Graph-Localized Offline Federated Multi-Agent Reinforcement Learning for Wireless Networks

Anonymous Authors[1]

Abstract

Wireless networks motivate offline multi-agent reinforcement learning (MARL) as online exploration degrades service, and operator logs are rarely poolable across deployments. Existing offline MARL relies on global coverage that scales exponentially in the number of agents. We exploit the interaction graph induced by interference and contention to replace global coverage with κ-hop neighborhood coverage: a client contributes to agent i's local estimator only if it observes $\mathcal{N}_i^\kappa$. We prove a localized offline policy guarantee whose error decomposes into a locality bias decaying exponentially in κ, a near-neighbor shift bias, and a federated estimation term shrinking with the pooled observability-valid sample size. Our algorithm, F-GLOFF, matches a centralized raw-pooled oracle on multi-AP user association without sharing transitions and outperforms the engineered baselines by 12%.

1. Introduction

Wireless networks naturally fit multi-agent reinforcement learning (MARL), where distributed agents act on partial observations, interact through interference and shared resources, and optimize system-level objectives such as throughput and reliability (Guo et al., 2022; Sana et al., 2020; Yang et al., 2024a). However, online exploration can degrade service or violate latency and reliability requirements, motivating offline RL where policies are learned from logged trajectories without further interaction (Levine et al., 2020). Operational logs are also distributed across deployments and often cannot be pooled because of privacy or bandwidth (Hu et al., 2021; Zhu et al., 2021; Kwon et al., 2025). These constraints motivate offline federated MARL, where policies are learned from local data and clients share sufficient statistics rather than raw trajectories (Khodadadian et al., 2022; Zhou et al., 2024; Woo et al., 2024).

However, offline RL is reliable only where logged data support the learned policy. Pessimistic offline RL addresses this by penalizing uncertain state–action pairs and is often analyzed through single-policy concentrability (Rashidinejad et al., 2021; Xie et al., 2021; Jin et al., 2021; Shi et al., 2022; Li et al., 2024). In MARL, however, the joint state–action space grows exponentially with the number of agents, making global coverage assumptions restrictive. Existing offline MARL methods use conservative value learning, counterfactual regularization, implicit constraints, or generative modeling to reduce distribution shift (Yang et al., 2021; Pan et al., 2022; Shao et al., 2023; Wang et al., 2023b; Zhu et al., 2024; Eldeeb et al., 2024), but they generally do not use the interaction graph to define the relevant coverage object.

Wireless networks offer a structural way around this barrier. Interference and contention induce an interaction graph, and each agent is mainly affected by nearby agents. Networked MARL exploits this locality through κ-hop policies and exponential decay of influence (Qu et al., 2020; Lin et al., 2021; Zhang et al., 2023). This line of work reduces the effective control dimension but is mainly online. Federated RL and wireless federated RL study distributed client data and model aggregation (Jin et al., 2022; Lan et al., 2024; Woo et al., 2025; Kwon et al., 2025), but usually define client relevance by participation rather than graph observability.

We study offline federated MARL around graph-local neighborhoods so that the relevant coverage object becomes the κ-hop neighborhood projection rather than the full joint state–action space, and a client contributes to agent i at radius κ only when its observable region contains $\mathcal{N}_i^\kappa$. The radius κ controls a bias–coverage tradeoff. Our contributions are:

1. We formulate offline federated MARL on interaction graphs and introduce neighborhood concentrability and observability-valid client sets as the localized analogs of global coverage and client relevance.

2. We prove a localized offline policy guarantee whose error decomposes additively into a locality bias decay-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ing as $\rho^{\kappa+1}$, a κ-independent near-neighbor shift bias, and a federated estimation term that shrinks with the pooled observability-valid sample size.

3. We propose **F-GLOFF**, a tabular algorithm that aggregates projected statistics from observability-valid clients and applies count-based pessimism without sharing raw trajectories. In multi-AP user association, F-GLOFF matches a centralized raw-pooled oracle, improves over the engineered baseline by 12% and the non-localized federated variant by 22%, and exhibits the predicted bias–coverage tradeoff.

2. Graph-Structured Offline MARL

Let $\mathcal{G} = (\mathcal{N}, E)$ be an undirected graph over agents $\mathcal{N} = \{1, \dots, n\}$. Agent i has local state space $\mathcal{S}_i$ and action space $\mathcal{A}_i$, giving global spaces $\mathcal{S} = \prod_i \mathcal{S}_i$ and $\mathcal{A} = \prod_i \mathcal{A}_i$. The environment is a cooperative discounted MDP with transition kernel $P(\cdot \mid s, a)$ and additive reward $r(s, a) = \sum_i r_i(s, a)$. For policy π and initial distribution μ,

$$V^{\pi}(\mu) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 \sim \mu\right], \qquad \gamma \in (0, 1).$$

Let $\pi^\star$ be an optimal unrestricted policy and define the κ-hop neighborhood $\mathcal{N}_i^\kappa = \{j \in \mathcal{N} : d_{\mathcal{G}}(i, j) \le \kappa\}$.

Assumption 2.1 (Local interaction-radius factorization)**.** For every agent i, the local reward and marginal next state distribution are determined by the one-hop neighborhood:

$$r_i(s, a) = r_i(s_{\mathcal{N}_i^1}, a_{\mathcal{N}_i^1}), \quad P(s_i' \mid s, a) = P_i(s_i' \mid s_{\mathcal{N}_i^1}, a_{\mathcal{N}_i^1}).$$

This is the standard interaction-radius factorization of networked MARL (Lin et al., 2021; Zhang et al., 2023); it generalizes the purely local form $r_i(s_i, a_i)$ of (Qu et al., 2020) to allow contemporaneous neighbor effects, which is essential for collision style interactions in wireless networks. One-step locality does not imply that long horizon values are local; Sec. 3 controls long range propagation through exponential decay.

Assumption 2.2 (Markov sufficient local context)**.** For every agent i and every radius $\kappa \ge 0$, there exists a finite set $\mathcal{X}_i^\kappa$ and a measurable map $\phi_i^\kappa : \mathcal{S}_{\mathcal{N}_i^\kappa} \to \mathcal{X}_i^\kappa$ such that the projected context $x_i^\kappa = \phi_i^\kappa(s_{\mathcal{N}_i^\kappa})$ is Markov sufficient for its own one-step evolution given the interaction-radius action profile: for all (s, a) and all $x' \in \mathcal{X}_i^\kappa$,

$$\Pr\big[\phi_i^\kappa(s'_{\mathcal{N}_i^\kappa}) = x' \;\big|\; s, a\big] = P_i^\kappa\big(x' \;\big|\; x_i^\kappa, a_{\mathcal{N}_i^1}\big).$$

Assumption 2.2 makes the projected per-agent transition P_i^κ an environment property rather than a property of the behavior policy, so count-based projected estimators have well-defined population targets; Sec. 6 provides a concrete realization. The boundary $\mathcal{N}_i^{\kappa+1} \setminus \mathcal{N}_i^\kappa$ contributes residual approximation of order $\rho^{\kappa+1}$, on the same order as the locality bias of Sec. 3 (Appendix B.3).

For decentralized control, define the κ-local policy class

$$\Pi_\kappa = \Big\{\pi : \pi(a \mid s) = \prod_{i=1}^{n} \pi_i(a_i \mid x_i^\kappa)\Big\}.$$

Let $\pi_\kappa^\star \in \arg\max_{\pi \in \Pi_\kappa} V^\pi(\mu)$ and let $\hat{\pi}_\kappa$ be the offline learned policy. The suboptimality decomposes as

$$V^{\pi^\star}(\mu) - V^{\hat{\pi}_\kappa}(\mu) \le \underbrace{V^{\pi^\star}(\mu) - V^{\pi_\kappa^\star}(\mu)}_{\text{localization bias}} + \underbrace{V^{\pi_\kappa^\star}(\mu) - V^{\hat{\pi}_\kappa}(\mu)}_{\text{offline estimation error}}.$$

Sec. 3 further refines the second term into a decentralization gap and an estimation gap, since F-GLOFF performs *per-agent* independent optimization rather than joint optimization in Π_κ.

Offline data and federated structure. The learner receives a fixed dataset $\mathcal{D} = \{(s_t^{(m)}, a_t^{(m)}, r_t^{(m)}, s_{t+1}^{(m)})\}_{m=1}^{M}$ from behavior policies and no further interaction. Instead of requiring coverage over $\mathcal{S} \times \mathcal{A}$, the learner targets the per-agent projected space $\mathcal{Z}_i^\kappa = \mathcal{X}_i^\kappa \times \mathcal{A}_i$ via tuples $(x_i^\kappa, a_i, r_i, x_i^{\kappa\prime})$; the interaction-radius transition $P_i^\kappa(\cdot \mid x_i^\kappa, a_{\mathcal{N}_i^1})$ and the per-agent action resolution are connected by behavior marginalization, formalized in Sec. 3 as the per-agent population MDP $\mathcal{M}_{i,\kappa,\bar{\mu}}$.

In the federated setting, $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$, where client k owns agents $_k$ and observes a region $\mathcal{O}_k \supseteq_k$ (states and actions of all agents in $\mathcal{O}_k$). A client contributes to agent i at radius κ only when it observes the full radius-κ neighborhood needed to construct x_i^κ and $x_i^{\kappa\prime}$, yielding the *observability-valid client set* $\mathcal{R}_i^\kappa = \big\{k \in [K] : \mathcal{N}_i^\kappa \subseteq \mathcal{O}_k,\ \mathcal{D}_{k,i}^\kappa \ne \emptyset\big\}$. The federated learner aggregates projected sufficient statistics over $k \in \mathcal{R}_i^\kappa$, the single-client baseline uses only the owner of i when observability-valid.

3. Localized Offline Learning Theory

We now formalize the bias-coverage tradeoff. Decompose the global action-value function as $Q^\pi(s, a) = \sum_i Q_i^\pi(s, a)$, where

$$Q_i^\pi(s, a) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \,\middle|\, s_0 = s,\ a_0 = a\right],$$

and assume $|r_i(s, a)| \le R_{\max}$ for all i, s, a.

Assumption 3.1 (Exponential decay of local value influence)**.** Let $\Pi_{\text{stat}} := \{\pi : \mathcal{S} \to \Delta(\mathcal{A}) : \pi \text{ stationary}\}$. There exist $c_{\text{loc}} > 0$ and $\rho \in (0, 1)$ such that for every $\pi \in \Pi_{\text{stat}}$, every agent i, every $\ell \ge 0$, and every $(s, a), (\bar{s}, \bar{a}) \in \mathcal{S} \times \mathcal{A}$ with $(s_{\mathcal{N}_i^\ell}, a_{\mathcal{N}_i^\ell}) = (\bar{s}_{\mathcal{N}_i^\ell}, \bar{a}_{\mathcal{N}_i^\ell})$,

$$\big|Q_i^\pi(s, a) - Q_i^\pi(\bar{s}, \bar{a})\big| \le c_{\text{loc}} \rho^{\ell+1}.$$

This is the exponential decay condition of (Qu et al., 2020; Lin et al., 2021). Distant coordinates have an exponentially small effect on local long run value, uniformly over stationary policies. It is a property of the MDP, independent of the data, and applies to $\pi^\star$, every $\pi \in \bigcup_{\kappa\geq 0}\Pi_\kappa$, the joint behavior policy $\bar{\mu}$, and pathwise to any realization of $\widehat{\pi}_\kappa$. Since $|Q_i^\pi| \leq R_{\max}/(1-\gamma)$, the constant c_{loc} scales as $O(R_{\max}/(1-\gamma))$ and is independent of n; the factor n in the bounds below comes from summing per-agent terms. Sufficient conditions are known under uniform local mixing (Qu et al., 2020) and under graph contraction assumptions (Lin et al., 2021; Zhang et al., 2023).

