On Global Convergence Rates for Federated Softmax Policy Gradient Under Heterogeneous Environments

Safwan Labbi[1] **Paul Mangold**[1] **Daniil Tiapkin**[1,2] **Eric Moulines**[3,4]
[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
[4] LRE EPITA , 94270 Le Kremlin-Bicêtre, France

Abstract

We provide global convergence rates for vanilla and entropy-regularized federated softmax stochastic policy gradient (`FedPG`) with local training. We show that `FedPG` converges to a near-optimal policy in terms of the average agent value, with a gap controlled by the level of heterogeneity. Remarkably, we obtain the first convergence rates for entropy-regularized policy gradient *with explicit constants*, leveraging a projection-like operator. Our results build upon a new analysis of federated averaging for non-convex objectives, based on the observation that the Łojasiewicz-type inequalities from the single-agent setting (Mei et al., 2020) do not hold for the federated objective. This uncovers a fundamental difference between single-agent and federated reinforcement learning: while single-agent optimal policies can be deterministic, federated objectives may inherently require stochastic policies.

1 INTRODUCTION

In Federated Reinforcement Learning (FRL), multiple agents learn collaboratively by interacting with their own independent environments. As Reinforcement learning (RL) is known to be highly data-hungry (Akkaya et al., 2019), and generating such training data is very time-consuming (Nair et al., 2015), FRL constitutes a promising framework that can dramatically reduce the number of samples each agent must collect. Raw trajectories, i.e., state, action and reward sequences, are never exchanged. Instead, agents communicate intermediate computations such as policy gradients to a central server, which aggregates them to update a global policy (Qi et al., 2021; Zhuo et al., 2023; Khodadadian et al., 2022). While FRL can accelerate the training, it must overcome two important obstacles: environment heterogeneity (Jin et al., 2022) and limited communications (Zhu et al., 2022; Fan et al., 2023). Although these challenges are shared with federated learning (Kairouz et al., 2021), the solutions developed in the classical federated learning *do not generally apply to FRL*. Specifically, the convergence of `FedAVG` applied to the FRL objective under heterogeneity and local training remains poorly understood. Addressing these obstacles in this context is thus crucial for the large-scale deployment of FRL.

In this paper, we establish the first global convergence analysis of federated policy gradient methods with local training in heterogeneous environments. As such, we address a significant gap in the federated policy-gradient literature, which has essentially focused on proving convergence to first-order stationary points of the average value (Wang et al., 2024a; Jin et al., 2022). More precisely, we analyze federated softmax policy gradient (`FedPG`) with and without entropy regularization, and propose algorithmic strategies to learn stochastic stationary policies in tabular environments.

Our analysis leverages a local property of agent-specific value functions, building on recent single-agent results (Mei et al., 2020), which show that these functions satisfy a *non-uniform Łojasiewicz* condition, a generalization of gradient dominance. Remarkably, these local properties do not extend to the global federated RL objective, motivating the development of a tailored theoretical framework. As part of this framework, we provide a novel convergence analysis of `FedAVG` for non-convex objectives, which also offers new insights into the behavior of Federated Averaging

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

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*	Global convergence	Last iterate	Communication Complexity**	Sample Complexity**
PAvg (Jin et al. 2022)	✗	✗	✗	✗
FEDSVRPG-M (Wang et al. 2024a)	✗	✗	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}\left(1/(M\epsilon^{3/2})\right)$
FEDHAPG-M (Wang et al. 2024a)	✗	✗	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}\left(1/(M\epsilon^{3/2})\right)$
S-FedPG (our work)	✓	✓	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}\left(1/(M\epsilon^{3})\right)$
RS-FedPG (our work)	✓	✓	$\mathcal{O}(\log(1/\epsilon))$	$\mathcal{O}(\log(1/\epsilon)/(M\epsilon))$

* Note that all methods optimize the unregularized objective (3), except RS-FedPG, which optimizes the entropy-regularized objective (4); ** For methods that do not enjoy global convergence, the reported sample and communication complexities correspond to finding an ϵ-stationary point of the objective F, i.e., a parameter θ such that $\mathbb{E}\left[\|\nabla F(\theta)\|^2\right] \leq \epsilon$. In contrast, for our methods S-FedPG and RS-FedPG, the complexities are stated for obtaining an ϵ-optimal solution, i.e. a θ such that $\mathbb{E}[F^\star - F(\theta)] \leq \epsilon$. These guarantees hold for any target accuracy $\epsilon > \epsilon_{\min}$, where $\epsilon_{\min}$ is the heterogeneity floor (equal to 0 in the homogeneous case); see Corollaries 5.1 and 5.3 for the exact expressions.

under non-uniform Łojasiewicz conditions.

Finally, we show that the differences between single-agent RL and heterogeneous FRL are intrinsic to FRL: they are not artifacts of gradient methods nor of our analysis. Specifically, we show that heterogeneity breaks classical RL properties, as the optimal *common* policy can be inherently *stochastic* or even *non-stationary*. This contrasts with the single-agent case, where there always exists an optimal deterministic and stationary policy (Agarwal et al., 2019).

Our contributions can be summarized as follows:

- We provide a novel analysis of federated averaging for objectives that satisfy a *local* Łojasiewicz-type conditions, showing how client-side regularity can be leveraged even when the global federated objective fails to satisfy any Łojasiewicz inequality.
- We present the *first* global convergence guarantees for entropy-regularized policy gradients in heterogeneous FRL. Exploiting the non-uniform Łojasiewicz property of the *local* objective, we show that FedPG converges to near-optimal policies and achieves linear speed-up in the number of agents.
- We reveal fundamental gaps between federated and classical RL, motivating our new analytical framework. A surprising observation is that, unlike in classical RL, optimal federated policies can be non-deterministic or time-varying.
- We validate the theory on two FRL benchmarks, showing the predicted scaling of FedPG with heterogeneity and robust empirical performance.

We compare to prior work in Table 1, review related literature in Section 2, introduce the problem setting in Section 3, present our main results in Section 4 and Section 5, describe specific FRL properties in Section 6, and provide experiments in Section 7.

2 RELATED WORK

Policy Gradient Methods. Policy gradient (PG) methods (Williams, 1992; Sutton et al., 1999) are well understood in tabular, single-agent discounted RL. For softmax policies with exact gradients, recent analyses characterize the optimization landscape of RL via Łojasiewicz-type inequalities, establishing global convergence with sublinear rates, and linear rates with entropy regularization (Mei et al., 2020; Zhang et al., 2020; Xiao, 2022; Agarwal et al., 2020). Stochastic PG is subtler: early results proved convergence to first-order stationary points (Zhang et al., 2021b,a; Yuan et al., 2022), and later work clarified when deterministic and stochastic updates align to recover global guarantees (Mei et al., 2021; Ding et al., 2025; Wang et al., 2026; Labbi et al., 2026).

Federated RL. FRL theory literature is growing fast (Zhuo et al., 2023). Under homogeneous dynamics, federated Q-learning variants have been shown to reduce sample complexity (Salgia and Chi, 2024; Zheng et al., 2025; Jin et al., 2022). With heterogeneous dynamics, non-asymptotic analyses reveal inherent trade-offs: collaboration gives speedups, but with an unavoidable bias scaling with heterogeneity (Wang et al., 2024b; Zhang et al., 2024; Labbi et al., 2025; Mangold et al., 2025). A notable exception is for federated policy evaluation, where such bias can be mitigated using control variate-type methods Mangold et al. (2024). Variants of federated PG have been analyzed in homogeneous settings, with global convergence and improved communication (Lan et al., 2023; Ganesh et al., 2024; Wang et al., 2024a), but rely on strong assumptions. In contrast, we derive global con-

vergence guarantees for federated PG with heterogeneous agents, stochastic gradients, and classical RL assumptions.

`FedAVG` under PL. Non-convex federated optimization under Polyak–Łojasiewicz (PL) conditions has received limited attention. Haddadpour and Mahdavi (2019) proved `FedAVG`'s convergence with deterministic gradients and controlled gradient diversity, while Haddadpour et al. (2019) proved linear speedups in the stochastic setting. A more recent analysis (Demidovich et al., 2025) still requires the PL condition on the global objective. Unfortunately, such a PL condition does not hold for the global objective with heterogeneous agents. By contrast, our analysis relies on *local* client objectives' regularity, showing that local PL-like structure can be leveraged even when the *global* objective fails to satisfy a Łojasiewicz-type property.

3 PRELIMINARIES

Problem Setting. We consider a FRL setting with M agents, where each agent $c \in [M]$ has its own independent Markov Decision Process (MDP), $\mathcal{M}_c \triangleq (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_c, \mathsf{r}_c, \rho)$, with a state space $\mathcal{S}$, an action space $\mathcal{A}$, and discount factor $\gamma < 1$, but distinct rewards r_c and transition kernels P_c. Following common practice in FRL, we define kernel and reward heterogeneity as

$$\varepsilon_{\mathsf{P}} \triangleq \max_{s,a \in \mathcal{S}\times\mathcal{A}} \max_{c,c' \in [M]^{\otimes 2}} \|\mathsf{P}_c(\cdot|s,a) - \mathsf{P}_{c'}(\cdot|s,a)\|_1 \ , \quad (1)$$

$$\varepsilon_{\mathsf{r}} \triangleq \max_{c,c' \in [M]^{\otimes 2}} \|\mathsf{r}_c - \mathsf{r}_{c'}\|_\infty \ . \quad (2)$$

We consider policies with *softmax* parameterization

$$\pi_\theta(a|s) \triangleq \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))} \text{ for } \theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \ , \quad (3)$$

and aim to minimize the averaged objective

$$\max_{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}} J(\theta) \triangleq \tfrac{1}{M} \textstyle\sum_{c=1}^M J_c(\theta) \ , \quad (4)$$

$$\text{where } J_c(\theta) \triangleq \mathbb{E}^{\pi_\theta}_{c,\rho}\left[\textstyle\sum_{t=0}^\infty \gamma^t \mathsf{r}_c(S_c^t, A_c^t)\right] \ , \quad (5)$$

and where $\mathbb{E}^{\pi_\theta}_{c,\rho}[\cdot]$ is the expectation over random trajectories generated by following the softmax policy π_θ parametrized by θ: the initial state is sampled from a distribution $S_c^0 \sim \rho(\cdot)$ and for all $t \geq 0$: $A_c^t \sim \pi_\theta(\cdot|S_c^t), S_c^{t+1} \sim \mathsf{P}_c(\cdot|S_c^t, A_c^t)$. We define

$$J_c^\star \triangleq \sup_{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}} J_c(\theta) \ , \quad J^\star \triangleq \tfrac{1}{M} \textstyle\sum_{c=1}^M J_c^\star \ , \quad (6)$$

the maximum value and its average over all agents. Similarly, we define the entropy-regularized FRL objective, for a temperature $\lambda > 0$, as

$$\max_{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}} \tilde{J}_\lambda(\theta) \triangleq \tfrac{1}{M} \textstyle\sum_{c=1}^M \tilde{J}_{c,\lambda}(\theta) \ ,$$

with $\tilde{J}_{c,\lambda}(\theta) \triangleq \mathbb{E}^{\pi_\theta}_{c,\rho}\left[\sum_{t=0}^\infty \gamma^t \big(\mathsf{r}_c(S_c^t, A_c^t) - \lambda h_\theta(A_c^t, S_c^t)\big)\right]$ and $h_\theta(A_c^t, S_c^t) \triangleq \log(\pi_\theta(A_c^t|S_c^t))$. Finally, let

$$\tilde{J}_{c,\lambda}^\star \triangleq \sup_{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}} \tilde{J}_{c,\lambda}(\theta), \quad \tilde{J}_\lambda^\star \triangleq \tfrac{1}{M} \textstyle\sum_{c=1}^M \tilde{J}_{c,\lambda}^\star \ , \quad (7)$$

be the maximum value and its average over agents.

Single-Agent Regularity. Taking $M = 1$, we have $J^\star = J_1^\star$ and $\tilde{J}_\lambda^\star = \tilde{J}_{1,\lambda}^\star$. In this setting, for a stationary policy π, we define the discounted state occupancy

$$d_\rho^\pi(s) \triangleq (1-\gamma) \textstyle\sum_{t=0}^\infty \gamma^t \cdot \rho \mathsf{P}_\pi^t(s) \ , \quad (8)$$

where $\mathsf{P}_\pi(s'|s) \triangleq \sum_{a \in \mathcal{A}} \pi(a|s) \mathsf{P}(s'|s,a)$. We also assume the initial state distribution ρ has strictly positive coefficients (see Section 5.1).

(Unregularized Case). Under these assumptions, in the single-agent setting, Mei et al. (2020) proved that J is smooth with constant $L_{2,s} \triangleq 8/(1-\gamma)^3$ (see its Lemma 7). Moreover, J satisfies a (non-uniform) *Łojasiewicz inequality* (Lemma 8 in Mei et al. 2020)

$$\|\nabla J(\theta)\|_2^2 \ \geq \ 2\,\mu_s(\theta)\left[J^\star - J(\theta)\right]^2 \text{ for } \forall \theta \in \Theta, \quad (9)$$

with $\mu_s(\theta) \triangleq \min_s \pi_\theta(a^\star(s)|s)^2 \cdot \left\|d_\rho^{\pi^\star}/d_\rho^{\pi_\theta}\right\|_\infty^{-2} / (2|\mathcal{S}|)$.

(Regularized Case). Under the above assumptions, in the single-agent setting, Mei et al. (2020) proved that $\tilde{J}_\lambda$ is $\tilde{L}_{2,\lambda}$-smooth with $\tilde{L}_{2,\lambda} \triangleq (8 + \lambda(4 + 8\log(|\mathcal{A}|))/(1-\gamma)^3$. Moreover, the regularized objective satisfies a stronger non-uniform Łojasiewicz inequality (Williams and Peng, 1991; Mnih et al., 2016; Schulman et al., 2017; Ahmed et al., 2019), with a *linear* suboptimality gap,

$$\|\nabla \tilde{J}_\lambda(\theta)\|_2^2 \geq 2\tilde{\mu}_\lambda(\theta)\left[\tilde{J}_\lambda^\star - \tilde{J}_\lambda(\theta)\right] \ , \quad (10)$$

for $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, with

$$\tilde{\mu}_\lambda(\theta) \triangleq \frac{\lambda \min_s d_\rho^{\pi_\theta}(s) \min_{s,a} \pi_\theta(a|s)^2}{|\mathcal{S}|(1-\gamma)} \left\|\frac{d_\rho^{\pi_\lambda^\star}}{d_\rho^{\pi_\theta}}\right\|_\infty^{-1} \ .$$

FRL (Non-)Regularity. From the single-agent case, for any agent $c \in [M]$, J_c, and $\tilde{J}_{c,\lambda}$ are smooth (hence J and $\tilde{J}_\lambda$ are smooth). Furthermore, we show in Section 5.1 that for any $c \in [M]$, local functions J_c and $\tilde{J}_{c,\lambda}$ satisfy non-uniform Łojasiewicz inequalities. Unfortunately, averaging such functions *does not preserve* such non-uniform Łojasiewicz property in general; the federated objective J (or $\tilde{J}_\lambda$), even when each client objective enjoys single-agent geometry (see Lemma D.9 where we provide a counter example). Consequently, analyses of federated averaging methods leveraging global Łojasiewicz conditions (e.g.,

Demidovich et al. 2025) cannot be directly applied in our context. We remedy this problem by proposing a novel analytical framework for FedAVG-style methods, tailored to global objectives that are sums of locally Łojasiewicz functions, as arise in heterogeneous FRL.

4 FEDAVG UNDER LOJASIEWICZ CONDITIONS

Federated Averaging. In this section, we provide convergence bounds on a general class of distributed non-convex optimization problems of the form

$$\max_{\theta\in\mathbb{R}^d} F(\theta) \triangleq \frac{1}{M}\sum_{c=1}^{M} f_c(\theta), \quad f_c(\theta) \triangleq \mathbb{E}_{Z_c}[f_c^{Z_c}(\theta)], \quad (11)$$

where each Z_c is a random variable sampled 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. Each function f_c is only available to the client c through *biased* stochastic gradients $\mathrm{g}_c^z(\theta)$, whose expected value is

$$\mathrm{g}_c(\theta) \triangleq \mathbb{E}_{Z_c\sim\xi_c(\theta)}[\mathrm{g}_c^{Z_c}(\theta)] \ , \quad (12)$$

but is typically different from the gradient of f_c. To solve (11), we use `proj-FedAVG`, a variant of federated averaging with a projection-like *improvement operator* (see Algorithm 1). Each communication round of `proj-FedAVG` 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:

$$\theta_c^{r,h+1} = \theta_c^{r,h} + \eta\cdot \mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h}), \quad \theta_c^{r,0}=\theta^r \quad (13)$$

where $\eta>0$ is a learning rate shared by the agents, and the $Z_c^{r,h}$ for $c\in[M]$, $r\in[R]$, and $h\in[H]$ are independent from an agent to another, and are a martingale with respect to the filtration

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

After H local steps, parameters are averaged as $\bar\theta^{r+1} = \frac{1}{M}\sum_{c=1}^M \theta_c^{r,H}$, and followed by a projection-like step $\theta^{r+1}=\mathcal{T}(\bar\theta^{r+1})$, where $\mathcal{T}\colon\mathbb{R}^d\to\mathbb{R}^d$.

Descent Lemma. We start by establishing a descent lemma, under the following assumption. We stress that this assumption is not restrictive, as we will show in Section 5 that this assumption is satisfied by PG methods under a standard RL assumption.

A-1. *For all $c\in[M]$, f_c is three times differentiable, and there exists $L_1, L_2, L_3, \zeta, \beta, \sigma_2^2, \sigma_4^4 \geq 0$ such that*

Algorithm 1 `proj-FedAVG`

Initialization: Learning rate $\eta>0$, parameter θ^0, Improvement operator $\mathcal{T}$
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: $\theta^{r+1}=\mathcal{T}(\bar\theta^{r+1})$ where $\bar\theta^{r+1}=\frac{1}{M}\sum_{c=1}^M\theta_c^{r,H}$

1. *Smoothness: for any $(\theta,u)\in\mathbb{R}^d\times\mathbb{R}^d$,*
$$\|\nabla f_c(\theta)\|\leq L_1 \ , \quad \|\nabla^2 f_c(\theta)u\|\leq L_2\|u\| \ , \quad \|\nabla^3 f_c(\theta)u^{\otimes 2}\|\leq L_3\|u\|^2 \ ,$$
and $\mathrm{g}_c(\theta)=\mathbb{E}_{Z_c\sim\nu_c(\theta)}[\mathrm{g}_c^{Z_c}(\theta)]$ is L_2-Lipschitz.
2. *Heterogeneity: $\|\nabla F(\theta)-\nabla f_c(\theta)\|_2\leq\zeta$*
3. *Bias and variance: for any $\theta\in\mathbb{R}^d$,*
$$\|\nabla f_c(\theta)-\mathrm{g}_c(\theta)\|_2\leq\beta \, ,$$
$$\mathbb{E}_{Z_c\sim\nu_c(\theta)}[\|\mathrm{g}_c^{Z_c}(\theta)-\mathrm{g}_c(\theta)\|_2^p]\leq\sigma_p^p \ , \ \text{for } p\in\{2,4\} \ .$$

Assumption A-1 captures standard regularity conditions on the local objectives, such as smoothness, bounded gradient bias and variance, and bounded gradient heterogeneity, which are commonly used in federated learning. Under this assumption, we derive the following descent lemma for non-convex objectives.

