A Method

A.1 Other Efficient Evaluation Techniques

The techniques for efficiently evaluating implicit gradients can be referred to (Koh and Liang, 2017; Grazzi et al., 2020). As computing the inverse second-order derivatives is the most computation-intensive operation, we will focus on it. Here, we briefly summarize two supplementary techniques introduced in Section 3.1.

Conjucate gradient. In Section 3.1, we use the trick of least square to compute the $(JJ^\top)^{-1}J\delta$. When $JJ^\top \succ 0$, we can solve the least square problem by the conjugate gradient (CG) method, which only needs $\mathcal{O}(d)$ time to converge. However, the algorithm will be unstable when the matrix $JJ^\top$ is ill-conditioned. As observed in Fig. 5, the $JJ^\top$ is likely to be ill-conditioned for deep networks. But here we provide this alternative for linear models such that J can be evaluated faster.

Neumann series. We can leverage the Neumann series to compute the matrix inverse. By the Neumann series, we have $(JJ^\top)^{-1}J\delta = \lim_{t\to\infty}\sum_{i=0}^{t}(I - JJ^\top)^i J\delta$. Let $s_t \triangleq \sum_{i=0}^{t}(I - JJ^\top)^i J\delta$ and $s_0 \triangleq J\delta$. Then the computation can be done by iteration $s_{t+1} = (I - JJ^\top)s_t + J\delta$ which only includes Jacobian-vector products.

B Proofs

B.1 Proof of the Approximation by Implicit Gradients

Here, we provide the proof for $\frac{\partial G_r(g_0)}{\partial g_0} = (JJ^\top)^{-1}J$. Recall Eq. (1) as

$$x^* = G_r(g) = \arg\min_x L_I(x;g) \triangleq \left\|\nabla_\theta L(x,\theta) - g\right\|^2. \tag{8}$$

The stationary condition of the minimization gives

$$\frac{\partial L_I(x^*;g)}{\partial x^*} = 0.$$

Given a small perturbation $\Delta_g \to 0$ on the gradient, we can estimate corresponding perturbation $\Delta_{x^*} \to 0$ as a function of Δ_g. Thus, we can approximate $\frac{\partial x^*}{\partial g}$ by $\frac{\partial \Delta_{x^*}}{\partial \Delta_g}$. Use Taylor expansion to show approximately

$$\begin{aligned}
\frac{\partial L_I(x^*;g)}{\partial x^*} + \frac{\partial^2 L_I(x^*;g)}{\partial g \partial x^*}\Delta_g + \frac{\partial^2 L_I(x^*;g)}{\partial {x^*}^2}\Delta_{x^*} &\approx 0 \\
\Delta_{x^*} &\approx -\left(\frac{\partial^2 L_I(x^*;g)}{\partial {x^*}^2}\right)^{-1}\frac{\partial L_I(x^*;g)}{\partial x^*}\Delta_g \\
\frac{\partial x^*_g}{\partial g} &\approx -\left(\frac{\partial^2 L_I(x^*;g)}{\partial {x^*}^2}\right)^{-1}\frac{\partial^2 L_I(x^*;g)}{\partial g \partial x^*}
\end{aligned} \tag{9}$$

where we drop higher-order perturbations. The above derivations can be rigorously proved using the Implicit Function Theorem. Since $\frac{\partial L_I(x^*;g)}{\partial x^*} = 2(\nabla_\theta L(x^*,\theta) - g)\nabla_x \nabla_\theta L(x^*,\theta)$, we can derive

$$\frac{\partial^2 L_I(x^*;g)}{\partial g \partial x^*} = -2\nabla_x \nabla_\theta L(x^*,\theta)$$

and

$$\frac{\partial^2 L_I(x^*;g)}{\partial {x^*}^2} = 2(\nabla_\theta L(x^*,\theta) - g)\nabla_x^2 \nabla_\theta L(x^*,\theta) + 2\nabla_x \nabla_\theta L(x^*,\theta)(\nabla_x \nabla_\theta L(x^*,\theta))^\top.$$

As $x_0 = x^* = G_r(g_0)$ and $g_0 = \nabla_\theta L(x^*,\theta)$, we can substitute them to obtain

$$\frac{\partial^2 L_I(x_0;g)}{\partial g \partial x_0} = -2\nabla_x \nabla_\theta L(x_0,\theta) := -2J(x^*(g_0),\theta), \tag{10}$$

$$\frac{\partial^2 L_I(x_0;g)}{\partial {x_0}^2} = 2(g_0 - g)\nabla_x^2 \nabla_\theta L(x_0,\theta) + 2JJ^\top. \tag{11}$$

Let $g = g_0$. Combine Eqs. (9) to (11) to get

$$\frac{\partial G_r(g_0)}{\partial g_0} = (JJ^\top)^{-1}J.$$

B.2 Proof of Theorem 3.1

Before we prove our main theorem, we prove several essential lemmas as below.