Proposition 3.2 (Localization gap)**.** *Under Assumption 3.1,*

$$V^{\pi^\star}(\mu) - V^{\pi^\star_\kappa}(\mu) \leq \frac{2nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma}.$$

The proof truncates each $Q_i^{\pi^\star}$ to a κ-hop function and applies a one-step approximate greedification argument (Appendix C).

Per-agent population MDP and comparator. F-GLOFF's empirical model conditions on the agent's own action a_i rather than on the full one-hop neighbor profile $a_{\mathcal{N}_i^1}$ used in Assumption 2.1. Under behavior policy $\bar{\mu}$, the projected transition $P_i^\kappa(\cdot \mid x, a_{\mathcal{N}_i^1})$ from Assumption 2.2 therefore induces a per-agent transition by marginalization,

$$P^\kappa_{i,\bar{\mu}}(x' \mid x, a_i) := \sum_{a_{-i}^{(1)}} \bar{\nu}_i\big(a_{-i}^{(1)} \mid x, a_i\big)\, P_i^\kappa\big(x' \mid x, (a_i, a_{-i}^{(1)})\big),$$

where $\bar{\nu}_i(\cdot \mid x, a_i)$ is the conditional behavior distribution of $a_{\mathcal{N}_i^1\setminus\{i\}}$ given (x_i^κ, a_i). Defining $r^\kappa_{i,\bar{\mu}}$ analogously, the pair $(P^\kappa_{i,\bar{\mu}}, r^\kappa_{i,\bar{\mu}})$ defines the *per-agent population MDP* $\mathcal{M}_{i,\kappa,\bar{\mu}}$ with state space $\mathcal{X}_i^\kappa$ and action space $\mathcal{A}_i$. Let $\bar{\pi}^\star_{i,\kappa} \in \arg\max_{\pi_i} V^{\pi_i}_{i,\kappa,\bar{\mu}}$ and $\bar{\pi}^\star_\kappa := \bigotimes_{i=1}^n \bar{\pi}^\star_{i,\kappa} \in \Pi_\kappa$. F-GLOFF's per-agent independent optimization targets $\bar{\pi}^\star_\kappa$, not the joint Π_κ-optimum $\pi^\star_\kappa$; the gap between these comparators is the third source of suboptimality. To control this gap we need a structural assumption on how far the comparator class differs from behavior on the one-hop interaction neighborhood.

Assumption 3.3 (Bounded near-neighbor shift)**.** There exists $\tau \geq 0$ such that for every agent i and every $\pi \in \{\pi^\star_\kappa, \bar{\pi}^\star_\kappa, \widehat{\pi}_\kappa\}$,

$$\sup_{(x,a_i)\in\mathcal{X}_i^\kappa\times\mathcal{A}_i} \big\|\pi_{-i}(\cdot \mid s_{\mathcal{N}_i^1\setminus\{i\}}) - \bar{\nu}_i(\cdot \mid x, a_i)\big\|_{\mathrm{TV}} \leq \tau,$$

where $\pi_{-i}(\cdot \mid s_{\mathcal{N}_i^1\setminus\{i\}})$ denotes the joint distribution of $a_{\mathcal{N}_i^1\setminus\{i\}}$ induced by π.

The constant τ is a property of the MDP–comparator–behavior triple, independent of the offline data, and does not vanish in the infinite data limit. It is small whenever the behavior policy is factorized over $\mathcal{N}_i^1$ and the comparator's per-agent factors are pointwise close to behavior. In the wireless instantiation of Sec. 6 all four logging policies are radius-0 factorized, so $\bar{\nu}_i$ is a product distribution and Assumption 3.3 holds with small τ. A formal discussion is in Appendix C.4.

Proposition 3.4 (Decentralization gap)**.** *Under Assumptions 3.1 and 3.3,*

$$V^{\pi^\star_\kappa}(\mu) - V^{\bar{\pi}^\star_\kappa}(\mu) \leq \frac{2nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{4nR_{\max}\tau}{(1-\gamma)^2}.$$

The proof mirrors Proposition 3.2, applied agent-wise on $\mathcal{M}_{i,\kappa,\bar{\mu}}$: truncation of $Q_i^{\pi^\star_\kappa}$ to a κ-hop function followed by one-step approximate greedification. The two contributions are the locality bias from boundary truncation and the near-neighbor shift bias from comparator–behavior mismatch on $\mathcal{N}_i^1$, identified by Corollary C.3 (Appendix C.4).

Bellman-consistent localized concentrability. Pessimism in F-GLOFF must control distribution shift between the deployed per-agent policy and the projected behavior on $\mathcal{X}_i^\kappa \times \mathcal{A}_i$. We use the localized version of the Bellman-consistent coefficient of (Xie et al., 2021), evaluated on $\mathcal{M}_{i,\kappa,\bar{\mu}}$. Let $\mathcal{F}_i^\kappa := \big\{f : \mathcal{X}_i^\kappa \times \mathcal{A}_i \to [0, R_{\max}/(1-\gamma)]\big\}$, $\mathcal{T}^\pi_{i,\kappa,\bar{\mu}}$ the per-agent Bellman operator, $d^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}$ the discounted occupancy of $\bar{\pi}^\star_{i,\kappa}$, and $\bar{d}_{i,\kappa,\bar{\mu}}$ the projected behavior occupancy.

Definition 3.5 (Localized Bellman-consistent concentrability)**.**

$$C^\star_{\kappa,\mathrm{BE}} := \max_{i\in\mathcal{N}} \sup_{f\in\mathcal{F}_i^\kappa} \frac{\mathbb{E}_{(x,a_i)\sim d^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}}\big[(f - \mathcal{T}^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}} f)^2(x,a_i)\big]}{\mathbb{E}_{(x,a_i)\sim \bar{d}_{i,\kappa,\bar{\mu}}}\big[(f - \mathcal{T}^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}} f)^2(x,a_i)\big]},$$

with the convention that the ratio is $+\infty$ if the denominator vanishes and the numerator does not.

This is a Bellman-residual ratio rather than a state–action density ratio, capturing distribution shift transmitted through the projected dynamics on $\mathcal{M}_{i,\kappa,\bar{\mu}}$, including the dependence of next state transitions on the behavior policy's neighbor-action distribution. By (Xie et al., 2021), finiteness of standard density-ratio single policy concentrability implies finiteness of $C^\star_{\kappa,\mathrm{BE}}$ up to a $1/(1-\gamma)^2$ factor, but the converse can fail.

Let $S_\kappa := \max_i |\mathcal{X}_i^\kappa|$, $A_{\max} := \max_i |\mathcal{A}_i|$, and $N_\kappa := \min_i N_{i,\kappa}$. Applying the Bellman-consistent pessimistic bound to each $\mathcal{M}_{i,\kappa,\bar{\mu}}$ with comparator $\bar{\pi}^\star_{i,\kappa}$ and union-

bounding over agents gives, with probability at least $1-\delta$,

$$\sum_i \Big(V^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i) - V^{\widehat{\pi}_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i) \Big) \le C_1\, n \sqrt{\frac{C^\star_{\kappa,\mathrm{BE}}\, S_\kappa\, A_{\max}\, \log(n/\delta)}{(1-\gamma)^3\, N_\kappa}}.$$

where C_1 is the absolute constant .Relating the per-agent population sum on the LHS to the global value gap $V^{\bar{\pi}^\star_\kappa}(\mu) - V^{\widehat{\pi}_\kappa}(\mu)$ requires a local-to-global passage (Corollary C.3), which contributes $2nc_{\mathrm{loc}}\rho^{\kappa+1}/(1-\gamma) + 4nR_{\max}\tau/(1-\gamma)^2$ to the bias. Details are in Appendix C.5.

Theorem 3.6 (Localized offline policy guarantee)**.** *Suppose Assumptions 2.1, 2.2, 3.1, and 3.3 hold and $C^\star_{\kappa,\mathrm{BE}} < \infty$. Then the localized pessimistic policy $\widehat{\pi}_\kappa$ satisfies, with probability at least $1-\delta$,*

$$V^{\pi^\star}(\mu) - V^{\widehat{\pi}_\kappa}(\mu) \le \frac{6nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{8nR_{\max}\tau}{(1-\gamma)^2} + C_1\, n \sqrt{\frac{C^\star_{\kappa,\mathrm{BE}}\, S_\kappa\, A_{\max}\, \log(n/\delta)}{(1-\gamma)^3\, N_\kappa}}.$$

The bound has three structurally distinct sources, contributing additively. The *locality bias* $6nc_{\mathrm{loc}}\rho^{\kappa+1}/(1-\gamma)$ aggregates equal contributions from the localization gap (Proposition 3.2), the decentralization gap (Proposition 3.4), and the local-to-global passage in the estimation argument (Corollary C.3); it decays exponentially in κ. The *near-neighbor shift bias* $8nR_{\max}\tau/(1-\gamma)^2$ collects the two τ-terms from the decentralization gap and the local-to-global passage; it is independent of κ and small whenever Assumption 3.3 holds with small τ. The *estimation term* grows with κ because S_κ enlarges and per-cell coverage worsens. Federation modifies only the estimation term. Sec. 4 replaces $\bar{d}_{i,\kappa,\bar{\mu}}$ and $N_{i,\kappa}$ by their observability valid federated counterparts, leaving the bias terms unchanged.

4. Federated Graph-Localized Estimation

Once the control problem is localized, federation should also be localized: for each agent i and radius κ, only clients that observe $\mathcal{N}^\kappa_i$ may contribute to its local estimator. Federation therefore affects the bound of Theorem 3.6 through the observability-valid client set $\mathcal{R}^\kappa_i$ (Sec. 2), the induced coverage coefficient, and the effective local sample count.

4.1. Federated estimator

Define client k's projected dataset for agent i at radius κ as

$$\mathcal{D}^\kappa_{k,i} = \big\{ (x_t, a_{i,t}, r_{i,t}, x'_{t+1}) : x_t = \phi^\kappa_i(s_{t,\mathcal{N}^\kappa_i}) \big\},$$

admissible only when $k \in \mathcal{R}^\kappa_i$. Each $k \in \mathcal{R}^\kappa_i$ computes the per-agent action-resolution sufficient statistics $N^\kappa_{k,i}(x,a_i)$, $N^\kappa_{k,i}(x,a_i,x')$, $R^\kappa_{k,i}(x,a_i)$; context-action counts, transition counts, and reward sums and shares them with the server; raw transitions and policy parameters are not shared. The neighbor action profile $a_{\mathcal{N}^1_i\setminus\{i\},t}$ is integrated out implicitly by counting at the (x,a_i) resolution.

