^{\rho,\pi_\lambda^{\star,c}}}{d_c^{\rho,\theta}}\right\|_\infty^{-1} \ .$$

In case of small heterogeneity between the agents, i.e **A**-1, this precedent Lemma, combined with Lemma D.6 allows us to establish that the global objective satisfies a Relaxed Łojasiewicz-type inequality.

Lemma E.11. *Assume **A**-1. For any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\zeta_{\mathrm{r}}^2 + \|\nabla J_{\mathrm{r}}(\theta)\|_2^2 \geq 2 \min_{c\in[M]} \mu_{\mathrm{r},c}^{\lambda}(\theta)(J_{\mathrm{r}}^{\star} - J_{\mathrm{r}}(\theta)) \ .$$

Proof. Let $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Using Lemma E.10 and the triangle inequality, we have for any $c \in [M]$

$$\begin{aligned}
&\min_{c\in[M]} 2\mu_{\mathrm{r},c}^{\lambda}(\theta)[J_{\mathrm{r},c}^{\star} - J_{\mathrm{r},c}(\theta)] \\
&\quad\leq 2\mu_{\mathrm{r},c}^{\lambda}(\theta)[J_{\mathrm{r},c}^{\star} - J_{\mathrm{r},c}(\theta)] \leq \|\nabla J_{\mathrm{r},c}(\theta)\|_2^2 \\
&\quad= \|\nabla J_{\mathrm{r},c}(\theta) - \nabla J_{\mathrm{r}}(\theta) + \nabla J_{\mathrm{r}}(\theta)\|_2^2 \\
&\quad\leq \|\nabla J_{\mathrm{r},c}(\theta) - \nabla J_{\mathrm{r}}(\theta)\|_2^2 + 2\langle \nabla J_{\mathrm{r},c}(\theta) - \nabla J_{\mathrm{r}}(\theta), \nabla J_{\mathrm{r},}(\theta)\rangle + \|\nabla J_{\mathrm{r}}(\theta)\|_2^2 \\
&\quad\leq \zeta_{\mathrm{r}} + 2\langle \nabla J_{\mathrm{r},c}(\theta) - \nabla J_{\mathrm{r}}(\theta), \nabla J_{\mathrm{r}}(\theta)\rangle + \|\nabla J_{\mathrm{r}}(\theta)\|_2^2 \ ,
\end{aligned}$$

where in the last inequality we used Lemma E.6. Finally, averaging the preceding inequality over all the agents concludes the proof. □

E.2 Convergence rates, sample and communication complexities

Theorem E.12 (Convergence rates of `RS-FedPG`)**.** *Assume **A**-1 and **A**-2 and no projection (i.e., set $\mathcal{T} = \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$). Additionally, assume that there exists $\mu_{\mathrm{r}}^{\lambda} \in (0,1)$ such that, with probability 1, $\inf_{r\in\mathbb{N}} \mu_{\mathrm{r}}^{\lambda}(\theta^r) \geq \mu_{\mathrm{r}}^{\lambda} > 0$. For any $\eta > 0$ such that $\eta H L_{2,\mathrm{r}} \leq 1/74$, the iterates of `RS-FedPG` satisfy*

$$\begin{aligned}
J_{\mathrm{r}}^{\star} - \mathbb{E}[J_{\mathrm{r}}(\theta^R)] \leq \left(1 - \frac{\eta H \mu_{\mathrm{r}}^{\lambda}}{2}\right)^R (J_{\mathrm{r}}^{\star} - J_{\mathrm{r}}(\theta^0) + \frac{3\eta L_{2,\mathrm{r}}\sigma_{\mathrm{r},2}^2}{M\mu_{\mathrm{r}}^{\lambda}} + \frac{\zeta_{\mathrm{r}}^2}{\mu_{\mathrm{r}}^{\lambda}} \\
+ 4\frac{\beta_{\mathrm{r}}^2}{\mu_{\mathrm{r}}^{\lambda}} + \frac{8\cdot 12^3\eta^4 L_{3,\mathrm{r}}^2 H(H-1)\sigma_{\mathrm{r},4}^4}{\mu_{\mathrm{r}}^{\lambda}} \ .
\end{aligned}$$

Proof. First note that no projection is used, which implies that $\bar{\theta}^{r+1} = \theta^{r+1}$. Under **A**-1 and **A**-2,all the results of Appendix E.1 hold, satisfying thereby the assumptions of Lemma C.3. Importantly note that if $\eta H L_{2,\mathrm{r}} \leq 1/74$ then it holds that $32\eta^2 H^2 L_{3,\mathrm{r}}^2 L_{1,\mathrm{r}}^2 \leq L_{2,\mathrm{r}}^2$ (as $32L_{3,\mathrm{r}}^2 L_{1,\mathrm{r}}^2 \leq 74^2 L_{2,\mathrm{r}}^4$ by Lemma E.3, Lemma E.4, and Lemma E.5). Applying Lemma C.3 yields

$$\begin{aligned}
-\mathbb{E}\left[J_{\mathrm{r}}(\theta^{r+1})\middle|\mathcal{F}^r\right] \leq -J_{\mathrm{r}}(\theta^r) - \frac{\eta H}{4}\|\nabla J_{\mathrm{r}}(\theta^r)\|_2^2 + \frac{3\eta^2 L_{2,\mathrm{r}} H\sigma_{\mathrm{r},2}^2}{2M} \\
+ 2\eta H\beta_{\mathrm{r}}^2 + 8\eta^3 L_{2,\mathrm{r}}^2 H^2(H-1)\zeta_{\mathrm{r}}^2 + 4\cdot 12^3\eta^5 L_{3,\mathrm{r}}^2 H^2(H-1)\sigma_{\mathrm{r},4}^4
\end{aligned} \tag{59}$$

Adding $J_{\mathrm{r}}^{\star}$, using Lemma E.11, and using the fact that $\eta H L_{2,\mathrm{r}} \leq 1/74$ to simplify the heterogeneity terms yields

$$\begin{aligned}
J_{\mathrm{r}}^{\star} - \mathbb{E}\left[J_{\mathrm{r}}(\theta^{r+1})\middle|\mathcal{F}^r\right] \leq (1 - \frac{\eta H\mu_{\mathrm{r}}^{\lambda}}{2})(J_{\mathrm{r}}^{\star} - J_{\mathrm{r}}(\theta)) + \frac{\eta H}{2}\zeta_{\mathrm{r}}^2 + \frac{3\eta^2 L_{2,\mathrm{r}} H\sigma_{\mathrm{r},2}^2}{2M} \\
+ 2\eta H\beta_{\mathrm{r}}^2 + 4\cdot 12^3\eta^5 L_{3,\mathrm{r}}^2 H^2(H-1)\sigma_{\mathrm{r},4}^4
\end{aligned} \tag{60}$$

The result follows from taking the expectation and unrolling the recursion. □

Recall that

$$\begin{aligned}
L_{1,\mathrm{r}} = \frac{1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^2} \ , \ L_{2,\mathrm{r}} = \frac{8+\lambda(4+8\log(|\mathcal{A}|)}{(1-\gamma)^3} \ , \ L_{3,\mathrm{sm}} = \frac{480+832\lambda\log|\mathcal{A}|}{(1-\gamma)^4} \ , \\
\zeta_{\mathrm{r}}^2 = \frac{38(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} \ , \ \beta_{\mathrm{r}} = \frac{2(1+\lambda\log(|\mathcal{A}|))\gamma^T T}{1-\gamma} + \frac{2(1+\lambda\log(|\mathcal{A}|))\gamma^T}{(1-\gamma)^2} \ , \\
\sigma_{\mathrm{r},2}^2 = \frac{12+24\lambda^2(\log(|\mathcal{A}|))^2}{(1-\gamma)^4 B} \ , \ \sigma_{\mathrm{r},4}^4 = \frac{(1120+4480\lambda^4\log(|\mathcal{A}|)^4)}{(1-\gamma)^8 B^2} \ ,
\end{aligned}$$

which are defined respectively in Lemmas E.2 and E.4 to E.8. We obtain the following result.

Corollary E.13 (Explicit Convergence Rate of RS-FedPG)**.** *Under the assumptions of Theorem E.12, let $\eta > 0$ such that $\eta H \leq 888^{-1}(1-\gamma)^3(1+\lambda\log(|\mathcal{A}|))^{-1}$, and $T \geq 1/(1-\gamma)$. Then, the iterates of RS-FedPG satisfy*

$$J_{\mathrm{r}}^\star - \mathbb{E}[J_{\mathrm{r}}(\theta^R)] \leq \left(1 - \frac{\eta H \mu_{\mathrm{r}}^\lambda}{2}\right)^R (J_{\mathrm{r}}^\star - J_{\mathrm{r}}(\theta^0) + \frac{864\eta(1+\lambda\log(|\mathcal{A}|)^3}{BM\mu_{\mathrm{r}}^\lambda(1-\gamma)^7} + \frac{38(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{\mu_{\mathrm{r}}^\lambda(1-\gamma)^6} + \frac{16(1+\lambda\log(|\mathcal{A}|))^2\gamma^{2T}T^2}{\mu_{\mathrm{r}}^\lambda(1-\gamma)^2} + \frac{51^8\eta^4 H(H-1)(1+\lambda\log(|\mathcal{A}|))^6}{\mu_{\mathrm{r}}^\lambda B^2(1-\gamma)^{16}} \ .$$

Corollary E.14 (Sample and Communication Complexity of RS-FedPG)**.** *Under the assumptions of Theorem E.12, let $\epsilon \geq 190(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2(\mu_{\mathrm{r}}^\lambda)^{-1}(1-\gamma)^{-6}$. Then, for a properly chosen truncation horizon, a properly chosen step size and number of local updates, RS-FedPG learns an ϵ-approximation of the optimal objective with a number of communication rounds*

$$R \geq \frac{888(1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^3\mu_{\mathrm{r}}^\lambda}\log\Big(\frac{5(J_{\mathrm{r}}^\star - J_{\mathrm{r}}(\theta^0))}{\epsilon}\Big) \ ,$$

for a total number of samples per agent of

$$RHB \geq \max\Big(\frac{24(1+\lambda\log(|\mathcal{A}|))B}{\mu_{\mathrm{r}}^\lambda(1-\gamma)^3}, \frac{8640(1+\lambda\log(|\mathcal{A}|))^3}{(\mu_{\mathrm{r}}^\lambda)^2\epsilon M(1-\gamma)^7}, \frac{2\cdot 12^4(1+\lambda\log(|\mathcal{A}|))^2}{\epsilon^{1/2}(\mu_{\mathrm{r}}^\lambda)^{3/2}(1-\gamma)^5}\Big)\log\Big(\frac{5(J_{\mathrm{r}}^\star - J_{\mathrm{r}}(\theta^0))}{\epsilon}\Big) \ .$$

Proof. First, we fix the truncation horizon T to

$$T \geq (1-\gamma)^{-1}\max\left(\frac{4}{1-\gamma}, \log\left(\frac{80(1+\lambda\log(|\mathcal{A}|))^2}{\epsilon\mu_{\mathrm{r}}^\lambda(1-\gamma)^2}\right)\right) \ .$$

Secondly, we require (i) that $\eta \leq L_{2,\mathrm{r}} \leq 12^{-1}(1-\gamma)^3(1+\lambda\log(|\mathcal{A}|))^{-1}$, and that (ii) each variance terms to be smaller than $\epsilon/5$, which gives the condition on the step size

$$\eta \leq \min\Big(\frac{(1-\gamma)^3}{12(1+\lambda\log(|\mathcal{A}|))}, \frac{\mu_{\mathrm{r}}^\lambda\epsilon MB(1-\gamma)^7}{4320(1+\lambda\log(|\mathcal{A}|))^3}, \frac{B\epsilon^{1/2}(\mu_{\mathrm{r}}^\lambda)^{1/2}(1-\gamma)^5}{12^4(1+\lambda\log(|\mathcal{A}|))^2}\Big) \ , \tag{61}$$

where the third element of the $\min$ comes from

$$\frac{51^8\eta^4 H(H-1)(1+\lambda\log(|\mathcal{A}|))^6}{\mu_{\mathrm{r}}^\lambda B^2(1-\gamma)^{16}} \leq \frac{51^8\eta^2(1+\lambda\log(|\mathcal{A}|))^2}{888^2\mu_{\mathrm{r}}^\lambda B^2(1-\gamma)^{10}} \leq \frac{\epsilon}{5} \ .$$

Then, H has to satisfy $\eta H \leq 888^{-1}(1-\gamma)^3(1+\lambda\log(|\mathcal{A}|))^{-1}$. This requires

$$H \leq \frac{1}{888\eta(1+\lambda\log(|\mathcal{A}|))^{-1}} \ .$$

Finally, we require that the number of communications is at least

$$R \geq \frac{2}{\eta H\mu_{\mathrm{r}}^\lambda}\log\Big(\frac{5(J_{\mathrm{r}}^\star - J_{\mathrm{r}}(\theta^0))}{\epsilon}\Big) = \frac{888(1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^3\mu_{\mathrm{r}}^\lambda}\log\Big(\frac{5(J_{\mathrm{r}}^\star - J_{\mathrm{r}}(\theta^0))}{\epsilon}\Big) \ .$$

The sample complexity follows from $RHB \geq \frac{2B}{\eta\mu_{\mathrm{r}}^\lambda}\log\Big(\frac{5(J_{\mathrm{r}}^\star - J_{\mathrm{r}}(\theta^0))}{\epsilon}\Big)$ and (61). □

E.3 Establishing a bound on μ_{r}^λ when $|\mathcal{A}| = 2$

The goal of this subsection is to show that when $|\mathcal{A}| = 2$, the field $\nabla J_{\mathrm{r},c}$ presents a particular structure (which does not hold for larger action spaces) that allows us to show the existence of a bounded set on which projecting the global iterates of RS-FedPG provably increases the value of the objective J_{r}. This property is crucial, as by projecting at every round the parameter on this bounded set, we constrain the sequence of policies to remain uniformly away from the boundary of the simplex, while guaranteeing that we do not degrade the value of the objective J_{r}. This allow us to derive an explicit lower bound on $\inf_{r\in[R]}\mu_{\mathrm{r}}^\lambda(\theta^r)$ by using Lemma E.10.

