Defending Against Data Reconstruction Attacks in Federated Learning: An Information Theory Approach

Qi Tan[a], Qi Li[b], Yi Zhao[c, ✉], Zhuotao Liu[b], Xiaobing Guo[d], and Ke Xu[a, ✉]

[a]*Department of Computer Science and Technology, Tsinghua University*
[b]*Institute for Network Science and Cyberspace, Tsinghua University*
[c]*School of Cyberspace Science and Technology, Beijing Institute of Technology*
[d]*Lenovo Research*

Abstract

Federated Learning (FL) trains a black-box and high-dimensional model among different clients by exchanging parameters instead of direct data sharing, which mitigates the privacy leak incurred by machine learning. However, FL still suffers from membership inference attacks (MIA) or data reconstruction attacks (DRA). In particular, an attacker can extract the information from local datasets by constructing DRA, which cannot be effectively throttled by existing techniques, e.g., Differential Privacy (DP).

In this paper, we aim to ensure a strong privacy guarantee for FL under DRA. We prove that reconstruction errors under DRA are constrained by the information acquired by an attacker, which means that constraining the transmitted information can effectively throttle DRA. To quantify the information leakage incurred by FL, we establish a channel model, which depends on the upper bound of joint mutual information between the local dataset and multiple transmitted parameters. Moreover, the channel model indicates that the transmitted information can be constrained through data space operation, which can improve training efficiency and the model accuracy under constrained information. According to the channel model, we propose algorithms to constrain the information transmitted in a single round of local training. With a limited number of training rounds, the algorithms ensure that the total amount of transmitted information is limited. Furthermore, our channel model can be applied to various privacy-enhancing techniques (such as DP) to enhance privacy guarantees against DRA. Extensive experiments with real-world datasets validate the effectiveness of our methods.

1 Introduction

Federated learning (FL) [41, 58, 61, 62] is a new form of machine learning (ML), which protects privacy by transmitting gradients or parameters to avoid sharing raw data. Specifically, the parameters form a *communication channel* between the server and each client, so the server gets information from local datasets via such channels. Based on the Data Processing Inequality (DPI) [15], communication by parameters, which is a deterministic mapping of local data, instead of raw data, reduces the risk of data privacy. However, recent studies reveal that the parameter channel of FL still leaks privacy. For example, multiple literature indicates that adversaries can conduct membership inference attacks (MIA) with uploaded model parameters [10, 13, 40, 42, 44, 48], which breaks the anonymity of data privacy. Moreover, adversaries can completely steal training data by data reconstruction attacks (DRA) [11, 12, 24, 28, 63], resulting in serious privacy issues in FL.

In order to enhance privacy protection for FL, dimension reduction [36, 46, 52] or differential privacy (DP) [3, 34, 35] are widely adopted approaches. However, dimension reduction lacks theoretical guarantees of the defense ability against DRA, so it cannot flexibly configure defense capabilities according to different privacy requirements. DP's privacy protection can provide theoretical guarantees (e.g., the privacy budget ε) for MIA attacks [7, 22]. It aims to guarantee that changing any data point will not significantly affect the output distribution of a system. This goal is different from the one in defending against DRA, which focuses on preventing the attacker from reconstructing the whole distribution of the local dataset. Thus DP still cannot defend against DRA attacks [7, 14, 22, 26, 38]. For instance, in DP-SGD, algorithms with identical privacy budget but different training hyper-parameters (e.g., different batch size B) cannot guarantee the same success rate for DRA [26]. Moreover, quantifying information leakage is the basis for defending against DRA attacks. Previous technique like Quantitative Information Flow (QIF)[1] [5] quantifies information leak under a white-box and time-invariant setting[47]. It requires the knowledge of the correlation between the distributions of inputs and outputs, which cannot hold in FL systems.

[1]QIF focuses on a special security concern, namely the probability of guessing a secret in one try [5, 51]. This security concern is different from the one in DRA, where DRA focuses on reconstructing the whole distribution of local data.

It is difficult to defend against DRA attacks in FL due to the following challenges. (i) **the black-box model.** For DNN-based models, the correlation between the input distribution and output distribution is extremely complex, and we cannot obtain the exact mapping function between the two distributions, making theoretical analysis impossible. (ii) **the high-dimensional parameter space.** Traditional mathematical tools (e.g., eigen-decomposition) cannot process high-dimensional parameter spaces on the scale of thousands of millions due to the requirements for large-scale storage and high-performance computing. (iii) **the time-variant system.** During the training process, constant parameter updates change the model at each step, leading to a time-variant system in FL. The time-variant system, which changes the output distributions accordingly, leads to dynamic information leakage, requiring continuously changing quantifications.

To address the above challenges, we develop a theoretical framework based on mutual information (MI)[2] to evaluate privacy leakage caused by DRA in FL, and design methods to constrain information leakage to defend against DRA attacks. Specifically, we demonstrate that the lower bound of mean squared error (MSE), which serves as an indicator of DRA's precision in reconstruction (i.e., the smaller MSE means the higher precision for the attacker), is determined by the amount of acquired information, i.e., the MI between the local dataset and the shared parameters. Thus, MI can be utilized as the indicator for quantifying the information leakage in FL. Then we build a channel model to analyze information leaks under the black-box setting of FL. Through our proposed channel model, we find that the transmitted information (i.e., the information leakage) is decided by two factors: the channel capacity C, which represents the maximal ability to transmit information in a single training round; and the optimization rounds n, which is correlated to the information accumulation. For example, if the channel capacity is bounded by a threshold κ, and the number of optimization rounds is less than n, then the total amount of information leakage is less than $n \cdot \kappa$. Furthermore, our channel model can analyze various privacy-enhancing methods in defending against DRA, including DP, gradient compression, and utilizing large batch size.

Based on the channel model, we utilize DPI to transform the operations (e.g., eigen-decomposition and adding noise) of constraining channel capacity from the parameter space to the data space. This transformation significantly improves the training efficiency and the model accuracy of the high dimensional and time-variant model under constrained information leakage. Specifically, our *protecting goal* is to decide the covariance matrix for the added noise according to a given data distribution $\boldsymbol{D}$, which ensures privacy protection by constraining the reconstruction error above a certain threshold. Compared to conventional protection techniques, which directly deal with parameters after gradient mapping, constraining in the data space has two distinct advantages: firstly, it makes the computational complexity independent of the optimization rounds n and the model's dimensionality d_m, reducing it from $O(n \cdot d_m)$ to $O(d_D)$, where d_D denotes the dimensionality of the data and $d_D \ll d_m$. Secondly, data space preserves the correlations between data attributes, hence we can leverage the prior knowledge of relative importance to implement stronger safeguards for the critical attributes, which enhances the capability to balance the utility and the privacy.

Finally, according to the theoretical results, we propose three implementations for constraining the channel capacity. These implementations incorporate different prior knowledge in defending against DRA, which can be employed to flexibly balance the utility and the privacy.

In summary, the contributions of our paper are as follows:

- We demonstrate that the amount of transmitted information decides the lower bound of the reconstruction error for DRA attacks.
- We establish a channel model to quantify the information leakage of the black-box model in FL, which can be applied to analyze various privacy-enhancing methods for defending against DRA.
- We theoretically constrain the transmitted information through the operation in data space instead of parameter space for the first time and demonstrate that it significantly improves the training efficiency and the model accuracy under constrained information leakage.
- By incorporating different prior knowledge, we propose three implementations to constrain channel capacity, which can be utilized to flexibly balance the utility and the privacy.
- Extensive experiments demonstrate that the newly proposed methods effectively enhance the safety, efficiency, and flexibility of FL.

2 Background and Preliminary

This paper studies the FL problem based on information theory. Specifically, in the FL scenario, the server and clients communicate by sending model parameters. Even without direct data sharing, the MI between shared parameters and the local dataset grows accordingly, which enhances the ability for attackers to conduct DRA attacks. For clarity purposes, Tab. 1 lists major notations used in the paper, and we will describe the remaining variables when they are utilized.

2.1 Federated Learning

Regarding FL, a specific client, namely the victim, receives the initial parameter $\boldsymbol{W}_1$ from the server in a specific communication round and conducts the optimization process with

[2] Mutual information $I(X;Y)$ [16, 29, 49] represents the uncertainty decrement of X when we observe Y.

Table 1: Major Notation Explanation

Notations	Explanation
$F(\cdot)$, η	Loss function and learning rate for local optimization
$\boldsymbol{W}_i^{(t)}$, $\boldsymbol{W}_o^{(t)}$	The received (input) and shared (output) parameters of the victim at time t
$\boldsymbol{D}$	Random variable follows data distribution of the victim
B	Batch size for optimization
E, n	Local steps for one communication round and the total local steps for all communication rounds
$\mathcal{A}^{(t)}(\cdot)$	The aggregation method at time t
$\boldsymbol{V}^{(t)}$	Variables (e.g., gradients, parameters) collected by the server from clients other than the victim
$C^{(t)}$	The channel capacity (maximum transmitted information) of the victim at time t
κ	The threshold of $C^{(t)}$, i.e., the setted channel capacity
$\boldsymbol{\Sigma}_*$	The covariance matrix of random variable
λ and σ	Eigenvalues for the covariance matrix and noise variables
$I(\boldsymbol{X};\boldsymbol{Y})$	The mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\boldsymbol{\xi}$	The noise variable that is subject to Gaussian distribution

the victim's dataset as

$$\boldsymbol{W}_{t+1} \leftarrow \boldsymbol{W}_t - \eta \cdot \nabla_{\boldsymbol{W}} F(\boldsymbol{W}_t; \boldsymbol{D}),\ t = 1, \cdots, E, \tag{1}$$

where E is the number of local steps. Then the victim sends $\boldsymbol{W}_{E+1}$ back to the server. If we rewrite $\boldsymbol{W}_1$ and $\boldsymbol{W}_{E+1}$ as $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$ respectively, the two parameters form a communication channel for information transmission (as illustrated in Fig. 2), and the local optimization process defined in Eq. (1) loads information from the dataset to the communication channel. Finally, the server collects the parameters from different clients (including the victim) for aggregation as

$$\boldsymbol{W}_i = \mathcal{A}(\boldsymbol{W}_o; \boldsymbol{V}). \tag{2}$$

2.2 Information Theory

Differential Entropy. The differential entropy of a random variable $\boldsymbol{X}$ is defined as follows

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{X}} f(\boldsymbol{x}) \log f(\boldsymbol{x})\, d\boldsymbol{x}, \tag{3}$$

which is utilized to describe the degree of random uncertainty. For a dataset with high information entropy, i.e., a dataset with plentiful information, it is difficult for an attacker to conduct DRA attacks. While for the dataset with low information entropy, the opposite is true.

However, calculating the differential entropy by Eq. (3) is infeasible in practice since we cannot obtain the distribution function of the target variable. Therefore, we use the maximum entropy distribution to analyze the worst-case scenario. Specifically, with $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$, the maximum entropy distribution is the Gaussian distribution, i.e., $h(\boldsymbol{X}) \leq h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$.

Mutual Information. For two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, the mutual information between them is

$$I(\boldsymbol{X};\boldsymbol{Y}) = h(\boldsymbol{X}) - h(\boldsymbol{X}|\boldsymbol{Y}) = h(\boldsymbol{Y}) - h(\boldsymbol{Y}|\boldsymbol{X}). \tag{4}$$

Specifically, the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ is a symmetric function, which describes the random uncertainty decrement of $\boldsymbol{X}$ when we observe $\boldsymbol{Y}$, and vice versa.

Specifically, in this paper, we utilize $I(\boldsymbol{D};\boldsymbol{W}_i,\boldsymbol{W}_o)$ to quantify the information leakage in FL. The intuition is that when an attacker observes the communication parameters of a victim, the random uncertainty of the victim's local dataset will decrease, which implies the attacker can extract information from the victim's local dataset to achieve more precise DRA.

Channel Capacity. The key factor in describing a communication channel is channel capacity. In information theory, traditional channel capacity is defined by the maximum MI between the sending variable $\boldsymbol{X}$ and the receiving variable $\boldsymbol{Y}$, i.e., $C = \max_{p(\boldsymbol{x})} I(\boldsymbol{X};\boldsymbol{Y})$, which is the maximum information that we can send by information coding.

In this work, the sending variable is $\boldsymbol{D}$, while the receiving variable is $\boldsymbol{W}_o$, which is decided by the variables $\boldsymbol{W}_i$ and $\boldsymbol{D}$ according to Eq. (1), hence the channel capacity can be formalized as $C = \max_{p(\boldsymbol{w}_o)} I(\boldsymbol{D};\boldsymbol{W}_o|\boldsymbol{W}_i)$, which is the upper bound of the transmitted information.

The reasons for choosing MI to measure the information leakage. Prior research like QIF and g-leakage utilizes min-entropy [5, 51] to measure the information leakage, which can only work in a white-box and time-invariant system [47]. However, FL is a black-box (e.g., deep neural networks) and time-variant (e.g., parameter updating in each round) system. Hence, these techniques are not applicable. Moreover, min-entropy is unsuitable for modeling the attack against FL, e.g., DRA. The min-entropy focuses on measuring the uncertainty of guessing the most likely output of random sources [23, 30], which is more related to cryptographic systems. However, in DRA, the attacker's target is to reconstruct the whole data distribution based on the victim's sharing parameters, instead of guessing a most likely data point. Therefore, the precision of DRA attacks depends on the difference between two distributions, which must take all data of the distributions into consideration. In this scenario, the Shannon entropy, i.e., MI, which is based on the expectation metric, can accurately measure the correlation between the whole distributions, thus is more suitable for analyzing the information leakage issue under DRA. Therefore, we choose MI to measure the information leakage in FL.

