# Reinforcement Learning-Based Personalized Differentially Private Federated Learning

Xiaozhen Lu, *Member, IEEE,* Zihan Liu, Liang Xiao, *Senior Member, IEEE,* Huaiyu Dai, *Fellow, IEEE*

***Abstract*—Due to the different privacy and local model quality requirements for each participant, federated learning (FL) is vulnerable to membership inference attacks. To solve this issue, we propose a risk-aware reinforcement learning (RL)-based personalized differentially private FL framework. This framework uses local model accuracy and privacy loss as the constraints to satisfy the user's personalized requirements. By designing a multi-agent RL, this framework optimizes perturbation policy including perturbation mechanisms and parameters (such as privacy budget and probabilistic relaxation). The goal of each participant is to improve global accuracy and reduce privacy loss, attack success rate, and short-term risk value. Firstly, the framework designs a two-level hierarchical policy selection module to choose the perturbation policy to accelerate learning speed. Secondly, our proposed framework designs a punishment function to evaluate short-term risk and an R-network to estimate long-term risk, which guarantees safe exploration. Thirdly, this framework formulates an improved Boltzmann policy distribution to increase the impact of risk, thus avoiding risky policies that may cause severe privacy leakage or local task failure. We also analyze the convergence performance and provide privacy analysis for both Gaussian and Laplace mechanisms. Experimental results based on the MNIST dataset demonstrate the effectiveness of our framework compared with benchmarks.**

***Index Terms*—Federated learning, local differential privacy, personalized privacy, reinforcement learning, safe exploration.**

## I. INTRODUCTION

Due to the challenges of data growth and model training requirements, federated learning (FL) significantly speeds up training efficiency. This technique utilizes local and parallel computing with user privacy protection to facilitate collaborative model training across diverse organizations and devices [1], [2], [3]. Nevertheless, each participant has different demands for privacy and local model quality due to the variant local task types [4], [5]. For example, some of the vehicular participants focus on task execution quality such as objective detection accuracy, while others prefer protecting user privacy. In this case, each participant has personalized privacy and local model quality requirements. However, this issue makes the transmission process of local model parameters vulnerable to membership inference attacks. The attacker can wiretap the transmission status to infer that the underlying transmitted parameters come from which participants [6]. Existing methods mainly apply homomorphic encryption [7] or secure multi-party computation [8] to defend against membership inference attacks. For example, a privacy-preserving FL framework in [7] utilizes symmetric homomorphic encryption to ensure training confidentiality. However, this method induces high computational overhead due to the complex encryption and decryption operations.

Local differential privacy (LDP) techniques such as Laplace and Gaussian mechanisms have emerged as a promising solution for FL to protect user privacy [9]–[11]. Specifically, users add noise to perturb their local model parameters and then upload the perturbed parameters to the central server. In this way, the attacker cannot obtain the actual model parameters of users [12]. For example, the FL-enabled mobile system in [13] uses a fixed privacy budget to perturb local model parameters with noise sampled from Laplace distribution. On the other hand, a communication constraint-aware FL system proposed in [14] perturbs local models with a given Gaussian privacy budget and probabilistic relaxation. However, the privacy and the utility in the FL process are highly dependent on the adopted perturbation mechanisms and the corresponding privacy budget and probabilistic relaxation.

Therefore, reinforcement learning (RL) has been applied to choose the privacy budget for recommendation systems [15] and perturbation angle for indoor location protection [16]. For instance, a deep RL-based location privacy protection scheme was proposed in [16] to improve location protection and quality of service. This method applies a dueling double deep Q-network and asynchronous advantage actor-critic algorithms to optimize privacy budget and angles. By applying the Boltzmann policy distribution adopted in [17], the scheme also considers risky explorations that cause server location privacy leakage. The challenges of LDP-based FL can be summarized as follows:

- How to balance user privacy and utility in the personalized FL process should be investigated, to adaptively optimize user perturbation mechanisms and parameters.
- Without considering the joint impact of local model accuracy and user-specific privacy requirements, existing RL-based privacy protection schemes are prone to exploring risky policies.

This work was supported in part by the National Natural Science Foundation of China under Grants 62202222, U22B2062, U21A20444 and U22B2029, in part by the National Natural Science Foundation of Jiangsu Province under Grant BK20220880, in part by the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grant BK20222001, in part by the National Key Research and Development Program of China under Grant 2023YFB3107600, and in part by the Fundamental Research Funds for the Central Universities under Grant NJ202403. This work was also supported in part by the US National Science Foundation under grant ECCS-2203214. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation. (*Corresponding Author: Xiaozhen Lu.*)

X. Lu and Z. Liu are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Email: luxiaozhen@nuaa.edu.cn.

L. Xiao is with the Department of Informatics and Communication Engineering, Xiamen University, Xiamen, China. Email: lxiao@xmu.edu.cn.

H. Dai is with the Department of Electrical and Computer Engineering, North Carolina State University, NC, USA. Email: hdai@ncsu.edu.

- Most RL-based privacy protection schemes focus on the individual observation of each user, thus suffering from slow learning speed.

To solve the challenges, this paper proposes a risk-aware RL-based personalized differentially private FL framework (named RARL-PDPFL) against membership inference attacks. This framework enables mobile devices with personalized requirements to perturb their local model parameters following an LDP guarantee before uploading them to the central server. By designing a risk-aware RL algorithm for mobile devices, this framework jointly optimizes perturbation mechanism, privacy budget, and probabilistic relaxation to satisfy requirements of user-specific privacy and local model accuracy.

The designed risk-aware RL algorithm applies a hierarchical structure and observation-sharing mechanism to accelerate learning speed. More specifically, the hierarchical structure divides the perturbation policy into two sub-policies (i.e., the perturbation mechanism and parameters) to compress the action set, thus facilitating learning speed compared with [16] and [17]. Specifically, the first level in the hierarchical structure chooses the perturbation mechanism, and the second level selects the privacy budget and probabilistic relaxation to facilitate learning. This algorithm designs an observation-sharing mechanism for each user to extract the historical performance of neighboring users, to improve perturbation policy optimization efficiency.

Our RARL-PDPFL explores users' local model accuracy and privacy loss as the basis to avoid risky policies that cannot satisfy personalized user requirements. Different from the scheme in [16], our scheme uses a punishment function and an R-network to estimate both the short-term and long-term risks to reduce dangerous explorations. An improved Boltzmann policy distribution is designed to balance safe exploration and exploitation, which considers the impact of reward and risk on the selection of perturbation mechanisms and parameters. To our knowledge, our proposed framework is the first work that uses user-specific privacy and local model accuracy requirements in the design of RL to resist membership inference attacks.

We provide the convergence analysis of our framework and prove that our scheme satisfies the LDP guarantee under both Laplace and Gaussian mechanisms. Experiments are executed on the MNIST dataset, with results showing that our framework outperforms the benchmarks SFAC in [13] and Privatized FedPaq in [14]. Further, we provide an ablation study to verify the effectiveness of our proposed hierarchical policy selection module, punishment function, and R-network.

The main contributions of this work are summarized as follows:

- We propose a personalized differentially private FL framework to choose the perturbation policy for mobile devices against membership inference attacks. This framework innovatively uses privacy loss and local model accuracy as the criteria to satisfy the personalized user requirements.
- We design a multi-agent risk-aware RL algorithm to improve training accuracy and reduce user privacy loss. The designed algorithm avoids exploring risky policies that cannot satisfy user-specific requirements by evaluating both short-term and long-term risks. This algorithm also uses a hierarchical structure and observation sharing among mobile devices to accelerate learning speed.
- To verify the effectiveness of our framework, we prove that our framework has convergence performance and satisfies the LDP guarantee for both Gaussian and Laplace mechanisms. We also provide a comparison with three benchmarks via dynamic and personalized performance, the impact of participants, and the ablation study.

The rest of this paper is organized as follows. We provide the preliminaries in Section II and the system model in Section III. A risk-aware RL-based personalized differentially private FL framework is introduced in Section IV. Theoretical analysis is provided in Section V and experimental results are discussed in Section VI. We provide an overview of related works in Section VII, followed by the summary and future work in Section VIII.

## II. PRELIMINARIES

In this section, we provide preliminaries for local differential privacy, risk-aware reinforcement learning, and hierarchical reinforcement learning.

### A. Local Differential Privacy for Federated Learning

According to [18], $(\varepsilon_i^{(k)}, \delta_i^{(k)})$-LDP provides a rigorous privacy notion for the local model parameters in the FL training process. Specifically, the privacy budget $\varepsilon_i^{(k)} \in (0, 1]$ indicates the similarities between the original parameters and the perturbed parameters. On the other hand, the probabilistic relaxation $\delta_i^{(k)} \in (0, 1]$ represents the probability that the LDP has been violated. In the FL training process, LDP helps mobile device $i$ (with $1 \leq i \leq N$) perturb local model parameters $\boldsymbol{\omega}_i^{(k)}$ with $M$ dimensions by using Laplace or Gaussian mechanisms denoted by $\mathcal{M}(\boldsymbol{\omega}_i^{(k)})$.

*Definition 1:* ($(\varepsilon_i^{(k)},\delta_i^{(k)})$-LDP [19]). By applying a perturbation mechanism $\mathcal{M}(\cdot)$ in local model parameters $\boldsymbol{\omega}_i^{(k)}$, mobile device $i$ can achieve $(\varepsilon_i^{(k)}, \delta_i^{(k)})$-LDP if the perturbed parameters $\tilde{\boldsymbol{\omega}}_i^{(k)}$ satisfy:

$$\begin{aligned} &\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_i^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right] \\ &\leq \exp\left(\varepsilon_i^{(k)}\right) \Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_j^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right] + \delta_i^{(k)}. \end{aligned} \tag{1}$$

According to [20] and [21], the Gaussian mechanism is suitable for large-scale datasets but does not satisfy a strict LDP guarantee with a lighter tail than the Laplace mechanism. On the contrary, the Laplace mechanism strictly satisfies the LDP guarantee but may affect the data utility.

