Federated Posterior Sharing for Multi-Agent Systems in Uncertain Environments

Yuxi Wang WANG.YUXI4@NORTHEASTERN.EDU

Peng Wu WU.P@NORTHEASTERN.EDU

Mahdi Imani M.IMANI@NORTHEASTERN.EDU

Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts, USA

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

The use of artificial intelligence (AI) agents is increasingly growing in complex, dynamic environments such as disaster response, search and rescue, and law enforcement. These domains are often only partially known, requiring agents to learn and adapt as they gather more information. In multi-agent settings, where agents operate independently and possess diverse, partial views of the environment, sharing their environmental knowledge is essential for enhancing operational efficiency and safety. Existing federated learning approaches focus primarily on policy sharing without modeling environmental uncertainty. To address this gap, this paper presents a framework that enables multiple agents to collaboratively share their probabilistic knowledge of the environment, building a global, shared understanding that efficiently guides their policies. Unlike existing data fusion techniques that exchange raw data—posing privacy risks and increasing communication costs—our method fuses agents' local posterior distributions as an abstract representation of their past data. We provide both single-step and multi-step synchronization, enabling recursive aggregation of agents' knowledge to support informed and adaptive decision-making. Numerical experiments show that our method achieves superior accuracy and decision efficiency compared to several existing methods, particularly in settings with heterogeneous priors and greater uncertainty.

Keywords: Federated Learning, Multi-Agent Systems, Collaborative Knowledge Sharing, Bayesian Fusion, Reinforcement Learning.

1. Introduction

The rapid advancement of artificial intelligence (AI) has led to the widespread deployment of autonomous agents across diverse and complex domains, such as disaster response, search and rescue missions, law enforcement, and environmental monitoring (Goodfellow et al., 2016; LeCun et al., 2015). These environments are often only partially known, requiring agents to continuously learn and adapt their strategies as new information emerges (Mnih et al., 2015; Sutton and Barto, 2018). For instance, in disaster relief operations, autonomous robots and vehicles may be deployed to locate survivors, deliver medical aid, or clear blocked pathways. In such scenarios, if one agent identifies a blocked route or hazardous area, sharing this information with others becomes essential to improve agents' decision-making (McMahan et al., 2017; Lin et al., 2024; Wang et al., 2019).

Existing federated learning approaches (Yang et al., 2019; Asadi et al., 2024; Wu et al., 2021) focus on policy sharing among agents performing similar tasks and goals, aiming to enhance sample efficiency and accelerate learning by leveraging information from other agents (Qi et al., 2021;

Jin et al., 2022; Khodadadian et al., 2022; Wang et al., 2024). In recent years, federated reinforcement learning (FRL) techniques have been developed to facilitate knowledge sharing across multiple independent agents in the same environment (Cha et al., 2020; Li et al., 2020). For example, the FedRL framework assumes that agents do not directly share their observations but instead aggregate their local models—such as Q-networks—to collaboratively improve policies (Jin et al., 2022; McMahan et al., 2017). Other frameworks adapt classical reinforcement learning methods to federated settings by incorporating techniques like policy distillation (Czarnecki et al., 2019), temporal-difference learning (Wang et al., 2023), and mixup augmentation (Jang et al., 2023).

Beyond these centralized FRL methods, additional categories address unique challenges in multi-agent environments. Federated Multi-Agent Reinforcement Learning (FMARL) approaches, for example, enable federated updates among teams of AI agents (Zhang et al., 2022). Policy Sharing and Knowledge Distillation techniques focus on exchanging specific learned components, such as sub-policies or value functions, rather than entire policy models, to make sharing more efficient (Stanton et al., 2021). Decentralized Federated Learning methods enable coordination without a central aggregator, which is particularly useful in ad-hoc or dynamic agent networks (Lian et al., 2017; Liu and Wu, 2022). Personalized Federated Reinforcement Learning tailors learning to individual or groups of agents, accounting for local data and environmental conditions to improve overall performances (Nadiger et al., 2019; Wu et al., 2024a).

However, existing approaches face limitations in partially known environments, where AI agents must adapt their strategies as they gather new information (Padakandla, 2021; Kazeminajafabadi and Imani, 2024; Zhou et al., 2020a). In such settings, agents typically maintain heterogeneous and localized knowledge of the environment. This partial knowledge directly impacts the efficiency of agents' decisions, potentially delaying operations and posing safety risks. Given that all agents share the same environment, integrating their environmental knowledge is essential for enabling more robust, context-aware decisions that extend beyond each agent's limited experience. Existing solutions for environmental knowledge sharing typically rely on data fusion, where raw data from multiple agents is aggregated to form a global perspective (Zhou et al., 2020a,b). However, data fusion requires the exchange of raw data, introducing several significant drawbacks: (1) It is resource-intensive and often impractical in communication-constrained environments; (2) It poses privacy risks and could expose sensitive information to adversarial threats; (3) As the number of agents grows, the computational and communication overhead associated with data fusion becomes prohibitive (Wu et al., 2024b; Geyer et al., 2017; Kairouz et al., 2021; Wu, 2024). Consequently, there is a pressing need for alternative approaches that enable effective environmental knowledge sharing without the heavy demands of raw data exchange.

