Exploration via Exploitation: The Blessing of Reward Diversity in Personalized Federated RL

Jiin Woo*
Carnegie Mellon University

Gauri Joshi
Carnegie Mellon University

Yuejie Chi
Yale University

Abstract

While federated reinforcement learning (RL) has traditionally guaranteed high efficiency in homogeneous environments, collaboration is often hindered when agents possess distinct goals. In this paper, we flip this paradigm, framing reward heterogeneity not as a hurdle to consensus, but as a structural asset that catalyzes exploration. We first propose Personalized Federated Upper Confidence Bound Value Iteration (PF-UCBVI), which achieves optimal linear speedup by decoupling shared dynamics from personalized objectives. To eliminate the risks of explicit exploration, we then introduce Personalized Federated Exploration-Free Value Iteration (PF-EFVI), a purely greedy algorithm that leverages reward diversity to ensure state-action coverage. We prove that PF-EFVI attains logarithmic regret without explicit exploration bonuses under sufficient reward diversity. Our results reveal that agent disagreement shifts the exploration burden from temporal complexity to spatial diversity, enabling efficient collective learning without explicit exploration.

1 Introduction

Reinforcement Learning (RL) has achieved remarkable success across various domains, yet its sample complexity remains a fundamental bottleneck in real-world applications (Chen et al., 2026). Federated Reinforcement Learning (FedRL) has emerged as a promising paradigm to alleviate this issue, enabling multiple agents to collaboratively learn policies by sharing experiences without centralizing raw trajectories (Khodadadian et al., 2022; Woo et al., 2023; Zheng et al., 2024).

However, most FedRL literature assumes homogeneous environments where all agents share identical dynamics and rewards (Khodadadian et al., 2022; Woo et al., 2025), while in practice agents often pursue distinct, personalized goals. For instance, in a fleet of autonomous vehicles navigating the same city, each vehicle has a different destination, and recommendation systems on a common platform serve users with vastly different interests. Existing approaches for heterogeneous FedRL (Labbi et al., 2025b; Xiong et al., 2025) largely treat reward diversity as an obstacle to be tolerated, averaged out, or mitigated through restrictive structural assumptions, and consequently struggle to deliver personalized learning under arbitrary heterogeneity with provable guarantees. This paper takes the opposite view: the diversity of rewards is not a hurdle, but a structural resource that can both accelerate collaborative learning and reshape the exploration-exploitation trade-off itself.

To put this view on rigorous footing, we consider a federated setting where each agent interacts with an episodic Markov Decision Process (MDP) with shared dynamics but diverse rewards. We first propose Personalized Federated Upper Confidence Bound Value Iteration (PF-UCBVI), which decouples the shared dynamics from the personalized rewards: the central server globally aggregates empirical transitions to estimate dynamics, while each agent maintains strictly local UCB bonuses

*Work done while at Carnegie Mellon University. Correspondence to Jiin Woo (jiinw@andrew.cmu.edu).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and Bellman backups tailored to its own reward. We prove that PF-UCBVI achieves an optimal linear speedup in average per-agent regret, fully agnostic to the level of reward heterogeneity. However, in domains such as robotics or clinical trials, exploratory actions can be risky, unethical, or expensive. This concern motivates our central insight: in a multi-agent system, the heterogeneity of goals is itself a structural catalyst for exploration. If a population of agents possesses sufficiently diverse rewards, their collective greedy executions naturally span the state-action space, lifting the exploration burden from any individual agent and placing it on the diversity of the population as a whole. Building on this insight, we propose Personalized Federated Exploration-Free Value Iteration (PF-EFVI), a purely greedy algorithm with neither UCB bonuses nor optimistic initialization.

It is worth noting that PF-EFVI is not a free lunch but a conditional blessing: when rewards lack diversity, the ensemble's greedy executions can collapse. The two algorithms are therefore complementary: PF-UCBVI serves as a worst-case safety net paying an explicit-exploration burn-in regardless of reward structure, while PF-EFVI delivers the blessing of reward diversity precisely when the collective agent ensemble is rich enough. Our experiments confirm a clean phase transition consistent with this picture. We now summarize our main contributions below. Throughout, M denotes the number of agents, S the size of the state space, A the size of the action space, H the horizon, K the number of episodes, λ_{min} the reward diversity parameter, and Δ_{min}^m the minimum suboptimality gap of agent m, $1 \le m \le M$.

- **Personalized federated UCB achieving linear speedup:** We propose PF-UCBVI, establishing that sharing transition dynamics while maintaining personalized UCB bonuses yields a near-optimal average regret $\widetilde{\mathcal{O}}(\sqrt{SAKH^3/M})$ per agent. This proves that linear speedup can be achieved without any heterogeneity penalty in the bound, regardless of the severity of reward heterogeneity.
- **Diversity-induced global coverage:** We introduce the *Reward Diversity* assumption (Assumption 1), which posits that whenever agents act greedily under any estimated transition dynamics, the ensemble jointly visits every state-action pair with probability at least $\lambda_{\mathsf{min}} > 0$. Unlike standard coverage assumptions that rely on environment stochasticity or active exploration, ours is driven entirely by the diversity of agents' rewards. We identify two complementary structural regimes under which the assumption is provably feasible and support it empirically in random MDPs with diverse rewards across agents.
- **Exploration-free federated RL with reward diversity:** We propose PF-EFVI, the first exploration-free algorithm for federated RL. We prove that under sufficient reward diversity, purely greedy agents achieve an instance-dependent regret bound of $\widetilde{\mathcal{O}}\left(\frac{H^5 \log K}{M\lambda_{\mathsf{min}}} \sum_{m=1}^{M} \frac{1}{\Delta_{\mathsf{min}}^m}\right)$, where Δ_{min}^m is agent m's minimum suboptimality gap. The $\log K$ dependency is scaled by $1/(M\lambda_{\mathsf{min}})$, which shrinks as the ensemble grows and diversifies, signifying that the temporal exploration penalty is effectively absorbed by the diversity of the ensemble, fundamentally reshaping the exploration-exploitation trade-off.

2 Related works

We briefly position our work along three thrusts; an extended discussion with full citations and a summary table (Table 1) is deferred to Appendix A.

Federated RL with homogeneous environments. A substantial body of work establishes linear speedup for federated RL when all agents share the same MDP, both for sample complexity under synchronous or Markovian sampling (Salgia and Chi, 2024; Khodadadian et al., 2022; Woo et al., 2025) and for online regret in episodic tabular MDPs (Zheng et al., 2024, 2025; Zhang et al., 2025). These results form the natural baseline, but they do not address agents with personalized rewards.

Federated RL with heterogeneous environments. Heterogeneous federated RL has been pursued along two complementary directions:

1. **Consensus-based methods** (Jin et al., 2022; Wang et al., 2024a; Zhang et al., 2024; Labbi et al., 2025b; Wang et al., 2024b; Mangold et al., 2025; Zhu et al., 2025; Yang et al., 2024; Labbi et al., 2025a; Chen et al., 2023) learn a single global policy that maximizes the average value across environments. Although elegant, all such methods inherently sacrifice individual utility

for global agreement: the average value can be far from any agent's optimum, and the benefits of collaboration shrink, and may even turn into negative transfer, as heterogeneity grows.

2. **Personalized methods** (Zhang and Azizan, 2026; Xiong et al., 2025; Wang et al., 2026) instead allow each agent to learn its own policy, achieving linear speedup under similarity, low-rank, or shared-representation assumptions. However, these analyses either rely on cross-agent structural assumptions, or assume a fixed behavior policy that already ensures sufficient coverage, and most focus on policy evaluation rather than online regret minimization.

Our work removes both restrictions: PF-UCBVI is fully agnostic to the level of reward heterogeneity, and PF-EFVI sources coverage from reward diversity rather than from a prescribed behavior policy.

Exploration-free learning and covariate diversity. In safety-critical or resource-constrained domains, exploratory actions are often risky, unethical, or expensive, motivating purely greedy approaches. Foundational results by Bastani et al. (2021) and Kannan et al. (2018) show that, under sufficient covariate diversity, greedy selection achieves sublinear regret in linear contextual bandits without any explicit exploration. Civitavecchia and Papini (2025) extend this perspective to single-agent RL, but their guarantee fundamentally relies on strong environment stochasticity to ensure state-action coverage, which is a stringent requirement for any individual agent. To our knowledge, no prior work simultaneously addresses (i) the federated heterogeneous setting, (ii) online regret minimization, and (iii) exploration-free learning; PF-EFVI fills this gap by sourcing coverage from cross-agent reward diversity rather than from environment stochasticity.

3 Problem formulation

We consider M agents operating in M finite-horizon MDPs $\{\mathcal{M}^m\}_{m=1}^M$ with shared dynamics and diverse rewards. Each agent m is characterized by a local MDP

$$\mathcal{M}^m = \big(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h^m\}_{h=1}^H\big).$$

Here, $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, and H is the episode horizon. The transition kernel $P_h(\cdot \mid s,a) \in \Delta(\mathcal{S})$ specifies the distribution of the next state at step h after taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. The agents share the same transition kernels $\{P_h\}_{h=1}^H$, but each agent has its own personalized reward function $r_h^m : \mathcal{S} \times \mathcal{A} \to [0,1]$. In this work, we assume the reward functions are known and fixed at each agent. A policy $\pi = \{\pi_h\}_{h=1}^H$ is a sequence of mappings $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ that specifies the action selection rule at step h for each state s. The value function of policy π for agent m at step h is defined as

$$V_h^{m,\pi}(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}^m(s_{h'}, a_{h'}) \mid s_h = s\right], \tag{1}$$

where the expectation is taken over the randomness in the trajectory generated by following policy π starting from state s at step h. The optimal value function for agent m is defined as $V_h^{m,\star}(s) = \sup_\pi V_h^{m,\pi}(s)$, and the optimal policy $\pi^{m,\star}$ is any policy that achieves this optimal value function at every (h,s).

Goal: regret minimization in online RL. In online RL, for K episodes, agents interact with their local environments using behavior policies updated based on the data collected previously. At episode k, each agent m selects a policy π_k^m based on the data collected from previous episodes and executes π_k^m independently of the other agents to generate a trajectory $\{s_{k,h}^m, a_{k,h}^m, r_{k,h}^m\}_{h=1}^H$, where the initial state $s_{k,1}^m$ is drawn from ρ.

Unlike prior works that sacrifice individual utility for global consensus (Labbi et al., 2025b; Zhang et al., 2024), in our setting, the goal of each agent is to learn their personalized optimal policy $\pi^{m,\star}$. The overall system aims to minimize the total regret across all agents over K episodes:

$$\mathsf{Regret}(K) = \sum_{m=1}^{M}\sum_{k=1}^{K} \mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right], \tag{2}$$

where π_k^m is the policy of agent m at episode k.

4 Personalized federated online RL with UCB exploration

To solve the personalized regret minimization problem efficiently, we must leverage the shared structure of the environments without forcing a single global objective. Since the transition kernel P_h is common across all agents, sharing state-action visitation data allows the ensemble to estimate the dynamics significantly faster than any single agent could alone.

However, because each agent pursues a distinct reward function r_h^m, their respective value functions, and consequently, their uncertainty about future returns, differ significantly. A standard federated approach with a single global exploration bonus would be heavily suboptimal here; it might force an agent to explore regions of the state space that are entirely irrelevant to its personalized goals. To bridge this gap, we propose Personalized Federated Upper Confidence Bound Value Iteration (PF-UCBVI).

4.1 Algorithm description

The core principle of PF-UCBVI is to globally aggregate empirical transition counts to build an accurate shared dynamics model, while strictly maintaining *personalized* UCB bonuses and local Bellman backups tailored to each agent's individual objectives. The algorithm proceeds as follows.

Initialization: the server initializes all global counts $N_h(s,a)$ and $N_h(s,a,s')$ to zero, and each agent $m \in [M]$ initializes $V_h^m(s) = Q_h^m(s,a) = H - h + 1$ for all (s,a,h) with $V_{H+1}^m = 0$.

Federated learning loop: at each episode $k = 1, \dots, K$, the algorithm performs three steps:

1. *Optimistic execution.* Agent m follows the greedy policy
$$\pi_{k,h}^m(s) = \arg\max_a Q_h^m(s,a)$$
to collect transitions, with ties broken uniformly at random once at the start of each episode.
2. *Global aggregation.* The server updates the global counts via $N_h(s,a) \leftarrow N_h(s,a) + \sum_m \mathbb{I}\{(s_{k,h}^m, a_{k,h}^m) = (s,a)\}$ and $N_h(s,a,s')$ analogously, then computes and broadcasts the empirical transition kernel $\widehat{P}_h(s' \mid s,a) = N_h(s,a,s') / \max\{1, N_h(s,a)\}$.
3. *Local Bellman backup.* For $h = H, \dots, 1$, each agent m first computes its personalized bonus
$$B_h^m(s,a) = \sqrt{\frac{4\mathsf{Var}_{\widehat{P}_{h,s,a}}(V_{h+1}^m)\iota}{\max\{N_h(s,a),1\}}} + \frac{25SH^2\iota}{\max\{N_h(s,a),1\}}, \quad \iota = \log(28M^2SAKH/\delta), \tag{3}$$
and then performs the value iteration update for $h = H, \dots, 1$:
$$Q_h^m(s,a) = \min\Big\{H - h + 1,\ r_h^m(s,a) + B_h^m(s,a) + \sum_{s' \in \mathcal{S}} \widehat{P}_h(s' \mid s,a) V_{h+1}^m(s')\Big\},$$
$$V_h^m(s) = \max_{a \in \mathcal{A}} Q_h^m(s,a).$$

To our knowledge, PF-UCBVI is the first algorithm that combines online UCB exploration with personalization in federated RL. Prior personalized federated RL works focus on policy evaluation (Xiong et al., 2025; Wang et al., 2026) or assume a fixed behavior policy with sufficient coverage (Zhang and Azizan, 2026), sidestepping the exploration–exploitation tradeoff entirely. The federated UCBVI algorithm (Labbi et al., 2025b), on the other hand, performs online exploration but only learns a consensus policy. PF-UCBVI sits at the intersection: a personalized extension of UCBVI (Azar et al., 2017) for online settings where each agent performs a personalized exploration and learns its own optimal policy under arbitrary reward heterogeneity.

4.2 Regret analysis of PF-UCBVI

We now present the main theoretical result for PF-UCBVI, which establishes a regret bound that achieves optimal linear speedup with respect to the number of agents M while allowing for arbitrary reward heterogeneity.

Theorem 1 (Regret of personalized federated UCBVI)**.** *Consider the setting of Section 3 with M agents, S states, A actions, and a planning horizon H. Let the agents perform the* PF-UCBVI *algorithm with a confidence parameter $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the total regret satisfies:*

$$\mathsf{Regret}(K) \leq C \cdot \left(\sqrt{SAMKH^3\iota^2} + S^2AH^3\iota^2 \; + \; MSAH^2\right), \tag{4}$$

where $C > 0$ is a universal constant and $\iota = \log(28M^2SAKH/\delta)$.

The proof of Theorem 1 is provided in Appendix C. The key implications of this result are as follows.

Optimal linear speedup and comparison with baselines. The total regret $\widetilde{\mathcal{O}}(\sqrt{SAMKH^3})$ in Theorem 1 translates into an average per-agent regret of $\widetilde{\mathcal{O}}(\sqrt{SAKH^3/M})$, so the number of episodes needed to reach a target average regret shrinks by a factor of M, the optimal linear speedup over isolated learning. Beyond the rate itself, PF-UCBVI differs from the closest baseline Fed-UCBVI (Labbi et al., 2025b) in both formulation and the role of heterogeneity. Fed-UCBVI targets a fundamentally different objective, learning a single *consensus* policy whose suboptimality with respect to each agent's personalized optimum grows with the level of reward heterogeneity. In contrast, our regret bound is fully agnostic to the level of reward heterogeneity: it holds no matter how different the agents' rewards are, and in the homogeneous setting its leading rate matches that of Labbi et al. (2025b) while each agent learns its own optimal policy. This shows that PF-UCBVI exploits the shared transition structure to accelerate personalized learning, without paying any heterogeneity penalty in the bound.

5 Personalized federated online RL without exploration

While PF-UCBVI navigates reward heterogeneity through explicit UCB-based exploration, such exploratory actions remain a significant burden in safety-critical or resource-constrained domains. In fact, the same mechanism that enables heterogeneity-agnostic collaboration suggests a deeper opportunity: if agents pursuing different objectives naturally visit different regions of the environment, their collective self-interest can eliminate the need for explicit exploration altogether. This shifts our perspective on heterogeneity from a condition to be tolerated to a potent resource that provides exploration "for free" across agents, motivating the exploration-free paradigm we develop next.

5.1 Key assumptions

To enable exploration-free learning in this heterogeneous setting, we introduce a key structural assumption. While single-agent RL typically requires strong environment stochasticity (e.g., covariate diversity) (Civitavecchia and Papini, 2025), we show that reward heterogeneity can itself induce sufficient coverage. Let $d_h^\pi(s,a)$ denote the marginal visitation probability of observing state-action pair (s,a) at step h under policy π.