### B. Risk-Aware Reinforcement Learning

Compared with typical RL algorithms, risk-aware RL explores policies such as perturbation parameters in a constrained Markov decision process (CMDP) with security constraints. Specifically, CMDP consists of the state space, action set, reward, punishment, and transition probability. For mobile device $1 \leq i \leq N$, the corresponding CMDP is modeled as a tuple $\mathcal{M}\langle \mathbf{S}_i, \mathbf{A}_i, r_i^{(k)}, \psi_i^{(k)}, \mathcal{P}\rangle$, where

- $\mathbf{S}_i$ is the state space including all the possible states.
- $\mathbf{A}_i$ is the action set consisting of all the available policies.
- $r_i^{(k)}$ is the reward obtained from the environment if mobile device $i$ executes action $a_i^{(k)}$ at state $s_i^{(k)}$.
- $\psi_i^{(k)}$ is the punishment function that indicates the short-term risk of chosen action $a_i^{(k)}$ evaluated by the constraint metrics such as privacy loss and local model accuracy.
- $\mathcal{P}(\mathbf{S}_i \times \mathbf{A}_i \times \mathbf{S}_i) \in [0,1]$ represents the transition probability from $s_i^{(k)}$ to $s_i^{(k+1)}$ after performing $a_i^{(k)}$.

### C. Hierarchical Reinforcement Learning

As an extension of RL, hierarchical RL compresses large-scale action sets to accelerate the learning speed and considers the policy selection priority in the learning process [17]. For example, the selection priority of the perturbation mechanism is higher than that of perturbation parameters. Taking a two-level hierarchical structure as an example, an agent divides the original action set $\hat{\mathbf{A}}_i$ into two sub-action set $\hat{\mathbf{A}}_{i,1}$ and $\hat{\mathbf{A}}_{i,2}$. In this case, the agent uses the first level to choose the first sub-policy $a_{i,1}^{(k)} \in \hat{\mathbf{A}}_{i,1}$, which is used as the basis to select the second sub-policy $a_{i,2}^{(k)} \in \hat{\mathbf{A}}_{i,2}$ with the second level. Due to the FL system involving a large number of participants, hierarchical RL can be used to facilitate the optimization of perturbation policies.

## III. SYSTEM MODEL

### A. Network Model

As shown in Fig. 1, we consider an FL system that consists of a central server and $N$ mobile devices without sufficient computing resources. At time slot $k \in [1, K]$, each mobile device executes tasks with local model parameters $\boldsymbol{\omega}_i^{(k)}$. To avoid privacy leakage, mobile device $i$ decides how to perturb its model parameters $\boldsymbol{\omega}_i^{(k)}$, including perturbation mechanism $a_{i,1}^{(k)} \in \{0,1\}$ and parameters $\boldsymbol{x}_i^{(k)} = [\varepsilon_i^{(k)}, \delta_i^{(k)}]$. Specifically, $a_{i,1}^{(k)} = 0$ represents that mobile device $i$ applies the Gaussian mechanism and chooses the Laplace mechanism otherwise. Let $[u_{i,m}^{(k)}]_{1\le m\le M}$ and $[y_{i,m}^{(k)}]_{1\le m\le M}$ be the Gaussian and Laplace noise vector, respectively, where $u_{i,m}^{(k)}$ follows a Gaussian distribution $g(\cdot)$ and $y_{i,m}^{(k)}$ follows a Laplace distribution $\bar{g}(\cdot)$. Thus, mobile device $i$ perturbs its parameters to obtain perturbed model parameters $\tilde{\boldsymbol{\omega}}_i^{(k)}$ with Gaussian or Laplace noise via

$$\tilde{\boldsymbol{\omega}}_i^{(k)} = \boldsymbol{\omega}_i^{(k)} + \begin{cases} \left[u_{i,m}^{(k)}\right]_{1\le m\le M}, & \text{Gaussian noise} \\ \left[y_{i,m}^{(k)}\right]_{1\le m\le M}, & \text{Laplace noise.} \end{cases} \tag{2}$$

According to [22], privacy budget $\varepsilon_i^{(k)}$ and probabilistic relaxation $\delta_i^{(k)}$ are used to evaluate the perturbation noise scale $\tau_i^{(k)}$. After obtaining the perturbed parameters $\tilde{\boldsymbol{\omega}}_i^{(k)}$, mobile device $i$ sends $\tilde{\boldsymbol{\omega}}_i^{(k)}$ to the central server. As a metric to evaluate the privacy protection level, the privacy loss $\zeta_i^{(k)}$ is evaluated via

$$\zeta_i^{(k)} = \left(1 - \frac{\tau_i^{(k)} - \min\limits_{1\le k\le K} \tau_i^{(k)}}{\max\limits_{1\le k\le K} \tau_i^{(k)} - \min\limits_{1\le k\le K} \tau_i^{(k)}}\right) \varepsilon_i^{(k)}. \tag{3}$$

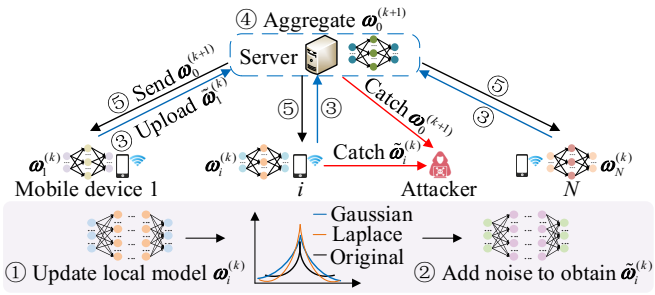



Fig. 1: Illustration of the system model, where $N$ mobile devices perturb their local model parameters to participate in the global aggregation.

By applying the FedAvg method as proposed in [23], the server gathers the perturbed local parameters of the $N$ mobile devices to aggregate a global model via

$$\boldsymbol{\omega}_0^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\boldsymbol{\omega}}_i^{(k)}. \tag{4}$$

The global model parameters $\boldsymbol{\omega}_0^{(k+1)}$ are distributed to the $N$ mobile devices.

Each mobile device uses the global model parameters $\boldsymbol{\omega}_0^{(k+1)}$ to update its local model parameters $\boldsymbol{\omega}_i^{(k+1)}$, with $1 \le i \le N$. For example, mobile device $i$ has a dataset with $|\mathcal{D}_i|$ pairs of training and test data, in which $p_{i,j}$ represents the $j$-th testing data and $q_{i,j}$ is the training data. The goal of mobile device $i$ is minimizing the loss function given by

$$\boldsymbol{\omega}_i^{(k+1)} = \boldsymbol{\omega}_0^{(k+1)} - \nabla_{\boldsymbol{\omega}_0^{(k+1)}} \frac{1}{|\mathcal{D}_i|} \sum_{j=1}^{|\mathcal{D}_i|} f_i\left(p_{i,j}, q_{i,j}; \boldsymbol{\omega}_0^{(k+1)}\right) \tag{5}$$

where $f_i(p_{i,j}, q_{i,j}; \boldsymbol{\omega}_0^{(k)})$ represents the prediction error of testing data $p_{i,j}$ under global model parameters $\boldsymbol{\omega}_0^{(k)}$. In this case, the error of local model $\boldsymbol{\omega}_i^{(k)}$ for mobile device $i$ can be modeled as

$$\tilde{l}_i^{(k)} = \frac{1}{|\mathcal{D}_i|} \sum_{j=1}^{|\mathcal{D}_i|} f_i\left(p_{i,j}, q_{i,j}; \boldsymbol{\omega}_i^{(k)}\right) \tag{6}$$

and the local model accuracy equals one minus the prediction error, i.e., $\rho_i^{(k)} = 1 - \tilde{l}_i^{(k)}$.

### B. Attack Model

In this work, an attacker performs membership inference attacks to capture the transmitted local model parameters. Specifically, it can obtain information such as gradients, change of hyper-parameters, and training times of the local model [24]. By analyzing the obtained information, the attacker can infer whether a specific piece of data (e.g., person, videos, or photos) belongs to the local dataset of a specific mobile device. What's worse, the attacker can even identify the source mobile device of the obtained local model parameters. This type of attack severely degrades the privacy of FL systems.

At time slot $k$, the attacker injects into the FL systems by launching malicious software, code, or scripts to monitor the

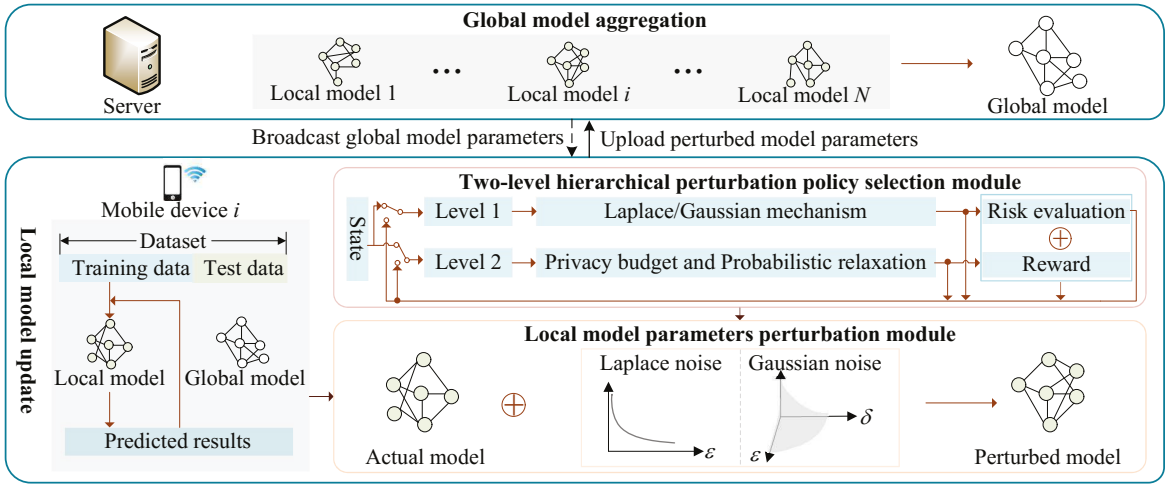



Fig. 2: Illustration of risk-aware RL-based personalized differentially private FL framework, with $1 \leq i \leq N$.