This paper addresses the decision-making processes of multiple independent agents operating in partially known environments. Each agent's decision-making is modeled as a partially known Markov Decision Process (MDP), where agents hold varying, incomplete prior knowledge of the environment (Puterman, 2014). To bridge these knowledge gaps, we propose a probabilistic knowledge-sharing method based on agents' posterior distributions. Our federated posterior sharing framework enables agents to periodically share their local posterior distributions, which are aggregated to construct a global posterior distribution. We demonstrate that this global posterior achieves the exact solution of optimal Bayesian data fusion (Wu et al., 2024b; Tresp, 2000), without requiring the exchange of any agent's raw data. By sharing knowledge in the form of abstract posteriors, agents can seamlessly and in real time exchange insights, enhancing safe and efficient operations. Both single-step and multi-step synchronization mechanisms are introduced in

the proposed method. Numerical experiments confirm the superiority of our approach in terms of knowledge gained by agents and the rewards achieved across various conditions, including differing levels of uncertainty, varying numbers of agents, and heterogeneous priors.

2. Federated Learning for Sharing Posterior of Environments

2.1. Problem Formulation: Partially-Known Markov Decision Process

Consider n agents performing tasks independently in a partially known environment, with uncertainty represented by a finite set Θ. The decision-making process of the i-th agent is modeled as a partially known Markov Decision Process (MDP), defined by the 5-tuple $\langle \Theta, \mathcal{S}^i, \mathcal{A}^i, \mathcal{P}^i_\theta, R^i \rangle$, where $\mathcal{S}^i$ denotes the state space of agent i, $\mathcal{A}^i$ represents the action space, $\mathcal{P}^i_\theta : \mathcal{S}^i \times \mathcal{A}^i \times \mathcal{S}^i \to [0, 1]$ is the state transition probability function under model $\theta \in \Theta$, such that $\mathcal{P}^i_\theta(s^i, a^i, s'^i) = p(s'^i | s^i, a^i, \theta)$, $R^i : \mathcal{S}^i \times \mathcal{A}^i \to \mathbb{R}$ is the reward function, encoding the reward $R^i(s^i, a^i, s'^i)$ earned when agent i takes action a^i in state s^i and transitions to state s'^i. In this setup, agents operate within the same environment but pursue independent objectives. For instance, multiple robots might be deployed for distinct tasks in the aftermath of a disaster, each with a unique objective but operating within a shared, uncertain environment. For example, in a disaster response scenario, multiple robots may be deployed, with each robot independently developing its own perception of the environment as it operates.

2.2. Problem Formulation: Decision-Making in Uncertain Environments

If agents were aware of the underlying true environment, they could operate according to policies that maximize the expected performance/reward in that environment. However, the true environment θ^* is unknown and hidden among a set of possible environment models Θ. Let θ represent the i-th agent's knowledge of the environment, and let $\pi^i_\theta : \mathcal{S}^i \to \mathcal{A}^i$ be a deterministic policy mapping the i-th agent's states to actions. The optimal policy for the i-th agent, given environment θ, can then be formulated as:

$$\pi^{i*}_\theta(s^i) = \underset{\pi^i}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{k=0}^{h} \gamma^t R^i(s^i_k, a^i_k, s^i_{k+1}) \mid s^i_0 = s^i, a^i_{0:h} \sim \pi^i, \theta\right], \text{for all } s^i \in \mathcal{S}^i, \quad (1)$$

where the expectation is taken with respect to state transitions under the environment θ, γ is the discount factor representing the impact of early stage reward compared to future ones, h denotes the time horizon, which may be finite or infinite depending on the problem, and the maximization is performed over all deterministic policies. Depending on the size of state and action spaces, the optimal policy in Equation (1) can be computed using dynamic programming (Puterman, 2014) or approximated using reinforcement learning techniques (Mnih et al., 2015).

Decision-making in partially known environments presents significant challenges, primarily due to the need for agents to continuously update their knowledge and adapt strategies as new information becomes available. Let $\Theta = (\theta^1, \ldots, \theta^L)$ represent the set of possible environment models, with the true model θ^* hidden within this set. Given policies $\pi^{i*}_{\theta^1}, \ldots, \pi^{i*}_{\theta^L}$ for each possible model, agent i must decide which policy to follow at any given time to optimize its performance. Bayesian and probabilistic approaches are widely adopted in such settings, allowing agents to recursively build a probabilistic model of environmental uncertainty (Rigter et al., 2021; Alali and Imani, 2025;

Ravari et al., 2024; Budd et al., 2023; Alali and Imani, 2024). By leveraging this probabilistic knowledge, agents can make more informed and efficient decisions, dynamically adapting their actions based on updated beliefs about the true environment.