The server aggregates $N^{\mathrm{fed}}_{i,\kappa}(x,a_i) = \sum_{k\in\mathcal{R}^\kappa_i} N^\kappa_{k,i}(x,a_i)$, and analogously for transition counts and reward sums, yielding;

$$\widehat{r}^\kappa_i(x,a_i) = \frac{R^\kappa_{i,\mathrm{fed}}(x,a_i)}{\max\{1, N^{\mathrm{fed}}_{i,\kappa}(x,a_i)\}},$$
$$\widehat{P}^\kappa_i(x' \mid x,a_i) = \frac{N^\kappa_{i,\mathrm{fed}}(x,a_i,x')}{\max\{1, N^{\mathrm{fed}}_{i,\kappa}(x,a_i)\}}.$$

Because aggregation is at the (x,a_i) resolution, this is an unbiased count-based estimator of the per-agent population MDP $\mathcal{M}_{i,\kappa,\bar{\mu}}$ from Sec. 3, not of the full interaction-radius dynamics $P^\kappa_i(\cdot \mid x, a_{\mathcal{N}^1_i})$. The single client baseline uses only agent i's owner when observability-valid, recovering asynchronous offline pessimism (Yan et al., 2023); the two estimators differ in which observability-valid statistics are aggregated.

F-GLOFF subtracts the count-based pessimism bonus

$$b^\kappa_i(x,a_i) = \min\left\{ c_b \sqrt{\frac{\log(2|\mathcal{X}^\kappa_i||\mathcal{A}_i|/\delta)}{\max\{1, N^{\mathrm{fed}}_{i,\kappa}(x,a_i)\}}}, b_{\max} \right\}$$

from the empirical reward.

4.2. Federated coverage and guarantee

The pooled projected behavior occupancy on $\mathcal{X}^\kappa_i \times \mathcal{A}_i$ is

$$\bar{d}^{\mathrm{fed}}_{i,\kappa}(x,a_i) = \sum_{k\in\mathcal{R}^\kappa_i} \frac{N^\kappa_{k,i}}{N^{\mathrm{fed}}_{i,\kappa}}\, \bar{d}^\kappa_{k,i}(x,a_i),$$

and the federated localized Bellman-consistent coefficient is the analog of Definition 3.5 with $\bar{d}_{i,\kappa,\bar{\mu}}$ replaced by $\bar{d}^{\mathrm{fed}}_{i,\kappa}$:

$$C^\star_{\kappa,\mathrm{BE,fed}} := \max_{i\in\mathcal{N}} \sup_{f\in\mathcal{F}^\kappa_i} \frac{\mathbb{E}_{(x,a_i)\sim \bar{d}^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}} \big[(f - \mathcal{T}^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}} f)^2(x,a_i) \big]}{\mathbb{E}_{(x,a_i)\sim \bar{d}^{\mathrm{fed}}_{i,\kappa}} \big[(f - \mathcal{T}^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}} f)^2(x,a_i) \big]}.$$

Effect of federation on the coefficient. Federation affects $C^\star_{\kappa,\mathrm{BE,fed}}$ relative to its single client counterpart in two ways. First, the pooled support equals the union of per-client supports, $\mathrm{supp}(\bar{d}^{\mathrm{fed}}_{i,\kappa}) = \bigcup_{k\in\mathcal{R}^\kappa_i} \mathrm{supp}(\bar{d}^\kappa_{k,i})$. Cells supported by some non-owner client but not by the owner have a nonzero federated denominator, removing otherwise-infinite ratios. Second, on the common support, averaging across clients can dilute density at the comparator's distribution and increase the supremum at a point. Whether the

coefficient tightens or loosens depends on whether the support gain dominates within support dilution; empirically, it does.

Define $N_\kappa^{\mathrm{fed}} := \min_{i\in\mathcal{N}} N_{i,\kappa}^{\mathrm{fed}}$. Substituting $C^\star_{\kappa,\mathrm{BE,fed}}$ and N_κ^{fed} into Theorem 3.6 yields, with probability at least $1-\delta$,

$$V^{\pi^\star}(\mu) - V^{\widehat{\pi}_\kappa^{\mathrm{fed}}}(\mu) \le \underbrace{\frac{6nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma}}_{\text{locality bias}} + \underbrace{\frac{8nR_{\max}\tau}{(1-\gamma)^2}}_{\text{near-neighbor shift bias}} + \varepsilon_{\mathrm{fed}}(\kappa), \tag{1}$$

where $\varepsilon_{\mathrm{fed}}(\kappa) := C_1\, n\sqrt{\frac{C^\star_{\kappa,\mathrm{BE,fed}}\, S_\kappa\, A_{\max}\, \log(n/\delta)}{(1-\gamma)^3\, N_\kappa^{\mathrm{fed}}}}$. Federation acts only on the estimation term, by increasing sample counts ($N_\kappa^{\mathrm{fed}} \ge N_\kappa^{\mathrm{single}}$) and potentially shrinking $C^\star_{\kappa,\mathrm{BE,fed}}$ through expanded support; the bias terms depend only on the MDP, the policy class Π_κ, and Assumption 3.3, not on the data. The exponent $\log(n/\delta)$ rather than $\log(nK/\delta)$ reflects that $N_{i,\kappa}^{\mathrm{fed}}$ is a single sum of independent samples and $\mathcal{R}_i^\kappa$ is determined by graph observability, so no union bound over K is needed.

4.3. Out-of-support diagnostic

Pessimism discourages unsupported context-action pairs at training time. We also report a deployment-time out-of-support diagnostic. For threshold $n_{\min}$, let $\mathcal{S}_{i,\kappa}^{\sup} := \{(x,a_i) : N_{i,\kappa}^{\mathrm{fed}}(x,a_i) \ge n_{\min}\}$. For deployed policy π,

$$\mathrm{OOS}(\pi) := \frac{1}{n}\sum_{i=1}^{n} \Pr_{x_i^\kappa\sim d_i^\pi,\, a_i\sim\pi_i(\cdot|x_i^\kappa)}\big[(x_i^\kappa,a_i)\notin\mathcal{S}_{i,\kappa}^{\sup}\big].$$

OOS captures how often the learned policy visits cells that were poorly supported in the offline data. It is a deployment diagnostic, not a training objective.

5. F-GLOFF Algorithm

F-GLOFF builds a κ-local empirical model for each agent from observability valid clients, evaluates candidate policies pessimistically, and selects the best per agent. Each agent solves an offline RL problem on its own per-agent population MDP $\mathcal{M}_{i,\kappa,\bar{\mu}}$ (Sec. 3); the joint output $\widehat{\pi}_\kappa = \bigotimes_i \pi_{i,\widehat{\theta}_i}$ is the empirical per-agent best response policy, whose population-level analog $\bar{\pi}_\kappa^\star$ is the comparator in (1).

Empirical model and policy class. For agent i, F-GLOFF aggregates projected transitions (x,a_i,r_i,x') over $k\in\mathcal{R}_i^\kappa$ to obtain $\widehat{r}_i^\kappa$, $\widehat{P}_i^\kappa$, and counts $N_i^\kappa(x,a_i)$ as in Sec. 4.1. Candidate policies $\pi_{i,\theta}$ are softmax over four local features: $s_\theta(y;x) = \theta_\eta\eta_y - \theta_{\mathrm{load}}\ell_y + \theta_{\mathrm{queue}}q_{\mathrm{own}}$, $s_\theta(\perp;x) = \theta_{\mathrm{defer}} - \theta_{\mathrm{queue}}q_{\mathrm{own}}$, where q_{own} is the UE queue state, ℓ_y is the local load around AP y, and η_y is the service probability of AP y. Then $\pi_{i,\theta}(a\mid x)\propto\exp(s_\theta(a;x)/\tau)$, with empty queues forced to defer. The candidate set Θ is a finite grid over $(\theta_\eta,\theta_{\mathrm{load}},\theta_{\mathrm{queue}},\theta_{\mathrm{defer}})$.

Algorithm 1 Graph-Localized Offline Federated MARL

Require: Graph $\mathcal{G}$, radius κ, client data $\{\mathcal{D}_k\}$, observable regions $\{\mathbb{O}_k\}$, grid Θ
for each agent $i\in\mathcal{N}$ **do**
  Construct $\mathcal{N}_i^\kappa$ and $\mathcal{R}_i^\kappa$
  Aggregate projected statistics over $k\in\mathcal{R}_i^\kappa$
  Form $\widehat{P}_i^\kappa$ and $\widehat{r}_i^\kappa$
  for each $\theta\in\Theta$ **do**
    Compute pessimistic rewards and fallback values
    Run H Bellman backup iterations
    Compute d^{π_θ} by power iteration
    Compute $V_i^{\mathrm{lb}}(\theta)$
  end for
  $\widehat{\theta}_i \leftarrow \arg\max_{\theta\in\Theta} V_i^{\mathrm{lb}}(\theta)$
end for$\widehat{\pi}_\kappa = \bigotimes_i \pi_{i,\widehat{\theta}_i}$

Pessimistic evaluation and selection. For each θ, supported cells receive the pessimistic reward

$$\widehat{r}_{i,\mathrm{pess}}^\kappa(x,a) = \widehat{r}_i^\kappa(x,a) - \min\left\{c_b\sqrt{\frac{\log(2|\mathcal{X}_i^\kappa||\mathcal{A}_i|/\delta)}{N_i^\kappa(x,a)}},\, b_{\max}\right\},$$

and unsupported actions use a self-loop fallback with reward $-b_{\max}$. F-GLOFF applies H Bellman iterations to obtain $V_i^{\theta,H}$, then scores each candidate under its empirical stationary distribution d^{π_θ}, $V_i^{\mathrm{lb}}(\theta) = \frac{\sum_x d^{\pi_\theta}(x)V_i^{\theta,H}(x)}{(1-\gamma^H)/(1-\gamma)}$, $\widehat{\theta}_i = \arg\max_{\theta\in\Theta} V_i^{\mathrm{lb}}(\theta)$. Implementation details such as on-policy aggregation, computational cost, are in Appendix D.

6. Wireless Instantiation and Evaluation Setup

We instantiate F-GLOFF on multi-AP user association: each UE chooses whether to defer or transmit to an in-range AP. UEs interact only when they share APs, so the access-set conflict graph is the natural locality structure. The simulator is slot-level and abstracts PHY/MAC details: each AP has a service success probability, and simultaneous transmissions to the same AP collide. UEs and APs are placed uniformly in $[0,L]^2$. UE i can access $Y_i = \{y\in[M] : \|z_i - w_y\| \le R\}$, and the conflict graph satisfies $\{i,j\}\in E \iff Y_i\cap Y_j\neq\emptyset$. A nonempty UE chooses $a_i(t)\in\{\perp\}\cup Y_i$; an empty UE defers. A unique transmission to AP y succeeds with probability η_y; collisions yield zero reward. Successful UEs receive reward one and remove their earliest deadline packet; unserved packets expire at their deadline. The local context is $x_i^\kappa = (q_i,(\ell_y)_{y\in Y_i})$, where q_i is UE i's queue id and $\ell_y = \min\big(\ell_{\max}, |\{j\in\mathcal{N}_i^\kappa : y\in Y_j,\ q_j\neq 0\}|\big)$ is the capped local load around AP y. The context size is $|\mathcal{X}_i^\kappa| = 2^d(\ell_{\max}+1)^{|Y_i|}$.