Define $B(\lambda)$ as the ℓ_∞ ball of radius $R(\lambda) \triangleq (1 + \lambda \log(2))/(\lambda(1-\gamma))$, and let $\text{proj}_{B(\lambda)}(\cdot)$ denote the ℓ_2-projection operator onto this ball. We also denote by $\bar{B}(\lambda)$ the complement of $B(\lambda)$ in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. From now on and until the end of the subsection, unless the inverse is explicitly mentioned, we fix the projection set $\mathcal{T}$ of `RS-FedPG` to $B(\lambda)$, the initial parameter to (0) and assume that the action space consists of exactly two actions, i.e., $|\mathcal{A}| = 2$. We denote these actions by a_1 and a_2.

To prove that projecting on $B(\lambda)$ increases the objective value, we first prove a property on the trajectory of the iterates of `RS-FedPG`. Define the vector $(1) \in \mathbb{R}^{|\mathcal{A}|}$ as the all-ones vector and define the set

$$P \triangleq \{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \forall s \in \mathcal{S}, \langle \theta(s,\cdot), (1)\rangle = 0\} \ . \tag{62}$$

We begin by proving that the iterates of the `RS-FedPG` algorithm almost surely remain within P.

Lemma E.15. *Let $\theta \in P$ and define the projected parameter $\theta' = \text{proj}_{B(\lambda)}(\theta)$. It holds that*

$$\theta'(s,\cdot) = \kappa(s)\theta(s,\cdot) \ \text{, where } \kappa(s) := \min\left(\frac{R(\lambda)}{|\theta(s,a_1)|}, 1\right).$$

In particular, we have $\theta' \in P$.

Proof. Fix $s \in \mathcal{S}$. We distinguish two cases depending on the value of $\theta(s,a_1)$.

Case 1. If $|\theta(s,a_1)| \leq R(\lambda)$, then by the assumption $\theta(s,a_1) + \theta(s,a_2) = 0$, it follows that $|\theta(s,a_2)| = |\theta(s,a_1)| \leq R(\lambda)$. Thus, no truncation occurs under the ℓ_∞ projection, and we have $\theta'(s,\cdot) = \theta(s,\cdot)$. Therefore, $\theta'(s,\cdot) = \kappa(s)\theta(s,\cdot)$ with $\kappa(s) = 1$.

Case 2. Suppose now that $|\theta(s,a_1)| > R(\lambda)$. Without loss of generality, assume $\theta(s,a_1) > R(\lambda)$ (the case $\theta(s,a_1) < -R(\lambda)$ is symmetric). Since $\theta(s,a_2) = -\theta(s,a_1)$, it follows that $\theta(s,a_2) < -R(\lambda)$. The projection onto the ℓ_∞ ball then truncates both components to the bounds $\pm R(\lambda)$, yielding

$$\theta'(s,a_1) = R(\lambda) \ , \quad \theta'(s,a_2) = -R(\lambda) \ .$$

This implies

$$\theta'(s,\cdot) = \frac{R(\lambda)}{|\theta(s,a_1)|}\theta(s,\cdot) \ ,$$

so again we have $\theta'(s,\cdot) = \kappa(s)\theta(s,\cdot)$ with $\kappa(s) = \frac{R(\lambda)}{|\theta(s,a_1)|} < 1$.

In both cases, the projection preserves the antisymmetry of the original vector: $\theta'(s,a_1) + \theta'(s,a_2) = 0$. Equivalently, $\theta'(s,\cdot) = \kappa(s)\theta(s)$, as desired. □

Lemma E.16. *Consider the iterates $(\theta_c^{r,h})$ generated by `RS-FedPG`. For all $r \in [R]$, we have $\theta^r \in P$.*

Proof. First, observe that P is convex and thus stable by convex combinations. Thereby, if for a given $r \in [R]$, it holds that for all $c \in [M]$, we have that $\theta_c^{r,H} \in P$ then $\bar{\theta}^{r+1} \in P$ and thereby by Lemma E.15, we have $\theta^{r+1} = \text{proj}_{B(\lambda)}(\theta) \in P$.

In particular, if we prove that $\theta^r \in P$ implies that $\theta_c^{r,H} \in P$, then an immediate recursion on r would prove the desired result. Subsequently, we prove by induction that if for some $r \in [R]$, we have $\theta^r \in P$ then for any $h \in [H]$, and for any $c \in [M]$ we have $\theta_c^{r,h} \in P$. Let us fix $r \in [R]$ such that $\theta^r \in P$.

Base case: As we assume $\theta^r \in P$, then by definition it holds for all $c \in [M]$ that $\theta_c^{r,0} \in P$.

Recursion: Let $h \in [H-1]$ such that for all $c \in [M], \theta_c^{r,h} \in P$. Fix $c \in [M], s \in \mathcal{S}$, and recall the definition of the estimator of the gradient (48):

$$\mathrm{g}_{\mathrm{r},c}^{Z_c} := \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t\left(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(A_{c,b}^\ell \mid S_{c,b}^\ell)\right)\left[\mathrm{r}(S_{c,b}^t, A_{c,b}^t) - \lambda\log(\pi_\theta(A_{c,b}^t, S_{c,b}^t))\right] \ .$$

Importantly, using that for any $(s', a') \in \mathcal{S} \times \mathcal{A}$, we have $\sum_{a\in\mathcal{A}} \frac{\partial \log(\pi_\theta(a'|s'))}{\partial\theta(s,a)} = 0$, we get

$$\sum_{a\in\mathcal{A}} \mathrm{g}_{\mathrm{r},c}^{Z_c}(s,a) \tag{63}$$
$$= \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t \left(\sum_{\ell=0}^{t}\sum_{a\in\mathcal{A}} \frac{\partial \log \pi_\theta(A_{c,b}^{\ell} \mid S_{c,b}^{\ell})}{\partial\theta(s,a)}\right)\left[\mathsf{r}(S_{c,b}^t, A_{c,b}^t) - \lambda \log(\pi_\theta(A_{c,b}^t, S_{c,b}^t))\right] = 0.$$

Combining the previous result and the induction hypothesis implies $\langle \theta_c^{r,h+1}(s,\cdot),(1)\rangle = 0$ validating thereby the recursion. □

Next, we establish for any $c \in [M]$ that the field $\nabla J_{\mathrm{r},c}$ satisfies a radiality-type property in $\bar{B}(\lambda)\cap P$.

Lemma E.17. *For any $c \in [M]$, the gradient field is radial in $\bar{B}(\lambda)\cap P$, i.e for any $\theta \in \bar{B}(\lambda)\cap P$, for any $s \in \mathcal{S}$ such that $|\theta(s,a_1)| > R(\lambda)$, we have*

$$\langle \theta(s,\cdot), \frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,\cdot)}\rangle \le 0 \ .$$

Proof. According to the definition of $J_{\mathrm{r},c}(\theta)$, we have

$$J_{\mathrm{r},c}(\theta) = \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\left[\widetilde{Q}_c^\theta(s,a) - \lambda\log(\pi_\theta(a|s))\right]\right] \ .$$

Taking the derivative w.r.t θ,

$$\begin{aligned}\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta} &= \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \frac{\partial\pi_\theta(a|s)}{\partial\theta}\left[\widetilde{Q}_c^\theta(s,a) - \lambda\log(\pi_\theta(a|s))\right]\right] \\ &\quad + \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\left[\frac{\partial\widetilde{Q}_c^\theta(s,a)}{\partial\theta} - \lambda\frac{1}{\pi_\theta(a|s)}\frac{\partial\pi_\theta(a|s)}{\partial\theta}\right]\right] \\ &= \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \frac{\partial\pi_\theta(a|s)}{\partial\theta}\left[\widetilde{Q}_c^\theta(s,a) - \lambda\log(\pi_\theta(a|s))\right]\right] + \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\frac{\partial\widetilde{Q}_c^\theta(s,a)}{\partial\theta}\right] \ ,\end{aligned}$$

where in the last equality we used that $\pi_\theta(\cdot|s)$ lies on the simplex. Now using (50), we obtain

$$\begin{aligned}\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta} &= \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \frac{\partial\pi_\theta(a|s)}{\partial\theta}\left[\widetilde{Q}_c^\theta(s,a) - \lambda\log(\pi_\theta(a|s))\right]\right] \\ &\quad + \gamma\mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\sum_{s'\in\mathcal{S}} \mathsf{P}_c(s'|s,a)\frac{\partial\widetilde{V}_c^\theta(s)}{\partial\theta}\right] \ .\end{aligned}$$

Unrolling the recursion yields

$$\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta} = \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d_c^{\rho,\theta}(s)\sum_{a\in\mathcal{A}} \frac{\partial\pi_\theta(a|s)}{\partial\theta}\left[\widetilde{Q}_c^\theta(s,a) - \lambda\log(\pi_\theta(a|s))\right] \ . \tag{64}$$

Now fix $s \in \mathcal{S}$. It holds that

$$\begin{aligned}\frac{\partial\pi_\theta(a_1|s)}{\partial\theta(s,a_1)} &= \pi_\theta(a_1|s)\pi_\theta(a_2|s) \ , \quad \frac{\partial\pi_\theta(a_2|s_1)}{\partial\theta(s,a_1)} = -\pi_\theta(a_1|s)\pi_\theta(a_2|s) \ , \\ \frac{\partial\pi_\theta(a_1|s)}{\partial\theta(s,a_2)} &= -\pi_\theta(a_1|s)\pi_\theta(a_2|s) \ , \quad \frac{\partial\pi_\theta(a_2|s_1)}{\partial\theta(s,a_1)} = \pi_\theta(a_1|s)\pi_\theta(a_2|s) \ .\end{aligned}$$

Note that for $s' \neq s$ and $a, a' \in \mathcal{A}$, we have $\frac{\partial\pi_\theta(a|s)}{\partial\theta(s',a')} = 0$. Thus, (64) simplifies to

$$\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,a_1)} = \frac{1}{1-\gamma}\cdot d_c^{\rho,\theta}(s)\cdot\left[\frac{\partial\pi_\theta(a_1|s)}{\partial\theta(s,a_1)}\cdot\left[\widetilde{Q}_c^\theta(s,a_1) - \lambda\log(\pi_\theta(a_1|s))\right]\right]$$

$$+\frac{1}{1-\gamma}\cdot d_c^{\rho,\theta}(s)\cdot\left[\frac{\partial\pi_\theta(a_2|s)}{\partial\theta(s,a_1)}\cdot\left[\widetilde{Q}_c^\theta(s,a_2)-\lambda\log(\pi_\theta(a_2|s))\right]\right]$$
$$=\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_2|s)\left[\widetilde{Q}_c^\theta(s,a_1)-\widetilde{Q}_c^\theta(s,a_2)-\lambda(\theta(a_1|s)-\theta(a_2|s))\right]\ .$$

Similarly, computing the derivative with respect to $\theta(s,a_2)$ yields

$$\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,a_2)}=\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_1|s)\left[\widetilde{Q}_c^\theta(s,a_2)-\widetilde{Q}_c^\theta(s,a_1)-\lambda(\theta(a_2|s)-\theta(a_1|s))\right]\ .$$

Now using that $\theta(s,a_1)+\theta(s,a_2)=0$, we get

$$\langle\theta(s,\cdot),\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,\cdot)}\rangle=\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_1|s)\left[2\theta(s,a_1)\left(\widetilde{Q}_c^\theta(s,a_2)-\widetilde{Q}_c^\theta(s,a_1)\right)\right]$$
$$-2\lambda\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_1|s)\theta(s,a_1)^2\ .$$

Finally using that for any $(s,a,c)\in\mathcal{S}\times\mathcal{A}\times[M]$, $\widetilde{Q}_c^\theta(s,a)\le\frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}$ concludes the proof. □

Remark E.18. The precedent property is not satisfied when the number of actions is strictly larger than 2.