transmitted $\tilde{\boldsymbol{\omega}}_i^{(k)}$ and $\boldsymbol{\omega}_0^{(k)}$, with $1 \leq i \leq N$. To balance the attack overhead and profit, the attacker can apply an RL model such as deep Q-network or soft Actor-Critic to change its attack probability $\xi^{(k)}$ that indicates the attack frequency on $N$ mobile devices within a time slot. More specifically, the attacker estimates the previous attack success rate and measures its attack energy consumption in the last time slot, which are used to formulate its state. The attacker aims to increase its attack success rate denoted by $\phi^{(k)}$ among the $N$ participated mobile devices and save attack overhead. Similar to [25], attack success rate $\phi^{(k)}$ is estimated every time slot, which depends on the number of mobile devices, attack probability, and privacy loss given by

$$\phi^{(k)} = \xi^{(k)} \frac{\sum_{i=1}^{N} \zeta_i^{(k)}}{N}. \tag{7}$$

### C. Problem Formulation

In the FL system, each mobile device has personalized requirements of privacy and local model accuracy due to the different types and importance of their underlying local tasks. For example, mobile device $i$ has privacy loss requirement smaller than $\hat{\zeta}_i$ and local model accuracy constraint larger than $\hat{\rho}_i$, with $1 \leq i \leq N$. Thus, mobile device $i$ formulates its goal as an optimization function based global model accuracy denoted by $\rho_0^{(k)}$, local model accuracy $\rho_i^{(k)}$, privacy loss $\zeta_i^{(k)}$ and attack success rate $\phi^{(k)}$, i.e.,

$$\max_{a_i^{(k)} \in \{0,1\}, \varepsilon_i^{(k)} \in (0,1], \delta_i^{(k)} \in (0,1]} \mathbb{E}\left[\rho_0^{(k-1)} + \rho_i^{(k)} - \zeta_i^{(k)} - \phi^{(k-1)}\right] \tag{8}$$

$$\text{s.t.} \quad 1 \leq i \leq N, \rho_i^{(k)} \geq \hat{\rho}_i, \zeta_i^{(k)} \leq \hat{\zeta}_i.$$

## IV. Risk-Aware RL-based Personalized Differentially Private FL framework

We design a personalized differentially private FL framework with a risk-aware RL algorithm named RARL-PDPFL to optimize the perturbation mechanism and parameters. The framework explores the privacy loss and the accuracy of local models to satisfy the personalized requirements of each mobile device. As shown in Fig. 2, this framework includes the global model aggregation and local task execution, parameter perturbation, and perturbation policy selection module. By designing a multi-agent RL algorithm, this framework formulates a CMDP for each mobile device. This framework uses a two-level hierarchical perturbation policy selection module to accelerate learning efficiency. Further, a risk-aware policy distribution that relies on both short-term and long-term risks is used to avoid exploring potentially dangerous perturbation policies.

### A. Constrained Markov Decision Process

This framework formulates the personalized differentially private FL process as a CMDP, with details introduced as follows.

**State Space:** Mobile device $i$ (with $1 \leq i \leq N$) estimates privacy loss $\zeta_i^{(k-1)}$ and local model accuracy $\rho_i^{(k-1)}$, and obtains global model accuracy $\rho_0^{(k-1)}$ from the central server. By analyzing the number of received spam or advertising times, and the shared information of other devices and the server, mobile device $i$ estimates attack success rate $\phi^{(k-1)}$. This framework enables each mobile device to share observations to reduce unnecessary random exploration. By exploiting the shared observations of the rest $N-1$ mobile devices, mobile device $i$ builds its state as

$$\boldsymbol{s}_i^{(k)} = \left[\rho_0^{(k-1)}, \phi^{(k-1)}, \zeta_i^{(k-1)}, \left[\rho_j^{(k-1)}\right]_{1 \leq j \leq N}\right] \in \mathbf{S}_i. \tag{9}$$

Particularly, mobile device $i$ extracts the shared observations from the other devices and the previous $\rho_j^{(k-2)}$ to formulate its state, if mobile device $j$ does not share its local mode accuracy $\rho_j^{(k-1)}$, with $1 \leq i \neq j \leq N$. Due to the usage of a hierarchical structure, our proposed algorithm can accelerate the learning speed even if one or more devices do not share their local model accuracy.

**Action set:** Each mobile device determines its perturbation policy to add noise to its local parameters. Taking mobile device $i$ as an example, perturbation policy $\boldsymbol{a}_i^{(k)} = [a_{i,1}^{(k)}, \boldsymbol{x}_i^{(k)}] \in \mathbf{A}_i$ includes perturbation mechanism $a_{i,1}^{(k)}$ with two available

choices, and perturbation parameters $\boldsymbol{x}_i^{(k)}$. Specifically, the perturbation parameters contain privacy budget $\varepsilon_i^{(k)}$ having $\Omega_{i,1}$ levels and probabilistic relaxation $\delta_i^{(k)}$ having $\Omega_{i,2}$ levels. Thus, $\mathbf{A}_i$ has $2\Omega_{i,1}\Omega_{i,2}$ available perturbation policies.

**Reward:** After executing the chosen perturbation policy $\boldsymbol{a}_i^{(k)}$, mobile device $i$ perturbs $\boldsymbol{\omega}_i^{(k)}$ with mechanism $a_{i,1}^{(k)}$, privacy budget $\varepsilon_i^{(k)}$ and probabilistic relaxation $\delta_i^{(k)}$ to obtain $\tilde{\boldsymbol{\omega}}_i^{(k)}$. Based on the prediction results of local tasks, mobile device $i$ estimates local model accuracy $\rho_i^{(k)}$ via Eq. (6) and privacy loss $\zeta_i^{(k)}$ via Eq. (3). This scheme uses the reward $r_i^{(k)}$ to represent the immediate profit obtained from the environment after performing the chosen perturbation policy, which is composed of the model accuracy, privacy loss, and attack success rate. The reward is evaluated via

$$r_i^{(k)} = \rho_0^{(k-1)} + v_1\rho_i^{(k)} - v_2\zeta_i^{(k)} - v_3\phi^{(k-1)}. \tag{10}$$

**Punishment function:** Each mobile device uses the personalized local model accuracy and privacy loss requirements as security constraints to measure the risk of a chosen perturbation policy. To make a trade-off between privacy protection and training accuracy, mobile device $i$ can tolerate a maximum privacy loss bounded by $\hat{\zeta}_i$ and a minimum local model accuracy bounded by $\hat{\rho}_i$. This scheme designs a punishment function (i.e., the risk value) to evaluate whether the chosen perturbation policy satisfies the user-specific privacy and local model quality requirements. The risk value $\psi_i^{(k)}$ is calculated based on an indicator function $\mathbb{I}(\cdot)$ in terms of privacy loss and local model accuracy, i.e.,

$$\psi_i^{(k)} = \mathbb{I}\left(\rho_i^{(k)} < \hat{\rho}_i\right) + v_4\mathbb{I}\left(\zeta_i^{(k)} > \hat{\zeta}_i\right), \tag{11}$$

where $v_4$ parameterizes the importance of privacy and training performance for risk formulation.

This framework modifies reward $r_i^{(k)}$ with risk value $\psi_i^{(k)}$ and weight $v_5$ as the short-term reward, to avoid the immediate dangerous exploration. The modified reward $\hat{r}_i^{(k)}$ is calculated by

$$\hat{r}_i^{(k)} = r_i^{(k)} - v_5\psi_i^{(k)}. \tag{12}$$

Different from [26], this algorithm uses the modified reward $\hat{r}_i^{(k)}$ to update both the perturbation policy distribution and weights of Q-networks.

### B. Two-Level Hierarchical Policy Selection Module

As illustrated in Fig. 3, our designed two-level hierarchical module consists of two Q-networks that estimate long-term expected reward (also called Q-values), and two R-networks that estimate long-term risk (i.e., R-values). With state $\boldsymbol{s}_i^{(k)}$ that has $N+3$ dimensions as the input, the first level is used to select perturbation mechanism $a_{i,1}^{(k)}$ with 2 dimensions. The second level chooses perturbation parameters $\boldsymbol{x}_i^{(k)}$ with state $\boldsymbol{s}_i^{(k)}$ and $a_{i,1}^{(k)}$ as the input, in which the output have $\Omega_{i,1}\Omega_{i,2}$ dimensions.

In first-level, Q-network 1 using weights $\boldsymbol{\theta}_{i,1}^{(k)}$ has an input layer with size of $N+3$, a hidden layer with $f_{i,1}$ neurons, and an output layer that outputs two Q-values $\mathbf{Q}_{i,1}(\boldsymbol{s}_i^{(k)}, \cdot; \boldsymbol{\theta}_{i,1}^{(k)})$. Similarly, R-network 1 with weights $\boldsymbol{\varphi}_{i,1}^{(k)}$ involves the same network architecture as Q-network 1, which outputs two R-values $\mathbf{R}_{i,1}(\boldsymbol{s}_i^{(k)}, \cdot; \boldsymbol{\varphi}_{i,1}^{(k)})$. In second-level, Q-network 2 with weights $\boldsymbol{\theta}_{i,2}^{(k)}$ uses state $\boldsymbol{s}_i^{(k)}$ and chosen perturbation mechanism $a_{i,1}^{(k)}$ from first-level as input. The corresponding input layer has $N+4$ neurons, hidden layer involves $f_{i,2}$ neurons, and output layer outputs $\Omega_{i,1}\Omega_{i,2}$ number of Q-values $\mathbf{Q}_{i,2}(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}, \cdot; \boldsymbol{\theta}_{i,2}^{(k)})$. By using a same network architecture as Q-network 2, R-network 2 involving weights $\boldsymbol{\varphi}_{i,2}^{(k)}$ outputs $\Omega_{i,1}\Omega_{i,2}$ number of R-values $\mathbf{R}_{i,2}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}, \cdot; \boldsymbol{\varphi}_{i,2}^{(k)}\right)$.