Lemma B.1. $\left\|\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - \nabla_x \|\nabla_\theta L(x',\theta)\|^2\right\| \leq (\mu_L \|\nabla_x \nabla_\theta L(x,\theta)\| + \mu_J \|\nabla_\theta L(x',\theta)\|) \|x - x'\|$

Proof.

$$\begin{aligned}
&\left\|\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - \nabla_x \|\nabla_\theta L(x',\theta)\|^2\right\| \\
&= \|\nabla_x \nabla_\theta L(x,\theta)\nabla_\theta L(x,\theta) - \nabla_x \nabla_\theta L(x,\theta)\nabla_\theta L(x',\theta) \\
&\quad + \nabla_x \nabla_\theta L(x,\theta)\nabla_\theta L(x',\theta) - \nabla_x \nabla_\theta L(x',\theta)\nabla_\theta L(x',\theta)\| \\
&\leq \|\nabla_x \nabla_\theta L(x,\theta)\| \|\nabla_\theta L(x,\theta) - \nabla_\theta L(x',\theta)\| + \|\nabla_x \nabla_\theta L(x',\theta) - \nabla_x \nabla_\theta L(x,\theta)\| \|\nabla_\theta L(x',\theta)\|.
\end{aligned}$$

Using Assumption 3.4 and 3.5 directly lead to

$$\left\|\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - \nabla_x \|\nabla_\theta L(x',\theta)\|^2\right\| \leq (\mu_L \|\nabla_x \nabla_\theta L(x,\theta)\| + \mu_J \|\nabla_\theta L(x',\theta)\|) \|x - x'\|.$$

□

Lemma B.2. $\left\|\nabla_x(\nabla_\theta^\top L(x,\theta)g) - \nabla_x(\nabla_\theta^\top L(x',\theta)g)\right\| \leq \mu_J \|g\| \|x - x'\|$

Proof. By Assumption 3.4, we have

$$\begin{aligned}
&\left\|\nabla_x(\nabla_\theta^\top L(x,\theta)g) - \nabla_x(\nabla_\theta^\top L(x',\theta)g)\right\| \\
&\leq \|\nabla_x \nabla_\theta L(x,\theta) - \nabla_x \nabla_\theta L(x',\theta)\| \|g\| \\
&\leq \mu_J \|g\| \|x - x'\|.
\end{aligned}$$

□

Lemma B.3. *The inversion loss $L_I(x;g)$ defined satisfies $\|\nabla_x L_I(x;g) - \nabla_x L_I(x';g)\| \leq \mu \|x - x'\|$ where*

$$\mu(x,x',\theta,g) = \mu_L \|\nabla_x \nabla_\theta L(x,\theta)\| + \mu_J \|\nabla_\theta L(x',\theta)\| + \mu_J \|g\|. \tag{12}$$

Proof. Since

$$\begin{aligned}
\nabla_x L_I(x;g) &= 2\nabla_x \nabla_\theta L(x,\theta)(\nabla_\theta L(x,\theta) - g) \\
&= 2\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - 2\nabla_x \nabla_\theta L(x,\theta)g,
\end{aligned}$$

we can derive

$$\begin{aligned}
&\|\nabla_x L_I(x;g) - \nabla_x L_I(x';g)\| \\
&= 2\left\|\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - \nabla_x \|\nabla_\theta L(x',\theta)\|^2 - \left[\nabla_x \nabla_\theta^\top L(x,\theta) - \nabla_x \nabla_\theta^\top L(x',\theta)\right] g\right\| \\
&\leq 2\left\|\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - \nabla_x \|\nabla_\theta L(x',\theta)\|^2\right\| + 2\left\|\left[\nabla_x \nabla_\theta^\top L(x,\theta) - \nabla_x \nabla_\theta^\top L(x',\theta)\right] g\right\|.
\end{aligned}$$

By Lemma B.1 and Lemma B.2, we have

$$\begin{aligned}
\left\|\nabla_x \|\nabla_\theta L(x,\theta)\|^2 - \nabla_x \|\nabla_\theta L(x',\theta)\|^2\right\| &\leq \mu_1 \|x - x'\|, \\
\left\|\nabla_x(\nabla_\theta^\top L(x,\theta)g) - \nabla_x(\nabla_\theta^\top L(x',\theta)g)\right\| &\leq \mu_2 \|x - x'\|.
\end{aligned}$$

where $\mu_1 = \mu_L \|\nabla_x \nabla_\theta L(x,\theta)\| + \mu_J \|\nabla_\theta L(x',\theta)\|$ and $\mu_2 = \mu_J \|g\|$. Let $\mu = 2\mu_1 + 2\mu_2$. Then we can get

$$\|\nabla_x L_I(x;g) - \nabla_x L_I(x';g)\| \leq \mu \|x - x'\|.$$

□

Theorem B.1 (Restated from Theorem 3.1)**.** *Let x_0 be the private data and $g_0 \triangleq \nabla_\theta L(x_0, \theta)$ be its corresponding gradient which is treated as a constant. If Assumption 3.4 and 3.5 hold, then the square root of the recovery RMSE satisfies:*

$$\|x_0 - G_r(g_0 + \delta)\| \geq \frac{\|J\delta\|}{\mu_L \|J\| + 2\mu_J \|g_0 + \delta\|}, \tag{13}$$

where $J = \nabla_x \nabla_\theta L(x_0, \theta)$.