Lemma 4.1. *Assume A-1. Then, for any $\eta>0$ such that $\eta H L_2\leq 1/6$ and $32\eta^2H^2L_3^2L_1^2\leq L_2^2$, the iterates of `proj-FedAVG` satisfy*

$$F(\theta^r)-\mathbb{E}\left[F(\bar\theta^{r+1})\middle|\mathcal{F}^r\right]\lesssim -\tfrac{\eta H}{4}\|\nabla F(\theta^r)\|_2^2+\tfrac{\eta^2L_2H\sigma_2^2}{M} + \eta H\beta^2+\eta^3L_2^2H^3\zeta^2+\eta^5L_3^2H^3\sigma_4^4 \ .$$

Sketch of proof. Let $\kappa=\frac{1}{\sqrt{\eta H}}$. Using the L_2-smoothness of each f_c, taking the expectation conditionally on $\mathcal{F}^r$ and using the 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

$$-\mathbb{E}\left[F(\bar\theta^{r+1})\middle|\mathcal{F}^r\right]+F(\theta^r)\leq -\tfrac{1}{2\kappa^2}\|\nabla F(\theta^r)\|_2^2 + \underbrace{\tfrac{1}{2\kappa^2}\|\nabla F(\theta^r)+\kappa^2\mathbb{E}\left[\theta^r-\bar\theta^{r+1}\middle|\mathcal{F}^r\right]\|_2^2}_{(\mathbf{A})} + \underbrace{\tfrac{L_2}{2}\mathbb{E}\left[\|\bar\theta^{r+1}-\theta^r\|_2^2\middle|\mathcal{F}^r\right]-\tfrac{\kappa^2}{2}\|\mathbb{E}\left[\bar\theta^{r+1}-\theta^r\middle|\mathcal{F}^r\right]\|_2^2}_{(\mathbf{B})} .$$

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

Bounding (A). Using Jensen's inequality, combined with Young's inequality and the bound on the bias of the stochastic estimator (A-1), we get

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

The term on the right-hand side captures the expected drift induced by local updates under client heterogeneity. We bound it via a third-order Taylor expansion combined with Burkholder's inequality (Theorem 8.6 in Osękowski, 2012); see Lemma B.1 for details, which yields

$$(\mathbf{A}) \lesssim \frac{\eta^3 L_2^2 H^2 (H-1)}{M} \sum_{c=1}^{M} \|\nabla f_c(\theta^r)\|_2^2 + L_3^2 \eta^5 H^2 (H-1)\sigma_4^4 + (1 + \eta^2 L_2^2 H(H-1))\eta H \beta^2 \ .$$

Bounding (B). We decompose (**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

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

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

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

Combining $\frac{1}{M}\sum_{c=1}^{M} \|\nabla f_c(\theta^r)\|^2 \leq \|\nabla F(\theta^r)\|^2 + \zeta^2$, with the bounds on (**A**) and (**B**), concludes the proof □

We provide a full statement and proof of this lemma in Lemma B.3. This result generalizes Theorem 4.2 of Glasgow et al. (2022) to biased oracles and relaxes the previously required pointwise third-order smoothness to only in-expectation third-order smoothness. In the following, we will use the following quantities

$$f_c^\star \triangleq \sup_{\theta \in \mathbb{R}^d} f_c(\theta) \ , \quad F^\star \triangleq \frac{1}{M}\sum_{c=1}^{M} f_c^\star \ . \tag{14}$$

We now apply Lemma 4.1 to two cases where the local functions satisfy a type of non-uniform Łojasiewicz condition. Using this approach, we will derive global convergence bounds, thereby generalizing the analyses of Glasgow et al. (2022) and Demidovich et al. (2025).

Quadratic Łojasiewicz (QL) inequality. For any $(c, \theta) \in [M] \times \mathbb{R}^d$, we define the non-uniform "quadratic-type" Łojasiewicz constant by

$$\mu_c(\theta) := \sup\{x \in \mathbb{R}^+, \|\nabla f_c(\theta)\|_2^2 \geq 2x \left(f_c^\star - f_c(\theta)\right)^2\}.$$

We first prove the convergence of `proj-FedAVG` under A-1 and the following two assumptions.

QL-1. *For any* $(c, \theta) \in [M] \times \mathbb{R}^d$, $\mu_c(\theta) > 0$.

QL-2. *There exists* $\underline{\mu} > 0$, *such that for any* $\theta \in \mathbb{R}^d$, *we have* $\min_{c \in [M]} \mu_c(\mathcal{T}(\theta)) \geq \underline{\mu}$ *and* $F(\mathcal{T}(\theta)) \geq F(\theta)$.

Assumption **QL-** 1 characterizes the landscape of each local objective by imposing a weaker, *quadratic* Łojasiewicz inequality, as opposed to the standard Polyak–Łojasiewicz condition Polyak (1963). Such structure arises in unregularized FRL; see (9). Assumption **QL**-2 ensures that applying $\mathcal{T}$ increases the value of the objective and keeps the iterates away from regions where the objective is ill-conditioned.

Theorem 4.2 (Convergence rates of `proj-FedAVG`)**.** *Assume A-1, **QL-** 1 and **QL-** 2. 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*

$$\mathbb{E}[F^\star - F(\theta^R)] \lesssim \frac{F^\star - F(\theta^0)}{1 + R \cdot (F^\star - F(\theta^0))\eta H \underline{\mu}} + \frac{\zeta + \beta}{\sqrt{\underline{\mu}}} + \frac{\zeta^2 + \beta^2}{L_2} + \sqrt{\frac{\eta L_2 \sigma_2^2}{M \underline{\mu}}} + \frac{\eta \sigma_2^2}{M} + \frac{\eta^2 L_3 H \sigma_4^2}{\sqrt{\underline{\mu}}} + \frac{\eta^4 L_3^2 H^2 \sigma_4^4}{L_2} \ .$$

Sketch of proof. Firstly, using **QL-** 1, we have for any $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$

$$\sqrt{\min_{c \in [M]} 2\mu_c(\theta)} \left[f_c^\star - f_c(\theta)\right] \leq \|\nabla f_c(\theta)\|_2 \ .$$

We then decompose $\nabla f_c(\theta) = \nabla f_c(\theta) - \nabla F(\theta) + \nabla F(\theta)$ and use triangle inequality and A-1 to bound

$$\|\nabla f_c(\theta)\|_2 \leq \zeta + \|\nabla F(\theta)\|_2 \ .$$

Averaging this inequality over all the agents, taking the square, and applying Young's inequality allows to derive a quasi-QL inequality for the global objective:

$$\zeta^2 + \|\nabla F(\theta)\|_2^2 \geq \min_{c \in [M]} \mu_c(\theta) (F^\star - F(\theta))^2. \tag{15}$$

Using **QL** 2, we obtain $\min_{r \geq 0} \min_{c \in [M]} \mu_c(\theta^r) \geq \underline{\mu}$, as $\theta^r = \mathcal{T}(\bar{\theta}^r)$. Next, using the fact that $F(\theta^{r+1}) \geq F(\bar{\theta}^{r+1})$ in (15), plugging the result in Lemma 4.1, and unrolling the recursion gives the result. □

We prove this theorem in Appendix B.2. This theorem shows that under the non-uniform Łojasiewicz inequality, the averaged objective's optimality gap decays sublinearly, converging to a residual floor determined by the gradient bias β, the heterogeneity ζ, and stochastic errors of order σ_2^2/M, which decrease linearly with the number of agents M (linear speed-up). Higher-order contributions scale as η^2. The homogeneous setting is recovered by setting $\zeta = 0$, while unbiased gradients correspond to $\beta = 0$. In the unbiased homogeneous case, the residual floor disappears. We now present the sample and communication complexity result.

Corollary 4.3 ((Simplified) Sample and Communication Complexity)**.** *Under the assumptions of Theorem 4.2, let* $\epsilon \gtrsim \frac{\zeta}{\underline{\mu}^{1/2}} + \frac{\beta}{\underline{\mu}^{1/2}} + \frac{\zeta^2}{2L_2} + \frac{\beta^2}{L_2}$, *and*

$\eta \leq \min\left(\frac{1}{6L_2}, \frac{\underline{\mu} M\epsilon^2}{216L_2\sigma_2^2}, \frac{\underline{\mu}^{1/2}\epsilon L_2}{13^2 L_3\sigma_4^2}, \frac{2\epsilon M}{\sigma_2^2}, \frac{\epsilon^{1/2}L_2^{3/2}}{24L_3\sigma_4^2}\right)$, *then* proj-FedAVG *achieves* $\mathbb{E}[F^\star - F(\theta^R)] \leq \epsilon$, *with*

$$R \gtrsim \frac{[F^\star - F(\theta^0) - \epsilon]}{(F^\star - F(\theta^0))\underline{\mu}\epsilon} \cdot \max\left(L_2, \frac{L_3 L_1}{L_2}\right) ,$$

communications and a number of samples per agent of

$$RH \gtrsim \frac{[F^\star - F(\theta^0) - \epsilon]}{(F^\star - F(\theta^0))\underline{\mu}\epsilon} \max\left(L_2, \frac{L_2\sigma_2^2}{\underline{\mu} M\epsilon^2}, \frac{L_3\sigma_4^2}{\mu^{1/2}\epsilon L_2}, \frac{L_3\sigma_4^2}{\epsilon^{1/2}L_2^{3/2}}\right).$$

See Appendix B.2 for a proof of this corollary. This result shows that FedAVG achieves linear speedup in the number of agents, provided that desired precision is not too large. Moreover, the number of communication rounds scales with $O(1/\epsilon)$.

Polyak-Łojasiewicz (PL) inequality. For any $(c, \theta) \in [M] \times \mathbb{R}^d$, we define the non-uniform Polyak-Łojasiewicz constant by:

$$\tilde{\mu}_c(\theta) := \sup\{x \in \mathbb{R}^+, \|\nabla f_c(\theta)\|_2^2 \geq 2x\,(f_c^\star - f_c(\theta))\}.$$

Next, we assume the following PL-type conditions.

PL-1. *For any* $(c, \theta) \in [M] \times \mathbb{R}^d$, $\tilde{\mu}_c(\theta) > 0$.

PL-2. *There exists* $\underline{\tilde{\mu}} > 0$, *such that for any* $\theta \in \mathbb{R}^d$, *we have* $\min_{c\in[M]} \tilde{\mu}_c(\mathcal{T}(\theta)) \geq \underline{\tilde{\mu}}$ *and* $F(\mathcal{T}(\theta)) \geq F(\theta)$.

Assumption **PL**-1 is closer to the standard Polyak–Łojasiewicz condition Polyak (1963), yet it remains more general since the Łojasiewicz coefficient is allowed to depend on θ, which can create highly ill-conditioned regions. We avoid such regions by assuming that the operator $\mathcal{T}$ confines the iterates of proj-FedAVG to a well-conditioned set; see **PL**-2.

Theorem 4.4 (Convergence rates of proj-FedAVG)**.** *Assume A-1, **PL**-1 and **PL**-2. 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*

$$\mathbb{E}\left[F^\star - F(\theta^R)\right] \lesssim \left(1 - \frac{\eta H\underline{\tilde{\mu}}}{2}\right)^R \cdot (F^\star - F(\theta^0)) + \frac{\zeta^2+\beta^2}{\underline{\tilde{\mu}}} + \frac{\eta L_2\sigma_2^2}{M\underline{\tilde{\mu}}} + \frac{\eta^4 L_3^2 H^2\sigma_4^4}{\underline{\tilde{\mu}} L_2} .$$

We prove this theorem in Appendix B.2. This theorem shows that FedAVG converges faster under these assumptions, although a residual floor term remains, depending on the gradient bias and heterogeneity. We now give a sample and complexity result.

Corollary 4.5 (Sample and Communication Complexity of proj-FedAVG)**.** *Under the assumptions of Theorem 4.4, let* $\epsilon > 4\zeta^2/\underline{\tilde{\mu}} + 16\beta^2/\underline{\tilde{\mu}}$ *and* $\eta \leq \min(\frac{1}{6L_2}, \frac{\underline{\tilde{\mu}}\epsilon M}{12L_2\sigma_2^2}, \frac{\underline{\tilde{\mu}}^{1/2}L_2^{3/2}\epsilon^{1/2}}{5L_3\sigma_4^2})$, *Then* proj-FedAVG *achieves* $\mathbb{E}[F^\star - F(\theta^R)] \leq \epsilon$, *with*

$$R \gtrsim \frac{L_2}{\underline{\tilde{\mu}}} \max\left(1, \frac{L_3 L_1}{L_2^2}\right) \log\left(\frac{4(F^\star - F(\theta^0))}{\epsilon}\right) ,$$

Algorithm 2 (S, RS)-FedPG

Initialization: Learning rate $\eta > 0$, parameter θ^0, improvement projector $\mathcal{T}$
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: $Z_c^{r,h+1} \triangleq (S_{c,b}^{r,h,1:T}, A_{c,b}^{r,h,1:T})_{b=1}^B$ 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 (16) for S-FedPG, and (17) for RS-FedPG.
    Server updates parameter: $\theta^{r+1} = \mathcal{T}(\bar{\theta}^{r+1})$ where $\bar{\theta}^{r+1} = \frac{1}{M}\sum_{c=1}^M \theta_c^{r,H}$

communications and a number of samples per agent of

$$RH \gtrsim \frac{L_2}{\underline{\tilde{\mu}}} \max\left(1, \frac{12L_2\sigma_2^2}{\underline{\tilde{\mu}}\epsilon M}, \frac{5L_3\sigma_4^2}{\underline{\tilde{\mu}}^{1/2}L_2^{5/2}\epsilon^{1/2}}\right) \log\left(\frac{4\Delta^0}{\epsilon}\right) ,$$

where $\Delta^0 = F^\star - F(\theta^0)$.

We prove this corollary in Appendix B.2. As in the previous case, this result proves that FedAVG converges with $O(\log(1/\epsilon))$ communication rounds, with linear speedup in the number of agents, up to higher-order terms. Next, we apply these results to federated policy gradient with heterogeneous agents.

5 ANALYSIS OF FEDPG

We introduce two federated extensions of policy gradient, S-FedPG and RS-FedPG, designed for (4) and (3), respectively (see Mei et al., 2020; Agarwal et al., 2021). These algorithms can be viewed as particular instances of the general proj-FedAVG framework (see Section 4). Both S-FedPG and RS-FedPG leverage a REINFORCE-like estimator (Williams, 1992) that uses a batch of independent B trajectories of length T. For completeness, the pseudo-code of S-FedPG and RS-FedPG, are provided in Algorithm 2. Next, we check that A-1, **QL**-1 and **QL**-2 holds for S-FedPG, and A-1,**PL**-1 and **PL**-2, holds for RS-FedPG. All proofs of subsequent results are carried in Appendix C and Appendix D.

5.1 Convergence Analysis of S-FedPG

For a batch of B trajectories $z \in (\mathcal{S} \times \mathcal{A})^{T\cdot B}$, the REINFORCE estimator is

$$\mathrm{g}_{c,\mathrm{s}}^z(\theta) \triangleq \frac{1}{B}\sum_{b=1}^B \sum_{t=0}^{T-1} \gamma^t \times \left(\sum_{\ell=0}^t \nabla \log \pi_\theta(a_b^\ell \mid s_b^\ell)\right) \mathsf{r}_c(s_b^t, a_b^t) . \quad (16)$$

Condition A-1 holds with the following constants $L_1 \simeq (1-\gamma)^{-2}$, $L_2 \simeq (1-\gamma)^{-3}$, $L_3 \simeq (1-\gamma)^{-4}$, $\zeta^2 \simeq \varepsilon_{\mathsf{P}}^2(1-\gamma)^{-6} + \varepsilon_{\mathsf{r}}^2(1-\gamma)^{-4}$, $\beta \simeq \frac{\gamma^T T}{1-\gamma} + \frac{\gamma^T}{(1-\gamma)^2}$, $\sigma_2^2 \simeq (1-\gamma)^{-4}B^{-1}$, and $\sigma_4^4 \simeq (1-\gamma)^{-8}B^{-2}$. Consider the sufficient exploration condition

Assumption $\mathbf{A}_\rho$. *ρ satisfies* $\rho_{\min} \triangleq \min_{s\in\mathcal{S}} \rho(s) > 0$.

Under $\mathbf{A}_\rho$, and assuming the uniform minorization condition $\underline{\mu_{\mathsf{s}}} \in (0,1)$ such that $\inf_{r\in\mathbb{N}} \mu_{\mathsf{s}}(\theta^r) \geq \underline{\mu_{\mathsf{s}}}$ almost surely, assumptions **QL**-1 and **QL**-2 are satisfied. While restrictive, this condition is unavoidable in our setting, as we neither employ projection methods nor consider entropy-regularized objectives. Similar requirements arise in the non-federated case ($M = 1$), e.g., (Lu et al., 2024, Theorem 3). We will later show that entropy regularization removes this requirement.

Applying Corollary 4.3, we obtain the following sample complexity and Communication Complexity bound:

Corollary 5.1 ((Simplified) Sample and Communication Complexity)**.** *Assume $\mathbf{A}_\rho$ and set $\mathcal{T} = \mathrm{Id}$. Additionally, assume that there exists $1 > \underline{\mu_{\mathsf{s}}} > 0$ such that $\inf_{r\in[\mathbb{N}]} \mu_{\mathsf{s}}(\theta^r) \geq \underline{\mu_{\mathsf{s}}}$, for any , $T \geq 4(1-\gamma)^{-2}$, and $M \cdot B \geq (1-\gamma)^{-1}$. Let $\epsilon \gtrsim \frac{\varepsilon_{\mathsf{P}}}{(1-\gamma)^3 \underline{\mu_{\mathsf{s}}}^{1/2}} + \frac{\varepsilon_{\mathsf{r}}}{(1-\gamma)^2\underline{\mu_{\mathsf{s}}}^{1/2}} + \frac{\gamma^T T}{(1-\gamma)\underline{\mu_{\mathsf{s}}}^{1/2}}$ and $\eta \lesssim \min\Big((1-\gamma)^3, (1-\gamma)^7\underline{\mu_{\mathsf{s}}}BM\epsilon^2, \underline{\mu_{\mathsf{s}}}^{1/2}\epsilon(1-\gamma)^5 B\Big)$. In this case S-FedPG achieves $\mathbb{E}[J^\star - J(\theta^R)] \leq \epsilon$, with a number of communication*

$$R \gtrsim \frac{[J^\star - J(\theta^0) - \epsilon/6]}{(J^\star - J(\theta^0))\underline{\mu_{\mathsf{s}}}\epsilon} \cdot \frac{1}{(1-\gamma)^3} \quad ,$$

for a total number of sampled trajectories per agent of

$$RHB \gtrsim \frac{[J^\star - J(\theta^0) - \epsilon/6]}{(J^\star - J(\theta^0))\underline{\mu_{\mathsf{s}}}\epsilon} \max\left(\frac{B}{(1-\gamma)^3}, \frac{(1-\gamma)^{-7}}{\underline{\mu_{\mathsf{s}}}M\epsilon^2}, \frac{(1-\gamma)^{-5}}{\underline{\mu_{\mathsf{s}}}^{1/2}\epsilon}\right).$$

This shows that S-FedPG has linear speedup as long as $M \lesssim \min\left(\frac{1}{\underline{\mu_{\mathsf{s}}}^{1/2}\epsilon(1-\gamma)^2}, \frac{1}{\underline{\mu_{\mathsf{s}}}B(1-\gamma)^4\epsilon^2}\right)$.

5.2 Convergence Analysis of RS-FedPG

The stochastic gradient estimator of RS-FedPG is given by (16), with the distinction that the reward is replaced with the entropy penalized reward

$$\mathrm{g}_{c,\mathsf{s}}^z(\theta) \triangleq \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t\Big(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(a_b^\ell|s_b^\ell)\Big) \times \big[\mathsf{r}_c(s_b^t,a_b^t) - \lambda\log(\pi_\theta(a_b^t|s_b^t))\big]. \tag{17}$$

We introduce on the central server side a projection-like operator $\mathcal{T}_\tau$, analogous to the projection step in projected gradient descent (Bertsekas, 2003). This operator restricts optimization to a region of interest by excluding policies with excessively large entropy penalization. For a policy π and $s \in \mathcal{S}$, define $a_{\max}^\pi(s) = \arg\max_{a\in\mathcal{A}}\{\pi(a|s)\}$, choosing at random in the arg max in case of ties. For $0 < \tau < 1/(2|\mathcal{A}|^2)$, set

$$\mathcal{A}_\tau^\pi(s) \triangleq \{a \in \mathcal{A}, \pi(a|s) \leq \tau/2\} \quad .$$

We define the operator $\mathcal{U}_\tau$ which acts as a projection in the policy space as follows: for each $(s,a) \in \mathcal{S}\times\mathcal{A}$: $\mathcal{U}_\tau(\pi)(a|s) = \tau$, if $\pi(a|s) \leq \tau/2$, $\mathcal{U}_\tau(\pi)(a|s) = \pi(a|s) - \sum_{b\in\mathcal{A}_\tau^\pi(s)}(\tau - \pi(b|s))$, if $a = a_{\max}^\pi(s)$, $\mathcal{U}_\tau(\pi)(a|s)\pi(a|s)$, otherwise. This operator prevents policies from becoming too deterministic: for any $s,a \in \mathcal{S}\times\mathcal{A}$, if $\pi(a|s)$ approaches zero, it is raised above a τ-dependent threshold. The operator $\mathcal{U}_\tau$ acts in policy space; we denote by $\mathcal{T}_\tau$ the associated operator in the logit space, i.e. for all θ, $\pi_{\mathcal{T}_\tau(\theta)} \triangleq \mathcal{U}_\tau(\pi_\theta)$ (see Appendix D.2 for details). With a suitable choice of τ, applying $\mathcal{T}_\tau$ yields logits with a higher regularized value.