### C. Risk-Aware Policy Distribution

This framework designs a risk-aware policy distribution to help each mobile device enable safety during the learning process. More specifically, the long-term expected reward in both the two levels is updated with $\hat{r}_i^{(k)}$ using the Bellman iteration equation, and the corresponding long-term risk is updated with $\psi_i^{(k)}$. A learning rate $\alpha_i \in (0, 1]$ in the Bellman iteration function balances the importance of the future reward/risk in the learning process. Taking the first level as an example, the long-term expected reward and risk are updated with

$$\begin{aligned}\mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}; \boldsymbol{\theta}_{i,1}^{(k)}\right) \leftarrow (1-\alpha_i)\,\mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}; \boldsymbol{\theta}_{i,1}^{(k)}\right) \\ + \alpha_i\left(\hat{r}_i^{(k)} + \arg\max_{a^*\in\{0,1\}} \mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(k+1)}, a^*; \boldsymbol{\theta}_{i,1}^{(k)}\right)\right),\end{aligned} \tag{13}$$

$$\begin{aligned}\mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}; \boldsymbol{\varphi}_{i,1}^{(k)}\right) \leftarrow (1-\alpha_i)\,\mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}; \boldsymbol{\varphi}_{i,1}^{(k)}\right) \\ + \alpha_i\left(\psi_i^{(k)} + \arg\min_{a^*\in\{0,1\}} \mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(k+1)}, a^*; \boldsymbol{\varphi}_{i,1}^{(k)}\right)\right).\end{aligned} \tag{14}$$

This framework formulates the policy distribution (i.e., the probability to choose perturbation mechanism $a_{i,1}^{(k)}$ or parameters $\boldsymbol{x}_i^{(k)}$) via Eq. (15). In this case, our framework increases the impact of risk on the perturbation policy selection, thus reducing risky explorations compared with [17].

### D. Network Update

Similar to [27], this framework applies Adam as the gradient descent algorithm in the experience replay technique. Compared with the stochastic gradient descent algorithm, the Adam algorithm uses a running average of the first and second moment for the gradient to update the network weights with fewer iterations, which is more efficient for large-scale FL systems. Specifically, mobile device $i$ saves the current state, chosen perturbation policy, short-term reward, and risk value to formulate an experience sequence into a replay buffer $\mathfrak{D}$, i.e., $\mathfrak{D} \leftarrow \mathfrak{D} \cup \{\boldsymbol{s}_i^{(k)}, \boldsymbol{a}_i^{(k)}, \hat{r}_i^{(k)}, \psi_i^{(k)}\}$. By randomly sampling $Z$ experiences from replay buffer $\mathfrak{D}$ as

$$\mathfrak{B} = \{\boldsymbol{s}_i^{(h(n))}, \boldsymbol{a}_i^{(h(n))}, \hat{r}_i^{(h(n))}, \psi_i^{(h(n))}\}_{1\le n\le Z}, \tag{16}$$

mobile device $i$ updates the networks, where $h(n)$ follows a uniform distribution $\mathrm{U}(1, k)$ that relies on the built replay buffer. The goal is to minimize the loss of Q-network and R-network of each level, i.e., the difference between estimated Q/R-values and target values.

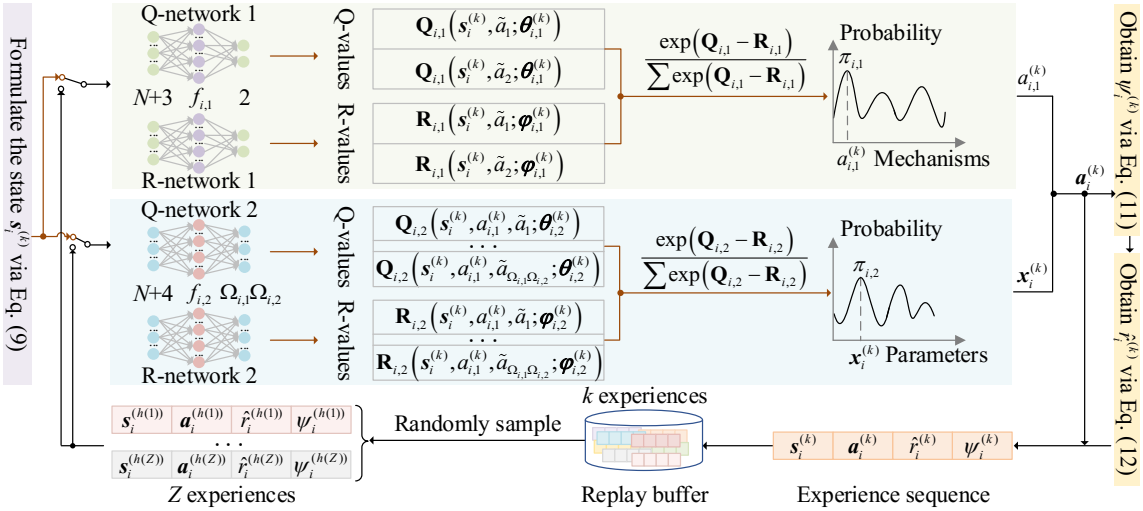



Fig. 3: Flowchart of the risk-aware RL-based perturbation policy selection for mobile device $i$, with $1 \le i \le N$.

$$\pi_{i,j}\left(\boldsymbol{s}_i^{(k)}, \boldsymbol{a}^*; \boldsymbol{\theta}_{i,j}^{(k)}, \boldsymbol{\varphi}_{i,j}^{(k)}\right) = \begin{cases} \frac{\exp\left(\mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}; \boldsymbol{\theta}_{i,1}^{(k)}\right) - \mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}; \boldsymbol{\varphi}_{i,1}^{(k)}\right)\right)}{\sum_{\hat{a}_{i,1}\in\{0,1\}} \exp\left(\mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(k)}, \hat{a}_{i,1}; \boldsymbol{\theta}_{i,1}^{(k)}\right) - \mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(k)}, \hat{a}_{i,1}; \boldsymbol{\varphi}_{i,1}^{(k)}\right)\right)}, & \text{if } j = 1, \boldsymbol{a}^* = a_{i,1}^{(k)} \\ \frac{\exp\left(\mathbf{Q}_{i,2}\left(\boldsymbol{s}_i^{(k)}, \boldsymbol{a}_i^{(k)}; \boldsymbol{\theta}_{i,2}^{(k)}\right) - \mathbf{R}_{i,2}\left(\boldsymbol{s}_i^{(k)}, \boldsymbol{a}_i^{(k)}; \boldsymbol{\varphi}_{i,2}^{(k)}\right)\right)}{\sum_{\hat{\mathbf{a}}\in\mathbf{A}_{i,2}} \exp\left(\mathbf{Q}_{i,2}\left(\boldsymbol{s}_i^{(k)}, \hat{\mathbf{a}}; \boldsymbol{\theta}_{i,2}^{(k)}\right) - \mathbf{R}_{i,2}\left(\boldsymbol{s}_i^{(k)}, \hat{\mathbf{a}}; \boldsymbol{\varphi}_{i,2}^{(k)}\right)\right)}, & \text{if } j = 2, \boldsymbol{a}^* = \boldsymbol{a}_i^{(k)}. \end{cases} \tag{15}$$

Thus, the update of weights for Q-network 1 by minimizing the loss function is given by

$$\mathcal{L}_{i,1}\left(\boldsymbol{\theta}_{i,1}^{(k)}\right) = \min_{\hat{\boldsymbol{\theta}}} \frac{1}{Z}\sum_{n=1}^{Z}\left[\left(\hat{r}_i^{(h(n))} - \mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(h(n))}, a_{i,1}^{(h(n))}; \hat{\boldsymbol{\theta}}\right) + \gamma_{i,1}\max_{\hat{a}\in\{0,1\}} \mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(h(n)+1)}, \hat{a}_{i,1}; \boldsymbol{\theta}_{i,1}^{-}\right)\right)^2\right], \tag{17}$$

where $\gamma_{i,1}$ is a discount factor and $\boldsymbol{\theta}_{i,1}^{-}$ is the weights of target Q-network 1. Similarly, the update of the weights for R-network 1 is given by

$$\bar{\mathcal{L}}_{i,1}\left(\boldsymbol{\varphi}_{i,1}^{(k)}\right) = \min_{\hat{\boldsymbol{\varphi}}} \frac{1}{Z}\sum_{n=1}^{Z}\left[\left(\psi_i^{(h(n))} - \mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(h(n))}, a_{i,1}^{(h(n))}; \hat{\boldsymbol{\varphi}}\right) + \gamma_{i,1}\min_{\hat{a}_{i,1}\in\{0,1\}} \mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(h(n)+1)}, \hat{a}_{i,1}; \boldsymbol{\varphi}_{i,1}^{-}\right)\right)^2\right], \tag{18}$$

where $\boldsymbol{\varphi}_{i,1}^{-}$ is the weights of target R-network 1. The update of weights for Q-network 2 and R-network 2 are similar to Eqs. (17) and (18), respectively.

According to [17], the computational complexity of the proposed RARL-PDPFL is given by $O(\sum_{i=1}^{N} Z\sqrt{k^3\Omega_{i,1}\Omega_{i,2}}(N+\Omega_{i,1}\Omega_{i,2}))$, which highly depends on the number of mobile devices and learning samples. Our RARL-PDPFL method is summarized in **Algorithm 1**.

## V. THEORETICAL ANALYSIS

In this section, we analyze the convergence performance and prove that our scheme satisfies the LDP guarantee under both Gaussian and Laplace mechanisms.