Proof. Utilize the stationary condition $\nabla_x L_I(x_g^*; g) = 0$ and Lemma B.3 to obtain

$$\|\nabla_x L_I(x; g)\| \leq \mu(x, x_g^*, \theta, g) \left\|x - x_g^*\right\|, \ \forall x.$$

As x_0 is the private sample whose gradient is $g_0 \triangleq \nabla_\theta L(x_0, \theta)$, then we have

$$\|x_0 - G_r(g_0 + \delta)\| \geq \frac{1}{\mu(x_0, x^*_{g_0+\delta}, \theta, g_0 + \delta)} \|\nabla_x L_I(x_0; g_0 + \delta)\|$$

Because

$$\begin{aligned} \nabla_x L_I(x_0; g_0 + \delta) &= 2\nabla_x \nabla_\theta L(x_0, \theta)(\nabla_\theta L(x_0, \theta) - g_0 - \delta) \\ &= 2\nabla_x \nabla_\theta L(x_0, \theta)\delta, \end{aligned}$$

we can attain

$$\|x_0 - G_r(g_0 + \delta)\| \geq \frac{2}{\mu} \|\nabla_x \nabla_\theta L(x_0, \theta)\delta\|.$$

With $\nabla_\theta L(x^*_{g_0+\delta}, \theta) = g_0 + \delta$, we can obtain

$$\begin{aligned} \mu(x_0, x^*_{g_0+\delta}, \theta, g_0 + \delta) &= 2\mu_L \|\nabla_x \nabla_\theta L(x_0, \theta)\| + 2\mu_J \left\|\nabla_\theta L(x^*_{g_0+\delta}, \theta)\right\| + 2\mu_J \|g_0 + \delta\| \\ &= 2\mu_L \|J\| + 4\mu_J \|g_0 + \delta\|. \end{aligned}$$

□

C Experiments

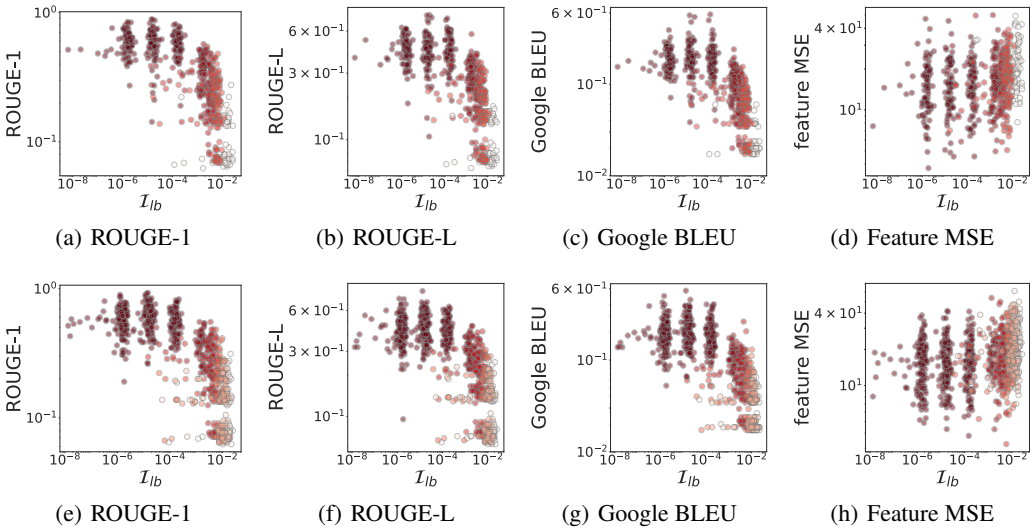
C.1 Experimental Details

Model architectures. The linear model we use is a matrix that maps the input data into a vector. The LeNet model is a convolutional neural network with 4 convolutional layers and 1 fully connected layer. We use the modified version following previous privacy papers Sun et al. (2020), whose detailed structure is in Table 1. ResNet18 is a popular deep convolutional network in computer vision with batch-normalization and residual layers (He et al., 2015a). Cross entropy loss is used as the loss function in all the experiments.

Experimental settings. We conduct two kinds of attacks in our paper: DGL and GS attacks. The learning rate of the two attacks is 0.1 and we use Adam as the optimizer. To consider a more powerful attack, only a single image is reconstructed in each inversion. When inverting LeNet, we uniformly initialize the model parameters in the range of $[-0.5, 0.5]$ as (Sun et al., 2020) to get a stronger attack. When inverting ResNet18, we use the default initialization method in PyTorch and follow Huang et al. (2021) to use BN statistics as an additional regularization term to conduct a stronger attack.

Table 1: A modified version of LeNet. Conv represents a convolutional layer. FC means a fully-connected layer.

Layers
Conv(in_channels=3, out_channels=12, kernel_size=5)
Conv(in_channels=12, out_channels=12, kernel_size=5)
Conv(in_channels=12, out_channels=12, kernel_size=5)
Conv(in_channels=12, out_channels=12, kernel_size=5)
Flattern
FC(out_features=10)



(a) ROUGE-1 (b) ROUGE-L (c) Google BLEU (d) Feature MSE

(e) ROUGE-1 (f) ROUGE-L (g) Google BLEU (h) Feature MSE

Figure 9: Evaluation of $\mathcal{I}_{lb}$ on BERT (top) and GPT-2 (bottom). A darker color means a larger noise variance. Four metrics are used to evaluate the semantic similarity between the original text and the recovered text. $\mathcal{I}_{lb}$ is linearly correlated to the four semantic metrics, which means $\mathcal{I}_{lb}$ can be used to estimate the privacy risk of the private text.