Formally, let $\mathcal{D}_k^i$ represent all data available to the i-th agent up to time step k, which may include interactions with the environment (i.e., state-action pairs). Given its local data $\mathcal{D}_k^i$, the agent's probabilistic knowledge of the environment is expressed through the following local posterior distribution:

$$p_k^i = [p_k^i(1), ..., p_k^i(l), ..., p_k^i(L)] = [P(\theta^1 \mid \mathcal{D}_k^i), ..., P(\theta^l \mid \mathcal{D}_k^i), ..., P(\theta^L \mid \mathcal{D}_k^i)], \tag{2}$$

where $p_k^i(l)$ denotes the i-th agent's belief that θ^l is the true environment model, and $\sum_{l=1}^{L} p_k^i(l) = 1$ ensures a valid posterior. In a probabilistic setting, agents can use these local posteriors to make adaptive decisions. One straightforward approach is to follow the policy corresponding to the model with the highest local posterior probability, given by:

$$a_k^i \sim \pi_{\hat{\theta}_k^i}^{i*}(s_k^i), \quad \text{where} \quad \hat{\theta}_k^i = \underset{\theta^l, l \in \{1,...,L\}}{\operatorname{argmax}} \; p_k^i(l), \tag{3}$$

where $\hat{\theta}_k^i$ represents the most probable model, known as the Maximum A Posterior (MAP) estimate, for agent i based on its local data.

The MAP estimates, $\hat{\theta}_k^i$, may vary across agents, as each agent's local posterior distribution is based solely on its own data, i.e., $p_k^i \neq p_k^j$ for $j \neq i$. For a given agent i, the probability that the agent follows a policy associated with an incorrect environment model can be quantified by:

$$c^i = P(\theta^* \neq \hat{\theta}_k^i) = 1 - P(\theta^* = \hat{\theta}_k^i) = 1 - \max_{\theta^l; l \in \{1,...,L\}} p_k^i(\theta), \tag{4}$$

where $c^i \in \left[0, \frac{L-1}{L}\right]$ indicates the probability that the MAP estimate of the model for agent i does not match the true model. Values of c^i close to $\frac{L-1}{L}$ indicate a nearly uniform posterior distribution among models, where the probabilities for each model are almost equal. This scenario implies a low confidence level in the MAP estimate, resembling a random choice. Conversely, values of c^i closer to zero suggest that the MAP estimate is significantly more probable than other models, indicating that the true environment model is more distinguishable from the local data.

Agents with higher c^i values are more likely to follow policies that do not correspond to the true model, potentially leading to delays in operations until the agent gathers enough information to refine its understanding of the environment and improve its decisions. This paper introduces a probabilistic knowledge-sharing framework that enables agents to share their individual knowledge to form a collective understanding of the environment, thereby enhancing both operational effectiveness and safety.

2.3. Federated Posterior Sharing for Collective Environmental Knowledge

Let $\mathcal{D}_k^1, \ldots, \mathcal{D}_k^n$ represent the local data of n agents operating within the same environment. Since each agent's experiences may vary—such as agents operating in different regions of the environment—combining their experiences to form a collective global posterior is essential. This global posterior can be represented as:

$$p_k^G = [P(\theta^1 \mid \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n), \ldots, P(\theta^L \mid \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n)], \tag{5}$$

where

$$P(\theta^l \mid \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n) = \frac{p(\mathcal{D}_k^1, \ldots, \mathcal{D}_k^n, \theta^l)}{p(\mathcal{D}_k^1, \ldots, \mathcal{D}_k^n)} = \frac{p(\mathcal{D}_k^1, \ldots, \mathcal{D}_k^n \mid \theta) \cdot P(\theta^l)}{p(\mathcal{D}_k^1, \ldots, \mathcal{D}_k^n)} = \frac{\left[\prod_{i=1}^n p(\mathcal{D}_k^i \mid \theta)\right] \cdot P(\theta^l)}{p(\mathcal{D}_k^1, \ldots, \mathcal{D}_k^n)}. \tag{6}$$

The final expression in Equation (6) is derived based on the assumption that each agent's state transitions are conditionally independent of other agents' states and actions, given its own state, action, and the environment model θ^l. This formulation shows that data from all agents are integrated to construct a global posterior, which offers a collective view of the environment.

However, constructing a global posterior through direct data sharing requires agents to share their local data, which introduces substantial security, privacy, and communication challenges. For instance, sharing data such as location information may compromise agents' privacy and expose sensitive details, especially in adversarial settings. Additionally, data fusion is computationally intensive and may not meet the real-time constraints necessary for scenarios involving multiple agents.

Let p_k^G be the global posterior in Equation (5), which aggregates the local data from all agents up to time step k. The data for agent i at time step k is represented by the set of state and action pairs $\mathcal{D}_k^i = \{a_{0:k-1}^i, s_{0:k}^i\}$. At time k, agent i takes action a_k^i, transitioning from state s_k^i to s_{k+1}^i. Using these updated local observations, the global posterior at time step $k+1$ can be computed as:

$$\begin{aligned} p_{k+1}^G(l) = P(\theta^l \mid \mathcal{D}_{k+1}^1, \ldots, \mathcal{D}_{k+1}^n) &= P(\theta^l \mid a_k^1, s_{k+1}^1, \mathcal{D}_k^1, \ldots, a_k^n, s_{k+1}^n, \mathcal{D}_k^n) \\ &= \frac{\left[\prod_{i=1}^n p(s_{k+1}^i \mid a_k^i, s_k^i, \theta^l)\right] p_k^G(l)}{p(\mathcal{D}_{k+1}^1, \ldots, \mathcal{D}_{k+1}^n)}. \end{aligned} \tag{7}$$

Since direct data sharing for recursive global posterior computation in Equation (7) poses significant privacy and security risks, we propose an alternative approach where agents compute and share their local posteriors instead of local data (Wu et al., 2024b). Unlike local data, local posteriors provide an abstract representation of the information available to each agent, minimizing the risk of data exposure, as it is typically infeasible to reconstruct the original data from the posterior alone. Further research to quantify the extent to which private information might be inferred from shared local or global posteriors remains an important area for future work.