Assumption 1 (Reward Diversity)**.** *Let $\mathcal{P}$ be the set of all valid transition kernels. For any assumed transition kernel $\widetilde{P} \in \mathcal{P}$ and each agent $m \in [M]$, let $\widetilde{Q}_h^m$ denote the action-value function computed using its true private reward r^m and the transition $\widetilde{P}$. Let $\widetilde{\pi}^m$ be the corresponding greedy policy induced by $\widetilde{Q}_h^m$ with uniform tie-breaking. We assume there exists a strictly positive constant $\lambda_{\mathsf{min}} > 0$ such that, for any assumed transition kernel $\widetilde{P} \in \mathcal{P}$ and each step $h \in [H]$:*

$$\min_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} \frac{1}{M}\sum_{m=1}^{M} d_h^{\widetilde{\pi}^m}(s,a) \geq \lambda_{\mathsf{min}}, \tag{5}$$

where $\mathcal{S}_h$ is the set of states reachable at step h, and $d_h^{\widetilde{\pi}^m}(s,a)$ is the marginal visitation probability of state-action pair (s,a) at step h when executing policy $\widetilde{\pi}^m$ in the true environment P.

This structural assumption posits that the diversity of true goals is potent enough to prevent the ensemble's policies from collapsing into a narrow subspace, regardless of early estimation errors in the dynamics. Since the sum of visitation probabilities across all state-action pairs is bounded, the most ideal scenario of perfectly uniform coverage yields $\lambda_{\mathsf{min}} \leq 1/(|\mathcal{S}_h|A)$. Consequently, $1/\lambda_{\mathsf{min}}$ absorbs the state-action space complexity.

Reward-diversity-induced coverage relaxation. Assumption 1 significantly relaxes the stringent requirements of single-agent RL by relying on the macroscopic diversity of the agents' goals rather than environment stochasticity. Intuitively, if $\{r^m\}_{m=1}^M$ are sufficiently diverse, their collective greedy executions will naturally span the state-action space, achieving exploration *for free* collectively. The parameter λ_{min} captures the degree of this system-wide heterogeneity; a larger λ_{min} indicates broader collective coverage, facilitating faster convergence, whereas a smaller λ_{min} implies more concentrated trajectories over limited state-action space.

5.2 The feasibility of the reward diversity-induced coverage

At the heart of Assumption 1 lies a simple mechanism: when agents' value functions point in *sufficiently different directions* across the state space, any perturbation in the dynamics estimate $\widetilde{P}$ pushes their greedy policies apart rather than together. To see this concretely, consider any assumed transition kernel $\widetilde{P} \in \mathcal{P}$ and the Q-value gap between two actions $a, a' \in \mathcal{A}$:

$$\widetilde{Q}_h^m(s,a) - \widetilde{Q}_h^m(s,a') = \underbrace{\left[r_h^m(s,a) - r_h^m(s,a')\right]}_{\text{action preference}} + \underbrace{\left[\widetilde{P}_h(\cdot \mid s,a) - \widetilde{P}_h(\cdot \mid s,a')\right]^\top \widetilde{V}_{h+1}^m}_{\text{dynamics-sensitive term}}. \quad (6)$$

The first term reflects each agent's intrinsic preference between actions and is invariant to dynamics estimation errors. The second term depends on $\widetilde{P}$ through the agent's value function $\widetilde{V}_{h+1}^m$ and is the source of $\widetilde{P}$-dependent policy variation. Crucially, this second term is shaped by the geometry of $\widetilde{V}_{h+1}^m$ and reward heterogeneity controls this geometry *differently across agents*. When agents' value functions are diverse, the same dynamics perturbation steers different agents toward different actions, automatically maintaining collective coverage regardless of estimation errors.

We illustrate two complementary regimes where Assumption 1 is naturally satisfied, both ensuring that agents collectively cover all state-action pairs despite perturbations in $\widetilde{P}$. The two regimes are distinguished by which term dominates the Q-value gap in (6), and accordingly by how collective coverage is preserved. In one regime, each agent's reward gap dominates the dynamics-sensitive term, so individual choices stay strong under $\widetilde{P}$-perturbations and agents with different reward preferences keep selecting different actions. In the other regime, the dynamics-sensitive term can be large and individual choices do shift with $\widetilde{P}$, but cross-agent anti-correlation forces agents to shift in opposite directions so that different actions are always covered between them. In both cases, no matter how $\widetilde{P}$ moves, the ensemble continues to spread its visits across the state-action space.

Regime I: strong individual choices. Suppose each agent's reward gap $r_h^m(s,a) - r_h^m(s,a')$ is large relative to the spread of $\widetilde{V}_{h+1}^m$ across states. Even when $\widetilde{P}$ is perturbed and the dynamics-sensitive term shifts, the reward gap continues to dictate the sign of each agent's Q-gap (the left-hand side of (6), which determines the greedy choice between a and a'), so each agent's greedy action does not flip. If different agents have reward gaps pointing toward different actions, they persistently select different actions across all $\widetilde{P}$ and at every step h, and the ensemble jointly covers all reachable state-action pairs provided that each agent reaches every reachable state with positive probability. The defining feature is that no single agent's choice is sensitive to dynamics estimation errors. Figure 1 (top) illustrates this with action-dependent state-independent rewards on $\{x, Y_1, Y_2\}$: $\widetilde{V}_{h+1}^m$ takes the same value at every state for any $\widetilde{P}$, the dynamics-sensitive term vanishes identically, and each agent's Q-gap reduces to its reward gap.

Regime II: coordinated opposition. Now suppose individual reward gaps are small while $\widetilde{V}_{h+1}^m$ varies substantially across states. Each agent's greedy action *does* respond to $\widetilde{P}$, since the dynamics-sensitive term now controls the sign of the Q-gap. Coverage instead arises from cross-agent structure: when agents' value functions are *anti-correlated*, in the sense that $\sum_m \widetilde{V}_{h+1}^m$ is constant across states, the dynamics-sensitive terms of different agents are forced to sum to zero across agents so their Q-gaps become mirror images. The two agents may both flip their preferred action as $\widetilde{P}$ changes, but they flip in lockstep opposition and always select different actions, again covering all reachable state-action pairs jointly. Figure 1 (bottom) illustrates this with state-dependent action-independent rewards on $\{x, Y_1, Y_2\}$, where agent A values Y_1 and agent B values Y_2: their Q-gaps are exact mirror images for any $\widetilde{P}_1$, and they swap actions in perfect opposition as $\widetilde{P}_1$ varies.

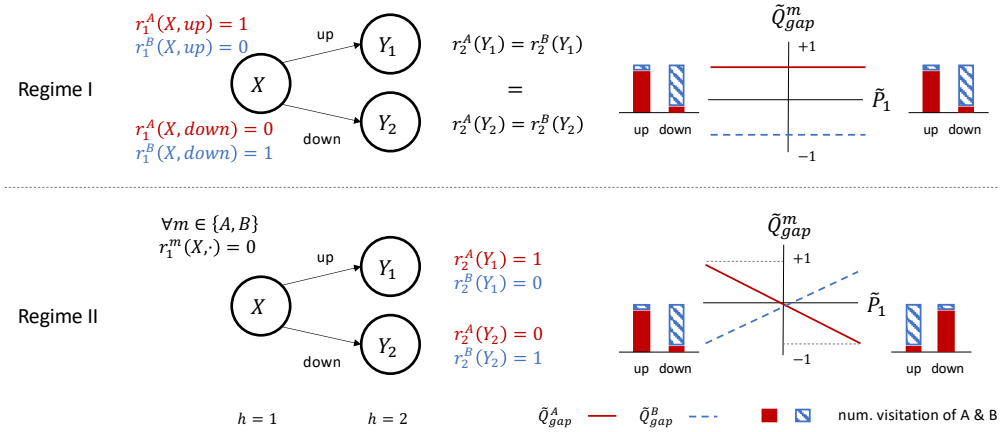

Figure 1: Two regimes of reward heterogeneity that induce global coverage in Assumption 1, illustrated on a common two-step MDP with $\mathcal{S} = \{x, Y_1, Y_2\}$, $\mathcal{A} = \{\text{up}, \text{down}\}$, $M = 2$ (agent A and agent B). In each row, the right panel plots each agent's step-1 *Q-gap* $\widetilde{Q}^m_{gap} = \widetilde{Q}^m_1(x, \text{up}) - \widetilde{Q}^m_1(x, \text{down})$, whose sign determines the agent's greedy choice at x, as the assumed kernel $\widetilde{P}_1$ varies. **Top (Regime I):** action-dependent state-independent rewards make $\widetilde{V}^m_2$ take the same value at every state, so each agent's Q-gap is fixed regardless of $\widetilde{P}_1$ (horizontal lines). Both agents persistently select different actions, and the cross-agent average visitation remains $1/2$ on each action at both extremes of $\widetilde{P}_1$. **Bottom (Regime II):** state-dependent action-independent rewards anti-correlated across agents make $\widetilde{V}^A_2 + \widetilde{V}^B_2$ take the same value at every state, so the two agents' Q-gaps are exact mirror images for any $\widetilde{P}_1$ (crossing lines). Individual visitations swap between the two extremes (blue and red bars exchange positions), but the cross-agent average remains $1/2$ on each action throughout.

The two regimes show that the seemingly strong requirement of Assumption 1 is nonetheless attainable through plausible structures of reward heterogeneity. In Section 6, we find empirical support for the assumption in general random MDPs with diverse rewards.

5.3 Algorithm description

Based on the reward diversity assumption, we propose Personalized Federated Exploration-Free Value Iteration (PF-EFVI). Unlike standard federated RL algorithms that require meticulous tuning of exploration bonuses, PF-EFVI relies solely on the empirical average of the global transitions and pure greedy exploitations at the local level. The algorithm proceeds as follows.

Initialization: the server initializes all global counts $N_h(s, a)$ and $N_h(s, a, s')$ to zero, and each agent m initializes $V^m_h(s) = Q^m_h(s, a) = 0$ for all (s, a, h) with $V^m_{H+1}(s) = 0$. Ties are broken uniformly at random.

Federated learning loop: at each episode $k = 1, \ldots, K$, the algorithm performs three steps:

1. *Local execution (pure exploitation).* Each agent $m \in [M]$ executes its current greedy policy $\pi^m_{k,h}(s) = \arg\max_a Q^m_h(s, a)$ to collect a local trajectory $\{s^m_{k,h}, a^m_{k,h}, r^m_h(s^m_{k,h}, a^m_{k,h})\}_{h=1}^H$.
2. *Global aggregation.* The server updates the global counts via $N_h(s, a) \leftarrow N_h(s, a) + \sum_m \mathbb{I}\{(s^m_{k,h}, a^m_{k,h}) = (s, a)\}$ and $N_h(s, a, s')$ analogously, then computes and broadcasts the empirical transition kernel $\widehat{P}_h(s' \mid s, a) = N_h(s, a, s')/N_h(s, a)$ if $N_h(s, a) \geq 1$, and $\widehat{P}_h(s' \mid s, a) = 1/S$ otherwise.
3. *Local Bellman backup.* Each agent m performs backward value iteration: for $h = H, \ldots, 1$,

$$Q^m_h(s, a) = r^m_h(s, a) + \sum_{s'} \widehat{P}_h(s' \mid s, a) V^m_{h+1}(s'),$$
$$V^m_h(s) = \max_a Q^m_h(s, a).$$

5.4 Regret analysis of PF-EFVI

In this section, we present the theoretical guarantees of the PF-EFVI algorithm. We show that under the reward diversity condition, a purely greedy approach without any explicit exploration or optimistic initialization achieves logarithmic cumulative regret.

Theorem 2 (Instance-dependent regret bound for PF-EFVI)**.** *Suppose Assumption 1 holds with a global coverage parameter $\lambda_{\min} > 0$. When M agents follow the* PF-EFVI *algorithm with zero-initialization ($V_{1,h}^m = 0$), then with probability at least $1-\delta$, the total system regret over K episodes is bounded by:*

$$\mathsf{Regret}(K) \le C \cdot \left(\frac{H^5 \log(4SAKMH/\delta)}{M\lambda_{\min}} \sum_{m=1}^{M} \frac{1}{\Delta_{\min}^m} + \frac{H \log(4SAKMH/\delta)}{\lambda_{\min}} + MH \right), \quad (7)$$

where $C > 0$ is a universal constant, and the instance-dependent gap $\Delta_{\min}^m$ for agent m is defined as the minimum suboptimality gap across all state-action pairs, i.e., $\Delta_{\min}^m = \min_{(s,a,h):\Delta_h^m(s,a)>0} \left(V_h^{m,\star}(s) - Q_h^{m,\star}(s,a)\right)$.

The proof of Theorem 2 is provided in Appendix D. The key implications of this result are as follows.

Logarithmic regret without exploration. Theorem 2 establishes that PF-EFVI attains an $\mathcal{O}(\log K)$ regret while every agent acts in a strictly greedy manner with respect to its own reward, without any UCB-style bonus, optimistic initialization, or random exploration. To our knowledge, this is the first gap-dependent, exploration-free regret analysis in the federated RL literature. The closest existing analyses leave a clear gap that our result fills. Zhang et al. (2025) establish a gap-dependent regret of $\widetilde{\mathcal{O}}(H^6 SA \log T/\Delta_{\min})$ for federated online Q-learning, but only for *homogeneous* MDPs and crucially through a UCB-based exploration mechanism. PF-EFVI achieves a comparable $\widetilde{\mathcal{O}}(H^5 \log K/(\lambda_{\min}\Delta_{\min}))$ instance-dependent rate (here, we take $\Delta_{\min}^m = \Delta_{\min}$ for simplicity of comparison), without any exploration bonus and under arbitrary reward heterogeneity, with the state-action complexity SA replaced by the effective coverage $1/\lambda_{\min}$ induced by reward diversity. A direct comparison with single-agent exploration-free RL is more delicate due to differences in assumptions: Civitavecchia and Papini (2025) achieve a worst-case guarantee for linear MDPs with a single agent, but their result fundamentally relies on strong environment stochasticity to ensure state-action coverage, which is a stringent requirement for any single agent. PF-EFVI removes this requirement entirely by sourcing coverage from cross-agent reward diversity rather than from the environment, so that even in deterministic or near-deterministic environments where their assumption is essentially vacuous, our $\log K$ guarantee continues to hold as long as $\lambda_{\min} > 0$.

From temporal exploration to spatial diversity. The coverage parameter $\lambda_{\min}$ enters the leading term of the regret bound only through the factor $1/(M\lambda_{\min})$, which can be interpreted as the inverse of an *effective sample size* per episode: at every episode, the ensemble's diverse greedy choices already cover the state-action space at rate $M\lambda_{\min}$, so that the factor SA in the $SA/\Delta_{\min}$ dependency of UCB-based methods is replaced by $1/\lambda_{\min}$. The two factors M and $\lambda_{\min}$ are complementary rather than interchangeable: increasing M alone, with all agents sharing the same reward, leaves $\lambda_{\min}$ at the coverage of a single greedy policy, which is typically zero, as Figure 4 illustrates empirically. The exploration cost is thus not eliminated but transferred from time to the population's reward diversity. The same $1/M$ prefactor also explains why the $\sum_m 1/\Delta_{\min}^m$ summation does not inflate the regret as one might fear: even an agent facing a hard task with very small $\Delta_{\min}^m$ inherits a federation-driven shrinkage of the global estimation error rather than relying on its own trajectories alone, and the per-agent cost of resolving a small gap decreases proportionally with M.

Complementary regimes with PF-UCBVI. An order-wise regret comparison between PF-EFVI and PF-UCBVI is not directly meaningful: PF-UCBVI enjoys an unconditional $\widetilde{\mathcal{O}}(\sqrt{SAMKH^3})$ guarantee that holds for any reward configuration, whereas PF-EFVI's $\mathcal{O}(\log K)$ bound is conditioned on Assumption 1: when $\lambda_{\min} = 0$, the ensemble's greedy executions can collapse onto a narrow subspace and PF-EFVI may fail to identify the optimal policy at all. The two algorithms therefore occupy complementary regimes: PF-UCBVI serves as a worst-case safety net, paying an explicit-exploration burn-in regardless of how rich the rewards are, while PF-EFVI enjoys a free lunch precisely when the population's reward diversity is strong enough that Assumption 1 is satisfied with

non-trivial $\lambda_{\min}$. In this regime, PF-EFVI matches or outperforms PF-UCBVI not by improving the regret but by avoiding the burn-in cost of its explicit UCB bonus, a phase transition empirically observed in Section 6.3.

Visitation-weighted regret bound. Let $\lambda_h(s,a) := \inf_{\widetilde{P}\in\mathcal{P}} \frac{1}{M}\sum_{m=1}^{M} d_h^{\widetilde{\pi}^m}(s,a)$ be the per-pair coverage, with $\widetilde{\pi}^m$ as in Assumption 1, so that $\min_{h,s,a}\lambda_h(s,a) \geq \lambda_{\min}$. Define

$$\Lambda^m := \sup_{\widetilde{P}\in\mathcal{P}} \left\{ \frac{1}{H}\sum_h \mathbb{E}_{(s,a)\sim d_h^{\widetilde{\pi}^m}}\left[\frac{1}{\lambda_h(s,a)}\right] + \frac{1}{H^2}\sum_h \mathbb{E}_{s\sim d_h^{\widetilde{\pi}^m}} \mathbb{E}^{\pi^{m,\star}}\left[\sum_{h'\geq h}\frac{1}{\lambda_{h'}(s_{h'},a_{h'})}\,\middle|\, s_h = s\right]\right\}, \tag{8}$$

where $\mathbb{E}^{\pi}[\cdot \mid s_h = s]$ is over the trajectory of π from $s_h = s$ in the true environment and $\Lambda^m \leq \frac{2}{\lambda_{\min}}$. Λ^m averages the inverse coverage along agent m's greedy trajectories and the optimal continuations from the states it reaches.