3 Key Observation and Method Overview

3.1 Key Observation

Regarding FL, different clients jointly optimize the model by passing parameters to the server instead of raw data. Since the parameters are the mapping of the original data, the attacker can still reconstruct the private data from the parameters, thereby stealing privacy. Correspondingly, the client can preserve privacy by perturbing the transmitted parameters. Therefore, we build a channel model to calculate the privacy

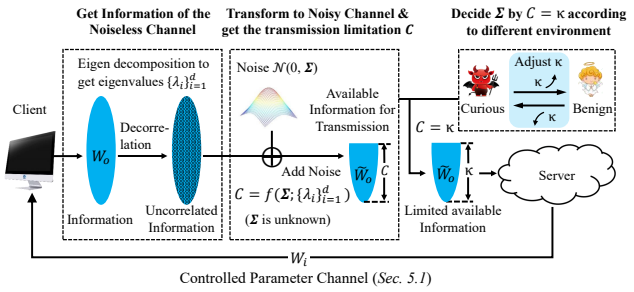

Figure 1: To enhance the capability for defending against DRA in FL, we develop techniques to constrain the amount of transmitted information below a certain threshold κ.

data contained in parameters. Moreover, based on the quantitative results, the privacy data in parameters can be flexibly adjusted to meet various privacy requirements.

In this section, we establish a formal correlation between the transmitted information, i.e., the MI, and the reconstruction error of DRA by the following theorem.

Theorem 1 (Lower bound for reconstruction error)**.** *For any random variable* $\boldsymbol{D}$, $\boldsymbol{D} \in \mathbb{R}^d$ *and* $\boldsymbol{W}$, $\boldsymbol{W} \in \mathbb{R}^m$, *we have*

$$\mathbb{E}[\|\boldsymbol{D} - \hat{\boldsymbol{D}}(\boldsymbol{W})\|^2/d] \geq \frac{e^{2h(\boldsymbol{D})/d}}{2\pi e} e^{-2I(\boldsymbol{D};\boldsymbol{W})/d}, \tag{5}$$

where $\hat{\boldsymbol{D}}(\boldsymbol{W})$ *is an estimator of* $\boldsymbol{D}$ *constructed by* $\boldsymbol{W}$.

Specifically, Eq. (5) denotes the lower bound of MSE in DRA, which represents the optimal error for data reconstruction. In the FL scenario, we denote $\boldsymbol{D}$ as the target data distribution and $\boldsymbol{W}$ as the shared parameter. Therefore $h(\boldsymbol{D})$ is the entropy of the target data distribution, which is a constant during machine learning. Moreover, the lower bound is negatively correlated to MI, i.e., $I(\boldsymbol{D};\boldsymbol{W})$, which indicates that if $\boldsymbol{W}$ contains more information of target data $\boldsymbol{D}$, i.e., a larger $I(\boldsymbol{D};\boldsymbol{W})$, the attacker can achieve a more precise reconstruction of $\boldsymbol{D}$. Consequently, large $I(\boldsymbol{D};\boldsymbol{W})$ exacerbates the privacy issue.

In FL, MI increases as the number of optimization rounds increases, thereby enhancing the precision of DRA. Therefore, constraining the overall MI in FL is the way to restrict the precision of DRA. To this end, we build a channel model to measure the increase of MI in Sec. 4, and propose three implementation methods to limit MI within a certain threshold.

3.2 Method Overview

In this study, we develop techniques to enhance FL's capability of privacy protection, in particular, defending against DRA. Normally, FL transmits information from the local dataset to the server by model parameters, i.e., the parameter channel. The transmitted information can be employed by an attacker to conduct various attacks. Hence, the objective of privacy protection is to constrain the amount of transmitted information. Specifically, for DRA, the reconstruction error is lower bounded by a function of MI (Thm. 1), which indicates that a smaller MI leads to a larger reconstruction error. Therefore, our technique is devoted to limiting the reconstruction ability by constraining the total MI in FL.

Threat Model. In this work, we focus on privacy leaks incurred by DRA in FL. Specifically, the attacker aims to reconstruct the data distribution $\boldsymbol{D}$ of a specific client (i.e., the victim) via parameters shared by the victim. We assume that the attacker can get the transmitted parameters by the victim (i.e., $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$). Then the attacker can reconstruct a data distribution $\hat{\boldsymbol{D}}(\boldsymbol{W}_i, \boldsymbol{W}_o)$ to approximate the target data distribution $\boldsymbol{D}$ (Appendix A explains the details of DRA).

Our goal is to constrain the private data that the attacker can obtain according to the *transmitted parameters* in FL. Therefore, the attacker's ability to construct the DRA attack is limited.

Controlled parameter channel. As illustrated in Fig. 1, to restrict the transmitted information, we transform the noiseless parameter $\boldsymbol{W}_o$ to a noisy Gaussian channel by adding Gaussian noise $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ to it. Based on the theorem of Gaussian channel in information theory [15], the noisy Gaussian channel has limited capability for information transmission, which is the channel capacity. Thus, with the eigenvalues of $\boldsymbol{W}_o$, we derive a formula $f(\boldsymbol{\Sigma})$ to characterize the channel capacity of the Gaussian channel by the maximum entropy distribution. Finally, we solve the equation $f(\boldsymbol{\Sigma}) = \kappa$ to decide $\boldsymbol{\Sigma}$ for constraining the transmitted information within a threshold κ. We will explain how the controlled parameter channel constrains transmitted information in Sec. 4 and Sec. 5.1.

Constraining channel capacity in the data space. To overcome the efficiency issue caused by the high dimensional and time-variant model in FL, we theoretically transform operations of constraining channel capacity from the parameter space to the data space. Specifically, the information contained in the resulting parameter is decided by the input data, thus constraining channel capacity can be achieved by restricting the information contained in the input data. Moreover, this transformation significantly improves the training efficiency and the model accuracy under constrained information leakage. We will explain the transformation in Sec. 5.2.

Theoretically, when an attacker observes $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$, the random uncertainty of the local data $\boldsymbol{D}$ decreases, which means the attacker gets more information to conduct DRA attacks, leading to more precise reconstruction. The amount of random uncertainty reduction ΔI (i.e., information leakage) in a round can be formalized as

$$\Delta I = I(\boldsymbol{D};\boldsymbol{W}_i, \boldsymbol{W}_o) - I(\boldsymbol{D};\boldsymbol{W}_i) = I(\boldsymbol{D};\boldsymbol{W}_o|\boldsymbol{W}_i),$$

which means the attacker gets ΔI information from $\boldsymbol{D}$ (as illustrated in Fig. 2). Moreover, this privacy leak occurs in each optimization round, which results in an increase in MI and consistently increases the risk of privacy.

As aforementioned, for defending against DRA, our technique constructs a controlled parameter channel by limiting

ΔI to less than a threshold κ for all optimization rounds. Then together with the bounded optimization rounds n, we provide $n \cdot \kappa$ guarantee for the total information leakage, thereby constraining the attack precision of DRA.

4 Channel Model of the Information Leakage

In this section, we formalize the problem of FL into a communication process based on information theory and then unfold the recurrent communication process into a time-dependent Markov Chain. Finally, we build a channel model to calculate the MI according to the unfolded communication process.

4.1 Accumulation of Mutual Information

As illustrated in Fig. 2, in the FL scenario, a specific client, i.e., the victim, communicates with the server through a logic channel: the parameters $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$. Specifically, There are three different information flows: *the ingress flow*, i.e., the received parameter $\boldsymbol{W}_i$, which determines the background knowledge possessed by the server (i.e., the attacker); *the egress flow*, which is the information contained in the sharing parameter $\boldsymbol{W}_o$; and *the internal flow*, i.e., local optimization process, which loads the information contained in the local dataset to the egress flow. Particularly, due to privacy requests, the victim only communicates with the server by $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$.

The information leakage of the communication channel depends on the MI increment when the server observes $\boldsymbol{W}_o$. Thus, it can be formalized as

$$I(\boldsymbol{D};\boldsymbol{W}_i,\boldsymbol{W}_o) = \underbrace{I(\boldsymbol{D};\boldsymbol{W}_i)}_{\text{Prior}} + \underbrace{I(\boldsymbol{D};\boldsymbol{W}_o|\boldsymbol{W}_i)}_{\text{Information Leakage } (\Delta I)}. \quad (6)$$

Eq. (6) is an immediate result according to the chain rule of MI. It indicates that the MI between $\boldsymbol{D}$ and the joint distribution $(\boldsymbol{W}_i,\boldsymbol{W}_o)$ can be divided into two parts: the prior knowledge and the information leakage ΔI. In the rest of this paper, we will utilize $I(\boldsymbol{D};\boldsymbol{W}_o|\boldsymbol{W}_i)$ instead of ΔI for more comprehensible analysis.

However, $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$ are joint distributions of different rounds, which contain multiple local learning processes. Hence, connecting Eq. (6) to the internal flow is difficult. To resolve this issue, we unfold the recurrent process to a time-dependent Markov chain. As illustrated in Fig. 2, there is only one local optimization process within a round from $\boldsymbol{W}_i^{(t)}$ to $\boldsymbol{W}_i^{(t+1)}$, hence we can analyze MI increment at round t.

Specifically, according to Eq. (1) and Eq. (2), the relationship between $\boldsymbol{W}_i^{(t)}$ and $\boldsymbol{W}_o^{(t)}$ in Fig. 2 is

$$\boldsymbol{W}_o^{(t)} = \boldsymbol{W}_i^{(t)} - \eta \cdot \nabla_{\boldsymbol{W}} F(\boldsymbol{W}_i^{(t)};\boldsymbol{D}), \quad (7)$$

which represents the local learning process with SGD. While for $\boldsymbol{W}_o^{(t)}$ and $\boldsymbol{W}_i^{(t+1)}$, the relationship is

$$\boldsymbol{W}_i^{(t+1)} = \mathcal{A}^{(t)}(\boldsymbol{W}_o^{(t)};\boldsymbol{V}^{(t)}), \quad (8)$$

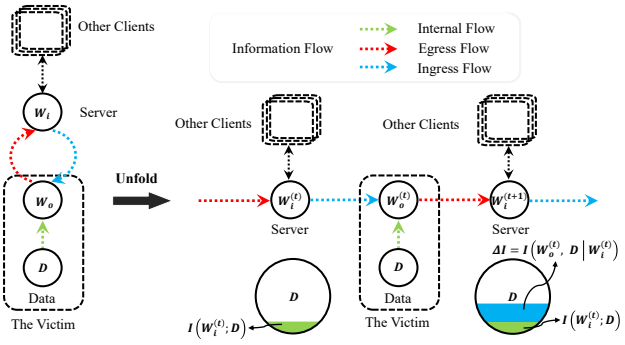

Figure 2: The process of FL can be unfolded to a time-dependent Markov chain. Hence we can analyze the mutual information in a round from $\boldsymbol{W}_i^{(t)}$ to $\boldsymbol{W}_i^{(t+1)}$.

where $\boldsymbol{V}^{(t)}$ are the variables uploaded by clients other than the victim.

Then based on the unfolded process, we transform the MI in Eq. (6) to the joint mutual information as

$$\begin{aligned} & I(\boldsymbol{D};\boldsymbol{W}_i^{(0)},\boldsymbol{W}_o^{(0)},\cdots,\boldsymbol{W}_i^{(n)}) \\ = \;& \textstyle\sum_{t=0}^{n} I(\boldsymbol{D};\boldsymbol{W}_i^{(t)}|\boldsymbol{W}_o^{(t-1)},\boldsymbol{W}_i^{(t-1)},\cdots,\boldsymbol{W}_o^{(0)},\boldsymbol{W}_i^{(0)}) \\ & + \textstyle\sum_{t=0}^{n-1} I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)},\boldsymbol{W}_o^{(t-1)},\cdots,\boldsymbol{W}_o^{(0)},\boldsymbol{W}_i^{(0)}) \\ = \;& \underbrace{\textstyle\sum_{t=0}^{n-1} I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)})}_{\Gamma_{client}} + \underbrace{\textstyle\sum_{t=0}^{n-1} I(\boldsymbol{D};\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)})}_{\Gamma_{server}} \\ & + \underbrace{I(\boldsymbol{D};\boldsymbol{W}_i^{(0)})}_{\text{Prior}}, \end{aligned} \quad (9)$$

where the first equality depends on the chain rule of MI and the last equality is an immediate consequence of the Markov property of the learning process.

Specifically, Eq. (9) indicates that the overall MI is comprised of three parts: the prior, Γ_{server}, and Γ_{client}.

4.2 Analysis of Mutual Information

Regarding Eq. (9), the information can be divided into three parts: the prior, Γ_{server}, and Γ_{client}. The prior knowledge $I(\boldsymbol{D};\boldsymbol{W}_i^{(0)})$ is decided by the background knowledge of the attacker before the learning process.

For Γ_{server}, it represents the aggregation process on the server, hence it cannot be controlled by the local learning process. On the contrary, it can be exploited by an attacker. Without loss of generality, we focus on the general term $I(\boldsymbol{D};\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)})$, $t \in \{0,\cdots,n-1\}$, which is the information increment on the server. Based on the relationship between the MI and the entropy, we rewrite it as

$$\begin{aligned} & I(\boldsymbol{D};\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)}) \\ = \;& h(\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)}) - h(\boldsymbol{W}_i^{(t+1)}|\boldsymbol{D},\boldsymbol{W}_o^{(t)}) \\ = \;& h(\mathcal{A}^{(t)}(\boldsymbol{W}_o^{(t)};\boldsymbol{V}^{(t)})|\boldsymbol{W}_o^{(t)}) - h(\mathcal{A}^{(t)}(\boldsymbol{W}_o^{(t)};\boldsymbol{V}^{(t)})|\boldsymbol{D},\boldsymbol{W}_o^{(t)}), \end{aligned}$$

the last equality is a substitution according to Eq. (8). The formula indicates that $I(\boldsymbol{D};\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)})$ is decided by $\boldsymbol{V}^{(t)}$, which are variables independent of local learning process.