2.3.1. SINGLE-STEP SYNCHRONIZATION FOR FEDERATED POSTERIOR SHARING

Let p_k^G be the global posterior among agents at time step k. After agent i takes action a_k^i and transitions to s_{k+1}^i, its local posterior distribution can be recursively updated as follows:

$$p_{k+1}^i = \left[P(\theta^1 \mid a_k^i, s_{k+1}^i, \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n), \ldots, P(\theta^L \mid a_k^i, s_{k+1}^i, \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n)\right], \tag{8}$$

where p_{k+1}^i is the local posterior of agent i, incorporating its latest local transition along with the prior global information. The l-th element of this local posterior can be expressed as:

$$\begin{aligned} p_{k+1}^i(l) &= P(\theta^l \mid a_k^i, s_{k+1}^i, \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n) \\ &\propto p(s_{k+1}^i \mid a_k^i, s_k^i, \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n, \theta^l) P(\theta^l \mid a_k^i, \mathcal{D}_k^1, \ldots, \mathcal{D}_k^n) \\ &= p(s_{k+1}^i \mid a_k^i, s_k^i, \theta^l) p_k^G(l), \end{aligned} \tag{9}$$

where the likelihood $p(s_{k+1}^i \mid a_k^i, s_k^i, \theta^l)$, derived from the private data of agent i, is combined with the prior global posterior $p_k^G(l)$ to form the updated local posterior for agent i.

Let $p^1_{k+1}, \ldots, p^n_{k+1}$ represent the local posteriors computed by the agents according to Equations (8) and (9). The global posterior at time step $k+1$ can then be computed by aggregating the local posteriors, rather than sharing raw data or individual likelihoods, as follows:

$$\begin{aligned} p^G_{k+1}(l) &= P(\theta^l \mid \mathcal{D}^1_{k+1}, \ldots, \mathcal{D}^n_{k+1}) \propto \left[\prod_{i=1}^{n} p(s^i_{k+1} \mid a^i_k, s^i_k, \theta^l)\right] p^G_k(l) \\ &= \frac{\prod_{i=1}^{n} \left[p(s^i_{k+1} \mid a^i_k, s^i_k, \theta^l) p^G_k(l)\right]}{p^G_k(l)^{n-1}} \propto \frac{\prod_{i=1}^{n} p^i_{k+1}(l)}{[p^G_k(l)]^{n-1}}, l \in \{1, \ldots, L\}. \end{aligned} \tag{10}$$

The denominator in the final expression ensures that the prior global posterior p^G_k is not redundantly applied multiple times, thus preventing biases that could arise in posterior-sharing methods. Despite relying on shared local posteriors rather than raw data, the approach in Equation (10) achieves the same optimality as Bayesian data fusion in (7).

2.3.2. T-STEP SYNCHRONIZATION FOR FEDERATED POSTERIOR SHARING

In settings where frequent sharing of local posteriors is impractical due to communication delays or stringent time constraints, periodic synchronization of local posteriors offers an effective solution for achieving precise knowledge fusion. With synchronization occurring at every time step, as shown in Equation (10), agents maintain a shared posterior at all times. We introduce a T-step synchronization method, ensuring robustness and adaptability in such scenarios.

Suppose the last synchronization occurred at time step k. Each agent will continue updating its local posterior independently between time steps $k+1$ and $k+T$ until the next synchronization. Let p^G_k denote the current shared global posterior at time step k. For $t \leq T$, the local posterior of agent i can be recursively updated as:

$$\begin{aligned} p^i_{k+t}(l) &= P(\theta^l \mid a^i_{k+t-1}, s^i_{k+1+t}, \mathcal{D}^1_k, \ldots, \mathcal{D}^n_k) \\ &\propto \left[\prod_{r=1}^{t} p(s^i_{k+r} \mid a^i_{k+r-1}, s^i_{k+r-1}, \mathcal{D}^1_k, \ldots, \mathcal{D}^n_k, \theta^l)\right] p^G_k(l) \\ &= p(s^i_{k+t} \mid a^i_{k+t-1}, s^i_{k+t-1}, \theta^l) \cdot p^i_{k+t-1}(l), \end{aligned} \tag{11}$$

where p^i_{k+t-1} is the local posterior of agent i, calculated based on the global posterior at time k and its most recent local data. At the next synchronization point, $k+T$, agents can use their latest local posteriors $p^1_{k+T}, \ldots, p^n_{k+T}$ to reconstruct the global posterior as follows:

$$\begin{aligned} p^G_{k+T}(l) &= P(\theta^l \mid \mathcal{D}^1_{k+T}, \ldots, \mathcal{D}^n_{k+T}) \propto \left[\prod_{i=1}^{n}\prod_{t=1}^{T} p(s^i_{k+t} \mid a^i_{k+t-1}, s^i_{k+t}, \theta^l)\right] p^G_k(l) \\ &= \frac{\prod_{i=1}^{n} \left[\prod_{t=1}^{T} p(s^i_{k+t} \mid a^i_{k+t-1}, s^i_{k+t-1}, \theta^l)\right] p^G_k(l)}{[p^G_k(l)]^{n-1}} = \frac{\prod_{i=1}^{n} p^i_{k+T}(l)}{[p^G_k(l)]^{n-1}}, \end{aligned} \tag{12}$$

where the denominator corresponds to the previous shared global posterior at time step k. This formulation ensures that the previous global posterior is accounted for only once, thus avoiding multiple applications that could bias the update. By tracking only the latest local posteriors, this approach enables the exact computation of the global posterior while preserving the accuracy of the federated knowledge-sharing process.

The global posterior encapsulates the collective knowledge of all agents about the environment, yielding a more accurate representation of uncertainty. As a result, aggregating information through

the global posterior increases the likelihood of correctly identifying the true environment model θ^*. According to Equation (4), we can show that the MAP estimate derived from the global and local posterior yields:

$$P\left(\theta^* \neq \arg\max_{\theta\in\Theta} P(\theta \mid \mathcal{D}^1_{k+T}, \ldots, \mathcal{D}^n_{k+T})\right) \leq P\left(\theta^* \neq \arg\max_{\theta\in\Theta} P(\theta \mid \mathcal{D}^i_{k+T})\right), \text{for } i = 1, \ldots, n, \quad (13)$$

where the global posterior incorporates data from all agents, whereas each local posterior uses only a subset (the local data). By properties of Bayesian inference, the inclusion of more data brings the posterior closer to the true distribution. Consequently, the MAP estimate from the global posterior is at least as accurate as any estimate derived from an individual local posterior.

This shared global posterior enables agents to make more informed decisions that are better aligned with the policy under the true environment. The MAP-based policy in Equation (3), when informed by the global posterior distribution, is given by:

$$a^i_{k+T} \sim \pi^{i*}_{\hat{\theta}_{k+T}}(s^i_{k+T}), \quad \text{for } i = 1, \ldots, n, \quad \text{where } \hat{\theta}_{k+T} = \underset{\theta^l, l\in\{1,\ldots,L\}}{\operatorname{argmax}} \; p^G_{k+T}(l), \quad (14)$$

where $\hat{\theta}_{k+T}$ represents the MAP model (*i.e.*, the model with the highest posterior probability) shared among all agents. The unified global posterior creates a shared understanding of the environment, giving all agents a consistent model. This alignment helps agents anticipate each other's states and trajectories, improving coordination and safety, especially in scenarios where diverse knowledge could cause misunderstandings or conflicts.

An illustrative diagram of the proposed framework is shown in Figure 1. In this framework, agents operating independently within a partially-known environment share their local posterior distributions with a central aggregator at each T-step synchronization interval. This aggregation produces a global posterior that reflects the collective knowledge of all agents, allowing each agent to use the MAP estimate from the global posterior to guide its policy. The agent-color-coded arrows represent the flow of local posterior updates, while blue arrows indicate shared global information. This approach maintains a real-time complexity of $O(L \times n)$ per step, where L is the number of possible environment models and n is the number of agents. Although MAP-based local update policies are demonstrated here for simplicity, the framework can integrate alternative policies that leverage the full posterior distribution, such as Bayesian or robust policies, offering flexibility for diverse operational needs.

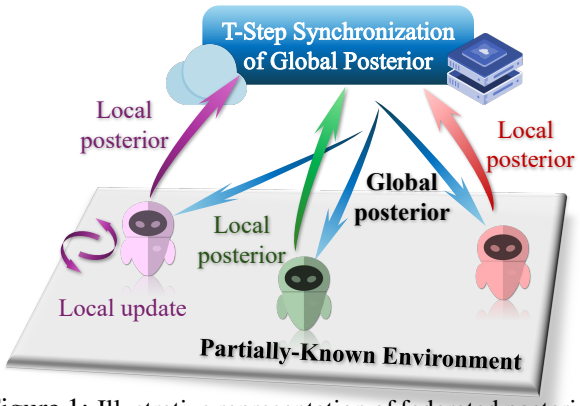

Figure 1: Illustrative representation of federated posterior sharing with T-step synchronization.

3. Numerical Experiments

In this section, we evaluate the performance of the proposed framework by comparing it with several existing methods using a grid-world environment, depicted in Figure 2. The environment consists of three independent agents operating within the same environment containing five unknown elements,

highlighted in yellow cells. Each unknown cell can independently be in one of two states: *Empty* or *Wall*, resulting in a total of 32 possible environment models. The true environment model is initially unknown to the agents. Each agent has a prior knowledge of the true environments, represented by a Dirichlet distribution with a single shared parameter β. Here, β determines the nature of the agents' priors: larger values of β correspond to scenarios with uniform, non-informative priors, indicating that agents share similar knowledge of environment uncertainty. Conversely, smaller values of β represent more peaked priors, with individual agents potentially having skewed knowledge that may vary among agents.