Lemma 5.2. *Assume that ρ satisfies $\mathbf{A}_\rho$. Let $\tau_\lambda \triangleq \min\left(\frac{1}{3}\exp\left(-\frac{16+8\gamma\lambda\log(|\mathcal{A}|)}{\lambda(1-\gamma)^2\rho_{\min}}\right), \frac{1}{3^8|\mathcal{A}|^4}\right)$. Then, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and for $\widetilde{\theta} = \mathcal{T}_{\tau_\lambda}(\theta)$, it holds that $\tilde{J}_\lambda(\widetilde{\theta}) \geq \tilde{J}_\lambda(\theta)$ and that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\pi_{\widetilde{\theta}}(a|s) \geq \tau_\lambda$.*

In this setting, Condition A-1 holds with: $L_1 \simeq \frac{1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^2}$, $L_2 \simeq \frac{1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^3}$, $L_3 \simeq \frac{1+\lambda\log|\mathcal{A}|}{(1-\gamma)^4}$, $\zeta^2 \simeq \frac{(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} + \frac{\varepsilon_r^2}{(1-\gamma)^4}$, $\beta \simeq \frac{1+\lambda\log(|\mathcal{A}|)\gamma^T T}{1-\gamma} + \frac{1+\lambda,\log(|\mathcal{A}|)\gamma^T}{(1-\gamma)^2}$, $\sigma_2^2 \simeq \frac{1+\lambda^2\log(|\mathcal{A}|)^2}{(1-\gamma)^4 B}$, $\sigma_4^4 \simeq \frac{1+\lambda^4\log(|\mathcal{A}|)^4}{(1-\gamma)^8 B^2}$.

By (10), each client satisfies the condition **PL**-1. Moreover, when using $\mathcal{T}_{\tau_\lambda}$ as the improvement operator, Lemma 5.2 shows that the corresponding PL constant is uniformly bounded from below by

$$\underline{\tilde{\mu}} = \underline{\tilde{\mu}}_\lambda := \lambda(1-\gamma)\,\rho_{\min}^2\tau_\lambda^2/|\mathcal{S}|,$$

where τ_λ is defined in Lemma 5.2. Thus, **PL**-2 also holds. We can therefore apply Corollary 4.5:

Corollary 5.3 ((Simplified) Sample and Communication Complexity of RS-FedPG)**.** *Assume $\mathbf{A}_\rho$ and that the projection operator is $\mathcal{T}_{\tau_\lambda}$. Let $\epsilon \gtrsim \frac{(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{\underline{\tilde{\mu}}_\lambda(1-\gamma)^6} + \frac{\varepsilon_r^2}{(1-\gamma)^4\underline{\tilde{\mu}}_\lambda} + \frac{(1+\lambda\log(|\mathcal{A}|))^2\gamma^{2T}T^2}{(1-\gamma)^2}$ and $\eta \lesssim \min\left(\frac{(1-\gamma)^3}{1+\lambda\log(|\mathcal{A}|)}, \frac{\underline{\tilde{\mu}}_\lambda\epsilon MB(1-\gamma)^7}{(1+\lambda\log(|\mathcal{A}|))^3}, \frac{\underline{\tilde{\mu}}_\lambda^{1/2}B(1-\gamma)^{7/2}\epsilon^{1/2}}{(1+\lambda\log(|\mathcal{A}|))^{3/2}}\right)$. Then RS-FedPG achieves $\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda\theta^R] \leq \epsilon$, with a number of communication*

$$R \gtrsim \frac{1}{\underline{\tilde{\mu}}_\lambda}\log\left(\frac{4(\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda\theta^0])}{\epsilon}\right)\frac{1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^3} \quad ,$$

for a total number of sampled trajectories per agent of

$$RHB \gtrsim \frac{1}{\underline{\tilde{\mu}}_\lambda}\log\Big(\frac{\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda\theta^0]}{\epsilon}\Big)\max\left(\frac{(1+\lambda\log(|\mathcal{A}|))B}{(1-\gamma)^3}, \frac{(1+\lambda\log(|\mathcal{A}|))^3}{\underline{\tilde{\mu}}_\lambda\epsilon M(1-\gamma)^7}, \frac{(1+\lambda\log(|\mathcal{A}|))^{3/2}}{\underline{\tilde{\mu}}_\lambda^{1/2}(1-\gamma)^{7/2}\epsilon^{1/2}}\right).$$

This proves that RS-FedPG achieve a logarithmic communication complexity in the desired accuracy while guaranteeing linear speedup as long as $M \lesssim \min\left(\frac{(1+\lambda\log(|\mathcal{A}|))^{3/2}}{\underline{\tilde{\mu}}_\lambda^{1/2}\epsilon^{1/2}(1-\gamma)^{3/2}}, \frac{(1+\lambda\log(|\mathcal{A}|))^2}{\underline{\tilde{\mu}}_\lambda\epsilon B(1-\gamma)^4}\right)$.

6 STRUCTURE OF HETEROGENEOUS FRL

In this section, we examine the structure of optimal policies in tabular FRL under heterogeneity. We provide only a brief overview here; a detailed analysis and proof of the theorem below are given in Appendix E. In the non-federated setting ($M = 1$), it is classical that optimal policies can always be chosen deterministic (Agarwal et al., 2019, Theorem 1.7). In contrast, for the federated case ($M > 1$), this property no longer holds: optimal policies may need to be stochastic, and in some cases even non-stationary. We provide an example of such an FRL instance in Figure 1.

A *history-dependent policy* is a sequence of mappings $(\pi_c^t)_{t\in\mathbb{N}}$, where each π_c^t selects actions based on the entire trajectory observed by agent c up to time t. The class of such policies is denoted by Π_ℓ. A *stationary stochastic policy* is a mapping $\pi\colon \mathcal{S} \to \mathcal{P}(\mathcal{A})$, assigning to each state a distribution over actions; this class is denoted by Π_{sta}. A *deterministic policy* is a mapping $\pi\colon \mathcal{S} \to \mathcal{A}$, choosing a single action for each state; these form the class Π_{det}. By construction, $\Pi_{\text{det}} \subset \Pi_{\text{sta}} \subset \Pi_\ell$. Restricting the policy class can only decrease (or preserve) the supremum of the objective J. Unlike the single-agent setting, where Π_{det} always contains an optimal policy, this property does not extend to FRL with heterogeneous dynamics.

Theorem 6.1. *There exists an FRL instance with two infinite-horizon discounted MDPs such that*

$$\max_{\pi\in\Pi_{\text{det}}} J(\pi) < \max_{\pi\in\Pi_{\text{sta}}} J(\pi), \qquad \max_{\pi\in\Pi_{\text{sta}}} J(\pi) < \max_{\pi\in\Pi_\ell} J(\pi).$$

The key difficulty stems from heterogeneity in the transition kernels: while homogeneous transitions (even with heterogeneous rewards) reduce to a standard RL problem with averaged rewards (see Appendix E.1), heterogeneity in dynamics fundamentally alters the structure of optimal policies.

Remark 6.2 (Algorithmic implications)**.** *The hierarchy of policy classes directly impacts algorithm design. Methods restricted to deterministic policies, such as Fed-Q-learning (Jin et al., 2022), are provably suboptimal in heterogeneous FRL (see Theorem 6.1). At the other extreme, history-dependent policies are too complex for practical implementation. Stationary stochastic policies thus strike a natural balance.*

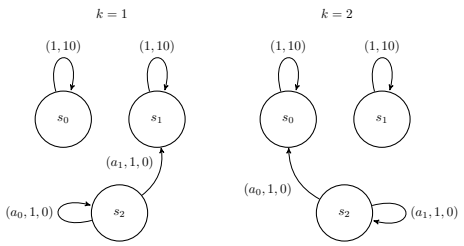


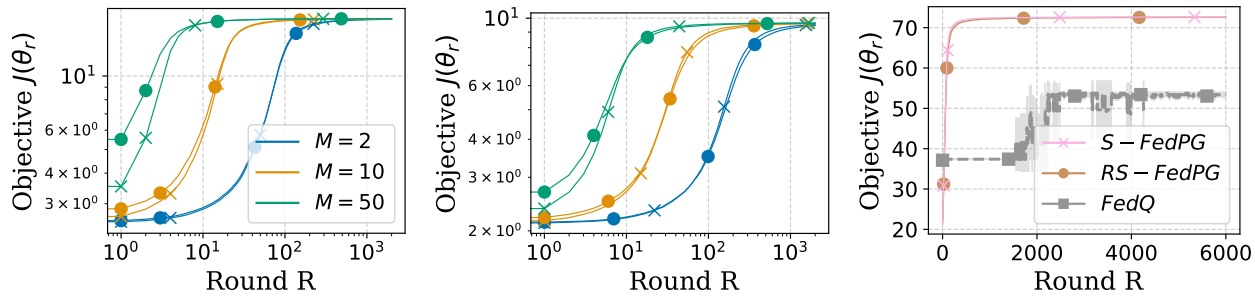
Figure 1: FRL task with no optimal deterministic policy. The triplet means (action, probability, reward). If the action is unspecified, it means that all actions give the same reward and lead to the same state.

7 EXPERIMENTS

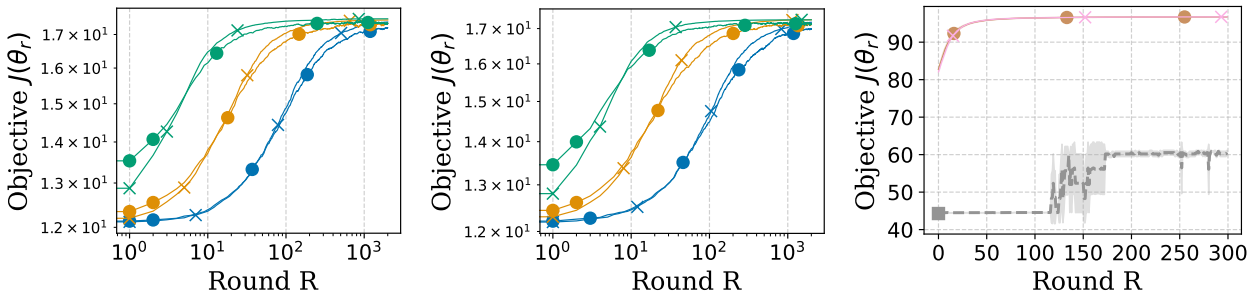
We study the empirical performance of the two proposed methods on two environments, and illustrate their advantage over `Fed-Q-learning` (Jin et al., 2022) in heterogeneous settings. Experiments were conducted on a computer with an Intel Xeon 6534 CPU with 196GB RAM. We report the average over 4 runs and the standard deviation in all the plots. Our code is available on https://github.com/Labbi-Safwan/FedPolicy-gradient

In the two environments, the transition kernel of agent c is 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 shared by all agents, and $\mathsf{P}_c^{\mathsf{ind}}$ is agent-specific. In both environments, the agent starts randomly from a uniformly sampled position, and $\gamma = 0.95$.

Synthetic. In the `synthetic` environment Zheng et al. (2023), 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 $\mathsf{P}^{\mathsf{com}}(\cdot|s,a)$ and $\mathsf{P}_c^{\mathsf{ind}}(\cdot|s,a)$ are drawn uniformly and randomly from the $|\mathcal{S}|$-dimensional simplex. 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 ($\varepsilon_{\mathsf{P}} \gg 0.3$) 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 $+1$ in every timestep, and the 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 1, to maximize their rewards.



(a) `GridWorld`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.0$ (b) `GridWorld`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.3$ (c) `GridWorld`, $H = 5$, $\varepsilon_{\mathsf{P}} \gg 0.3$



(d) `Synthetic`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.0$ (e) `Synthetic`, $H = 5$, $\varepsilon_{\mathsf{P}} = 0.3$ (f) `Synthetic`, $H = 5$, $\varepsilon_{\mathsf{P}} \gg 0.3$

Figure 2: Comparison of `S-FedPG` (crosses), `RS-FedPG` (circles), and `Fed-Q-learning` (squares) in two environments: `Synthetic` (below) and `Gridword` (above). From left to right: Value of the global objective $J(\theta^r)$ achieved by `S-FedPG`, and `RS-FedPG`, for $\varepsilon_{\mathsf{P}} = 0$, and for different numbers of agents $M \in \{2, 10, 50\}$ as a function of the number of rounds R; Value of the global objective $J(\theta^r)$ achieved by `S-FedPG`, and `RS-FedPG`, for $\varepsilon_{\mathsf{P}} = 0.3$, and for different numbers of agents $M \in \{2, 10, 50\}$ as a function of the number of rounds R; (c) Value of $J(\theta^r)$, comparing all three algorithms.

GridWorld (Domingues et al., 2021). In the `GridWorld` environment, an agent navigates 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. 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.

`FedPG` has linear speedup. In Figures 2a, 2b, 2d and 2e, we illustrate the *linear speedup* property by evaluating `S-FedPG` and `RS-FedPG` in both homogeneous and slightly heterogeneous environments. Specifically, we report the global objective J during the learning process for various numbers of agents, using the theoretically motivated step size, and $\lambda = 0.05$ for `RS-FedPG`. Both algorithms show that using a larger number of agents in the federation reduces the number of iterations per agent to reach convergence, highlighting the benefits of collaboration even among heterogeneous agents.

`FedPG` outperforms `Fed-Q-learning`. In Figures 2c and 2f, we compare the performance of `Fed-Q-learning` with `S-FedPG` and `RS-FedPG` on two highly heterogeneous FRL problems. `S-FedPG` and `RS-FedPG` *learn better policies*, demonstrating, as suggested by Theorem 6.1, the advantage of leveraging methods that learn a stochastic policy.

8 CONCLUSION

This work extends the theoretical foundations of FRL in heterogeneous environments. Our main contribution is the first global convergence guarantee for both non-regularized and entropy-regularized policy gradient methods in heterogeneous FRL. We also highlight structural differences that challenge classical RL properties, showing in particular that deterministic and stationary policies can be suboptimal. An important direction for future work is to address the *heterogeneity bias* that arises in federated objectives. A natural candidate is to adapt bias-reduction techniques developed in the convex setting, such as SCAFFOLD (Karimireddy et al., 2020), to the non-convex FRL landscape.

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.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. (2019). Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*.

Bertsekas, D. P. (2003). On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on automatic control*, 21(2):174–184.

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 non-convex federated optimization. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R., editors, *The Thirteenth International Conference on Learning Representations*, pages 80916–80983.

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.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR.

Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. (2022). Federated reinforcement learning: Linear speedup under markovian sampling. In *International Conference on Machine Learning*, pages 10997–11057. PMLR.

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.

Labbi, S., Tiapkin, D., Mangold, P., and Moulines, E. (2026). Beyond softmax and entropy: Convergence rates of policy gradients with f-softargmax parameterization & coupled regularization. In *The Fourteenth International Conference on Learning Representations*.

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.

Lu, M., Aghaei, M., Raj, A., and Vaswani, S. (2024). Towards principled, practical policy gradient for bandits and tabular MDPs. In *Reinforcement Learning Conference*.

Mangold, P., Berthier, E., and Moulines, E. (2025). Convergence guarantees for federated sarsa with local training and heterogeneous agents. *arXiv preprint arXiv:2512.17688*.

Mangold, P., Samsonov, S., Labbi, S., Levin, I., Alami, R., Naumov, A., and Moulines, E. (2024). Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning.

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.

Nair, A., Srinivasan, P., Blackwell, S., Alcicek, C., Fearon, R., De Maria, A., Panneershelvam, V., Suleyman, M., Beattie, C., Petersen, S., et al. (2015). Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*.

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.

Polyak, B. (1963). Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878.

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, H., Toso, L. F., Mitra, A., and Anderson, J. (2026). Model-free learning with heterogeneous dynamical systems: A federated LQR approach. *Transactions on Machine Learning Research*.

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*, pages 22423–22478.

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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Answer: **Yes**
   Justification: **The mathematical setting, the assumptions, and the algorithm are extensively described in Sections 3 to 5**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Answer: **Yes**
   Justification: **Sample and communication complexity results are derived in Sections 4 and 5.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Answer: **Yes**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Answer: **Yes**
   Justification: **Assumptions required explicitly stated in Sections 4 and 5 and referenced in the main theorems.**
   (b) Complete proofs of all theoretical results.
   Answer: **Yes**
   Justification: **The proofs of all results are provided in the appendix.**
   (c) Clear explanations of any assumptions.
   Answer: **Yes**
   Justification: **We provide intuition and justification for assumptions in Sections 4 and 5.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   Answer: **Yes**
   Justification: **The complete code is provided in the supplementary material.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Answer: *Yes*
   Justification: **Training details and parameters are described in Section 7.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Answer: **Yes**
   Justification: **Results report averages over 4 runs with standard deviations Section 7.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   Answer: **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Answer: **Yes**
   (b) The license information of the assets, if applicable.
   Answer: **Not Applicable**
   Justification: **The environments used are standard benchmarks with no licensing restrictions.**
   (c) New assets either in the supplemental material or as a URL, if applicable.
   Answer: **Yes**
   (d) Information about consent from data providers/curators.
   Answer: **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   Answer: **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   Answer: **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   Answer: **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   Answer: **Not Applicable**

Supplementary Materials

Contents

A Notations

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_c	Reward function of agent c	Section 3
γ	Discount factor	Section 3
ρ	Initial distribution over the state space	Section 3
ε_{P}	Heterogeneity on the transition kernels	Equation (1)
ε_{r}	Heterogeneity on the Rewards	Equation (2)
J	Global Non regularized FRL objective	Equation (4)
J_c	Local Non regularized objective of agent c	Equation (5)
λ	Regularization temperature	Section 3
$\tilde{J}_\lambda$	Global regularized FRL objective	Section 3
$\tilde{J}_{c,\lambda}$	Local regularized objective of agent c	Section 3
R	Number of Communication Rounds of `FedPG`/`proj-FedAVG`	Section 4
H	Number of local steps of `FedPG`/`proj-FedAVG`	Section 4
T	Length of the sampled trajectories by `FedPG`	Section 5
B	Number of trajectories collected per iteration	Section 5
$\mathcal{T}$	Projection operator used in `FedPG`/`proj-FedAVG`	Section 4
π_θ	Softmax policy parametrized by $\theta \in \mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{A}\rvert}$	Section 3
F	Global objective optimised by `proj-FedAVG`	Section 4
f_c	Local function of agent c in `proj-FedAVG`	Section 4
$\mathrm{g}_c^Z(\theta)$	Stochastic estimators of `proj-FedAVG`	Section 4
$\mathrm{g}_c(\theta)$	Expected value of the stochastic estimators of `proj-FedAVG`	Section 4
L_1	Bound on the gradients of f_c	Section 4
L_2	Smoothness constants of f_c and g_c	Section 4
L_3	Bounds on the third-order derivative tensors of f_c	Section 4
σ_p^p	Bounds on the p-th central moments of $\mathrm{g}_c^Z(\theta)$ for $p \in \{2,4\}$	Section 4
β	Bounds on bias of $\mathrm{g}_c^Z(\theta)$	Section 4
ζ	Bound on gradient heterogeneity of F	Section 4
μ_c	'Quadratic' Łojasiewicz coefficient of agent c	Section 4
$\tilde{\mu}_c$	Polyak-Łojasiewicz coefficient of agent c	Section 4

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

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

$\langle \cdot, \cdot \rangle$ 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 Descent lemma

In this section, we study the following federated stochastic optimization problem

$$\max_{\theta\in\mathbb{R}^d} F(\theta) = \frac{1}{M}\sum_{c=1}^{M} f_c(\theta) \ , \quad \text{where } f_c(\theta) \triangleq \mathbb{E}_{Z_c\sim\xi_c(\theta)}[f_c^{Z_c}(\theta)] \ , \tag{18}$$

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

$$\mathrm{g}_c(\theta) \triangleq \mathbb{E}_{Z_c\sim\xi_c(\theta)}[\mathrm{g}_c^{Z_c}(\theta)] \ , \tag{19}$$

but is typically different from the gradient of f_c.