Proof. In what follows, we consider the FRL instance where all the agents share the same following MDP. This MDP has four states $\mathcal{S}=\{s_0,s_1,s_2,s_3\}$ and three actions $\mathcal{A}=\{a_1,a_2,a_3\}$. If the agent is at state s_0 and picks action a_i, then he moves deterministically to the state s_i. In any other state, the agent stays in the same state, whatever action he takes. Additionally, we set $\mathsf{r}(s_0,a_1)=1$ and zero elsewhere. We set the initial distribution to be the uniform distribution on the state space.

Fix $s\in\mathcal{S}$. It holds that

$$\frac{\partial\pi_\theta(a_1|s)}{\partial\theta(s,a_1)}=\pi_\theta(a_1|s)\left(\pi_\theta(a_2|s)+\pi_\theta(a_3|s)\right)\ ,\quad\frac{\partial\pi_\theta(a_2|s_1)}{\partial\theta(s,a_1)}=-\pi_\theta(a_1|s)\pi_\theta(a_2|s)\ ,$$
$$\frac{\partial\pi_\theta(a_3|s)}{\partial\theta(s,a_1)}=-\pi_\theta(a_1|s)\pi_\theta(a_3|s)\ ,$$

with similar expressions for the derivative with respect to $\theta(s,a_2)$ and $\theta(s,a_3)$. Starting from (64) and rearranging the terms, we get

$$\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,a_1)}=\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\sum_{a\in\mathcal{A}}\frac{\partial\pi_\theta(a|s)}{\partial\theta(s,a_1)}\left[\widetilde{Q}_c^\theta(s,a)-\lambda\log(\pi_\theta(a|s))\right]$$
$$=\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_2|s)\left[\widetilde{Q}_c^\theta(s,a_1)-\widetilde{Q}_c^\theta(s,a_2)-\lambda\left(\theta(s,a_1)-\theta(s,a_2)\right)\right]$$
$$+\frac{1}{1-\gamma}d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_3|s)\left[\widetilde{Q}_c^\theta(s,a_1)-\widetilde{Q}_c^\theta(s,a_3)-\lambda\left(\theta(s,a_1)-\theta(s,a_3)\right)\right]\ ,$$

with similar expressions for the partial derivative of $J_{\mathrm{r},c}(\theta)$ with respect to $\theta(s,a_2)$ and $\theta(s,a_3)$. Thus, we get

$$(1-\gamma)\langle\theta(s,\cdot),\frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,\cdot)}\rangle$$
$$=d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_2|s)\left[(\widetilde{Q}_c^\theta(s,a_1)-\widetilde{Q}_c^\theta(s,a_2))\theta(s,a_1)-\lambda\theta(s,a_1)\left(\theta(s,a_1)-\theta(s,a_2)\right)\right]$$
$$+d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_3|s)\left[(\widetilde{Q}_c^\theta(s,a_1)-\widetilde{Q}_c^\theta(s,a_3))\theta(s,a_1)-\lambda\theta(s,a_1)\left(\theta(s,a_1)-\theta(s,a_3)\right)\right]$$
$$+d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_1|s)\left[(\widetilde{Q}_c^\theta(s,a_2)-\widetilde{Q}_c^\theta(s,a_1))\theta(s,a_2)-\lambda\theta(s,a_2)\left(\theta(s,a_2)-\theta(s,a_1)\right)\right]$$
$$+d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_3|s)\left[(\widetilde{Q}_c^\theta(s,a_2)-\widetilde{Q}_c^\theta(s,a_3))\theta(s,a_2)-\lambda\theta(s,a_2)\left(\theta(s,a_2)-\theta(s,a_3)\right)\right]$$
$$+d_c^{\rho,\theta}(s)\pi_\theta(a_3|s)\pi_\theta(a_1|s)\left[(\widetilde{Q}_c^\theta(s,a_3)-\widetilde{Q}_c^\theta(s,a_1))\theta(s,a_3)-\lambda\theta(s,a_3)\left(\theta(s,a_3)-\theta(s,a_1)\right)\right]$$

$$+ d_c^{\rho,\theta}(s)\pi_\theta(a_3|s)\pi_\theta(a_2|s)\left[(\widetilde{Q}_c^\theta(s,a_3) - \widetilde{Q}_c^\theta(s,a_2))\theta(s,a_3) - \lambda\theta(s,a_3)\left(\theta(s,a_3) - \theta(s,a_2)\right)\right] .$$

Rearranging the terms gives

$$\begin{aligned}(1-\gamma)\langle\theta(s,\cdot), \frac{\partial J_{\mathrm{r},c}(\theta)}{\partial\theta(s,\cdot)}\rangle &= d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_2|s)(\widetilde{Q}_c^\theta(s,a_1) - \widetilde{Q}_c^\theta(s,a_2))(\theta(s,a_1) - \theta(s,a_2)) \\ &\quad - \lambda d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_2|s)\left(\theta(s,a_1) - \theta(s,a_2)\right)^2 \\ &\quad + d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_3|s)(\widetilde{Q}_c^\theta(s,a_1) - \widetilde{Q}_c^\theta(s,a_3))(\theta(s,a_1) - \theta(s,a_3)) \\ &\quad - \lambda d_c^{\rho,\theta}(s)\pi_\theta(a_1|s)\pi_\theta(a_3|s)\left(\theta(s,a_1) - \theta(s,a_3)\right)^2 \\ &\quad + d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_3|s)(\widetilde{Q}_c^\theta(s,a_2) - \widetilde{Q}_c^\theta(s,a_3))(\theta(s,a_2) - \theta(s,a_3)) \\ &\quad - \lambda d_c^{\rho,\theta}(s)\pi_\theta(a_2|s)\pi_\theta(a_3|s)\left(\theta(s,a_2) - \theta(s,a_3)\right)^2 \quad .\end{aligned}$$

For $x > 0$, define $\theta_x \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ such that for any $i \in \{0,1,2,3\}$, we have: $\theta_x(s_i,a_1) = x + 1/x, \theta_x(s_i,a_2) = x$, and $\theta_x(s_i,a_3) = -2x - 1/x$. Note that there exists $b\colon \mathbb{R}\to\mathbb{R}$ and $b\colon \mathbb{R}\to\mathbb{R}$ such that $\widetilde{Q}_c^{\theta_x}(s_0,a_1) = 1 + b(x)$, $\widetilde{Q}_c^{\theta_x}(s_0,a_2) = c(x)$, and $\lim_{x\to\infty} b(x) = \lim_{x\to\infty} c(x) = 0$. In this case, all the terms that contain $\pi_{\theta_x}(a_3|s_0)$ are negligible for sufficiently large x. Thereby we have

$$(1-\gamma)\langle\theta_x(s_0,\cdot), \frac{\partial J_{\mathrm{r},c}(\theta_x)}{\partial\theta(s_0,\cdot)}\rangle \sim_{x\to\infty} d_c^{\rho,\theta_x}(s_0)\pi_{\theta_x}(a_1|s_0)\pi_{\theta_x}(a_2|s_0)\left[\frac{1 + b(x) - c(x) - \lambda/x}{x}\right] .$$

As the two equivalents must be of similar signs for sufficiently large x then we can conclude that there does not exist a ball on which the field is systematically radial outside of it. □

By combining Lemma E.16 and Lemma E.17, we show that, for any $c \in [M]$, projecting the iterates of `RS-FedPG` onto $B(\lambda)$ always results in increasing the objective $J_{\mathrm{r},c}$.

Lemma E.19. *Let $\theta \in P$ and define $\theta' = proj_{B(\lambda)}(\theta)$. It holds that*

$$J_{\mathrm{r},c}(\theta') \geq J_{\mathrm{r},c}(\theta) \quad .$$

Proof. We distinguish two cases:

Case 1. If $\theta \in B(\lambda)$, then $\theta' = \theta$ and thus the result follows.

Case 2. Now, we consider the case where $\theta \in \bar{B}(\lambda)$. In this case, there exists $s \in \mathcal{S}$ such that $|\theta(s,a_1)| \geq R(\lambda)$. Using Lemma E.15, we have $\theta'(s,\cdot) = \kappa(s)\theta(s,\cdot)$ where $\kappa(s) = R(\lambda)/|\theta(s,a_1)|$. Now, define the function

$$\begin{aligned} g\colon [0,1] &\to \mathbb{R} \\ t &\mapsto J_{\mathrm{r},c}((1-t)\theta + t\theta') \quad .\end{aligned}$$

We have $g'(t) = \langle\theta' - \theta, \nabla J_{\mathrm{r},c}((1-t)\theta + t\theta')\rangle$. Now using Taylor's expansion with integral rest, we get

$$J_{\mathrm{r},c}(\theta') = J_{\mathrm{r},c}(\theta) + \int_{t=0}^1 g'(t)dt = J_{\mathrm{r},c}(\theta) + \int_{t=0}^1 \langle\theta' - \theta, \nabla J_{\mathrm{r},c}((1-t)\theta + t\theta')\rangle dt \quad .$$

Defining $\mathcal{S}_+$ as the set of states where $|\theta(s,a_1)| \geq R(\lambda)$, we get by decomposing the scalar product

$$\begin{aligned} J_{\mathrm{r},c}(\theta') &= J_{\mathrm{r},c}(\theta) + \sum_{s\in\mathcal{S}_+}\int_{t=0}^1 \langle\theta'(s,\cdot) - \theta(s,\cdot), \frac{\partial J_{\mathrm{r},c}((1-t)\theta + t\theta')}{\partial\theta(s,\cdot)}\rangle dt \\ &= J_{\mathrm{r},}(\theta) + \sum_{s\in\mathcal{S}_+}\int_{t=0}^1 \frac{\kappa(s)-1}{1-t+t\kappa(s)}\langle(1-t)\theta(s,\cdot) + t\theta'(s,\cdot)), \frac{\partial J_{\mathrm{r},c}((1-t)\theta + t\theta')}{\partial\theta(s,\cdot)}\rangle dt\end{aligned}$$

As for any $t \in (0,1)$, we have $(1-t)\theta + t\theta' \in P \cap \bar{B}(\lambda)$, for any $s \in \mathcal{S}_+, 0 \leq \kappa(s)$ and $|(1-t)\theta(s,a_1) + t\theta'(s,a_1)| \geq R(\lambda)$, then applying Lemma E.17 proves the positivity of the integral term which completes the proof. □

Lemma E.20. *Assume **A**-2. It holds that*

$$\inf_{\theta\in B(\lambda)\cap P}\mu_{\mathrm{r}}^{\lambda}(\theta)\geq\frac{\lambda(1-\gamma)\min_s\rho(s)^2}{4|\mathcal{S}|}\mathrm{e}^{-4\cdot\frac{1+\lambda\log(2)}{\lambda(1-\gamma)}}\quad.$$

Proof. Using Lemma E.10, we have

$$\inf_{\theta\in B(\lambda)\cap P}\mu_{\mathrm{r}}^{\lambda}(\theta)=\min_{c\in[M]}\inf_{\theta\in B(\lambda)\cap P}\left\{\frac{\lambda}{|\mathcal{S}|}\frac{1}{1-\gamma}\min_s d_c^{\rho,\pi_\theta}(s)\cdot\min_{s,a}\pi_\theta(a|s)^2\cdot\left\|\frac{d_c^{\rho,\pi_\lambda^{\star,c}}}{d_c^{\rho,\theta}}\right\|_\infty^{-1}\right\}\quad,$$

where $\pi_\lambda^{\star,c}$ is the unique optimal regularized policy of agent c. Fix any $\theta\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $c\in[M]$. Under **A**-2, it holds that for any $s\in\mathcal{S}, d_c^{\rho,\theta}(s)\geq(1-\gamma)\rho(s)$. Additionally, as $\theta\in B(\lambda)$ then $\|\theta\|_\infty\leq\frac{1+\lambda\log(2)}{\lambda(1-\gamma)}$. Combining the two preceding inequalities yields

$$\min_{c\in[M]}\inf_{\theta\in B(\lambda)\cap P}\mu_{\mathrm{r},c}^{\lambda}(\theta)\geq\frac{\lambda(1-\gamma)\min_s\rho(s)^2}{4|\mathcal{S}|}\mathrm{e}^{-4\cdot\frac{1+\lambda\log(2)}{\lambda(1-\gamma)}}\quad,$$

which concludes the proof. □

Combining the previous results, we can prove an exactly similar convergence rate to the one in Theorem E.12, with the difference that we have a lower bound on the constant $\mu_{\mathrm{r}}^{\lambda}$.