### A. Convergence Analysis

By Eq. (5), mobile device $i$ has loss of local model given by

$$F_i\left(\boldsymbol{\omega}_i^{(k)}\right) \triangleq \frac{1}{|\mathcal{D}_i|}\sum_{j=1}^{|\mathcal{D}_i|} f_i\left(p_{i,j}, q_{i,j}; \boldsymbol{\omega}_0^{(k)}\right), 1 \le i \le N. \tag{19}$$

Similarly, the loss of the global model denoted by $F_0(\boldsymbol{\omega}_0^{(k)})$ is given by

$$F_0\left(\boldsymbol{\omega}_0^{(k)}\right) = \frac{1}{N}\sum_{i=1}^{N} F_i\left(\tilde{\boldsymbol{\omega}}_i^{(k)}\right). \tag{20}$$

*Assumption 1:* $F_0(\cdot)$ is assumed to satisfy the Polyak-Lojasiewicz inequality with positive parameter $d$ and $(\|\nabla F(\boldsymbol{\omega}_0^{(k)})\|_2)^2 \le \beta$, i.e.,

$$\mathbb{E}\left[F_0\left(\boldsymbol{\omega}_0^{(k)}\right) - F_0\left(\boldsymbol{\omega}_0^{*}\right)\right] \le \frac{1}{2d}\left(\left\|\nabla F_0\left(\boldsymbol{\omega}_0^{(k)}\right)\right\|_2\right)^2. \tag{21}$$

*Assumption 2:* $F_i(\cdot), 1 \le i \le N$ is satisfying $G$-Lipschitz smooth, and we have

$$F_i\left(\boldsymbol{\omega}_i^{(k+1)}\right) \le F_i\left(\boldsymbol{\omega}_i^{(k)}\right) + \nabla F_i\left(\boldsymbol{\omega}_i^{(k)}\right)^{\top}\left(\boldsymbol{\omega}_i^{(k+1)} - \boldsymbol{\omega}_i^{(k)}\right) + \frac{G}{2}\left(\left\|\boldsymbol{\omega}_i^{(k+1)} - \boldsymbol{\omega}_i^{(k)}\right\|_2\right)^2, 1 \le i \le N. \tag{22}$$

*Assumption 3:* For $\forall \boldsymbol{\omega}_i^{(k)}$ and $\boldsymbol{\omega}_0^{(k)}$, we have

$$\mathbb{E}\left[\left(\left\|\nabla F_i\left(\boldsymbol{\omega}_i^{(k)}\right)\right\|_2\right)^2\right] \le \left(\left\|\nabla F_0\left(\boldsymbol{\omega}_0^{(k)}\right)\right\|_2\right)^2 B^2. \tag{23}$$

**Algorithm 1** Risk-aware RL-based personalized differentially private FL algorithm

Initialize $N$, $K$, $Z$, $\boldsymbol{\omega}_0^{(0)}$, $\rho_0^{(0)}$, $\phi^{(0)}$, $\mathcal{D} = \varnothing$, $\left\{\boldsymbol{\theta}_{i,j}^{(0)}, \boldsymbol{\varphi}_{i,j}^{(0)}, \rho_i^{(0)}, \zeta_i^{(0)}\right\}_{1\le j\le 2, 1\le i\le N}$
**for** $k = 1, 2, \cdots, K$ **do**
  **for** $i = 1, 2, \cdots, N$ **do**
    Download global model parameters $\boldsymbol{\omega}_0^{(k)}$
    Obtain $\left[\rho_j^{(k-1)}\right]_{1\le j\ne i\le N}$ from $N-1$ mobile devices
    Formulate state $\boldsymbol{s}_i^{(k)}$ via Eq. (9)
    Input $\boldsymbol{s}_i^{(k)}$ to the first-level
    Obtain $\mathbf{Q}_{i,1}\left(\boldsymbol{s}_i^{(k)}, \cdot; \boldsymbol{\theta}_{i,1}^{(k)}\right)$ and $\mathbf{R}_{i,1}\left(\boldsymbol{s}_i^{(k)}, \cdot; \boldsymbol{\varphi}_{i,1}^{(k)}\right)$
    Choose perturbation mechanism $a_{i,1}^{(k)}$ via Eq. (15)
    Input $\boldsymbol{s}_i^{(k)}$ and $a_{i,1}^{(k)}$ into the second-level
    Second level outputs $\mathbf{Q}_{i,2}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}, \cdot; \boldsymbol{\theta}_{i,2}^{(k)}\right)$ and $\mathbf{R}_{i,2}\left(\boldsymbol{s}_i^{(k)}, a_{i,1}^{(k)}, \cdot; \boldsymbol{\varphi}_{i,2}^{(k)}\right)$
    Choose perturbation parameters $\boldsymbol{x}_i^{(k)}$ via Eq. (15)
    Perform local task to obtain $\boldsymbol{\omega}_i^{(k)}$
    Perturb local model parameters as $\tilde{\boldsymbol{\omega}}_i^{(k)}$ via Eq. (2)
    Upload $\tilde{\boldsymbol{\omega}}_i^{(k)}$ to the central sever
    Evaluate $\rho_i^{(k)}$ via Eq. (6)
    Compute privacy loss $\zeta_i^{(k)}$ via Eq. (3)
    Estimate attack success rate $\phi^{(k)}$ via Eq. (7)
    Compute immediate reward $r_i^{(k)}$ via Eq. (10)
    Compute short-term risk $\psi_i^{(k)}$ via Eq. (11)
    Modify reward $\hat{r}_i^{(k)}$ via Eq. (12)
    $\mathfrak{D} \leftarrow \mathfrak{D} \cup \left\{\boldsymbol{s}_i^{(k)}, \boldsymbol{a}_i^{(k)}, \hat{r}_i^{(k)}, \psi_i^{(k)}\right\}$
    Sample $Z$ experiences from replay buffer $\mathfrak{D}$
    Update network weights via Eqs. (17) and (18)
  **end for**
**end for**

*Theorem 1:* Our proposed RARL-PDPFL has convergence performance given by

$$\mathbb{E}\left[F_0\left(\boldsymbol{\omega}_0^{(K)}\right) - F_0\left(\boldsymbol{\omega}_0^*\right)\right] \le \left(1 + 2d\left(-\frac{1}{\mu} + \frac{BG}{\mu(\mu+d)} + \frac{GB^2}{2(\mu+d)^2}\right)\right)^K \mathbb{E}\left[F_0\left(\boldsymbol{\omega}_0^{(0)}\right) - F_0\left(\boldsymbol{\omega}_0^*\right)\right] + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\beta}{\mu} + \frac{GB\beta}{\mu+d} + \frac{G}{2}\right)\frac{\left(1 + 2d\left(-\frac{1}{\mu} + \frac{BG}{\mu(\mu+d)} + \frac{GB^2}{2(\mu+d)^2}\right)\right)^K - 1}{2d\left(-\frac{1}{\mu} + \frac{BG}{\mu(\mu+d)} + \frac{GB^2}{2(\mu+d)^2}\right)} \max\left\{\frac{M \max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(K)} - \boldsymbol{\omega}_j^{(K)}\right\|_2\right\}\sqrt{2\ln\left(\frac{1.25}{\min\limits_{\hat{\delta}_i\in(0,1]}\hat{\delta}_i}\right)}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}, 2M\left(\frac{\max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(K)} - \boldsymbol{\omega}_j^{(K)}\right\|_1\right\}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}\right)^2\right\}. \quad (24)$$

*Proof 1:* If mobile device $i$ chooses perturbs local model parameters with Gaussian noise at time slot $k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left[u_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right] \le \mathbb{E}\left[\left(\left\|\left[u_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right)^2\right] \le \mathbb{E}\left[\sum_{m=1}^{M}\left(u_{i,m}^{(k)}\right)^2\right] = M\left(\sigma_i^{(k)}\right)^2 \\
&\le \frac{M \max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right\}\sqrt{2\ln\left(\frac{1.25}{\min\limits_{\hat{\delta}_i\in(0,1]}\hat{\delta}_i}\right)}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}. \quad (25)
\end{aligned}$$

If mobile device $i$ chooses Laplace mechanism, we have

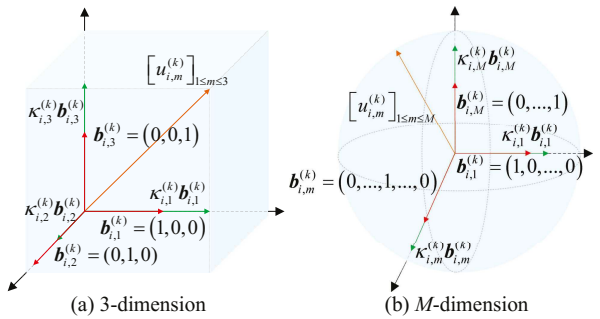
$$\begin{aligned}
&\mathbb{E}\left[\left\|\left[y_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right] \le \mathbb{E}\left[\left(\left\|\left[y_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right)^2\right] \\
&\le 2M\left(\frac{\max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_1\right\}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}\right)^2. \quad (26)
\end{aligned}$$

Let $\boldsymbol{\eta}_i^{(k)}$ denote noise vector, which equals to $[u_{i,m}^{(k)}]_{1\le m\le M}$ with Gaussian mechanism while is $[y_{i,m}^{(k)}]_{1\le m\le M}$ with Laplace mechanism. Thus, we have

$$\mathbb{E}\left[\left\|\boldsymbol{\eta}_i^{(k)}\right\|_2\right] \le \max\left\{2M\left(\frac{\max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_1\right\}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}\right)^2, \frac{M \max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right\}\sqrt{2\ln\left(\frac{1.25}{\min\limits_{\hat{\delta}_i\in(0,1]}\hat{\delta}_i}\right)}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}\right\}. \quad (27)$$

Thus, according to [28] and Eq. (27), we have

$$\mathbb{E}\left[F_0\left(\boldsymbol{\omega}_0^{(k)}\right) - F_0\left(\boldsymbol{\omega}_0^*\right)\right] \le \left(1 + 2d\left(-\frac{1}{\mu} + \frac{BG}{\mu(\mu+d)} + \frac{GB^2}{2(\mu+d)^2}\right)\right)\mathbb{E}\left[F_0\left(\boldsymbol{\omega}_0^{(k-1)}\right) - F_0\left(\boldsymbol{\omega}_0^*\right)\right] + \left(\frac{\beta}{\mu} + \frac{GB\beta}{\mu+d} + \frac{G}{2}\right)\left(\frac{1}{N}\sum_{i=1}^{N} \max\left\{\frac{M \max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right\}\sqrt{2\ln\left(\frac{1.25}{\min\limits_{\hat{\delta}_i\in(0,1]}\hat{\delta}_i}\right)}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}, 2M\left(\frac{\max\limits_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_1\right\}}{\min\limits_{\hat{\varepsilon}_i\in(0,1]}\hat{\varepsilon}_i}\right)^2\right\}\right). \quad (28)$$




(a) 3-dimension (b) $M$-dimension

Fig. 4: Decomposition of Gaussian noise vector $[u_{i,m}^{(k)}]_{1\le m\le M}$.

Thus, we can obtain Eq. (24).