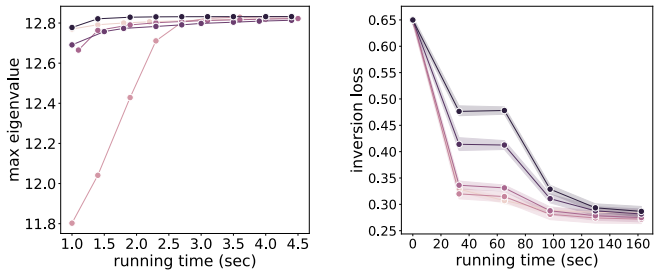
C.2 Empirical Validation on Language Data

We evaluate the proposed I^{2}F metric on BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019), which are popular language models in natural language processing. We use TAG (Deng et al., 2021), which is an attack on Transformer-based language models based on the L_1 and L_2 distance between the original gradient and the dummy gradient, and follow the code in `https://github.com/JonasGeiping/breaching`. We use the default setting in the code and iteratively update the input embedding 12,000 times. We randomly sample 70 sentences from WikiText-103 (Merity et al., 2016) as the private text. We use *ROUGE-1*, *ROUGE-L* (Lin, 2004), *Google BLEU* (Wu et al., 2016) and *feature MSE* to measure the semantic similarity between the original text and the recovered text. ROUGE-1 measures the overlap of 1-grams in the original and recovered text, while ROUGE-L measures the length of the longest common subsequence between two sentences. While ROUGE metrics calculate the reconstruction recall, the Google BLEU score uses as the output the smaller value of the precision and recall of the original and recovered text and has a broader range of the overlap of n-grams, where $n = 1, 2, 3, 4$. Since the above three metrics are discrete and not consistent with our assumptions, we include a continuous metric, the feature MSE, which measures the distance between the final layer's feature of the original and recovered text.

The results are presented in Fig. 9. A darker color indicates a larger noise variance. For each noise variance, we randomly sample the perturbation from a zero-mean Gaussian distribution 5 times and repeat this for 3 different random seeds. It shows that I^{2}F is correlated to these four metrics, which means I^{2}F can be used to estimate the privacy risk of text datasets with large language models. The correlation of BERT and GPT-2 between the four metrics has a similar mode. Though ROUGE-1, ROUGE-L and Google BLEU measure the structural similarity of sentences, which consists of discrete tokens, I^{2}F presents a clear correlation. For the feature MSE, although I^{2}F has a less distinguished correlation with feature MSE, it can still be utilized to estimate the privacy risk.

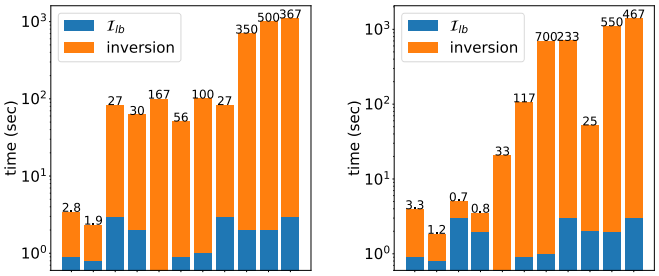
C.3 Efficiency

To show the efficiency of computing the I^{2}F values, we compare it with the GS attack on randomly-picked CIFAR10 images with ResNet18. As the major complexity comes from evaluating the maximal eigenvalue via the power iteration method, we compare the time required for the power iteration against the inversion loss of GS to converge. We notice that the convergence depends on the initialization of the power iteration and the learning rate for DGL. Thus, we repeat power iteration with



(a) Power iteration (b) Inversion attack

Figure 10: Evaluation of the efficiency of computing $\lambda_{\max}(JJ^\top)$ (our method) by power iteration and inversion attack by minimizing inversion loss (L_I). Colors in (a) indicate different seeds. Darker colors in (b) indicate larger learning rates. Our method using power iteration can converge faster than direct inversion attacks.

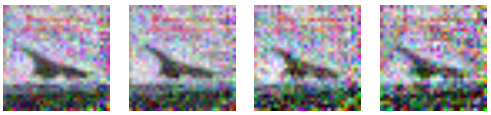


(a) DGL attack (b) GS attack

Figure 11: Comparison of the efficiency of computing $\mathcal{I}_{lb}$ (our method) by power iteration and **inversion** attack by minimizing inversion loss (L_I). Blue bars indicate the time of computing $\mathcal{I}_{lb}$ while orange bars indicate the time of minimizing inversion loss by DGL and GS. The time ratio of computing $\mathcal{I}_{lb}$ versus minimizing inversion loss is present above the orange bars. The x-axis are model-dataset pairs sorted by the model scales: MLP-MNIST, MLP-CIFAR10, LeNet-MNIST, LeNet-CIFAR10, RN18-MNIST, RN18-CIFAR10, RN34-CIFAR10, RN50-CIFAR10, RN101-CIFAR10, RN152-CIFAR10, RN152-ImageNet. For large models and datasets, where minimizing inversion loss needs a huge computation overhead, $\mathcal{I}_{lb}$ can provide an efficient estimation of the privacy risk.

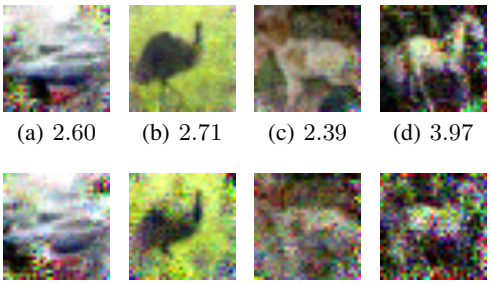
5 different seeds. For the inversion attack, we evaluate multiple learning rates $(1, 0.5, 0.1, 0.05, 0.01)$ to show the fastest convergence. Each experiment is repeated 5 times with different random seeds. As shown in Fig. 10, the power iteration method can converge to the maximal eigenvalue within 50 iterations (5 seconds at most). In comparison, the inversion loss demands 3000 more iterations in 150 seconds to fully converge, which is 20 times larger than the power iteration method. Thus, our method can give an accurate and fast approximation of the recovery MSE without the exhaustive whole inversion process.