Theorem E.21. *Assume **A**-1, set $\theta^0=(0)$ and $\mathcal{T}=B(\lambda)$. For any $\eta>0$ such that $\eta H\leq 888^{-1}(1-\gamma)^3(1+\lambda\log(|\mathcal{A}|))^{-1}$, and $T\geq 1/(1-\gamma)$the iterates of `RS-FedPG` satisfy*

$$\begin{aligned}J_{\mathrm{r}}^{\star}-\mathbb{E}[J_{\mathrm{r}}(\theta^R)]\leq\left(1-\frac{\eta H\mu_{\mathrm{r}}^{\lambda}}{2}\right)^R(J_{\mathrm{r}}^{\star}-J_{\mathrm{r}}(\theta^0)+\frac{38(1+\lambda\log(2))^2\varepsilon_{\mathsf{P}}^2}{\mu_{\mathrm{r}}^{\lambda}(1-\gamma)^6}+\frac{864\eta(1+\lambda\log(2)^3}{BM\mu_{\mathrm{r}}^{\lambda}(1-\gamma)^7}\\+\frac{16(1+\lambda\log(2))^2\gamma^{2T}T^2}{\mu_{\mathrm{r}}^{\lambda}(1-\gamma)^2}+\frac{51^8\eta^4H(H-1)(1+\lambda\log(2))^6}{\mu_{\mathrm{r}}^{\lambda}B^2(1-\gamma)^{16}}\quad.\end{aligned}$$

*where $\mu_{\mathrm{r}}^{\lambda}=\inf_{\theta\in B(\lambda)\cap P}\mu_{\mathrm{r}}^{\lambda}(\theta)$. Additionally, if one assume **A**-2 then*

$$\mu_{\mathrm{r}}^{\lambda}\geq\frac{\lambda(1-\gamma)\min_s\rho(s)^2}{4|\mathcal{S}|}\mathrm{e}^{-4\cdot\frac{1+\lambda\log(2)}{\lambda(1-\gamma)}}\quad.$$

Proof. The proof is exactly similar to that of Theorem E.12. The only difference is that we include an additional step at the very beginning, that is

$$-\mathbb{E}\left[J_{\mathrm{r}}(\theta^{r+1})\middle|\mathcal{F}^r\right]\leq-\mathbb{E}\left[J_{\mathrm{r}}(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right]\quad,$$

which holds by Lemma E.16 and Lemma E.19. The bound on $\mu_{\mathrm{r}}^{\lambda}$ holds by Lemma E.20. □

F Analysis of `b-RS-FedPG`

F.1 Bit-level auto-regressive softmax parametrization

Let us consider an FRL instance $(\mathcal{M}_c)_{c\in[M]}$, where the number of actions is a power of two, i.e, there exists k such that $|\mathcal{A}|=2^k$. This can be assumed without any loss of generality by adding artificial actions to the MDP that have the same effect as any fixed action $a\in\mathcal{A}$ for instance. In what follows, we aim to construct an *bit-level* FRL instance $(\bar{\mathcal{M}}_c:=(\bar{\mathcal{S}},\bar{\mathcal{A}},\bar{\gamma},\bar{\mathsf{P}}_c,\bar{\mathsf{r}}))_{c\in[M]}$ with 2 actions and show that we can reformulate the FRL task associated with $(\mathcal{M}_c)_{c\in[M]}$ over the class of stationary non-deterministic policies by solving the FRL task associated with $(\bar{\mathcal{M}}_c)_{c\in[M]}$.

Define the alphabet $\Sigma:=\{0,1\}$ and define the space of all words in alphabet Σ as $\Sigma^*=\bigcup_{k=0}^{\infty}\Sigma^k$ including the empty word, the length of a word w is denoted by $|w|$. Additionally, we define the operation of concatenation of two words w and w' as $w\circ w'$ and we define a prefix of length k of the word $w=w_1\circ w_2\ldots\circ w_{|w|}$ as $w_{:k}=w_1\circ w_2\circ\ldots\circ w_k$ for $k\leq|w|$, where $w_i\in\Sigma$ are individual characters. Then we define $\Sigma^{<k}$ as a set of words of length smaller than k.

Then, since $|\mathcal{A}| = 2^k$, we can associate the action space of the FRL instance $(\mathcal{M}_c)_{c\in[M]}$ with a set Σ^k of binary words of length exactly k, and define the corresponding action as a_w for $w \in \Sigma^k$. Conversely, we define $w(a)$ as the word associated with action a.

Now, consider an FRL instance with $|\mathcal{S}| \times 2^{k-1}$ states, which we denote by $\bar{\mathcal{S}}$ for $\bar{\mathcal{S}} := \mathcal{S} \times \Sigma^{<k}$.

The set of actions of this FRL is given by our binary alphabet $\bar{\mathcal{A}} := \Sigma$.

For a given agent $c \in [M]$, the transition kernel of agent c is defined as:

$$\bar{\mathsf{P}}_c((s', w')|(s, w), \bar{a}) := \begin{cases} \mathsf{P}_c(s'|s, a_{w\circ\bar{a}}) \cdot 1(w' = \emptyset) & \text{if } |w| = k-1, \\ 1((s', w') = (s, w \circ \bar{a})) & \text{otherwise .} \end{cases}$$

Similarly, we define the reward function as follows:

$$\bar{\mathsf{r}}((s, w), \bar{a}) := \begin{cases} \gamma^{-\frac{k-1}{k}} \mathsf{r}(s_i, a_{w\circ\bar{a}}), & \text{if } |w| = k-1\,, \\ 0, & \text{otherwise .} \end{cases}$$

For a given logit $\theta \in \mathbb{R}^{|\bar{\mathcal{S}}|\times|\bar{\mathcal{A}}|}$, we define the following softmax policy in the extended environment as

$$\bar{\pi}_\theta(0|(s, w)) := \frac{\exp(\theta((s, w), 0))}{\exp(\theta((s, w), 1)) + \exp(\theta((s, w), 0))}\ ,$$
$$\bar{\pi}_\theta(1|(s, w)) := \frac{\exp(\theta((s, w), 1))}{\exp(\theta((s, w), 0)) + \exp(\theta((s, w), 1))}\ .$$

Drawing inspiration for auto-regressive sequence modeling, we can define the following corresponding policy in the original FRL instance

$$\pi_\theta(a_w|s) := \prod_{p=1}^{k} \bar{\pi}_\theta(w_p|(s, w_{:p})) = \prod_{p=1}^{k} \frac{\exp(\theta((s, w_{:p}), w_p))}{\exp(\theta((s, w_{:p}), 0)) + \exp(\theta((s, w_{:p}), 1))}\ .$$

Compared to a usual softmax parameterization, this bit-level softmax parameterization allows to execute a policy π_θ using only $k = \log_2(|\mathcal{A}|)$ operations instead of $|\mathcal{A}|$, which is useful when the action space is large.

In this bit-level FRL instance, the discount factor γ must be rescaled to reflect the fact that states embedding the original FRL instance's states are k times farther apart. We define the rescaled discount factor as $\bar{\gamma} := \gamma^{1/k}$. Define the bit-entropy regulariser as

$$\mathcal{H}^{\rho}_{\mathrm{b},c}(\theta) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t h^\theta_{\mathrm{b}}(S^t_c, A^t_c)\middle| S^0_c \sim \rho\right]\ , \quad \text{where} \tag{65}$$
$$h^\theta_{\mathrm{b}}(s, a) := -\sum_{p=0}^{k-1} \gamma^{p/k} \log \bar{\pi}_\theta(w(a)_p|(s, w(a)_{:p}))\ .$$

Finally, denote by $\widetilde{V}^\theta_{\mathrm{b},c}(s) = V^{\pi_\theta}_c(s) + \lambda\mathcal{H}^\theta_{\mathrm{b},c}(s)$ and by $\bar{V}_c$ the entropy-regularized value function in this bit-level MDP associated to the c-the agent. The following proposition links the entropy-regularised value function associated with a policy $\bar{\pi}_\theta$ in the bit-level MDP to the bit-entropy regularised value function associated with the policy $\bar{\pi}_\theta$ in the original MDP.

Proposition F.1. *Let $(\mathcal{M}_c)_{c\in[M]}$ be an FRL instance and let $(\bar{\mathcal{M}}_c)_{c\in[M]}$ be the corresponding bit-level FRL instance. For any $c \in [M]$, for any $s \in \mathcal{S}$, it holds that*

$$\bar{V}^{\bar{\pi}_\theta}_c((s, \emptyset)) = \widetilde{V}^\theta_{\mathrm{b},c}(s)\ .$$

Proof. Fix $c \in [M]$ and let $s \in \mathcal{S}$. Let $((\bar{S}^t_c, \bar{W}^t_c), \bar{A}^t_c)$ be the trajectory followed by an agent starting from state $(s, \emptyset)$ and following the policy $\bar{\pi}_\theta$ in the MDP $\bar{\mathcal{M}}_c$. Similarly, let (S^t_c, A^t_c) be the trajectory followed by an agent following the policy π_θ in the MDP $\mathcal{M}^c$. By construction, it holds that $(S^t_c, A^t_c) \sim (\bar{S}^{kt+k-1}_c, a_{\bar{W}^{kt+k-1}_c \circ \bar{A}^{kt+k-1}_c})$. Note that the agent obtains non-null-reward

only when he is in a state of type (s, w) where w is a word of length $k-1$ and which happens deterministically every k iterations. Thus, we have

$$\begin{aligned}
\bar{V}_c^{\bar{\pi}_\theta}((s,\emptyset)) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\bar{\gamma}^t\left[\bar{\mathsf{r}}((\bar{S}_c^t,\bar{W}_c^t),\bar{A}_c^t) - \lambda\log\bar{\pi}_\theta(\bar{A}_c^t \mid (\bar{S}_c^t,\bar{W}_t^c))\right]\middle|(\bar{S}_c^0,\bar{W}_c^0)=(s,\emptyset)\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\bar{\gamma}^{kt+k-1}\bar{\mathsf{r}}((\bar{S}_c^{kt+k-1},\bar{W}_c^{kt+k-1}),\bar{A}_{kt+k-1}^c)\middle|(\bar{S}_c^0,\bar{W}_c^0)=(s,\emptyset)\right] \\
&- \lambda\sum_{t=0}^{\infty}\bar{\gamma}^{kt}\mathbb{E}_\pi\left[\sum_{p=0}^{k-1}\bar{\gamma}^p\log\bar{\pi}_\theta(\bar{A}_c^{kt+p} \mid (\bar{S}_c^{kt+p},\bar{W}_c^{kt+p}))\middle|(\bar{S}_c^0,\bar{W}_c^0)=(s,\emptyset)\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\mathsf{r}(S_c^t,A_c^t)\middle|S_c^0=s\right] \\
&- \lambda\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_\pi\left[\sum_{p=0}^{k-1}\bar{\gamma}^p\log\bar{\pi}_\theta(w(A_c^t)_p|(S_c^t,w(A_c^t)_{:p}))\middle|S_c^0=s\right] = \widetilde{V}_{\mathrm{b},c}^{\theta}(s)\ ,
\end{aligned}$$

where in the before last inequality, we used that for any $p \in \{0,\dots k-1\}$, $\bar{W}_c^{kt+p}$ is a deterministic function of $\bar{W}_c^{kt+k-1} \circ \bar{A}_c^{kt+t-1}$ and that $a_{\bar{W}_c^{kt+k-1}\circ\bar{A}_c^{kt+k-1}} \sim A_c^t$ combined with the fact that $\bar{S}_c^{kt+p} = \bar{S}_c^{kt}$. □

F.2 Designing and analyzing `b-RS-FedPG`

`b-RS-FedPG` is a special instance of `proj-FedAVG` in which, the local objective function is $f_c = J_{\mathrm{b},c} \triangleq J_{\mathrm{sm},c} + \lambda\mathcal{H}_{\mathrm{b},c}^{\rho}$, We define the global objective of this algorithm as $J_{\mathrm{b}} := 1/M\sum_{c=1}^{M} J_{\mathrm{b},c}$. The projection set for this algorithm is $\mathcal{T} = B_{\mathrm{b}}(\lambda)$ defined as the ℓ_∞ ball of radius $R_{\mathfrak{b}}(\lambda) \triangleq (1+\lambda\log(2))/(\lambda(1-\bar{\gamma}))$. Finally, the data distribution $\xi_c(\theta)$ corresponds to the distribution $\nu_c(\theta)$. Subsequently, we aim to show that `b-RS-FedPG` satisfies the same properties of `RS-FedPG` in the case where $|\mathcal{A}| = 2$.