Each agent in this environment has two distinct goals, represented by cells marked in the same color as the corresponding agent in Figure 2. Each agent can take one of four actions: *Up*, *Down*, *Left*, or *Right*. Movement is probabilistic: The agent successfully moves in the intended direction with probability p, while with probability $(1-p)/2$, it instead moves toward one of the perpendicular directions. If an agent attempts to move into a wall, it remains in its current position. To encourage task completion, each agent receives a reward of +200 upon reaching each goal cell and incurs a penalty of -1 for each action taken, incentivizing efficient task completion.

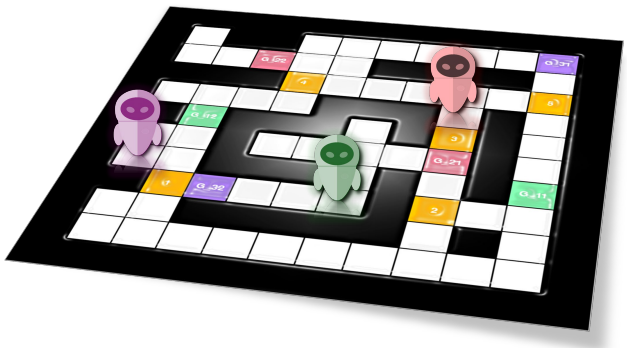

Figure 2: Grid environment with three agents, each with two color-matched goal cells, and five unknown cells (yellow).

For policy training, we employ the Q-Learning algorithm offline to find the agents' policies across all 32 possible environments. The performance of our proposed framework is benchmarked against two alternative approaches: (1) the conditionally independent posterior (CIP) method (Wu et al., 2023), and (2) the local posterior approach, where agents only rely on their individual past information without federated posterior sharing. Experimental parameters are set as follows: $p = 0.9$ and $\beta = 100$.

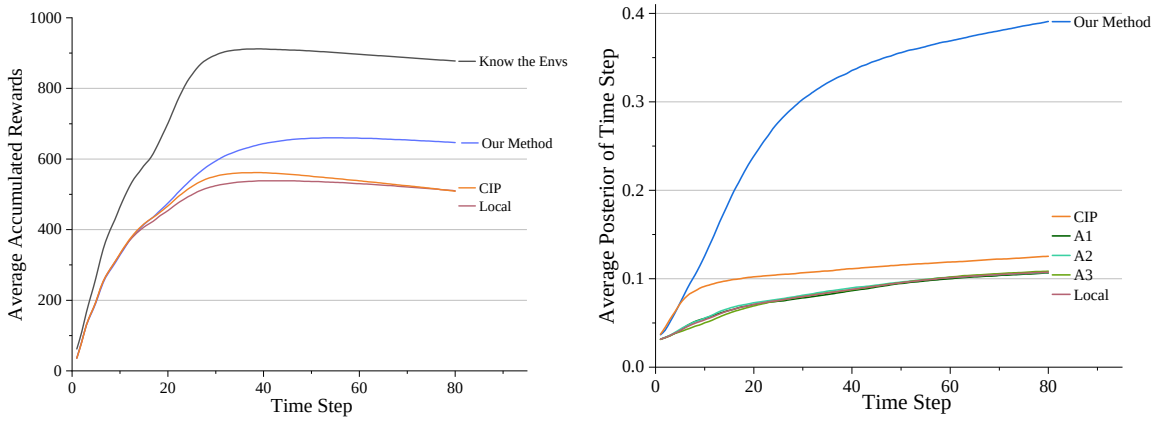

Figure 3: Average accumulated rewards (left) and posterior of the true environment (right) over time for the proposed method and competing methods.

Figure 3 shows a comparison of the proposed method against competing methods over 10,000 independent trials, with each trial selecting a true environment model at random from the 32 pos-

sible configurations. The left plot illustrates the average accumulated reward for all three agents. The black curve represents the baseline performance if the true environments were fully known, serving as a performance upper bound for methods that rely on partial environment knowledge. As expected, the proposed method, represented by the blue curve, performs closer to the baseline. In contrast, the CIP and local policies exhibit significantly lower accumulated rewards. For the CIP method, this underperformance arises due to excessive reliance on local prior information during global posterior computation, which introduces bias. This biased global posterior adversely impacts decision-making. The local policies also underperform because agents are confined to different areas of the environment and thus hold diverse, partial knowledge of the unknown regions. The right plot illustrates all three agents' average posterior by time step when utilizing different methods. Without shared knowledge, each agent relies solely on its experience, limiting its ability to make effective decisions across the entire environment. By contrast, the proposed method iteratively corrects for this bias, achieving an unbiased, precise posterior that significantly improves overall performance.