Specifically, if $\boldsymbol{V}^{(t)}$ is independent of $\boldsymbol{D}$, then $\boldsymbol{D} \to \boldsymbol{W}_o^{(t)} \to \boldsymbol{W}_i^{(t+1)}$ forms a Markov chain. Therefore, given the observation of $\boldsymbol{W}_o^{(t)}$, $\boldsymbol{D}$ is conditional independent of $\boldsymbol{W}_i^{(t+1)}$, i.e., $I(\boldsymbol{D};\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)}) = 0$. In this situation, $\Gamma_{server} = 0$, which means the server (i.e., the attacker) cannot affect DRA attacks. On the contrary, if $\boldsymbol{V}^{(t)}$ has the information of $\boldsymbol{D}$ that is independent of $\boldsymbol{W}_o^{(t)}$, i.e., the attacker can get auxiliary information other than the local learning process, the mutual information $I(\boldsymbol{D};\boldsymbol{W}_i^{(t+1)}|\boldsymbol{W}_o^{(t)}) > 0$, which increases the risk of information leakage for the rest communication rounds when $T \geq t+1$. In this situation, the server (i.e., the attacker) can increase the risk of privacy by utilizing the information collected by means other than FL.

However, the analysis of Γ_{server} is beyond the protection of the local learning process in FL, since Γ_{server} is related to the auxiliary information for the attacker to conduct the DRA attack. Particularly, our target is to bound Γ_{client}, which is the information leakage of the local learning process in FL. Sec. 5 indicates that regardless of $\Gamma_{server} > 0$ or not, Γ_{client} is constrained by our methods.

Finally, we put emphasis on Γ_{client}, the most important part that is correlated to information leakage of the local learning process in FL. Similarly, we focus on the general term of Γ_{client}, i.e., $I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)})$, which is the MI increment at round t, then we have

$$I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)}) = h(\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)}) - h(\boldsymbol{W}_o^{(t)}|\boldsymbol{D},\boldsymbol{W}_i^{(t)}). \quad (10)$$

Based on Eq (7), if $\boldsymbol{W}_i^{(t)}$ is observed, $\boldsymbol{W}_o^{(t)}$ is decided by a deterministic function of $\boldsymbol{D}$, which has a finite entropy. Moreover, if $\boldsymbol{W}_i^{(t)}$ and $\boldsymbol{D}$ are both observed, $\boldsymbol{W}_o^{(t)}$ is deterministic, thereby $h(\boldsymbol{W}_o^{(t)}|\boldsymbol{D},\boldsymbol{W}_i^{(t)}) \to -\infty$. The result is reasonable since the volume of a constant's support set[3] goes to 0, i.e., $2^{-\infty} = 0$. In this case, $I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \to +\infty$, which means a noiseless channel results in unlimited risk of privacy leaks.

To limit the information leakage, we add a Gaussian noise to $\boldsymbol{W}_o^{(t)}$, which transforms Eq. (10) to a noisy egress flow. Specifically, we turn to analyze

$$\widetilde{\boldsymbol{W}}_o^{(t)} = \boldsymbol{W}_o^{(t)} + \boldsymbol{\xi}, \quad (11)$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$, and $\boldsymbol{W}_o^{(t)} = \lim_{\boldsymbol{\Sigma}\to\boldsymbol{0}} \widetilde{\boldsymbol{W}}_o^{(t)}$. Then the property of $\widetilde{\boldsymbol{W}}_o^{(t)}$ is implied by following lemma.

Lemma 1 (Maximum entropy distribution)**.** *Let $\boldsymbol{X}$ be a continuous random vector with $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}_{\boldsymbol{X}}$, $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. Let $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Y}},\boldsymbol{\Sigma}_{\boldsymbol{Y}})$ be a Gaussian random variable that is independent with $\boldsymbol{X}$, then $h(\boldsymbol{X}+\boldsymbol{Y})$ achieves its maximum when $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}})$.*

[3]The volume of support set for a random variable $\boldsymbol{X}$ is $2^{h(\boldsymbol{X})}$ [15].

To analyze the noisy egress flow after transformation, we rewrite Eq. (10) as

$$I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)}) = h(\boldsymbol{W}_o^{(t)}+\boldsymbol{\xi}|\boldsymbol{W}_i^{(t)}) - h(\boldsymbol{W}_o^{(t)}+\boldsymbol{\xi}|\boldsymbol{D},\boldsymbol{W}_i^{(t)}). \quad (12)$$

As $\boldsymbol{\xi}$ is a Gaussian variable, Lemma 1 implies that the first term on the right-hand side of Eq. (12), i.e., $h(\boldsymbol{W}_o^{(t)}+\boldsymbol{\xi}|\boldsymbol{W}_i^{(t)})$, is upper bounded by the entropy of the Gaussian distribution. Moreover, when $\boldsymbol{D}$ and $\boldsymbol{W}_i^{(t)}$ are both observed, $\boldsymbol{W}_o^{(t)}$ is a constant, which means the only randomness of $\widetilde{\boldsymbol{W}}_o^{(t)} = \boldsymbol{W}_o^{(t)} + \boldsymbol{\xi}$ comes from $\boldsymbol{\xi}$, thus the second term on the right-hand side of Eq. (12) is $h(\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}))$. Let $\mathbb{E}[\boldsymbol{W}_o^{(t)}] = \boldsymbol{\mu}_{\boldsymbol{W}}$ and $Cov(\boldsymbol{W}_o^{(t)}) = \boldsymbol{\Sigma}_{\boldsymbol{W}}$, we have the upper bound of Eq. (12) as

$$I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \leq h(\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{W}},\boldsymbol{\Sigma}_{\boldsymbol{W}}+\boldsymbol{\Sigma})) - h(\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})). \quad (13)$$

In practice, the distribution of $\boldsymbol{W}_o^{(t)}$ is extremely complex and time-variant, so we utilize upper bound (13) to limit the information leakage. The important parts of Eq. (13) are the covariance matrixes of $\boldsymbol{W}_o^{(t)}$ and $\boldsymbol{\xi}$. Moreover, Sec. 5 indicates that regardless of $\boldsymbol{\Sigma}_{\boldsymbol{W}}$, we can restrict $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \leq \kappa$, $\forall \kappa > 0$, by deciding $\boldsymbol{\Sigma}$.

Finally, for multiple local updates, the scenario is slightly different. We denote the number of local steps in one communication round as E, and the total number of local steps for all communication rounds as n, where n is divisible into E. These notations imply that the number of communications is $T = \frac{n}{E}$. Then the issue can be transformed to the one-step case by a time-dependent aggregation method as

$$\boldsymbol{W}_i^{(t)} = \begin{cases} \mathcal{A}^{(t)}(\boldsymbol{W}_o^{(t-1)};\boldsymbol{V}^{(t-1)}), & \text{if } E \,|\, t \\ \boldsymbol{W}_o^{(t-1)}, & \text{otherwise.} \end{cases} \quad (14)$$

Where $E \,|\, t$ represents we make aggregation on the server every E steps. If we set $E = n$, the FL problem reduces to classical ML without collaboration.

Based on the former analysis, the new aggregation rule only changes Γ_{server} in Eq. (9), while Γ_{client} remains the same. Hence, the privacy analysis for the local optimization process is identical to classical ML. Therefore, our analysis of Γ_{client} will focus on classical ML in the remainder of this paper.

5 Controlled Parameter Channel

Based on the former analysis, the important part of defending against DRA is $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)})$, which represents the MI increment at round t. The key parameters to restrict the MI increment are $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ and $\boldsymbol{\Sigma}$. In this section, we first propose a method for deciding $\boldsymbol{\Sigma}$ that ensures $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \leq \kappa$. Then, we transform the operations for constraining MI from the parameter space to the data space and propose three implementation methods for constraining the channel capacity.

Finally, we analyze existing techniques with our theoretical results in defending against DRA.

5.1 Controlled Channel Capacity

Channel capacity is the maximum ability to transmit information within a single round, i.e., max $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)})$, which is the key parameter for constraining the information leakage.

To constrain the channel capacity, we have Thm 2.

Theorem 2 (Channel capacity)**.** *Let $\widetilde{\boldsymbol{W}}_o^{(t)} = \boldsymbol{W}_o^{(t)} + \sqrt{\sigma}\cdot\boldsymbol{\xi}$, where $\sigma \geq 0$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, if $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ are the mean vector and covariance matrix of $\boldsymbol{W}_o^{(t)}$ when $\boldsymbol{W}_i^{(t)}$ is observed, we have $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \leq f^{(t)}(\sigma)$, where*

$$f^{(t)}(\sigma) := \frac{1}{2}\sum_{i=1}^{d} \ln\frac{\lambda_i^{(t)}+\sigma}{\sigma},\ \sigma\in(0,\ +\infty). \tag{15}$$

where $\lambda_i^{(t)}$ is the i-th eigenvalue of the covariance matrix $\boldsymbol{\Sigma}^{(t)}$ and d represents the dimension of $\boldsymbol{W}_o^{(t)}$.

Based on Lemma 1, $I(\boldsymbol{D},\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)})$ achieves its maximum $f^{(t)}(\sigma)$ when $\boldsymbol{W}_o^{(t)}$ conforms to the Gaussian distribution. According to the Central Limit Theorem, if we use mini-batch SGD, the distribution of $\boldsymbol{W}_o^{(t)}$ converges to the Gaussian distribution, which means upper bound (15) becomes tighter when we utilize larger batch size for local training.

Regarding Γ_{client}, if we denote $C^{(t)} = f^{(t)}(\sigma^{(t)})$, where $\sigma^{(t)}$ represents σ at time t, we have $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \leq C^{(t)}$, which indicates $\boldsymbol{W}_i^{(t)}$ and $\widetilde{\boldsymbol{W}}_o^{(t)}$ form a communication channel with channel capacity $C^{(t)}$.

Specifically, according to Eq. (15), $C^{(t)}$ is decided by two components: $\{\lambda_i^{(t)}\}_{i=1}^d$ and $\sigma^{(t)}$. First, $\{\lambda_i^{(t)}\}_{i=1}^d$ represent the eigenvalues of $\boldsymbol{\Sigma}^{(t)}$. Based on Eq. (7), $\boldsymbol{W}_o^{(t)}$ is related to $\boldsymbol{W}_i^{(t)}$, which is the parameter received from the server. Therefore, the server (i.e., the attacker) can craft $\boldsymbol{W}_i^{(t)}$ to get more information from the local dataset (as displayed in Fig. 3).

Second, $\sigma^{(t)}$ is related to the added noise. Based on the monotonicity of $C^{(t)} = f^{(t)}(\sigma^{(t)})$, we conclude that for any $\boldsymbol{W}_i^{(t)}$, there exists an unique $\sigma^{(t)}$ that satisfies $f^{(t)}(\sigma^{(t)}) = \kappa$, where $\kappa \geq 0$ is a certain threshold.

If we denote $\lambda_i^{(t)} = \lambda^{(t)}$, $i \in \{1,\cdots,d\}$, we have $C^{(t)} = \frac{d}{2}\ln(\frac{\lambda^{(t)}+\sigma^{(t)}}{\sigma^{(t)}})$. In this scenario, the channel capacity can be displayed as Fig. 3. Specifically, the channel capacity is an increasing function of $\lambda^{(t)}$ and a decreasing function of $\sigma^{(t)}$. Moreover, it is worth noting that even the server can change the channel capacity by crafting $\boldsymbol{W}_i^{(t)}$, the victim can constrain the transmitted information within κ by solving $f^{(t)}(\sigma^{(t)}) = \kappa$ after receiving $\boldsymbol{W}_i^{(t)}$, which decides the added noise $\sigma^{(t)}$.

5.2 Limiting Channel Capacity in Data Space

For a controlled parameter channel, our target is to constrain the transmitted information at round t, i.e., $I(\boldsymbol{D};\widetilde{\boldsymbol{W}}_o^{(t)}|\boldsymbol{W}_i^{(t)})$. Based on Thm. 2, there are two steps for solving the equation $C^{(t)} = \kappa$ in the parameter space: the eigen-decomposition of $\boldsymbol{\Sigma}^{(t)}$ and solving the high order equation with corresponding eigenvalues. Whereas, $\boldsymbol{W}_o^{(t)}$ is a high-dimensional and time-variant parameter, which leads to an extremely large number of calculations. Hence, the implementation of the aforementioned method in the parameter space is computationally expensive.

Regarding $\boldsymbol{W}_o^{(t)}$ in Fig. 2, when $\boldsymbol{W}_i^{(t)}$ is observed, it is conditional independent with the previous parameters $\boldsymbol{W}_*^{(s)}$, $\forall s < t$, and all of the previous local learning processes. Therefore, if $\boldsymbol{W}_i^{(t)}$ is observed, $\boldsymbol{W}_o^{(t)}$ is purely decided by the local learning process at round t, i.e., $\boldsymbol{D} \to \boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)}$. Then based on these properties, we design a random function $M(\cdot)$ to map the raw data $\boldsymbol{D}$ to the noisy data $\widetilde{\boldsymbol{D}}$, and then utilize the noisy data for the local training, i.e., $\widetilde{\boldsymbol{W}}_o^{(t)} = \boldsymbol{W}_i^{(t)} - \eta\cdot\nabla_{\boldsymbol{W}}F(\boldsymbol{W}_i^{(t)};\widetilde{\boldsymbol{D}})$.

If we use the random function $M(\cdot)$ before the local learning process, the variables form a Markov Chain, i.e., $\boldsymbol{D} \to \widetilde{\boldsymbol{D}} \to \boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)}$. According to the DPI [15], we have

$$I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)}) \leq I(\boldsymbol{D};\widetilde{\boldsymbol{D}}|\boldsymbol{W}_i^{(t)}) = I(\boldsymbol{D};\widetilde{\boldsymbol{D}}), \tag{16}$$

where the last equality results from the independence between $M(\cdot)$ and $\boldsymbol{W}_i^{(t)}$. Therefore, we can bound $I(\boldsymbol{D};\widetilde{\boldsymbol{D}})$ so as to restrict $I(\boldsymbol{D};\boldsymbol{W}_o^{(t)}|\boldsymbol{W}_i^{(t)})$. Moreover, bounding $I(\boldsymbol{D};\widetilde{\boldsymbol{D}})$ enables us to limit the channel capacity in the data space. Specifically, we use $\widetilde{\boldsymbol{D}} = M(\boldsymbol{D}) = \boldsymbol{D} + \boldsymbol{\xi}$ as the random function, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})$, and the key parameter is $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$.

