FedHPD: Heterogeneous Federated Reinforcement Learning via Policy Distillation

Extended Abstract

Wenzheng Jiang
National University of Defense Technology
China
jiangwenzheng@nudt.edu.cn

Ji Wang
National University of Defense Technology
China
wangji@nudt.edu.cn

Xiongtao Zhang
National University of Defense Technology
China
zhangxiongtao14@nudt.edu.cn

Weidong Bao
National University of Defense Technology
China
wdbao@nudt.edu.cn

Cheston Tan
CFAR, A*STAR
Singapore
cheston-tan@i2r.a-star.edu.sg

Flint Xiaofeng Fan
National University of Singapore
Singapore
fxf@u.nus.edu

ABSTRACT

This paper focuses on Federated Reinforcement Learning (FedRL) in black-box settings with heterogeneous agents. Existing studies mostly assume agent homogeneity and knowability of internal details. To tackle these issues, we propose Federated Heterogeneous Policy Distillation (FedHPD). FedHPD uses action probability distributions as a medium for knowledge sharing among heterogeneous agents. Extensive experiments show that FedHPD achieves significant improvements across various benchmark tasks.

KEYWORDS

Federated Reinforcement Learning, Agent Heterogeneity, Knowledge Distillation

ACM Reference Format:
Wenzheng Jiang, Ji Wang, Xiongtao Zhang, Weidong Bao, Cheston Tan, and Flint Xiaofeng Fan. 2025. FedHPD: Heterogeneous Federated Reinforcement Learning via Policy Distillation: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement Learning (RL) faces low sample efficiency [19] and privacy leakage [13] in real-world applications. *Federated Learning* (FL) [12] enables clients to collaboratively improve training efficiency while preserving data privacy, which has led to the emergence of *Federated Reinforcement Learning* (FedRL) [14]. This fusion offers a promising approach for intelligent decision-making in distributed environments [1]. Recognizing its potential, the research community has extensively explored FedRL under various settings [6, 9, 11, 15, 17, 20].

Despite its promise, most typical FedRL frameworks [3, 5, 8] operate under the assumption of agent homogeneity (i.e., identical policy networks and training configurations), which significantly limits effective training in resource-constrained environments [18]. In addition, existing FedRL frameworks typically operate under a white-box paradigm, where internal details could be shared with the server. However, in certain business-oriented domains, the disclosure of internal details is prohibited due to commercial sensitivities and regulatory compliance [10]. These practical constraints lead us to formulate a more challenging question: ***How can we effectively perform FedRL when each agent employs a unique model that remains a black box to the server?***

In light of the above question, we propose *Federated Heterogeneous Policy Distillation* (FedHPD). Compared to FedHQL [4], a pivotal feature of FedHPD is its elimination of reliance on the server-side MDP during training. Unlike traditional policy distillation [16], FedHPD periodically extracts knowledge from local policies to form a global consensus, which ensures the continuity and stability of local training and helps to improve policy generalization. Through distillation, we achieve collaborative training among heterogeneous agents. Through the alternating process of multiple rounds of local training and periodic collaborative training, FedHPD could balance communication overhead and training efficacy [1].

2 PROBLEM FORMULATION

Assuming the system contains a set of K heterogeneous and mutually black-box agents, where agent k interacts in a separate copy of the MDP $\mathcal{M}$ according to policy π_k, generating their own local private data $D_k \triangleq \{(s, a, s', r)_i\}_{i=1}^{|D_k|}$. The policy $\pi_k(a|\phi_k(s;\theta_k))$ is composed of a nonlinear function $\phi_k(s;\theta_k)$ that predicts the probability of taking action a given the state s. The nonlinear function $\phi_k(s;\theta_k)$ is parameterized by a set of parameters θ_k and is learned using private experience data D_k.

Agent heterogeneity [2] is often manifested in the differences in neural network architectures and training configurations [4]. We aim to develop a new FedRL framework that allows heterogeneous agents to share their knowledge from local policies in a black-box

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1] For an extended version of this paper, refer to [7].
[2] In this paper, "agent heterogeneity" indicates that policy networks and training configurations of agents in FedRL are heterogeneous [4].

manner, without being constrained by the server-side MDP process as in FedHQL [4]. The objective functions of the problem can be formulated as follows:

Assuming that all agents are trained synchronously, given the number of training rounds T, let ψ'_k and ψ_k represent the training performance of agent k under the FedRL framework and its performance when trained independently, respectively. We have:
Under the same training rounds,
for system:

$$\sum_{k=1}^{K} \psi'_k \geq \sum_{k=1}^{K} \psi_k; \tag{1}$$

for agent $k, k \in \{1, 2, \ldots, K\}$:

$$\psi'_k \geq \psi_k. \tag{2}$$

3 FEDHPD

Firstly, we set up a simulated environment on the server and train a virtual agent within it. Although optimal performance is not required, reasonable parameterization is needed to ensure stable operation in the environment. For example, in autonomous driving, the simulated environment can be based on an existing simulation platform (such as CARLA [2]), where the virtual agent undergoes preliminary training to learn basic driving strategies. Next, we conduct multiple tests with different initial state inputs to simulate the diverse scenarios that the agent may encounter. In each test, the virtual agent generates a series of states, from which a subset of states is randomly selected to form the public state set $\mathcal{S}_p$.

Notably, the synthetic state set $\mathcal{S}_p$ generated by the virtual agent is independent of local training and serves solely for KD. Of course, for tasks like autonomous driving, robot control, and games, there is an abundance of available experience data that can directly serve as the public state set $\mathcal{S}_p$. In conclusion, obtaining the public state set is reasonable and convenient in real-world. For communication considerations, $\mathcal{S}_p$ is distributed to each local agent before the start of training, so that the communication only involves the upload and distribution of the knowledge.

To handle tasks both in discrete and continuous action spaces, we use action probability distribution as a bridge for communication. Agent k generates its private experience data D_k by interacting with the environment for local training. During the collaborative training process, action probability distribution is distilled from the policies of heterogeneous agents using the public state set $\mathcal{S}_p$. Subsequently, agents digest knowledge by comparing the distributions output by themselves with the global consensus. In other words, local agents use public state set $\mathcal{S}_p$ and private data D_k to train and improve their policy ϕ_k, surpassing individual efforts. Our framework is illustrated in Fig.1. The process for FedHPD is described as follows:

(1) Local Training: In each training round, all local agents must interact with the environment, using the generated data D_k to update their own policy parameters θ_k^i, resulting in θ_k^{i+1}. Note that no collaboration is required at this stage.

(2) Collaborative Training: In the collaborative training phase (conducted every d training rounds), agents share their knowledge based on the output distributions under the public state set $\mathcal{S}_p$. The detailed steps are as follows:

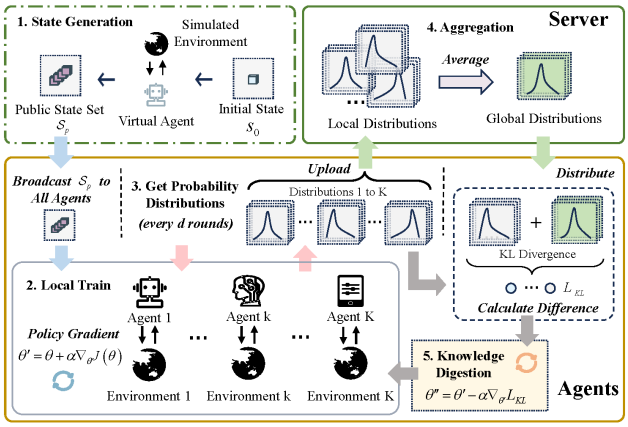

Figure 1: Illustration of FedHPD.

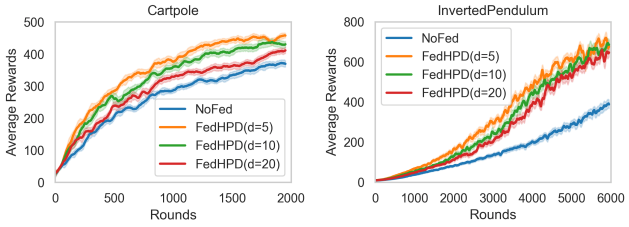

Figure 2: Comparisons of system performance under different distillation intervals (d = 5, 10, 20).

- Get Probability Distributions: Each local agent obtains the action distributions $\mathcal{P}_k^{i+1}$ under state set $\mathcal{S}_p$ based on its own parameterized policy $\pi(\theta_k^{i+1})$, and uploads $\mathcal{P}_k^{i+1}$ to server;
- Knowledge Aggregation: Server aggregates the uploaded probability distributions to obtain global consensus $\overline{\mathcal{P}^{i+1}}$ and sends $\overline{\mathcal{P}^{i+1}}$ to local agents;
- Knowledge Digestion: Local agents calculate the KL divergence between their predicted distributions $\mathcal{P}_k^{i+1}$ and global distributions $\overline{\mathcal{P}^{i+1}}$ to update their policy parameters θ_k^{i+1}.