In Fig. 11, we compare the computation cost of computing $\mathcal{I}_{lb}$ with minimizing inversion loss. We show that for almost all the models and datasets we evaluate, the time ratio is larger than 1, which means it is more efficient to compute $\mathcal{I}_{lb}$ than minimize the inversion loss. It indicates that our method is a more efficient way to estimate the privacy risk for most models and datasets, than the empirical inversion attack. Another key point is that, for large models and datasets, such as models larger than RN18 with CIFAR10 or ImageNet, the time ratio is even much larger than 500. When the time consumption of inversion attacks on these models and datasets is huge (about 16 minutes or even longer), our method significantly reduces the computation overheads for estimating the privacy risks.



(a) $\lambda = 3.90$ (b) $\lambda = 0.12$ (c) $\lambda = 0.03$ (d) $\lambda = 0.0006$

Figure 12: Same perturbation sizes but different protection effects by different eigenvectors of LeNet. Recovered CIFAR10 images associated with different eigenvectors are present. When perturbing with eigenvectors with smaller eigenvalues, the recovered images are more noisy.



(a) 2.60 (b) 2.71 (c) 2.39 (d) 3.97

(e) 0.15 (f) 0.28 (g) 0.32 (h) 0.33

Figure 13: Same perturbation sizes but different protection effects by different eigenvectors of ResNet18. Recovered CIFAR10 images associated with different eigenvalues are present. When perturbing with eigenvectors with smaller eigenvalues, the recovered images are more noisy and lack some semantic information.

C.4 More Visual Results

More images of unequal perturbations. We present in Fig. 12 the recovered CIFAR10 images when perturbing the gradient with eigenvectors with different eigenvalues. When perturbing with eigenvectors with smaller eigenvalues, the recovered images are more noisy, which is consistent with our former observation.

We also present the unequal perturbations with different eigenvectors of ResNet18 on the CIFAR10 dataset in Fig. 13. Even with the same perturbation scale, the eigenvectors with larger eigenvalues provide stronger protection, where the corresponding recovered images are more noisy and lose some semantic information.

More images of unfair privacy protection. We show more results of unfair privacy protection in Fig. 14. The images of digits 5 and 8 can still be recognized by their outlines, while images of digits 7 and 9 are unrecognizable noise.

We also show the unfair privacy protection of ResNet18 on the CIFAR10 dataset in Fig. 15. In this experiment, we also observe a large variance of recovery MSE among samples, indicating sample-wise unfairness. At the class level, we still can find gradients of a few classes to be easily inverted. For example, class 8 has most MSEs lower than the average value.

D I^{2}F with Gradient Pruning Defense

We present the relationship between the RMSE and $\mathcal{I}_{lb}$ in Fig. 16. The y-axis is RMSE and the x-axis is $\mathcal{I}_{lb}$. It shows $\mathcal{I}_{lb}$ can be used to estimate the worst-case privacy risk with gradient pruning defense.

E Comparison of I^{2}F with More Metrics

MSE is a pixel-wise distance that lacks semantic and structural information. To evaluate the effectiveness of I^{2}F on more metrics, we consider SSIM and LPIPS (Zhang et al., 2018) to measure the

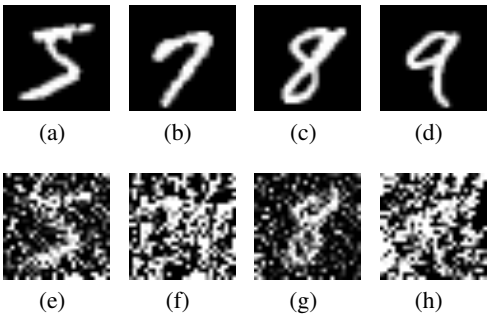


Figure 14: Original (top) and corresponding recovered (bottom) images of LeNet on the MNIST dataset. The gradients are perturbed with Gaussian noise of variance 10^{-3}. The defense is unfair as images of digit 5 and digit 8 can be recognized by the outline.