Using Proposition F.1, it holds that for any $c \in [M]$, $J_{\mathrm{b},c}$ inherit similar properties to the one established on $J_{\mathrm{r},c}$. In particular, we have that:

- $J_{\mathrm{b},c}$ is $L_{2,\mathrm{b}} \triangleq (8+\lambda(4+8\log(2))/(1-\bar{\gamma})^3$-smooth.
- ∇J_{b} is bounded by $L_{1,\mathrm{b}} \triangleq (1+\lambda\log(2))(1-\bar{\gamma})^2$.
- J_{b} is three times differentiable and has a third derivative tensor bounded by $L_{3,\mathrm{b}} \triangleq (480 + 832\lambda\log(2))/(1-\bar{\gamma})^4$.
- the gradient heterogeneity is uniformly bounded by $\zeta_{\mathrm{r}}^2 \triangleq \frac{38(1+\lambda\log(2))^2\varepsilon_{\mathsf{F}}^2}{(1-\bar{\gamma})^6}$.
- $J_{\mathrm{b},c}(\mathrm{proj}_{B(\lambda)}(\theta)) \geq J_{\mathrm{b},c}(\theta)$, for any $\theta \in P$, where

$$P_{\mathrm{b}} \triangleq \{\theta \in \mathbb{R}^{\bar{\mathcal{S}}\times\bar{\mathcal{A}}}, \forall(s,w)\in\bar{\mathcal{S}}, \langle\theta((s,w),\cdot),(1)\rangle = 0\}\ . \tag{66}$$

We define the following distribution on $\bar{\mathcal{S}}$.

$$\bar{\rho}((s,w)) := \begin{cases}\rho(s), & \text{if } w=\emptyset\,,\\ 0, & \text{otherwise}\,.\end{cases}$$

By Proposition F.1, it also holds that

$$\|\nabla J_{\mathrm{b},c}(\theta)\|_2^2 \geq 2\mu_{\mathrm{b},c}^{\lambda}(\theta)\left[J_{\mathrm{r},c}^{\star} - J_{\mathrm{r},c}(\theta)\right],\ \text{where}$$

$$\mu_{\mathrm{b},c}^{\lambda}(\theta) \triangleq \frac{\lambda}{|\mathcal{S}|(1-\bar{\gamma})}\min_{(s,w)\in\bar{\mathcal{S}}}\bar{d}_c^{\bar{\rho},\theta}((s,w))\min_{((s,w),a)\in\bar{\mathcal{S}}\times\bar{\mathcal{A}}}\bar{\pi}_\theta(a|(s,w))^2\left\|\frac{\bar{d}_c^{\bar{\rho},\bar{\pi}_c^{\star,\lambda}}}{\bar{d}_c^{\bar{\rho},\theta}}\right\|_\infty^{-1}, \tag{67}$$

where $\bar{\pi}_c^{\star,\lambda}$ is the unique optimal regularized policy of agent c in the bit-level FRL instance, and $\bar{d}_c^{\bar{\rho},\pi_\theta}$ is the state occupancy distribution of agent c in the bit-level FRL instance when using the policy $\bar{\pi}_\theta$.

We now cautiously design the stochastic estimator of the gradient so that it matches the entropy-regularized stochastic estimator that would have been used in the bit-level MDP by `RS-FedPG`. For any parameter $\theta \in \mathbb{R}^{|\bar{\mathcal{S}}|\times|\bar{\mathcal{A}}|}$, and given an observation $Z_c \sim \nu_c(\theta)$, the biased estimator of the stochastic gradient is set to be:

$$\mathrm{g}_{\mathrm{b},c}^{Z_c} := \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{k(T-1)} \bar{\gamma}^t \left(\sum_{\ell=0}^{t} \nabla \log \bar{\pi}_\theta(w(A_{c,b}^{\mathrm{p}_t})_{\mathrm{q}_t}|(S_{c,b}^{\mathrm{p}_t}, w(A_{c,b}^{\mathrm{p}_t})_{:\mathrm{q}_t}))\right) R_{c,b}^t \ , \tag{68}$$

where

$$R_{c,b}^t \triangleq \left[1_{\mathrm{q}_t=k-1}\mathsf{r}(S_{c,b}^{\mathrm{p}_t}, A_{c,b}^{\mathrm{p}_t}) - \lambda \log \bar{\pi}_\theta(w(A_{c,b}^{\mathrm{p}_t})_{\mathrm{q}_t}|(S_{c,b}^{\mathrm{p}_t}, w(A_{c,b}^{\mathrm{p}_t})_{:\mathrm{q}_t}))\right] \ ,$$

$\mathrm{p}_t \triangleq \lfloor t/k \rfloor$, and $\mathrm{q}_t \triangleq t - k\lfloor t/k \rfloor$. Importantly, note that the distribution of $((S_{c,b}^{\mathrm{p}_t}, w(A_{c,b}^{\mathrm{p}_t})_{:\mathrm{q}_t}), w(A_{c,b}^{\mathrm{p}_t})_{\mathrm{q}_t})$ is $\nu_{\mathrm{b},c}(\theta)$ where for any $c \in [M]$, $\theta \in \mathbb{R}^{|\bar{\mathcal{S}}|\times|\bar{\mathcal{A}}|}$, and $z \in (\bar{\mathcal{S}}\times\bar{\mathcal{A}})^{kT}$, we have

$$\nu_{\mathrm{b},c}(\theta; z) \triangleq \bar{\rho}((s^0, w^0))\bar{\pi}_\theta(a^0|(s^0, w^0)) \cdot \prod_{t=1}^{kT-1} \bar{\mathsf{P}}_c((s^t, w^t) \mid (s^{t-1}, w^{t-1}), a^{t-1})\bar{\pi}_\theta(a^t|(s^t, w^t)) \ .$$

This remark is fundamental as it allows us to derive the following properties of the stochastic estimator

- The second and fourth central moment of $\mathrm{g}_{\mathrm{b},c}^{Z_c}(\theta)$ are respectively bounded by
$$\sigma_{\mathrm{b},2}^2 \triangleq \frac{12 + 24\lambda^2(\log(2))^2}{B(1-\bar{\gamma})^4}, \ \ \sigma_{\mathrm{b},4}^4 \triangleq \frac{1120 + 4480\lambda^4 \log(2)^4}{B^2(1-\bar{\gamma})^8} \ .$$
- The bias of the estimator is bounded by $\beta_{\mathrm{b}} \triangleq \frac{2(1+\lambda\log(2))\bar{\gamma}^{kT}}{1-\bar{\gamma}}\left(kT + \frac{1}{1-\bar{\gamma}}\right)$.
- $\mathrm{g}_{\mathrm{b},c} \triangleq \mathbb{E}_{Z_c\sim\nu_c(\theta)}[\mathrm{g}_{\mathrm{b},c}^{Z_c}]$ is $L_{2,\mathrm{b}} \triangleq (8 + \lambda(4 + 8\log(2))/(1-\bar{\gamma})^3$-smooth.
- For all $r \in [R]$, we have $\theta^r \in P_{\mathrm{b}}$ where $(\theta_c^{r,h})$ are the iterates generated by `b-RS-FedPG`.

Before applying Theorem E.21 to derive the convergence for `b-RS-FedPG`, it remains to prove that $\inf_{\theta\in B_{\mathrm{b}}(\lambda)\cap P_{\mathrm{b}}} \mu_{\mathrm{b}}^\lambda(\theta) \triangleq \min_{c\in[M]} \mu_{\mathrm{b},c}^\lambda(\theta)$ is strictly positive. This is not straightforward as $\bar{\rho}$ does not cover the whole state $\bar{\mathcal{S}}$, and thus it is not clear why $\bar{d}_c^{\bar{\rho},\pi_\theta}((s,w))$ is positive for all $(s,w) \in \bar{\mathcal{S}}$. However, by exploiting our knowledge on the transitions of the bit-level FRL instance, we can guarantee such a property, which is what we do subsequently.

Lemma F.2. *Assume A-2. It holds that*

$$\min_{c\in[M]} \inf_{\theta\in B(\lambda)\cap P} \mu_{\mathrm{r},c}^\lambda(\theta) \geq \frac{\bar{\gamma}^{3k-3}\lambda(1-\bar{\gamma})}{4^k|\mathcal{S}|} \min_s \rho(s)^2 \mathrm{e}^{-4k\cdot\frac{1+\lambda\log(2)}{\lambda(1-\bar{\gamma})}} \ .$$

Proof. Fix any $\theta \in \mathbb{R}^{|\bar{\mathcal{S}}|\times|\bar{\mathcal{A}}|}$ and $c \in [M]$. Under **A**-2, it holds that for any $(s,\emptyset) \in \bar{S}$, $\bar{d}_{\bar{\rho}}^{c,\theta}((s,\emptyset)) \geq (1-\bar{\gamma})\rho(s)$. Using the flow conservation constraints for occupancy measures (Puterman, 1994), for any agent $c \in [M]$, and $(s,w) \in \bar{\mathcal{S}}$ such that $k = |w| \neq \emptyset$, it holds that

$$\begin{aligned}\bar{d}_{\bar{\rho}}^{c,\theta}((s,w)) &= (1-\gamma)\bar{\rho}(s,w) + \bar{\gamma}\sum_{(s',w'),a'} \bar{\mathsf{P}}_c((s,w)|(s',w'),a')\bar{\pi}_\theta(a'|(s',w'))\bar{d}_c^{\bar{\rho},\theta}((s',w')) \\ &\geq \bar{\gamma}\pi_\theta(w_k|(s,w_{:k-1}))\bar{d}_c^{\bar{\rho},\theta}((s,w_{:k-1})) \geq \bar{\gamma}\min_{(s,w),a} \bar{\pi}_\theta(a|(s,w))\bar{d}_c^{\bar{\rho},\theta}((s,w_{:k-1})) \ .\end{aligned}$$

Unrolling the recursion on k, implies that

$$\min_{(s,w),a} \bar{d}_{\bar{\rho}}^{c,\theta}((s,w)) \geq \bar{\gamma}^{k-1} \min_{(s,w),a} \bar{\pi}_\theta(a|(s,w))^{k-1}(1-\bar{\gamma})\rho(s) \ . \tag{69}$$

Additionally, as $\theta \in B_{\mathrm{b}}(\lambda) \cap P_{\mathrm{b}}$ then $\min_{(s,w),a} \bar{\pi}_\theta(a|(s,w)) \geq \mathrm{e}^{-2\cdot\frac{1+\lambda\log(2)}{\lambda(1-\bar{\gamma})}}/2$. Combining the two preceding inequalities and applying (67) concludes the proof. □

Finally, we can directly apply Theorem E.21 to derive the convergence for `b-RS-FedPG`.

Theorem F.3. *Assume **A**-1 and **A**-2, set $\theta^0 = (0)$ and $\mathcal{T} = B_{\mathrm{b}}(\lambda)$. For any $\eta > 0$ such that $\eta H \leq 888^{-1}(1-\bar{\gamma})^3(1+\lambda\log(2))^{-1}$, and $T \geq 1/(1-\bar{\gamma})$the iterates of `b-RS-FedPG` satisfy*

$$J_{\mathrm{b}}^{\star} - \mathbb{E}[J_{\mathrm{b}}(\theta^R)] \leq \left(1 - \frac{\eta H \mu_{\mathrm{b}}^{\lambda}}{2}\right)^R (J_{\mathrm{b}}^{\star} - J_{\mathrm{b}}(\theta^0) + \frac{38(1+\lambda\log(2))^2\varepsilon_{\mathsf{P}}^2}{\mu_{\mathrm{b}}^{\lambda}(1-\bar{\gamma})^6} + \frac{864\eta(1+\lambda\log(2)^3}{BM\mu_{\mathrm{b}}^{\lambda}(1-\bar{\gamma})^7} + \frac{16(1+\lambda\log(2))^2\bar{\gamma}^{2T}T^2}{\mu_{\mathrm{b}}^{\lambda}(1-\bar{\gamma})^2} + \frac{51^8\eta^4 H(H-1)(1+\lambda\log(2))^6}{\mu_{\mathrm{b}}^{\lambda}B^2(1-\bar{\gamma})^{16}} \ .$$

where

$$\mu_{\mathrm{b}}^{\lambda} \triangleq \frac{\bar{\gamma}^{3k-3}\lambda(1-\bar{\gamma})}{4^k|\mathcal{S}|}\min_s \rho(s)^2 \mathrm{e}^{-4k\cdot\frac{1+\lambda\log(2)}{\lambda(1-\bar{\gamma})}} \ .$$

Corollary F.4 (Sample and Communication Complexity of `b-RS-FedPG`)**.** *Assume **A**-1 and **A**-2. Set $\theta^0 = (0), \mathcal{T} = B_{\mathrm{b}}(\lambda)$. Define*

$$\mu_{\mathrm{b}}^{\lambda} \triangleq \frac{\bar{\gamma}^{3k-3}\lambda(1-\bar{\gamma})}{4^k|\mathcal{S}|}\min_s \rho(s)^2 \mathrm{e}^{-4k\cdot\frac{1+\lambda\log(2)}{\lambda(1-\bar{\gamma})}} \ .$$

Let $\epsilon \geq 190(1+\lambda\log(2))^2\varepsilon_{\mathsf{P}}^2(\mu_{\mathrm{r}}^{\lambda})^{-1}(1-\bar{\gamma})^{-6}$. Then, for a properly chosen truncation horizon, a properly chosen step size and number of local updates, `b-RS-FedPG` learns an ϵ-approximation of the optimal objective with a number of communication rounds

$$R \geq \frac{888(1+\lambda\log(2)}{(1-\bar{\gamma})^2\mu_{\mathrm{b}}^{\lambda}}\log\left(\frac{5(J_{\mathrm{b}}^{\star} - J_{\mathrm{b}}(\theta^0))}{\epsilon}\right) \ ,$$

for a total number of samples per agent of

$$RHB \geq \max\Big(\frac{24(1+\lambda\log(2))B}{\mu_{\mathrm{b}}^{\lambda}(1-\bar{\gamma})^3}, \frac{8640(1+\lambda\log(2))^3}{(\mu_{\mathrm{b}}^{\lambda})^2\epsilon M(1-\bar{\gamma})^7}, \frac{2\cdot 12^4(1+\lambda\log(2))^2}{\epsilon^{1/2}(\mu_{\mathrm{b}}^{\lambda})^{3/2}(1-\bar{\gamma})^5}\Big)\log\left(\frac{5(J_{\mathrm{b}}^{\star} - J_{\mathrm{b}}(\theta^0))}{\epsilon}\right) \ .$$

G Technical lemmas

G.1 Basic Lemmas

For completeness, we state withtout proofs basic results which are routinely used in our proofs.