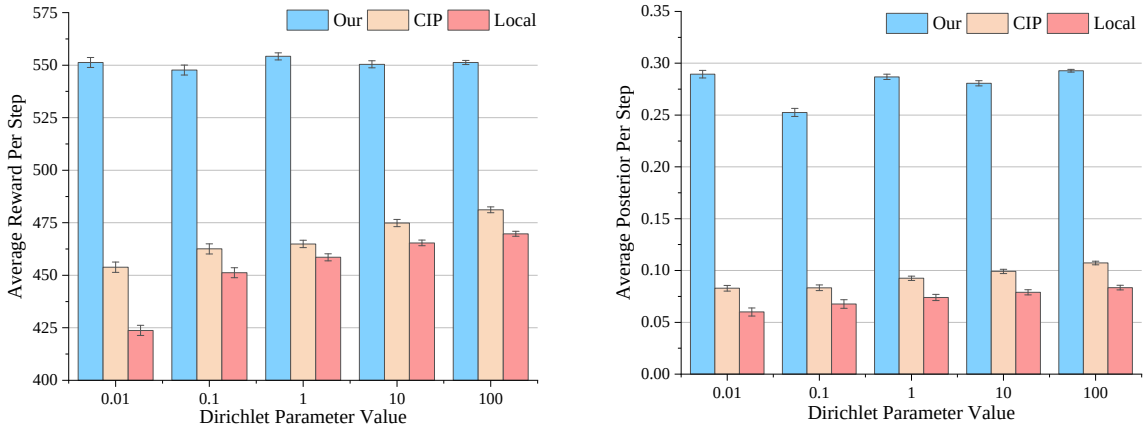

Figure 4: Effect of agents' Dirichlet prior on average reward per step (left) and posterior probability of the true model (right) across different methods.

To study the effect of different initial priors, we vary the Dirichlet parameter β from 0.01 to 100, as shown in Figure 4. Lower Dirichlet parameter values indicate more imbalanced priors among agents, suggesting a higher chance of incorrect initial assumptions about the environment. In contrast, higher values result in priors closer to a uniform distribution, reflecting more accurate initial knowledge. The left plot in Figure 4 shows the average reward per step over 10,000 runs, each lasting 80 time steps. The results show that the proposed method consistently outperforms competing methods across different prior conditions. Notably, our method maintains consistent performance regardless of the level of prior imbalance, highlighting its robustness in effectively sharing knowledge in scenarios with varying levels of prior imbalance. As expected, the other two methods exhibit increased average rewards when agents have more uniform prior knowledge, compared to when they have imbalanced and potentially incorrect priors. Moreover, local policies perform worse than the CIP method because they lack access to the global posterior, resulting in slower updates to the entire environment. This difference is especially noticeable under lower Dirichlet parameters due to the higher chance of incorrect initial assumptions. The right plot shows a similar trend, with the average posterior probability of the true model increasing as priors become more uniform. Across all conditions, the proposed method demonstrates superior and robust performance in accurately identifying

the true environment model, underscoring its effectiveness in leveraging collective knowledge even when agents start with diverse prior beliefs.

Table 1: Average reward per step across synchronization intervals ($T = 1, 3, 5$) and levels of uncertainty (1, 3, 5 unknowns).

Agents synchronization interval	Number of unknown cells in the environment								
	2			3			5		
	Our method	CIP	Local	Our method	CIP	Local	Our method	CIP	Local
1	673.97	659.21		648.23	632.16		551.29	481.13	
3	666.80	654.02	604.00	637.60	626.74	551.81	533.41	470.00	469.72
5	656.50	653.01		622.26	621.92		517.74	461.37	

Finally, we examine the performance of the proposed method in relation to synchronization frequency and the number of unknown elements (or uncertainties) in the environment. We test three synchronization intervals ($T = 1, 3, 5$) and three levels of environmental uncertainty (1, 3, and 5 unknowns), with unknown elements randomly selected from the five possible unknowns shown in Figure 2. Table 1 presents the average reward per step for all methods under these conditions. Since local policies do not involve any information sharing, their results remain consistent across different synchronization intervals. The proposed method, however, consistently outperforms all other methods across scenarios, particularly as the number of unknowns increases. This improvement is attributed to the effectiveness of knowledge sharing, which becomes increasingly valuable in environments with higher uncertainty. As the synchronization interval T increases, agents share knowledge less frequently, resulting in a slight reduction in average reward for both the proposed and CIP methods. Nevertheless, the proposed method demonstrates robust performance across all conditions, highlighting its adaptability and effectiveness in varied synchronization and uncertainty settings.

4. Conclusion and Future Work

In this paper, we introduce a federated learning framework that enables knowledge sharing among independent autonomous agents operating in partially known environments. Unlike traditional federated learning methods that focus on sharing policies or rewards, our approach facilitates the sharing of environmental knowledge through the agents' posterior distributions. The proposed method avoids raw data sharing, instead aggregating local posterior distributions into a global posterior, thereby preserving privacy and enhancing computational efficiency. This approach empowers agents to make adaptive and informed decisions while reducing communication overhead and safeguarding sensitive information. Our extensive numerical experiments demonstrate that the proposed framework consistently outperforms existing methods under diverse conditions, including settings with high uncertainty and heterogeneous priors among agents. Future work will explore privacy considerations related to posterior-based knowledge sharing, such as incorporating differential privacy techniques. Additionally, we plan to extend the framework to multi-agent cooperative settings and domains with partial observability.

Acknowledgments

The authors acknowledge the support of the National Science Foundation award IIS-2311969, ARMY Research Laboratory award W911NF-23-2-0207, Office of Naval Research award N00014-23-1-2850, ARMY Research Office awards W911NF-24-2-0166 and W911NF-21-1-0299.

References

Mohammad Alali and Mahdi Imani. Bayesian reinforcement learning for navigation planning in unknown environments. *Frontiers in Artificial Intelligence*, 7:1308031, 2024.