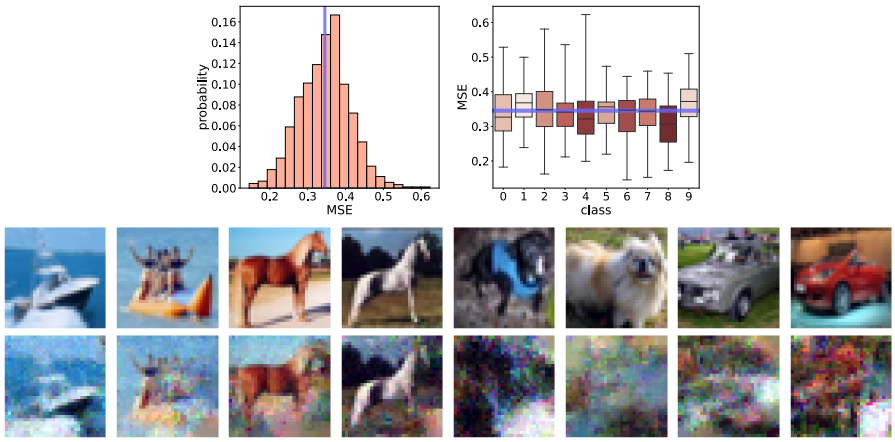
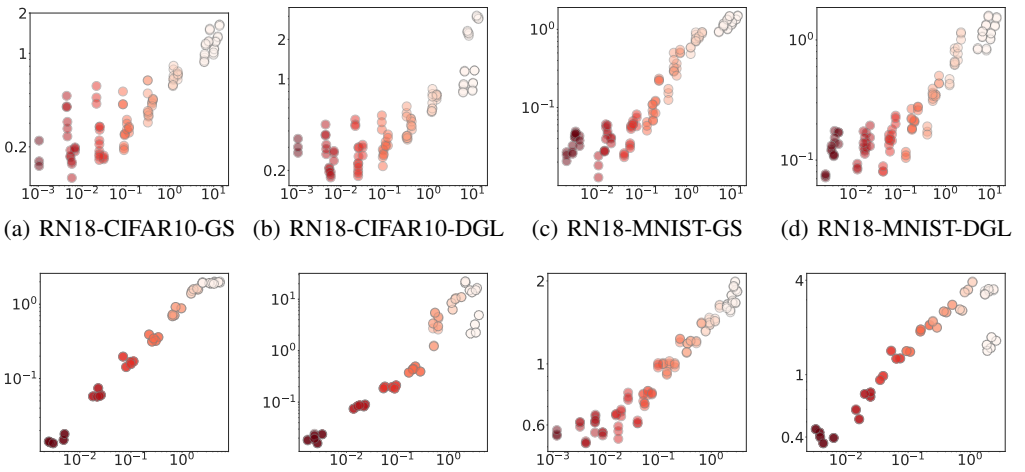


Figure 15: The sample-wise and class-wise statistics of the GS MSE on the CIFAR10 dataset of ResNet18. The purple lines indicate the average values. Large variances are observed among samples. The original (first row) and recovered (second row) images for the well- and poorly-protected classes are depicted at the bottom. The defense is unfair as some classes, e.g., class 7 (horse) and class 8 (ship), are more vulnerable to inversion attacks.

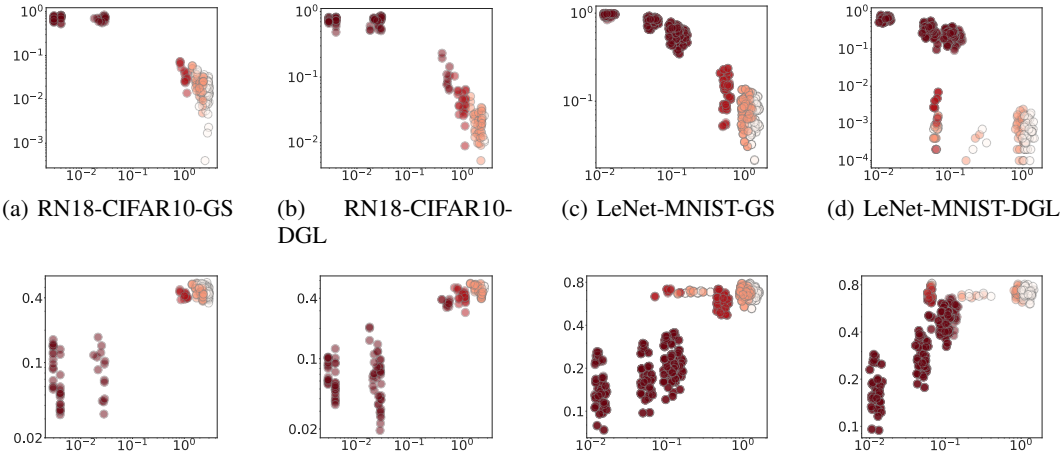


(a) RN18-CIFAR10-GS (b) RN18-CIFAR10-DGL (c) RN18-MNIST-GS (d) RN18-MNIST-DGL

(e) LeNet-MNIST-GS (f) LeNet-MNIST-DGL (g) LeNet-CIFAR10-GS (h) LeNet-CIFAR10-DGL

Figure 16: RMSE (y-axis) vs. $\mathcal{I}_{lb}$ (x-axis) with gradient pruning. A darker color indicates a smaller pruning ratio. It shows that $\mathcal{I}_{lb}$ is a good estimator of RMSE.

structural similarity and semantic distance between the original and recovered images, respectively. The relationship between SSIM and LPIPS is shown in Fig. 17. Since $\mathcal{I}_{lb}$ aims to lower bound the privacy risk in terms of RMSE, $\mathcal{I}_{lb}$ does not have the same scale as SSIM and LPIPS. Nevertheless, $\mathcal{I}_{lb}$ also has a positive correlation between SSIM and LPIPS, which implies that it is a good estimator for the structural similarity and semantic distance between the original and recovered images.



(a) RN18-CIFAR10-GS (b) RN18-CIFAR10-DGL (c) LeNet-MNIST-GS (d) LeNet-MNIST-DGL

(e) RN18-CIFAR10-GS (f) RN18-CIFAR10-DGL (g) LeNet-MNIST-GS (h) LeNet-MNIST-DGL

Figure 17: $\mathcal{I}_{lb}$ is positively correlated with these two metrics and is a good estimator for the structural similarity and semantic distance between the original and recovered images. Darker color indicates higher variance. **Top ((a)-(d)):** SSIM (y-axis) vs. $\mathcal{I}_{lb}$ (x-axis). **Bottom ((e)-(h)):** LPIPS (y-axis) vs. $\mathcal{I}_{lb}$ (x-axis). A higher SSIM and a lower LPIPS indicate a higher privacy risk.