Table 1. Reference comparison. Tx: per-UE per-slot transmit attempts; Eff: deliveries / arrivals.

Method	Reward	Tx	Eff	OOS
Engineered baselines				
`round_robin`	2.34 ± 0.10	0.31	0.68	0.00
`best_q`	1.81 ± 0.36	0.32	0.52	0.00
`behavior_mix`	2.14 ± 0.12	0.25	0.62	0.00
Non-localized				
fed (observability)	2.15 ± 0.32	0.24	0.63	0.78
raw-pooled (oracle)	2.63 ± 0.08	0.29	0.76	0.01
F-GLOFF (ours)				
$\kappa = 0$ federated	2.38 ± 0.10	0.31	0.69	0.00
$\kappa = 0$ single	2.38 ± 0.11	0.31	0.69	0.08
$\kappa = 1$ federated	$\mathbf{2.62 \pm 0.11}$	0.29	0.75	0.03
$\kappa = 1$ single	2.58 ± 0.11	0.29	0.75	0.12
$\kappa = 2$ federated	2.54 ± 0.15	0.29	0.73	0.15
$\kappa = 2$ single	2.43 ± 0.14	0.28	0.70	0.35

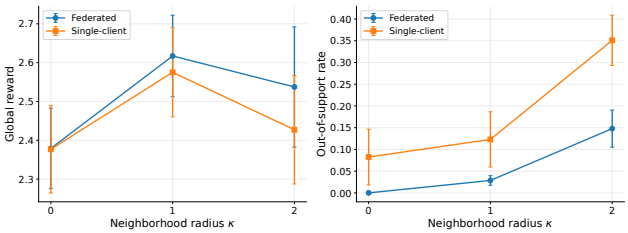

Figure 1. Effect of κ. Reward is non-monotone with optimum at $\kappa = 1$; OOS rises with κ. Federation reduces OOS for $\kappa > 0$.

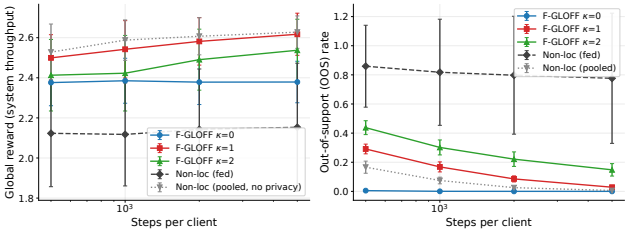

Figure 2. Reward and OOS vs. steps/client. F-GLOFF ($\kappa = 1$) approaches the raw-pooled oracle with more data, while the non-localized variant stays high OOS.

Evaluation setup. Logs are generated by four rule-based behavior policies (`uniform`, `best_q`, `round_robin`, `defer_heavy`) and partitioned across clients by balanced spatial k-means with halo h. Reference configuration is $n = 16$ UEs, $M = 12$ APs, $K = 4$ clients, $h = 60$, budget 5000 slots/client, $\gamma = 0.95$, four random seeds. We compare F-GLOFF to the engineered rules, the logging behavior mixture, single client and behavior-distribution ablations, and non-localized variants. Full configuration, dynamics, and logging policies are in Appendix E.

7. Results

Table 1 compares engineered baselines, non-localized variants, and F-GLOFF at the reference budget. F-GLOFF with $\kappa = 1$ achieves reward 2.62 ± 0.11, matching the centralized raw pooled oracle (2.63 ± 0.08) without sharing raw transitions, and improves over the strongest engineered baseline (`round_robin`) by 12% and over the non-localized federated variant by 22%. The gain over `round_robin` comes without a higher transmit rate, so the learned policy improves AP selection rather than transmitting more aggressively. The κ-sweep exhibits the predicted bias–coverage tradeoff: $\kappa = 0$ is roughly at `round_robin` (graph-local context is needed), $\kappa = 1$ gives the best reward, and $\kappa = 2$ has higher out-of-support (OOS) risk.

7.1. Bias–coverage tradeoff

Fig. 1 shows the effect of the locality radius. Federated rewards are non-monotone in κ (2.38, 2.62, 2.54 for $\kappa = 0, 1, 2$) while OOS rises monotonically (0.000, 0.029, 0.148), consistent with Thm. 3.6: increasing κ reduces locality bias but makes offline coverage harder. The single-client variant has higher OOS at every κ, so federation provides a safer supported policy that becomes increasingly important as the local context space grows.

7.2. Sample complexity and non-localized comparison

Fig. 2 sweeps the offline data budget. F-GLOFF with $\kappa = 1$ improves from 2.50 to 2.62 as the budget grows, with OOS dropping from 0.291 to 0.029. The non-localized observability-respecting federated variant stagnates near 2.12–2.15 with high OOS at all budgets, while the raw-pooled non-localized oracle improves with data, showing that the non-localized failure is tied to the observability constraint, not to the global summary alone. Locality is therefore needed for federated offline estimation under partial observability. Clients cannot reliably estimate global context models when no client observes the full global state. Federation diagnostics, a client-count sweep over $K \in \{2, 4, 8\}$ and an IID/non-IID logging comparison, confirm that the federation benefit grows with K at $\kappa = 1$ and that F-GLOFF is stable across logging regimes; details are in Appendix F.1.

8. Conclusion

F-GLOFF matches a centralized raw-pooling oracle on multi-AP user association without sharing transitions, showing that graph locality and observability-valid federation suffice to recover centralized performance under partial observability. The supporting theory replaces global coverage with neighborhood projected coverage and decomposes the offline error into structurally distinct sources. Future work should scale the approach with graph-structured function approximation (Ren et al., 2025), adaptive κ from OOS diagnostics, and add formal privacy to the shared statistics.

9. Impact Statement

This work studies offline federated learning for wireless network control. Since the method learns from logged data and avoids online exploration, it may reduce service disruption during policy development. Potential risks include deployment under insufficient support and unfair performance across clients with different observability. We mitigate these risks through pessimistic evaluation, observability-valid aggregation, and out-of-support diagnostics.

References

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pp. 1042–1051. PMLR, 2019.

Eldeeb, E., Sifaou, H., Simeone, O., Shehab, M., and Alves, H. Conservative and risk-aware offline multi-agent reinforcement learning. *IEEE Transactions on Cognitive Communications and Networking*, 11(3):1913–1926, 2024.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

Guo, Z., Chen, Z., Liu, P., Luo, J., Yang, X., and Sun, X. Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks. *IEEE Journal on Selected Areas in Communications*, 40 (5):1587–1599, 2022. doi: 10.1109/JSAC.2022.3143251.

Hou, I.-H. and Kumar, P. Utility-optimal scheduling in time-varying wireless networks with delay constraints. In *Proceedings of the eleventh ACM international symposium on Mobile ad hoc networking and computing*, pp. 31–40, 2010.

Hu, S., Chen, X., Ni, W., Hossain, E., and Wang, X. Distributed machine learning for wireless communication networks: Techniques, architectures, and applications. *IEEE Communications Surveys & Tutorials*, 23(3):1458–1493, 2021.

Jaramillo, J. J., Srikant, R., and Ying, L. Scheduling for optimal rate allocation in ad hoc networks with heterogeneous delay constraints. *IEEE Journal on Selected Areas in Communications*, 29(5):979–987, 2011.

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

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pp. 5084–5096. PMLR, 2021.

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

Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. Federated reinforcement learning: Linear speedup under markovian sampling. In *International conference on machine learning*, pp. 10997–11057. PMLR, 2022.

Kwon, D., YeonseoJeong, Hong, S., and Hong, S. Federated model-based offline multi-agent reinforcement learning for wireless networks. In *NeurIPS 2025 Workshop: AI and ML for Next-Generation Wireless Communications and Networking*, 2025. URL https://openreview.net/forum?id=nFztuJRkJw.

Lan, G., Han, D.-J., Hashemi, A., Aggarwal, V., and Brinton, C. G. Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. *arXiv preprint arXiv:2404.08003*, 2024.

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

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

Lin, Y., Qu, G., Huang, L., and Wierman, A. Multi-agent reinforcement learning in stochastic networked systems. *Advances in neural information processing systems*, 34: 7825–7837, 2021.

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

Oliehoek, F. A., Amato, C., et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

Pan, L., Huang, L., Ma, T., and Xu, H. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International conference on machine learning*, pp. 17221–17237. PMLR, 2022.

Qu, G., Lin, Y., Wierman, A., and Li, N. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33:2074–2086, 2020.

Qu, G., Wierman, A., and Li, N. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628, 2022.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Ren, Z., Zhang, R., Dai, B., and Li, N. Scalable spectral representations for multiagent reinforcement learning in network mdps. In Li, Y., Mandt, S., Agrawal, S., and Khan, E. (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 550–558. PMLR, 03–05 May 2025.

Sana, M., De Domenico, A., Yu, W., Lostanlen, Y., and Strinati, E. C. Multi-agent reinforcement learning for adaptive user association in dynamic mmwave networks. *IEEE Transactions on Wireless Communications*, 19(10): 6520–6534, 2020.

Shao, J., Qu, Y., Chen, C., Zhang, H., and Ji, X. Counterfactual conservative q learning for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:77290–77312, 2023.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*, pp. 19967–20025. PMLR, 2022.

Singh, S. P. and Yee, R. C. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Wang, H., Mitra, A., Hassani, H., Pappas, G. J., and Anderson, J. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023a.

Wang, X., Xu, H., Zheng, Y., and Zhan, X. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. *Advances in Neural Information Processing Systems*, 36:52413–52429, 2023b.

Woo, J., Shi, L., Joshi, G., and Chi, Y. Federated offline reinforcement learning: Collaborative single-policy coverage suffices. *arXiv preprint arXiv:2402.05876*, 2024.

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

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Yan, Y., Li, G., Chen, Y., and Fan, J. The efficacy of pessimism in asynchronous q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219, 2023.

Yang, K., Shi, C., Shen, C., Yang, J., Yeh, S.-P., and Sydir, J. J. Offline reinforcement learning for wireless network optimization with mixture datasets. *IEEE Transactions on Wireless Communications*, 23(10):12703–12716, 2024a.

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

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., Yang, J., and Zhao, Q. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021.

Zhang, C., Wang, H., Mitra, A., and Anderson, J. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *arXiv preprint arXiv:2401.15273*, 2024.

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

Zhang, Y., Qu, G., Xu, P., Lin, Y., Chen, Z., and Wierman, A. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–51, 2023.

Zheng, Z., Gao, F., Xue, L., and Yang, J. Federated q-learning: Linear regret speedup with low communication cost. *arXiv preprint arXiv:2312.15023*, 2023.

Zhou, D., Zhang, Y., Sonabend-W, A., Wang, Z., Lu, J., and Cai, T. Federated offline reinforcement learning. *Journal of the American Statistical Association*, 119(548):3152–3163, 2024.