The rationale of constraining transmitted information in data space. In addition to the upper bound derived by DPI, we can explain the rationale of constraining in data space by Taylor's expansion for a more comprehensible analysis. Specifically, if we use a Gaussian noise in the data space to constrain the information leakage, we can expand the gradient mapping as follows

$$\begin{aligned}\nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D}+\boldsymbol{\xi}) \quad &= \nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D}) + \nabla_{\boldsymbol{D}}\nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D})^{\mathrm{T}}\cdot\boldsymbol{\xi} \\ &\quad + O(\|\boldsymbol{\xi}\|^2).\end{aligned} \tag{17}$$

Eq. (17) indicates that constraining the transmitted information in the data space is equivalent to adding an adaptive noise $\nabla_{\boldsymbol{D}}\nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D})^{\mathrm{T}}\cdot\boldsymbol{\xi}$ to the parameter.

Moreover, the coefficient $\nabla_{\boldsymbol{D}}\nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D})$is the variation of $\nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D})$. If $\nabla_{\boldsymbol{W}}F(\boldsymbol{W};\boldsymbol{D})$ changes significantly according to the data $\boldsymbol{D}$, which means the gradient has a high distinction degree with regard to the data, i.e., the gradient leaks more information, the large coefficient will provide strong privacy protection. Otherwise, the small coefficient leads to better utility without violating the privacy requirements.

Additionally, Fig. 4 provides a toy example for understand-

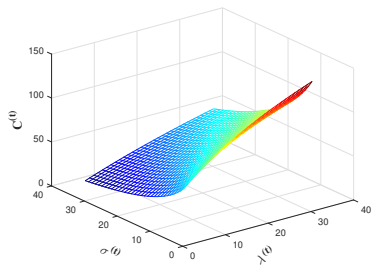

Figure 3: The channel capacity $C^{(t)}$ is the maximum MI increment at round t. It is an increasing function of $\lambda^{(t)}$ and a decreasing function of $\sigma^{(t)}$.

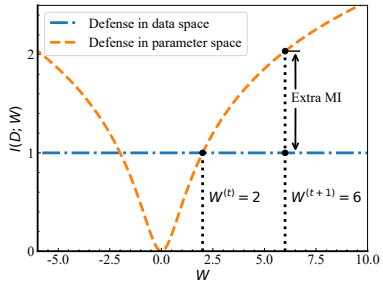

Figure 4: A toy example to explain the rationale for constraining in data space, which is equivalent to adding an adaptive noise to the parameter.

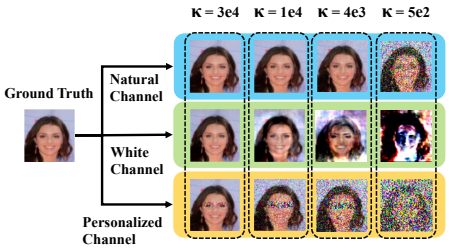

Figure 5: Visualizations for CelebA when we apply different channel implementations (Natural, White, and Personalized) and utilize different channel capacities.

ing the rationale of constraining in the data space. Specifically, we set $D \sim \mathcal{N}(-1,1)$ and our target is to decide the model W that minimizes $F(W;D) = W^2 D$ through gradient descent. Particularly, we require the information leakage to be less than 1, i.e., $C^{(t)} = 1$. As illustrated in Fig. 4, when we use the method in the parameter space to limit $C^{(t)} = 1$ at $W^{(t)} = 2$, i.e., adding the noise $\xi \sim \mathcal{N}(0, \frac{16}{e^2-1})$ to the gradient $\nabla_W F(W;D) = 2WD$, the resulting noise cannot guarantee $C^{(t+1)} = 1$ at next round $W^{(t+1)} = 6$ (we set the learning rate to 1), which implies that we need to recalculate the covariance matrix for the noise at each round. On the contrary, constraining in the data space (adding the noise $\xi \sim \mathcal{N}(0, \frac{1}{e^2-1})$ to the data D) results in an adaptive noise to guarantee $C^{(t)} = 1$ for all t regardless of $W^{(t)}$, leading to the $O(1)$ time-complexity for achieving the privacy requirement.

By incorporating different prior knowledge, we propose three implementation methods for deciding $\boldsymbol{\Sigma_\xi}$.

Natural Channel. For Natural Channel, the relative importance of different data attributes is naturally decided by the data itself. Specifically, based on Thm. 2, with substituting $\widetilde{\boldsymbol{D}}$ for $\widetilde{\boldsymbol{W}}_o^{(t)}$, we have

$$I(\boldsymbol{D};\widetilde{\boldsymbol{D}}) \le f(\sigma) = \frac{1}{2}\sum_{i=1}^{d} \ln \frac{\lambda_i + \sigma}{\sigma},\ \sigma \in (0, +\infty), \tag{18}$$

where λ_i and d represents the i-th eigenvalue and the dimension of data $\boldsymbol{D}$, respectively. In this case, we chose $\boldsymbol{\Sigma_\xi} = \sigma \mathbf{I}$. Moreover, $f(\sigma)$ is a monotone function of σ, hence there is a unique σ satisfying $f(\sigma) = \kappa$. Meanwhile, this σ guarantees $I(\boldsymbol{D};\widetilde{\boldsymbol{D}}) \le f(\sigma) = \kappa$. Combining it with Eq. (16), we conclude that the information leakage at time t is less than κ. However, $f(\sigma) = \kappa$ is a polynomial equation of order d, hence we need to solve it with numerical methods (e.g., binary search). In summary, the process of Natural Channel is

1. Make the eigen-decomposition of $\boldsymbol{\Sigma_D}$, which is the covariance matrix of data $\boldsymbol{D}$, to get $\Lambda = diag(\lambda_1, \cdots, \lambda_d)$
2. Solve the equation $f(\sigma) = \kappa$ to get σ with binary search
3. Get $\widetilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I})$

White Channel. In the method of the White Channel, we treat the relative importance of all attributes to be equal, which provides much stronger protection for the local dataset. For such a purpose, we add a constraint to Eq. (18) as

$$f(\boldsymbol{\Psi}) = \frac{1}{2}\sum_{i=1}^{d} \ln \frac{\lambda_i + \sigma_i}{\sigma_i} = \kappa \tag{19}$$
$$s.t. \quad \ln \frac{\lambda_i + \sigma_i}{\sigma_i} = \ln \frac{\lambda_j + \sigma_j}{\sigma_j},\ \text{for } 1 \le i < j \le d,$$

where $\boldsymbol{\Psi} = \text{diag}(\sigma_1, \cdots, \sigma_d)$ represents the eigenvalues of $\boldsymbol{\Sigma_\xi}$. Then we can get $\sigma_i = \frac{\lambda_i}{\exp(2\kappa/d)-1}$ by solving Eq (19). Finally, we need to transform $\boldsymbol{\Psi}$ back to $\boldsymbol{\Sigma_\xi}$. According to Eq. (25) in the proof of Thm. 2, we have $\boldsymbol{\Sigma_D} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, where $\boldsymbol{Q}$ is an orthogonal matrix composed with the eigenvectors of $\boldsymbol{\Sigma_D}$. Then with a similar process, we have $\boldsymbol{\Sigma_\xi} = \boldsymbol{Q}\boldsymbol{\Psi}\boldsymbol{Q}^T$. Therefore, the typical process of the White Channel is

1. Make the eigen-decomposition of $\boldsymbol{\Sigma_D}$ to get $\Lambda = diag(\lambda_1, \cdots, \lambda_d)$ and the eigenspace $\boldsymbol{Q}$
2. Get $\boldsymbol{\Psi} = \text{diag}(\sigma_1, \cdots, \sigma_d)$ by $\sigma_i = \frac{\lambda_i}{\exp(2\kappa/d)-1}$
3. Get $\boldsymbol{\Sigma_\xi} = \boldsymbol{Q}\boldsymbol{\Psi}\boldsymbol{Q}^T$
4. Get $\widetilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_\xi})$

Personalized Channel. In practice, the relative importance of different attributes in data is different. For example, [59] claims that for online diagnosis, our target is to extract disease-relevant features but remove identity features from the facial images of patients, which means we add more noise to the identity features compared to the disease-relevant features. For analyzing the problem, we assume the relative importance of different dimensions is $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{d-1})^T$, e.g., (height, weight)=(1, 2) represents the relative importance of the height to the weight is $1:2$.

The relative importance decides the level of noise addition. In other words, if the attribute is more important, we add more noise to it for stronger protection. To utilize the prior knowledge $\boldsymbol{\beta}$, we chose $\boldsymbol{\Sigma_\xi} = \sigma \cdot \text{diag}(\boldsymbol{\beta})$. According to the proof of Thm. 2, σ is decided by

$$f(\sigma) = \frac{\ln\det[\mathbf{\Sigma_D} + \sigma \cdot \mathrm{diag}(\boldsymbol{\beta})]}{2\sum_{i=1}^{d} \ln(\sigma \cdot \beta_i)} = \kappa. \tag{20}$$

However, the intrinsic correlations of local data bring difficulties in solving Eq. (20), so we derive the following theorem for simplifying the calculation.

Theorem 3 (Upper bound of Personalized Channel.)**.** *Let* $\mathbf{\Sigma} = \mathbf{\Sigma}_{d-1}$*, we can rewrite* $\mathbf{\Sigma}$ *as*

$$\mathbf{\Sigma}_i = \begin{pmatrix} \mathbf{\Sigma}_{i-1} & \boldsymbol{\rho}_i \\ \boldsymbol{\rho}_i^{\mathrm{T}} & c_{i,i} \end{pmatrix}, \ i \in \{0, \cdots, d-1\} \tag{21}$$

where d *represent the dimension of* $\mathbf{\Sigma}$*, then we have*

$$\ln\det(\mathbf{\Sigma} + \mathrm{diag}(\boldsymbol{\beta})) \leq \min[\sum_{i=0}^{d-1} \ln(c_{i,i} + \beta_i), \sum_{i=0}^{d-1} \ln(u_i + k_i)],$$

$$u_i = c_{i,i} - \boldsymbol{\rho}_i^{\mathrm{T}} \mathbf{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i, \ k_i = \beta_i + (\mathbf{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i)^{\mathrm{T}} \mathrm{diag}(\boldsymbol{\beta}) (\mathbf{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i).$$

Combining Thm. 3 with Eq. (20), we can get an upper bound of the channel capacity for the Personalized Channel

$$U(\sigma) = \frac{\min[\sum_{i=0}^{d-1} \ln(c_{i,i} + \sigma\beta_i), \sum_{i=0}^{d-1} \ln(u_i + \sigma k_i)]}{2\sum_{i=1}^{d} \ln(\sigma \cdot \beta_i)}.$$

Moreover, $U(\sigma)$ decouples $\mathbf{\Sigma}$ and $\mathrm{diag}(\boldsymbol{\beta})$,hence we can use a pre-processing to reduce the calculation of solving $U(\sigma) = \kappa$. The process of the Personalized Channel is

1. Get u_i and k_i according to Thm. (3)
2. Solve the equation $U(\sigma) = \kappa$ to get σ with binary search
3. Get $\mathbf{\Sigma_\xi} = \sigma \cdot \mathrm{diag}(\boldsymbol{\beta})$
4. Get $\widetilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_\xi})$

Finally, we visualize $\widetilde{\boldsymbol{D}}$ of different implementations according to different channel capacities in Fig. 5.

Furthermore, constraining in the data space brings two advantages: first, compared to the parameter space, data space is white-box, low-dimensional, and time-invariant. Second, in the data space, the relative importance of attributes is preserved, which makes it easier to leverage prior knowledge.

Guidelines for noise injection. In summary, for defending against DRA, the important target is to restrict the transmitted information by noise addition. Here we provide guidelines for the noise injection:

- For privacy-enhancing techniques (e.g., DP and gradient compression) in FL, the reconstruction error for DRA is theoretically above a threshold (Thm. 1) when we restrict the transmitted information within κ. It can be achieved by solving $f(\sigma) = \kappa$, where $f(\sigma)$ is defined in Eq. (18). This equation decides the injection noise with corresponding statistics.
- For defending against DRA in FL, we can inject noise to the training data instead of transmitted parameters. This transformation can produce an adaptive noise, which reduces the computational complexity brought by the high dimensional and time-variant system in FL.
- For any pre-processing process such as embedding, we can get the same theoretical guarantee by substituting $\mathbf{\Sigma}_{embedding}$ for $\mathbf{\Sigma_D}$ in Eq. (18), where $\mathbf{\Sigma}_{embedding}$ and $\mathbf{\Sigma_D}$ are covariance matrixes calculated by embeddings and data, respectively.
- In our method, a larger batch size leads to a tighter upper bound and a stronger ability to defend against DRA.

5.3 Channel Capacity for Existing Methods

With Thm. 1 and Thm. 2, we can analyze existing methods for defending against DRA by the channel capacity $C^{(t)}$.

Existing privacy-enhancing methods in FL can be divided into three categories: perturbation, compression, and utilizing large batch size. The method of perturbation is represented by DP. Moreover, we demonstrate that for defending against DRA, the methods of gradient compression and utilizing large batch size are equivalent to adding more noise for perturbation, and the intrinsic mechanism of them are restricting the transmitted information.