F Dynamics of $\mathcal{I}_{lb}$ During Training

Previous existing empirical results show that privacy risk decreases by training epochs (Balunović et al., 2022; Geiping et al., 2020). We evaluate the dynamics of $\mathcal{I}_{lb}$, RMSE with DGL attack and RMSE during training in Fig. 18. It shows that as the training epoch increases, the $\mathcal{I}_{lb}$ also has an increasing trend. While the RMSE of RN18 on the CIFAR10 dataset has a similar trend as $\mathcal{I}_{lb}$, that of LeNet on the MNIST dataset has a significant rise at the epoch 60, which is due to the slower learning speed of LeNet than RN18. Moreover, almost for all the epochs, there is a sample with low $\mathcal{I}_{lb}$, which again emphasizes the unfairness in privacy protection.

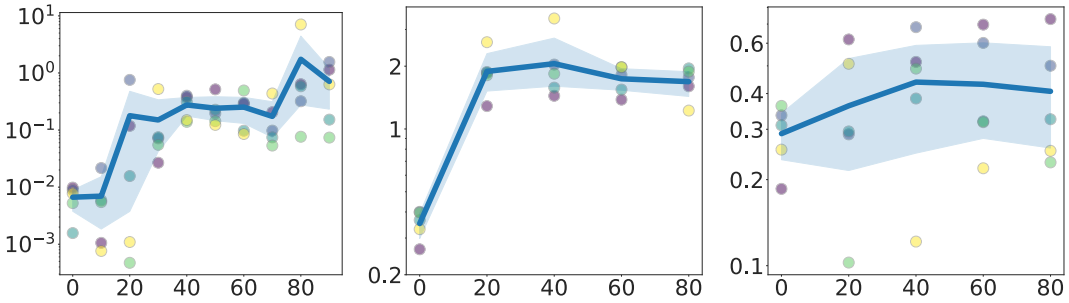
G The Impact of ϵ on Efficient Matrix Inversion

In Fig. 19, we study the impact of ϵ on efficient matrix inversion proposed in Section 4. We evaluate the impact on the LeNet with the MNIST dataset. The y-axis is the RMSE. $\mathcal{I}$ (matrix inversion) is calculated as defined in Eq. (4). $\mathcal{I}_{lb}$ (matrix norm) is calculated as defined in Eq. (5). It shows with $\epsilon \in [1, 10]$, $\mathcal{I}_{lb}$ is a lower bound of the RMSE. It also shows that $\mathcal{I}_{lb}$ is an accurate estimator of I^2F. Thus, we can directly use $\mathcal{I}_{lb}$ in practice to lower bound the privacy risk to avoid fine-tuning the ϵ.

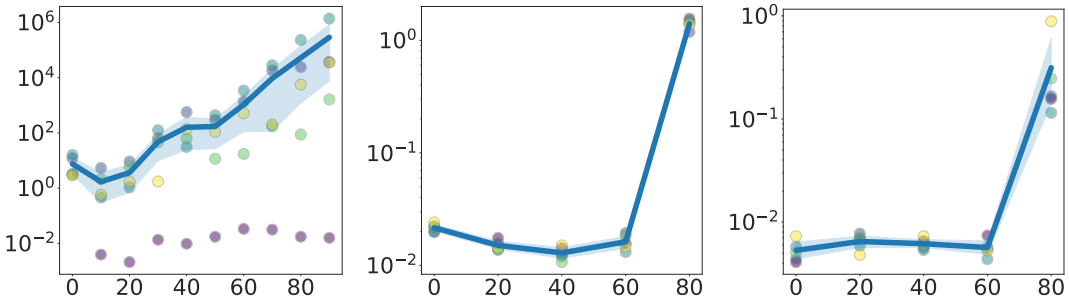
H Discussion

H.1 Validity of Assumption 3.4-3.5

We make assumptions about the Lipschitz continuous Jacobian and gradient in Assumption 3.4 and Assumption 3.5, respectively. These two assumptions are not necessary for I^2F but are only used to provide a theoretical validation of I^2F when the noise δ is not infinitesimal. To discuss the validity of these two assumptions in practice, we calculate the value of μ_J and μ_L of LeNet with two datasets.



(a) RN18 on CIFAR10



(b) LeNet on MNIST

Figure 18: Privacy risks decrease by training epochs (x-axis). Different colors indicate different samples. The y-axis from left to right: $\mathcal{I}_{lb}$, RMSE w/ DGL attack, RMSE w/ GS attack whose smaller values indicate lower risks. The blue line indicates the mean value and the shadow is the variance (some outliers are dropped). The noise is sampled from a Gaussian distribution with a mean zero and variance 10^{-3}.

For the CIFAR10 dataset, $\mu_L = 0.5014$ and $\mu_J = 1.7 \times 10^{-13}$. For the MNIST dataset, $\mu_L = 0.7192$ and $\mu_J = 3.7 \times 10^{-13}$. These values are not so large that they are reasonable in practice.