To solve (18), 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 using a projection-like operator $\mathcal{T}\colon \mathbb{R}^d\to\mathbb{R}^d$. For completeness, we give the pseudo-code for this algorithm in Algorithm 1.

B.1 Descent Lemma

Assumptions. We derive our new ascent lemma 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. *There exists $L_1>0$, such that for all $c\in[M]$ and $\theta\in\mathbb{R}^d$,*

$$\|\nabla f_c(\theta)\| \le L_1 \ , \quad \text{for all } c\in[M] \ , \ \theta\in\mathbb{R}^d \ . \tag{20}$$

FL-2. *For any $c\in[M]$, the functions ∇f_c and the biased gradients g_c are L_2-Lipschitz, that is*

$$\|\nabla^2 f_c(\theta)u\| \le L_2\|u\| \ , \quad \text{for all } \theta\in\mathbb{R}^d, u\in\mathbb{R}^d \ , \tag{21}$$

$$\|\mathrm{g}_c(\theta)-\mathrm{g}_c(\theta')\| \le L_2\|\theta-\theta'\| \ , \quad \text{for all } \theta,\theta'\in\mathbb{R}^d \ . \tag{22}$$

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*

$$\|\nabla^3 f_c(\theta)u^{\otimes 2}\| \le L_3\|u\|^2 \ , \quad \text{for all } \theta\in\mathbb{R}^d \ , \ u\in\mathbb{R}^d \ . \tag{23}$$

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

$$\|\nabla f_c(\theta)-\nabla F(\theta)\| \le \zeta \ , \quad \text{for all } c\in[M] \ , \ \theta\in\mathbb{R}^d \ . \tag{24}$$

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

$$\mathbb{E}_{Z_c\sim\xi_c(\theta)}[\|\mathrm{g}_c^{Z_c}(\theta)-\mathrm{g}_c(\theta)\|^p] \le \sigma_p^p \ , \quad \text{for all } c\in[M] \ , \ \theta\in\mathbb{R}^d \ . \tag{25}$$

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

$$\|\mathrm{g}_c(\theta)-\nabla f_c(\theta)\| \le \beta \ , \quad \text{for all } c\in[M] \ , \ \theta\in\mathbb{R}^d \ . \tag{26}$$

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

In the following, we define the filtration adapted to the global and local iterates of Algorithm 1 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 1.

Lemma B.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****-2 and* ***FL****-1. respectively. Then the iterates of* `proj-FedAVG` *satisfy*

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

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

$$\mathrm{D}_{3,c}^r(\theta_c^{r,h}) = \int_0^1 (1-t)\nabla^3 f_c(\theta^r + t(\theta_c^{r,h}-\theta^r))\mathrm{d}t \ . \tag{27}$$

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

$$\begin{aligned}\mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} &\leq 2\|\nabla^2 f_c(\theta^r)\mathbb{E}\left[\theta_c^{r,h}-\theta^r\middle|\mathcal{F}^r\right]\|_2^2 + 2\|\mathbb{E}\left[\mathrm{D}_{3,c}^r(\theta_c^{r,h})(\theta_c^{r,h}-\theta^r)^{\otimes 2}\middle|\mathcal{F}^r\right]\|_2^2 \\ &\leq 2L_2^2\|\mathbb{E}\left[\theta_c^{r,h}-\theta^r\middle|\mathcal{F}^r\right]\|_2^2 + 2L_3^2\mathbb{E}\left[\|\theta_c^{r,h}-\theta^r\|_2^4\middle|\mathcal{F}^r\right] \ ,\end{aligned} \tag{28}$$

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 (26) to write

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

Completing the sum until $\ell = H-1$ and plugging this inequality in (28), 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

$$\frac{1}{MH}\sum_{c=1}^{M}\sum_{h=1}^{H-1}\mathbf{U}_{\boldsymbol{c}}^{\boldsymbol{h}} \leq \frac{6\eta^2 L_2^2 H(H-1)}{M}\sum_{c=1}^{M}\|\nabla f_c(\theta^r)\|_2^2$$

$$+\frac{4L_3^2}{MH}\sum_{c=1}^{M}\sum_{h=0}^{H-1}\mathbb{E}\left[\|\theta_c^{r,h}-\theta^r\|_2^4\middle|\mathcal{F}^r\right]+6\eta^2L_2^2H(H-1)\beta^2 \ . \tag{29}$$

(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

$$\begin{aligned}\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \le 4^3\eta^4\underbrace{\mathbb{E}\left[\left\|\sum_{\ell=0}^{h-1}\mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell})-\mathrm{g}_c(\theta_c^{r,\ell})\right\|^4\middle|\mathcal{F}^r\right]}_{T_1}+4^3\eta^4\underbrace{\mathbb{E}\left[\left\|\sum_{\ell=0}^{h-1}\mathrm{g}_c(\theta_c^{r,\ell})-\nabla f_c(\theta_c^{r,\ell})\right\|^4\middle|\mathcal{F}^r\right]}_{T_2}\\ +4^3\eta^4\underbrace{\mathbb{E}\left[\left\|\sum_{\ell=0}^{h-1}\nabla f_c(\theta_c^{r,\ell})-\nabla f_c(\theta^r)\right\|^4\middle|\mathcal{F}^r\right]}_{T_3}+4^3\eta^4\underbrace{h^4\|\nabla f_c(\theta^r)\|^4}_{T_4} \ .\end{aligned} \tag{30}$$

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

$$T_1 \le 3^4\left\{\sum_{\ell=0}^{h-1}\mathbb{E}^{1/2}\left[\|\mathrm{g}_c^{Z_c^{r,\ell+1}}(\theta_c^{r,\ell})-\mathrm{g}_c(\theta_c^{r,\ell})\|^4\middle|\mathcal{F}^r\right]\right\}^2 \le 3^4h^2\sigma_4^4 \ . \tag{31}$$

The term T_2 is a bias term, which we bound using (26),

$$T_2 \le h^3\sum_{\ell=0}^{h-1}\mathbb{E}\left[\|\mathrm{g}_c(\theta_c^{r,\ell})-\nabla f_c(\theta_c^{r,\ell})\|^4\middle|\mathcal{F}^r\right] \le h^4\beta^4 \ . \tag{32}$$

Then, we bound T_3 using (21)

$$T_3 \le h^3\sum_{\ell=0}^{h-1}\mathbb{E}\left[\|\nabla f_c(\theta_c^{r,\ell})-\nabla f_c(\theta^r)\|^4\middle|\mathcal{F}^r\right] \le L_2^4h^3\sum_{\ell=0}^{h-1}\mathbb{E}\left[\|\theta_c^{r,\ell}-\theta^r\|^4\middle|\mathcal{F}^r\right] \ . \tag{33}$$

Finally, we bound T_4 using gradient's boundedness (20),

$$T_4 \le L_1^2h^4\|\nabla f_c(\theta^r)\|^2 \ . \tag{34}$$

Plugging (31), (32), (33), (34) in (30), we obtain

$$\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \le 4^3\eta^4h^4L_1^2\|\nabla f_c(\theta^r)\|^2+4^3\eta^4h^4\beta^4+4^3\eta^4L_2^4h^3\sum_{\ell=0}^{h-1}\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{\ell}}+3\cdot12^3\eta^4h^2\sigma_4^4 \ .$$

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

$$\begin{aligned}\frac{1}{H}\sum_{h=0}^{H-1}\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \le \frac{4^3\eta^4L_2^4H^3(H-1)}{5H}\sum_{h=0}^{H-1}\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}}+\frac{3\cdot12^3\eta^4H(H-1)}{3}\sigma_4^4\\ +\frac{4^3\eta^4H^2(H-1)^2}{5}\left(L_1^2\|\nabla f_c(\theta^r)\|^2+\beta^4\right) \ .\end{aligned}$$

Using $\eta HL_2 \le 1/6$, 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

$$\frac{1}{MH}\sum_{c=1}^{M}\sum_{h=0}^{H-1}\mathbf{V}_{\boldsymbol{c}}^{\boldsymbol{h}} \le \frac{2\cdot4^3\eta^4H^2(H-1)^2}{5}\left\{\beta^4+L_1^2\|\nabla f_c(\theta^r)\|^2\right\}+\frac{6\cdot12^3\eta^4H(H-1)}{3}\sigma_4^4 \ . \tag{35}$$

(Final Result.) Plugging (35) back in (29) and using $L_3^2\eta^2H^2L_1^2 \le L_2^2/32$ and $\beta \le L_1$ gives

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

and the result follows. □

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

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

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 $\mathbb{E}\left[\mathrm{g}_c^{Z_c^{r,h+1}}(\theta_c^{r,h})\middle|\mathcal{F}^r\right]=\mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]$,

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

Since $\{Z_c^{r,h}\}_{h=1}^{H}$ are independent conditional to $\mathcal{F}^r$, we have, using (25),

$$\mathbf{A_c^{r,h+1}}=\mathbb{E}\Big[\Big\|\theta_c^{r,h}-\mathbb{E}\left[\theta_c^{r,h}\middle|\mathcal{F}^r\right]+\eta\big(\mathrm{g}_c(\theta_c^{r,h})-\mathbb{E}\left[\mathrm{g}_c(\theta_c^{r,h})\middle|\mathcal{F}^r\right]\big)\Big\|^2\Big]+\eta^2\sigma_2^2. \tag{36}$$

Then, by Young's inequality, we have

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

Finally, we have, by Young's inequality and (21),

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

where we used Jensen's inequality in the last inequality. Then, notice that $4(\eta^2+\eta/L_2)L_2^2=4(\eta^2L_2^2+\eta L_2)\le 5\eta L_2$ since $\eta L_2\le 1/4$. Plugging this in (36), we obtain

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

And unrolling this inequality gives

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

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

Lemma B.3 (Descent Lemma)**.** *Assume* ***FL-*** *1 to* ***FL-*** *6. 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* `proj-FedAVG` *satisfy*

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

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

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

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

$$\begin{aligned}-\mathbb{E}\left[F(\bar\theta^{r+1})\middle|\mathcal{F}^r\right] + F(\theta^r) &\le -\langle \kappa^{-1}\nabla F(\theta^r), \kappa \mathbb{E}\left[\bar\theta^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\rangle + \frac{L_2}{2}\mathbb{E}\left[\|\bar\theta^{r+1} - \theta^r\|_2^2\middle|\mathcal{F}^r\right] \\ &= -\frac{1}{2\kappa^2}\|\nabla F(\theta^r)\|_2^2 + \underbrace{\frac{1}{2\kappa^2}\|\nabla F(\theta^r) + \kappa^2 \mathbb{E}\left[\theta^r - \bar\theta^{r+1}\middle|\mathcal{F}^r\right]\|_2^2}_{(\mathbf{A})} \\ &\quad + \underbrace{\frac{L_2}{2}\mathbb{E}\left[\|\bar\theta^{r+1} - \theta^r\|_2^2\middle|\mathcal{F}^r\right] - \frac{\kappa^2}{2}\|\mathbb{E}\left[\bar\theta^{r+1} - \theta^r\middle|\mathcal{F}^r\right]\|_2^2}_{(\mathbf{B})} \ . \end{aligned} \tag{37}$$

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

Bounding (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

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

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 (26), we obtain

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

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

$$\begin{aligned}(\mathbf{A}) \le \frac{8\eta^3 L_2^2 H^2 (H-1)}{M}\sum_{c=1}^M \|\nabla f_c(\theta^r)\|_2^2 \\ + 4\cdot 12^3 \cdot 2 L_3^2 \eta^5 H^2 (H-1)\sigma_4^4 + (1 + 8\eta^2 L_2^2 H (H-1))\eta H \beta^2 \ .\end{aligned} \tag{38}$$

Bounding (B). We decompose (**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 B.2, which gives

$$(\mathbf{B}) \leq \frac{3\eta^2 L_2 H \sigma_2^2}{2M} \ . \tag{39}$$

Bound on (37). Plugging in the bounds (38) and (39) on (**A**) and (**B**) in (37) yields

$$\begin{aligned}-\mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right] + F(\theta^r) \leq &-\frac{\eta H}{2}\|\nabla F(\theta^r)\|_2^2 + \frac{8\eta^3 L_2^2 H^2(H-1)}{M}\sum_{c=1}^M\|\nabla f_c(\theta^r)\|_2^2 \\ &+ 4\cdot 12^3\eta^5 L_3^2 H^2(H-1)\sigma_4^4 + 2\eta H\beta^2 + \frac{3\eta^2 L_2 H\sigma_2^2}{2M} \ ,\end{aligned} \tag{40}$$

where we used $\eta H L_2 \leq 1/6$ to bound $(1 + 8\eta^2 L_2^2 H(H-1))\eta H\beta^2 \leq 2\eta H\beta^2$. Moreover, we have $8\eta^2 L_2^2 H^2 \leq 1/4$ 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

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

and the result of the lemma follows from plugging this inequality in (40). □

B.2 Convergence under Local Non-Uniform Łojasiewicz inequality

Firstly, define

$$f_c^\star = \sup_{\theta\in\mathbb{R}^d} f_c(\theta) \ , \quad F^\star = \frac{1}{M}\sum_{c=1}^M f_c^\star \ .$$

Convergence under local 'quadratic' Non-Uniform Łojasiewicz inequalities For any $(c,\theta) \in [M]\times\mathbb{R}^d$, define

$$\mu_c(\theta) := \sup\{x \in \mathbb{R}^+, \|\nabla f_c(\theta)\|_2^2 \geq 2x\left(f_c^\star - f_c(\theta)\right)^2\} \ .$$

We assume the following additional condition

QL- 1. *For any $c \in [M]$, we have $f_c^\star < \infty$. Additionnally, for any $c \in [M]$,and $\theta \in \mathbb{R}^d$, there exists $\mu_c(\theta) > 0$ such that $\|\nabla f_c(\theta)\|_2^2 \geq 2\mu_c(\theta)\left(f_c^\star - f_c(\theta)\right)^2$.*

For any parameter $\theta \in \mathbb{R}^d$, define $\mu(\theta) \triangleq \min_{c\in[M]}\mu_c(\theta)$.

QL- 2. *For any $\theta \in \mathbb{R}^d$, it holds that $F(\mathcal{T}(\theta)) \geq F(\theta)$. Additionally, there exists $\underline{\mu} > 0$, such that we have $\mu(\mathcal{T}(\theta)) \geq \underline{\mu}$.*

Under these two additional assumptions, we can derive global convergence rates for `proj-FedAVG`. We preface the proof with two elementary Lemmas.

Lemma B.4. *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 $v_k \triangleq w_k - M$ 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

$$\frac{v_{k+1} - v_k}{v_k^2} \leq -\kappa \ . \tag{41}$$

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 (41)

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

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 $w_k \leq B + M$. □

Lemma B.5. *Assume **FL**-4 and **QL**-1. For any $\theta \in \mathbb{R}^d$, it holds that*

$$\zeta^2 + \|\nabla F(\theta)\|_2^2 \geq \mu(\theta)(F^\star - F(\theta))^2 \ .$$

Proof. Let $\theta \in \mathbb{R}^d$. Using **QL**-1, we have for any $c \in [M]$

$$\sqrt{\min_{c \in [M]} 2\mu_c(\theta)} \, [f_c^\star - f_c(\theta)] \leq \sqrt{2\mu_c(\theta)} \, [f_c^\star - f_c(\theta)] \leq \|\nabla f_c(\theta)\|_2 \ .$$

We then decompose $\nabla f_c(\theta) = \nabla f_c(\theta) - \nabla F(\theta) + \nabla F(\theta)$ and use triangle inequality and **FL**-4 to bound

$$\|\nabla f_c(\theta)\|_2 \leq \|\nabla f_c(\theta) - \nabla F(\theta)\|_2 + \|\nabla F(\theta)\|_2 \leq \zeta + \|\nabla F(\theta)\|_2 \ .$$

Averaging the resulting inequality over all the agents, taking the square, and applying Young's inequality concludes the proof. □

Theorem B.6 (Convergence rates of `proj-FedAVG`)**.** *Assume **FL**-1 to **FL**-6, **QL**-1 and **QL**-2. 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*

$$\begin{aligned} F^\star - \mathbb{E}[F(\theta^R)] \leq{}& \frac{F^\star - F(\theta^0)}{1 + R \cdot (F^\star - F(\theta^0)) \cdot (\eta H \underline{\mu}/4)} + \left(\frac{6\eta L_2 \sigma_2^2}{M\underline{\mu}}\right)^{1/2} \\ &+ \left(\frac{16 \cdot 12^3 \eta^4 L_3^2 H(H-1)\sigma_4^4}{\underline{\mu}}\right)^{1/2} + \left(\frac{2\zeta^2}{\underline{\mu}}\right)^{1/2} + \left(\frac{8\beta^2}{\underline{\mu}}\right)^{1/2} \\ &+ \frac{\zeta^2}{12 L_2} + \frac{\eta \sigma_2^2}{4M} + \frac{\beta^2}{3L_2} + \frac{12^3 \eta^4 L_3^2 H(H-1)\sigma_4^4}{L_2} \ . \end{aligned}$$

Proof. Firstly, using **QL**-2 note by an immediate recursion that

$$\inf_{r\geq 0}\mu(\theta^r)\geq \underline{\mu}\ .$$

Applying Lemma B.3 yields

$$-\mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right]\leq -F(\theta^r)-\frac{\eta H}{4}\|\nabla F(\theta^r)\|_2^2+\frac{3\eta^2L_2H\sigma_2^2}{2M}\\ +2\eta H\beta^2+8\eta^3L_2^2H^2(H-1)\zeta^2+4\cdot 12^3\eta^5L_3^2H^2(H-1)\sigma_4^4\ .$$

Adding $F^\star$, and using Lemma B.5 combined with **QL**-2 yields

$$F^\star-\mathbb{E}\left[F(\theta^{r+1})\middle|\mathcal{F}^r\right]\leq F^\star-F(\theta^r)-\frac{\eta H\underline{\mu}}{4}\left(F^\star-F(\theta^r)\right)^2+\frac{\eta H}{4}\zeta^2+\frac{3\eta^2L_2H\sigma_2^2}{2M}\\ +2\eta H\beta^2+8\eta^3L_2^2H^2(H-1)\zeta^2+4\cdot 12^3\eta^5L_3^2H^2(H-1)\sigma_4^4\ .$$

Taking the expectation with respect to all the stochasticity, applying Jensen's inequality, and using that $\eta H L_2\leq 1/6$ to simplify the heterogeneity terms gives

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

where we defined $\delta^r=F^\star-\mathbb{E}[F(\theta^r)]$, $\kappa=\frac{\eta H\underline{\mu}}{4}$, and

$$B=\frac{\eta H}{2}\zeta^2+\frac{3\eta^2L_2H\sigma_2^2}{2M}+2\eta H\beta^2+4\cdot 12^3\eta^5L_3^2H^2(H-1)\sigma_4^4\ .$$

Finally, applying Lemma B.4 on the sequence δ^r concludes the proof. □

Corollary B.7 (Sample and Communication Complexity)**.** *Under the assumptions of Theorem B.6, let*

$$\epsilon>\frac{12\zeta}{\underline{\mu}^{1/2}}+\frac{18\beta}{\underline{\mu}^{1/2}}+\frac{\zeta^2}{2L_2}+\frac{2\beta^2}{L_2}\ ,$$

and

$$\eta\leq\min\left(\frac{1}{6L_2},\frac{\underline{\mu}M\epsilon^2}{216L_2\sigma_2^2},\frac{\mu^{1/2}\epsilon L_2}{13^2L_3\sigma_4^2},\frac{2\epsilon M}{\sigma_2^2},\frac{\epsilon^{1/2}L_2^{3/2}}{24L_3\sigma_4^2}\right)\ .$$