*Remark 1:* After $K$ time slots, the difference or gap between the training global model $\boldsymbol{\omega}_0^{(K)}$ and the optimal model $\boldsymbol{\omega}_0^*$ is lower than a bound, which relies on the Lipschitz smooth coefficient $G$, number of mobile devices $N$, and dimensions of local model parameters (i.e., $M$). According to Eq. (5), the optimal local model parameters is given by

$$\boldsymbol{\omega}_i^* = \boldsymbol{\omega}_0^* - \frac{1}{|\mathcal{D}_i|}\sum_{j=1}^{|\mathcal{D}_i|} f_i\left(p_{i,j}, q_{i,j}; \boldsymbol{\omega}_0^*\right). \tag{29}$$

In this case, the optimal global model accuracy $\rho_0^*$ and local model accuracy $\rho_i^*$ are calculated by

$$\rho_0^* = 1 - F\left(\boldsymbol{\omega}_0^*\right), \tag{30}$$

$$\rho_i^* = 1 - \frac{1}{|\mathcal{D}_i|}\sum_{j=1}^{|\mathcal{D}_i|} f_i\left(p_{i,j}, q_{i,j}; \boldsymbol{\omega}_i^*\right), 1 \le i \le N. \tag{31}$$

### B. Privacy Analysis

As shown in Fig. 4, the Gaussian noise vector $[u_{i,m}^{(k)}]_{1\le m\le M}$ can be decomposed into a linear combination of $M$ unit vectors $[\boldsymbol{b}_{i,m}]_{1\le m\le M}$ as

$$\begin{aligned}&\left[u_{i,m}^{(k)}\right]_{1\le m\le M} = \kappa_{i,1}^{(k)}\boldsymbol{b}_{i,1}^{(k)} + \cdots + \kappa_{i,M}^{(k)}\boldsymbol{b}_{i,M}^{(k)}, 1 \le i \le N\\ &\text{s.t.} \quad \forall \kappa_{i,m}^{(k)} \sim \mathcal{N}\left(0, \left(\sigma_i^{(k)}\right)^2\right), \left\langle \boldsymbol{b}_{i,m}^{(k)}, \boldsymbol{b}_{i,\tilde{m}}^{(k)} \right\rangle = 0.\end{aligned} \tag{32}$$

*Theorem 2:* For mobile device $i$, our proposed framework satisfies $(\varepsilon_0^*, \delta_0^*)$-LDP guarantee, where $\varepsilon_0^*$ and $\delta_0^*$ are upper bounds of privacy budget and probabilistic relaxation after $K$ time slots.

*Proof 2:* By Eq. (32), if mobile device $i$ chooses the Gaussian mechanism, for $\forall 1 \le k \le K$, we have

$$\left(\left\|\left[u_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right)^2 = \left(\kappa_{i,1}^{(k)}\right)^2 + \left(\left\|\sum_{m=1}^{M}\kappa_{i,m+1}^{(k)}\boldsymbol{b}_{i,m+1}^{(k)}\right\|_2\right)^2, \tag{33}$$

$$\begin{aligned}&\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)} + \left[u_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right)^2\\ &= \left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2 + \kappa_{i,1}^{(k)}\right)^2 + \left(\left\|\sum_{m=1}^{M}\kappa_{i,m+1}^{(k)}\boldsymbol{b}_{i,m+1}^{(k)}\right\|_2\right)^2.\end{aligned} \tag{34}$$

By Eq. (1) and [20], for $1 \le i \ne j \le N$, we have

$$\begin{aligned}&\left|\ln\left(\frac{\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_i^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right]}{\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_j^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right]}\right)\right|\\ &= \left|\ln\left(\frac{g\left(\left\|\left[u_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right)}{g\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)} + \left[u_{i,m}^{(k)}\right]_{1\le m\le M}\right\|_2\right)}\right)\right|\\ &= \left|\frac{1}{2\left(\sigma_i^{(k)}\right)^2}\left(\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right)^2 + 2\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\kappa_{i,1}^{(k)}\right)\right|\\ &\le \left|\frac{1}{2\left(\sigma_i^{(k)}\right)^2}\left(\max_{1\le i\ne j\le N}\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right)^2\right.\right.\\ &\left.\left.+2\kappa_{i,1}^{(k)}\max_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right\}\right)\right|.\end{aligned} \tag{35}$$

Thus, we have

$$\left|\kappa_{i,1}^{(k)}\right| \le \frac{2\left(\sigma_i^{(k)}\right)^2\varepsilon_i^{(k)} - \max_{1\le i\ne j\le N}\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right)^2}{2\max_{1\le i\ne j\le N}\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2}. \tag{36}$$

According to [29], we have

$$\begin{aligned}&\Pr\left[\left|\kappa_{i,1}^{(k)}\right| > \frac{2\left(\sigma_i^{(k)}\right)^2\varepsilon_i^{(k)} - \max_{1\le i\ne j\le N}\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right)^2}{2\max_{1\le i\ne j\le N}\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2}\right]\\ &< \frac{\sqrt{2\pi}\sigma_i^{(k)}}{\pi}\exp\left(\frac{-\left(\frac{2\left(\sigma_i^{(k)}\right)^2\varepsilon_i^{(k)} - \max_{1\le i\ne j\le N}\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right)^2}{2\max_{1\le i\ne j\le N}\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2}\right)^2}{2\left(\sigma_i^{(k)}\right)^2}\right).\end{aligned} \tag{37}$$

According to [20], we have

$$\sigma_i^{(k)} \ge \max_{1\le i\ne j\le N}\left\{\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right\}\frac{\sqrt{2\ln\left(\frac{1.25}{\delta_i^{(k)}}\right)}}{\varepsilon_i^{(k)}}. \tag{38}$$

Thus, we have

$$\Pr\left[\left|\kappa_{i,1}^{(k)}\right| > \frac{2\left(\sigma_i^{(k)}\right)^2\varepsilon_i^{(k)} - \max_{1\le i\ne j\le N}\left(\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2\right)^2}{2\max_{1\le i\ne j\le N}\left\|\boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right\|_2}\right] \le \delta_i^{(k)}. \tag{39}$$

By Eq. (1), we have

$$\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_i^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right] \le \exp\left(\varepsilon_i^{(k)}\right)\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_j^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right] + \delta_i^{(k)}. \tag{40}$$

According to Eq. (1) and [21], if mobile device $i$ chooses Laplace mechanism, for $\forall 1 \le k \le K$, we have

$$\left|\ln\left(\frac{\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_i^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right]}{\Pr\left[\mathcal{M}\left(\boldsymbol{\omega}_j^{(k)}\right) = \tilde{\boldsymbol{\omega}}_i^{(k)}\right]}\right)\right|$$

TABLE I: Parameters of hierarchical policy selection module for each mobile device

| Networks | Input size | Hidden neurons | Output size | Activation | Size of replay buffer | Sampling size | Learning rate |
|---|---|---|---|---|---|---|---|
| Q/R-network 1 | 6 | 32 | 2 | ReLU | 5000 | 16 | 0.0001 |
| Q/R-network 2 | 7 | 32 | 9 | | | | |

$$
\begin{aligned}
&= \left| \ln \left( \frac{\bar{g}\left(\tilde{\boldsymbol{\omega}}_i^{(k)} - \boldsymbol{\omega}_i^{(k)}\right)}{\bar{g}\left(\tilde{\boldsymbol{\omega}}_i^{(k)} - \boldsymbol{\omega}_j^{(k)}\right)} \right) \right| \\
&= \exp\left( \frac{\varepsilon_i^{(k)} \left( \left\| \tilde{\boldsymbol{\omega}}_i^{(k)} - \boldsymbol{\omega}_i^{(k)} \right\|_1 - \left\| \tilde{\boldsymbol{\omega}}_i^{(k)} - \boldsymbol{\omega}_j^{(k)} \right\|_1 \right)}{\max\limits_{1 \le i \ne j \le N} \left\| \boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)} \right\|_1} \right) \\
&\le \exp\left( \frac{\varepsilon_i^{(k)} \left( \left\| \boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)} \right\|_1 \right)}{\max\limits_{1 \le i \ne j \le N} \left\| \boldsymbol{\omega}_i^{(k)} - \boldsymbol{\omega}_j^{(k)} \right\|_1} \right) \\
&\le \exp\left( \varepsilon_i^{(k)} \right). \qquad (41)
\end{aligned}
$$

Thus, by Eqs. (40) and (41), we can that our proposed RARL-PDPFL satisfies $(\varepsilon_0^*, \delta_0^*)$-LDP guarantee with both Gaussian and Laplace mechanisms.

*Remark 2:* $\forall 1 \le k \le K$, if the sum of privacy budget $\varepsilon_i^{(k)}$ and that of probabilistic relaxation $\delta_i^{(k)}$ are smaller than $\varepsilon_0^*$ and $\delta_0^*$ respectively, each mobile device applies our proposed RARL-PDPFL method can satisfy LDP guarantee under both Gaussian and Laplace mechanisms.

## VI. Experiments Results

### A. Experiments Settings

Experiments were performed based on Pytorch 1.12.1 and GPU NVIDIA GeForce RTX 4060, including a central server and three mobile devices, and a membership inference attacker. According to [30], privacy protection can be divided into three types, (i.e., low, middle, and high levels). The corresponding requirements are given by: 1) Low level with privacy loss $\zeta_i^{(k)}$ ranging from 15% to 18%; 2) Middle level with $2\% < \zeta_i^{(k)} \le 15\%$; 3) High level with $\zeta_i^{(k)} \le 2\%$. In the experiments, the three mobile devices have to satisfy privacy protection levels from low to high, i.e., mobile device 1 has the lowest level, and mobile device 3 has the highest level requirements.

Standard MNIST dataset[1] is used in the experiments, which consists of 60000 training data and 10000 testing data for handwritten digit recognition. The local model uses a convolutional neural network with two convolutional layers and two fully connected layers and uses a rectified linear unit (ReLu) function to activate the model. In the experiments, mobile devices sample 469 training data and 500 testing data to train the local model parameters, with the Adam algorithm and 1 local epoch. The privacy budget is chosen from $0.1$ to $0.5$ and the probabilistic relaxation is set from $e^{-3}$ to $e^{-1}$.

All three mobile devices have the same parameter settings for their Q/R networks, with the learning parameters illustrated in Table I. Specifically, the input size of Q/R-network 1 equals the dimensions of the state. Both the second Q-network and R-network have an input size of 7. In the experiments, ReLU is used as the activation function in our proposed RARL-PDPFL, as it can overcome the vanishing gradient problem resulting from differential operations and increase the nonlinearity to accelerate the learning speed. Besides, Q-network 1 outputs two Q-values, and Q-network 2 outputs Q-values with 9 dimensions, which equals the number of available perturbation policies. The attacker uses a deep Q-network to determine its attack probability from a set of $[0.1, 0.9]$, which is quantified into three available levels. The random exploration rate $\epsilon$ linearly decreases from $0.1$ to $10^{-4}$ within 200 time slots, with each time slot decreasing by $4.9 \times 10^{-4}$.