Proposition 1. *Under the conditions of Theorem 2, with probability at least* $1-\delta$*,*

$$\mathsf{Regret}(K) \leq C\left[\log(1+K_0)\left(\frac{H^5\iota}{M}\sum_{m=1}^{M}\frac{\Lambda^m}{\Delta_{\min}^m} + \frac{H^4\sqrt{S}\,\iota}{\lambda_{\min}}\right) + \frac{H\iota}{\lambda_{\min}} + MH\right], \tag{9}$$

where C is a universal constant, $K_0 = \max_m K_0^m$, and $K_0^m = \lceil C_0 H^4\iota/(M\lambda_{\min}(\Delta_{\min}^m)^2)\rceil$ for a sufficiently large universal constant C_0.

Hence the gap-dependent term no longer hinges on the worst-covered pair: it depends on the per-pair coverage $\lambda_h(s,a)$ only through its average Λ^m along agent m's greedy trajectories and their optimal continuations, so poorly covered pairs off these trajectories no longer enter this term. The proof is given in Appendix E.

6 Experiments

We complement the theoretical results with two empirical studies on a single unified family of synthetic MDPs: Section 6.2 provides empirical support for Assumption 1 under generic reward heterogeneity beyond the structural regimes of Section 5.2, and Section 6.3 compares PF-UCBVI and PF-EFVI as M varies, exhibiting a phase transition qualitatively consistent with our theory.

6.1 Experimental setup

We use synthetic tabular MDPs with $P_h(\cdot \mid s,a) \sim \text{Dirichlet}(\alpha\cdot\mathbf{1})$ and sparse heterogeneous rewards: each agent receives reward 1 on an independently drawn random subset of $\lceil \kappa\cdot SAH\rceil$ cells (h,s,a) and 0 elsewhere, where small α and κ make myopic behavior unlikely to reach the rewarding cells. The parameters and implementation details are given in Appendix G.

6.2 Empirical feasibility under generic reward heterogeneity

The two regimes in Section 5.2 establish that Assumption 1 is attainable under specific structural conditions. We now examine it empirically under *generic* reward heterogeneity, where each agent's preferred reward cells are drawn independently and neither regime applies by construction. To probe robustness to dynamics estimation error, we perturb the true kernel as $\widetilde{P} = (1-\epsilon)P + \epsilon U$ with $U_h(\cdot \mid s,a) \sim \text{Dirichlet}(\mathbf{1})$, and measure $\widehat{\lambda}_{\min}(\widetilde{P}) = \min_{(s,a,h)} \frac{1}{M}\sum_{m=1}^{M} d_h^{\widetilde{\pi}^m}(s,a)$.

Figure 2 reports the result. The curves show no downward trend in ϵ, indicating that collective coverage is not destroyed by dynamics estimation error, and from $M = 16$ on $\widehat{\lambda}_{\min}$ grows with M with curves overlapping across ϵ, indicating that coverage is driven by reward heterogeneity rather than by the kernel perturbation. To rule out that this simply reflects the number of policies executed, Figure 4 compares two reward setups at $M = 64$: identical rewards yield identical policies and highly uneven visitation, whereas i.i.d. rewards spread visits nearly uniformly across (s,h). The contrast isolates reward diversity, not ensemble size, as the source of coverage.

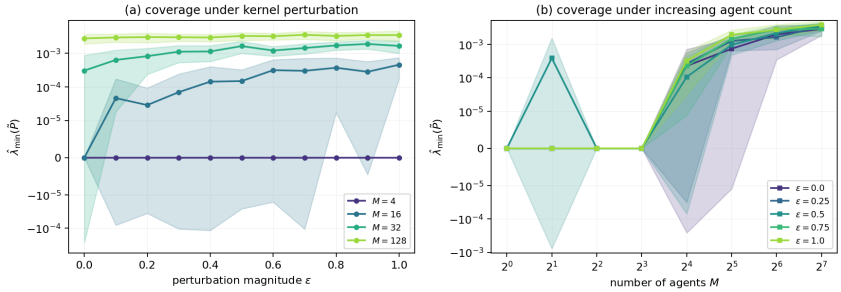

Figure 2: Empirical feasibility of Assumption 1 on random tabular MDPs. **(a)** $\widehat{\lambda}_{\min}(\widetilde{P})$ shows no downward trend in ϵ. **(b)** $\widehat{\lambda}_{\min}$ becomes positive from $M = 16$ and grows with M, overlapping across ϵ, indicating that reward heterogeneity drives coverage independently of $\widetilde{P}$.

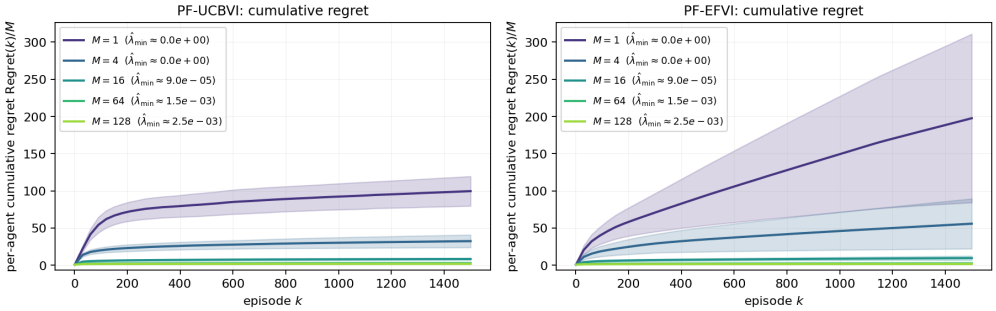

Figure 3: Per-agent cumulative regret $\text{Regret}(k)/M$ of PF-UCBVI (left) and PF-EFVI (right), with ± 1 std bands. PF-UCBVI converges sub-linearly at every M with a visible burn-in. PF-EFVI fails at small M ($\widehat{\lambda}_{\min} = 0$ for $M = 1, 4$), where its cumulative regret keeps growing linearly, but matches or beats PF-UCBVI once $M \geq 16$ by avoiding the explicit-exploration burn-in.

6.3 Performance of PF-UCBVI and PF-EFVI

We compare the per-agent cumulative regret of PF-UCBVI and PF-EFVI on the unified MDP family of Section 6.1, in which the myopic r-greedy policy is far from optimal. We sweep $M \in \{1, 4, 16, 64, 128\}$ over $K = 1500$ episodes and report results averaged over 8 MDP draws. Figure 3 displays the per-agent cumulative regret $\text{Regret}(k)/M$.

The results support the theory along three axes. First, PF-UCBVI's cumulative regret flattens at every M, faster for larger M, qualitatively consistent with the linear speedup of Theorem 1. Second, at $M = 1, 4$ the empirical $\widehat{\lambda}_{\min}$ is identically zero across all 8 MDP draws: PF-EFVI's greedy agents cannot escape the narrow region induced by their own policies, its cumulative regret keeps growing linearly, reaching roughly twice that of PF-UCBVI by $K = 1500$ at $M = 1$, indicating that exploitation alone can fail when reward heterogeneity is insufficient. Third, once $M \geq 16$, $\widehat{\lambda}_{\min}$ becomes positive and grows with M, and PF-EFVI's per-agent regret matches PF-UCBVI (within standard deviations at $M = 16, 64$, strictly below at $M = 128$). The advantage at large M is precisely the burn-in cost PF-EFVI avoids by relying on heterogeneous greedy behaviours instead of explicit bonuses, in line with the temporal-to-spatial trade-off of Theorem 2. A comparison with Fed-UCBVI is in Appendix G.

7 Discussion

This work argues that reward heterogeneity is a structural advantage in FedRL. PF-UCBVI demonstrates that decoupling shared dynamics from personalized rewards ensures optimal linear speedup with no heterogeneity penalty in the regret bound. Moving beyond linear speedup, PF-EFVI reveals that spatial diversity can eliminate the need for risky, explicit exploration altogether. This blessing, however, is a conditional one: when the population's rewards are not diverse enough, the ensemble's greedy executions can collapse and exploitation alone is no longer sufficient.

Acknowledgments and Disclosure of Funding

The work of G. Joshi is supported in part by the grants NSF-CCF 2007834, CCF-2045694, CNS-2112471, and ONR N00014-23-1-2149. The work of Y. Chi is supported in part by NSF under grants ECCS-2537189 and CNS-2148212, and in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program. The authors declare no competing interests.

References

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. JMLR. org.

Bastani, H., Bayati, M., and Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349.

Chen, Y., Chi, Y., Fan, J., Li, G., Wei, Y., and Yan, Y. (2026). Statistical foundations of reinforcement learning: A non-asymptotic perspective. *arXiv preprint arXiv:2610.41903*.

Chen, Y., Zhang, X., Zhang, K., Wang, M., and Zhu, X. (2023). Byzantine-robust online and offline distributed reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3230–3269. PMLR.

Civitavecchia, L. and Papini, M. (2025). Exploration-free reinforcement learning with linear function approximation. In *Reinforcement Learning Conference*.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*.

Jiao, Y., Woo, J., Li, G., Joshi, G., and Chi, Y. (2026). Sample complexity of average-reward Q-learning: From single-agent to federated reinforcement learning. *arXiv preprint arXiv:2601.13642*.

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.

Kannan, S., Morgenstern, J., Roth, A., Waggoner, B., and Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

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.

Labbi, S., Mangold, P., Tiapkin, D., and Moulines, E. (2025a). On global convergence rates for federated policy gradient under heterogeneous environment. *arXiv preprint arXiv:2505.23459*.

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

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

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. In *Proc. Computational Learning Theory Conference*.

Salgia, S. and Chi, Y. (2024). The sample-communication complexity trade-off in federated Q-learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wang, H., He, S., Zhang, Z., Miao, F. M., and Anderson, J. (2024a). Momentum for the win: collaborative federated reinforcement learning across heterogeneous environments. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wang, L. M., Yang, P., and Su, L. (2026). Personalized multi-agent average reward TD-learning via joint linear approximation. *arXiv preprint arXiv:2603.02426*.

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

Weissman, T., Ordentlich, E., Seroussi, G., Verdú, S., and Weinberger, M. J. (2003). Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97, Hewlett-Packard Labs.

Woo, J., Joshi, G., and Chi, Y. (2023). The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37157–37216. PMLR.

Woo, J., Joshi, G., and Chi, Y. (2025). The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *Journal of Machine Learning Research*, 26(26):1–85.

Woo, J., Shi, L., Joshi, G., and Chi, Y. (2024). Federated offline reinforcement learning: Collaborative single-policy coverage suffices. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 53165–53201. PMLR.

Xiong, G., Wang, S., Jiang, D., and Li, J. (2025). On the linear speedup of personalized federated reinforcement learning with shared representations. In *The Thirteenth International Conference on Learning Representations*.

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhang, C. and Azizan, N. (2026). Personalized collaborative learning with affinity-based variance reduction. In *The Fourteenth International Conference on Learning Representations*.

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, H., Zheng, Z., and Xue, L. (2025). Gap-dependent bounds for federated q-learning. In *Proceedings of the 42nd International Conference on Machine Learning*.

Zheng, Z., Gao, F., Xue, L., and Yang, J. (2024). 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 *ICLR*.

Zhu, F., Mitra, A., and Heath, R. W. (2025). Achieving tighter finite-time rates for heterogeneous federated stochastic approximation under markovian sampling. *arXiv preprint arXiv:2504.11645*.

A Related Works

To situate our work within the broader context of federated RL, we review three critical research thrusts: homogeneous federated RL, heterogeneous federated RL, and exploration-free RL.

Federated RL with homogeneous environments. A large body of work on federated RL assumes homogeneous environments where all agents share the same MDP. For sample complexity, Khodadadian et al. (2022); Woo et al. (2025) establish linear speedup under generative or Markovian sampling. Woo et al. (2024) extend this line to the offline setting under single-policy coverage, and Jiao et al. (2026) extended to the average-reward setting. For online regret minimization in episodic tabular MDPs, Zheng et al. (2024) propose a federated Q-learning algorithm with Hoeffding and Bernstein style UCB bonuses, achieving linear regret speedup with low communication cost, which Zheng et al. (2025) subsequently improve via reference-advantage decomposition to attain near-optimal regret. Salgia and Chi (2024) characterize the fundamental sample-communication trade-off, while Zhang et al. (2025) establish the gap-dependent regret bounds for online federated Q-learning. These results form the natural baseline, but they do not address agents with personalized rewards.

Federated RL with heterogeneous environments. While the majority of federated RL literature assumes homogeneous environments, recent works have begun to explore the more realistic setting where agents possess different transition dynamics and reward functions. Existing literature in this area primarily focuses on two directions:

1. **Consensus-based methods** (Wang et al., 2024a; Zhang et al., 2024; Labbi et al., 2025b) learn a single global policy optimizing the average value across environments. Early baselines along this line include Jin et al. (2022), who introduce QAvg and PAvg under transition heterogeneity. Wang et al. (2024b) subsequently provide a fine-grained analysis of federated Q-learning showing that local updates introduce a non-vanishing heterogeneity bias, and Mangold et al. (2025) prove that FedSARSA does not converge to the optimum of a virtual averaged environment but to an environment-coupled fixed point with a similar bias. From a stochastic approximation perspective, Zhu et al. (2025) achieve linear speedup under Markovian sampling without projection. On the policy gradient side, Yang et al. (2024) develop federated natural policy gradient and actor critic methods under decentralized topology and reward heterogeneity, Labbi et al. (2025a) reveal that under heterogeneity, optimal federated policies can be non-deterministic or even time-varying, and establish global convergence under local Łojasiewicz conditions. Beyond cooperative settings, Chen et al. (2023) consider Byzantine-robust distributed RL where a fraction of agents may report adversarial information. Despite their progress, all consensus-based methods inherently sacrifice individual utility for global agreement: as the average value can be far from any agent's optimum, the resulting policy can be sub-optimal for individual agents, and the benefits of collaboration shrink, and may turn into negative transfer, as heterogeneity grows.
2. **Personalized methods** instead allow agents to learn individualized policies. Zhang and Azizan (2026) prove linear speedup via affinity-based variance reduction under similarity conditions, while Xiong et al. (2025) and Wang et al. (2026) achieve linear speedup through shared low-dimensional representations for federated TD and average-reward TD learning, respectively. However, these methods either rely on similarity, low-rank, or shared-representation assumptions across agents, or assume a fixed behavior policy that already ensures sufficient coverage, and most analyses focus on policy evaluation rather than online regret minimization.

Despite these advances, most personalized algorithms assume a degree of similarity between agents or rely on a fixed behavior policy that ensures sufficient coverage of the state-action space. Our work removes both restrictions: PF-UCBVI is fully agnostic to the level of reward heterogeneity, and PF-EFVI sources coverage from reward diversity rather than from a prescribed behavior policy.

Exploration-free learning and covariate diversity. In practice, many RL applications avoid explicit exploration in favor of a greedy approach, particularly in domains such as robotics or clinical trials where explorative actions may be risky, unethical, or prohibitively expensive. Beyond safety concerns, explorative solutions are often harder to implement, and their outcomes are less predictable or interpretable. Consequently, common sense often suggests a greedy strategy in real-world scenarios, even when standard theory dictates the necessity of exploration. Foundational studies by Bastani et al. (2021) and Kannan et al. (2018) have investigated this phenomenon in linear contextual bandits. They established that under a sufficient diversity of contexts, known as the covariate-diversity assumption, a purely greedy selection can achieve sublinear regret without any explicit exploration mechanisms. Recently, Civitavecchia and Papini (2025) further characterized the covariate-diversity condition for RL. However, their results still require strong environment stochasticity to ensure state-action coverage, which remains a stringent requirement for a single agent. To our knowledge, no prior work simultaneously addresses (i) the federated heterogeneous setting, (ii) online regret minimization, and (iii) exploration-free learning; PF-EFVI fills this gap by sourcing coverage from cross-agent reward diversity rather than from environment stochasticity.

B Notation

Throughout the analysis, $X_{k,h}$ for $X \in \{N, \widehat{P}, B^m, Q^m, V^m\}$ denotes the value of X at the *beginning of episode* k (i.e., before episode k's execution step). All these values are computed using data from episodes $1, \dots, k-1$. In particular:

- $X_{1,h}$ refers to the initialization (no data collected yet);

Table 1: Comparison with related lines of work.