In this case `proj-FedAVG` *achieves* $F^\star-\mathbb{E}[F(\theta^R)]$*, with a number of communication*

$$R\geq\frac{144\left[F^\star-F(\theta^0)-\epsilon/6\right]}{(F^\star-F(\theta^0))\underline{\mu}\epsilon}\max\left(L_2,\frac{L_3L_1}{L_2}\right)\ ,$$

for a total number of samples per agent of

$$RH\geq\frac{144[F^\star-F(\theta^0)-\epsilon/6]}{(F^\star-F(\theta^0))\underline{\mu}\epsilon}\max\left(L_2,\frac{36L_2\sigma_2^2}{\underline{\mu}M\epsilon^2},\frac{29L_3\sigma_4^2}{\mu^{1/2}\epsilon L_2},\frac{\sigma_2^2}{12M\epsilon},\frac{4L_3\sigma_4^2}{\epsilon^{1/2}L_2^{3/2}}\right)\ .$$

Proof. First, we require (i) that $\eta L_2\leq 1/6$, and that (ii) each variance terms to be smaller than $\epsilon/6$, which gives the condition on the step size

$$\eta\leq\min\left(\frac{1}{6L_2},\frac{\underline{\mu}M\epsilon^2}{216L_2\sigma_2^2},\frac{\mu^{1/2}\epsilon L_2}{13^2L_3\sigma_4^2},\frac{2\epsilon M}{\sigma_2^2},\frac{\epsilon^{1/2}L_2^{3/2}}{24L_3\sigma_4^2}\right)\ . \tag{42}$$

Then, H has to satisfy $\eta H L_2\leq 1/6$ and $32\eta^2H^2L_3^2L_1^2\leq L_2^2$, which requires

$$H\leq\frac{1}{\eta}\min\left(\frac{1}{6L_2},\frac{L_2}{6L_3L_1}\right)\ .$$

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

$$R\geq\frac{F^\star-F(\theta^0)-\epsilon/6}{(F^\star-F(\theta^0))\eta H\underline{\mu}\epsilon/24}=\frac{144\left[F^\star-F(\theta^0)-\epsilon/6\right]}{(F^\star-F(\theta^0))\underline{\mu}\epsilon}\max\left(L_2,\frac{L_3L_1}{L_2}\right)\ .$$

The sample complexity follows from $RH\geq\frac{F^\star-F(\theta^0)-\epsilon/6}{(F^\star-F(\theta^0))\eta\underline{\mu}\epsilon/24}$ and (42). □

Convergence under local 'linear' Non-Uniform Łojasiewicz inequalities For any $(c,\theta) \in [M] \times \mathbb{R}^d$, define

$$\tilde{\mu}_c(\theta) := \sup\{x \in \mathbb{R}^+, \|\nabla f_c(\theta)\|_2^2 \geq 2x\,(f_c^\star - f_c(\theta))\} \ .$$

We assume the following additional condition

PL-1. *For any $c \in [M]$, we have $f_c^\star < \infty$. Additionnally, for any $c \in [M]$,and $\theta \in \mathbb{R}^d$, there exists $\tilde{\mu}_c(\theta) > 0$ such that*

$$\|\nabla f_c(\theta)\|_2^2 \geq 2\tilde{\mu}_c(\theta)\,(f_c^\star - f_c(\theta)) \ .$$

For any parameter $\theta \in \mathbb{R}^d$, define

$$\tilde{\mu}(\theta) \triangleq \min_{c\in[M]} \tilde{\mu}_c(\theta) \ .$$

PL-2. *For any $\theta \in \mathbb{R}^d$, it holds that*

$$F(\mathcal{T}(\theta)) \geq F(\theta) \ .$$

Additionally, there exists $\underline{\tilde{\mu}} > 0$, such that we have $\tilde{\mu}(\mathcal{T}(\theta)) \geq \underline{\tilde{\mu}}$.

Under these two additional assumptions, we can derive global convergence rates for `proj-FedAVG`. We preface the proof with an elementary Lemma.

Lemma B.8. *Assume **FL**-4 and **PL**-1. For any $\theta \in \mathbb{R}^d$, it holds that*

$$\zeta^2 + \|\nabla F(\theta)\|_2^2 \geq 2\tilde{\mu}(\theta)(F^\star - F(\theta)) \ .$$

Proof. Let $\theta \in \mathbb{R}^d$. Using **PL**-1 and the triangle inequality, we have for any $c \in [M]$

$$\begin{aligned}
&\min_{c\in[M]} 2\tilde{\mu}_c(\theta)[f_c^\star - f_c(\theta)] \\
&\quad\leq 2\tilde{\mu}_c(\theta)[f_c^\star - f_c(\theta)] \leq \|\nabla f_c(\theta)\|_2^2 \\
&\quad= \|\nabla f_c(\theta) - \nabla F(\theta) + \nabla F(\theta)\|_2^2 \\
&\quad= \|\nabla f_c(\theta) - \nabla F(\theta)\|_2^2 + 2\langle \nabla f_c(\theta) - \nabla F(\theta), \nabla F(\theta)\rangle + \|\nabla F(\theta)\|_2^2 \\
&\quad\leq \zeta^2 + 2\langle \nabla f_c(\theta) - \nabla F(\theta), \nabla F(\theta)\rangle + \|\nabla F(\theta)\|_2^2 \ ,
\end{aligned}$$

where in the last inequality we used **FL**-4. Finally, averaging the preceding inequality over all the agents concludes the proof. □

Theorem B.9 (Convergence rates of `proj-FedAVG`)**.** *Assume **FL**-1 to **FL**-6, **PL**-1 and **PL**-2. 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*

$$\begin{aligned}
F^\star - \mathbb{E}\left[F(\theta^R)\right] \leq \left(1 - \frac{\eta H \underline{\tilde{\mu}}}{2}\right)^R (F^\star - F(\theta^0)) + \frac{3\eta L_2 \sigma_2^2}{M\underline{\tilde{\mu}}} + \frac{\zeta^2}{\underline{\tilde{\mu}}} \\
+ 4\frac{\beta^2}{\underline{\tilde{\mu}}} + \frac{8^2 \cdot 12\eta^4 L_3^2 H(H-1)\sigma_4^4}{\underline{\tilde{\mu}} L_2} \ .
\end{aligned}$$

Proof. Firstly, using **PL**-2 note by an immediate recursion that

$$\inf_{r\geq 0} \tilde{\mu}(\theta^r) \geq \underline{\tilde{\mu}} \ .$$

Applying Lemma B.3 yields

$$-\mathbb{E}\left[F(\bar{\theta}^{r+1})\middle|\mathcal{F}^r\right] \leq -F(\theta^r) - \frac{\eta H}{4}\|\nabla F(\theta^r)\|_2^2 + \frac{3\eta^2 L_2 H \sigma_2^2}{2M}$$

$$+ 2\eta H\beta^2 + 8\eta^3 L_2^2 H^2 (H-1)\zeta^2 + 8\cdot 12^2 \eta^5 L_3^2 H^2 (H-1)\sigma_4^4 \ .$$

Adding $F^\star$, and using Lemma B.8 combined with **PL**-2 yields

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

Taking the expectation with respect to all the stochasticity, yield

$$F^\star - \mathbb{E}\left[F(\theta^{r+1})\right] \le \left(1 - \frac{\eta H \underline{\tilde{\mu}}}{2}\right)(F^\star - \mathbb{E}\left[F(\theta^r)\right]) + \frac{\eta H}{4}\zeta^2 + \frac{3\eta^2 L_2 H \sigma_2^2}{2M} + 2\eta H\beta^2 + 8\eta^3 L_2^2 H^2 (H-1)\zeta^2 + 8\cdot 12^2 \eta^5 L_3^2 H^2 (H-1)\sigma_4^4 \ .$$

The result follows from unrolling the recursion. □

Corollary B.10 (Sample and Communication Complexity of `proj-FedAVG`)**.** *Under the assumptions of Theorem B.9, let*

$$\epsilon > \frac{4\zeta^2}{\underline{\tilde{\mu}}} + \frac{16\beta^2}{\underline{\tilde{\mu}}} \ ,$$

and

$$\eta \le \min\left(\frac{1}{6L_2}, \frac{\underline{\tilde{\mu}}\epsilon M}{12 L_2 \sigma_2^2}, \frac{\underline{\tilde{\mu}}^{1/2} L_2^{3/2} \epsilon^{1/2}}{5 L_3 \sigma_4^2}\right) \ .$$

Then `proj-FedAVG` achieves $F^\star - \mathbb{E}[F(\theta^R)]$, with a number of communication

$$R \ge \frac{12}{\underline{\tilde{\mu}}} \log\left(\frac{4(F^\star - F(\theta^0))}{\epsilon}\right) \max\left(L_2, \frac{L_3 L_1}{L_2}\right) \ ,$$

for a total number of samples per agent of

$$RH \ge \frac{2}{\underline{\tilde{\mu}}} \log\left(\frac{4F^\star - F(\theta^0))}{\epsilon}\right) \max\left(6L_2, \frac{12 L_2 \sigma_2^2}{\underline{\tilde{\mu}}\epsilon M}, \frac{5 L_3 \sigma_4^2}{\underline{\tilde{\mu}}^{1/2} L_2^{3/2} \epsilon^{1/2}}\right) \ .$$

Proof. Firstly, we require (i) that $\eta L_2 \le 1/6$, and that (ii) each variance terms to be smaller than $\epsilon/4$, which gives the condition on the step size

$$\eta \le \min\left(\frac{1}{6L_2}, \frac{\underline{\tilde{\mu}}\epsilon M}{12 L_2 \sigma_2^2}, \frac{\underline{\tilde{\mu}}^{1/2} L_2^{3/2} \epsilon^{1/2}}{5 L_3 \sigma_4^2}\right) \ , \tag{43}$$

Then, H has to satisfy $\eta H \le 1/6L_2$ and $32\eta^2 H^2 L_3^2 L_1^2 \le L_2^2$. This requires

$$H \le \frac{1}{\eta}\min\left(\frac{1}{6L_2}, \frac{L_2}{6L_3 L_1}\right) \ .$$

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

$$R \ge \frac{2}{\eta H \underline{\tilde{\mu}}} \log\left(\frac{4(F^\star - F(\theta^0))}{\epsilon}\right) = \frac{12}{\underline{\tilde{\mu}}} \log\left(\frac{4(F^\star - F(\theta^0))}{\epsilon}\right) \max\left(L_2, \frac{L_3 L_1}{L_2}\right) \ .$$

The sample complexity follows from $RH \ge \frac{2}{\eta \underline{\tilde{\mu}}} \log\left(\frac{4F^\star - F(\theta^0))}{\epsilon}\right)$ and (43). □

C Analysis of S-FedPG

S-FedPG can be interpreted as a specific instance of proj-FedAVG, where the projection set is the identity function, i.e. $\mathcal{T}\colon \theta \to \theta$, the local objective is defined as $f_c = J_c$, and the agent data distribution $\xi_c(\theta)$ corresponds to $[\nu_c(\theta)]^{\otimes B}$, where $\nu_c(\theta)$ is the distribution induced by sampling B truncated trajectories from the policy π_θ, defined by

$$.\nu_c(\theta; z) = \rho(s^0)\pi_\theta(a^0|s^0) \prod_{t=0}^{T-1} \mathsf{P}(s^t \mid s^{t-1}, a^{h-1})\pi(a^h|s^h) \ . \tag{44}$$

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 (16)) for the stochastic gradient:

$$\mathrm{g}_{c,\mathrm{s}}^{Z_c}(\theta) \triangleq \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t \left(\sum_{\ell=0}^{t} \nabla \log \pi_\theta(A_{c,b}^\ell \mid S_{c,b}^\ell)\right) \mathsf{r}_c(S_{c,b}^t, A_{c,b}^t) \ . \tag{45}$$

Define also

$$\mathrm{g}_{c,\mathrm{s}}(\theta) = \mathbb{E}_{Z_c\sim\nu_c(\theta)}[\mathrm{g}_{c,\mathrm{s}}^{Z_c}(\theta)]. \tag{46}$$

To apply the convergence results of Appendix B, it remains to verify that Assumptions **FL**-1 to **FL**-6 and **QL**-1 hold (**QL**-2 will be assumed to hold for S-FedPG). We establish these conditions in the following.

C.1 Checking the assumptions

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

$$V_c^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t \mathsf{r}_c(S_c^t, A_c^t)\middle| S_c^0 = s\right], \tag{47}$$

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

$$Q_c^\pi(s,a) \triangleq \mathsf{r}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathsf{P}_c(s'|s,a)V_c^\pi(s') \ . \tag{48}$$

This allows to define the advantage function $A_c^\pi(s,a) = Q_c^\pi(s,a) - V_c^\pi(s)$. Define $J_c(\theta) \triangleq J_c(\pi_\theta)$. We define the advantage function of a policy π_θ as

$$A_c^{\pi_\theta}(s,a) \triangleq Q_c^{\pi_\theta}(s,a) - V_c^{\pi_\theta}(s) \ , \quad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A} \ . \tag{49}$$

The occupancy measure of agent $c \in [M]$, is defined as

$$d_c^{\rho,\pi}(s) \triangleq (1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t \rho \mathsf{P}_{c,\pi}^t(s) \ , \text{where} \qquad \mathsf{P}_{c,\pi}(s'|s) \triangleq \textstyle\sum_{a\in\mathcal{A}}\pi(a|s)\mathsf{P}_c(s'|s,a)$$

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

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

$$\frac{\partial J_c(\theta)}{\partial\theta(s,a)} = \frac{1}{1-\gamma}\cdot d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s)A_c^{\pi_\theta}(s,a) \ , \tag{50}$$

where $A_c^{\pi_\theta}$ is defined in (49).

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

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

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

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

Lemma C.3. *For $c \in [M]$, the function J_c is $L_{2,\mathrm{s}} = 8/(1-\gamma)^3$-smooth and J is also $L_{2,\mathrm{s}}$-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 C.4. *For all $c \in [M]$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds*

$$\|\nabla J_c(\theta)\|_2 \leq L_{1,\mathrm{s}} \ , \quad \text{where } L_{1,\mathrm{s}} \triangleq \frac{1}{(1-\gamma)^2} \ .$$

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

$$\|J_c(\theta)\|_2 \leq \|J_c(\theta)\|_1 \leq \frac{1}{1-\gamma}\sum_{s,a} d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s)|A_c^{\pi_\theta}(s,a)| \ .$$

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

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

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

Proof. By Lemma F.10 with $\lambda = 0$ we have for any $u, v, w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

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

Next, we notice that

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

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

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

$$\|\nabla J(\theta) - \nabla J_c(\theta)\|_2^2 \leq \zeta_{\mathrm{s}}^2 \ , \quad \text{where } \zeta_{\mathrm{s}}^2 \triangleq \frac{56\varepsilon_{\mathsf{p}}^2}{(1-\gamma)^6} + \frac{36\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4} \ .$$

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

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

Lemma C.7 (Lemmas 6 and 7 from Ding et al. (2025))**.** *Consider the stochastic gradient defined in (45). For any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\|\nabla J_c(\theta) - \mathrm{g}_{c,\mathrm{s}}(\theta)\|_2 \leq \beta_{\mathrm{s}} \triangleq \frac{2\gamma^T}{1-\gamma}\left(T + \frac{1}{1-\gamma}\right) \ ,$$
$$\mathrm{Var}(\mathrm{g}_{c,\mathrm{s}}^{Z_c}(\theta)) \leq \sigma_{2,\mathrm{s}}^2 \triangleq \frac{12}{B(1-\gamma)^4} \ .$$

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

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

$$\mathbb{E}_{Z_c\sim\nu_c(\theta)}\left[\|\mathrm{g}_{c,\mathrm{s}}^{Z_c}(\theta) - \mathrm{g}_{c,\mathrm{s}}(\theta)\|_2^4\right] \le \sigma_{4,\mathrm{s}}^4 \triangleq \frac{1120}{B^2(1-\gamma)^8} \ . \tag{51}$$

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

The preceding lemmas conclude to prove that **FL**-1 to **FL**-6 hold. The following lemma establishes that **QL**-1 hold.

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

$$\|\nabla J_c(\theta)\|_2^2 \ge 2\mu_{c,\mathrm{s}}(\theta)\cdot[J_c^\star - J_c(\theta)]^2 \ ,$$

where

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

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.

C.2 Convergence rates, sample, and communication complexities

Using Theorem B.6, and Corollary B.7, we derive the following convergence rates.

Theorem C.10 (Convergence rates of `S-FedPG`)**.** *Assume $\boldsymbol{A}_\rho$ and no projection ($\mathcal{T}\colon \theta \to \theta$). Additionally, assume that there exists $1 > \underline{\mu_{\mathrm{s}}} > 0$ such that such that $\inf_{r\in\mathbb{N}} \mu_{\mathrm{s}}(\theta^r) \ge \underline{\mu_{\mathrm{s}}}$. For any $\eta > 0$ such that $\eta H L_{2,\mathrm{s}} \le 1/74$ the iterates of `S-FedPG` satisfy*

$$\begin{aligned} J^\star - \mathbb{E}[J(\theta^R)] \le{}& \frac{J^\star - J(\theta^0)}{1 + R\cdot(J^\star - J(\theta^0))\cdot(\eta H\underline{\mu_{\mathrm{s}}}/4)} + \left(\frac{6\eta L_{2,\mathrm{s}}\sigma_{2,\mathrm{s}}^2}{M\underline{\mu_{\mathrm{s}}}}\right)^{1/2} \\ &+ \left(\frac{16\cdot 12^3\eta^4 L_{3,\mathrm{s}}^2 H(H-1)\sigma_{4,\mathrm{s}}^4}{\underline{\mu_{\mathrm{s}}}}\right)^{1/2} + \left(\frac{2\zeta_{\mathrm{s}}^2}{\underline{\mu_{\mathrm{s}}}}\right)^{1/2} + \left(\frac{8\beta_{\mathrm{s}}^2}{\underline{\mu_{\mathrm{s}}}}\right)^{1/2} \\ &+ \frac{\zeta_{\mathrm{s}}^2}{12L_{2,\mathrm{s}}} + \frac{\eta\sigma_{2,\mathrm{s}}^2}{4M} + \frac{\beta_{\mathrm{s}}^2}{3L_{2,\mathrm{s}}} + \frac{12^3\eta^4 L_{3,\mathrm{s}}^2 H(H-1)\sigma_{4,\mathrm{s}}^4}{L_{2,\mathrm{s}}} \ . \end{aligned}$$

Proof. First note that the combination of lemmas of Appendix C.1 and the assumption made on the trajectory of the iterates implies that Assumptions **FL**-1 to **FL**-6, **QL**-1, and **QL**-2 hold. Importantly note that if $\eta H L_{2,\mathrm{s}} \le 1/74$ then it holds that $32\eta^2 H^2 L_{3,\mathrm{s}}^2 L_{1,\mathrm{s}}^2 \le L_{2,\mathrm{s}}^2$ (as $32 L_{3,\mathrm{s}}^2 L_{1,\mathrm{s}}^2 \le 74^2 L_{2,\mathrm{s}}^4$ by Lemma C.3, Lemma C.4, and Lemma C.5). Thus, applying Theorem B.6 concludes the proof. □

Recall that

$$\begin{gathered} L_{1,\mathrm{s}} = \frac{1}{(1-\gamma)^2} \ , \quad L_{2,\mathrm{s}} = \frac{8}{(1-\gamma)^3} \ , \quad L_{3,\mathrm{s}} = \frac{480}{(1-\gamma)^4} \ , \quad \zeta_{\mathrm{s}}^2 = \frac{56\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} + \frac{36\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4} \ , \\ \beta_{\mathrm{s}} = \frac{2\gamma^T T}{1-\gamma} + \frac{2\gamma^T}{(1-\gamma)^2} \ , \quad \sigma_{2,\mathrm{s}}^2 = \frac{12}{(1-\gamma)^4 B} \ , \quad \sigma_{4,\mathrm{s}}^4 = \frac{1120}{(1-\gamma)^8 B^2} \ , \end{gathered}$$