Perturbation. As a method with theoretical guarantee, DP focuses on the problem of protecting individual information. In this work, we consider the event-level DP [35, 37] because we focus on the privacy issue for a specific client (i.e., the victim). That is, whether the attacker can reconstruct the local dataset with the victim's shared parameters. Specifically, the widely used DP technique in FL is the Gaussian mechanism, thus we analyze the Gaussian mechanism in this section. A typical process of the Gaussian mechanism for (ε, δ)-DP consists of two stages: gradient clipping, which guarantees the sensitivity of the gradient is bounded; noise addition, which provides the (ε, δ)-DP guarantee based on the bounded sensitivity. To analyze DP based on information theory, we have the following theorem.

Theorem 4 (Channel capacity for DP)**.** *In FL, if the sensitivity of the gradient mapping is upper bounded by S, i.e.,* $\|\boldsymbol{g}\|_2 \leq S$*, the channel capacity of* (ε, δ)*-DP, i.e.,* C_{DP}*, is upper bounded by following formulas:*

(1) $C_{DP} \leq \frac{B \cdot S^2}{\sigma}$,

(2) $C_{DP} \leq \frac{B \cdot \varepsilon^2}{2\log(1.25/\delta)}$,

where B *represents the batch size and* σ *is the noise scale.*

Compared to the conventional DP theorem, Thm. 4 indicates that batch size B is a key factor of the defense ability to defend against DRA, which has been overlooked by prior literature. Specifically, increasing B reduces DP's ability to defend against DRA, which has been validated by experiments in Fig. 6 (the details are explained in Appendix D).

Compression. Another defense technique is compression, it intuitively focuses on reducing transmitted information by reducing the dimension of shared parameters. However, most of these methods lack theoretical guarantees.

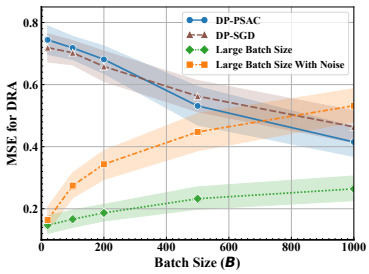

Figure 6: Defensive capabilities of DP and utilizing large batch size according to different batch sizes.

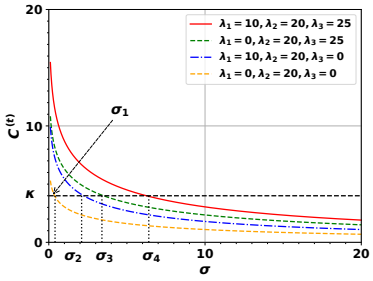

Figure 7: Reducing the dimensionality of a gradient is equivalent to adding more noise into the original gradient.

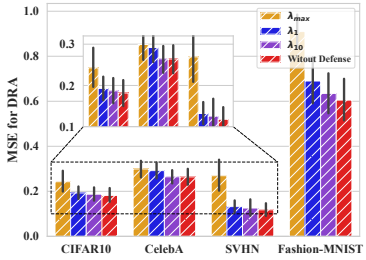

Figure 8: DRA becomes more difficult when we compress the dimension with a large eigenvalue.

Specifically, information is contained in each dimension of the gradient (i.e., the parameter), and the channel capacity decreases accordingly when we reduce the dimension of it. If we analyze the compression in the eigenspace, dimension reduction can be theoretically described by Thm. 2. Specifically, the total channel capacity is $C^{(t)} = \frac{1}{2}\sum_{i=1}^{d} \ln \frac{\lambda_i^{(t)}+\sigma}{\sigma}$, and the general term $\frac{1}{2}\ln \frac{\lambda_i^{(t)}+\sigma}{\sigma}$ represents the channel capacity of the i-th dimension. Compressing the i-th dimension in the eigenspace, i.e., setting $\lambda_i^{(t)} = 0$, leads the channel capacity of i-th dimension to be 0, thereby reducing the total channel capacity. As illustrated in Fig. 7, when we compress the gradient, i.e., setting eigenvalues to be 0, the results are equivalent to adding more noise for perturbation. As displayed in Fig. 8, we also conduct experiments to validate our theoretical result, the experiments indicate that the improvement of the defense ability becomes more significant when we compress the dimension with a larger eigenvalue (the details are explained in Appendix D).

Large Batch Size. Another defense strategy for privacy protection in FL is utilizing large batch size [63]. Our model can theoretically formalize this strategy. If we denote the batch size as B, then the gradient in mini-batch SGD is $\frac{1}{B}\sum_{i=1}^{B} \nabla_{\boldsymbol{W}} F(\boldsymbol{W}_i^{(t)}; \boldsymbol{D}_i)$, where $\{\boldsymbol{D}_i\}_{i=1}^{B}$ represents the set of iid data points sampled from the dataset. The covariance matrix with large batch size is scaled by B, i.e., $\boldsymbol{\Sigma}_B^{(t)} = \frac{1}{B}\boldsymbol{\Sigma}^{(t)}$. Hence, with substituting $\boldsymbol{\Sigma}_B^{(t)}$ for $\boldsymbol{\Sigma}^{(t)}$ in Thm. 2, we have

$$C_B^{(t)} = f_B^{(t)}(\sigma) := \frac{1}{2}\sum_{i=1}^{d} \ln \frac{(\lambda_i^{(t)}/B+\sigma)}{\sigma}. \tag{22}$$

With the constant noise addition, $C^{(t)}$ is a decreasing function of B, which means we can enhance the ability to defend against DRA by increasing B. Fig. 6 experimentally validates this theory and we put the details of Fig. 6 in Appendix D due to the space limitation.

Guidelines for hyper-parameters. Finally, to improve existing defensive algorithms, we provide several guidelines for choosing the hyper-parameters in FL:

- A smaller batch size in DP algorithm is more effective to defend against DRA.
- Compressing the dimension with larger eigenvalue results in the stronger ability to defend against DRA.
- A larger batch size leads to a stronger defense ability to defend against DRA when we add a constant noise to the parameter during training.

6 Experiment

In this section, we conduct experiments with various models and datasets to validate our theories and compare our methods with other privacy-enhancing techniques. All experiments are performed upon a Supermicro SYS-420GP-TNR server with two Intel(R) Xeon(R) Gold 6348 CPUs (2×28 cores), Ubuntu 18.04.1, 10GB memory, and four NVIDIA A100 PCIe 80GB GPUs. Meanwhile, to eliminate the impact of randomness, each experiment is repeated 10 times.

6.1 Experimental Settings

Datasets. Dataset is a task-dependent factor, which means we cannot change the dataset when the training task is decided. In our experiments, we resize the data to 32*32 for comparison, and utilize four classical datasets, including CIFAR10 [31], CelebA [39], SVHN [45], and Fashion-MNIST [57].

Basic Models. In this work, we experiment with four classical model architectures, including LeNet [33], AlexNet [32], VGG16 [50], and ResNet10 [27]. Moreover, model architectures are task-independent, we can select architectures according to different goals, e.g., utility or privacy.

Attacks. We test our methods on two typical attacks in FL: *MIA* and *DRA*, the details of different attacks are as follows.

Membership Inference Attack. For MIA, we use the white-box attack [38, 44]. Meanwhile, we employ a partial knowledge attacker, which means the attacker can access to part of the training dataset. In this case, the attacker has much stronger background knowledge. Moreover, we use four inputs for attacking [38]: the samples' ranked posteriors, classification loss, gradients of the last layer, and one-hot encoding of the true label. These inputs are fed into different neural networks to get different embeddings, then we concatenate all embeddings as the input of a 4-layer MLP to get the inference.

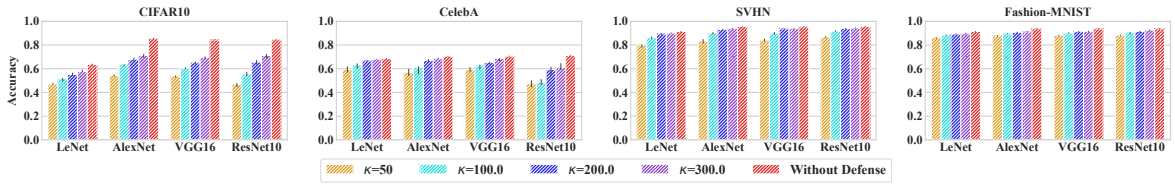

Figure 9: The effect of channel capacities (κ) for model accuracy (Natural Channel).

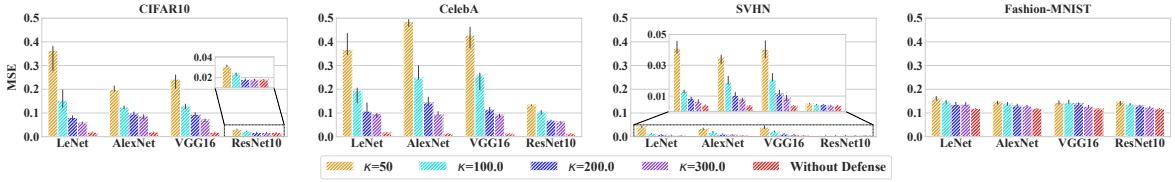

Figure 10: The effect of channel capacities (κ) for DRA (Natural Channel).

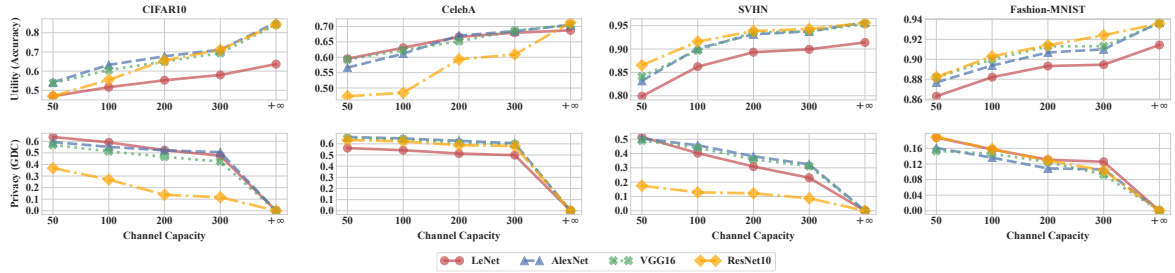

Figure 11: Utility privacy tradeoff according to different channel capacities.

Data Reconstruction Attack. For DRA, we employ two representative attacks: model inversion attack and gradient inversion attack. For model inversion attack [20, 38], we first construct a dummy input with auxiliary information (i.e., the mean value of images that are not in the training dataset) as the input for the target model, then utilize different target labels to optimize the dummy input. We use Adam optimizer with a learning rate of 1e-2 for 600 iterations. Additionally, we employ the settings proposed by Geiping et al. [21] to investigate the defense ability for gradient inversion attacks.

Metric. We evaluate our design with various criteria, including the model utility, the defense capability against different attacks, the utility-privacy trade-offs, and the efficiency. To this end, we employ the following metrics for evaluation.

Test Accuracy. FL searches for an accurate model for classification, hence we use test accuracy as the metric for utility.

AUC. We use the attack AUC to measure the attack performance for MIA. It's worth noting that a smaller AUC means a stronger defense capability.

MSE. We use MSE as the main metric for DRA since MSE is a general metric that indicates the convergence of random variables. That is, if the MSE of two random variables is 0, we conclude that they have identical distributions [55], which means a perfect reconstruction. Additionally, a large MSE means a stronger ability to defend against DRA. Specifically, for model inversion attacks, we calculate the metric between the reversed data and the center of the corresponding class for different classes. We use the median of all classes as the final metric. For gradient inversion attacks, we employ the mean value of the metric between the reconstructed data and the target data.

General Defense Capability (GDC). For evaluating the comprehensive defense capability, we use the improvements of the aforementioned attacks. Specifically, for MIA, we define the improvement as

$$IMP_{MIA} = (AUC_{without\_def} - AUC)/AUC_{without\_def}.$$

While the improvement for DRA is

$$IMP_{DRA} = (\frac{1}{MSE_{without\_def}} - \frac{1}{MSE})/\frac{1}{MSE_{without\_def}}.$$

Finally, we define general defense capability as

$$GDC = (IMP_{MIA} + IMP_{DRA})/2,$$

and a larger *GDC* means a stronger defense capability.

Default parameter configuration. In our experiments, we mainly utilize natural channel to investigate the channel parameters. Specifically, we fix the optimization rounds as $n = 1 \times 10^4$ and utilize κ in {50, 100, 200, 300} to investigate the effect of different channel capacities. Additionally, we fix $\kappa = 300$ and utilize $TotalInfo$ in $\{5 \times 10^5, 1 \times 10^6, 2 \times 10^6, 3 \times 10^6\}$ to investigate the effect of different n, where $n = \lfloor \frac{TotalInfo}{\kappa} \rfloor$ and $\lfloor \cdot \rfloor$ is the floor function. We will explicitly explain it when we utilize different configurations.

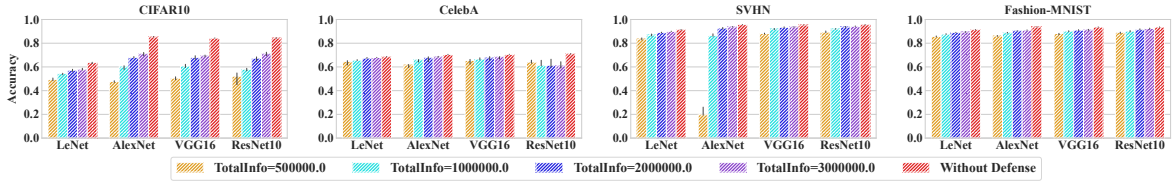

Figure 12: The effect of optimization number (n) for model accuracy when $\kappa = 300$ (Natural Channel).

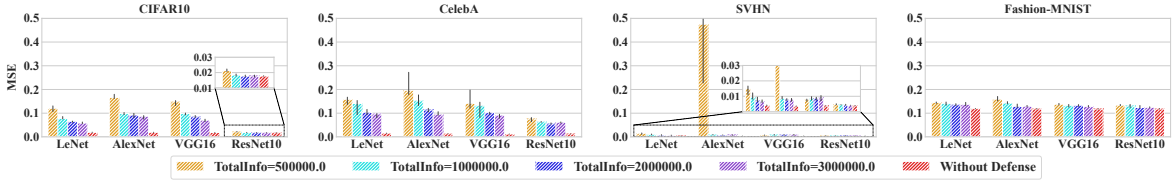

Figure 13: The effect of optimization number (n) for DRA when $\kappa = 300$ (Natural Channel).