Zhu, H., Xu, J., Liu, S., and Jin, Y. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

Zhu, Z., Liu, M., Mao, L., Kang, B., Xu, M., Yu, Y., Ermon, S., and Zhang, W. Madiff: Offline multi-agent learning with diffusion models. *Advances in Neural Information Processing Systems*, 37:4177–4206, 2024.

A. Related Work

This appendix provides additional references that contextualize the framework.

Foundations of offline value-function approximation. The single-policy concentrability used in pessimistic offline RL descends from the all-policy concentrability and error propagation analyses developed for fitted value iteration and approximate policy iteration (Munos & Szepesvári, 2008; Farahmand et al., 2010), which characterize how Bellman residuals accumulate under distribution shift along the discounted occupancy. Information theoretic perspectives on the structural limits of batch RL are developed in (Chen & Jiang, 2019), complementing the Bellman-consistent viewpoint we adopted.

Federated RL beyond offline settings. A complementary line of work studies federated reinforcement learning under environmental or behavioral heterogeneity, including federated temporal-difference learning with linear function approximation (Wang et al., 2023a), federated natural policy gradient methods for multi-task RL on graph topologies (Yang et al., 2024b), federated Q-learning with regret guarantees and low communication (Zheng et al., 2023), and finite-time analysis of on-policy heterogeneous federated RL (Zhang et al., 2024).

Networked MARL extensions. The exponential decay framework (Qu et al., 2020; Lin et al., 2021; Zhang et al., 2023) employed in this paper has a journal counterpart that consolidates the scalable actor–critic results (Qu et al., 2022). Recent work has extended the κ-hop locality principle beyond tabular settings using scalable spectral local representations for network MDPs (Ren et al., 2025), suggesting a path to function approximation extensions of graph-localized offline learning.

Cooperative MARL and decentralized formulations. Cooperative MARL, including value decomposition, centralized training with decentralized execution paradigms, and game-theoretic perspectives, is broadly surveyed in (Zhang et al., 2021). The decentralized partially observable formulation that underlies federated agents with bounded observable regions is treated comprehensively in (Oliehoek et al., 2016); the per-client observability set $\mathcal{O}_k$ in our setting can be viewed as a graph-induced specialization of the local observation models studied in that framework.

B. Additional Details for the Graph-Structured Offline MARL Formulation

This appendix expands on the graph-structured offline MARL formulation in Sec. 2.

B.1. Discounted value and throughput

The theory uses the discounted value

$$V^{\pi}(\mu) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 \sim \mu\right], \qquad \gamma \in (0,1),$$

whereas the wireless experiments report average per-slot system throughput. These criteria induce the same policy ordering in stationary regimes. If a policy induces a stationary distribution with expected per-slot reward $\bar{r}^{\pi}$, then its discounted value is approximately $\bar{r}^{\pi}/(1-\gamma)$, up to transient effects from the initial distribution. Thus, when transients are negligible, maximizing discounted value is equivalent to maximizing long-run average throughput.

B.2. Wireless interpretation of the interaction-radius factorization

The symmetric one-hop dependence on $a_{\mathcal{N}_i^1}$ in Assumption 2.1 is what makes the factorization usable for wireless networks: a UE's per-slot transmission outcome at AP y depends on whether any neighboring UE on the access-set conflict graph also transmits to y in the same slot, which is a function of $a_{\mathcal{N}_i^1}$ rather than a_i alone. Under this factorization, collision indicators and contention outcomes appear directly in the instantaneous reward $r_i(s_{\mathcal{N}_i^1}, a_{\mathcal{N}_i^1})$ without needing to be folded into the next state, while the state-side dependence on $s_{\mathcal{N}_i^1}$ accommodates queue and contention indicators of neighboring agents.

B.3. Latent states and abstracted local contexts

Assumption 2.2 posits a finite local context $x_i^{\kappa} = \phi_i^{\kappa}(s_{\mathcal{N}_i^{\kappa}})$ and a projected transition $P_i^{\kappa}(x' \mid x_i^{\kappa}, a_{\mathcal{N}_i^1})$ such that the conditional law of $\phi_i^{\kappa}(s'_{\mathcal{N}_i^{\kappa}})$ given (s,a) depends on (s,a) only through $(x_i^{\kappa}, a_{\mathcal{N}_i^1})$. We clarify the scope of this idealization.

Under Assumption 2.1, each component s'_j for $j \in \mathcal{N}_i^\kappa$ depends on $(s_{\mathcal{N}_j^1}, a_{\mathcal{N}_j^1})$, and since $\mathcal{N}_j^1 \subseteq \mathcal{N}_i^{\kappa+1}$, the joint next-step state $s'_{\mathcal{N}_i^\kappa}$ is a function of $(s_{\mathcal{N}_i^{\kappa+1}}, a_{\mathcal{N}_i^{\kappa+1}})$ alone. Strict Markov-sufficiency of $(x_i^\kappa, a_{\mathcal{N}_i^1})$ therefore requires the abstraction ϕ_i^κ to absorb the influence of the boundary state $s_{\mathcal{N}_i^{\kappa+1} \setminus \mathcal{N}_i^\kappa}$ and of out-of-radius actions $a_{\mathcal{N}_i^{\kappa+1} \setminus \mathcal{N}_i^1}$. Assumption 2.2 is the idealized setting in which this absorption is exact; the residual boundary dependence is of order $\rho^{\kappa+1}$ under the exponential decay condition of Sec. 3, matching the locality bias already absorbed elsewhere in the analysis.

In the multi-AP user-association experiment, ϕ_i^κ retains the agent's queue state and capped local contention measures around its accessible APs; under the simulator's slot-level dynamics, these statistics carry the relevant information for one-step queue and access evolution up to the load cap $\ell_{\max}$. The framework treats this context as the operational local state, while the theory describes the idealized Markov setting.

B.4. Theoretical and operational projected spaces

Graph locality changes the relevant notion of support. Under Assumptions 2.1 and 2.2, evaluating a κ-local policy only requires coverage of neighborhood-level context-action pairs at the per-agent action resolution. We distinguish two related projected spaces:

- The *theoretical* projected space at the interaction-radius resolution,

$$\mathcal{Z}_i^{\kappa,\text{theory}} = \mathcal{X}_i^\kappa \times \mathcal{A}_{\mathcal{N}_i^1}, \qquad \mathcal{A}_{\mathcal{N}_i^1} = \prod_{j \in \mathcal{N}_i^1} \mathcal{A}_j,$$

  on which Assumption 2.2 states that the projected transition $P_i^\kappa(\cdot \mid x_i^\kappa, a_{\mathcal{N}_i^1})$ is a property of the environment.

- The *operational* projected space at the per-agent action resolution,

$$\mathcal{Z}_i^\kappa = \mathcal{X}_i^\kappa \times \mathcal{A}_i,$$

  which is the resolution at which F-GLOFF actually estimates the local model and runs Bellman backups.

The empirical model is estimated from data tuples $(x_i^\kappa, a_i, r_i, x_i^{\kappa\prime})$ at the per-agent resolution; conditional on (x_i^κ, a_i), the joint neighbor profile $a_{\mathcal{N}_i^1 \setminus \{i\}}$ is integrated out under behavior, yielding the per-agent population MDP $\mathcal{M}_{i,\kappa,\bar{\mu}}$ of Sec. 3. The two projections are connected by behavior-marginalization.

B.5. Federated observability and admissible clients

The federated setup with client-owned agent sets $_k \subseteq \mathcal{O}_k$, observable regions $\mathcal{O}_k$, and observability valid client set $\mathcal{R}_i^\kappa = \{k \in [K] : \mathcal{N}_i^\kappa \subseteq \mathcal{O}_k, \mathcal{D}_{k,i}^\kappa \neq \emptyset\}$. We add two clarifications.

Action observability. The data tuples $(x_i^\kappa, a_i, r_i, x_i^{\kappa\prime})$ require the client to observe the radius-κ neighborhood states (to construct x_i^κ and $x_i^{\kappa\prime}$) and the focal agent's own action; observability of one-hop neighbor actions $a_{\mathcal{N}_i^1 \setminus \{i\}}$ is not required, because the algorithm marginalizes over them implicitly via the per-agent counting at the (x_i^κ, a_i) resolution. This prevents observability leakage in simulation: a deployed client should not contribute statistics involving variables it would not observe, even if the simulator stores the full global trajectory.

Monotonicity in κ. Two opposing monotonicities drive the bias–coverage tradeoff:

- $|\mathcal{R}_i^\kappa|$ is *monotone non-increasing* in κ: as κ grows, $\mathcal{N}_i^\kappa$ grows, so the inclusion $\mathcal{N}_i^\kappa \subseteq \mathcal{O}_k$ is more restrictive and fewer clients qualify.
- $|\mathcal{X}_i^\kappa|$ is *monotone non-decreasing* in κ: a larger neighborhood admits more distinct contexts, since ϕ_i^κ is determined by states over $\mathcal{N}_i^\kappa$.

Increasing κ therefore both enlarges the local context space (worsening per-cell coverage) and reduces the number of admissible clients (reducing total per-cell samples). Federation is useful precisely because it pools valid local statistics across all clients in $\mathcal{R}_i^\kappa$, which mitigates the second of these effects whenever multiple clients are observability-valid.

C. Additional Details for the Localized Offline Learning Theory

This appendix gives the supporting derivations for Sec. 3: the truncation lemma, the localized policy-difference lemma and local-to-global decomposition, the proofs of the localization and decentralization gaps, and the application of Bellman-consistent pessimistic offline RL bounds to per-agent population MDPs.

C.1. Neighborhood truncation

Lemma C.1 (Neighborhood truncation of Q_i^π)**.** *Under Assumption 3.1, for every policy $\pi \in \Pi_{\mathrm{stat}}$, every agent i, and every $\ell \geq 0$, there exists a function*

$$\widetilde{Q}_{i,\ell}^{\pi} : \mathcal{S}_{\mathcal{N}_i^\ell} \times \mathcal{A}_{\mathcal{N}_i^\ell} \to \mathbb{R}$$

such that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| Q_i^\pi(s,a) - \widetilde{Q}_{i,\ell}^{\pi}(s_{\mathcal{N}_i^\ell}, a_{\mathcal{N}_i^\ell}) \right| \leq c_{\mathrm{loc}} \rho^{\ell+1}.$$

Proof. Fix an arbitrary reference assignment $(s^{\mathrm{ref}}, a^{\mathrm{ref}})$ on $(\mathcal{N}_i^\ell)^c$. For each local pair (x, u), define

$$\widetilde{Q}_{i,\ell}^{\pi}(x,u) = Q_i^\pi\Big((x, s^{\mathrm{ref}}_{(\mathcal{N}_i^\ell)^c}), (u, a^{\mathrm{ref}}_{(\mathcal{N}_i^\ell)^c})\Big).$$

For any global pair (s, a), the pair (s, a) and its reference completion agree on $\mathcal{N}_i^\ell$. Assumption 3.1 then gives the desired bound. □

C.2. Localized policy-difference lemma and local-to-global decomposition

The following lemma is the discounted analog of the bounded difference results of (Qu et al., 2020) and (Lin et al., 2021) adapted to our setting.