To verify the effectiveness of our framework, SFAC in [13], Privatized FedPaq in [14] and SHRL in [17] are chosen as benchmarks, with details showing as follows:

- **SFAC** in [13] uses the Laplace mechanism to perturb local model parameters with a fixed privacy budget of $0.5$.
- **Privatized FedPaq** in [14] adds Gaussian noise to local models for each of the three mobile devices using a fixed privacy budget of $0.5$ and probabilistic relaxation of $e^{-1}$.
- **SHRL** in [17] can be applied to help FL optimize the perturbation mechanism and parameters with a typical Boltzmann policy distribution.

### B. Personalized and Dynamic Performance

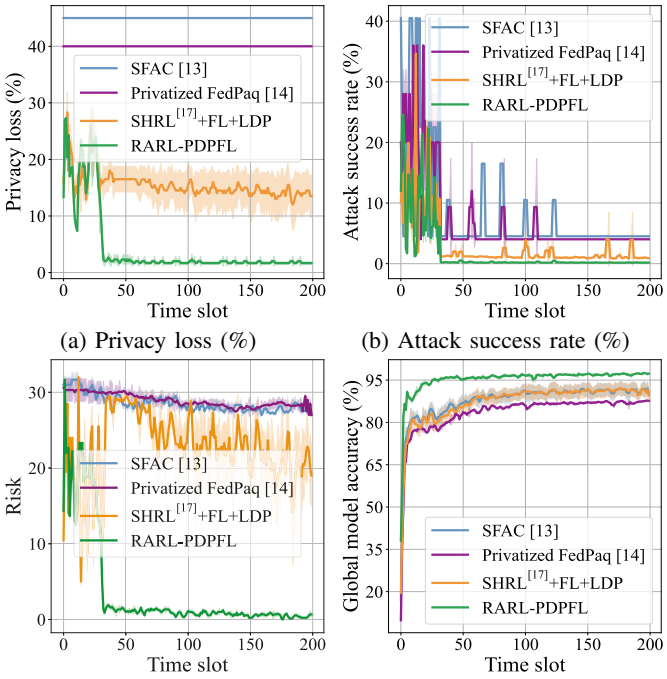
The privacy protection and training performance averaged over $50$ time slots after convergence is shown in Table II. The results show that our scheme enables each mobile device to satisfy its privacy requirements as well as achieves a $97.58\%$ global accuracy. That is because RARL-PDPFL uses privacy loss and local model accuracy as the basis to avoid risky policies that cannot satisfy user-specific privacy requirements.

As shown in Fig. 5, the privacy loss, attack success rate, and risk decrease with time while the global model accuracy increases with time. For example, our RARL-PDPFL decreases the privacy loss to $1.73\%$, attack success rate to $1.70‰$ and risk to $0.67$, and improves the global model accuracy to $97.30\%$ after 200 time slots. Besides, our scheme outperforms benchmarks SFAC in [13], Privatized FedPaq in [14] and SHRL in [17] with lower privacy loss, attack success rate, and risk, and higher global model accuracy. For instance, RARL-PDPFL reduces $95.82\%$ privacy loss and $96.77\%$ attack success rate, increases $10.80\%$ accuracy, and decreases $97.60\%$ risk compared with Privatized FedPaq [14]. The performance gain results from the joint optimization of perturbation mechanism and parameters as well as the consideration of risk-aware method in the learning process. Further, the proposed framework has $87.40\%$ lower privacy loss, $96.24\%$ lower attack success rate, $8.52\%$ higher global model accuracy, and $3.30\%$ less risk than SHRL [17]. The reason is that our scheme considers both short-term and long-term risks to reduce dangerous explorations and improves the impact of risks on the policy distribution formulation.

[1] http://yann.lecun.com/exdb/mnist/

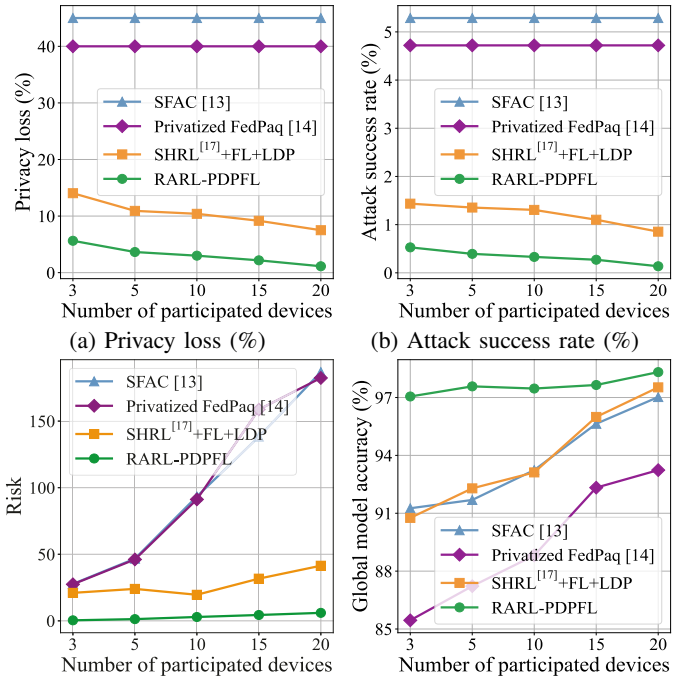
TABLE II: Performances of RARL-PDPFL based on user-personalized privacy requirements of three mobile devices

| Metrics / Device ID | Privacy loss (%) | Privacy requirements guarantee | Accuracy (%) | Convergence time slot |
|---|---|---|---|---|
| **1** | 12.00 | ✓ | 97.58 | 34 |
| **2** | 3.90 | ✓ | | |
| **3** | 1.00 | ✓ | | |




(a) Privacy loss (%) (b) Attack success rate (%) (c) Risk (d) Global model accuracy (%)

Fig. 5: Privacy and training performance averaged over 3 episodes with three mobile devices based on the MNIST dataset.




(a) Privacy loss (%) (b) Attack success rate (%) (c) Risk (d) Global model accuracy (%)

Fig. 6: Average performance under $3 \sim 20$ participated mobile devices on MNIST dataset.

### C. Impact of Participated Mobile Devices

The impact of participated device numbers on privacy and training performance is shown in Fig. 6, in which the privacy loss requirements of low, middle, and high levels are 10%, 20%, and 70% proportions among the total number of mobile devices. The privacy loss and attack success rate almost decrease with the number of participated mobile devices, while the risk and global model accuracy increase with it. Our scheme is more robust than the three baselines, as the number of participated mobile devices changes from 3 to 20. On the other hand, the risk of our scheme slightly increases the number of participated mobile devices. The corresponding performance gain of our scheme is larger than 96.78%, 96.71%, and 85.19% compared with SFAC, Privatized FedPaq, and SHRL respectively for $3 \sim 20$ mobile devices.

### D. Ablation Study

We also perform a series of ablation experiments to evaluate the performance of our designed punishment function, R-network, and hierarchical structure, as shown in Table III. The results show that our designed punishment function can guide mobile devices to avoid immediate risky policies, thus reducing privacy loss, attack success rate, and risk. Besides, the designed R-network estimates the long-term risk to formulate perturbation policy distribution, thus guiding each mobile device to explore the optimal policies quickly. Further, our designed two-level hierarchical structure takes the policy selection priority into account to accelerate convergence speed and thus further improve privacy protection performance.

## VII. RELATED WORK

Recently, blockchain, homomorphic encryption, secure aggregation, and trust formulation methods are used in FL to resist membership inference attacks [31]–[34]. For example, an elude secure aggregation method is proposed in [31], which proves that FL is vulnerable to attacks due to incorrect usage of secure aggregation. The blockchain-enabled FL privacy protection scheme presented in [33] designs a verification mechanism to help a central server select the honest participating clients. Further, a detection and aggregation algorithm is designed for FL [34], in which the penultimate layer representations are used to improve the defense performance, and the discrepancies are extracted to update the trust values.

Differential privacy has been applied to help enhance the privacy of FL [28], [35]–[39]. For instance, a Gaussian mechanism-based FL privacy protection framework proposed in [28] proposes a random participant scheduling method to improve privacy level. To solve the vulnerability of the

TABLE III: Performance of ablation study with three mobile devices

| Metrics / Algorithms | Accuracy (%) | Privacy loss (%) | Attack success rate (‰) | Risk | Convergence time slot |
|---|---|---|---|---|---|
| **RARL-PDPFL** | **97.36** | **1.76** | **1.78** | **0.47** | **34** |
| **RARL-PDPFL w/o punishment function** | 97.28 | 7.90 | 8.23 | 6.37 | 50 |
| **RARL-PDPFL w/o R-network** | 97.27 | 6.50 | 6.91 | 5.17 | 137 |
| **RARL-PDPFL w/o hierarchical structure** | 97.14 | 4.33 | 5.01 | 5.03 | 43 |

stochastic gradient descent algorithm in FL, an LDP-enabled FL framework is proposed in [38]. The framework perturbs local model parameters with Gaussian noise to make a trade-off among user privacy loss, global model accuracy, and transmission rate. A fine-grained differentially private FL scheme is then presented in [39], which uses the importance of fully connected layers to allocate Laplace noise for higher protection level and training accuracy.

The choice of perturbation policy including privacy budget and probabilistic relaxation is highly related to privacy and training performance of FL framework [15], [40]–[42]. For example, a Gaussian DP-based FL framework in [42] applies a gradient-boosting decision tree to update the local model and changes the privacy budget based on user contributions. To further improve the privacy level, the deep RL-based privacy-aware scheme in [15] applies deep Q-network to choose the privacy budget to reduce the privacy loss against inference attackers. The privacy-preserving FL framework designed in [40] proposes an adaptive privacy decomposition mechanism to dynamically decay the Gaussian noise to resist gradient leakage attacks.