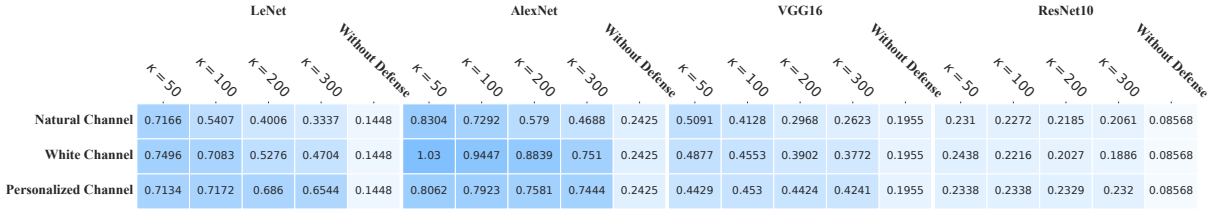

Figure 14: Heatmaps of MSE for the gradient inversion attacks on CIFAR-10.

6.2 The Effect of Controlled Channel

Impact of channel capacity (κ). κ represents the upper bound of information leakage in a single training round. As illustrated in Fig. 9, when we enhance the channel, i.e., increase κ, the accuracy increases accordingly regardless of the datasets and the model architectures, which means a wider channel leads to a better utility. However, the increasing utility is at the cost of reducing data privacy. Fig. 10 displays the corresponding defense capability against DRA. Specifically, for DRA attacks, the MSE decreases according to the increasing κ, which means the reconstructed data is closer to the target data. These results indicate that in the practical scenario, we can adjust channel capacity by κ to balance the utility and the privacy according to various requirements.

Moreover, we explicitly display the utility-privacy tradeoff in Fig. 11. Specifically, the utility increases with the increase of channel capacity κ, while the ability of privacy protection decreases with the increase of κ. These results are reasonable since channel capacity constrains the transmitted information, and the reconstruction error, i.e., the MSE, can be theoretically restricted when the transmitted information is limited based on Thm. 1. Meanwhile, more available information results in a more accurate model, indicating that the adjustable channel capacity can be utilized to balance the utility-privacy tradeoff.

Impact of optimization rounds (n). Different from the channel capacity C, the number of optimization rounds n affects the parameter channel through another dimension: information accumulation. Similarly, the results in Fig. 12 and 13 imply that n is another parameter to influence the utility and the defense ability of the controlled channel. Increasing n results in obtaining more information from the local dataset, thereby enhancing the utility while reducing the defense ability.

Defense ability against gradient inversion attack. We randomly select 30 images from each dataset as the target data and evaluate the defense abilities to defend against gradient inversion attacks. As shown in Fig. 14, the MSE of data reconstruction consistently increases as we decrease the threshold κ, which means a smaller κ leads to a stronger ability to defend against DRA. The MSE of the White Channel and the Personalized Channel are larger than the Natural Channel, indicating that the other two methods provide stronger defense capability compared to the Natural Channel.

6.3 Comparing with Other Methods

In this section, we compare our methods with two classic perturbation mechanisms: DP-SGD [1] and DP-PSAC [56]. Specifically, DP-SGD first utilizes DP for ML and DP-PSAC is the state-of-art method to improve DP-SGD by adaptively clipping the gradients.

Utility-Privacy trade-off. We first compare our methods with DP-SGD and DP-PSAC in the utility-privacy plane. The plane is formed by two axes: the x-axis represents GDC, which is the general ability for privacy protection. While the y-axis represents the accuracy, which represents the utility. For all methods, we use the model of LeNet and fix the rounds as $n = 1 \times 10^4$. The results are displayed in Fig. 15. Specifically, in our methods, we use channel capacity in {50, 100, 200,

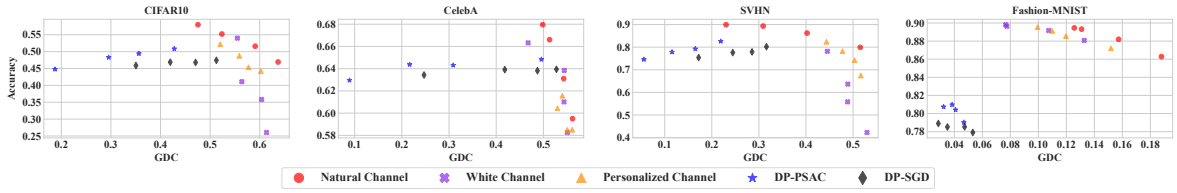

Figure 15: Evaluating different methods in the Utility-Privacy plane.

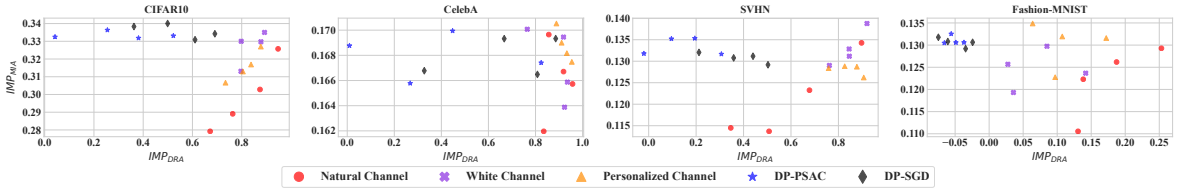

Figure 16: Evaluating different methods in the Defense-Defense plane.

300} for the Natural Channel, and {500, 800, 1000, 1500} for the White Channel and the Personalized Channel. Additionally, we use 1:50 in $\boldsymbol{\beta}$ for the Personalized Channel to protect the attributes around the eyes (Appendix C indicates the details of $\boldsymbol{\beta}$). For DP-SGD and DP-PSAC, we use the clipping bound $S = 1.0$, $\delta = 1 \times 10^{-5}$ ($\delta = 3 \times 10^{-6}$ for CelebA), and the multiplier $\sqrt{\sigma}$ in {0.8, 0.57, 0.46, 0.2066} (the corresponding privacy budget ε for these DP training are {1.705, 5.120, 12.026, 331.668}, respectively) . According to Thm. 4, these σ result in channel capacities in {100, 200, 300, 1500} when we utilize $B = 64$, which are identical to the channel capacities of our methods. As in Fig. 15, the Natural Channel achieves the best utility-privacy trade-off. The reason is that it utilizes the original importance to maintain the information in data attributes, and the constrained channel capacity ensures the ability to defend against DRA.

Comparison of the details in defense ability. To investigate the detailed defense ability of different methods, we decouple the defense capabilities to display them in the defense-defense plane. Specifically, x-axis represents IMP_{DRA} and y-axis represents IMP_{MIA}. As displayed in Fig. 16, DP specializes in defending against MIA but is weaker in defending against DRA, this is reasonable since DP focuses on protecting individual information. Compared to DP, our methods specialize in defending against DRA. Moreover, our methods achieve the best comprehensive defense capability for both attacks.

Table 2: Training time (s) for 25 epochs of different models

Model	Parameters	SGD	DP-SGD	Ours
LeNet	6.20×10^4	283.43±0.95	445.45±20.41 (+57.16%)	312.99±1.40 (+10.43%)
ResNet10	4.90×10^6	624.23±8.66	1397.34±17.38 (+123.85%)	651.29±10.39 (+4.33%)
AlexNet	3.59×10^7	374.13±2.68	1023.83±2.85 (+173.66%)	406.02±1.04 (+8.52%)
VGG16	1.34×10^8	581.69±5.35	3402.91±4.72 (+485.00%)	612.99±4.62 (+5.38%)

Efficiency of constraining in the data space. As displayed in Tab. 2, we compare the efficiency of our method with DP-SGD, which is a method that works in the parameter space. We use OPACUS 1.1.2 [60] for the implementation of DP-SGD. The results indicate that when we transform the operations to the data space, it significantly reduces the amount of calculation, i.e., the training time, especially when the dimension of the parameter increases.

7 Related Work

Differential Privacy. Differential Privacy [18] is an important technique for privacy protection. For the definition of differential privacy, Mironov [43] proposes to utilize Rényi divergence, which leads to compact and accurate privacy loss. Moreover, Abadi et al. [1] propose an empirical algorithm DP-SGD for applying DP to ML training, then multiple literature tries to improve the performance of DP-SGD [3, 6, 9, 56]. Most of them focus on the utility, trying to adaptively clip the gradients or add suitable perturbations. For the application of DP, Levy et al. [35] propose to protect user-level DP instead of ensuring the privacy of individual samples, which makes DP more reliable to FL. Truex et al. [53] present a protocol LDP-Fed to formally ensure data privacy in collecting local parameters with high precision. However, DP is still vulnerable to DRA, several researchers indicate that they can reconstruct the data without violating the requirements of DP [14, 17], implying that the algorithms are still vulnerable to DRA. Compared to DP, our work aims to defend against DRA based on information theory.

Limiting Mutual Information. Another related concept of privacy protection depends on MI. Several studies connect MI with DP by deriving the upper bound of MI for a distinct DP mechanism [4, 8, 16]. For privacy protection, Li et al. [36] propose to train a feature extractor that minimizes MI between the output features and the assigned label while maximizing the MI between output features and the original

data. Similarly, Osia et al. [46] propose to train an extractor with the variational lower bound for MI estimation. Moreover, Hannun et al. [25] utilize Fisher Information to measure the information leakage, and the conclusions are consistent with our theories. Finally, we can also model FL based on information theory [2, 54], which measures the MI between different variables in FL. Compared to the former methods, our technique models the communication channel of FL directly based on information theory and utilizes its upper bound to derive practical methods for constraining the transmitted information, which leads to the strong ability to defend against DRA.

8 Discussion and Conclusion

Our method constrains the information leakage of the black-box model according to an upper bound derived by maximum entropy distribution (Lemma 1). We can tighten the upper bound by utilizing a large batch size or incorporating domain knowledge. Among them, utilizing large batch size for training causes the output distribution to be closer to a Gaussian distribution, while incorporating domain knowledge enables us to get more properties of the output distribution. If we get a tighter upper bound, the tradeoff between utility and privacy can be further improved.

In summary, as the reconstruction error of DRA is decided by the transmitted information, we build a channel model to measure the information leakage of the black-box model in FL. The model indicates that the amount of transmitted information is decided by the channel capacity C and the number of optimization rounds n. Guided by the model, we develop methods to constrain the channel capacity within a threshold κ. Combining it with the limited optimization rounds n, the upper bound of the total transmitted information remains below $n \cdot \kappa$, which ensures the ability to defend against DRA. Furthermore, we transform the operations of constraining channel capacity from the parameter space to the data space. The transformation significantly improves the training efficiency and the model accuracy under constrained information leakage. Finally, extensive experiments with real-world datasets validate the benefit of our methods.

9 Acknowledgments

We thank our shepherd and anonymous reviewers for their thoughtful comments. This work was supported in part by the National Science Foundation for Distinguished Young Scholars of China under No. 61825204, National Natural Science Foundation of China under No. 62202258, No. 62132011, No. 61932016, Beijing Outstanding Young Scientist Program under No. BJJWZYJH01201910003011, China Postdoctoral Science Foundation under No. 2021M701894, China National Postdoctoral Program for Innovative Talents, Shuimu Tsinghua Scholar Program, Lenovo Young Scientist Program, and the Beijing National Research Center for Information Science and Technology key projects. Yi Zhao and Ke Xu are the corresponding authors.

References

[1] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proc. of CCS*, 2016.

[2] L. Adilova, J. Rosenzweig, and M. Kamp. Information-theoretic perspective of federated learning. *CoRR*, abs/1911.07652, 2019.

[3] N. Agarwal, A. T. Suresh, F. X. Yu, S. Kumar, and B. McMahan. cpsgd: Communication-efficient and differentially-private distributed SGD. In *Proc. NeurIPS*, 2018.

[4] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi. Differential privacy: On the trade-off between utility and information leakage. In *Proc. FAST*, 2011.

[5] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith. *The Science of Quantitative Information Flow*. Information Security and Cryptography. Springer, 2020.

[6] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy. Differentially private learning with adaptive clipping. In *Proc. NeurIPS*, pages 17455–17466, 2021.

[7] B. Balle, G. Cherubin, and J. Hayes. Reconstructing training data with informed adversaries. In *Proc. S&P*, 2022.

[8] G. Barthe and B. Köpf. Information-theoretic bounds for differentially private mechanisms. In *Proc. CSF*, 2011.

[9] Z. Bu, Y. Wang, S. Zha, and G. Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *CoRR*, abs/2206.07136, 2022.

[10] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *Proc. S&P*, 2022.

[11] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. In *Proc. USENIX Security*, 2021.

[12] J. Chen, Y. Zhao, Q. Li, X. Feng, and K. Xu. Feddef: Defense against gradient leakage in federated learning-based network intrusion detection systems. *IEEE TIFS*, 2023.

[13] C. A. Choquette-Choo, F. Tramèr, N. Carlini, and N. Papernot. Label-only membership inference attacks. In *Proc. ICML*, 2021.

[14] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *Proc. SIGKDD*, 2011.

[15] T. M. Cover and J. A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[16] P. Cuff and L. Yu. Differential privacy as a mutual information constraint. In *Proc. SIGSAC*, 2016.

[17] T. Dick, C. Dwork, M. Kearns, T. Liu, A. Roth, G. Vietri, and Z. S. Wu. Confidence-ranked reconstruction of census microdata from published statistics. *The National Academy of Sciences*, 2023.

[18] C. Dwork. Differential privacy. In *Proc. ICALP*, 2006.

[19] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

[20] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of SIGSAC on computer and communications security*, pages 1322–1333, 2015.

[21] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Proc. NeurIPS*, 2020.