Line of work	Task	Exploration	Personalized
Homogeneous federated RL (Khodadadian et al., 2022; Woo et al., 2025, 2024; Zheng et al., 2024, 2025; Zhang et al., 2025; Salgia and Chi, 2024)	control, evaluation	via UCB bonuses, or none (prescribed behavior policies)	×
Heterogeneous federated RL, consensus (Jin et al., 2022; Wang et al., 2024a,b; Mangold et al., 2025; Zhu et al., 2025; Yang et al., 2024; Labbi et al., 2025b,a)	control	via UCB bonuses, or none (prescribed behavior policies)	×
Heterogeneous federated RL, personalized (Zhang and Azizan, 2026; Xiong et al., 2025; Wang et al., 2026)	evaluation	none (fixed behavior policy)	✓
Single-agent exploration-free RL (Bastani et al., 2021; Kannan et al., 2018; Civitavecchia and Papini, 2025)	control	free, via context or environment stochasticity	×
PF-UCBVI (ours)	control	via personalized UCB bonuses	✓
PF-EFVI (ours)	control	free, via reward diversity	✓

- $N_{k,h}(s,a) = \sum_{m=1}^{M} \sum_{i=1}^{k-1} \mathbb{I}\{(s_{i,h}^m, a_{i,h}^m) = (s,a)\}$;
- $\widehat{P}_{k,h}(s'|s,a) = N_{k,h}(s,a,s') / \max\{1, N_{k,h}(s,a)\}$. For PF-EFVI, $\widehat{P}_{k,h}(\cdot|s,a)$ is instead the uniform distribution when $N_{k,h}(s,a) = 0$
- $\widehat{P}_{k,h,s,a} = \widehat{P}_{k,h}(\cdot|s,a)$ and $P_{h,s,a} = P_h(\cdot|s,a)$.
- $\iota = \log(28M^2SAKH/\delta)$.

The policy executed in episode k is $\pi_{k,h}^m(s) = \arg\max_a Q_{k,h}^m(s,a)$. For PF-UCBVI, the tie-breaking draw of episode k is made before the episode and is included in the filtration $\mathcal{F}_\tau$ of Step 4 from the start of episode k, so π_k^m is a deterministic policy given the history. For PF-EFVI, ties are broken uniformly at random, and $\pi_{k,h}^m(\cdot|s)$ denotes the resulting uniform distribution over $\arg\max_a Q_{k,h}^m(s,a)$.

C Proof of Theorem 1

To analyze the regret of PF-UCBVI, we bound the total expected regret as follows:

$$\text{Regret}(K) = \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{E}_{s_1 \sim \rho} \left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1) \right],$$

where, for each agent m and episode k, we write $\Delta_{k,h}^m(s) = V_{k,h}^m(s) - V_h^{m,\star}(s)$ and $\Delta_{k,h}^{m,\pi_k^m}(s) = V_{k,h}^m(s) - V_h^{m,\pi_k^m}(s)$. We first prove the optimism of the value function, i.e., $\Delta_{k,h}^m(s) \geq 0$ for all $(k,h,m,s) \in [K] \times [H] \times [M] \times \mathcal{S}$.

Step 1: Optimism. We want to show that $\Delta_{k,h}^m(s) = V_{k,h}^m(s) - V_h^{m,\star}(s) \geq 0$ for all k,h,m,s. We can prove this by induction on h. For $h = H+1$, $V_{k,H+1}^m(s) = V_{H+1}^{m,\star}(s) = 0$ for all m,k,s. We can then use backward induction to show that the optimism holds for all $h \in [H]$. For any (s,h) at episode k of agent m, denoting $a^\star = \pi_h^{m,\star}(s)$ for notational simplicity, since $Q_{k,h}^m = \min\{H-h+1,\ r_h^m + \widehat{P}V_{k,h+1}^m + B_{k,h}^m\}$, we first prove that if the minimum in $Q_{k,h}^m(s,a^\star)$ is attained by $H-h+1$, then $Q_{k,h}^m(s,a^\star) = H-h+1 \geq Q_h^{m,\star}(s,a^\star)$; otherwise, we can prove the optimism by the estimation error decomposition as follows:

$$\Delta_{k,h}^m(s) = V_{k,h}^m(s) - V_h^{m,\star}(s) \overset{(i)}{\geq} Q_{k,h}^m(s,a^\star) - Q_h^{m,\star}(s,a^\star)$$

$$
\begin{aligned}
&= \left(r_h^m(s,a^\star) + \widehat{P}_{k,h,s,a^\star} V_{k,h+1}^m + B_{k,h}^m(s,a^\star)\right) \\
&\quad - \left(r_h^m(s,a^\star) + P_{h,s,a^\star} V_{h+1}^{m,\star}\right) \\
&= \left(\widehat{P}_{k,h,s,a^\star} - P_{h,s,a^\star}\right) V_{k,h+1}^m + P_{h,s,a^\star}\left(V_{k,h+1}^m - V_{h+1}^{m,\star}\right) \\
&\quad + B_{k,h}^m(s,a^\star) \\
&\overset{(ii)}{\geq} -B_{k,h}^m(s,a^\star) - \frac{2}{H} P_{h,s,a^\star} \Delta_{k,h+1}^m \\
&\quad + P_{h,s,a^\star} \Delta_{k,h+1}^m + B_{k,h}^m(s,a^\star) \\
&= \left(1 - \frac{2}{H}\right) P_{h,s,a^\star} \Delta_{k,h+1}^m \geq 0.
\end{aligned} \tag{10}
$$

where (i) holds as $V_{k,h}^m(s) \geq Q_{k,h}^m(s,a^\star)$ and $V_h^{m,\star}(s) = Q_h^{m,\star}(s,a^\star)$ and (ii) follows from Lemma 1 and the induction hypothesis $\Delta_{k,h+1}^m(s) \geq 0$ for all s; for $H = 1$, the last line is 0 since $\Delta_{k,2}^m = 0$.

Lemma 1. *(Transition estimation error bound) With probability at least $1 - 2\delta/7$, simultaneously for all episodes k, steps h, state-action pairs (s,a), and agents m such that $\Delta_{k,h+1}^m \geq 0$, we have*

$$
-B_{k,h}^m(s,a) - \frac{2}{H} P_{h,s,a} \Delta_{k,h+1}^m \leq \left(\widehat{P}_{k,h,s,a} - P_{h,s,a}\right) V_{k,h+1}^m \leq B_{k,h}^m(s,a) + \frac{2}{H} P_{h,s,a} \Delta_{k,h+1}^m \tag{11}
$$

Once the optimism is established, for each agent m at episode k, we can decompose the regret as follows:

$$
\mathbb{E}_{s_1 \sim \rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right] \leq \mathbb{E}_{s_1 \sim \rho}\left[V_{k,1}^m(s_1) - V_1^{m,\pi_k^m}(s_1)\right] \tag{12}
$$

Step 2: Establishing the Recursion. To further bound $\Delta_{k,h}^{m,\pi_k^m} = V_{k,h}^m - V_h^{m,\pi_k^m}$ in (12), for any step h and state-action pair (s_h, a_h) induced by ρ and $a_h = \pi_{k,h}^m(s_h) = \arg\max_{a \in \mathcal{A}} Q_{k,h}^m(s_h,a)$, we decompose the estimation error as follows:

$$
\begin{aligned}
&\Delta_{k,h}^{m,\pi_k^m}(s_h) = V_{k,h}^m(s_h) - V_h^{m,\pi_k^m}(s_h) = Q_{k,h}^m(s_h,a_h) - Q_h^{m,\pi_k^m}(s_h,a_h) \\
&\leq \min\left\{H, r_h^m(s_h,a_h) + \widehat{P}_{k,h,s_h,a_h} V_{k,h+1}^m + B_{k,h}^m(s_h,a_h) \; - \left(r_h^m(s_h,a_h) + P_{h,s_h,a_h} V_{h+1}^{m,\pi_k^m}\right)\right\} \\
&\leq \min\left\{H, B_{k,h}^m(s_h,a_h) + \left(\widehat{P}_{k,h,s_h,a_h} - P_{h,s_h,a_h}\right) V_{k,h+1}^m + P_{h,s_h,a_h} \Delta_{k,h+1}^{m,\pi_k^m}\right\} \\
&\overset{(i)}{\leq} \min\left\{H, 2B_{k,h}^m(s_h,a_h) + \frac{2}{H} P_{h,s_h,a_h} \Delta_{k,h+1}^m + P_{h,s_h,a_h} \Delta_{k,h+1}^{m,\pi_k^m}\right\} \\
&\overset{(ii)}{\leq} \min\{H, 2B_{k,h}^m(s_h,a_h)\} + \left(1 + \frac{2}{H}\right) P_{h,s_h,a_h} \Delta_{k,h+1}^{m,\pi_k^m},
\end{aligned} \tag{13}
$$

where (i) holds due to Lemma 1, which yields:

$$
\left(\widehat{P}_{k,h,s_h,a_h} - P_{h,s_h,a_h}\right) V_{k,h+1}^m \leq B_{k,h}^m(s_h,a_h) + \frac{2}{H} P_{h,s_h,a_h} \Delta_{k,h+1}^m, \tag{14}
$$

(ii) holds because $V_{h+1}^{m,\star} \geq V_{h+1}^{m,\pi_k^m}$ and $\Delta_{k,h+1}^m \leq \Delta_{k,h+1}^{m,\pi_k^m}$ according to the optimism shown in Step 1.

Step 3: Variance Transfer. Recall the definition of the exploration bonus $B_{k,h}^m(s,a)$, which utilizes the empirical variance of $V_{k,h+1}^m$:

$$
B_{k,h}^m(s,a) = \sqrt{\frac{4\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V_{k,h+1}^m)\iota}{\max\{1, N_{k,h}(s,a)\}}} + \frac{25SH^2\iota}{\max\{1, N_{k,h}(s,a)\}}, \tag{15}
$$

where ι is defined in the Notation. To bound the bonus term, we need to transfer the empirical variance of V_k to that of $V^\star$, and then transfer the empirical variance $\mathsf{Var}_{\widehat{P}}$ to the true variance Var_P.

- $\mathsf{Var}_{\widehat{P}}(V^m_{k,h+1}) \to \mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1})$: First, we transfer the empirical variance of the estimated value function to that of the optimal value function. Writing $V^m_{k,h+1} = V^{m,\star}_{h+1} + \Delta^m_{k,h+1}$ with $\Delta^m_{k,h+1} = V^m_{k,h+1} - V^{m,\star}_{h+1}$, and denoting $\widehat{P} = \widehat{P}_{k,h,s,a}$ for brevity, we have

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}}(V^m_{k,h+1}) &= \mathsf{Var}_{\widehat{P}}\big(V^{m,\star}_{h+1} + \Delta^m_{k,h+1}\big) \\
&= \mathbb{E}_{s'\sim\widehat{P}}\Big[\Big(\big(V^{m,\star}_{h+1}(s') - \widehat{P}V^{m,\star}_{h+1}\big) + \big(\Delta^m_{k,h+1}(s') - \widehat{P}\Delta^m_{k,h+1}\big)\Big)^2\Big] \\
&\overset{(i)}{\le} 2\,\mathbb{E}_{s'\sim\widehat{P}}\Big[\big(V^{m,\star}_{h+1}(s') - \widehat{P}V^{m,\star}_{h+1}\big)^2\Big] + 2\,\mathbb{E}_{s'\sim\widehat{P}}\Big[\big(\Delta^m_{k,h+1}(s') - \widehat{P}\Delta^m_{k,h+1}\big)^2\Big] \\
&= 2\mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1}) + 2\mathsf{Var}_{\widehat{P}}(\Delta^m_{k,h+1}) \\
&\overset{(ii)}{=} 2\mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1}) + 2\Big(\widehat{P}\big(\Delta^m_{k,h+1}\big)^2 - \big(\widehat{P}\Delta^m_{k,h+1}\big)^2\Big) \\
&\overset{(iii)}{\le} 2\mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1}) + 2\,\widehat{P}\big(\Delta^m_{k,h+1}\big)^2 \\
&\overset{(iv)}{\le} 2\mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1}) + 2H\,\widehat{P}\Delta^m_{k,h+1},
\end{aligned} \tag{16}$$

  where (i) follows from the elementary inequality $(x+y)^2 \le 2x^2 + 2y^2$ applied pointwise, (ii) is the definition of variance, (iii) drops the non-negative term $(\widehat{P}\Delta^m_{k,h+1})^2 \ge 0$, and (iv) holds since the optimism property (Step 1) and $V^m_{k,h+1} \le H$ and $V^{m,\star}_{h+1} \ge 0$ guarantee $0 \le \Delta^m_{k,h+1}(s') \le H$ for all s'.
- $\mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1}) \to \mathsf{Var}_P(V^{m,\star}_{h+1})$: Next, we convert the empirical variance of the optimal value function to the true variance by applying Lemma 2 as follows:

$$\mathsf{Var}_{\widehat{P}}(V^{m,\star}_{h+1}) \le 2\mathsf{Var}_P(V^{m,\star}_{h+1}) + \frac{10H^2\iota}{\max\{N_{k,h}(s,a),1\}}. \tag{17}$$

  Lemma 2. *(Empirical Variance Gap) With probability at least $1-2\delta/7$, for all (k,h,s,a,m), the deviation between the empirical and true variance of the fixed function $V^{m,\star}_{h+1} \in [0,H]$ is bounded by:*

$$\big|\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V^{m,\star}_{h+1}) - \mathsf{Var}_{P_{h,s,a}}(V^{m,\star}_{h+1})\big| \le \mathsf{Var}_{P_{h,s,a}}(V^{m,\star}_{h+1}) + \frac{10H^2\iota}{\max\{N_{k,h}(s,a),1\}}. \tag{18}$$

- **Bound on $\widehat{P}\Delta^m_{k,h+1}$**: Invoking the bound on $|(\widehat{P}-P)\Delta^m_{k,h+1}|$ established in the proof of Lemma 1, we have

$$\begin{aligned}
\widehat{P}\Delta^m_{k,h+1} \le P\Delta^m_{k,h+1} + \big|(\widehat{P}-P)\Delta^m_{k,h+1}\big| &\le \Big(1+\frac{1}{H}\Big)P\Delta^m_{k,h+1} + \frac{9SH^2\iota}{\max\{N_{k,h}(s,a),1\}} \\
&\le 2P\Delta^m_{k,h+1} + \frac{9SH^2\iota}{\max\{N_{k,h}(s,a),1\}}.
\end{aligned} \tag{19}$$

Combining the variance transfer above, (17) and (19),

$$\mathsf{Var}_{\widehat{P}}(V^m_{k,h+1}) \le 4\mathsf{Var}_P(V^{m,\star}_{h+1}) + 4H\,P_{h,s,a}\Delta^m_{k,h+1} + \frac{38\,SH^3\iota}{\max\{N_{k,h}(s,a),1\}}. \tag{20}$$

Substituting this into the bonus definition (15),

$$\begin{aligned}
B^m_{k,h}(s,a) \le \sqrt{\frac{16\mathsf{Var}_{P_{h,s,a}}(V^{m,\star}_{h+1})\iota}{\max\{N_{k,h}(s,a),1\}}} + \sqrt{\frac{16H\iota\cdot P_{h,s,a}\Delta^m_{k,h+1}}{\max\{N_{k,h}(s,a),1\}}} \\
+ \frac{13\sqrt{S}H^{3/2}\iota}{\max\{N_{k,h}(s,a),1\}} + \frac{25SH^2\iota}{\max\{N_{k,h}(s,a),1\}}.
\end{aligned} \tag{21}$$

For the second term, applying Young's inequality $\sqrt{xy} \le \alpha x + \frac{y}{4\alpha}$ with $\alpha = \frac{1}{4e^2H}$, $x = P_{h,s,a}\Delta^m_{k,h+1}$, and $y = \frac{16H\iota}{\max\{N_{k,h}(s,a),1\}}$:

$$\sqrt{\frac{16H\iota\cdot P_{h,s,a}\Delta^m_{k,h+1}}{\max\{N_{k,h}(s,a),1\}}} \le \frac{1}{4e^2H}P_{h,s,a}\Delta^m_{k,h+1} + \frac{16e^2H^2\iota}{\max\{N_{k,h}(s,a),1\}}. \tag{22}$$

Since $S, H \geq 1$, the three lower-order terms satisfy $13\sqrt{S}H^{3/2} + 25SH^2 + 16e^2H^2 \leq 157SH^2$, and we arrive at

$$B^m_{k,h}(s,a) \leq \underbrace{\sqrt{\frac{16\mathsf{Var}_{P_{h,s,a}}(V^{m,\star}_{h+1})\iota}{\max\{N_{k,h}(s,a),1\}}} + \frac{157SH^2\iota}{\max\{N_{k,h}(s,a),1\}}}_{=:\widetilde{B}^m_{k,h}(s,a)} + \frac{1}{4e^2H}P_{h,s,a}\Delta^m_{k,h+1}$$
$$\leq \widetilde{B}^m_{k,h}(s,a) + \frac{1}{4e^2H}P_{h,s,a}\Delta^{m,\pi^m_k}_{k,h+1}, \tag{23}$$

defining the residual-free bonus $\widetilde{B}^m_{k,h}(s,a)$ that depends only on the true variance of the optimal value function $V^{m,\star}_{h+1}$ and the last term holds since $\Delta^m_{k,h+1} \leq \Delta^{m,\pi^m_k}_{k,h+1}$ by optimism (Step 1).