## VIII. Summary and Future Work

In this paper, we have proposed a risk-aware RL-based personalized differentially private FL framework to satisfy user-specific requirements against membership inference attacks. This framework designs a two-level hierarchical structure to jointly optimize the perturbation mechanism, privacy budget, and probabilistic relaxation. A punishment function has been used to avoid immediate dangerous policies. Further, we have designed an R-network to estimate the long-term risk of each chosen perturbation policy, and a policy distribution to increase the impact of both short-term and long-term risks. We also have analyzed the convergence performance of our framework and proved that our framework satisfies the LDP guarantee. Experimental results based on the MNIST datasets show that our proposed framework outperforms three benchmarks. For example, our scheme has $7.96\%$ higher global model accuracy, $96.22\%$ lower privacy loss, and $97.52\%$ less risk than the benchmark SFAC in [13] at time slot 200.

In the future, we will further consider the mobility of mobile devices in the design of model parameters transmission. We plan to apply the proposed framework in unmanned aerial vehicles or the Internet of Vehicles that perform objective detection local tasks. Another efficient mechanism is to combine the age of information with LDP techniques, which can potentially save unnecessary interactions among mobile devices and thus further protect user privacy.

## References


[1] J. Mills, J. Hu, and G. Min, "Multi-task federated learning for personalised deep neural networks in edge computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 3, pp. 630–641, Jul. 2022.

[2] Q. Li, Z. Wen, Z. Wu *et al.*, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3347–3366, Nov. 2023.

[3] Y. Xu, L. Wang, H. Xu *et al.*, "Enhancing federated learning with server-side unlabeled data by adaptive client and data selection," *IEEE Trans. Mobile Comput.*, vol. 23, no. 4, pp. 2813–2831, Apr. 2024.

[4] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, pp. 3093–3106, Los Angeles, CA, USA, Nov. 2022.

[5] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy (S&P)*, pp.739–753, San Francisco, CA, USA, May 2019.

[6] S. Mahloujifar, E. Ghosh, and M. Chase, "Property inference from poisoning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, pp. 1120–1137, San Francisco, CA, USA, May 2022.

[7] J. Zhao, H. Zhu, F. Wang *et al.*, "PVD-FL: A privacy-preserving and verifiable decentralized federated learning framework," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2059–2073, May 2022.

[8] H. Chaudhari, R. Rachuri, and A. Suresh, "Trident: Efficient 4PC framework for privacy preserving machine learning," in *Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, Feb. 2020.

[9] S. Wang, L. Huang, Y. Nie *et al.*, "Local differential private data aggregation for discrete distribution estimation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 9, pp. 2046–2059, Feb. 2019.

[10] A. Kolluri, T. Baluta, and P. Saxena, "Private hierarchical clustering in federated networks," in *Proc. ACM. Conf. Computer Commun. Secur. (CCS)*, pp. 2342–2360, Virtual Event, Republic of Korea, Nov. 2021.

[11] H. Guo, H. Wang, T. Song *et al.*, "SIREN+: Robust federated learning with proactive alarming and differential privacy," *IEEE Trans. Dependable Secure Comput.*, pp. 1–18, Feb. 2024.

[12] R. Xue, K. Xue, B. Zhu *et al.*, "Differentially private federated learning with an adaptive noise mechanism," *IEEE Trans. Inf. Forensics Security*, pp. 74–87, Sept. 2023.

[13] Y. Wang, Z. Su, N. Zhang, and A. Benslimane, "Learning in the air: Secure federated learning for UAV-assisted crowdsensing," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1055–1069, Aug. 2021.

[14] N. Mohammadi, J. Bai, Q. Fan, Y. Song, Y. Yi *et al.*, "Differential privacy meets federated learning under communication constraints," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22 204–22 219, Aug. 2022.

[15] Y. Xiao, L. Xiao, X. Lu *et al.*, "Deep-reinforcement-learning-based user profile perturbation for privacy-aware recommendation," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4560–4568, Sept. 2020.

[16] M. Min, S. Yang, H. Zhang *et al.*, "Indoor semantic location privacy protection with safe reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 5, pp. 1385–1398, Jul. 2023.

[17] X. Lu, L. Xiao, G. Niu, X. Ji, and Q. Wang, "Safe exploration in wireless security: A safe reinforcement learning algorithm with hierarchical structure," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 732–743, Feb. 2022.

[18] C. Wei, S. Ji, C. Liu, W. Chen, and T. Wang, "AsgLDP: Collecting and generating decentralized attributed graphs with local differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3239–3254, Apr. 2020.

[19] Q. Chen, Z. Wang, H. Wang, and X. Lin, "FedDual: Pair-wise gossip helps federated learning in large decentralized networks," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 335–350, Nov. 2023.

[20] M. Lee, G. Yu, and H. Dai, "Privacy-preserving decentralized inference with graph neural networks in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 543–558, May 2023.

[21] L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, "Privacy-preserving collaborative deep learning with unreliable participants," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1486–1500, Sept. 2019.

[22] F. Boenisch, C. Mühl, A. Dziedzic, R. Rinberg, and N. Papernot, "Have it your way: Individualized privacy assignment for DP-SGD," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, pp. 19073–19103, New Orleans, LA, USA, Dec. 2023.

[23] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Art. Intell. Stat. (AISTATS)*, pp. 1273–1282, Fort Lauderdale, FL, USA, Feb. 2017.

[24] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, pp. 3-18, San Jose, CA, USA, May 2017.

[25] Y. Alufaisan, M. Kantarcioglu, and Y. Zhou, "Robust transparency against model inversion attacks," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 5, pp. 2061–2073, Aug. 2020.

[26] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Repr. (ICLR)*, pp. 1–15, San Diego, CA, USA, May 2015.

[28] K. Wei, J. Li, M. Ding *et al.*, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, Apr. 2020.

[29] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014.

[30] X. Pang, Z. Wang, D. Liu *et al.*, "Towards personalized privacy-preserving truth discovery over crowdsourced data streams," *IEEE/ACM Trans. Netw.*, vol. 30, no. 1, pp. 327–340, Sept. 2021.

[31] D. Pasquini, D. Francati, and G. Ateniese, "Eluding secure aggregation in federated learning via model inconsistency," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, pp. 2429–2443, Los Angeles, CA, USA, Nov. 2022.

[32] X. Liu, H. Li, G. Xu *et al.*, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4574–4588, Aug. 2021.

[33] S. Awan, F. Li, B. Luo, and M. Liu, "Poster: A reliable and accountable privacy-preserving federated learning framework using the blockchain," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, pp. 2561–2563, London, United Kingdom, Nov. 2019.

[34] N. Wang, Y. Xiao, Y. Chen *et al.*, "Flare: defending federated learning against model poisoning attacks via latent space representations," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, pp. 946–958, Nagasaki, Japan, May 2022.

[35] T. Nguyen and M. T. Thai, "Preserving privacy and security in federated learning," *IEEE/ACM Trans. Netw.*, vol. 32, no. 1, pp. 833–843, Aug. 2023.

[36] R. Hu, Y. Guo, and Y. Gong, "Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy," *IEEE Trans. Mobile Comput.*, pp. 1–14, Dec. 2023.

[37] K. Wei, J. Li, C. Ma *et al.*, "Personalized federated learning with differential privacy and convergence guarantee," *IEEE Trans. Inf. Forensics Security*, pp. 4488–4503, Jul. 2023.

[38] M. Kim, O. Günlü, and R. F. Schaefer, "Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, pp. 2650-2654, Toronto, ON, Canada, Jun. 2021.

[39] L. Zhu, X. Liu, Y. Li *et al.*, "A fine-grained differentially private federated learning against leakage from gradients," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11 500–11 512, Nov. 2022.

[40] J. Hu, Z. Wang, Y. Shen *et al.*, "Shield against gradient leakage attacks: Adaptive privacy-preserving federated learning," *IEEE/ACM Trans. Netw.*, vol. 32, no. 2, pp. 1407–1422, Sept. 2023.

[41] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, pp. 332-349, San Francisco, CA, USA, May 2019.

[42] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2134–2143, Sept. 2019.


**Xiaozhen Lu** (Member, IEEE) received the B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2017, and the Ph.D. degree in communication and information systems from Xiamen University, Xiamen, China, in 2021. She is currently an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. She was a recipient of the Best Student Paper Award for ML4CS 2019, and the Excellent Paper Award for CWSN 2020. Her research interests include reinforcement learning, federated learning, network security, and wireless communications.

**Zihan Liu** received the B.S. degree in software engineering from the Nantong University, Nantong, China, in 2023. She is currently working toward the M.S. degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her current research interests include reinforcement learning and federated learning.

**Liang Xiao** (Senior Member, IEEE) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, China, in 2000, the M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and the Ph.D. degree in electrical engineering from Rutgers University, NJ, USA, in 2009. She was a Visiting Professor with Princeton University, Virginia Tech, and the University of Maryland, College Park. She is currently a Professor with the Department of Informatics and Communication Engineering, Xiamen University, Xiamen, China. She was a recipient of the Best Paper Award for 2016 INFOCOM Big Security WS and 2017 ICC. She has served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.

**Huaiyu Dai** (Fellow, IEEE) received the B.E. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2002.

He was with Bell Labs, Lucent Technologies, Holmdel, NJ, in summer 2000, and with AT&T Labs-Research, Middletown, NJ, in summer 2001. He is currently a Professor of Electrical and Computer Engineering with NC State University, Raleigh, holding the title of University Faculty Scholar. His research interests are in the general areas of communications, signal processing, networking, and computing, with over 280 peer-reviewed journal/conference papers published. His current research focuses on machine learning and artificial intelligence for communications and networking, multilayer and interdependent networks, dynamic spectrum access and sharing, as well as security and privacy issues in the above systems.

He has served as an area editor for IEEE Transactions on Communications, a member of the Executive Editorial Committee for IEEE Transactions on Wireless Communications, and an editor for IEEE Transactions on Signal Processing. Currently he serves as an area editor for IEEE Transactions on Wireless Communications. He will serve as the Editor in Chief for IEEE Transactions of Signal and Information Processing over Networks in 2025. He was a co-recipient of best paper awards at 2010 IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS 2010), 2016 IEEE INFOCOM BIGSECURITY Workshop, and 2017 IEEE International Conference on Communications (ICC 2017). He received Qualcomm Faculty Award in 2019, and IEEE Communications Society William R. Bennett Prize in 2024. He is a Fellow of IEEE, and of Asia-Pacific Artificial Intelligence Association.