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

Corollary C.11 (Simplified convergence rates of `S-FedPG`)**.** *Under the assumptions of Theorem C.10, for any $\eta > 0$ such that $\eta H \le (1-\gamma)^3/592$, $T \ge 4(1-\gamma)^2$, and $M\cdot B \ge (1-\gamma)^{-1}$, the iterates of `S-FedPG` satisfy*

$$\begin{aligned} J^\star - \mathbb{E}[J(\theta^R)] \le{}& \frac{J^\star - J(\theta^0)}{1 + R\cdot(J^\star - J(\theta^0))\cdot\eta H\mu_{\mathrm{s}}/4} + \frac{24\eta^{1/2}}{\mu_{\mathrm{s}}^{1/2} M^{1/2} B^{1/2}\cdot(1-\gamma)^{3.5}} \\ &+ \frac{14^7\eta^2 H^{1/2}(H-1)^{1/2}}{\mu_{\mathrm{s}}^{1/2}(1-\gamma)^8 B} + \frac{13T\gamma^T}{\mu_{\mathrm{s}}^{1/2}(1-\gamma)} + \frac{13\varepsilon_{\mathsf{P}}}{\mu_{\mathrm{s}}^{1/2}(1-\gamma)^3} + \frac{4\varepsilon_{\mathsf{r}}}{\mu_{\mathrm{s}}^{1/2}(1-\gamma)^2} \ . \end{aligned}$$

Corollary C.12 (Sample and Communication Complexity)**.** *Under the assumptions of Theorem C.10, for any* $T \geq 4(1-\gamma)^{-2}$*, and* $M \cdot B \geq (1-\gamma)^{-1}$ *let*

$$\epsilon > \frac{94\varepsilon_{\mathsf{P}}}{(1-\gamma)^3 \underline{\mu_{\mathsf{s}}}^{1/2}} + \frac{75\varepsilon_{\mathsf{r}}}{(1-\gamma)^2 \underline{\mu_{\mathsf{s}}}^{1/2}} + \frac{90\gamma^T T}{(1-\gamma)\underline{\mu_{\mathsf{s}}}^{1/2}} \ ,$$

and

$$\eta \leq \min\left(\frac{(1-\gamma)^3}{48}, \frac{(1-\gamma)^7 \underline{\mu_{\mathsf{s}}} BM\epsilon^2}{20736}, \frac{\underline{\mu_{\mathsf{s}}}^{1/2}\epsilon(1-\gamma)^5 B}{2008}\right) \ .$$

In this case `S-FedPG` *achieves* $J^\star - J(\theta^0) \leq \epsilon$*, with a number of communication*

$$R \geq \frac{8640\left[J^\star - J(\theta^0) - \epsilon/6\right]}{(J^\star - J(\theta^0))\underline{\mu_{\mathsf{s}}}\epsilon} \cdot \frac{1}{(1-\gamma)^3} \ ,$$

for a total number of trajectories sampled per agent of

$$RHB \geq \frac{144[J^\star - J(\theta^0) - \epsilon/6]}{(J^\star - J(\theta^0))\underline{\mu_{\mathsf{s}}}\epsilon} \max\left(\frac{8B}{(1-\gamma)^3}, \frac{3464}{(1-\gamma)^7 \underline{\mu_{\mathsf{s}}} M\epsilon^2}, \frac{335}{\underline{\mu_{\mathsf{s}}}^{1/2}\epsilon(1-\gamma)^5}\right) \ .$$

D Analysis of RS-FedPG

`RS-FedPG` is a special instance of `proj-FedAVG` in which, the local objective function is $f_c = \tilde{J}_{c,\lambda} =: J_c + \lambda \mathcal{H}_c^\rho$, where for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ we have

$$\mathcal{H}_c^\rho(\theta) \triangleq -\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \log(\pi_\theta(A_c^t|S_c^t)) \mid S_c^0 \sim \rho\right] \ . \tag{52}$$

We additionally define the global objective of the algorithm as $\tilde{J}_\lambda \triangleq \frac{1}{M}\sum_{c=1}^M \tilde{J}_{c,\lambda}$. The client-specific data distribution $\xi_c(\theta)$ corresponds to $\nu_c(\theta)$, as defined in Eq. (44). 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 $\tilde{J}_{c,\lambda}$ as:

$$\tilde{\mathsf{g}}_{c,\lambda}^{Z_c}(\theta) \triangleq \frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T-1}\gamma^t\left(\sum_{\ell=0}^{t}\nabla\log\pi_\theta(a_{c,b}^\ell \mid S_{c,b}^\ell)\right)\left[\mathsf{r}_c(S_{c,b}^t, A_{c,b}^t) - \lambda\log(\pi_\theta(A_{c,b}^t, S_{c,b}^t))\right] \ . \tag{53}$$

We also define

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

To apply the convergence results of Appendix B, it remains to verify that Assumptions **FL**-1 to **FL**-6 , **PL**-1, and **PL**-2 hold. We establish these conditions in the following.

D.1 Checking the assumptions

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):

$$\widetilde{V}_c^{\pi_\theta}(s) \triangleq V_c^{\pi_\theta}(s) + \lambda\mathcal{H}_c^s(\theta) \ , \ \text{ where } \mathcal{H}_c^s(\theta) = -\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\log(\pi_\theta(A_c^t|S_c^t)) \mid S_c^0 = s\right] \tag{54}$$

$$\widetilde{Q}_c^{\pi_\theta}(s,a) \triangleq \mathsf{r}(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}_c(s'|s,a)\widetilde{V}_c^{\pi_\theta}(s') \ , \tag{55}$$

$$\widetilde{A}_c^{\pi_\theta}(s,a) \triangleq \widetilde{Q}_c^{\pi_\theta}(s,a) - \lambda\log(\pi_\theta(a \mid s)) - \widetilde{V}_c^{\pi_\theta}(s) \ , \quad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A} \ . \tag{56}$$

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*

$$\frac{\partial \tilde{J}_{c,\lambda}(\theta)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s)\widetilde{A}_c^{\pi_\theta}(s,a) \ , \tag{57}$$

where $\widetilde{A}_c^{\pi_\theta}$ *is defined in* (56).

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

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

$$\|\tilde{\mathrm{g}}_{c,\lambda}(\theta') - \tilde{\mathrm{g}}_{c,\lambda}(\theta)\| \leq \tilde{L}_{2,\lambda}\|\theta' - \theta\|_2 \ .$$

Proof. Fix any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $c \in [M]$. Let $\mathfrak{T} \triangleq (S^0, A^0, \dots, S^{T-1}, A^{T-1})$ be a random variable distributed according to $\nu_c(\theta)$, as defined in (44). Then, $\tilde{\mathrm{g}}_{c,\lambda}(\theta)$ can be equivalently expressed as

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

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 F.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

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

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

$$\frac{\partial \pi_\theta(a \mid s)}{\partial \theta(\bar{s}, \bar{a})} = \mathsf{1}_{\bar{s}}(s)(\mathsf{1}_{\bar{a}}(a)\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(\bar{a}|s)) \ . \tag{58}$$

Bounding $\mathrm{F}_\ell^t(s, a, \bar{s}, \bar{a})$**.** Using (58), note that, for any (s, a, s^ℓ, a^ℓ), we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which concludes the proof. □

Lemma D.3. *For any $c \in [M]$, $\tilde{J}_{c,\lambda}$ and $\tilde{J}_\lambda$ are $\tilde{L}_{2,\lambda} \triangleq (8 + \lambda(4 + 8\log(|\mathcal{A}|))/(1-\gamma)^3$-smooth.*

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

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

$$\|\nabla \tilde{J}_{c,\lambda}(\theta)\|_2 \le \tilde{L}_{1,\lambda} \ , \qquad \text{where } \tilde{L}_{1,\lambda} \triangleq \frac{1 + \lambda \log(|\mathcal{A}|)}{(1-\gamma)^2} \ .$$

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

$$\|\tilde{J}_{c,\lambda}(\theta)\|_2 \le \|\tilde{J}_{c,\lambda}(\theta)\|_1 = \frac{1}{1-\gamma} \sum_{s,a} d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s)|\widetilde{A}_c^{\pi_\theta}(s,a)| \ .$$

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

$$\|\tilde{J}_{c,\lambda}(\theta)\|_2 \le \|\tilde{J}_{c,\lambda}(\theta)\|_1 \le \frac{1 + \lambda\log(|\mathcal{A}|)}{(1-\gamma)^2} \sum_{s,a} d_c^{\rho,\pi_\theta}(s)\pi_\theta(a|s) = \frac{1 + \lambda\log(|\mathcal{A}|)}{(1-\gamma)^2} \ .$$

which concludes the proof. □

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

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

Proof. By Lemma F.10 we have for any $u, v, w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

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

Next, we notice that

$$\mathrm{d}^3\tilde{J}_{c,\lambda}(\theta)[u,v,w] = \rho^\top \mathrm{d}^3 V_c^{\pi_\theta}[u,v,w] \, ,$$

thus

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

□

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

$$\|\nabla\tilde{J}_\lambda(\theta) - \nabla\tilde{J}_{c,\lambda}(\theta)\|_2^2 \le \tilde{\zeta}_\lambda^2 \ , \qquad \text{where } \tilde{\zeta}_\lambda^2 \triangleq \frac{56(1 + \lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathrm{P}}^2}{(1-\gamma)^6} + \frac{36\varepsilon_{\mathrm{r}}^2}{(1-\gamma)^4} \ .$$

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

$$\left|\frac{\partial\tilde{J}_\lambda(\theta)}{\partial\theta(s,a)} - \frac{\partial\tilde{J}_{c,\lambda}(\theta)}{\partial\theta(s,a)}\right| \le \frac{1}{M}\frac{1}{1-\gamma}\sum_{k=1}^M \pi_\theta(a|s)\left|d_k^{\rho,\pi_\theta}(s)\widetilde{A}_k^{\pi_\theta}(s,a) - d_c^{\rho,\pi_\theta}(s)\widetilde{A}_c^{\pi_\theta}(s,a)\right|$$

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

We bound each of (**X**) and (**Y**) separately.

Bounding (X). First note that we have

$$0 \leq \widetilde{V}_c^{\pi_\theta}(s) \leq \frac{1+\lambda \log(|\mathcal{A}|)}{1-\gamma} \ , \tag{59}$$

where $\widetilde{V}_c^{\pi_\theta}(s)$ is defined in (54). Combining the previous inequality and applying the triangle inequality yields

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

Bounding (Y). Using the triangle inequality, we get

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

where in the last inequality, we used that $\varepsilon_{\mathsf{r}} = \max_{(c,c') \in [M]^2} \|\mathsf{r}_c - \mathsf{r}_{c'}\|$ (defined in Section 3). Using $\varepsilon_{\mathsf{P}} = \max_{(s,a,(c,c')) \in \mathcal{S} \times \mathcal{A} \times [M]^2} \|\mathsf{P}_c(\cdot|s,a) - \mathsf{P}_{c'}(\cdot|s,a)\|_1$, combined with (59), we obtain

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

Using (54), 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 F.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 F.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 (**Y**)

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

Combining the bounds on (**X**) and (**Y**) yields

$$\begin{aligned}&\left| \frac{\partial \tilde{J}_\lambda(\theta)}{\partial \theta(s,a)} - \frac{\partial \tilde{J}_{c,\lambda}(\theta)}{\partial \theta(s,a)} \right| \\ &\leq \frac{1}{M} \sum_{k=1}^{M} \left[\frac{2(1+\lambda \log(|\mathcal{A}|))}{1-\gamma} + \lambda |\log(\pi(a|s))| \right] \left| d_k^{\rho,\theta}(s) - d_c^{\rho,\theta}(s) \right| \frac{\pi_\theta(a|s)}{1-\gamma}\end{aligned}$$

$$+\left[\frac{3(1+\lambda\log(|\mathcal{A}|))\varepsilon_{\mathsf{P}}}{(1-\gamma)^2}+\frac{3\varepsilon_{\mathsf{r}}}{1-\gamma}\right]\cdot\frac{d_c^{\rho,\theta}(s)\pi_\theta(a|s)}{1-\gamma}$$

Thus, we get by Young's inequality

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

Now applying Jensen's inequality yields

$$\begin{aligned}
&\|\nabla\tilde{J}_\lambda(\theta)-\nabla\tilde{J}_{c,\lambda}(\theta)\|_2^2\\
&\quad\le\frac{2}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^M\left[\frac{2(1+\lambda\log(|\mathcal{A}|))}{1-\gamma}+\lambda|\log(\pi(a|s))|\right]^2\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|^2\frac{\pi_\theta(a|s)^2}{(1-\gamma)^2}\\
&\quad+\frac{2}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^M\left[\frac{18(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}+\frac{18\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4}\right]\cdot d_c^{\rho,\theta}(s)^2\pi_\theta(a|s)^2\\
&\quad\le\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^M\frac{16(1+\lambda\log(|\mathcal{A}|))^2}{(1-\gamma)^2}\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|^2\frac{\pi_\theta(a|s)^2}{(1-\gamma)^2}\\
&\quad+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^M4\lambda^2|\log(\pi(a|s))|^2\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|^2\frac{\pi_\theta(a|s)^2}{(1-\gamma)^2}\\
&\quad+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^M\left[\frac{36(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}+\frac{36\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4}\right]\cdot d_c^{\rho,\theta}(s)^2\pi_\theta(a|s)^2\ ,
\end{aligned}$$

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

$$\begin{aligned}
\|\nabla\tilde{J}_\lambda(\theta)-\nabla\tilde{J}_{c,\lambda}(\theta)\|_2^2\quad&\le\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^M\frac{16(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}}{(1-\gamma)^3}\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|\frac{\pi_\theta(a|s)}{(1-\gamma)^2}\\
&+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{c=1}^M4\lambda^2|\log(\pi(a|s))|\left|d_k^{\rho,\theta}(s)-d_c^{\rho,\theta}(s)\right|\frac{\pi_\theta(a|s)\varepsilon_{\mathsf{P}}}{(1-\gamma)^3}\\
&+\frac{1}{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^M\left[\frac{36(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}+\frac{36\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4}\right]\cdot d_c^{\rho,\theta}(s)\pi_\theta(a|s)\ ,
\end{aligned}$$

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 F.6 for both the first and second term yields

$$\|\nabla\tilde{J}_\lambda(\theta)-\nabla\tilde{J}_{c,\lambda}(\theta)\|_2^2\le\frac{56(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6}+\frac{36\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4}\ ,$$

which concludes the proof. □

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

Lemma D.7 (Lemma 6 from Ding et al. (2025)). *Consider the stochastic gradient defined in (53). We have*

$$\|\tilde{\mathrm{g}}_{c,\lambda}(\theta) - \nabla \tilde{J}_{c,\lambda}(\theta)\|_2 \leq \tilde{\beta}_\lambda \triangleq \frac{2(1+\lambda\log(|\mathcal{A}|))\gamma^T}{1-\gamma}\left(T + \frac{1}{1-\gamma}\right) \ ,$$
$$\mathrm{Var}(\tilde{\mathrm{g}}_{c,\lambda}^{Z_c}) \leq \tilde{\sigma}_{2,\lambda}^2 \triangleq \frac{12 + 24\lambda^2(\log(|\mathcal{A}|))^2}{B(1-\gamma)^4} \ .$$

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

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

$$\mathbb{E}_{Z_c\sim\nu_c(\theta)}\left[\|\tilde{\mathrm{g}}_{c,\lambda}^{Z_c}(\theta) - \tilde{\mathrm{g}}_{c,\lambda}(\theta)\|_2^4\right] \leq \tilde{\sigma}_{4,\lambda}^4 \triangleq \frac{1120 + 4480\lambda^4\log(|\mathcal{A}|)^4}{B^2(1-\gamma)^8} \ . \tag{60}$$

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

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

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 $\tilde{\mathrm{g}}_{c,\lambda}^{Z_c}(\theta)$ by the fourth central moment of $\mathrm{u}(Z_{c,b})$. Indeed, expanding the norm to the fourth power yields

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

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:

$$\mathbb{E}_{Z_c}\left[\|\tilde{\mathrm{g}}_{c,\lambda}^{Z_c}(\theta) - \tilde{\mathrm{g}}_{c,\lambda}(\theta)\|_2^4\right] = \frac{1}{B^3}\left[\underbrace{\mathbb{E}\left[\|\mathrm{u}(Z_{c,1})-\bar{\mathrm{u}}\|_2^4\right]}_{(\mathbf{M})} + 3(B-1)\mathbb{E}\left[\|\mathrm{u}(Z_{c,1})-\bar{\mathrm{u}}\|^2\right]^2\right] \ . \tag{61}$$

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

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

Thus, using the triangle inequality, we have

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

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 $\|\nabla \log(\pi_\theta(a \mid s))\|_2 \le 2$ (see, e.g., proof of Lemma 7 in Ding et al. (2025)), gives

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

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

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

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

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

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

Next, applying Cauchy-Schwarz inequality gives

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

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

$$(\mathbf{M_2}) \le \lambda^4 \frac{64}{(1-\gamma)^6} \mathbb{E}_{Z_c} \left[\frac{1-\gamma^T}{1-\gamma} \sum_{t=0}^{T-1} \gamma^t |\log(\pi_\theta(A_{c,1}^t, S_{c,1}^t))|^4 \right]$$

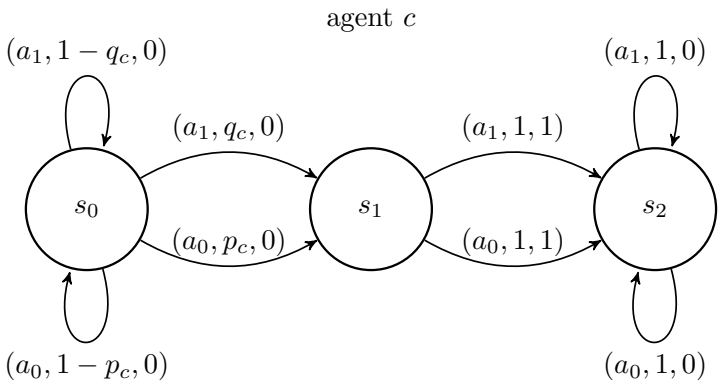


Figure 3: FRL task with an objective that admits strictly local minimas. The triplet means (action, probability, reward) , $\gamma = 0.999$, and $\lambda = 1$. If the action is not specified, it means that all the actions give the same reward and lead to the same state.

$$\leq \lambda^4 \frac{64}{(1-\gamma)^7} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{Z_c} \left[|\log(\pi_\theta(A_{c,1}^t \mid S_{c,1}^t))|^4 \right] \ . \tag{62}$$

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

$$\max_{\pi \in \mathcal{P}(\mathcal{A})} - \sum_{a \in \mathcal{A}} \pi(a \mid s) \log(\pi(a \mid s))^4 = (\log(|\mathcal{A}|))^4 \ . \tag{63}$$

Plugging in the previous bound in (62) 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 (61) concludes the proof. □

We first show that, in general, the objective $\tilde{J}_\lambda$ does not have Łojasiewicz structure.

Lemma D.9. *There exists an FRL instance where the objective $\tilde{J}_\lambda$ admits a strict local minima.*

Proof. Consider the FRL task defined in Figure 3. Define $x \triangleq \pi_\theta(a_1|s_0)$. From the flow conservation constraints for occupancy measures for any agent $c \in [M]$, it holds that

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

Rearranging the precedent terms yields

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

which implies

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

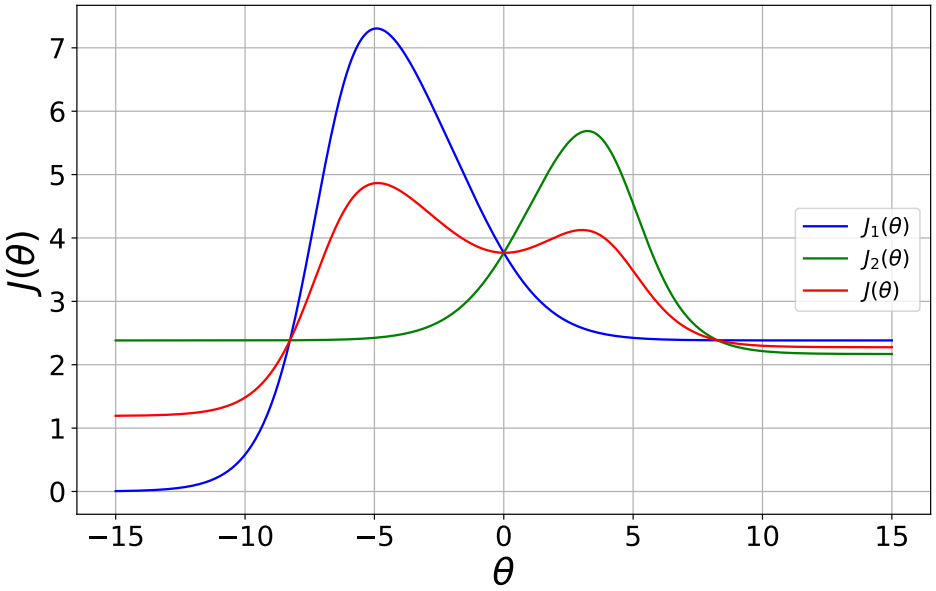


Figure 4: An example that shows that FRL objective $\tilde{J}_\lambda$ does not necessarily have a Łojasiewicz structure.