Step 4: Bounding the Total Regret via Martingale Concentration. Integrating the variance transfer (23) from Step 3 into the estimation error bound (13) of Step 2, substituting the bonus term $B^m_{k,h}$ with $\widetilde{B}^m_{k,h}$, we have

$$\Delta^{m,\pi^m_k}_{k,h}(s_h) \leq \min\{H, 2\widetilde{B}^m_{k,h}(s_h,a_h)\} + \left(1+\frac{2}{H}\right)P_{h,s_h,a_h}\Delta^{m,\pi^m_k}_{k,h+1} + \frac{1}{2e^2H}P_{h,s_h,a_h}\Delta^{m,\pi^m_k}_{k,h+1}$$
$$\leq \min\{H, 2\widetilde{B}^m_{k,h}(s_h,a_h)\} + \left(1+\frac{3}{H}\right)P_{h,s_h,a_h}\Delta^{m,\pi^m_k}_{k,h+1}. \tag{24}$$

For the realized trajectory of agent m in episode k, let $h \in [H]$ and $s^m_{k,h}, a^m_{k,h}$ be the state and action at time h and $P = P_{h,s^m_{k,h},a^m_{k,h}}$,

$$\Delta^{m,\pi^m_k}_{k,h}(s^m_{k,h}) \leq \min\{H, 2\widetilde{B}^m_{k,h}(s^m_{k,h},a^m_{k,h})\}$$
$$+ \left(1+\frac{3}{H}\right)\Big(\Delta^{m,\pi^m_k}_{k,h+1}(s^m_{k,h+1}) + \underbrace{(P_{h,s^m_{k,h},a^m_{k,h}}\Delta^{m,\pi^m_k}_{k,h+1} - \Delta^{m,\pi^m_k}_{k,h+1}(s^m_{k,h+1}))}_{:=\xi^m_{k,h}}\Big) \tag{25}$$

where $\xi^m_{k,h}$ is a martingale difference, which has conditional mean zero given the past and $(s^m_{k,h}, a^m_{k,h})$, and $\xi^m_{k,h}$ lies in an interval of length $2H$ given the past, since $V^m_{k,h+1}, V^{m,\pi^m_k}_{h+1} \in [0,H]$.

Unrolling the recursion along the realized trajectory from $h=1$ to H, with $\Delta^{m,\pi^m_k}_{k,H+1} = 0$, and then using $(1+3/H)^{h-1} \leq e^3$ and $\min\{H, 2x\} \leq 2\min\{H,x\}$,

$$\Delta^{m,\pi^m_k}_{k,1}(s^m_{k,1}) \leq \sum_{h=1}^{H}\left(1+\frac{3}{H}\right)^{h-1}\min\{H, 2\widetilde{B}^m_{k,h}(s^m_{k,h},a^m_{k,h})\} + \sum_{h=1}^{H}\left(1+\frac{3}{H}\right)^h\xi^m_{k,h}$$
$$\leq 2e^3\sum_{h=1}^{H}\min\{H, \widetilde{B}^m_{k,h}(s^m_{k,h},a^m_{k,h})\} + \sum_{h=1}^{H}\left(1+\frac{3}{H}\right)^h\xi^m_{k,h}. \tag{26}$$

We now bound the total regret by summing over all agents $m \in [M]$ and all episodes $k \in [K]$ using the bounds from the previous steps, and applying the martingale difference sequence (MDS) concentration argument to the sum of martingale differences $\xi^m_{k,h}$:

$$\mathsf{Regret}(K) \leq \sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho}\big[\Delta^{m,\pi^m_k}_{k,1}(s_1)\big]$$
$$= \sum_{m=1}^{M}\sum_{k=1}^{K}\Big(\Delta^{m,\pi^m_k}_{k,1}(s^m_{k,1}) + \underbrace{\mathbb{E}_{s_1\sim\rho}[\Delta^{m,\pi^m_k}_{k,1}(s_1)] - \Delta^{m,\pi^m_k}_{k,1}(s^m_{k,1})}_{=:\zeta^m_k}\Big)$$
$$\leq 2e^3\sum_{m,k,h}\min\{H, \widetilde{B}^m_{k,h}(s^m_{k,h},a^m_{k,h})\} + \sum_{m=1}^{M}\sum_{k=1}^{K}\Big(\zeta^m_k + \sum_{h=1}^{H}\left(1+\frac{3}{H}\right)^h\xi^m_{k,h}\Big)$$

$$= 2e^3 \sum_{m,k,h} \min\{H, \widetilde{B}^m_{k,h}(s^m_{k,h}, a^m_{k,h})\} + \sum_{\tau=1}^{MK(H+1)} X_\tau,$$

where $X_\tau = \zeta^m_k$ for $h = 0$, $X_\tau = (1 + 3/H)^h \xi^m_{k,h}$ for $h \geq 1$ for each τ corresponding to exactly one (k, m, h) with $h \in \{0, \ldots, H\}$.

To apply the Azuma-Hoeffding inequality to $\{X_\tau\}$, we order the steps of all agents and episodes. We define a global rollout index $\tau \in [1, MK(H+1)]$ that strictly flattens the multi-agent episodes into a step-by-step sequential timeline:

$$\tau = (k-1)M(H+1) + (m-1)(H+1) + h + 1 \quad \text{for} \quad k \in [K],\ m \in [M],\ h \in \{0, \ldots, H\}. \tag{27}$$

Let $\mathcal{F}_\tau$ be the filtration (a σ-algebra) capturing all information up to and including the state-action pair of the τ-th step; the slot $h = 0$ carries no state-action pair. Formally, $\mathcal{F}_\tau$ is defined as:

$$\mathcal{F}_\tau = \sigma\left(\bigcup_{i=1}^{\tau} \left\{s^{(i)}, a^{(i)}\right\}\right), \quad \mathcal{F}_1 = \{\emptyset, \Omega\}, \tag{28}$$

where the superscript (i) refers to the state-action pair at the specific global step index i. Since the global estimator $\widehat{P}$ and the counter N are updated strictly using past data contained in the episode-start filtration $\mathcal{F}_{(k-1)M(H+1)+1} \subset \mathcal{F}_\tau$, the policy π^m_k and the bonus function $\widetilde{B}^m_{k,h}$ evaluated during step τ are $\mathcal{F}_\tau$-measurable.

As X_τ is $\mathcal{F}_{\tau+1}$-measurable, with $\mathbb{E}[X_\tau \mid \mathcal{F}_\tau] = 0$ and X_τ lying in an $\mathcal{F}_\tau$-measurable interval of length $2e^3 H$, applying the Azuma-Hoeffding inequality to the bounded MDS $\sum_{\tau=1}^{MK(H+1)} X_\tau$, we guarantee that with probability at least $1 - \delta/7$:

$$\sum_{\tau=1}^{MK(H+1)} X_\tau \leq \sqrt{\frac{1}{2} \sum_{\tau=1}^{MK(H+1)} \left(2e^3 H\right)^2 \iota} \leq 2e^3 \sqrt{H^3 MK\iota}. \tag{29}$$

Substituting the concentration bound (29) into the total expected regret, we have

$$\begin{aligned}
\mathsf{Regret}(K) &\leq \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{E}_{s_1 \sim \rho}\left[\Delta^{m, \pi^m_k}_{k,1}(s_1)\right] \\
&\leq 2e^3 \sum_{m,k,h} \min\{H, \widetilde{B}^m_{k,h}(s^m_{k,h}, a^m_{k,h})\} + 2e^3 \sqrt{H^3 MK\iota},
\end{aligned} \tag{30}$$

Step 5: Finalizing the Regret Bound. Now, we substitute the upper bound of the empirical exploration bonus into the total expected regret.

$$\begin{aligned}
\mathsf{Regret}(K) &\leq 2e^3 \sum_{m,k,h} \min\{H, \widetilde{B}^m_{k,h}(s^m_{k,h}, a^m_{k,h})\} + 2e^3 \sqrt{H^3 MK\iota} \\
&\leq 4e^3 MSAH^2 + 2e^3 \sum_{N_{k,h} \geq M} \widetilde{B}^m_{k,h}(s^m_{k,h}, a^m_{k,h}) + 2e^3 \sqrt{H^3 MK\iota} \\
&\leq 4e^3 MSAH^2 + 8e^3 \sum_{N_{k,h} \geq M} \sqrt{\frac{\mathsf{Var}_{P_{h, s^m_{k,h}, a^m_{k,h}}}(V^{m,\star}_{h+1})\iota}{\max\{N_{k,h}(s^m_{k,h}, a^m_{k,h}), 1\}}} \\
&\quad + 314e^3 \sum_{N_{k,h} \geq M} \frac{SH^2\iota}{\max\{N_{k,h}(s^m_{k,h}, a^m_{k,h}), 1\}} + 2e^3 \sqrt{H^3 MK\iota} \\
&\leq 8e^3 \sqrt{\iota} \underbrace{\sqrt{\sum_{m,k,h} \mathsf{Var}_{P_{h, s^m_{k,h}, a^m_{k,h}}}(V^{m,\star}_{h+1})(s^m_{k,h}, a^m_{k,h})}}_{\text{(I)}} \cdot \sqrt{\sum_{N_{k,h} \geq M} \frac{1}{\max\{N_{k,h}(s^m_{k,h}, a^m_{k,h}), 1\}}}
\end{aligned}$$

$$+ 4e^3 MSAH^2 + 314e^3 SH^2\iota \underbrace{\sum_{N_{k,h}\geq M} \frac{1}{\max\{N_{k,h}(s^m_{k,h}, a^m_{k,h}), 1\}}}_{\text{(II)}} + 2e^3\sqrt{H^3MK\iota}. \tag{31}$$

where $\sum_{N_{k,h}\geq M}$ denotes the sum over (m, k, h) with $N_{k,h}(s^m_{k,h}, a^m_{k,h}) \geq M$. The second inequality uses $\min\{H, \widetilde{B}\} \leq H$ for the visits with $N_{k,h}(s^m_{k,h}, a^m_{k,h}) < M$; for each (h, s, a), there are fewer than $2M$ such visits, since the counter increases by at most M per episode. The last inequality uses the Cauchy-Schwarz inequality.

- **Bound of term (I):** The term (I) is the sum of the variances of the value functions, with $P = P_{h,s^m_{k,h},a^m_{k,h}}$. Since $0 \leq V^{m,\pi^m_k}_{h+1} \leq V^{m,\star}_{h+1} \leq H$, as in Azar et al. (2017, Lemma 9),

$$\mathsf{Var}_P(V^{m,\star}_{h+1}) - \mathsf{Var}_P(V^{m,\pi^m_k}_{h+1}) \leq P\big((V^{m,\star}_{h+1})^2 - (V^{m,\pi^m_k}_{h+1})^2\big) \leq 2H\, P\Delta^{m,\pi^m_k}_{k,h+1}. \tag{32}$$

  Hence $\text{(I)}^2 \leq \sum_{m,k,h} \mathsf{Var}_P(V^{m,\pi^m_k}_{h+1}) + 2H\sum_{m,k,h} P\Delta^{m,\pi^m_k}_{k,h+1}$, and we bound the two sums in turn.

  - **Bounding** $\sum_{m,k,h} \mathsf{Var}_P(V^{m,\pi^m_k}_{h+1})$: For each (m, k), conditionally on the history before the episode k of agent m, the law of total variance (Azar et al., 2017, Lemma 8) gives $\mathbb{E}[\sum_{h=1}^H \mathsf{Var}_P(V^{m,\pi^m_k}_{h+1})] \leq H^2$; the trajectories of the other agents in episode k do not affect that of agent m given π^m_k. Since $0 \leq \sum_{h=1}^H \mathsf{Var}_P(V^{m,\pi^m_k}_{h+1}) \leq H^3$, the Azuma-Hoeffding inequality over the MK pairs (m, k) gives, with probability at least $1 - \delta/7$,

$$\sum_{m,k,h} \mathsf{Var}_P(V^{m,\pi^m_k}_{h+1}) \leq MKH^2 + H^3\sqrt{2MK\iota}. \tag{33}$$

  - **Bounding** $\sum_{m,k,h} P\Delta^{m,\pi^m_k}_{k,h+1}$: As in Azar et al. (2017, Lemma 7), write $P\Delta^{m,\pi^m_k}_{k,h+1} = \Delta^{m,\pi^m_k}_{k,h+1}(s^m_{k,h+1}) + \xi^m_{k,h}$ and unroll (25) from step $h + 1$:

$$P\Delta^{m,\pi^m_k}_{k,h+1} \leq 2e^3 \sum_{i=1}^H \min\{H, \widetilde{B}^m_{k,i}(s^m_{k,i}, a^m_{k,i})\} + \sum_{i=h}^H \Big(1 + \frac{3}{H}\Big)^{i-h} \xi^m_{k,i}. \tag{34}$$

    Sum over (m, k, h). Each $\xi^m_{k,i}$ then appears with total weight $\sum_{h\leq i}(1 + 3/H)^{i-h} \leq e^3 H$, so the second term is a sum of MKH martingale differences each lying in an interval of length $2e^3H^2$ given the past. By the Azuma-Hoeffding inequality, it is at most $e^3H^2\sqrt{2MKH\iota}$ with probability at least $1 - \delta/7$. For the first term, the same steps as in (31), including the split at $N_{k,h} < M$, and the bound (39) on (II) give

$$\sum_{m,k,h} \min\{H, \widetilde{B}^m_{k,h}(s^m_{k,h}, a^m_{k,h})\} \leq 4\sqrt{2SAH\iota \log K} \cdot \text{(I)} + \mathcal{O}\big(MSAH^2 + S^2AH^3\iota \log K\big), \tag{35}$$

    where $\mathcal{O}(\cdot)$ hides absolute constants only. Therefore,

$$\sum_{m,k,h} P\Delta^{m,\pi^m_k}_{k,h+1} \leq 8e^3 H\sqrt{2SAH\iota \log K} \cdot \text{(I)} + e^3\sqrt{2H^5MK\iota} + \mathcal{O}\big(MSAH^3 + S^2AH^4\iota \log K\big). \tag{36}$$

  - **Combining.** Substitute (33) and (36). The square-root terms in (33) and (36) contribute $\mathcal{O}(MKH^2 + H^5\iota)$ by the AM-GM inequality. Hence

$$\text{(I)}^2 \leq 16e^3H^2\sqrt{2SAH\iota \log K} \cdot \text{(I)} + \mathcal{O}\big(MKH^2 + MSAH^4 + S^2AH^5\iota \log(4K)\big), \tag{37}$$

    and solving this quadratic inequality in (I),

$$\text{(I)} = \mathcal{O}\Big(\sqrt{MKH^2} + \sqrt{MSAH^4} + \sqrt{S^2AH^5\iota^2}\Big). \tag{38}$$

- **Bound of term (II):** The counter $N_{k,h}(s,a)$ is shared by all agents and fixed during episode k, and $N_{k+1,h}(s,a) = N_{k,h}(s,a) + \sum_{m=1}^{M} \mathbb{1}\{(s_{k,h}^m, a_{k,h}^m) = (s,a)\}$. Hence,

$$\begin{aligned}
\text{(II)} &= \sum_{k,h} \sum_{(s,a):\, N_{k,h}(s,a)\geq M} \frac{\sum_{m=1}^{M} \mathbb{1}\{(s_{k,h}^m, a_{k,h}^m) = (s,a)\}}{N_{k,h}(s,a)} \\
&= \sum_{h,s,a} \sum_{k:\, N_{k,h}(s,a)\geq M} \frac{N_{k+1,h}(s,a) - N_{k,h}(s,a)}{N_{k,h}(s,a)} \\
&\overset{(i)}{\leq} 2 \sum_{h,s,a} \sum_{j=M+1}^{MK} \frac{1}{j} \leq 2SAH \log K. \qquad (39)
\end{aligned}$$

In (i), $N_{k,h}(s,a) \geq M$ implies $N_{k+1,h}(s,a) \leq 2N_{k,h}(s,a)$, so each fraction is at most $2\sum_{j=N_{k,h}(s,a)+1}^{N_{k+1,h}(s,a)} 1/j$, and these ranges of j are disjoint over k and lie in $\{M+1, \ldots, MK\}$.

Substituting these bounds on (I) and (II) into (31), with probability at least $1-\delta$,

$$\mathsf{Regret}(K) = \mathcal{O}\Big(\sqrt{H^3SAMK\iota^2} + S^2AH^3\iota^2 + MSAH^2\Big). \qquad (40)$$

D Proof of Theorem 2

The proof proceeds in four major steps, demonstrating how the federated global coverage eliminates the need for action-wise exploration bonuses. Throughout this proof, (h,s,a) ranges over $s \in \mathcal{S}_h$; this is without loss of generality since, for $N_{k,h}(s,a) \geq 1$, $\widehat{P}_{k,h}(\cdot|s,a)$ is supported on observed next states, which lie in $\mathcal{S}_{h+1}$.

Step 1: Global Concentration via Reward Heterogeneity. Under Assumption 1, the heterogeneity of the agents' rewards guarantees that their greedy policies under the estimate $\widehat{P}_k$ used by the algorithm jointly cover the state-action space. Although individual agents act purely greedily, the federated global count $N_{k,h}(s,a)$ grows linearly with the episode k.

Lemma 3 (Global concentration via heterogeneity)**.** *Suppose Assumption 1 holds. Let $N_{k,h}(s,a)$ be the global visitation count for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ at step $h \in [H]$ until episode k. With probability at least $1 - \frac{\delta}{2}$, for all $k \geq 1$, $h \in [H]$, and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$:*

$$N_{k+1,h}(s,a) \geq \frac{1}{2}(k-1)M\lambda_{\mathsf{min}} - \ln\frac{2SAH}{\delta}. \qquad (41)$$

Consequently, with the burn-in threshold $K_{th} = \lceil 4\log(4M^2SAKH/\delta)/(M\lambda_{\mathsf{min}}) \rceil + 2$, for all $k \geq K_{th}$ the global visitation count is lower bounded by a linear rate: $N_{k+1,h}(s,a) \geq \frac{1}{4}kM\lambda_{\mathsf{min}}$.