[22] C. Guo, B. Karrer, K. Chaudhuri, and L. van der Maaten. Bounding training data reconstruction in private (deep) learning. In *Proc. ICML*, Proceedings of Machine Learning Research, 2022.

[23] P. Hagerty and T. Draper. Entropy bounds and statistical tests. In *Proc. on the NIST Random Bit Generation*, 2012.

[24] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani. Reconstructing training data from trained neural networks. In *Proc. NeurIPS*, 2022.

[25] A. Y. Hannun, C. Guo, and L. van der Maaten. Measuring data leakage in machine-learning models with fisher information (extended abstract). In *Proc. IJCAI*, 2022.

[26] J. Hayes, S. Mahloujifar, and B. Balle. Bounding training data reconstruction in DP-SGD. *CoRR*, abs/2302.07225, 2023.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[28] B. Hitaj, G. Ateniese, and F. Pérez-Cruz. Deep models under the GAN: information leakage from collaborative deep learning. In *Proc. CCS*, 2017.

[29] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *Proc. ICLR*, 2019.

[30] Y. Kim, C. Guyot, and Y. Kim. On the efficient estimation of min-entropy. *IEEE TIFS*, 2021.

[31] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NeurIPS*, 2012.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 1998.

[34] M. Lécuyer, R. Spahn, K. Vodrahalli, R. Geambasu, and D. Hsu. Privacy accounting and quality control in the sage differentially private ML platform. In *Proc. SOSP*, 2019.

[35] D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh. Learning with user-level privacy. In *Proc. NeurIPS*, 2021.

[36] A. Li, Y. Duan, H. Yang, Y. Chen, and J. Yang. TIPRDC: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In *Proc. KDD*, 2020.

[37] Y. Liu, A. T. Suresh, F. X. Yu, S. Kumar, and M. Riley. Learning discrete distributions: user vs item-level privacy. In *Proc. NeurIPS*, 2020.

[38] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. D. Cristofaro, M. Fritz, and Y. Zhang. Ml-doctor: Holistic risk assessment of inference attacks against machine learning models. In *Proc. USENIX Security*, 2022.

[39] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015.

[40] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen. A pragmatic approach to membership inferences on machine learning models. In *Proc. EuroS&P*, 2020.

[41] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. AISTATS*, 2017.

[42] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *Proc. S&P*, 2019.

[43] I. Mironov. Rényi differential privacy. In *Proc. CSF*, 2017.

[44] 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. S&P*, 2019.

[45] Y. Netzer, T. Wang, A. Coates, A. Bissacco, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. *Proc. NeurIPS*, 2011.

[46] S. A. Osia, A. Taheri, A. S. Shamsabadi, K. Katevas, H. Haddadi, and H. R. Rabiee. Deep private-feature extraction. *IEEE Trans. Knowl. Data Eng.*, 2020.

[47] M. Romanelli, K. Chatzikokolakis, C. Palamidessi, and P. Piantanida. Estimating g-leakage via machine learning. In *Proc. CCS*, 2020.

[48] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proc. ICML*, 2019.

[49] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *CoRR*,

abs/1703.00810, 2017.
[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, *Proc. ICLR*, 2015.
[51] G. Smith. On the foundations of quantitative information flow. In L. de Alfaro, editor, *Proc. FOSSACS*, 2009.
[52] C. Thapa, M. A. P. Chamikara, S. Camtepe, and L. Sun. Splitfed: When federated learning meets split learning. In *Proc. AAAI*, 2022.
[53] S. Truex, L. Liu, K. H. Chow, M. E. Gursoy, and W. Wei. Ldp-fed: federated learning with local differential privacy. In *Proc. on Edge Systems, Analytics and Networking*, 2020.
[54] M. P. Uddin, Y. Xiang, X. Lu, J. Yearwood, and L. Gao. Mutual information driven federated learning. *IEEE TPDS*, 2021.
[55] L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.
[56] T. Xia, S. Shen, S. Yao, X. Fu, K. Xu, X. Xu, X. Fu, and W. Wang. Differentially private learning with per-sample adaptive clipping. *CoRR*, abs/2212.00328, 2022.
[57] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.
[58] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 2019.
[59] Y. Yang, J. Lyu, R. Wang, Q. Wen, L. Zhao, W. Chen, S. Bi, J. Meng, K. Mao, Y. Xiao, et al. A digital mask to safeguard patient privacy. *Nature medicine*, 2022.
[60] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov. Opacus: User-friendly differential privacy library in pytorch. *CoRR*, abs/2109.12298, 2021.
[61] Y. Zhao, K. Xu, J. Chen, and Q. Tan. Collaboration-enabled intelligent internet architecture: Opportunities and challenges. *IEEE Netw.*, 2022.
[62] G. Zhou, Q. Li, Y. Liu, Y. Zhao, Q. Tan, S. Yao, and K. Xu. Fedpage: Pruning adaptively toward global efficiency of heterogeneous federated learning. *IEEE/ACM Transactions on Networking*, 2023.
[63] L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. In *Proc. NeurIPS*, 2019.

A The detail explanation of DRA

For DRA in FL, the attacker aims to build a reconstruction $\hat{\boldsymbol{D}}(\boldsymbol{W}_i, \boldsymbol{W}_o)$ to approximate the data $\boldsymbol{D}$, where $\boldsymbol{W}_i$ and $\boldsymbol{W}_o$ are the transmitted parameters in FL. Moreover, if the MSE between $\hat{\boldsymbol{D}}(\boldsymbol{W}_i, \boldsymbol{W}_o)$ and $\boldsymbol{D}$ achieves 0, i.e., $\mathbb{E}\|\hat{\boldsymbol{D}}(\boldsymbol{W}_i, \boldsymbol{W}_o) - \boldsymbol{D}\|^2 = 0$, we conclude that $\hat{\boldsymbol{D}}(\boldsymbol{W}_i, \boldsymbol{W}_o)$ and $\boldsymbol{D}$ have identical distributions [55].

Additionally, with the definition of $\boldsymbol{D}$, i.e., the random variable that follows the distribution of the local dataset, the MIA [10] can be viewed as inferring whether a particular data point belongs to the support set of $\boldsymbol{D}$.

B Proofs of Lemmas and Theorems

B.1 Proof of Theorem 1

Before the proof of Thm. 1, we have the following lemma.

Lemma 2. *For any d-dimensional semi-positive definite matrix $\boldsymbol{A}$, i.e., $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, we have* $\det(\boldsymbol{A}) \leq \left(\frac{\mathrm{tr}(\boldsymbol{A})}{d}\right)^d$.

Proof. Since $\boldsymbol{A}$ is a semi-positive definite matrix, we have its eigen values, i.e., $\{\lambda_i\}_{i=1}^d$, are non-negative. we can get

$$\det(\boldsymbol{A}) = \prod_{i=1}^{d} \lambda_i \leq \left(\frac{\sum_{i=1}^d \lambda_i}{d}\right)^d = \left(\frac{\mathrm{tr}(\boldsymbol{A})}{d}\right)^d,$$

where the inequality depends on the AM-GM inequality. □

Then based on Lemma 2, we can prove Thm. 1.

Proof. As $Cov(\boldsymbol{D})$ is the covariance matrix of $\boldsymbol{D}$, we have

$$\begin{aligned}
h(\boldsymbol{D}|\boldsymbol{W}) &\overset{(1)}{\leq} \mathbb{E}_{\boldsymbol{W}}[\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det(Cov(\boldsymbol{D}|\boldsymbol{W}))] \\
&\overset{(2)}{\leq} \frac{d}{2}\log(2\pi e) + \mathbb{E}[\frac{d}{2}\log(\frac{\mathrm{tr}(Cov(\boldsymbol{D}|\boldsymbol{W}))}{d})] \\
&\overset{(3)}{\leq} \frac{d}{2}\log(2\pi e) + \mathbb{E}[\frac{d}{2}\log(\frac{\mathbb{E}[\|\boldsymbol{D}-\hat{\boldsymbol{D}}(\boldsymbol{W})\|^2|\boldsymbol{W}]}{d})] \\
&\overset{(4)}{\leq} \frac{d}{2}\log(2\pi e) + \frac{d}{2}\log(\frac{\mathbb{E}[\mathbb{E}[\|\boldsymbol{D}-\hat{\boldsymbol{D}}(\boldsymbol{W})\|^2|\boldsymbol{W}]]}{d}) \\
&= \frac{d}{2}\log(2\pi e) + \frac{d}{2}\log(\frac{\mathbb{E}[\|\boldsymbol{D}-\hat{\boldsymbol{D}}(\boldsymbol{W})\|^2]}{d}),
\end{aligned}$$

where (1) is a consequence that with identical mean vector and covariance matrix, Gaussian distribution is the maximum entropy distribution. (2) depends on Lemma 2. The reason for (3) is that mean vector is the optimal estimator in terms of mean squared error, and (4) is a consequence of Jensen's inequality. Hence,

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{D}-\hat{\boldsymbol{D}}(\boldsymbol{W})\|^2/d]) &\geq \frac{1}{2\pi e} e^{2h(\boldsymbol{D}|\boldsymbol{W})/d} \\
&= \frac{e^{2h(\boldsymbol{D})/d}}{2\pi e} e^{-2I(\boldsymbol{D};\boldsymbol{W})/d},
\end{aligned}$$

where the last equality is a consequence of Eq. (4). □

B.2 Proof of Lemma 1

Proof. Firstly, as $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then $\mathbb{E}[\boldsymbol{X}+\boldsymbol{Y}] = \boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{\mu}_{\boldsymbol{Y}}$ and $Cov(\boldsymbol{X}+\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{X}} + \boldsymbol{\Sigma}_{\boldsymbol{Y}}$.

Secondly, with a specific mean vector and covariance matrix, the maximum entropy distribution is Gaussian, which implies that $h(\boldsymbol{X}+\boldsymbol{Y}) \leq h(\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\Sigma}_{\boldsymbol{X}} + \boldsymbol{\Sigma}_{\boldsymbol{Y}}))$.

Finally, if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_X})$, then $\boldsymbol{X}+\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu_X}+\boldsymbol{\mu_Y}, \boldsymbol{\Sigma_X}+\boldsymbol{\Sigma_Y})$, which concludes the proof. □

B.3 Proof of Theorem 2

Proof. According to the relationship between mutual information and differential entropy, we have

$$\begin{aligned} & I(\boldsymbol{D}, \widetilde{\boldsymbol{W}}_o^{(t)} \mid \boldsymbol{W}_i^{(t)}) \\ &= h(\widetilde{\boldsymbol{W}}_o^{(t)} \mid \boldsymbol{W}_i^{(t)}) - h(\widetilde{\boldsymbol{W}}_o^{(t)} \mid \boldsymbol{D}, \boldsymbol{W}_i^{(t)}) \\ &= h(\widetilde{\boldsymbol{W}}_o^{(t)} \mid \boldsymbol{W}_i^{(t)}) - h(\mathcal{N}(\boldsymbol{0}, \sigma \cdot \mathbf{I})) \\ &\leq h(\mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} + \sigma \cdot \mathbf{I})) - h(\mathcal{N}(\boldsymbol{0}, \sigma \cdot \mathbf{I})), \end{aligned} \tag{23}$$

where the second equality depends on the fact that $\boldsymbol{W}_o^{(t)}$ is a constant when $\boldsymbol{W}_i^{(t)}$ and $\boldsymbol{D}$ are both observed. The last inequality is an immediate consequence of Lemma 1 with $\widetilde{\boldsymbol{W}}_o^{(t)} = \boldsymbol{W}_o^{(t)} + \sqrt{\sigma} \cdot \boldsymbol{\xi}$.

Then with the differential entropy of Gaussian distribution, we can further transform Eq. (23) to

$$\begin{aligned} & h(\mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} + \sigma \cdot \mathbf{I})) - h(\mathcal{N}(\boldsymbol{0}, \sigma \cdot \mathbf{I})) \\ =& \frac{1}{2} \ln (2\pi e)^d \det(\boldsymbol{\Sigma}^{(t)} + \sigma \cdot \mathbf{I}) - \frac{1}{2} \ln (2\pi e)^d \sigma^d, \end{aligned} \tag{24}$$

where $\det(\cdot)$ denotes the determinant of a matrix and the second term of Eq. (24) is decided by $\det(\sigma \cdot \mathbf{I}) = \sigma^d$.

Next, we focus on the first term of Eq. (24). Note that $\boldsymbol{\Sigma}^{(t)}$ is the covariance matrix of $\boldsymbol{W}_o^{(t)}$, hence it's a real symmetric matrix, then according to eigen decomposition, we have

$$\boldsymbol{\Sigma}^{(t)} = \boldsymbol{Q}^{(t)} \boldsymbol{\Lambda}^{(t)} \boldsymbol{Q}^{(t)\mathrm{T}} = \boldsymbol{Q}^{(t)} \mathrm{diag}(\lambda_1^{(t)}, \cdots, \lambda_d^{(t)}) \boldsymbol{Q}^{(t)\mathrm{T}}, \tag{25}$$

where $\boldsymbol{Q}^{(t)}$ is an orthogonal matrix, i.e., $\boldsymbol{Q}^{(t)} \boldsymbol{Q}^{(t)\mathrm{T}} = \mathbf{I}$. Hence, $\sigma \cdot \mathbf{I} = \boldsymbol{Q}^{(t)} (\sigma \mathbf{I}) \boldsymbol{Q}^{(t)\mathrm{T}}$. Then we have

$$\begin{aligned} & \det(\boldsymbol{\Sigma}^{(t)} + \sigma \cdot \mathbf{I}) \\ =& \det(\boldsymbol{Q}^{(t)} \mathrm{diag}(\lambda_1^{(t)}, \cdots, \lambda_d^{(t)}) \boldsymbol{Q}^{(t)\mathrm{T}} + \boldsymbol{Q}^{(t)} (\sigma \mathbf{I}) \boldsymbol{Q}^{(t)\mathrm{T}}) \\ =& \det(\mathrm{diag}(\lambda_1^{(t)}, \cdots, \lambda_d^{(t)}) + \sigma \mathbf{I}) = \prod_{i=1}^{d} (\lambda_i^{(t)} + \sigma). \end{aligned}$$

Then we have $I(\boldsymbol{D}, \widetilde{\boldsymbol{W}}_o^{(t)} \mid \boldsymbol{W}_i^{(t)}) \leq \frac{1}{2} \ln \frac{\prod_{i=1}^{d} (\lambda_i^{(t)} + \sigma)}{\sigma^d}$, which immediately completes the proof. □

B.4 Proof of Theorem 3

Before the proof of Thm. 3, we have the following lemma.