Lemma G.1 (Theorem 2.1.5, Nesterov (2018))**.** *If $f\colon \mathbb{R}^d \to \mathbb{R}$ is a L-smooth function, then we have for any $x, y \in \mathbb{R}^d$*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|x-y\|_2^2 \ .$$

Lemma G.2 (Reinforce)**.** *Let $(\mathsf{Z}, \mathcal{Z})$ be a measurable space, let $\Theta \subset \mathbb{R}^d$ be open, and let μ be a σ-finite measure on $(\mathsf{Z}, \mathcal{Z})$. Suppose*

1. *$Y\colon \mathsf{Z}\times\Theta \to \mathbb{R}$ is $\mathcal{Z}\otimes\mathcal{B}(\Theta)$-measurable.*
2. *For each $z \in \mathsf{Z}$ and each $i = 1, \ldots, d$, the partial derivative*

$$\frac{\partial Y(z,\theta)}{\partial\theta_i}$$

exists for all $\theta \in \Theta$ and the map

$$\mathsf{Z}\times\Theta \ \ni \ (z,\theta) \ \mapsto \ \frac{\partial Y(z,\theta)}{\partial\theta_i}$$

is measurable.

3. *For each $\theta \in \Theta$, $\gamma_\theta \colon \mathsf{Z} \to [0,\infty)$ is a probability density w.r.t. μ, and for each $i = 1,\dots,d$ the map*

$$z \;\mapsto\; \frac{\partial \gamma_\theta(z)}{\partial \theta_i}$$

exists for all $\theta \in \Theta$ and is measurable on Z.

4. *(Dominating function.) For each $i = 1,\dots,d$ and each $\theta_0 \in \Theta$, there exist a neighborhood $U \subset \Theta$ of θ_0 and an integrable function $h_i \in L^1(\mu)$ such that for μ-a.e. $z \in \mathsf{Z}$ and all $\theta \in U$,*

$$\left| \frac{\partial}{\partial \theta_i} \Big[Y(z,\theta)\, \gamma_\theta(z) \Big] \right| \;=\; \left| \frac{\partial Y(z,\theta)}{\partial \theta_i}\, \gamma_\theta(z) + Y(z,\theta)\, \frac{\partial \gamma_\theta(z)}{\partial \theta_i} \right| \;\le\; h_i(z).$$

Define

$$J(\theta) \;=\; \int_{\mathsf{Z}} Y(z,\theta)\, \gamma_\theta(z)\, \mu(\mathrm{d}z).$$

Then $J \colon \Theta \to \mathbb{R}$ is continuously differentiable, and for each $i = 1,\dots,d$,

$$\frac{\partial J(\theta)}{\partial \theta_i} = \int_{\mathsf{Z}} \frac{\partial}{\partial \theta_i} \Big[Y(z,\theta)\, \gamma_\theta(z) \Big]\, \mu(dz).$$

Equivalently,

$$\frac{\partial J(\theta)}{\partial \theta_i} = \int_{\mathsf{Z}} \left[\frac{\partial Y(z,\theta)}{\partial \theta_i} + Y(z,\theta)\, \frac{\partial \ln \gamma_\theta(z)}{\partial \theta_i} \right] \gamma_\theta(z) \mu(\mathrm{d}z).$$

G.2 Performance difference lemma

Lemma G.3 (First performance-difference lemma, Kakade and Langford (2002))**.** *Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}, \mathsf{r})$ and let V^π and be the value function in this MDP. For any policies π_1 and π_2, it holds*

$$V^{\pi_1}(\rho) - V^{\pi_2}(\rho) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d^{\rho,\pi_1}(s) \sum_{a \in \mathcal{A}} \pi_1(a|s) \cdot A^{\pi_2}(s,a) \;\;,$$

where A^{π_2} is the advantage function.

Lemma G.4 (Second Performance difference lemma, Russo (2019))**.** *Let us consider two MDPs $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_1, \mathsf{r})$ and $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_2, \mathsf{r})$. Let V_1^π and V_2^π be respectively the two value functions in these two MDPs. It holds that*

$$V_1^\pi(s) - V_2^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (\mathsf{P}_1 - \mathsf{P}_2) V_2^\pi(S^t, A^t) \middle| s_0 = s \right] \;\;,$$

where the expectation is taken over the trajectories $(S^0, A^0, S^1, A^1 \dots)$ generated by a stationary policy π in the MDP $\mathcal{M}_2$.

Lemma G.5. *Let us consider two MDPs $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_1, \mathsf{r})$ and $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_2, \mathsf{r})$ such that $\sup_{s,a \in \mathcal{S}\times\mathcal{A}} \|\mathsf{P}_1(\cdot|s,a) - \mathsf{P}_2(\cdot|s,a)\|_1 \le \varepsilon_{\mathsf{P}}$. For a given stationary policy π, let V_1^π and V_2^π be respectively the two value functions of this policy in these two MDPs. If $\|V_1^\pi\|_\infty \le c$ and $\|V_2^\pi\|_\infty \le c$ then it holds that for all $s \in \mathcal{S}$*

$$|V_1^\pi(s) - V_2^\pi(s)| \le \frac{\varepsilon_{\mathsf{P}} c}{1-\gamma} \;\;.$$

Proof. Follows directly from a combination of Lemma G.4, Holder's inequality and the fact that $\|V_2^\pi\|_\infty \le c$ and $\|V_2^\pi\|_\infty \le c$. □

Lemma G.6. *Assume **A**-1. Then, for all $c, c' \in [M]$, it holds that*

$$\|d_{c'}^{\rho,\theta} - d_c^{\rho,\theta}\|_1 \le \frac{\gamma \varepsilon_{\mathsf{P}}}{1-\gamma} \;\;.$$

Proof. Let us start from the definition of flow conservation constraints for occupancy measures (Puterman, 1994) for any agent $c \in [M]$

$$d_c^{\rho,\theta}(s) = (1-\gamma)\rho(s) + \gamma \sum_{(s',a')} \mathsf{P}_c(s|s',a')\pi_\theta(a'|s')d_c^{\rho,\theta}(s') \ .$$

Then, we have

$$\begin{aligned}
\sum_s |d_{c'}^{\rho,\theta}(s) - d_c^{\rho,\theta}(s)| &\leq \gamma \sum_{s',a'} \sum_s \left| \mathsf{P}_{c'}(s|s',a')\pi_\theta(a'|s')d_{c'}^{\rho,\theta}(s') - \mathsf{P}_c(s|s',a')\pi_\theta(a'|s')d_c^{\rho,\theta}(s') \right| \\
&\leq \gamma \sum_{s',a'} \underbrace{\sum_s |\mathsf{P}_{c'}(s|s',a') - \mathsf{P}_c(s|s',a')|}_{\leq \varepsilon_{\mathsf{P}}} \pi_\theta(a'|s')d_{c'}^{\rho,\theta}(s') \\
&+ \gamma \sum_{s',a'} \underbrace{\sum_s \mathsf{P}_c(s|s',a')}_{=1} \pi_\theta(a'|s') \left| d_{c'}^{\rho,\theta}(s') - d_c^{\rho,\theta}(s') \right| \\
&\leq \gamma\varepsilon_{\mathsf{P}} + \gamma \sum_s |d_{c'}^{\rho,\theta}(s) - d_c^{\rho,\theta}(s)| \ ,
\end{aligned}$$

which concludes the proof. □

G.3 Properties of softmax parametrization and value

In this section, we derive useful technical inequalities that show bounds on the derivatives of the softmax parametrization. The results for the first two differentials could be extracted from Mei et al. (2020).

Lemma G.7. *For any $u, v, w \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, we have*

$$\begin{aligned}
|\mathrm{d}\pi_\theta[u](a|s)| &\leq 2\pi_\theta(a|s)\|u\|_\infty \ , \\
|\mathrm{d}^2\pi_\theta[u,v](a|s)| &\leq 8\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty \ , \\
|\mathrm{d}^3\pi_\theta[u,v,w](a|s)| &\leq 48\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty\|w\|_\infty \ .
\end{aligned}$$

Proof. Let us start from the expression for the derivative of parametrization (see, e.g., Lemma C.1. of Agarwal et al. (2020))

$$\frac{\partial \pi_\theta(a|s)}{\partial \theta(s,a_1)} = \pi_\theta(a|s)(\mathsf{1}_a(a_1) - \pi_\theta(a_1|s)) \, ,$$

thus

$$\mathrm{d}\pi_\theta[u](a|s) = \pi_\theta(a|s) \cdot (u(s,a) - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle) \ .$$

To simplify the following notation, we define a random variable $A \sim \pi_\theta(\cdot|s)$, then we have

$$\mathrm{d}\pi_\theta[u](a|s) = \pi_\theta(a|s) \cdot (u(s,a) - \mathbb{E}_{\pi_\theta}[u(s,A)]) \ .$$

Using the fact that $|u(s,a) - \mathbb{E}_{\pi_\theta}[u(s,A)]| \leq 2\|u\|_\infty$, we conclude the first statement.

Next, we continue by deriving the second derivative

$$\begin{aligned}
\frac{\partial^2 \pi_\theta(a|s)}{\partial\theta(s,a_1)\partial\theta(s,a_2)} &= \pi_\theta(a|s)(\mathsf{1}_a(a_2) - \pi_\theta(a_2|s))(\mathsf{1}_a(a_1) - \pi_\theta(a_1|s)) \\
&- \pi_\theta(a|s)\pi_\theta(a_1|s)(\mathsf{1}_{a_1}(a_2) - \pi_\theta(a_2|s)) \\
&= \pi_\theta(a|s)\left((\mathsf{1}_a(a_2) - \pi_\theta(a_2|s))(\mathsf{1}_a(a_1) - \pi_\theta(a_1|s)) - \pi_\theta(a_1|s)(\mathsf{1}_{a_1}(a_2) - \pi_\theta(a_2|s))\right) \ .
\end{aligned}$$

In particular, we have

$$\begin{aligned}
\mathrm{d}^2\pi_\theta[u,v](a|s) = \pi_\theta(a|s) &\sum_{a_1,a_2} ((\mathsf{1}_a(a_2) - \pi_\theta(a_2|s))(\mathsf{1}_a(a_1) - \pi_\theta(a_1|s))))u(s,a_1)u(s,a_2) \\
&- \pi_\theta(a|s) \sum_{a_1,a_2} \pi_\theta(a_1|s)(\mathsf{1}_{a_1}(a_2) - \pi_\theta(a_2|s))u(s,a_1)v(s,a_2)
\end{aligned}$$

$$= \pi_\theta(a|s)(u(s,a) - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle)(v(s,a) - \langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle)$$
$$- \pi_\theta(a|s)\left(\langle \pi_\theta(\cdot|s), u(s,\cdot)\cdot v(s,\cdot)\rangle - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle \cdot \langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle\right).$$

Using the same inequality, we have

$$|\mathrm{d}^2\pi_\theta[u,v](a|s)| \le 8\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty .$$

Finally, we continue with the computation of the third differential:

$$\mathrm{d}^3\pi_\theta[u,v,w](a|s) = \underbrace{\mathrm{d}\left[\pi_\theta(a|s)(u(s,a) - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle)(v(s,a) - \langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle)\right][w]}_{(\mathbf{D_1})}$$
$$- \underbrace{\mathrm{d}\left[\pi_\theta(a|s)\left(\langle \pi_\theta(\cdot|s), u(s,\cdot)\cdot v(s,\cdot)\rangle - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle \cdot \langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle\right)\right][w]}_{(\mathbf{D_2})} .$$