4 EMPIRICAL EVALUATION

We set up 10 heterogeneous agents to validate the effectiveness of FedHPD in the system performance improvement. For the entire system, we focus on the average performance of all agents, comparing the system performance under independent local training (NoFed) and FedHPD with different distillation intervals (d = 5, 10, 20). As presented in Fig. 2, we conduct experiments on the CartPole with a discrete action space and the InvertedPendulum with a continuous action space. It is evident that, compared to NoFed, FedHPD achieves superior results across both tasks. For Cartpole, the system typically needs around 1200 local training rounds to achieve a reward above 300 under NoFed, whereas with FedHPD, it requires only about 700 training rounds on average. For InvertedPendulum, NoFed can only obtain a reward of approximately 400, but the system consistently achieves rewards exceeding 600 with FedHPD, . The complete experiment settings and results could be found in [7].

REFERENCES

[1] Zhongxiang Dai, Flint Xiaofeng Fan, Cheston Tan, Trong Nghia Hoang, Bryan Kian Hsiang Low, and Patrick Jaillet. 2024. Chapter 14 - Federated sequential decision making: Bayesian optimization, reinforcement learning, and beyond. In *Federated Learning*. Academic Press, 257–279. https://doi.org/10.1016/B978-0-44-319037-7.00023-5

[2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *CoRL*. 1–16.

[3] Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. 2021. Fault-tolerant federated reinforcement learning with theoretical guarantee. In *Advances in Neural Information Processing Systems*, Vol. 34. 1007–1021.

[4] Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Cheston Tan, and Bryan Kian Hsiang Low. 2023. FedHQL: Federated Heterogeneous Q-Learning. In *International Conference on Autonomous Agents and Multiagent Systems*. 2810–2812.

[5] Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei-Tsang Ooi. 2024. FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF. *arXiv preprint arXiv:2412.15538* (2024).

[6] Bin Feng, Zhuping Liu, Gang Huang, and Chuangxin Guo. 2023. Robust federated deep reinforcement learning for optimal control in multiple virtual power plants with electric vehicles. *Applied Energy* 349 (2023), 121615.

[7] Wenzheng Jiang, Ji Wang, Xiongtao Zhang, Weidong Bao, Cheston Tan, and Flint Xiaofeng Fan. 2025. FedHPD: Heterogeneous Federated Reinforcement Learning via Policy Distillation. *arXiv preprint arXiv:2502.00870* (2025).

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

[9] Philip Jordan, Florian Grötschla, Flint Xiaofeng Fan, and Roger Wattenhofer. 2024. Decentralized Federated Policy Gradient with Byzantine Fault-Tolerance and Provably Fast Convergence. In *International Conference on Autonomous Agents and Multiagent Systems*. 964–972.

[10] Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581* (2019).

[11] Hei Yi Mak, Flint Xiaofeng Fan, Luca A Lanzendörfer, Cheston Tan, Wei Tsang Ooi, and Roger Wattenhofer. 2024. CAESAR: Enhancing Federated RL in Heterogeneous MDPs through Convergence-Aware Sampling with Screening. *arXiv preprint arXiv:2403.20156* (2024).

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*. 1273–1282.

[13] Xinlei Pan, Weiyao Wang, Xiaoshuai Zhang, Bo Li, Jinfeng Yi, and Dawn Song. 2019. How You Act Tells a Lot: Privacy-Leaking Attack on Deep Reinforcement Learning.. In *AAMAS 2019*, Vol. 19. 368–376.

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

[15] Jing Qiao, Zuyuan Zhang, Sheng Yue, Yuan Yuan, Zhipeng Cai, Xiao Zhang, Ju Ren, and Dongxiao Yu. 2024. BR-DeFedRL: Byzantine-Robust Decentralized Federated Reinforcement Learning with Fast Convergence and Communication Efficiency. In *IEEE Conference on Computer Communications*. IEEE, 141–150.

[16] Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy distillation. *arXiv preprint arXiv:1511.06295* (2015).

[17] Jiin Woo, Gauri Joshi, and Yuejie Chi. 2023. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. In *International Conference on Machine Learning*. 37157–37216.

[18] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. 2023. Heterogeneous federated learning: State-of-the-art and research challenges. *Comput. Surveys* 56, 3 (2023), 1–44.

[19] Yang Yu. 2018. Towards sample efficient reinforcement learning. In *International Joint Conference on Artificial Intelligence*. 5739–5743.

[20] Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. 2024. Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning. In *International Conference on Learning Representations*.