The value of the objective function is thus

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

where for any $y \in (0,1), H(y) \triangleq -y\log y - (1-y)\log(1-y)$. 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

$$\tilde{J}_{c,\lambda}(\theta) = \frac{\tau H(\sigma(\theta)) + \gamma \cdot f(\sigma(\theta); p_c, q_c) \cdot (1 + \tau\log 2 + \gamma\tau\log 2)}{1 - \gamma(1 - f(\sigma(\theta); p_c, q_c))}.$$

The plot of $\tilde{J}_\lambda$ (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 4 shows that this problem does not have a Łojasiewicz structure. □

However, each agent locally satisfies a Łojasiewicz-type property

Lemma D.10 (Lemma 15 of Mei et al. (2020)). *Assume* $\boldsymbol{A}_\rho$. *For any agent* $c \in [M]$, *denote by* $\pi^\star_{c,\lambda}$ *the unique optimal regularized policy (see e.g Nachum et al. (2017) for the proof of existence and unicity) of this agent. It holds*

$$\|\nabla \tilde{J}_{c,\lambda}(\theta)\|_2^2 \geq 2\tilde{\mu}_{c,\lambda}(\theta)\left[\tilde{J}^\star_{c,\lambda} - \tilde{J}_{c,\lambda}(\theta)\right] ,$$

where $\tilde{\mu}_{c,\lambda}(\theta)$ *is defined as*

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

D.2 Construction of the projection operator

The goal of this section is therefore to show the existence of an operator $\mathcal{U}$ with two crucial properties: (i) for any policy and agent, applying this operator produces a new policy with a higher regularized value, and; (ii) Every policy generated by this operator assigns at least a fixed minimum probability to every action. The main idea is to build the improvement operator such that it slightly augments the smallest probability weights, such that for any state action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ the probability $\pi(a|s)$ stays above a certain threshold. We will show below that this procedure improves the global objective while keeping the probabilities uniformly bounded away from 0 when the threshold is properly chosen. For any policy π, state $s \in \mathcal{S}$, $\tau < 1/(2|\mathcal{A}|^2)$, we respectively define $\mathcal{A}_\tau^\pi(s)$, and $a_{\max}^\pi(s)$ as

$$\mathcal{A}_\tau^\pi(s) \triangleq \{a \in \mathcal{A}, \pi(a|s) \leq \tau/2\} \quad , \quad a_{\max}^\pi(s) = \arg\max_{a \in \mathcal{A}} \{\pi(a|s)\} \ ,$$

where the arg max is chosen at random in the case of ties. Note that the definition of τ_λ ensures that $a_{\max}^\pi(s)$ does not belong to the set $\mathcal{A}_\tau^\pi(s)$ as

$$\max_{a \in \mathcal{A}} \pi(a|s) \geq 1/|\mathcal{A}| \ .$$

Finally, we define the improvement operator as follows:

$$\begin{aligned} \mathcal{U}_\tau : \mathcal{P}(\mathcal{A})^{\mathcal{S}} &\longrightarrow \mathcal{P}(\mathcal{A})^{\mathcal{S}}, \\ \pi &\longmapsto \mathcal{U}_\tau(\pi), \end{aligned} \tag{64}$$

where for every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{U}_\tau(\pi)(a|s) = \begin{cases} \tau, & \text{if } \pi(a|s) \leq \tau/2, \\ \pi(a|s) - \displaystyle\sum_{b \in \mathcal{A}_\tau^\pi(s)} \big(\tau - \pi(b|s)\big), & \text{if } a = a_{\max}^\pi(s), \\ \pi(a|s), & \text{otherwise.} \end{cases}$$

The operator $\mathcal{U}_\tau$ builds $\mathcal{U}_\tau(\pi)(a|s)$ by (statewise) raising each $a \in \mathcal{A}_\tau^\pi(s)$ to τ, substracting the total added mass from the single action $a_{\max}^\pi(s)$, and leaving other actions unchanged. If $\mathcal{A}_\tau^\pi(s) = \emptyset$, for all $s \in \mathcal{S}$, then $\mathcal{U}_\tau(\pi) = \pi$. Note that mass conservation is immediate from the definition and the fact that $\tau < 1/(2|\mathcal{A}|^2)$. Non-negativity of $\mathcal{U}_\tau(\pi)(a_{\max}^\pi(s)|s)$ follows because the removed mass is

$$\sum_{a \in \mathcal{A}_\tau^\pi(s)} \{\tau - \pi(a|s)\} \leq \tau \times |\mathcal{A}| \leq \frac{1}{2|\mathcal{A}|}$$

Since $\pi(a_{\max}^\pi(s)|s) \geq 1/|\mathcal{A}|$, we get that $\mathcal{U}(\pi)(a_{\max}^\pi(s)|s) \geq 1/(2|\mathcal{A}|)$. This in particular shows that $\mathcal{U}_\tau(\pi)$ is a policy. Next, define

$$\tau_\lambda \triangleq \min\left(\frac{1}{3}\exp\left(-\frac{16 + 8\gamma\lambda\log(|\mathcal{A}|)}{\lambda(1-\gamma)^2\rho_{\min}}\right), \frac{1}{3^8|\mathcal{A}|^4}\right) \ . \tag{65}$$

The following lemma establishes the crucial improvement property when $\tau = \tau_\lambda$.

Lemma D.11. *Assume that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. For any policy π, for any agent $c \in [M]$, it holds that*

$$\widetilde{V}_c^{\mathcal{U}_{\tau_\lambda}(\pi)}(\rho) \geq \widetilde{V}_c^{\pi}(\rho) \ .$$

Additionally, for any policy π, we have that

$$\mathcal{U}_{\tau_\lambda}(\pi)(a|s) \geq \tau_\lambda/2 \ .$$

Proof. Set an arbitrary policy π. For avoiding heavy notations, we will, through this proof, denote by $\mathcal{A}_\tau^\pi = \mathcal{A}_{\tau_\lambda}^\pi$. We consider the case where there is $s \in \mathcal{S}$ such that $\mathcal{A}_\tau^\pi(s) \neq \emptyset$ (alternatively $\mathcal{U}_{\tau_\lambda}(\pi) = \pi$, which makes the previous inequality immediately valid). Define $\tilde{\pi} = \mathcal{U}_{\tau_\lambda}(\pi)$. The following applies

$$\widetilde{V}_c^{\tilde{\pi}}(\rho) - \widetilde{V}_c^{\pi}(\rho) = \sum_{s \in \mathcal{S}} d_c^{\rho,\tilde{\pi}}(s) \sum_{a \in \mathcal{A}} \left[\tilde{\pi}(a|s)\mathsf{r}(s,a) - \lambda\tilde{\pi}(a|s)\log\left(\tilde{\pi}(a|s)\right)\right]$$

$$
\begin{aligned}
&\quad - \sum_{s\in\mathcal{S}} d_c^{\rho,\pi}(s) \sum_{a\in\mathcal{A}} \left[\pi(a|s)\mathsf{r}(s,a) - \lambda\pi(a|s)\log\left(\pi(a|s)\right)\right] \\
&= \underbrace{\sum_{s\in\mathcal{S}} \left(d_c^{\rho,\tilde{\pi}}(s) - d_c^{\rho,\pi}(s)\right) \sum_{a\in\mathcal{A}} \left[\tilde{\pi}(a|s)\mathsf{r}(s,a) - \lambda\tilde{\pi}(a|s)\log\left(\tilde{\pi}(a|s)\right)\right]}_{(\mathbf{I})} \\
&+ \underbrace{\sum_{s\in\mathcal{S}} d_c^{\rho,\pi}(s) \sum_{a\in\mathcal{A}} (\tilde{\pi}(a|s) - \pi(a|s))\mathsf{r}(s,a)}_{(\mathbf{II})} \\
&+ \underbrace{\lambda \sum_{s\in\mathcal{S}} d_c^{\rho,\pi}(s) \sum_{a\in\mathcal{A}} \left[\pi(a|s)\log\left(\pi(a|s)\right) - \tilde{\pi}(a|s)\log\left(\tilde{\pi}(a|s)\right)\right]}_{(\mathbf{III})} \quad .
\end{aligned}
$$

We now lower-bound each of the three terms separately.

Bounding (I). Using Lemma F.7, we have

$$
\begin{aligned}
(\mathbf{I}) &\geq -\|d_c^{\rho,\tilde{\pi}} - d_c^{\rho,\pi}\|_1 \max_{s\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}} \left[\tilde{\pi}(a|s)\mathsf{r}(s,a) - \lambda\tilde{\pi}(a|s)\log\left(\tilde{\pi}(a|s)\right)\right] \right| \\
&\geq -\frac{\gamma}{1-\gamma} \sup_{s\in\mathcal{S}} \|\tilde{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 \left(1 + \lambda\log(|\mathcal{A}|)\right) \\
&\geq -\frac{2\gamma}{1-\gamma}\tau_\lambda \max_{s\in\mathcal{S}} \left\{ \sum_{a\in\mathcal{A}_\tau^\pi(s)} 1 \right\} \left(1 + \lambda\log(|\mathcal{A}|)\right) \quad ,
\end{aligned}
$$

where in the last inequality we used that (because we increase the probability of the actions in $A_\tau^\pi(s)$ by τ_λ and remove the total added mass from the probability of $\pi(a_{\max}^\pi(s))$

$$
\sup_{s\in\mathcal{S}} \|\tilde{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 \leq 2 \max_{s\in\mathcal{S}} \left\{ \sum_{a\in A_\tau^\pi(s)} \right\} \tau_\lambda \quad .
$$

Bounding (II). Using the triangle inequality yields

$$
(\mathbf{II}) \geq -\sup_{s\in\mathcal{S}} \|\tilde{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 \geq -2 \max_{s\in\mathcal{S}} \left\{ \sum_{a\in\mathcal{A}_\tau^\pi(s)} 1 \right\} \tau_\lambda \quad .
$$

Bounding (III). All the state-action pairs on which the original π allocates the same probability then the policy $\tilde{\pi}$ are equal to 0 in (**III**) allowing us to simplify this term

$$
\begin{aligned}
(\mathbf{III}) &= \lambda \sum_{s\in\mathcal{S}} d_c^{\rho,\pi}(s) \sum_{a\in\mathcal{A}} \left[\pi(a|s)\log\left(\pi(a|s)\right) - \tilde{\pi}(a|s)\log\left(\tilde{\pi}(a|s)\right)\right] \\
&= \lambda \sum_{s\in\mathcal{S}} d_c^{\rho,\pi}(s) \sum_{a\in\mathcal{A}_\tau^\pi(s)} \left[\pi(a|s)\log\left(\pi(a|s)\right) - \tilde{\pi}(a|s)\log(\tilde{\pi}(a|s))\right] \\
&+ \lambda \sum_{s\in\mathcal{S}} \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) d_c^{\rho,\pi}(s) \left[\pi(a_{\max}^\pi(s)|s)\log\left(\pi(a_{\max}^\pi(s)|s)\right) - \tilde{\pi}(a_{\max}^\pi(s)|s)\log\left(\tilde{\pi}(a_{\max}^\pi(s)|s)\right)\right] \quad .
\end{aligned}
$$

Since $x \mapsto x\log(x)$ is convex, for all $u, v \in [0;1]$, $u\log(u) - v\log(v) \geq [\log(v)+1](u-v)$, we have

$$
\begin{aligned}
(\mathbf{III}) &\geq \lambda \sum_{s\in\mathcal{S}} d_c^{\rho,\pi}(s) \sum_{a\in\mathcal{A}_\tau^\pi(s)} \left(\pi(a|s) - \tilde{\pi}(a|s)\right) \left[\log(\tau_\lambda) + 1\right] \qquad (\text{since } \tilde{\pi}(a|s) = \tau_\lambda) \\
&+ \lambda \sum_{s\in\mathcal{S}} \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) d_c^{\rho,\pi}(s) \left[\pi(a_{\max}^\pi(s)|s) - \tilde{\pi}(a_{\max}^\pi(s)|s)\right] \left[\log\left(\tilde{\pi}(a_{\max}^\pi(s)|s)\right) + 1\right] \quad ,
\end{aligned}
$$

Next, using that

$$\tilde{\pi}(a^{\pi}_{\max}(s)|s) \geq \pi(a^{\pi}_{\max}(s)|s) - |\mathcal{A}|\tau_\lambda \geq \frac{1}{|\mathcal{A}|} - \frac{1}{2|\mathcal{A}|} = \frac{1}{2|\mathcal{A}|} \ ,$$

combined with the monotonicity of $x\colon \log(x)+1$ and the fact that $\pi(a^{\pi}_{\max}(s)|s) - \tilde{\pi}(a^{\pi}_{\max}(s)|s) \geq 0$ yields

$$\begin{aligned}(\mathbf{III}) \geq \lambda \sum_{s\in\mathcal{S}} d_c^{\rho,\tilde{\pi}}(s) \sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} (\pi(a|s) - \tilde{\pi}(a|s)) \left[\log(\tau_\lambda)+1\right] \\ + \lambda \sum_{s\in\mathcal{S}} 1(\mathcal{A}^{\pi}_{\tau}(s) \neq \emptyset) d_c^{\rho,\tilde{\pi}}(s) \left[\pi(a^{\pi}_{\max}(s)|s) - \tilde{\pi}(a^{\pi}_{\max}(s)|s)\right] \left[\log(\frac{1}{2|\mathcal{A}|})+1\right] \ ,\end{aligned}$$

Additionally, since

$$0 \leq \pi(a^{\pi}_{\max}(s)|s) - \tilde{\pi}(a^{\pi}_{\max}(s)|s) \leq \sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} (\pi(a|s) - \tilde{\pi}(a|s)) \leq \tau_\lambda \sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} 1 \ ,$$

implies

$$\begin{aligned}(\mathbf{III}) &\geq -\frac{\lambda}{2} \sum_{s\in\mathcal{S}} d_c^{\rho,\tilde{\pi}}(s) 1(\mathcal{A}^{\pi}_{\tau}(s) \neq \emptyset) \left(\sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} 1\right) \tau_\lambda \left[\log(\tau_\lambda)+1\right] \\ &\quad - \lambda \sum_{s\in\mathcal{S}} d_c^{\rho,\tilde{\pi}}(s) 1(\mathcal{A}^{\pi}_{\tau}(s) \neq \emptyset) \left(\sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} 1\right) \tau_\lambda [\log(2|\mathcal{A}|)+1] \ , \\ &\geq -\frac{\lambda}{4} \sum_{s\in\mathcal{S}} d_c^{\rho,\tilde{\pi}}(s) 1(\mathcal{A}^{\pi}_{\tau}(s) \neq \emptyset) \left(\sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} 1\right) \tau_\lambda \left[\log(\tau_\lambda)+1\right] \ ,\end{aligned}$$

where in the last inequality, we used that $\tau_\lambda \leq \frac{1}{3^8|\mathcal{A}|^4} \leq \exp(-4\log(2|\mathcal{A}|) - 5)$. Hence, by using $\mathbf{A}_\rho$, we can lower bound this term as follows

$$(\mathbf{III}) \geq -\frac{\lambda}{4}(1-\gamma)\rho_{\min} \max_{s\in\mathcal{S}} \left\{ \sum_{a\in\mathcal{A}^{\pi}_{\tau}(s)} 1 \right\} \tau_\lambda \left[\log(\tau_\lambda)+1\right] \ .$$

Collecting these lower bounds and using that

$$[\log(\tau_\lambda)+1] \leq -\frac{16 + 8\gamma\lambda\log(|\mathcal{A}|)}{\lambda(1-\gamma)^2\rho_{\min}}$$

concludes the proof. □

Finally, we define the operator that maps each policy to one corresponding parameter

$$\mathcal{L} : \Pi \ \to \ \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$$

by

$$\mathcal{L}(\pi)(s,a) \ \triangleq \ \log(\pi(a|s)), \quad \text{for all} (s,a) \in \mathcal{S}\times\mathcal{A} \, . \tag{66}$$

Finally, we define the improvement operator on the logitspace as

$$\mathcal{T}_\tau \triangleq \mathcal{L} \circ \mathcal{T}_\tau \ .$$

The following lemma shows that $\mathcal{L}_\tau$ successfully recovers a parameter that gives the policy and that $\mathcal{T}_\tau$ improves the value of the objective when $\lambda = \tau_\lambda$.

Lemma D.12. *Assume that the initial distribution ρ satisfies $\mathbf{A}_\rho$. For any policy π, it holds that*

$$\pi_{\mathcal{L}(\pi)} = \pi \ ,$$

Additionally, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have that

$$\widetilde{V}_{\mathcal{T}_{\tau_\lambda}(\theta)}(\rho) \geq \widetilde{V}_\theta(\rho) \ , \quad \pi_{\mathcal{T}_{\tau_\lambda}(\theta)} \geq \tau_\lambda \ .$$

Proof. The proof follows immediately from the definition of the softmax policy, from (66), and Lemma D.11. □

D.3 Convergence rates, sample and communication complexities

Firstly, define

$$\tilde{\mu}_\lambda(\theta) \triangleq \min_{c\in[M]} \tilde{\mu}_{c,\lambda}(\theta) \ , \quad \underline{\tilde{\mu}}_\lambda \triangleq \lambda(1-\gamma)\rho_{\min}^2\tau_\lambda^2/|\mathcal{S}| \ . \tag{67}$$

where $\tilde{\mu}_{c,\lambda}(\theta)$ is defined Lemma D.10. Applying Theorem B.9 and Corollary B.10 yields the following convergence results.