Next, we consider each term separately. First, we have

$$(\mathbf{D_1}) = \mathrm{d}\pi_\theta[w](a|s)\cdot(u(s,a) - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle)(v(s,a) - \langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle)[w]$$
$$- \pi_\theta(a|s)\langle \mathrm{d}\pi_\theta[w](\cdot|s), u(s,\cdot)\rangle)(v(s,a) - \langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle)$$
$$- \pi_\theta(a|s)(u(s,a) - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle)\langle \mathrm{d}\pi_\theta(\cdot|s)[w], v(s,\cdot)\rangle .$$

To bound this term, we notice that for any $x \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ it holds

$$\langle \mathrm{d}\pi_\theta(\cdot|s)[w], x(s,\cdot)\rangle = \sum_{a\in\mathcal{A}} \mathrm{d}\pi_\theta(a|s)[w]\cdot x(s,a)$$
$$= \sum_{a\in\mathcal{A}} \pi_\theta(a|s)(w(s,a) - \langle \pi_\theta(\cdot|s), w(s,\cdot)\rangle)x(s,a)$$
$$= \mathbb{E}\left[x(s,A)w(s,A)\right] - \mathbb{E}\left[x(s,A)\right]\mathbb{E}\left[w(s,A)\right] = \mathrm{Cov}(x(s,A), w(s,A)) ,$$

where a random variable A follows $\pi_\theta(\cdot|s)$. Using this relation, we have

$$|(\mathbf{D_1})| \le \pi_\theta(a|s)\cdot|w(s,a) - \mathbb{E}[w(s,A)]|\cdot|u(s,a) - \mathbb{E}[w(s,A)]|\cdot|v(s,a) - \mathbb{E}[w(s,A)]|$$
$$+ \pi_\theta(a|s)|\mathrm{Cov}(u(s,A), w(s,A))||v(s,a) - \mathbb{E}[v(s,A)]|$$
$$+ \pi_\theta(a|s)|\mathrm{Cov}(v(s,A), w(s,A))||u(s,a) - \mathbb{E}[u(s,A)]| .$$

Next, we notice that $|x(s,a) - \mathbb{E}[x(s,A)]| \le 2\|x\|_\infty$ for any $x \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and, as a result, $|\mathrm{Cov}(x(s,A), w(s,A))| \le 4\|x\|_\infty\|w\|_\infty$. Thus, we have

$$|(\mathbf{D_1})| \le 24\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty\|w\|_\infty .$$

Next, we analyze the second term. For this term, we have

$$(\mathbf{D_2}) = \mathrm{d}\pi_\theta(a|s)[w]\cdot\mathrm{Cov}(u(s,A), v(s,A)) + \pi_\theta(a|s)\Big(\langle \mathrm{d}\pi_\theta(\cdot|s)[w], u(s,\cdot)\cdot v(s,\cdot)\rangle$$
$$- \langle \mathrm{d}\pi_\theta(\cdot|s)[w], u(s,\cdot)\rangle\cdot\langle \pi_\theta(\cdot|s), v(s,\cdot)\rangle - \langle \pi_\theta(\cdot|s), u(s,\cdot)\rangle\cdot\langle \mathrm{d}\pi_\theta[w](\cdot|s), v(s,\cdot)\rangle\Big) .$$

By the same reasoning as for term $(\mathbf{D_1})$, we have

$$|(\mathbf{D_2})| \le 24\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty\|w\|_\infty ,$$

thus we have

$$|\mathrm{d}^3\pi_\theta[u,v,w]|(a|s) \le 48\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty\|w\|_\infty .$$

□

Lemma G.8. *Let $\mathcal{H}(\pi_\theta) \in \mathbb{R}^{\mathcal{S}}$ be a vector of entropies of policy π_θ. Then we have*

$$\|\mathrm{d}\mathcal{H}(\pi_\theta)[u]\|_\infty \le 2\log|\mathcal{A}|\cdot\|u\|_\infty ,$$
$$\|\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v]\|_\infty \le (4 + 8\log|\mathcal{A}|)\|u\|_\infty\|v\|_\infty ,$$
$$\|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\|_\infty \le (56 + 48\log|\mathcal{A}|)\|u\|_\infty\|v\|_\infty\|w\|_\infty$$

Proof. We recall that $\mathcal{H}(\pi_\theta)(s) = -\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\log\pi_\theta(a|s)$.

Define a function $h(x) = -x\log x$, then we have $h'(x) = -(\log x + 1), h''(x) = -1/x, h'''(x) = 1/x^2$. Thus, by Lemma G.7 we have

$$\begin{aligned}\mathrm{d}\mathcal{H}(\pi_\theta)[u](s) &= \sum_{a\in\mathcal{A}} \mathrm{d}h(\pi_\theta(a|s))[u] = \sum_{a\in\mathcal{A}} h'(\pi_\theta(a|s))\cdot \mathrm{d}\pi_\theta[u](a|s)\\ &= \sum_{a\in\mathcal{A}} -(\log\pi_\theta(a|s)+1)\cdot\pi_\theta(a|s)(u(s,a) - \langle\pi_\theta(\cdot|s), u(s,\cdot)\rangle).\end{aligned}$$

Notice that

$$\sum_{a\in\mathcal{A}} \pi_\theta(a|s)(u(s,a) - \langle\pi_\theta(\cdot|s), u(s,\cdot)\rangle) = 0\,,$$

thus, using $|u(s,a) - \langle\pi_\theta(\cdot,s), u\rangle| \le 2\|u\|_\infty$ and $\sum_{a\in\mathcal{A}} |\pi_\theta(a|s)\log\pi_\theta(a|s)| \le \log|\mathcal{A}|$, we conclude the first statement.

Next, we have to compute the second differential; here we have by a high-order chain rule

$$\begin{aligned}\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v](s) &= \sum_{a\in\mathcal{A}} \mathrm{d}^2 h(\pi_\theta(a|s))[u]\\ &= \sum_{a\in\mathcal{A}} h''(\pi_\theta(a|s))\mathrm{d}\pi_\theta(a|s)[u]\mathrm{d}\pi_\theta(a|s)[v] + \sum_{a\in\mathcal{A}} h'(\pi_\theta(a|s))\mathrm{d}^2\pi_\theta(a|s)[u,v]\\ &= \sum_{a\in\mathcal{A}} \left(-\frac{1}{\pi_\theta(a|s)}\right)\mathrm{d}\pi_\theta(a|s)[u]\mathrm{d}\pi_\theta(a|s)[v] - \sum_{a\in\mathcal{A}} (\log\pi_\theta(a|s)+1)\mathrm{d}^2\pi_\theta(a|s)[u,v]\,.\end{aligned}$$

Next, we see that by linearity

$$\sum_{a\in\mathcal{A}} \mathrm{d}\pi_\theta(a|s)[u] = \mathrm{d}\left(\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\right)[u] = 0\,,$$

thus the sum of second and third derivatives also should be equal to zero.

Using a bound from Lemma G.7, we have

$$|\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v](s)| \le \sum_{a\in\mathcal{A}} 4\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty + 8\sum_{a\in\mathcal{A}} |\log\pi_\theta(a|s)|\cdot\pi_\theta(a|s)\|u\|_\infty\|v\|_\infty\,.$$

By a bound on entropy, we conclude the second statement.

For the last statement, we also apply the high-order chain rule to have

$$\begin{aligned}\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w](s) = &\sum_{a\in\mathcal{A}} h'''(\pi_\theta(a|s))\mathrm{d}\pi_\theta(a|s)[u]\mathrm{d}\pi_\theta(a|s)[v]\pi_\theta(a|s)[w]\\ &+ \sum_{a\in\mathcal{A}} h''(\pi_\theta(a|s))\mathrm{d}^2\pi_\theta(a|s)[u,w]\mathrm{d}\pi_\theta(a|s)[v]\\ &+ \sum_{a\in\mathcal{A}} h''(\pi_\theta(a|s))\mathrm{d}\pi_\theta(a|s)[u]\mathrm{d}^2\pi_\theta(a|s)[v,w]\\ &+ \sum_{a\in\mathcal{A}} h''(\pi_\theta(a|s))\mathrm{d}\pi_\theta(a|s)[w]\mathrm{d}^2\pi_\theta(a|s)[u,v]\\ &+ \sum_{a\in\mathcal{A}} h'(\pi_\theta(a|s))\mathrm{d}^3\pi_\theta(a|s)[u,v,w]\,.\end{aligned}$$

Using a fact that $\sum_{a\in\mathcal{A}} \mathrm{d}^3\pi_\theta(a|s)[u,v,w] = 0$, we have the following from by Lemma G.7

$$|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w](s)| \le (56 + 48\log|\mathcal{A}|)\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

□

Lemma G.9. *Let $\widetilde{V}_c^{\pi_\theta}$ be a regularized value function in the MDP that corresponds to an agent $c \in [M]$. Then for any $u, v, w \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, its directional derivatives satisfy the following bounds*

$$\|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty \le \frac{8 + 10\lambda \log|\mathcal{A}|}{1-\gamma}\|u\|_\infty\,,$$
$$\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty \le \frac{40 + 60\lambda \log|\mathcal{A}|}{(1-\gamma)^3}\|u\|_\infty\|v\|_\infty\,,$$
$$\|\mathrm{d}^3\widetilde{V}_c^{\pi_\theta}[u,v,w]\|_\infty \le \frac{480 + 832\lambda \log|\mathcal{A}|}{(1-\gamma)^4}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

Proof. Let us start by writing down regularized Bellman equations (see, e.g., Geist et al. (2019)). In the following, we treat $\widetilde{Q}_c^\pi$ as a matrix of size $\mathcal{S}\times\mathcal{A}$ with elements $\widetilde{Q}_c^\pi(s,a)$ and π_θ as a matrix of size $\mathcal{A}\times\mathcal{S}$ with elements $\pi_\theta(a|s)$,

$$\widetilde{V}_c^{\pi_\theta} = \widetilde{Q}_c^{\pi_\theta}\cdot\pi_\theta + \lambda\mathcal{H}(\pi_\theta)\,, \qquad \widetilde{Q}_c^{\pi_\theta} = \mathsf{r} + \gamma\mathsf{P}_c\widetilde{V}_c^{\pi_\theta},$$

where P_c is a linear operator from a space of vectors of size $\mathcal{S}$ to a space of matrices of size $\mathcal{S}\times\mathcal{A}$, and $\mathcal{H}(\pi)\in\mathbb{R}^{\mathcal{S}}$ is a vector of policy entropies for each state.

First differential. We start as follows

$$\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u] = \widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u] + \mathrm{d}\widetilde{Q}_c^{\pi_\theta}[u]\cdot\pi_\theta + \lambda\mathrm{d}\mathcal{H}(\pi_\theta)[u], \qquad \mathrm{d}\widetilde{Q}_c^{\pi_\theta}[u] = \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\,.$$

Thus, we have

$$\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u] = \widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u] + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\cdot\pi_\theta + \lambda\mathrm{d}\mathcal{H}(\pi_\theta)[u]\,.$$

As a result, we have

$$\|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty \le \|\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u]\|_\infty + \gamma\|\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\cdot\pi_\theta\|_\infty + \lambda\|\mathrm{d}\mathcal{H}(\pi_\theta)[u]\|_\infty\,. \tag{70}$$

For the first term, we have for any $s\in\mathcal{S}$ by a simple bound on Q-value and Lemma G.7

$$|\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u]|(s) \le \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sum_{a\in\mathcal{A}}|\mathrm{d}\pi_\theta[u](a|s)| \le \frac{8(1+\lambda\log A)}{1-\gamma}\|u\|_\infty\,.$$

For the second term, we have for any $s\in\mathcal{S}$

$$\|\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\cdot\pi_\theta\|_\infty = \max_s\left|\sum_{a,s'}\mathsf{P}_c(s'|s,a)\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\cdot\pi_\theta(a|s)\right| \le \|\widetilde{V}_c^{\pi_\theta}[u]\|_\infty\,.$$

finally, by Lemma G.8 we have

$$\|\mathrm{d}\mathcal{H}(\pi_\theta)[u]\|_\infty \le 2\log|\mathcal{A}|\cdot\|u\|_\infty.$$

Thus, from (70) it holds

$$\|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty \le \gamma\|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty + \frac{8+10\lambda\log|\mathcal{A}|}{1-\gamma}\|u\|\,.$$

Rearranging the terms, we conclude the first statement.