This implies that the environment is thoroughly explored by the collective behaviors of the heterogeneous agents, completely bypassing the need for forced exploration.

Step 2: Shrinking Bellman Error. With the linearly growing global visitation counts for $k > K_{th}$, the empirical transition dynamics $\widehat{P}_{k,h}$ rapidly converge to the true dynamics P_h as follows:

Lemma 4 (Shrinking Bellman Error)**.** *Suppose the result of Lemma 3 holds. Let $Q_{k,h}^m(s,a)$ be the action-value function computed by agent m at episode k using the empirical transition $\widehat{P}_{k,h}$, and let $Q_h^{m,\star}(s,a)$ be the true optimal action-value function. With probability at least $1-\delta$, for all $k > K_{th}$, $m \in [M]$, $h \in [H]$, and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$:*

$$\left|Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a)\right| \leq \mathcal{O}\left(\sqrt{\frac{H^4\log(4KMHSA/\delta)}{M\lambda_{\mathsf{min}}k}} + \frac{H^2\log(4KMHSA/\delta)}{M\lambda_{\mathsf{min}}k}\right) \qquad (42)$$

Step 3: Burn-in Costs. Because PF-EFVI does not inject any UCB bonuses, an agent m will strictly choose the optimal action $\pi_h^{m,\star}(s)$ if its Q-value estimation error is strictly less than half of the minimum suboptimality gap, i.e., $\left|Q_{k,h}^m - Q_h^{m,\star}\right| < \Delta_{\min}^m/2$. We derive the critical burn-in episode K_0^m for agent m as follows:

Lemma 5 (The burn-in cost of pure exploitation)**.** *There exists a critical burn-in episode K_0^m such that for all episodes $k > \max(K_{th}, K_0^m)$, the purely greedy policy $\pi_{k,h}^m(s) = \arg\max_a Q_{k,h}^m(s,a)$ selects an optimal action at every state, yielding zero instantaneous regret. This tipping point is bounded by:*

$$K_0^m = \mathcal{O}\left(\frac{H^4 \log(KMHSA/\delta)}{M\lambda_{\min}(\Delta_{\min}^m)^2}\right) \tag{43}$$

Step 4: Summing the Regret. To compute the total regret for agent m, we split the learning process into the initial uncontrolled phase ($k \le K_{th}$) and the shrinking phase ($K_{th} < k \le K_0^m$). If a suboptimal action is chosen, its gap is bounded by $2\epsilon_k$. Since the instantaneous regret is strictly zero for all $k > \max(K_{th}, K_0^m)$ (by Lemma 5), we sum the regret across these two phases.

During the uncontrolled phase ($P1$), the instantaneous regret is trivially bounded by the horizon H. Thus:

$$\mathsf{Regret}_{P1}^m(K) \le \sum_{k=1}^{K_{th}} H = H \cdot K_{th} \le \frac{4H\log(4M^2SAKH/\delta)}{M\lambda_{\min}} + 3H \tag{44}$$

During the shrinking phase ($P2$), we integrate the bounded error ϵ_k:

$$\begin{aligned}
\mathsf{Regret}_{P2}^m(K) &= \sum_{k=K_{th}+1}^{K_0^m} \mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right] \\
&\overset{(i)}{=} \sum_{k=K_{th}+1}^{K_0^m} \mathbb{E}_{\pi_k^m, s_1\sim\rho}\left[\sum_{h=1}^{H}\left(Q_h^{m,\star}(s_h,\pi_h^{m,\star}(s_h)) - Q_h^{m,\star}(s_h,\pi_{k,h}^m(s_h))\right)\right] \\
&\le \sum_{k=1}^{K_0^m} \mathcal{O}\left(\sqrt{\frac{H^6\log(KMHSA/\delta)}{kM\lambda_{\min}}}\right) \\
&\le \mathcal{O}\left(\sqrt{\frac{H^6\log(KMHSA/\delta)}{M\lambda_{\min}}}\sum_{k=1}^{K_0^m}\frac{1}{\sqrt{k}}\right) \\
&= \mathcal{O}\left(\sqrt{\frac{H^6\log(KMHSA/\delta)}{M\lambda_{\min}}}\cdot 2\sqrt{K_0^m}\right)
\end{aligned} \tag{45}$$

where (i) follows from the Performance Difference Lemma. Since the estimated Q-value error at any step is bounded by ϵ_k, the suboptimality gap of the chosen greedy action at each step is at most $2\epsilon_k$. Summing this over the H steps of the trajectory yields an instantaneous regret bound of $2H\epsilon_k$.

We substitute the explicit value of $\sqrt{K_0^m} = \mathcal{O}\left(\frac{H^2\sqrt{\log(KMHSA/\delta)}}{\sqrt{M\lambda_{\min}}\Delta_{\min}^m}\right)$ into Equation (45):

$$\begin{aligned}
\mathsf{Regret}_{P2}^m(K) &\le \mathcal{O}\left(\sqrt{\frac{H^6\log(KMHSA/\delta)}{M\lambda_{\min}}}\cdot\frac{H^2\sqrt{\log(KMHSA/\delta)}}{\sqrt{M\lambda_{\min}}\Delta_{\min}^m}\right) \\
&= \mathcal{O}\left(\frac{H^5\log(KMHSA/\delta)}{M\lambda_{\min}\Delta_{\min}^m}\right)
\end{aligned} \tag{46}$$

Summing these independent bounds over all M agents yields the final total system regret:

$$\mathsf{Regret}(K) = \sum_{m=1}^{M}\left(\mathsf{Regret}_{P1}^m(K) + \mathsf{Regret}_{P2}^m(K)\right)$$

$$\leq \sum_{m=1}^{M} \left(\frac{4H \log(4M^2 SAKH/\delta)}{M\lambda_{\mathsf{min}}} + 3H \right) + \sum_{m=1}^{M} \mathcal{O}\left(\frac{H^5 \log(KMHSA/\delta)}{M\lambda_{\mathsf{min}} \Delta^m_{\min}} \right)$$
$$= \mathcal{O}\left(\frac{H \log(4SAKMH/\delta)}{\lambda_{\mathsf{min}}} + MH \right) + \mathcal{O}\left(\frac{H^5 \log(KMHSA/\delta)}{M\lambda_{\mathsf{min}}} \sum_{m=1}^{M} \frac{1}{\Delta^m_{\min}} \right). \tag{47}$$

E A visitation-weighted form of Theorem 2

In Theorem 2, the coverage enters only through λ_{min}, that is, through the worst-covered state-action pair. We show that the leading, gap-dependent term only needs coverage along the trajectories that matter for each agent.

Proof of Proposition 1. We work on the events of Lemmas 3 and 4 and of (53) below, which hold simultaneously with probability at least $1-\delta$. We write $\lambda_{h'}$ for $\lambda_{h'}(s_{h'}, a_{h'})$ inside trajectory expectations.

Step 1: where λ_{min} enters. In Step 4 of the proof of Theorem 2, the regret of each episode $k \in (K_{th}, K_0^m]$ is bounded by $2H\epsilon_k$, where $\epsilon_k = \max_{h,s,a} |Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a)|$ is governed by the worst-covered pair through Lemma 4. We keep the rest of Step 4 and replace only this bound.

Step 2: regret decomposition. Fix m and $k > K_{th}$. By step (i) of (45),

$$\mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right] = \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{\pi_k^m}}\left[\Delta_h^m(s,a)\right]. \tag{48}$$

If $d_h^{\pi_k^m}(s,a) > 0$, then $\pi^m_{k,h}(a|s) > 0$, so $a \in \arg\max_{a'} Q^m_{k,h}(s,a')$ and, with $a^\star = \pi_h^{m,\star}(s)$,

$$\Delta_h^m(s,a) \leq \left[Q_h^{m,\star} - Q^m_{k,h}\right](s,a^\star) + \left[Q^m_{k,h} - Q_h^{m,\star}\right](s,a). \tag{49}$$

Thus only the underestimation at the optimal action and the overestimation at the played action matter. Lemma 4 bounds both by ϵ_k.

Step 3: one-sided errors. Lemma 6 below bounds the overestimation along π_k^m and the underestimation along $\pi^{m,\star}$, with the local error at each pair governed by its own coverage. Applying Lemma 6 to (49), and using $\mathbb{E}^{\pi^{m,\star}}[\,\cdot \mid s_h = s, a_h = a^\star] = \mathbb{E}^{\pi^{m,\star}}[\,\cdot \mid s_h = s]$,

$$\Delta_h^m(s,a) \leq CH\sqrt{\frac{\iota}{kM}}\Big(\mathbb{E}^{\pi_k^m}\Big[\sum_{h'=h}^{H} \lambda_{h'}^{-1/2} \,\Big|\, s_h = s, a_h = a\Big]$$
$$+ \mathbb{E}^{\pi^{m,\star}}\Big[\sum_{h'=h}^{H} \lambda_{h'}^{-1/2} \,\Big|\, s_h = s\Big]\Big) + \frac{CH^3\sqrt{S}\,\iota}{kM\lambda_{\mathsf{min}}}. \tag{50}$$

Step 4: from pairs to Λ^m. If the first term of (50) is at least the second, then $\Delta_h^m(s,a)$ is at most twice the first term, so $\Delta_h^m(s,a) \leq (2\cdot\text{first term})^2/\Delta_h^m(s,a)$. We then use $(x+y)^2 \leq 2x^2+2y^2$, the bound $(\mathbb{E}^\pi[\sum_{h'=h}^H \lambda_{h'}^{-1/2}])^2 \leq H\,\mathbb{E}^\pi[\sum_{h'=h}^H \lambda_{h'}^{-1}]$ (Jensen and Cauchy–Schwarz), and $\Delta_h^m(s,a) \geq \Delta^m_{\min}$. Otherwise, $\Delta_h^m(s,a)$ is at most twice the second term. In both cases, for every (s,a) with $\pi^m_{k,h}(a|s) > 0$ and $\Delta_h^m(s,a) > 0$,

$$\Delta_h^m(s,a) \leq \frac{CH^3\iota}{kM\Delta^m_{\min}}\Big(\mathbb{E}^{\pi_k^m}\Big[\sum_{h'=h}^{H} \frac{1}{\lambda_{h'}} \,\Big|\, s_h = s, a_h = a\Big]$$
$$+ \mathbb{E}^{\pi^{m,\star}}\Big[\sum_{h'=h}^{H} \frac{1}{\lambda_{h'}} \,\Big|\, s_h = s\Big]\Big) + \frac{CH^3\sqrt{S}\,\iota}{kM\lambda_{\mathsf{min}}}. \tag{51}$$

We substitute this into (48). Pairs with $\Delta_h^m(s,a) = 0$ contribute nothing, and the right-hand side is nonnegative, so the restriction $\Delta_h^m(s,a) > 0$ can be dropped. The second expectation depends on (s,a) only through s. For the first, if $(s_h, a_h) \sim d_h^{\pi_k^m}$ and the trajectory continues with π_k^m, then $(s_{h'}, a_{h'}) \sim d_{h'}^{\pi_k^m}$ by the Markov property, so

$$\sum_{h=1}^{H}\sum_{h'=h}^{H}\mathbb{E}_{(s,a)\sim d_{h'}^{\pi_k^m}}\Big[\frac{1}{\lambda_{h'}(s,a)}\Big] = \sum_{h'=1}^{H} h'\,\mathbb{E}_{(s,a)\sim d_{h'}^{\pi_k^m}}\Big[\frac{1}{\lambda_{h'}(s,a)}\Big] \le H\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{\pi_k^m}}\Big[\frac{1}{\lambda_h(s,a)}\Big].$$

Hence the sum over h of the bracket in (51) is at most H^2 times the braces of (8) at $\widetilde{P} = \widehat{P}_k$. This is at most $H^2\Lambda^m$, because π_k^m is the uniformly tie-broken greedy policy for $\widehat{P}_k \in \mathcal{P}$. Therefore

$$\mathbb{E}_{s_1\sim\rho}\big[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\big] \le \frac{CH^5\iota\,\Lambda^m}{kM\Delta_{\min}^m} + \frac{CH^4\sqrt{S}\,\iota}{kM\lambda_{\mathsf{min}}} \qquad (k > K_{th}). \tag{52}$$

Step 5: summation. We split the episodes as in Step 4. Phase P1 is bounded by (44), hence by $4H\iota/(M\lambda_{\mathsf{min}}) + 3H$ since $\log(4M^2SAKH/\delta) \le \iota$, and episodes $k > \max\{K_{th}, K_0^m\}$ contribute nothing by Lemma 5, since K_0^m is at least its threshold for C_0 large enough. Phase P2 is where the argument departs: we sum (52) instead of $2H\epsilon_k$, and $\sum_{k=K_{th}+1}^{K_0^m}\frac{1}{k} \le \log(1+K_0^m)$ replaces $\sum_k k^{-1/2} \le 2\sqrt{K_0^m}$. This gives

$$\mathsf{Regret}_{P2}^m(K) \le C\log(1+K_0^m)\left[\frac{H^5\iota\,\Lambda^m}{M\Delta_{\min}^m} + \frac{H^4\sqrt{S}\,\iota}{M\lambda_{\mathsf{min}}}\right].$$

Summing over $m \in [M]$ and using $K_0^m \le K_0$ gives (9). □

Besides the events of Lemmas 3 and 4, we use the ℓ_1 deviation inequality of Weissman et al. (2003): for n i.i.d. samples from a distribution P on S points, $\mathbb{P}(\|\widehat{P}_n - P\|_1 \ge \varepsilon) \le (2^S - 2)e^{-n\varepsilon^2/2}$. We apply it to the first n next states observed at each (h,s,a), as in the proof of Lemma 4, with $\varepsilon = \sqrt{2(S\ln 2 + \iota)/n}$, and take a union bound over (h,s,a) and $n \in [MK]$. With probability at least $1-\delta/4$, for all k, h and (s,a) with $N_{k,h}(s,a) \ge 1$,

$$\big\|\widehat{P}_{k,h}(\cdot|s,a) - P_h(\cdot|s,a)\big\|_1 \le \sqrt{\frac{2(S\ln 2+\iota)}{N_{k,h}(s,a)}}. \tag{53}$$

Taking the Bernstein event of Lemma 4 at level $\delta/4$ instead of $\delta/2$ changes only the constant inside its logarithm. Hence all the events used here hold simultaneously with probability at least $1-\delta$.

Lemma 6 (One-sided per-pair estimation error)**.** *On the events of Lemmas 3 and 4 and of* (53)*, for all $k > K_{th}$, $m \in [M]$, $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,*

$$Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a) \le CH\sqrt{\frac{\iota}{kM}}\,\mathbb{E}^{\pi_k^m}\Big[\sum_{h'=h}^{H}\frac{1}{\sqrt{\lambda_{h'}(s_{h'},a_{h'})}}\,\Big|\,s_h = s, a_h = a\Big] + \frac{CH^3\sqrt{S}\,\iota}{kM\lambda_{\mathsf{min}}}, \tag{54}$$

$$Q_h^{m,\star}(s,a) - Q_{k,h}^m(s,a) \le CH\sqrt{\frac{\iota}{kM}}\,\mathbb{E}^{\pi^{m,\star}}\Big[\sum_{h'=h}^{H}\frac{1}{\sqrt{\lambda_{h'}(s_{h'},a_{h'})}}\,\Big|\,s_h = s, a_h = a\Big] + \frac{CH^3\sqrt{S}\,\iota}{kM\lambda_{\mathsf{min}}}. \tag{55}$$

F Proof of Lemmas

F.1 Proof of Lemma 1

By applying the empirical Bernstein inequality (Maurer and Pontil, 2009) with a union bound over (h,s,a,m) and $N_{k,h}(s,a) \in [MK]$, we can show the upper bound as follows for $N_{k,h}(s,a) \ge 2$, using $1/(N_{k,h}(s,a)-1) \le 2/N_{k,h}(s,a)$ for its sample-variance normalization and second-order term; for $N_{k,h}(s,a) \le 1$, the lemma holds trivially since $B_{k,h}^m(s,a) \ge H$:

$$\big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V_{k,h+1}^m = \big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V_{h+1}^{m,\star} + \big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)\big(V_{k,h+1}^m - V_{h+1}^{m,\star}\big)$$

$$\leq \sqrt{\frac{2\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V_{h+1}^{m,\star})\iota}{\max\{N_{k,h}(s,a),1\}}} + \frac{8H\iota}{\max\{N_{k,h}(s,a),1\}} + \left|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_{k,h+1}^{m}\right|. \tag{56}$$

Bounding $|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_{k,h+1}^{m}|$: To bound the last term $|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_{k,h+1}^{m}|$, we need to consider all possible $\Delta_{k,h+1}^{m}$ and apply the Bernstein inequality. For any fixed Δ_ϵ that approximates $\Delta_{k,h+1}^{m}$ by the error of ϵ, by applying the Bernstein inequality and using the fact that $0 \leq \Delta_{k,h+1}^{m} \leq H$, we can show that

$$\begin{aligned}|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_\epsilon| &\leq \min\left\{H, \sqrt{\frac{2\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon)\ln(2/\delta')}{N_{k,h}(s,a)}} + \frac{2H\log(2/\delta')}{3N_{k,h}(s,a)}\right\} \\ &\leq \sqrt{\frac{2\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon)\ln(2/\delta')}{\max\{N_{k,h}(s,a),1\}}} + \frac{2H\log(2/\delta')}{3\max\{N_{k,h}(s,a),1\}}.\end{aligned} \tag{57}$$