Lemma C.2 (Localized policy-difference)**.** *Under Assumption 3.1, for any pair of policies $\pi, \pi' \in \Pi_{\mathrm{stat}}$ that agree on every agent in $\mathcal{N}_i^\kappa$ (i.e., $\pi_j(\cdot \mid \cdot) = \pi'_j(\cdot \mid \cdot)$ for all $j \in \mathcal{N}_i^\kappa$),*

$$\left| V_i^\pi(s) - V_i^{\pi'}(s) \right| \leq \frac{c_{\mathrm{loc}} \rho^{\kappa+1}}{1-\gamma}, \qquad \forall s \in \mathcal{S}.$$

Proof. The argument uses a coupling between the trajectories of π and π' starting from $s_0 = s$. Inside $\mathcal{N}_i^\kappa$, the two policies agree by hypothesis, so the per-step actions on $\mathcal{N}_i^\kappa$ can be coupled to be identical as long as the trajectory's state on $\mathcal{N}_i^\kappa$ is the same. Coordinates outside $\mathcal{N}_i^\kappa$ may evolve differently under π and π', but their effect on agent i's per-step reward is governed by Assumption 3.1: since r_i depends on $a_{\mathcal{N}_i^1}$ only, variation of agent i's value across coordinates outside $\mathcal{N}_i^\ell$ for $\ell \geq \kappa + 1$ is bounded by $c_{\mathrm{loc}} \rho^{\kappa+1}$ at each step. Summing the discounted contributions yields the $1/(1-\gamma)$ factor. □

Behavior-comparator near-neighbor closeness. Lemma C.2 controls the value gap between policies that agree on $\mathcal{N}_i^\kappa$, but the corollary below requires comparing per-agent values when policies on the one-hop interaction neighborhood $\mathcal{N}_i^1 \setminus \{i\}$ change from comparator-induced to behavior-induced. This change is not controlled by Assumption 3.1 alone, since exponential decay only bounds variation across coordinates outside $\mathcal{N}_i^\ell$ for $\ell \geq \kappa + 1$. We therefore invoke Assumption 3.3 from the main text, which bounds in TV by τ the joint distribution of one-hop neighbor actions induced by $\pi \in \{\pi_\kappa^\star, \bar{\pi}_\kappa^\star, \widehat{\pi}_\kappa\}$ relative to the behavior conditional $\bar{\nu}_i(\cdot \mid x, a_i)$. The discussion of τ in Sec. 3 extends to a small forward note: condition (i) there (factorized behavior over $\mathcal{N}_i^1$) is achievable whenever $\bar{\mu} \in \Pi_0 \subseteq \Pi_\kappa$, in which case $\bar{\nu}_i$ inherits the factorization. When τ is small the localization-bias picture dominates; when τ is large, best response iteration (Appendix D.1) is the principled remedy.

Corollary C.3 (Local-to-global decomposition)**.** *Under Assumptions 3.1 and 3.3, for any $\pi \in \{\bar{\pi}_\kappa^\star, \widehat{\pi}_\kappa\}$,*

$$\left| V^\pi(\mu) - \sum_{i=1}^n V_{i,\kappa,\bar{\mu}}^{\pi_i}(\mu_i) \right| \leq \frac{n c_{\mathrm{loc}} \rho^{\kappa+1}}{1-\gamma} + \frac{2n R_{\max}\, \tau}{(1-\gamma)^2}.$$

Proof. Fix agent i. Decompose

$$V_i^{\pi}(\mu) - V_{i,\kappa,\bar{\mu}}^{\pi_i}(\mu_i) = \underbrace{\left[V_i^{\pi}(\mu) - V_i^{\pi^{(b)}}(\mu)\right]}_{\text{(I): near-neighbor shift}} + \underbrace{\left[V_i^{\pi^{(b)}}(\mu) - V_{i,\kappa,\bar{\mu}}^{\pi_i}(\mu_i)\right]}_{\text{(II): outside-}\mathcal{N}_i^{\kappa}\text{ truncation}},$$

where $\pi^{(b)}$ is the auxiliary stationary policy that uses π_i on the focal agent, draws the joint $a_{\mathcal{N}_i^1\setminus\{i\}}$ from $\bar{\nu}_i(\cdot \mid x_i^{\kappa}, a_i)$ on the one-hop interaction neighborhood at each step, and uses π_j for $j \notin \mathcal{N}_i^1$.

Term (II): outside-$\mathcal{N}_i^{\kappa}$ truncation. $\pi^{(b)}$ and the auxiliary policy that defines $V_{i,\kappa,\bar{\mu}}^{\pi_i}$ agree on $\mathcal{N}_i^{\kappa}$: both use π_i on $\{i\}$, both use $\bar{\nu}_i$ on $\mathcal{N}_i^1 \setminus \{i\}$, and both lie in Π_{stat} (extending appropriately on $\mathcal{N}_i^{\kappa} \setminus \mathcal{N}_i^1$). They differ only on $\mathcal{N} \setminus \mathcal{N}_i^{\kappa}$. By Lemma C.2, $|V_i^{\pi^{(b)}}(\mu) - V_{i,\kappa,\bar{\mu}}^{\pi_i}(\mu_i)| \le c_{\mathrm{loc}}\rho^{\kappa+1}/(1-\gamma)$.

Term (I): near-neighbor shift. π and $\pi^{(b)}$ agree on $\{i\} \cup (\mathcal{N} \setminus \mathcal{N}_i^1)$ and differ only on $\mathcal{N}_i^1 \setminus \{i\}$. Per step, agent i's expected reward differs by at most $2R_{\max}\,\tau$, since r_i is bounded by $R_{\max}$ and the joint distribution of $a_{\mathcal{N}_i^1\setminus\{i\}}$ differs by at most τ in TV by Assumption 3.3. The per-step transition kernel of s_i similarly differs by at most τ in TV (since P_i depends on $a_{\mathcal{N}_i^1}$); a standard performance-difference argument under TV-shift (Kakade & Langford, 2002; Rashidinejad et al., 2021) gives $|V_i^{\pi}(\mu) - V_i^{\pi^{(b)}}(\mu)| \le 2R_{\max}\,\tau/(1-\gamma)^2$.

□

C.3. Proof of the localization gap

Proof of Proposition 3.2. Set $\epsilon := nc_{\mathrm{loc}}\rho^{\kappa+1}$. Apply Lemma C.1 to $\pi^\star$ at radius κ: for each agent i, there exists a truncated function $\widetilde{Q}_{i,\kappa}^{\pi^\star} : \mathcal{S}_{\mathcal{N}_i^{\kappa}} \times \mathcal{A}_{\mathcal{N}_i^{\kappa}} \to \mathbb{R}$ with

$$\sup_{(s,a)} \left| Q_i^{\pi^\star}(s,a) - \widetilde{Q}_{i,\kappa}^{\pi^\star}(s_{\mathcal{N}_i^{\kappa}}, a_{\mathcal{N}_i^{\kappa}}) \right| \le c_{\mathrm{loc}}\rho^{\kappa+1}.$$

Define $\widetilde{Q}^{\pi^\star}(s,a) := \sum_{i=1}^n \widetilde{Q}_{i,\kappa}^{\pi^\star}(s_{\mathcal{N}_i^{\kappa}}, a_{\mathcal{N}_i^{\kappa}})$, so that $\|Q^{\pi^\star} - \widetilde{Q}^{\pi^\star}\|_\infty \le \epsilon$.

Each summand of $\widetilde{Q}^{\pi^\star}$ depends on $a_{\mathcal{N}_i^{\kappa}}$, so the joint argmax of $\widetilde{Q}^{\pi^\star}$ over $\mathcal{A}$ does not factor agent-wise in general. By the factorized-policy construction of (Qu et al., 2020) applied to the truncated $\widetilde{Q}^{\pi^\star}$ (see also (Lin et al., 2021; Zhang et al., 2023)), there exists a factorized policy $\bar{\pi}_\kappa \in \Pi_\kappa$ such that

$$\widetilde{Q}^{\pi^\star}(s, \pi^\star(s)) \le \widetilde{Q}^{\pi^\star}(s, \bar{\pi}_\kappa(s)) \quad \text{for every } s \in \mathcal{S}, \tag{2}$$

i.e., $\bar{\pi}_\kappa$ is at least as good as $\pi^\star$ when evaluated on the truncated function pointwise. We apply the one-step approximate greedification argument (Singh & Yee, 1994). For every $s \in \mathcal{S}$,

$$\begin{aligned} V^{\pi^\star}(s) - V^{\bar{\pi}_\kappa}(s) &= Q^{\pi^\star}(s,\pi^\star(s)) - Q^{\bar{\pi}_\kappa}(s,\bar{\pi}_\kappa(s)) \\ &= \underbrace{\left[Q^{\pi^\star}(s,\pi^\star(s)) - \widetilde{Q}^{\pi^\star}(s,\pi^\star(s))\right]}_{\le\epsilon\text{ (truncation)}} + \underbrace{\left[\widetilde{Q}^{\pi^\star}(s,\pi^\star(s)) - \widetilde{Q}^{\pi^\star}(s,\bar{\pi}_\kappa(s))\right]}_{\le 0\text{ by (2)}} \\ &\quad + \underbrace{\left[\widetilde{Q}^{\pi^\star}(s,\bar{\pi}_\kappa(s)) - Q^{\pi^\star}(s,\bar{\pi}_\kappa(s))\right]}_{\le\epsilon\text{ (truncation)}} + \underbrace{\left[Q^{\pi^\star}(s,\bar{\pi}_\kappa(s)) - Q^{\bar{\pi}_\kappa}(s,\bar{\pi}_\kappa(s))\right]}_{(\star)}. \end{aligned}$$

The last term satisfies

$$(\star) = \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,\bar{\pi}_\kappa(s))}\left[V^{\pi^\star}(s') - V^{\bar{\pi}_\kappa}(s')\right] \le \gamma\left\|V^{\pi^\star} - V^{\bar{\pi}_\kappa}\right\|_\infty,$$

where the equality is the standard Q-difference identity and the inequality uses $V^{\pi^\star}(s') \ge V^{\bar{\pi}_\kappa}(s')$ pointwise (since $\pi^\star$ is optimal). Combining,

$$V^{\pi^\star}(s) - V^{\bar{\pi}_\kappa}(s) \le 2\epsilon + \gamma\left\|V^{\pi^\star} - V^{\bar{\pi}_\kappa}\right\|_\infty.$$

Taking sup over s gives $\|V^{\pi^\star} - V^{\bar{\pi}_\kappa}\|_\infty \le 2\epsilon/(1-\gamma)$, and therefore

$$V^{\pi^\star}(\mu) - V^{\bar{\pi}_\kappa}(\mu) \le \left\|V^{\pi^\star} - V^{\bar{\pi}_\kappa}\right\|_\infty \le \frac{2\epsilon}{1-\gamma} = \frac{2nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma}.$$

Since $\pi^\star_\kappa = \arg\max_{\pi\in\Pi_\kappa} V^\pi(\mu)$ and $\bar{\pi}_\kappa \in \Pi_\kappa$, we have $V^{\pi^\star_\kappa}(\mu) \ge V^{\bar{\pi}_\kappa}(\mu)$, completing the proof. □

C.4. Proof of the decentralization gap

Proof of Proposition 3.4. The argument has the same structure as Proposition 3.2: truncation followed by one-step approximate-greedification. The new ingredient is that the comparator $\bar{\pi}^\star_\kappa$ is per-agent optimal on the population MDPs $\{\mathcal{M}_{i,\kappa,\bar{\mu}}\}_i$ rather than the joint Π_κ-optimum. We use Corollary C.3 to relate the global value of policies in Π_κ to per-agent population values, this is where Assumption 3.3 enters.