Second differential. For the second differential, we have

$$\begin{aligned}
\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v] &= \mathrm{d}\left(\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u]\right)[v] + \gamma\mathrm{d}\left(\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\cdot\pi_\theta\right)[v] + \lambda\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v]\\
&= \left(\mathrm{d}\widetilde{Q}_c^{\pi_\theta}[v]\right)\mathrm{d}\pi_\theta[u] + \widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}^2\pi_\theta[u,v] + \gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\cdot\pi_\theta\\
&\quad + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\mathrm{d}\pi_\theta[v] + \lambda\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v]\\
&= \widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}^2\pi_\theta[u,v] + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\mathrm{d}\pi_\theta[v] + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[v]\mathrm{d}\pi_\theta[u]\\
&\quad + \gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\cdot\pi_\theta + \lambda\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v]\,.
\end{aligned}$$

Next, to derive a bound, we apply the bound on the first differential of the value as well Lemma G.7 and Lemma G.8:

$$|\widetilde{Q}_c^{\pi_\theta} \cdot \mathrm{d}^2\pi_\theta[u,v]|(s) \leq \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sum_{a\in\mathcal{A}}|\mathrm{d}^2\pi_\theta(a|s)[u,v]| \leq \frac{8(1+\lambda\log|\mathcal{A}|))}{1-\gamma}\|u\|_\infty\|v\|_\infty\,,$$

$$|\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\mathrm{d}\pi_\theta[v]|(s) \leq \|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty\sum_{a\in\mathcal{A}}|\mathrm{d}\pi_\theta(a|s)[v]| \leq \frac{16+20\lambda\log|\mathcal{A}|}{(1-\gamma)^2}\|u\|_\infty\|v\|_\infty\,,$$

$$|\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\cdot\pi_\theta|(s) \leq \|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty\,,$$

$$|\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v]|(s) \leq (4+8\log|\mathcal{A}|)\|u\|_\infty\|v\|_\infty\,,$$

thus

$$\begin{aligned}\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty \leq{}& \gamma\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty \\ &+\left(\frac{32+40\lambda\log|\mathcal{A}|}{(1-\gamma)^2}+\frac{8(1+\lambda\log|\mathcal{A}|)}{1-\gamma}+\lambda(4+8\log|\mathcal{A}|)\right)\|u\|_\infty\|v\|_\infty\,.\end{aligned}$$

Since $|\mathcal{A}| \geq 2$, then $2\log|\mathcal{A}| \geq 1$, we can simplify it as follows

$$\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty \leq \frac{40+64\lambda\log|\mathcal{A}|}{(1-\gamma)^3}\|u\|_\infty\|v\|_\infty\,.$$

Third differential. Next, we proceed with the third differential as follows

$$\begin{aligned}\mathrm{d}^3\widetilde{V}_c^{\pi_\theta}[u,v,w] ={}& \widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w] + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[w]\cdot\mathrm{d}^2\pi_\theta[u,v] \\ &+\gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,w]\mathrm{d}\pi_\theta[v] + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\mathrm{d}^2\pi_\theta[v,w] \\ &+\gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[v,w]\mathrm{d}\pi_\theta[u] + \gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[v]\mathrm{d}^2\pi_\theta[u,w] \\ &+\gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\cdot\mathrm{d}\pi_\theta[w] + \gamma\mathsf{P}_c\mathrm{d}^3\widetilde{V}_c^{\pi_\theta}[u,v,w]\cdot\pi_\theta + \mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\,.\end{aligned}$$

By the triangle inequality

$$\begin{aligned}\|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty \leq{}& \|Q_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]\|_\infty + \gamma\|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[w]\cdot\mathrm{d}^2\pi_\theta[u,v]\|_\infty \\ &+\gamma\|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[u,w]\mathrm{d}\pi_\theta[v]\|_\infty + \gamma\|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[u]\mathrm{d}^2\pi_\theta[v,w]\|_\infty \\ &+\gamma\|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[v,w]\mathrm{d}\pi_\theta[u]\|_\infty + \gamma\|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[v]\mathrm{d}^2\pi_\theta[u,w]\|_\infty \\ &+\gamma\|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[u,v]\cdot\mathrm{d}\pi_\theta[w]\|_\infty + \gamma\|\mathsf{P}_c\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\cdot\pi_\theta\|_\infty \\ &+\|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\|_\infty\,.\end{aligned}$$

To simplify notation, let us define $R_{1,2}(u,v,w) := \|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[u,v]\mathrm{d}^2\pi_\theta[w]\|_\infty$ and $R_{2,1}(u,v,w) := \|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[u,v]\mathrm{d}\pi_\theta[w]\|_\infty$. Next, we notice that

$$\|\mathsf{P}_c\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\cdot\pi_\theta\|_\infty = \max_s\left|\sum_{s'}\pi_\theta(a|s)\mathsf{P}(s'|s,a)\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]_{s'}\right| \leq \|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty\,,$$

thus, we have a contraction argument that implies

$$\begin{aligned}\|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty \leq \frac{1}{1-\gamma}\Big(&\|Q_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]\|_\infty + \|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\|_\infty \\ &+\gamma(R_{1,2}(w,u,v)+R_{1,2}(u,v,w)+R_{1,2}(v,u,w)) \\ &+\gamma(R_{2,1}(u,w,v)+R_{2,1}(v,w,u)+R_{2,1}(u,v,w))\Big)\,.\end{aligned} \tag{71}$$

Next, we bound all terms that appear in the bound above. First, we apply Lemma G.7 for a fixed state $s \in \mathcal{S}$

$$|\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]|(s) \leq \sum_{a\in\mathcal{A}}|\widetilde{Q}_c^{\pi_\theta}(s,a)\cdot\mathrm{d}^3\pi_\theta[u,v,w](a|s)|$$

$$\leq \frac{1+\lambda \log|\mathcal{A}|}{1-\gamma} \sum_{a\in\mathcal{A}} |\mathrm{d}^3\pi_\theta[u,v,w](a|s)|$$
$$\leq \frac{48(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

Also, by Lemma G.8 we have

$$\|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\|_\infty \leq (56+48\log|\mathcal{A}|)\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

Next we bound $R_{1,2}$ as follows

$$R_{1,2}(u,v,w) = \max_{s\in\mathcal{S}}\left|\sum_{a\in\mathcal{A}}(\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u])(s,a)\mathrm{d}^2\pi_\theta[v,w](a|s)\right|$$
$$\leq \|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty \cdot \max_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left|\mathrm{d}^2\pi_\theta[v,w](a|s)\right|\,.$$

Applying the bound for the first differential as well as Lemma G.7

$$R_{1,2}(u,v,w) \leq \frac{8\cdot(8+10\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

Finally, using the same idea, we have the following bound for $R_{2,1}$:

$$R_{2,1}(u,v,w) \leq \|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty \cdot \max_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\mathrm{d}\pi_\theta[w](a|s)|$$
$$\leq \frac{2\cdot(40+64\lambda\log|\mathcal{A}|)}{(1-\gamma)^3}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

Overall, we can rewrite (71) as follows

$$\|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty \leq \frac{1}{1-\gamma}\Bigg(\frac{48(1+\lambda\log|\mathcal{A}|)}{1-\gamma} + \lambda(56+48\log|\mathcal{A}|)$$
$$+\frac{24\cdot(8+10\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + \frac{6\cdot(40+64\lambda\log|\mathcal{A}|)}{(1-\gamma)^3}\Bigg)\|u\|_\infty\|v\|_\infty\|w\|_\infty\,,$$

and, after rearranging the terms and using a bound $2\log|\mathcal{A}| \geq 1$, we have the following bound

$$\|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty \leq \frac{480+832\lambda\log|\mathcal{A}|}{(1-\gamma)^4}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

□

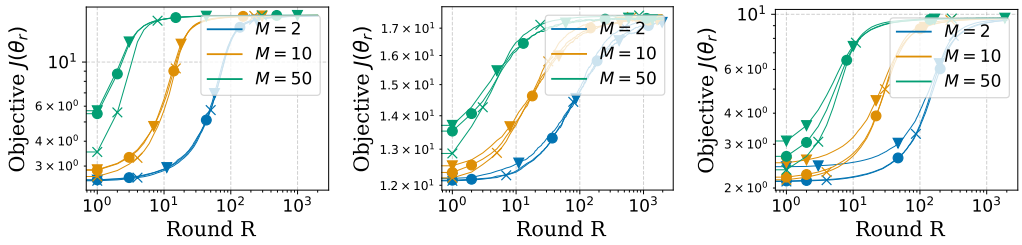
H Experiments

In this section, we provide further experimental details and additional experiments on the one described in Section 5. Experiments were conducted on a computer with an Intel Xeon 6534 and 196GB RAM.[2] We report the average over 4 runs and the standard deviation in all the plots.

We conduct experiments on synthetic and gridworld problems. In each problem, for each agent c, the transition kernel P_c is defined as a mixture of two kernels: $\mathsf{P}_c = (1-\varepsilon_{\mathsf{P}})\mathsf{P}^{\mathsf{com}} + \varepsilon_{\mathsf{P}}\mathsf{P}_c^{\mathsf{ind}}$ where $\mathsf{P}_c^{\mathsf{com}}$ is a common kernel, and $\mathsf{P}_c^{\mathsf{ind}}$ is an individual kernel specific to each agent. We now describe these kernel for the our synthetic and gridworld problems.

Synthetic. The `synthetic` environment was originally introduced by Zheng et al. (2023). In this setting, all agents share a common reward function r, where each reward value $\mathsf{r}(s,a)$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$ is independently sampled from the uniform distribution over $[0,1]$. For each (s,a), the transition kernels $\mathsf{P}^{\mathsf{com}}(\cdot \mid s,a)$ and $\mathsf{P}_c^{\mathsf{ind}}(\cdot \mid s,a)$ are drawn uniformly and randomly from the

[2] Our code is available online on GitHub: `https://github.com/Labbi-Safwan/FedPolicy-gradient`



(a) `GridWorld`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.0$ (b) `Synthetic`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.0$ (c) `GridWorld`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.3$

Figure 7: Comparison of `S-FedPG` (crosses), `RS-FedPG` (circles), and `b-RS-FedPG` (triangles): (a) Value of the global objective $J(\theta^r)$ in the `GridWorld` environment, for the three `FedPG` variants and different numbers of agents $M \in \{2, 10, 50\}$, shown on a log-log scale; (b) Value of the global objective $J(\theta^r)$ in the `Synthetic` environment, for the three `FedPG` variants and different numbers of agents $M \in \{2, 10, 50\}$, shown on a log-log scale; (c) Value of the global objective $J(\theta^r)$ in the `GridWorld` environment, for the three `FedPG` variants and different numbers of agents $M \in \{2, 10, 50\}$, shown on a log-log scale.

$|\mathcal{S}|$-dimensional simplex. The agent starts randomly from a uniformly sampled position. In the experiments with $\varepsilon_{\mathsf{P}} = 0$ or 0.3, we consider environments with $|\mathcal{S}| = 5$ states and $|\mathcal{A}| = 4$ actions.

The highly heterogeneous `synthetic` FRL instance extends the previous setup by adding two states, each reachable from one of the original five states. Once reached, these states yield a reward of $+1$ at every timestep, and agents remain there indefinitely. This instance includes two types of MDPs, differing in which high-reward state is accessible: in the first type, the first two actions deterministically lead to the rewarding state, while the last two deterministically make the agent stay in the same state; in the second type, this mapping is reversed. As a result, agents must take opposing actions, similar to Figure 4, to maximize their rewards. This conflict requires stochastic policies to ensure that, over time, all agents reach their respective high-reward states.

GridWorld. The `GridWorld` environment (Domingues et al., 2021) features an agent navigating a 3×3 grid to reach a goal state at $(2, 2)$, receiving a reward of $+1$ upon arrival and 0 otherwise. The agent can move in four directions, with intended actions succeeding with probability 0.8 under the shared dynamics $\mathsf{P}^{\mathsf{com}}$, and failing to a random neighbor with probability 0.2. The individual transition kernel $\mathsf{P}_c^{\mathsf{ind}}$ moves the agents to a neighboring cell with random probabilities that are specific to each agent. A wall at $(1, 1)$ results in $|\mathcal{S}| = 8$ reachable states. The discount factor is $\gamma = 0.95$, and the agent starts from a uniformly sampled position. We use this setup for experiments with $\varepsilon_{\mathsf{P}} = 0$ and $\varepsilon_{\mathsf{P}} = 0.3$. In the highly heterogeneous FRL instance, the target position is connected to two additional states, similarly to what has been described in the heterogeneous `synthetic` FRL instance, and does not yield any reward. Additionally, one of the two paths that leads to the previously targeted position is suppressed.

The `Fed-Q-learning` algorithm we consider corresponds to the version introduced in Jin et al. (2022), which operates under a generative model setting. Specifically, at each iteration, the algorithm updates its Q-table using $T \times B$ samples drawn uniformly from the state-action space. For `RS-FedPG` and `b-RS-FedPG`, we use a regularization temperature $\lambda = 0.05$. In addition to the results presented in the main text, we provide supplementary plots demonstrating that our proposed algorithms exhibit linear speedup in both the homogeneous and mildly heterogeneous regimes.