Theorem D.13 (Convergence rates of RS-FedPG). *Assume $\boldsymbol{A}_\rho$ and that the projection operator is $\mathcal{T}_{\tau_\lambda}$. For any $\eta > 0$ such that $\eta H \tilde{L}_{2,\lambda} \le 1/74$, the iterates of RS-FedPG satisfy*

$$\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda(\theta^R)] \le \left(1 - \frac{\eta H \underline{\tilde{\mu}}_\lambda}{2}\right)^R (\tilde{J}_\lambda^\star - \tilde{J}_\lambda(\theta^0) + \frac{3\eta \tilde{L}_{2,\lambda}\tilde{\sigma}_{2,\lambda}^2}{M\underline{\tilde{\mu}}_\lambda} + \frac{\tilde{\zeta}_\lambda^2}{\underline{\tilde{\mu}}_\lambda} + 4\frac{\tilde{\beta}_\lambda^2}{\underline{\tilde{\mu}}_\lambda} + \frac{8\cdot 12^3 \eta^4 \tilde{L}_{3,\lambda}^2 H(H-1)\tilde{\sigma}_{4,\lambda}^4}{\underline{\tilde{\mu}}_\lambda} \ .$$

Proof. First, note that the combination of all the lemmas of Appendix D.1 shows that Assumptions **FL**-1 to **FL**-6 , **PL**-1, and **PL**-2 holds. Next, note that if $\eta H \tilde{L}_{2,\lambda} \le 1/74$ then it holds that $32\eta^2 H^2 \tilde{L}_{3,\lambda}^2 \tilde{L}_{1,\lambda}^2 \le \tilde{L}_{2,\lambda}^2$ (as $32\tilde{L}_{3,\lambda}^2\tilde{L}_{1,\lambda}^2 \le 74^2 \tilde{L}_{2,\lambda}^4$ by Lemma D.3, Lemma D.4, and Lemma D.5). Thus, applying Theorem B.9 concludes the proof. □

Recall that

$$\tilde{L}_{1,\lambda} = \frac{1+\lambda\log(|\mathcal{A}|)}{(1-\gamma)^2} \ , \quad \tilde{L}_{2,\lambda} = \frac{8+\lambda(4+8\log(|\mathcal{A}|)}{(1-\gamma)^3} \ , \quad \tilde{L}_{3,\lambda} = \frac{480+832\lambda\log|\mathcal{A}|}{(1-\gamma)^4} \ ,$$
$$\tilde{\zeta}_\lambda^2 = \frac{56(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{(1-\gamma)^6} + \frac{36\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4} \ , \quad \tilde{\beta}_\lambda = \frac{2(1+\lambda\log(|\mathcal{A}|))\gamma^T T}{1-\gamma} + \frac{2(1+\lambda\log(|\mathcal{A}|))\gamma^T}{(1-\gamma)^2} \ ,$$
$$\tilde{\sigma}_{2,\lambda}^2 = \frac{12+24\lambda^2(\log(|\mathcal{A}|))^2}{(1-\gamma)^4 B} \ , \quad \tilde{\sigma}_{4,\lambda}^4 = \frac{(1120+4480\lambda^4\log(|\mathcal{A}|)^4)}{(1-\gamma)^8 B^2} \ ,$$

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

Corollary D.14 (Explicit Convergence Rate of RS-FedPG). *Under the assumptions of Theorem D.13, let $\eta > 0$ such that $\eta H \le 888^{-1}(1-\gamma)^3(1+\lambda\log(|\mathcal{A}|))^{-1}$, and $T \ge 1/(1-\gamma)$. Then, the iterates of RS-FedPG satisfy*

$$\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda(\theta^R)] \le \left(1 - \frac{\eta H \underline{\tilde{\mu}}_\lambda}{2}\right)^R (\tilde{J}_\lambda^\star - \tilde{J}_\lambda(\theta^0) + \frac{864\eta(1+\lambda\log(|\mathcal{A}|)^3}{BM\underline{\tilde{\mu}}_\lambda(1-\gamma)^7} + \frac{56(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{\underline{\tilde{\mu}}_\lambda(1-\gamma)^6}$$
$$+ \frac{36(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{r}}^2}{\underline{\tilde{\mu}}_\lambda(1-\gamma)^6} + \frac{16(1+\lambda\log(|\mathcal{A}|))^2\gamma^{2T}T^2}{\underline{\tilde{\mu}}_\lambda(1-\gamma)^2} + \frac{51^8\eta^4 H(H-1)(1+\lambda\log(|\mathcal{A}|))^6}{\underline{\tilde{\mu}}_\lambda B^2(1-\gamma)^{16}} \ ,$$

where we recall that $\underline{\tilde{\mu}}_\lambda$ is defined in (67).

Corollary D.15 (Sample and Communication Complexity of RS-FedPG). *Under the assumptions of Theorem D.13, let*

$$\epsilon \ge \frac{224(1+\lambda\log(|\mathcal{A}|))^2\varepsilon_{\mathsf{P}}^2}{\underline{\tilde{\mu}}_\lambda(1-\gamma)^6} + \frac{144\varepsilon_{\mathsf{r}}^2}{(1-\gamma)^4\underline{\tilde{\mu}}_\lambda} + \frac{256(1+\lambda\log(|\mathcal{A}|))^2\gamma^{2T}T^2}{(1-\gamma)^2} \ .$$

and

$$\eta \le \min\left(\frac{(1-\gamma)^3}{72(1+\lambda\log(|\mathcal{A}|))}, \frac{\underline{\tilde{\mu}}\epsilon MB(1-\gamma)^7}{3456(1+\lambda\log(|\mathcal{A}|))^3}, \frac{\underline{\tilde{\mu}}^{1/2}B(1-\gamma)^{7/2}\epsilon^{1/2}}{9^6(1+\lambda\log(|\mathcal{A}|))^{3/2}}\right) \ .$$

Then RS-FedPG achieves $\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda\theta^R] \le \epsilon$, with a number of communication

$$R \ge \frac{12}{\underline{\tilde{\mu}}}\log\left(\frac{4(\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda\theta^0])}{\epsilon}\right)\frac{104(1+\lambda\log(|\mathcal{A}|))}{(1-\gamma)^3} \ ,$$

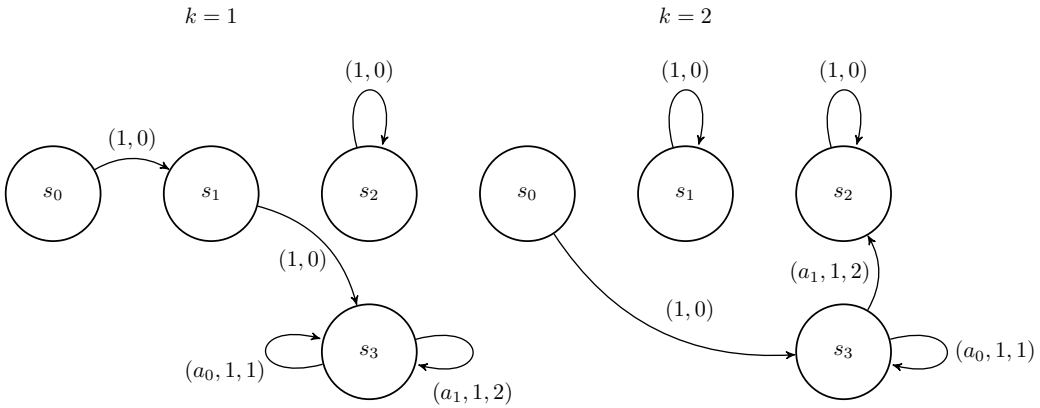


Figure 5: FRL task with no optimal stationary policy. The triplet means (action, probability, reward) and $\gamma = 0.9$. If the action is not specified, it means that all the actions give the same reward and lead to the same state

for a total number of sampled trajectories per agent of

$$RHB \geq \frac{2}{\underline{\tilde{\mu}}} \log\Big(\frac{4(\tilde{J}_\lambda^\star - \mathbb{E}[\tilde{J}_\lambda \theta^0])}{\epsilon}\Big) \max\Big(\frac{72(1+\lambda\log(|\mathcal{A}|))B}{(1-\gamma)^3}, \frac{3456(1+\lambda\log(|\mathcal{A}|))^3}{\underline{\tilde{\mu}}\epsilon M(1-\gamma)^7}, \frac{9^6(1+\lambda\log(|\mathcal{A}|))^{3/2}}{\underline{\tilde{\mu}}^{1/2}(1-\gamma)^{7/2}\epsilon^{1/2}}\Big) \ .$$

E On the different classes of policies

The goal of this section is to prove Theorem 6.1. 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

$$V_c^\pi(\rho) \triangleq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathsf{r}(S_c^t, A_c^t)\right] \ , \tag{68}$$

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 E.1. *There exists an FRL instance such that any stochastic stationary policy is suboptimal with respect to some history-dependent policy.*

Proof. We consider the FRL task described in Figure 5 with $\rho = (1, 0, 0, 0)$. We show here that it holds

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

Define the following 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

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

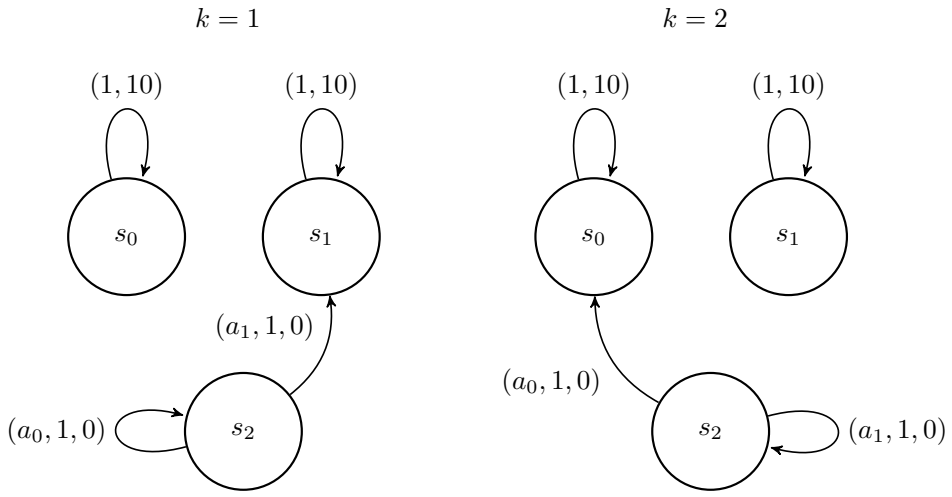


Figure 6: FRL task with no optimal deterministic policy. The triplet means (action, probability, reward) and $\gamma = 0.9$. If the action is not specified, it means that all the actions give the same reward and lead to the same state

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

$$V_1^{\pi^\star_{\mathrm{sta}}}(s_0) + V_2^{\pi^\star_{\mathrm{sta}}}(s_0) = \sum_{k=2}^{\infty} \gamma^k (1 \cdot p + 2 \cdot (1-p)) + V_2^{\pi^\star_{\mathrm{sta}}}(s_0) \ .$$

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

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

By gathering the two precedent expressions, we get

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

$$\leq \frac{1}{1-\gamma}\left[2\gamma^2 + (1-\gamma) - (1-2\gamma)\frac{1-\gamma}{1-\gamma+p\gamma}\right] \leq \frac{2\gamma^2}{1-\gamma} + 2\gamma \ ,$$

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 $\pi^\star_{\text{sta}}$ with respect to the history-dependent policy π_ℓ. □

Lemma E.2. *There exists an FRL instance such that any deterministic policy is suboptimal with respect to some stationary stochastic policy.*

Proof. We consider the FRL task of Figure 6, 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

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

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

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

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

Case $\pi^\star_{\text{det}}(s_2) = a_0$: 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

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

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

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

Combining the two previous lemmas concludes the proof of Theorem 6.1.

E.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 E.3. *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,*

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

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.

Proof. Consider an FRL instance where each agent's MDP is defined as $\mathcal{M}_c \triangleq (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}, \mathsf{r}_c, \rho)$. 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 $\bar{\mathsf{r}} \triangleq \frac{1}{M}\sum_{c=1}^{M}\mathsf{r}_c$ 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. □

F Technical lemmas

F.1 Basic Lemmas

For completeness, we state without proof basic results that are routinely used in our proofs.

Lemma F.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 F.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 $i = 1, \dots, d$, 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 $\theta \in U$,*

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

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,

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

F.2 Performance difference lemma

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

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

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 F.5. *Let us consider two MDPs* $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_1, \mathsf{r}_1)$ *and* $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}_2, \mathsf{r}_2)$ *such that* $\sup_{s,a \in \mathcal{S} \times \mathcal{A}} \|\mathsf{P}_1(\cdot|s,a) - \mathsf{P}_2(\cdot|s,a)\|_1 \leq \varepsilon_{\mathsf{P}}$ *and* $\|\mathsf{r}_1 - \mathsf{r}_2\|_\infty \leq \varepsilon_{\mathsf{r}}$*. 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* $\|V_1^\pi\|_\infty \leq c$ *and* $\|V_2^\pi\|_\infty \leq c$ *then it holds that for all* $s \in \mathcal{S}$

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

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

Lemma F.6. *For all* $c, c' \in [M]$*, it holds that*

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

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

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

which concludes the proof. □

Lemma F.7. *Consider any two policies* π_i*,* $i = 1, 2$*, and any agent* $c \in [M]$*. It holds that*

$$\|d_c^{\rho,\pi_1} - d_c^{\rho,\pi_2}\|_1 \leq \frac{\gamma}{1-\gamma} \sup_{s \in \mathcal{S}} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 \ .$$

Proof. Let us start from the definition of flow conservation constraints for the discounted state occupancy (Puterman, 1994), for $i \in \{1, 2\}$, we have

$$d_c^{\rho,\pi_i}(s) = (1-\gamma)\rho(s) + \gamma \sum_{s'} \mathsf{P}_{c,\pi_i}(s|s') d_c^{\rho,\pi_i}(s') \ .$$

Then, we have

$$\begin{aligned}
\sum_{s\in\mathcal{S}} |d_c^{\rho,\pi_2}(s) - d_c^{\rho,\pi_1}(s)| &\leq \gamma \sum_{(s',s')} \sum_{s} |\mathsf{P}_c(s|s',s')\pi_2(s'|s') d_c^{\rho,\pi_2}(s') - \mathsf{P}_c(s|s',s')\pi_1(s'|s') d_c^{\rho,\pi_1}(s')| \\
&\leq \gamma \sum_{s',s'} \sum_{s} \mathsf{P}_c(s|s',s') \left|\pi_2(a'|s') - \pi_1(a'|s')\right| d_c^{\rho,\pi_2}(s') \\
&+ \gamma \sum_{s',a'} \sum_{s} \mathsf{P}_c(s|s',a')\pi_1(a'|s') \left|d_c^{\rho,\pi_1}(s') - d_c^{\rho,\pi_2}(s')\right| \\
&\leq \gamma \sup_{s\in\mathcal{S}} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 + \gamma \sum_{s'} |d_c^{\rho,\pi_1}(s') - d_c^{\rho,\pi_2}(s')| \ ,
\end{aligned}$$

which concludes the proof. □

F.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 F.8. *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))

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

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

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

In particular, we have

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

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

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 F.9. *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 F.8 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 F.8, 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

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

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 F.8

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

□

Lemma F.10. *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

$$\|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty \le \|\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u]\|_\infty + \gamma\|\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\cdot\pi_\theta\|_\infty + \lambda\|\mathrm{d}\mathcal{H}(\pi_\theta)[u]\|_\infty\,. \tag{69}$$

For the first term, we have for any $s\in\mathcal{S}$ by a simple bound on Q-value and Lemma F.8

$$|\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}\pi_\theta[u]|(s) \le \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sum_{a\in\mathcal{A}}|\mathrm{d}\pi_\theta[u](a|s)| \le \frac{8(1+\lambda\log A)}{1-\gamma}\|u\|_\infty\,.$$

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 F.9 we have

$$\|\mathrm{d}\mathcal{H}(\pi_\theta)[u]\|_\infty \le 2\log|\mathcal{A}|\cdot\|u\|_\infty.$$

Thus, from (69) 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 F.8 and Lemma F.9:

$$
\begin{aligned}
|\widetilde{Q}_c^{\pi_\theta} \cdot \mathrm{d}^2\pi_\theta[u,v]|(s) &\le \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sum_{a\in\mathcal{A}}|\mathrm{d}^2\pi_\theta(a|s)[u,v]| \le \frac{8(1+\lambda\log|\mathcal{A}|))}{1-\gamma}\|u\|_\infty\|v\|_\infty\,,\\
|\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\mathrm{d}\pi_\theta[v]|(s) &\le \|\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\|_\infty\sum_{a\in\mathcal{A}}|\mathrm{d}\pi_\theta(a|s)[v]| \le \frac{16+20\lambda\log|\mathcal{A}|}{(1-\gamma)^2}\|u\|_\infty\|v\|_\infty\,,\\
|\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\cdot\pi_\theta|(s) &\le \|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty\,,\\
|\mathrm{d}^2\mathcal{H}(\pi_\theta)[u,v]|(s) &\le (4+8\log|\mathcal{A}|)\|u\|_\infty\|v\|_\infty\,,
\end{aligned}
$$

thus

$$
\begin{aligned}
\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty &\le \gamma\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty\\
&\quad+\left(\frac{32+40\lambda\log|\mathcal{A}|}{(1-\gamma)^2}+\frac{8(1+\lambda\log|\mathcal{A}|)}{1-\gamma}+\lambda(4+8\log|\mathcal{A}|)\right)\|u\|_\infty\|v\|_\infty\,.
\end{aligned}
$$

Since $|\mathcal{A}|\ge 2$, then $2\log|\mathcal{A}|\ge 1$, we can simplify it as follows

$$
\|\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\|_\infty \le \frac{40+64\lambda\log|\mathcal{A}|}{(1-\gamma)^3}\|u\|_\infty\|v\|_\infty\,.
$$

Third differential. Next, we proceed with the third differential as follows

$$
\begin{aligned}
\mathrm{d}^3\widetilde{V}_c^{\pi_\theta}[u,v,w] &= \widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]+\gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[w]\cdot\mathrm{d}^2\pi_\theta[u,v]\\
&\quad+\gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,w]\mathrm{d}\pi_\theta[v]+\gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[u]\mathrm{d}^2\pi_\theta[v,w]\\
&\quad+\gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[v,w]\mathrm{d}\pi_\theta[u]+\gamma\mathsf{P}_c\mathrm{d}\widetilde{V}_c^{\pi_\theta}[v]\mathrm{d}^2\pi_\theta[u,w]\\
&\quad+\gamma\mathsf{P}_c\mathrm{d}^2\widetilde{V}_c^{\pi_\theta}[u,v]\cdot\mathrm{d}\pi_\theta[w]+\gamma\mathsf{P}_c\mathrm{d}^3\widetilde{V}_c^{\pi_\theta}[u,v,w]\cdot\pi_\theta+\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\,.
\end{aligned}
$$

By the triangle inequality

$$
\begin{aligned}
\|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty &\le \|Q_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]\|_\infty+\gamma\|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[w]\cdot\mathrm{d}^2\pi_\theta[u,v]\|_\infty\\
&\quad+\gamma\|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[u,w]\mathrm{d}\pi_\theta[v]\|_\infty+\gamma\|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[u]\mathrm{d}^2\pi_\theta[v,w]\|_\infty\\
&\quad+\gamma\|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[v,w]\mathrm{d}\pi_\theta[u]\|_\infty+\gamma\|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[v]\mathrm{d}^2\pi_\theta[u,w]\|_\infty\\
&\quad+\gamma\|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[u,v]\cdot\mathrm{d}\pi_\theta[w]\|_\infty+\gamma\|\mathsf{P}_c\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\cdot\pi_\theta\|_\infty\\
&\quad+\|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\|_\infty\,.
\end{aligned}
$$

To simplify notation, let us define $R_{1,2}(u,v,w) \triangleq \|\mathsf{P}_c\mathrm{d}V_c^{\pi_\theta}[u,v]\mathrm{d}^2\pi_\theta[w]\|_\infty$ and $R_{2,1}(u,v,w) \triangleq \|\mathsf{P}_c\mathrm{d}^2V_c^{\pi_\theta}[u,v]\mathrm{d}\pi_\theta[w]\|_\infty$. Next, we notice that

$$
\|\mathsf{P}_c\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\cdot\pi_\theta\|_\infty = \max_s\left|\sum_{s'}\pi_\theta(a|s)\mathsf{P}(s'|s,a)\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]_{s'}\right| \le \|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty\,,
$$

thus, we have a contraction argument that implies

$$
\begin{aligned}
\|\mathrm{d}^3V_c^{\pi_\theta}[u,v,w]\|_\infty \le \frac{1}{1-\gamma}\Bigg(&\|Q_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]\|_\infty+\|\mathrm{d}^3\mathcal{H}(\pi_\theta)[u,v,w]\|_\infty\\
&+\gamma(R_{1,2}(w,u,v)+R_{1,2}(u,v,w)+R_{1,2}(v,u,w))\\
&+\gamma(R_{2,1}(u,w,v)+R_{2,1}(v,w,u)+R_{2,1}(u,v,w))\Bigg)\,.
\end{aligned}
\tag{70}
$$

Next, we bound all terms that appear in the bound above. First, we apply Lemma F.8 for a fixed state $s\in\mathcal{S}$

$$
|\widetilde{Q}_c^{\pi_\theta}\cdot\mathrm{d}^3\pi_\theta[u,v,w]|(s) \le \sum_{a\in\mathcal{A}}|\widetilde{Q}_c^{\pi_\theta}(s,a)\cdot\mathrm{d}^3\pi_\theta[u,v,w](a|s)|
$$

$$\leq \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sum_{a\in\mathcal{A}}|\mathrm{d}^3\pi_\theta[u,v,w](a|s)|$$
$$\leq \frac{48(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

Also, by Lemma F.9 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 F.8

$$R_{1,2}(u,v,w) \leq \frac{8\cdot(8+10\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}\|u\|_\infty\|v\|_\infty\|w\|_\infty\,.$$

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 (70) 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

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

□