Truncation: Set $\epsilon := nc_{\mathrm{loc}}\rho^{\kappa+1}$. Apply Lemma C.1 to $\pi^\star_\kappa \in \Pi_\kappa \subseteq \Pi_{\mathrm{stat}}$ at radius κ: there exist truncated functions $\widetilde{Q}^{\pi^\star_\kappa}_{i,\kappa}$ such that $\|Q^{\pi^\star_\kappa} - \widetilde{Q}^{\pi^\star_\kappa}\|_\infty \le \epsilon$, where $\widetilde{Q}^{\pi^\star_\kappa}(s,a) := \sum_i \widetilde{Q}^{\pi^\star_\kappa}_{i,\kappa}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})$.

Global value of $\bar{\pi}^\star_\kappa$ and $\pi^\star_\kappa$ via per-agent population MDPs: By Corollary C.3 applied to $\bar{\pi}^\star_\kappa \in \Pi_\kappa$ (and using Assumption 3.3, since $\bar{\pi}^\star_\kappa$ is one of the policies covered by Assumption 3.3), the global value of $\bar{\pi}^\star_\kappa$ is approximated by the sum of its per-agent population-MDP values:

$$\left|V^{\bar{\pi}^\star_\kappa}(\mu) - \sum_{i=1}^n V^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i)\right| \le \frac{nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{2nR_{\max}\tau}{(1-\gamma)^2}.$$

The same bound applies to $\pi^\star_\kappa$ (also covered by Assumption 3.3): writing each $\pi^\star_{\kappa,i}$ for the i-th component of the joint optimum,

$$\left|V^{\pi^\star_\kappa}(\mu) - \sum_{i=1}^n V^{\pi^\star_{\kappa,i}}_{i,\kappa,\bar{\mu}}(\mu_i)\right| \le \frac{nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{2nR_{\max}\tau}{(1-\gamma)^2}.$$

Per-agent optimality of $\bar{\pi}^\star_{i,\kappa}$: By definition, $\bar{\pi}^\star_{i,\kappa}$ is the optimal policy in $\mathcal{M}_{i,\kappa,\bar{\mu}}$, so

$$V^{\pi^\star_{\kappa,i}}_{i,\kappa,\bar{\mu}}(\mu_i) \le V^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i), \quad \forall i.$$

Summing over agents,

$$\sum_{i=1}^n V^{\pi^\star_{\kappa,i}}_{i,\kappa,\bar{\mu}}(\mu_i) \le \sum_{i=1}^n V^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i).$$

Combining the bounds from Steps 2 and 3,

$$\begin{aligned}
V^{\pi^\star_\kappa}(\mu) - V^{\bar{\pi}^\star_\kappa}(\mu) &\le \sum_i V^{\pi^\star_{\kappa,i}}_{i,\kappa,\bar{\mu}}(\mu_i) + \frac{nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{2nR_{\max}\tau}{(1-\gamma)^2} \\
&\quad - \Big(\sum_i V^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i) - \frac{nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} - \frac{2nR_{\max}\tau}{(1-\gamma)^2}\Big) \\
&\le \underbrace{\Big(\sum_i V^{\pi^\star_{\kappa,i}}_{i,\kappa,\bar{\mu}}(\mu_i) - \sum_i V^{\bar{\pi}^\star_{i,\kappa}}_{i,\kappa,\bar{\mu}}(\mu_i)\Big)}_{\le 0 \text{ by Step 3}} + \frac{2nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{4nR_{\max}\tau}{(1-\gamma)^2} \\
&\le \frac{2nc_{\mathrm{loc}}\rho^{\kappa+1}}{1-\gamma} + \frac{4nR_{\max}\tau}{(1-\gamma)^2}.
\end{aligned}$$

□

C.5. Bellman-consistent pessimistic estimation on the per-agent population MDP

For each agent i and radius κ, the projected federated data define an empirical version of the per-agent population MDP $\mathcal{M}_{i,\kappa,\bar{\mu}}$. Specifically, the empirical model $(\widehat{r}_i^{\kappa}, \widehat{P}_i^{\kappa})$ from Sec. 3 is an unbiased count-based estimator of $(r_{i,\bar{\mu}}^{\kappa}, P_{i,\bar{\mu}}^{\kappa})$, since the data are sampled from the joint behavior policy and conditional on $(x_t, a_{i,t})$ the neighbor profile $a_{-i,t}^{(1)}$ is distributed according to $\bar{\nu}_i$ by definition of conditional expectation.

Applying the Bellman-consistent pessimistic offline RL guarantee of (Xie et al., 2021) to each per-agent problem on $\mathcal{M}_{i,\kappa,\bar{\mu}}$ with the tabular function class $\mathcal{F}_i^{\kappa}$, the localized Bellman-consistent coefficient $C_{\kappa,\mathrm{BE}}^{\star}$ from Definition 3.5, and the single-policy comparator $\bar{\pi}_{i,\kappa}^{\star}$, gives, with probability at least $1-\delta/n$,

$$V_{i,\kappa,\bar{\mu}}^{\bar{\pi}_{i,\kappa}^{\star}}(\mu_i) - V_{i,\kappa,\bar{\mu}}^{\widehat{\pi}_{i,\kappa}}(\mu_i) \le C_1 \sqrt{\frac{C_{\kappa,\mathrm{BE}}^{\star}\,|\mathcal{X}_i^{\kappa}|\,|\mathcal{A}_i|\,\log(n/\delta)}{(1-\gamma)^3\,N_{i,\kappa}}},$$

where C_1 is an absolute constant. The single-agent counterpart of this bound, under asynchronous data collection with single policy concentrability, is treated in Yan et al. (2023). Union-bounding over agents and using $|\mathcal{X}_i^{\kappa}| \le S_{\kappa}$, $|\mathcal{A}_i| \le A_{\max}$, $N_{i,\kappa} \ge N_{\kappa}$ gives

$$\sum_{i=1}^{n}\Big(V_{i,\kappa,\bar{\mu}}^{\bar{\pi}_{i,\kappa}^{\star}}(\mu_i) - V_{i,\kappa,\bar{\mu}}^{\widehat{\pi}_{i,\kappa}}(\mu_i)\Big) \le C_1\, n \sqrt{\frac{C_{\kappa,\mathrm{BE}}^{\star}\,S_{\kappa}\,A_{\max}\,\log(n/\delta)}{(1-\gamma)^3\,N_{\kappa}}}. \tag{3}$$

Local-to-global passage. It remains to relate the LHS of (3), which is in terms of per-agent population values, to the global value gap $V^{\bar{\pi}_{\kappa}^{\star}}(\mu) - V^{\widehat{\pi}_{\kappa}}(\mu)$. Both $\bar{\pi}_{\kappa}^{\star}$ and $\widehat{\pi}_{\kappa}$ are covered by Assumption 3.3, so by Corollary C.3,

$$\Big|V^{\bar{\pi}_{\kappa}^{\star}}(\mu) - \sum_i V_{i,\kappa,\bar{\mu}}^{\bar{\pi}_{i,\kappa}^{\star}}(\mu_i)\Big| \le \frac{n c_{\mathrm{loc}} \rho^{\kappa+1}}{1-\gamma} + \frac{2nR_{\max}\,\tau}{(1-\gamma)^2},$$

$$\Big|V^{\widehat{\pi}_{\kappa}}(\mu) - \sum_i V_{i,\kappa,\bar{\mu}}^{\widehat{\pi}_{i,\kappa}}(\mu_i)\Big| \le \frac{n c_{\mathrm{loc}} \rho^{\kappa+1}}{1-\gamma} + \frac{2nR_{\max}\,\tau}{(1-\gamma)^2}.$$

Combining with (3),

$$V^{\bar{\pi}_{\kappa}^{\star}}(\mu) - V^{\widehat{\pi}_{\kappa}}(\mu) \le \frac{2n c_{\mathrm{loc}} \rho^{\kappa+1}}{1-\gamma} + \frac{4nR_{\max}\,\tau}{(1-\gamma)^2} + C_1\, n \sqrt{\frac{C_{\kappa,\mathrm{BE}}^{\star}\,S_{\kappa}\,A_{\max}\,\log(n/\delta)}{(1-\gamma)^3\,N_{\kappa}}}.$$

Adding the localization gap (Proposition 3.2, $\le 2nc_{\mathrm{loc}}\rho^{\kappa+1}/(1-\gamma)$), the decentralization gap (Proposition 3.4, $\le 2nc_{\mathrm{loc}}\rho^{\kappa+1}/(1-\gamma) + 4nR_{\max}\tau/(1-\gamma)^2$), and the per-agent estimation bound combined with the local-to-global passage above yields

$$V^{\pi^{\star}}(\mu) - V^{\widehat{\pi}_{\kappa}}(\mu) \le \frac{6n c_{\mathrm{loc}} \rho^{\kappa+1}}{1-\gamma} + \frac{8nR_{\max}\,\tau}{(1-\gamma)^2} + C_1\, n \sqrt{\frac{C_{\kappa,\mathrm{BE}}^{\star}\,S_{\kappa}\,A_{\max}\,\log(n/\delta)}{(1-\gamma)^3\,N_{\kappa}}},$$

matching the bound stated in Theorem 3.6.

C.6. Federated effective count and on-policy aggregation

Federation modifies only the estimation term of Theorem 3.6. The effect on the federated coefficient $C_{\kappa,\mathrm{BE,fed}}^{\star}$ relative to its single-client counterpart is described in Sec. 4.2; we record the explicit single-client effective count for comparison,

$$N_{i,\kappa}^{\mathrm{single}} = N_{i,\kappa}^{(\mathrm{owner}(i))}\mathbf{1}[\mathrm{owner}(i) \in \mathcal{R}_i^{\kappa}],$$

versus the federated $N_{i,\kappa}^{\mathrm{fed}} = \sum_{k \in \mathcal{R}_i^{\kappa}} N_{k,i}^{\kappa}$. The variance reduction $N_{i,\kappa}^{\mathrm{fed}} \ge N_{i,\kappa}^{\mathrm{single}}$ is the unconditional benefit of federation. On top of this, in non-IID regimes heterogeneous clients can also expand the support of $\bar{d}_{i,\kappa}^{\mathrm{fed}}$ and thereby shrink $C_{\kappa,\mathrm{BE,fed}}^{\star}$.

On-policy aggregation. Theorem 3.6 gives a worst-case error bound, while the algorithm aggregates per-state values using the candidate policy's on-policy stationary distribution. The uniform bound dominates any expectation over local states, including the expectation under the candidate policy's induced distribution, so the guarantee remains compatible with on-policy aggregation.

D. Additional Details for F-GLOFF

This appendix gives implementation details for F-GLOFF.