To cover all possible $\Delta_{k,h+1}^{m}$ with an error of ϵ, we need to consider all possible Δ_ϵ, splitting $[0,H]$ into $\lceil H/\epsilon \rceil$ intervals for all states, incurring covering number $\mathcal{N}(\epsilon) \leq \left(1+\frac{H}{\epsilon}\right)^S$. By taking a union bound over $\mathcal{N}(\epsilon)$, (h,s,a,m), and possible visitation counter $N_{k,h}(s,a) = N \in [MK]$, we can replace δ' with $\delta/(7M^2SAKH\mathcal{N}(\epsilon))$ in the above inequality, yielding

$$\begin{aligned}
&|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_{k,h+1}^{m}| \\
&\leq |(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_\epsilon| + |(\widehat{P}_{k,h,s,a} - P_{h,s,a})(\Delta_{k,h+1}^{m} - \Delta_\epsilon)| \\
&\leq |(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_\epsilon| + 2\epsilon \\
&\leq \sqrt{\frac{2\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon)\log(14M^2SAKH\mathcal{N}(\epsilon)/\delta)}{\max\{N_{k,h}(s,a),1\}}} + \frac{2H\log(14M^2SAKH\mathcal{N}(\epsilon)/\delta)}{3\max\{N_{k,h}(s,a),1\}} + 2\epsilon \\
&\overset{(i)}{\leq} \sqrt{\frac{2H(P\Delta_{k,h+1}^{m} + \epsilon)\log(14M^2SAKH\mathcal{N}(\epsilon)/\delta)}{\max\{N_{k,h}(s,a),1\}}} + \frac{2H\log(14M^2SAKH\mathcal{N}(\epsilon)/\delta)}{3\max\{N_{k,h}(s,a),1\}} + 2\epsilon \\
&\overset{(ii)}{\leq} \sqrt{\frac{8SHP\Delta_{k,h+1}^{m}\iota}{\max\{N_{k,h}(s,a),1\}}} + \frac{7HS\iota}{\max\{N_{k,h}(s,a),1\}} \\
&\overset{(iii)}{\leq} \frac{1}{H}P_{h,s,a}\Delta_{k,h+1}^{m} + \frac{9SH^2\iota}{\max\{N_{k,h}(s,a),1\}},
\end{aligned} \tag{58}$$

where (i) holds because

$$\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon) = P_{h,s,a}\Delta_\epsilon^2 - (P_{h,s,a}\Delta_\epsilon)^2 \leq P_{h,s,a}\Delta_\epsilon^2 \leq HP_{h,s,a}\Delta_\epsilon \leq HP_{h,s,a}\Delta_{k,h+1}^{m} + H\epsilon, \tag{59}$$

(ii) holds when choosing very small $\epsilon = 1/(HMK)^2$, since then $\log(14M^2SAKH\mathcal{N}(\epsilon)/\delta) \leq \iota + S\log(1+H^3M^2K^2) \leq 4S\iota$, and (iii) follows by applying the Young's inequality that $xy \leq cx^2 + \frac{y^2}{4c}$ for any $x, y > 0$ and $c > 0$, which yields

$$\sqrt{\frac{8SHP_{h,s,a}\Delta_{k,h+1}^{m}\iota}{\max\{N_{k,h}(s,a),1\}}} \leq \frac{1}{H}P_{h,s,a}\Delta_{k,h+1}^{m} + \frac{2SH^2\iota}{\max\{N_{k,h}(s,a),1\}} \tag{60}$$

when $c = 1/H$, $x = \sqrt{P\Delta_{k,h+1}^{m}}$, $y = \sqrt{\frac{8SH\iota}{\max\{N_{k,h}(s,a),1\}}}$.

By substituting the above bound into (56), we can show that

$$\left(\widehat{P}_{k,h,s,a} - P_{h,s,a}\right)V_{k,h+1}^{m} \leq \sqrt{\frac{2\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V_{h+1}^{m,\star})\iota}{\max\{N_{k,h}(s,a),1\}}} + \frac{17SH^2\iota}{\max\{N_{k,h}(s,a),1\}} + \frac{1}{H}P_{h,s,a}\Delta_{k,h+1}^{m}. \tag{61}$$

Since $V_{h+1}^{m,\star} = V_{k,h+1}^m - \Delta_{k,h+1}^m$ and $0 \le \Delta \le H$,

$$\mathsf{Var}_{\widehat{P}}(V_{h+1}^{m,\star}) \le 2\mathsf{Var}_{\widehat{P}}(V_{k,h+1}^m) + 2\mathsf{Var}_{\widehat{P}}(\Delta_{k,h+1}^m) \le 2\mathsf{Var}_{\widehat{P}}(V_{k,h+1}^m) + 2H\,\widehat{P}\Delta_{k,h+1}^m,$$

hence

$$\sqrt{\frac{2\mathsf{Var}_{\widehat{P}}(V_{h+1}^{m,\star})\iota}{N}} \le \sqrt{\frac{4\mathsf{Var}_{\widehat{P}}(V_{k,h+1}^m)\iota}{N}} + \sqrt{\frac{4H\iota\,\widehat{P}\Delta_{k,h+1}^m}{N}}.$$

Using $\widehat{P}\Delta \le 2P\Delta + 9SH^2\iota/N$ (from the bound on $|(\widehat{P}-P)\Delta|$ proved above) and $2\sqrt{xy} \le x+y$,

$$\sqrt{\frac{4H\iota\,\widehat{P}\Delta}{N}} \le \sqrt{\frac{8H\iota\,P\Delta}{N}} + \frac{6\sqrt{S}\,H^{3/2}\iota}{N} \le \frac{1}{H}P\Delta + \frac{8SH^2\iota}{N}.$$

Then, we can rewrite (61) as

$$\big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V_{k,h+1}^m \le \underbrace{\sqrt{\frac{4\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V_{k,h+1}^m)\iota}{\max\{N_{k,h}(s,a),1\}}} + \frac{25SH^2\iota}{\max\{N_{k,h}(s,a),1\}}}_{=B_{k,h}^m(s,a)} + \frac{2}{H}P_{h,s,a}\Delta_{k,h+1}^m. \tag{62}$$

Similarly, we can also show the lower bound as follows:

$$\big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V_{k,h+1}^m \ge -B_{k,h}^m(s,a) - \frac{2}{H}P_{h,s,a}\Delta_{k,h+1}^m. \tag{63}$$

F.2 Proof of Lemma 2

For notational simplicity, we omit the subscripts and superscripts in $\widehat{P}_{k,h,s,a}$ and $P_{h,s,a}$, and simply denote them as $\widehat{P}$ and P respectively. We also denote $V = V_{h+1}^{m,\star}$ for brevity. First, we decompose the variance difference as follows:

$$\begin{aligned}
\big|\mathsf{Var}_{\widehat{P}}(V) - \mathsf{Var}_P(V)\big| &= \big|(\widehat{P}V^2 - (\widehat{P}V)^2) - (PV^2 - (PV)^2)\big| \\
&= \big|(\widehat{P}V^2 - PV^2) + ((PV)^2 - (\widehat{P}V)^2)\big| \\
&\le \big|(\widehat{P}-P)V^2\big| + \big|(P+\widehat{P})V \cdot (\widehat{P}-P)V\big| \\
&\le \big|(\widehat{P}-P)V^2\big| + 2H\big|(\widehat{P}-P)V\big|
\end{aligned} \tag{64}$$

Then, applying the Bernstein inequality to both terms, we can show that

$$\begin{aligned}
\big|(\widehat{P}-P)V^2\big| &\le \sqrt{\frac{2\mathsf{Var}_P(V^2)\iota}{N}} + \frac{2H^2\iota}{3N} \quad (V^2 \in [0,H^2]) \\
&\le \sqrt{\frac{8H^2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H^2\iota}{3N}, \qquad (65) \\
\big|(\widehat{P}-P)V\big| &\le \sqrt{\frac{2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H\iota}{3N}, \qquad (66)
\end{aligned}$$

using the fact that

$$\begin{aligned}
\mathsf{Var}_P(V^2) &= \frac{1}{2}\mathbb{E}_{x,y}\left[(V(x)^2 - V(y)^2)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{x,y}\left[(V(x)-V(y))^2(V(x)+V(y))^2\right] \le 4H^2\mathsf{Var}_P(V).
\end{aligned} \tag{67}$$

Combining the above bounds, we have:

$$\big|\mathsf{Var}_{\widehat{P}}(V) - \mathsf{Var}_P(V)\big| \le \min\left\{H^2, \sqrt{\frac{32H^2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H^2\iota}{N}\right\}. \tag{68}$$

Applying the Young's inequality,

$$\sqrt{\frac{32H^2\mathsf{Var}_P(V)\iota}{N}} = 2\cdot\sqrt{\mathsf{Var}_P(V)}\cdot\sqrt{\frac{8H^2\iota}{N}} \le \mathsf{Var}_P(V) + \frac{8H^2\iota}{N} \tag{69}$$

Thus, we have

$$\left|\mathsf{Var}_{\widehat{P}}(V) - \mathsf{Var}_P(V)\right| \le \mathsf{Var}_P(V) + \frac{10H^2\iota}{\max\{N,1\}}. \tag{70}$$

F.3 Proof of Lemma 3

We clarify the relationship between the estimated dynamics $\widehat{P}_{\tau,h}$ and the coverage guarantee. While each agent m computes its greedy policy π_τ^m using the estimated transition kernel $\widehat{P}_{\tau,h} \in \mathcal{F}_{\tau-1}$, the marginal visitation probability $d_h^{\pi_\tau^m}(s,a)$ is always evaluated under the *true* transition kernel P_h, not under $\widehat{P}_{\tau,h}$. Since $\widehat{P}_{\tau,h}(\cdot|s,a)$ belongs to the set of all valid transition kernels $\mathcal{P}$, Assumption 1 applies directly with $\widetilde{P} = \widehat{P}_{\tau,h}$: regardless of how inaccurate $\widehat{P}_{\tau,h}$ may be relative to the true dynamics P_h, the diversity of the agents' reward functions $\{r^m\}_{m=1}^M$ structurally ensures that the resulting greedy policies $\{\pi_\tau^m\}_{m=1}^M$ collectively cover the state-action space under the true environment, yielding $\sum_{m=1}^M d_h^{\pi_\tau^m}(s,a) \ge M\lambda_{\mathsf{min}}$ for every episode τ. Here π_τ^m is the greedy policy with uniform tie-breaking, which is exactly the policy $\widetilde{\pi}^m$ of Assumption 1 for $\widetilde{P} = \widehat{P}_{\tau,h}$.

Let $\mathcal{F}_0$ be the trivial σ-field and, for $\tau \ge 1$, let

$$\mathcal{F}_\tau = \sigma\Big(\big\{s_{i,h}^m : m\in[M],\, i\le\tau,\, h\in[H+1]\big\} \cup \big\{a_{i,h}^m : m\in[M],\, i\le\tau,\, h\in[H]\big\}\Big)$$

be the σ-field generated by the states and actions of all agents in episodes $1,\dots,\tau$. At the beginning of episode τ, the central server broadcasts the empirical transition kernel $\widehat{P}_{\tau,h}$, which is fully determined by $\mathcal{F}_{\tau-1}$. Conditioned on the filtration $\mathcal{F}_{\tau-1}$, each agent $m\in[M]$ computes its purely greedy policy $\pi_{\tau,h}^m$ based on its local reward r_h^m and the shared estimated dynamics $\widehat{P}_{\tau,h}$. Let $d_h^{\pi_\tau^m}(s,a)$ be the exact marginal probability of visiting (s,a) at step h under this executed policy π_τ^m.

Fix (h,s,a) and order the indicators $\mathbb{I}(s_{\tau,h}^m = s, a_{\tau,h}^m = a)$ lexicographically in (τ,m), refining $(\mathcal{F}_\tau)$ by revealing the agents' episode-τ trajectories one at a time. Given $\mathcal{F}_{\tau-1}$, the episode-τ trajectories of the agents are independent, since the agents interact with independent copies of the environment, and agent m's trajectory follows π_τ^m in P from $s_1\sim\rho$. Hence, given everything preceding position (τ,m) in this order, the indicator at (τ,m) equals one with probability $d_h^{\pi_\tau^m}(s,a)$, which is $\mathcal{F}_{\tau-1}$-measurable since π_τ^m is determined by r^m and $\widehat{P}_\tau$. The lower-tail bound for adapted Bernoulli sequences of Dann et al. (2017, Lemma F.4) then gives, with probability at least $1-\delta'$, simultaneously for all $k\ge 1$,

$$N_{k+1,h}(s,a) \ge \frac{1}{2}\sum_{\tau=1}^{k}\sum_{m=1}^{M} d_h^{\pi_\tau^m}(s,a) - \ln\frac{1}{\delta'}. \tag{71}$$

We take a union bound over (h,s,a) with $\delta' = \delta/(2SAH)$.

As shown above, Assumption 1 applies with $\widetilde{P} = \widehat{P}_{\tau,h} \in \mathcal{P}$. Thus, for every episode $\tau \ge 2$ (in episode 1 all actions tie under the zero initialization, so π_1^m is uniform and we only use $d_h^{\pi_1^m} \ge 0$):

$$\sum_{m=1}^{M} d_h^{\pi_\tau^m}(s,a) \ge M\lambda_{\mathsf{min}}. \tag{72}$$

Substituting this lower bound into the concentration inequality above, we obtain, for all $k\ge 1$, $h\in[H]$ and (s,a),

$$N_{k+1,h}(s,a) \ge \frac{1}{2}(k-1)M\lambda_{\mathsf{min}} - \ln\frac{2SAH}{\delta}. \tag{73}$$

For $k \ge K_{th}$, we have $\ln(2SAH/\delta) \le \log(4M^2SAKH/\delta) \le \frac{1}{4}(k-2)M\lambda_{\mathsf{min}}$, and hence $N_{k+1,h}(s,a) \ge \frac{1}{2}(k-1)M\lambda_{\mathsf{min}} - \frac{1}{4}(k-2)M\lambda_{\mathsf{min}} \ge \frac{1}{4}kM\lambda_{\mathsf{min}}$. This completes the proof.

F.4 Proof of Lemma 4

We bound the absolute estimation error of the Q-value by backward induction on h. For the base case $h = H + 1$, the error is 0.

For any step $h \in [H]$, taking the absolute difference between the empirical Q-value update and the true optimal Bellman equation, we apply Hölder's inequality:

$$\left|Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a)\right| \le \left|\langle \widehat{P}_{k,h}(s,a) - P_h(s,a), V^{m,\star}_{h+1}\rangle\right| + \left|\langle \widehat{P}_{k,h}(s,a), V^m_{k,h+1} - V^{m,\star}_{h+1}\rangle\right| \tag{74}$$

Let $s_{1,h+1}, s_{2,h+1}, \ldots \overset{\text{i.i.d.}}{\sim} P_h(\cdot|s,a)$ be the successive next states observed at (h,s,a), so that $\langle \widehat{P}_{k,h}(s,a), V\rangle$ equals $\frac{1}{n}\sum_{i\le n} V(s_{i,h+1})$ on $\{N_{k,h}(s,a) = n\}$. For each fixed $n \in [MK]$, Bernstein's inequality gives, with probability at least $1-\delta'$,

$$\left|\langle \widehat{P}_{k,h} - P_h, V^{m,\star}_{h+1}\rangle\right| \le \sqrt{\frac{2\mathsf{Var}_{P_h}(V^{m,\star}_{h+1})\log(\frac{2}{\delta'})}{N_{k,h}(s,a)}} + \frac{2\|V^{m,\star}_{h+1}\|_\infty \log(\frac{2}{\delta'})}{3N_{k,h}(s,a)} \tag{75}$$

For $k > K_{th}$, Lemma 3 at $k-1 \ge K_{th}$ gives $N_{k,h}(s,a) \ge \frac{1}{4}(k-1)M\lambda_{\min} \ge \frac{1}{8}kM\lambda_{\min}$, and $N_{k,h}(s,a) \ge 1$ since $\frac{1}{4}K_{th}M\lambda_{\min} \ge \log(4M^2SAKH/\delta) > 1$. Substituting this lower bound and $\|V^{m,\star}_{h+1}\|_\infty \le H$, we have:

$$\langle \widehat{P}_{k,h} - P_h, V^{m,\star}_{h+1}\rangle \le \mathcal{O}\left(\sqrt{\frac{\mathsf{Var}_{P_h}(V^{m,\star}_{h+1})\log(\frac{2}{\delta'})}{M\lambda_{\min}k}} + \frac{H\log(\frac{2}{\delta'})}{M\lambda_{\min}k}\right) \tag{76}$$

The second term can be recursively unpacked as the maximum estimation error of the value function at step $h+1$. As the errors propagate backward until reaching the last step H, the maximum error accumulated over H steps is bounded by:

$$\begin{aligned}\max_{s,a}\left|Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a)\right| &\le \mathcal{O}\left(\sqrt{\frac{\log(\frac{2}{\delta'})}{M\lambda_{\min}k}}\sum_{h'=h}^{H}\sqrt{\mathsf{Var}_{P_{h'}}(V^{m,\star}_{h'+1})} + \frac{H^2\log(\frac{2}{\delta'})}{M\lambda_{\min}k}\right)\\ &\le \mathcal{O}\left(\sqrt{\frac{\log(\frac{2}{\delta'})}{M\lambda_{\min}k}}H^2 + \frac{H^2\log(\frac{2}{\delta'})}{M\lambda_{\min}k}\right)\end{aligned} \tag{77}$$

Here, we use the fact that the variance of the optimal value function is bounded by H^2, implying $\sqrt{\mathsf{Var}_{P_{h'}}(V^{m,\star}_{h'+1})} \le H$. Summing this worst-case bound over H steps yields H^2, which is then absorbed into the square root as H^4.