Mohammad Alali and Mahdi Imani. Deep Reinforcement Learning Data Collection for Bayesian Inference of Hidden Markov Models. *IEEE Transactions on Artificial Intelligence*, 2025.

Negar Asadi, Seyed Hamid Hosseini, Mahdi Imani, Daniel P Aldrich, and Seyede Fatemeh Ghoreishi. Privacy-preserved federated reinforcement learning for autonomy in signalized intersections. In *International Conference on Transportation and Development 2024*, pages 390–403, 2024.

Matthew Budd, Paul Duckworth, Nick Hawes, and Bruno Lacerda. Bayesian reinforcement learning for single-episode missions in partially unknown environments. In *Conference on Robot Learning*, pages 1189–1198. PMLR, 2023.

Han Cha, Jihong Park, Hyesung Kim, Mehdi Bennis, and Seong-Lyun Kim. Proxy experience replay: Federated distillation for distributed reinforcement learning. *IEEE Intelligent Systems*, 35 (4):94–101, 2020.

Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1331–1340. PMLR, 2019.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Junwoo Jang, Jungwoo Han, and Jinwhan Kim. K-mixup: Data augmentation for offline reinforcement learning using mixup in a Koopman invariant subspace. *Expert Systems with Applications*, 225:120136, 2023.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 18–37. PMLR, 2022.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Armita Kazeminajafabadi and Mahdi Imani. Optimal Joint Defense and Monitoring for Networks Security under Uncertainty: A POMDP-Based Approach. *IET Information Security*, 2024(1): 7966713, 2024.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *International Conference on Machine Learning*, pages 10997–11057. PMLR, 2022.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Qinbin Li, Zeyi Wen, and Bingsheng He. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4642–4649, 2020.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Yuxin Lin, Seyede Fatemeh Ghoreishi, Tian Lan, and Mahdi Imani. High-level human intention learning for cooperative decision-making. In *2024 IEEE Conference on Control Technology and Applications (CCTA)*, pages 209–216. IEEE, 2024.

Haotian Liu and Wenchuan Wu. Federated reinforcement learning for decentralized voltage control in distribution networks. *IEEE Transactions on Smart Grid*, 13(5):3840–3843, 2022. doi: 10.1109/TSG.2022.3169361.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Chetan Nadiger, Anil Kumar, and Sherine Abdelhak. Federated reinforcement learning for fast personalization. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 123–127, 2019. doi: 10.1109/AIKE.2019.00031.

Sindhu Padakandla. A survey of reinforcement learning algorithms for dynamically varying environments. *ACM Computing Surveys (CSUR)*, 54(6):1–25, 2021.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: Techniques, applications, and open challenges. *arXiv preprint arXiv:2108.11887*, 2021.

Amirhossein Ravari, Seyede Fatemeh Ghoreishi, and Mahdi Imani. Optimal inference of hidden Markov models through expert-acquired data. *IEEE Transactions on Artificial Intelligence*, 5(8): 3985–4000, 2024.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse Bayes-adaptive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 1142–1154. Curran Associates, Inc., 2021.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34:6906–6919, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Volker Tresp. A bayesian committee machine. *Neural computation*, 12(11):2719–2741, 2000.

Han Wang, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023.

Muxing Wang, Pengkun Yang, and Lili Su. On the convergence rates of federated Q-learning across heterogeneous environments. *arXiv preprint arXiv:2409.03897*, 2024.

Puming Wang, Laurence T Yang, Jintao Li, Jinjun Chen, and Shangqing Hu. Data fusion in cyber-physical-social systems: State-of-the-art and perspectives. *Information Fusion*, 51:42–57, 2019.

Peng Wu. *Bayesian data fusion for distributed learning*. PhD thesis, Northeastern University, 2024.

Peng Wu, Tales Imbiriba, Junha Park, Sunwoo Kim, and Pau Closas. Personalized federated learning over non-iid data for indoor localization. In *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 421–425, 2021. doi: 10.1109/SPAWC51858.2021.9593115.

Peng Wu, Tales Imbiriba, Víctor Elvira, and Pau Closas. Bayesian data fusion with shared priors. *IEEE Transactions on Signal Processing*, 2023.

Peng Wu, Tales Imbiriba, and Pau Closas. A Bayesian framework for clustered federated learning, 2024a. URL https://arxiv.org/abs/2410.15473.

Peng Wu, Tales Imbiriba, Víctor Elvira, and Pau Closas. Bayesian data fusion with shared priors. *IEEE Transactions on Signal Processing*, 72:275–288, 2024b. doi: 10.1109/TSP.2023.3343564.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Sai Qian Zhang, Jieyu Lin, and Qi Zhang. A multi-agent reinforcement learning approach for efficient client selection in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9091–9099, 2022.

Tongle Zhou, Mou Chen, Chenguang Yang, and Zhiqiang Nie. Data fusion using Bayesian theory and reinforcement learning method. *Science China. Information Sciences*, 63(7):170209, 2020a.

Tongle Zhou, Mou Chen, and Jie Zou. Reinforcement learning based data fusion method for multi-sensors. *IEEE/CAA Journal of Automatica Sinica*, 7(6):1489–1497, 2020b.