Lemma 3. *For any d-dimensional semi-positive definite matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, we have*

(1) $\boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{A}+\boldsymbol{B})^{-1} \boldsymbol{\alpha} \leq \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{A}^{-1} \boldsymbol{\alpha}$,
(2) $\boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{A}+\boldsymbol{B})^{-1} \boldsymbol{\alpha} \leq \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{\alpha}$,

where $\boldsymbol{\alpha}$ is a d-dimensional vector.

Proof. Here we only prove inequality (1), and the proof for inequality (2) is the same as inequality (1). First, we have

$$(\boldsymbol{A}+\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \boldsymbol{A}^{-1}, \tag{26}$$

hence, we can get following result.

$$\boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{A}+\boldsymbol{B})^{-1} \boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \boldsymbol{A}^{-1}) \boldsymbol{\alpha}.$$

Then we need to prove

$$\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{A}^{-1} (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \boldsymbol{A}^{-1} \boldsymbol{\alpha} \geq 0. \tag{27}$$

As $\boldsymbol{A}^{-1}$ is a symmetric matrix, we can rewrite Eq. (27) as

$$\tilde{\boldsymbol{\alpha}}^{\mathrm{T}} (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \tilde{\boldsymbol{\alpha}} \geq 0, \tag{28}$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{A}^{-1} \boldsymbol{\alpha}$. As $\boldsymbol{A}$ and $\boldsymbol{B}$ are semi-positive definite, we have $(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}$ is a semi-positive definite matrix either, hence inequality (28) holds, which concludes the proof. □

Then with Lemma 3, we can prove Thm. 3.

Proof. As we can rewrite $\boldsymbol{\Sigma}_i$, $i \in \{0, \cdots, d-1\}$, as

$$\begin{aligned} \boldsymbol{\Sigma}_i &= \begin{pmatrix} \boldsymbol{\Sigma}_{i-1} & \boldsymbol{\rho}_i \\ \boldsymbol{\rho}_i^{\mathrm{T}} & c_{i,i} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{I}_{i-1} & \boldsymbol{0} \\ \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{i-1} & \boldsymbol{\rho}_i \\ \boldsymbol{0} & c_{i,i} - \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i \end{pmatrix}, \end{aligned} \tag{29}$$

Hence, the determinant is

$$\det \boldsymbol{\Sigma}_i = (c_{i,i} - \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i) \det \boldsymbol{\Sigma}_{i-1}. \tag{30}$$

Based on this result, with substitution $\boldsymbol{\Sigma}_i + \boldsymbol{\beta}_i$ for $\boldsymbol{\Sigma}_i$, we have

$$\det(\boldsymbol{\Sigma}_i + \boldsymbol{\beta}_i) = v_i \cdot \det(\boldsymbol{\Sigma}_{i-1} + \mathrm{diag}(\boldsymbol{\beta}_{i-1})), \tag{31}$$

$$v_i = c_{i,i} + \beta_i - \boldsymbol{\rho}_i^{\mathrm{T}} (\boldsymbol{\Sigma}_{i-1} + \mathrm{diag}(\boldsymbol{\beta}_{i-1}))^{-1} \boldsymbol{\rho}_i, \tag{32}$$

where $\boldsymbol{\beta}_i = (\beta_0, \cdots, \beta_{i-1})$. Then as $(\boldsymbol{\Sigma}_{i-1} + \mathrm{diag}(\boldsymbol{\beta}_{i-1}))^{-1}$ is a semi-positive definite matrix, we have $v_i \leq c_{i,i} + \beta_i$.

Furthermore, we can rewrite v_i as

$$\begin{aligned} v_i = \; & c_{i,i} + \beta_i - \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i \\ & + \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} (\boldsymbol{\Sigma}_{i-1}^{-1} + \mathrm{diag}(\boldsymbol{\beta}_{i-1})^{-1})^{-1} \boldsymbol{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i \\ \leq \; & c_{i,i} + \beta_i - \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i + \boldsymbol{\rho}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i-1}^{-1} \mathrm{diag}(\boldsymbol{\beta}_{i-1}) \boldsymbol{\Sigma}_{i-1}^{-1} \boldsymbol{\rho}_i, \end{aligned}$$

where the first equality is based on Eq. (26), and the last inequality depends on Lemma 3. Finally, according to Eq. (30), we have $\ln \det \boldsymbol{\Sigma}_{d-1} = \sum_{i=0}^{d-1} \ln v_i$ by induction. Combining above results, we can immediately conclude the proof. □

B.5 Proof of Theorem 4

Proof. For the first stage, if we denote the threshold of gradient clipping as S and denote the added noise as $\sigma \cdot \mathbf{I}$, we get $\|\boldsymbol{g}^{(t)}\|_2 \leq S$, where $\boldsymbol{g}^{(t)} = \nabla_{\boldsymbol{W}} F(\boldsymbol{W}_i^{(t)}; \boldsymbol{D})$ is a random vector as a function of $\boldsymbol{D}$. Hence, we have

$$\mathrm{tr}(\boldsymbol{\Sigma}^{(t)}) \leq \mathrm{tr}(\mathbb{E}(\boldsymbol{g}^{(t)} (\boldsymbol{g}^{(t)})^T)) = \mathbb{E}(\|\boldsymbol{g}^{(t)}\|_2^2) \leq S^2, \tag{33}$$

where the first inequality depends on the relationship between covariance matrix and the auto-correlation matrix, i.e., $\mathrm{tr}[\mathbb{E}(\boldsymbol{g}^{(t)} (\boldsymbol{g}^{(t)})^T) - \boldsymbol{\Sigma}^{(t)}] = \mathrm{tr}[\mathbb{E}(\boldsymbol{g}^{(t)}) \mathbb{E}(\boldsymbol{g}^{(t)})^T)] \geq 0$.

Then we have $\sum_{i=1}^{d}\lambda_i^{(t)} = \mathrm{tr}(\mathbf{\Sigma}^{(t)}) \le S^2$. Based on AM-GM inequality and the fact $\lambda_i^{(t)} \ge 0$, the channel capacity derived by Eq. (15) has an upper bound as

$$f^{(t)}(\sigma) \le \hat{f}^{(t)}(\sigma) = d \cdot \ln \frac{\sigma + \frac{\sum_{i=1}^{d}\lambda_i^{(t)}}{d}}{\sigma} \le d \cdot \ln \frac{\sigma + S^2/d}{\sigma}.$$

Therefore, if we treat σ as a constant, and denote $u_\sigma(d) := d \cdot \ln \frac{\sigma+S^2/d}{\sigma}, d \ge 0$, we have $f^{(t)}(\sigma) \le u_\sigma(d)$. We observe that $u_\sigma^{'}(d) = \ln \frac{\sigma+S^2/d}{\sigma} + \frac{\sigma}{\sigma+S^2/d} - 1$. Hence

$$u''_\sigma(d) = -\frac{S^2(2\sigma d + S^2)}{d(\sigma d + S^2)^2} \le 0 \Rightarrow u_\sigma^{'}(d) \ge u_\sigma^{'}(+\infty) = 0,$$

which implies $u_\sigma(d)$ is an increasing function of the dimension d. Based on L'Hospital's rule, we can conclude that

$$C^{(t)} = f^{(t)}(\sigma) \le u_\sigma(d) \le u_\sigma(+\infty) = \frac{S^2}{\sigma}. \tag{34}$$

For the second stage, which provides a (ε, δ)-DP, the authors in [19] demonstrate that we need to choose $\sigma \ge S^2 \cdot \frac{2\log(1.25/\delta)}{\varepsilon^2}$, then with a substitution in Eq. (34), we have $C^{(t)} \le \frac{\varepsilon^2}{2\log(1.25/\delta)}$. Hence, the channel capacity of (ε, δ)-DP in the FL scenario is upper bounded by $\frac{\varepsilon^2}{2\log(1.25/\delta)}$.

Finally, as illustrated in [19], for mini-batch SGD in DP, i.e., the batch size $B > 1$, the mechanism becomes $\boldsymbol{g} = \frac{1}{B}(\sum_{i=1}^{B}\bar{\boldsymbol{g}}_i + \boldsymbol{\xi})$, let $\tilde{\boldsymbol{g}} = \frac{1}{B}\sum_{i=1}^{B}\bar{\boldsymbol{g}}_i$ and $\tilde{\boldsymbol{\xi}} = \frac{1}{B}\boldsymbol{\xi}$, we have $\mathbf{\Sigma}_{\tilde{\boldsymbol{g}}} = \frac{1}{B}\mathbf{\Sigma}_{\boldsymbol{g}_i}$ and $\mathbf{\Sigma}_{\tilde{\boldsymbol{\xi}}} = \frac{1}{B^2}\mathbf{\Sigma}_{\boldsymbol{\xi}}$. By substitution them into Eq. (33), we can immediately conclude the proof. □

C Details of the Prior Knowledge

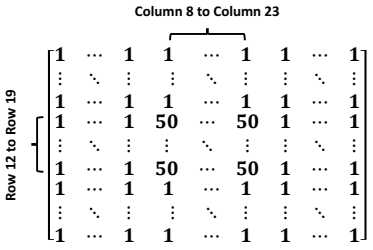

Figure 17: The prior knowledge $\boldsymbol{\beta}$ used for the Personalized Channel in Sec. 6. Specifically, the number in the matrix represents the coefficient for the added noise.

For the Personalized Channel, the prior knowledge $\boldsymbol{\beta}$ used in Sec. 6 is a $32 * 32 * c$ tensor, where c represents the number of image channel (e.g., $c = 1$ for Fashion-MNIST dataset and $c = 3$ for CIFAR-10 dataset, respectively). The number in $\boldsymbol{\beta}$ represents the coefficient for the added noise, which means we add more noise to the dimension if the coefficient of this dimension is larger. In these experiments, we employ position-based prior knowledge, which means the prior knowledge for all image channels are identical. Therefore, the important parameters are the numbers of one channel, which is a $32 * 32$ matrix (as displayed in Fig. 17).

Specifically, as we aim to protect the private information around eyes for the data in CelebA (as displayed in Fig. 5), we set the number of these positions (i.e., the sub-matrix with row 12 to 19 and column 8 to 23.) to be 50, and set the rest number of the matrix to be 1. Therefore, this prior knowledge adds 50 times noise to the chosen positions (i.e., the positions around the eyes) to provide stronger privacy protection for this area. Additionally, we also use this prior knowledge for other datasets, including Fashion-MNIST, CIFAR10, and SVHN.

D Validating the theories of existing methods

The impact of batch size for DP and utilizing large batch size in defending against DRA. To improve the existing defense algorithm for defending against DRA, we design experiments to validate the roles of B by gradient inversion attack on CIFAR-10. Specifically, to keep identical reconstruction difficulties for different B, we only reconstruct one image from the gradient as the mean restoration $\mathbb{E}[\hat{\boldsymbol{D}}]$. Then we calculate the MSE between $\mathbb{E}[\hat{\boldsymbol{D}}]$ and the mean value of local dataset $\mathbb{E}[\boldsymbol{D}]$, i.e., $\|\mathbb{E}[\hat{\boldsymbol{D}}] - \mathbb{E}[\boldsymbol{D}]\|^2$. Due to the convexity of MSE, the difference between $\hat{\boldsymbol{D}}$ and $\boldsymbol{D}$ is lower bounded, i.e., $\mathbb{E}\|\hat{\boldsymbol{D}} - \boldsymbol{D}\|^2 \ge \|\mathbb{E}[\hat{\boldsymbol{D}}] - \mathbb{E}[\boldsymbol{D}]\|^2$. Moreover, we utilize clipping bound $S = 1.0$ and the noise multiplier $\sigma = 1.3$ for the DP training. The results are displayed in Fig. 6. If we utilize DP for privacy protection, the performance for defending against DRA consistently decreases with an increasing B. This phenomenon is consistent with Thm. 4, indicating that the small B is beneficial for DP in defending against DRA.

Additionally, when we do not apply any defense technique for privacy, the defense ability increases according to the increasing B. This phenomenon is consistent with Eq. (22) due to the intrinsic noise in the collected data. Moreover, the performance improvement becomes more significant when we add a constant noise to the parameters. The conclusion is that when we add a constant noise to the parameter, a larger B leads to a stronger ability to defend against DRA.

The impact of eigenvalues in compression. We also validate our theories for compression on CIFAR-10 datasets. In these experiments, we utilize the gradient inversion attack to reconstruct data from the gradients. For the compression, we set the eigenvector to be $\mathbf{0}$ to reduce the information in the corresponding dimension. Then we map the training data to the compressed eigenspace. For comparison, we only compress one dimension according to the eigenvalue. The results are displayed in Fig. 8. Specifically, the eigenvalues $\{\lambda_{max}, \lambda_1, \lambda_{10}\}$ are as follows: $\{222.21, 85.08, 9.91\}$ for CIFAR10, $\{346.21, 71.01, 12.82\}$ for CelebA, $\{313.0, 28.04, 4.67\}$ for SVHN, and $\{101.88, 61.80, 3.21\}$ for Fashion-MNIST. The experiment results demonstrate that the compression on the dimension with a larger eigenvalue leads to a stronger defensive ability to defend against DRA, which is consistent with our theories.