We require this bound to hold uniformly for all $k \in [K]$, $m \in [M]$, $h \in [H]$, $s \in \mathcal{S}_h$, and $a \in \mathcal{A}$. Taking a union bound over (m,h,s,a) and $n \in [MK]$, which covers every $k \in [K]$, and allocating the remaining failure probability $\delta/2$, we set $\delta' = \delta/(2M^2KHSA)$, so that $\log(2/\delta') = \log(4M^2KHSA/\delta) \le 2\log(4KMHSA/\delta)$. Then,

$$\max_{s,a}\left|Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a)\right| \le \mathcal{O}\left(\sqrt{\frac{H^4\log(4KMHSA/\delta)}{M\lambda_{\min}k}} + \frac{H^2\log(4KMHSA/\delta)}{M\lambda_{\min}k}\right) \tag{78}$$

This completes the proof.

F.5 Proof of Lemma 5

Let $\epsilon_k = \max_{s,a,h}|Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a)|$ be the maximum absolute estimation error at episode k. By Lemma 4, we have $\epsilon_k \le \mathcal{O}\left(\sqrt{\frac{H^4\log(KMHSA/\delta)}{kM\lambda_{\min}}}\right)$. Here the second-order term of Lemma 4 is

dropped because that term is at most the first-order term if $kM\lambda_{\min} \geq \log(4M^2KHSA/\delta)$, which holds for $k > K_{th}$ since $kM\lambda_{\min} > K_{th}M\lambda_{\min} \geq 4\log(4M^2SAKH/\delta)$.

Because the agent acts greedily without any UCB bonuses, it will strictly prefer the optimal action $a^\star$ over any suboptimal action a' if:

$$Q^m_{k,h}(s,a^\star) > Q^m_{k,h}(s,a') \tag{79}$$

Using the absolute error bound, the estimated values satisfy $Q^m_{k,h}(s,a^\star) \geq Q^{m,\star}_h(s,a^\star) - \epsilon_k$ and $Q^m_{k,h}(s,a') \leq Q^{m,\star}_h(s,a') + \epsilon_k$. Therefore, the optimal action is guaranteed to be chosen if:

$$Q^{m,\star}_h(s,a^\star) - \epsilon_k > Q^{m,\star}_h(s,a') + \epsilon_k \implies 2\epsilon_k < Q^{m,\star}_h(s,a^\star) - Q^{m,\star}_h(s,a') = \Delta^m_h(s,a') \tag{80}$$

To ensure this holds for all suboptimal actions, we require $2\epsilon_k < \Delta^m_{\min}$. Substituting the upper bound of ϵ_k and solving for k, we find that the condition holds for all $k > K^m_0$. For $k > \max(K_{th}, K^m_0)$, no suboptimal actions are ever chosen.

F.6 Proof of Lemma 6

Fix $k > K_{th}$ and m. The proof of Lemma 3 applies to each fixed (h,s,a). Its coverage step $\frac{1}{M}\sum_m d^{\pi^m_\tau}_h(s,a) \geq \lambda_{\min}$ can be replaced by $\frac{1}{M}\sum_m d^{\pi^m_\tau}_h(s,a) \geq \lambda_h(s,a)$, which holds by the definition of λ_h because π^m_τ is the uniformly tie-broken greedy policy for $\widehat{P}_\tau \in \mathcal{P}$ for $\tau \geq 2$. Combining this with the count argument in the proof of Lemma 4, where K_{th} is unchanged because $\lambda_h(s,a) \geq \lambda_{\min}$, gives, for all (h,s,a),

$$N_{k,h}(s,a) \geq 1, \qquad N_{k,h}(s,a) \geq \frac{1}{8}kM\lambda_h(s,a). \tag{81}$$

Since $\widehat{P}_{k,h}(s,a)$ is a probability vector, the identity behind (74) reads $Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a) = \langle \widehat{P}_{k,h}(s,a) - P_h(s,a), V^{m,\star}_{h+1}\rangle + \langle \widehat{P}_{k,h}(s,a), V^m_{k,h+1} - V^{m,\star}_{h+1}\rangle$. For every s',

$$\begin{aligned} Q^m_{k,h+1}(s',\pi^{m,\star}_{h+1}(s')) - Q^{m,\star}_{h+1}(s',\pi^{m,\star}_{h+1}(s')) &\leq V^m_{k,h+1}(s') - V^{m,\star}_{h+1}(s') \\ &\leq \sum_{a'} \pi^m_{k,h+1}(a'|s')\big[Q^m_{k,h+1}(s',a') - Q^{m,\star}_{h+1}(s',a')\big]. \end{aligned} \tag{82}$$

The lower bound holds because $V^m_{k,h+1}(s') = \max_{a'} Q^m_{k,h+1}(s',a')$, which is at least $Q^m_{k,h+1}(s',\pi^{m,\star}_{h+1}(s'))$, and because $V^{m,\star}_{h+1}(s') = Q^{m,\star}_{h+1}(s',\pi^{m,\star}_{h+1}(s'))$. For the upper bound, $\pi^m_{k,h+1}(\cdot|s')$ is supported on $\arg\max_{a'} Q^m_{k,h+1}(s',a')$, so $V^m_{k,h+1}(s') = \sum_{a'} \pi^m_{k,h+1}(a'|s')Q^m_{k,h+1}(s',a')$, while $V^{m,\star}_{h+1}(s') \geq \sum_{a'} \pi^m_{k,h+1}(a'|s')Q^{m,\star}_{h+1}(s',a')$.

The substitution in the proof of Lemma 4, with the per-pair count (81) in place of the count of Lemma 3, gives $|\langle \widehat{P}_{k,h}(s,a) - P_h(s,a), V^{m,\star}_{h+1}\rangle| \leq CH\sqrt{\iota/(kM\lambda_h(s,a))} + CH\iota/(kM\lambda_{\min})$. In the same way, (53) gives $\|\widehat{P}_{k,h}(s,a) - P_h(s,a)\|_1 \leq C\sqrt{(S+\iota)/(kM\lambda_{\min})}$. For the maximal error ϵ_k of Lemma 5, Lemma 4 gives $\epsilon_k \leq C\sqrt{H^4\iota/(kM\lambda_{\min})}$ for $k > K_{th}$, since its second-order term is dominated there, because $kM\lambda_{\min} > K_{th}M\lambda_{\min} \geq 4\log(4M^2SAKH/\delta) \geq \iota$. Combining these bounds and using $\sqrt{(S+\iota)\iota} \leq 2\sqrt{S}\,\iota$ and $H\iota \leq H^2\sqrt{S}\,\iota$,

$$\big|\big\langle \widehat{P}_{k,h}(s,a) - P_h(s,a), V^{m,\star}_{h+1}\big\rangle\big| + \big\|\widehat{P}_{k,h}(s,a) - P_h(s,a)\big\|_1\epsilon_k \leq CH\sqrt{\frac{\iota}{kM\lambda_h(s,a)}} + \frac{CH^2\sqrt{S}\,\iota}{kM\lambda_{\min}}. \tag{83}$$

Start from the identity above and apply the upper half of (82), which is allowed because $\widehat{P}_{k,h}(s,a) \geq 0$. Then write $\widehat{P}_{k,h}(s,a) = P_h(s,a) + (\widehat{P}_{k,h}(s,a) - P_h(s,a))$, and bound the second part by $\|\widehat{P}_{k,h}(s,a) - P_h(s,a)\|_1\epsilon_k$, since $|Q^m_{k,h+1} - Q^{m,\star}_{h+1}| \leq \epsilon_k$. With (83), this gives

$$Q^m_{k,h}(s,a) - Q^{m,\star}_h(s,a) \leq CH\sqrt{\frac{\iota}{kM\lambda_h(s,a)}} + \frac{CH^2\sqrt{S}\,\iota}{kM\lambda_{\min}}$$

$$+ \sum_{s'} P_h(s'|s,a) \sum_{a'} \pi^m_{k,h+1}(a'|s') \big[Q^m_{k,h+1}(s',a') - Q^{m,\star}_{h+1}(s',a') \big]. \tag{84}$$

In the same way, the lower half of (82) gives

$$Q^{m,\star}_h(s,a) - Q^m_{k,h}(s,a) \le CH \sqrt{\frac{\iota}{kM\lambda_h(s,a)}} + \frac{CH^2\sqrt{S}\,\iota}{kM\lambda_{\min}} + \sum_{s'} P_h(s'|s,a) \big[Q^{m,\star}_{h+1}(s', \pi^{m,\star}_{h+1}(s')) - Q^m_{k,h+1}(s', \pi^{m,\star}_{h+1}(s')) \big]. \tag{85}$$

For $h = H$, the last sums are 0, because $V^m_{k,H+1} = V^{m,\star}_{H+1} = 0$.

We prove (54) with CH^3 in the second-order term replaced by $(H - h + 1)CH^2$, by backward induction on h. For $h = H$, the right-hand side of (84) is exactly the claim, because $\mathbb{E}^{\pi^m_k}[\lambda_H(s_H, a_H)^{-1/2} \mid s_H = s, a_H = a] = \lambda_H(s,a)^{-1/2}$. Suppose the claim holds at $h+1$ for all (s', a'), and substitute it into the last sum of (84). The weights $P_h(s'|s,a)\pi^m_{k,h+1}(a'|s')$ are nonnegative and sum to one, and, by the tower property,

$$\sum_{s'} P_h(s'|s,a) \sum_{a'} \pi^m_{k,h+1}(a'|s')\, \mathbb{E}^{\pi^m_k}\Big[\sum_{h'=h+1}^{H} \frac{1}{\sqrt{\lambda_{h'}(s_{h'}, a_{h'})}} \,\Big|\, s_{h+1} = s', a_{h+1} = a' \Big]$$
$$= \mathbb{E}^{\pi^m_k}\Big[\sum_{h'=h+1}^{H} \frac{1}{\sqrt{\lambda_{h'}(s_{h'}, a_{h'})}} \,\Big|\, s_h = s, a_h = a \Big].$$

Adding the step-h term $CH\sqrt{\iota/(kM)}\,\lambda_h(s,a)^{-1/2}$ completes the expectation from $h' = h$, and the second-order terms add up to $(H-h)CH^2 + CH^2 = (H-h+1)CH^2 \le CH^3$. This proves (54). The proof of (55) is identical, using (85) with $\pi^{m,\star}$ in place of π^m_k.

G Additional experimental results

We use synthetic finite-horizon tabular MDPs with heterogeneous rewards. We draw $P_h(\cdot \mid s,a) \sim \text{Dirichlet}(\alpha \cdot \mathbf{1})$, ρ is uniform, and small α yields concentrated dynamics that trap greedy policies. Each agent is assigned an independently drawn random subset of $\lceil \kappa \cdot SAH \rceil$ preferred (h,s,a) cells with reward 1 (and 0 elsewhere); small κ makes the rewards sparse, so myopic behaviour rarely reaches them. Unless stated otherwise we use $S = 10$, $A = 4$, $H = 5$, $\alpha = 0.05$, and $\kappa = 0.015$ (three reward cells out of 200); when sweeping M we share a base draw of $M_{\max}$ rewards and take the first M so only the federation size changes. Results average over multiple MDP draws; shaded bands show one standard deviation. We set $\widehat{P}_h(\cdot \mid s,a)$ to uniform at unvisited (h,s,a) to keep it a valid kernel, and scale the theoretical UCB bonus by a constant c_{bonus} tuned separately for each UCB algorithm (PF-UCBVI and Fed-UCBVI) on the same held-out MDP over $\{10^{-5}, \ldots, 10^{-1}\}$; the selected value is 10^{-5} for both.

Comparison with a consensus baseline. We also compare with Fed-UCBVI (Labbi et al., 2025b), which learns a single consensus policy maximizing the average value across agents, in the setup of Section 6.1. Table 2 reports the per-agent regret at $K = 1500$, obtained from a separate set of runs with different random seeds, so the values differ from those in Figure 3. With $M = 1$, where there is no reward heterogeneity, PF-UCBVI and Fed-UCBVI perform comparably. With $M = 4, 16$ agents of distinct rewards, the per-agent regret of Fed-UCBVI grows with M, since the consensus policy can be suboptimal for individual agents, whereas that of PF-UCBVI and PF-EFVI decreases with M.

Table 2: Per-agent regret at $K = 1500$ (mean $\pm$ std over 8 MDP draws).

M	Algorithm	Cumulative regret / agent	Instantaneous regret / agent
1	PF-UCBVI	87.8 ± 8.8	0.011 ± 0.007
	PF-EFVI	205.4 ± 130.2	0.113 ± 0.083
	Fed-UCBVI	107.0 ± 28.9	0.016 ± 0.011
4	PF-UCBVI	25.0 ± 5.3	0.002 ± 0.001
	PF-EFVI	33.1 ± 16.5	0.010 ± 0.009
	Fed-UCBVI	744.9 ± 132.1	0.481 ± 0.089
16	PF-UCBVI	7.3 ± 1.1	0.000 ± 0.000
	PF-EFVI	7.7 ± 1.3	0.001 ± 0.001
	Fed-UCBVI	1162.7 ± 130.9	0.771 ± 0.087

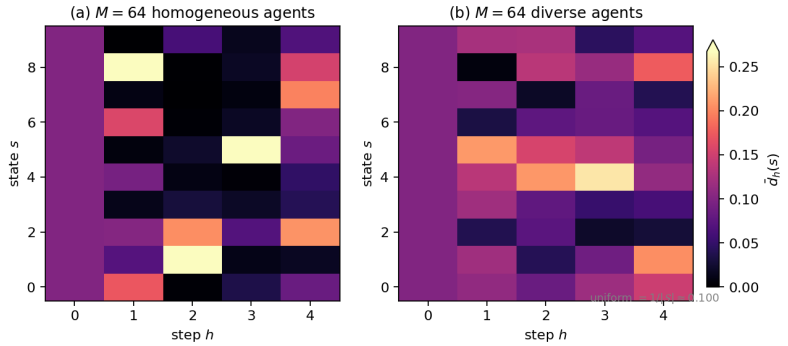

Figure 4: Reward diversity, not ensemble size, drives collective coverage. State visitation $\bar{d}_h(s)$ on a single MDP with $\widetilde{P} = P$. **(a)** $M = 64$ agents with identical rewards: visitation is highly uneven (brightest cells exceed the color scale). **(b)** $M = 64$ agents with i.i.d. rewards: visits spread nearly uniformly. The shared color scale is capped at the diverse panel's maximum.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provided theoretical and empirical results to support our claims in Section 5, Section 4, and Section 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the conditional nature of PF-EFVI's guarantee in Section 5, specifically, that the $\mathcal{O}(\log K)$ regret holds under Assumption 1 (see "Complementary regimes with PF-UCBVI" paragraph and the empirical demonstration at $M = 1, 4$ in Section 6.3).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provided the full set of assumptions (Assumption 1) and complete proofs for theoretical results in the main paper and the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided all the experiment setup needed to reproduce the main experimental results in Section 6 and Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

Justification: We do not release the code, but we provide complete pseudocode for both algorithms in Sections 4 and 5, and all experimental details (synthetic MDP construction, hyperparameters, sweeping protocol) in Section 6.1 and Appendix G, which we believe is sufficient to reproduce all results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided all the training and test details for our experiments in Section 6 and Appendix G, including data generation, hyperparameters, and how they were chosen.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We reported results averaged over multiple MDPs with error bars in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments use synthetic tabular MDPs with $|\mathcal{S}| = 10$, $|\mathcal{A}| = 4$, $H = 5$, $K = 1500$, and $M \leq 128$, which can be run on a standard laptop CPU (no GPU required). A single experimental sweep over $M \in \{1, 4, 16, 64, 128\}$ with 8 MDP draws completes in under 30 minutes on a modern laptop.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that our research conforms to it in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is a theoretical study of personalized federated reinforcement learning with synthetic experiments only. We do not deploy any model or release any dataset, and the contributions are foundational algorithmic and theoretical results. We do not foresee direct negative societal impacts; on the positive side, exploration-free federated RL has potential applications in safety-critical domains (e.g., personalized healthcare, robotics) where explicit exploration is costly or risky, but realizing such impact would require substantial additional work beyond the scope of this paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any pre-trained models, datasets, or assets that pose a risk for misuse. All experiments are conducted on synthetic tabular MDPs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We did not use any existing data or code assets in our paper, so this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new assets such as datasets, pre-trained models, or external code packages. The pseudocode for our algorithms is provided in Sections 4 and 5.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or any research with human subjects. All experiments use synthetic MDPs.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so no IRB approval is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This research does not use LLMs as an important, original, or non-standard component of the core methods. Any usage was limited to writing and editing assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.