Algorithm variants. The reference method aggregates projected statistics over all clients in $\mathcal{R}_i^\kappa$ and ranks candidates using on-policy aggregation under d^{π_θ}. We compare against the following variants:

- *Single-client.* Uses only $\{\mathrm{owner}(i)\} \cap \mathcal{R}_i^\kappa$ and falls back to a conservative default when the owner is not observability-valid.
- *Non-localized observability respecting.* Uses a global summary and admits a client only if it observes the required global information.
- *Raw-pooled.* Ignores observability and serves as a centralized oracle rather than a valid federated method.

Restricted policy class. F-GLOFF uses the four-parameter softmax policy class of Sec. 5 rather than a fully tabular policy with $|\mathcal{X}_i^\kappa|(|\mathcal{A}_i| - 1)$ parameters per agent. The restriction acts as an implicit regularizer that improves stability under offline optimization with sparse coverage. The resulting finite grid Θ is task-specific; we make no claim of universal optimality.

Pessimistic Bellman backup. For a fixed candidate θ, write $\pi_\theta = \pi_{i,\theta}$. The pessimistic Bellman operator on the per-agent population MDP $\mathcal{M}_{i,\kappa,\bar{\mu}}$ uses the empirical model $(\widehat{P}_i^\kappa, \widehat{r}_{i,\mathrm{pess}}^\kappa)$ on supported cells and routes the unsupported-action mass to a self-loop with reward $-b_{\max}$:

$$(\mathcal{T}_\theta V)(x) = \sum_{a: N_i^\kappa(x,a) \geq 1} \pi_\theta(a \mid x) \left[\widehat{r}_{i,\mathrm{pess}}^\kappa(x,a) + \gamma \sum_{x'} \widehat{P}_i^\kappa(x' \mid x, a) V(x') \right]$$

$$+ \left(1 - \sum_{a: N_i^\kappa(x,a) \geq 1} \pi_\theta(a \mid x) \right) \left(-b_{\max} + \gamma V(x) \right).$$

F-GLOFF iterates $\mathcal{T}_\theta$ for H steps from $V_0 \equiv 0$ to obtain $V_i^{\theta,H}$.

On-policy aggregation and OOS. The empirical policy-induced kernel under π_θ is

$$\widehat{P}_i^{\pi_\theta}(x' \mid x) = \sum_a \pi_\theta(a \mid x) \widehat{P}_i^\kappa(x' \mid x, a),$$

with unsupported action mass routed to the fallback self-loop as above. For numerical stability, the kernel is mixed with a small uniform component,

$$\widehat{P}_\varepsilon = (1 - \varepsilon) \widehat{P}_i^{\pi_\theta} + \varepsilon / |\mathcal{X}_i^\kappa|,$$

and d^{π_θ} is approximated by power iteration on $\widehat{P}_\varepsilon$. The candidate score $V_i^{\mathrm{lb}}(\theta)$ from Sec. 5 uses this approximation.

The deployment time OOS diagnostic of Sec. 4.3 admits the per-agent estimator

$$\widehat{\mathrm{OOS}}_i(\theta) = \sum_x d^{\pi_\theta}(x) \sum_{a: N_i^\kappa(x,a) < n_{\min}} \pi_\theta(a \mid x).$$

D.1. Inter-agent coordination and the decentralization gap

F-GLOFF selects $\widehat{\theta}_i = \arg\max_{\theta \in \Theta} V_i^{\mathrm{lb}}(\theta)$ independently for each agent and deploys $\widehat{\pi}_\kappa = \bigotimes_i \pi_{i,\widehat{\theta}_i}$, with population level comparator $\bar{\pi}_\kappa^\star = \bigotimes_i \bar{\pi}_{i,\kappa}^\star \in \Pi_\kappa$. The decentralization gap between $\pi_\kappa^\star$ and $\bar{\pi}_\kappa^\star$ is bounded by Proposition 3.4.

The restricted class Θ adds a further in-class approximation when $\bar{\pi}_{i,\kappa}^\star$ is not exactly representable; we fold this into the per-agent offline-RL bound in $\mathcal{M}_{i,\kappa,\bar{\mu}}$ alongside the pessimism penalty, following the policy-class restriction handling of (Xie et al., 2021).

Table 2. Default configuration parameters.

Parameter	Value	Parameter	Value
UEs n	16	APs M	12
Area $L \times L$	100×100	Range R	32
Arrival rates λ_i	$\mathrm{Unif}[0.10, 0.30]$	AP service rates η_y	$\mathrm{Unif}[0.60, 0.95]$
Deadline d	2	Clients K	4
Default halo radius	60	Data budgets	$\{500, 1000, 2000, 5000\}$
Reference budget	5000 slots/client	Discount γ	0.95
OPE horizon H	50	Power steps T_d	50
Mixing ε	0.01	Load cap $\ell_{\max}$	2
Policy candidates $\lvert\Theta\rvert$	108	Evaluation rollout	2000 slots
Seeds	$\{0, 1, 2, 3\}$		

In our setup all four logging policies are radius-0 factorized, so $\bar{\nu}_i$ is a product distribution and Assumption 3.3 holds with small τ. Empirically, additional best-response rounds did not change selected policies, consistent with the small-τ regime in which the locality term of Proposition 3.4 dominates. In denser networks or under more aggressive comparator policies, the decentralization gap can become significant through both the locality term (less geometric decay at finite κ) and the near-neighbor shift term (τ no longer small), and best response iteration is the principled remedy.

Computational complexity and choice of radius. For a fixed agent and candidate, the Bellman backup costs $O(H|\mathcal{X}_i^\kappa||\mathcal{A}_i|)$ in supported cells with sparse transitions. Power iteration costs $O(T_d|\mathcal{X}_i^\kappa|^2)$ in the dense case and less with sparse kernels. Across agents and candidates, the total cost is approximately

$$O\big(n|\Theta|\,\big[H|\mathcal{X}_i^\kappa||\mathcal{A}_i| + T_d|\mathcal{X}_i^\kappa|^2\big]\big)\,.$$

The radius κ is the main complexity driver. Larger radii reduce localization bias but enlarge $\mathcal{X}_i^\kappa$, reduce per-cell support, and increase OOS risk.

E. Additional Details for the Wireless Instantiation

This appendix supplements Sec. 6 with the experimental specification: task choice rationale, queue and arrival mechanics, the configuration table, logging policies, client partitioning, baselines, and reported metrics.

Task choice. We instantiate F-GLOFF on the multi-AP user association rather than single-channel distributed channel access. The multi-AP association has a sparse interaction structure because UEs with disjoint access sets do not interact, while UEs with overlapping access sets contend through shared APs. The Single-channel DCA places all transmitters in one collision domain and does not provide a meaningful sparse graph. Topologies that produce a UE with an empty access set are regenerated; AP service probabilities η_y and UE arrival rates λ_i are sampled from the uniform ranges in Table 2.

Arrival and queue mechanics. We specify the arrival and queue model used in the simulator. In each slot, UE i receives a new packet with probability λ_i (Bernoulli arrivals); the packet is inserted at the maximum deadline and is dropped if the corresponding queue slot is full. At the end of each slot, all packet deadlines decrement by one and any packet at deadline zero is discarded. The queue state $q_i \in \{0, 1, \ldots, 2^d - 1\}$ encodes the d-slot deadline buffer as a binary tuple of length d. The system reward in a slot is the total number of successful transmissions. This formulation follows the standard model of deadline-constrained wireless scheduling (Hou & Kumar, 2010; Jaramillo et al., 2011).

Default configuration. Table 2 summarizes the reference configuration. The total arrival rate is approximately $\sum_i \lambda_i \approx 3.3$ packets per slot, below the aggregate service capacity but high enough for AP contention to become a bottleneck.

Logging, clients, and baselines. Offline data are generated by four rule-based policies. The **uniform** policy samples uniformly from $\{\perp\} \cup Y_i$. The **best_q** policy chooses the accessible AP with the largest service probability. The **round_robin** policy cycles through accessible APs and defers when the queue is empty. The **defer_heavy** policy defers with probability $1/2$ and otherwise samples uniformly from Y_i.

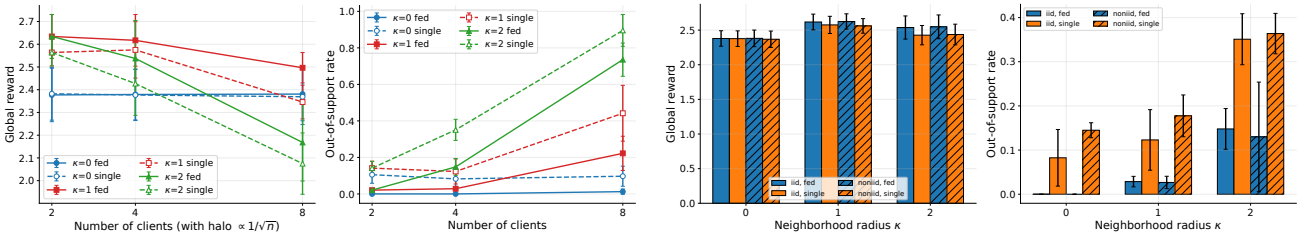

Figure 3. Federation diagnostics. Left panel shows the client count sweep. Right panel compares IID and non-IID logging.

In IID logging, each client uses a rotated mixture of the four behaviors across its UEs. In non-IID logging, client k uses a single behavior for all owned UEs. The default $K = 4$ clients are formed by balanced spatial k-means clustering. Client k observes its owned UEs and all UEs within halo distance h of any owned UE.

The baselines comprise the four engineered policies, the behavior mixture that generated the logs, a non-localized observability-respecting variant, and a non-localized raw-pooled oracle (which ignores the no-raw-sharing and observability constraints).

Metrics. The headline metric is average system reward $\bar{r} = \mathbb{E}[\sum_i r_i(t)]$. Delivery efficiency is expected deliveries divided by expected arrivals.

F. Additional Results and Ablations

Data budget sweep. Two interpretations follow directly from the data budget sweep of Sec. 7. First, the $\kappa = 0$ curve is essentially flat (2.38 across all four budgets), indicating a bias-limited regime in which additional samples cannot recover the locality bias. Second, $\kappa = 2$ remains more data-hungry than $\kappa = 1$ throughout the sweep ($\kappa = 1$ leads by 0.08–0.12 in reward at every budget), because its larger local context space spreads samples across more cells and slows per-cell convergence.

F.1. Client count and non-IID logging

Figure 3 reports additional federation diagnostics. The left panel shows the client-count sweep, and the right panel compares IID and non-IID logging. In the client count sweep with $K \in \{2, 4, 8\}$ and halo $h(K) = 1.2L/\sqrt{K}$, the benefit of federation grows with K at $\kappa = 1$: as clients own smaller regions, single client local coverage weakens and aggregation over $\mathcal{R}_i^\kappa$ becomes more important. At $K = 8$, $\kappa = 2$ becomes geometrically limited because the scaled halo is often too small to cover two-hop neighborhoods. Under non-IID logging, federated F-GLOFF remains stable across regimes while single client OOS rises because each owner observes a narrower behavior distribution, which supports the coverage-mixture role of federation.