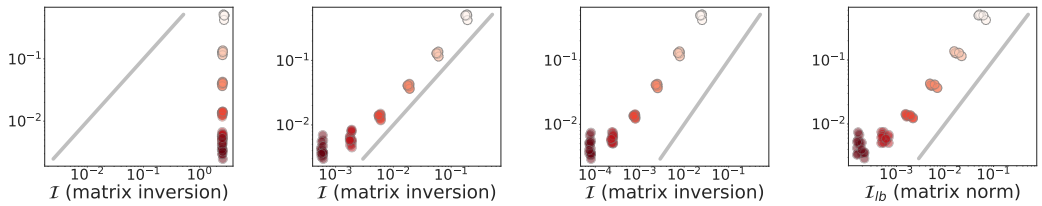
H.2 Extension of I^{2}F to the GS Attack

The derivation of I^{2}F is considered in the DGL attack as defined in Eq. (1), but I^{2}F can also be applied to the GS attack. Note that the minimizer of the DGL attack is one solution to the GS attack. That means the GS attack can be attained by an optimal DGL attack which is our assumption. Therefore, the DGL attack-based theorem is applicable to the GS attack.

Empirically, we evaluate $\mathcal{I}_{lb}$ on GS attack in Figs. 2 and 16. It shows that $\mathcal{I}_{lb}$ is linearly correlated to the metrics of RMSE, which proves the utility of I^{2}F under GS attack.

H.3 Extension of I^{2}F with Prior Knowledge

Our theorem of I^{2}F can be extended to take into account the prior knowledge. Consider the inversion optimization problem with prior knowledge as $\min_x L'_I(x; g) = L_I(x; g) + I_C(x)$ where $I_C(x)$ constrains x in the prior space C and $L_I(x; g) = \|\nabla_\theta L(x; \theta) - g\|$ defined in Eq. (1). Then the optimization problem can be rewritten as $\min_{x \in C} L_I(x; g)$. Thus, as long as the original image x_0 is in the feasible region defined by $I_C(x)$, our Assumption 3.1 and theorems are also applicable. Intuitively, a good regularization should satisfy the requirement, otherwise, it will unreasonably reject the correct recovery.



(a) DGL ($\epsilon = 0.1$) (b) DGL ($\epsilon = 1$) (c) DGL ($\epsilon = 10$) (d) DGL ($\mathcal{I}_{lb}$)

(e) GS ($\epsilon = 0.1$) (f) GS ($\epsilon = 1$) (g) GS ($\epsilon = 10$) (h) GS ($\mathcal{I}_{lb}$)

Figure 19: The impact of the value of ϵ is evaluated on LeNet with the MNIST dataset. The y-axis is the RMSE. (a)(b)(c)(d): DGL attack. (e)(f)(g)(h): GS attack. It is observed that $\mathcal{I}$ (matrix inversion) is effective with $\epsilon \in [1, 10]$ but not $\epsilon = 0.1$. It shows that (1) there exists a range of ϵ where $\mathcal{I}_{lb}$ can lower bound the RMSE; (2) $\mathcal{I}_{lb}$ is an accurate estimator of I^2F, thus we can avoid fine-tuning ϵ.

H.4 Discussion with Prior Works

Closest to our work, Hannun et al. (2021) provided a second-order worst-case metric for analyzing privacy attacks. However, our work provides novel contributions both on technique and implications which essentially root from the proposed I^2F metric.

Technical difference. First, we focus on a different scope against (Hannun et al., 2021). Hannun et al. (2021) proposed Fisher information loss (FIL) to measure the information leakage risk in the context of model inversion and attribute inference, e.g., only the attribute inference is considered in their experiments. Instead, we evaluate the privacy risk under gradient inversion attacks. Second, our metric is more scalable and applicable to large models. For example, in Eq. (18) in (Hannun et al., 2021), the inverse of the Hessian matrix needs to be calculated even when quantifying the information leakage of only one sample, which is inefficient and intractable for large models. Because of the computation inefficiency, only linear regression and logistic regression models are considered in their theories and experiments. Instead, we verify the feasibility of I^2F in much larger models like ResNet152 on ImageNet in Fig. 3.

Our new findings. First, though the unfairness of information leakage of different samples was discussed in (Hannun et al., 2021), we investigate the issue in a different attack method and justify the commonness of the unfairness in different attacks. Second, we additionally provide other insights than (Hannun et al., 2021). For instance, the influence on gradient inversion of different initialization methods is studied. We also find the influence of perturbations is not equivalent even in the same noise scale. We believe these insights are also critical to the privacy and security community, especially in the area of gradient inversion.

Besides, (Guo et al., 2022; Hayes et al., 2023) propose to bound the reconstruction attack in terms of the attack success rate and the expectation of the L_2 distance between the recovered and original image. Nevertheless, their conclusions are based on the differential privacy (DP) quantification framework so it is hard to analyze the influence of other defense mechanisms such as gradient pruning and arbitrary noise. Also, they bound the privacy risk from the statistical sense raised by the randomness of DP, while our work can evaluate the sample-wise worst-case privacy risk at any time during the model training. Moreover, they assume the access of the attacker to the remaining samples (except for the privacy sample) or the multi-round gradient, which is not practical in real gradient inversion scenarios.

H.5 Discussions about the Worst-case Assumption

Our work is mainly built upon the Assumption 3.1 that there is only a unique minimizer for $L_I(x; g)$ given a gradient vector g. Because of the hardness of optimizing a non-linear objective in attack (Eq. (1)), the worst-case may not be reachable in practice. Here, we discuss when the assumption has to be relaxed and why our method is still applicable. In addition, we emphasize that a stronger attacker exists theoretically, resulting in the necessity of a worst-case assumption.

Non-bijective inversion mapping $G_r(g)$. $G_r(g)$ could be non-bijective when the loss function is non-convex. In other words, given the same g, the output of $G_r(g)$ could include multiple choices. We want to argue that this case does not conflict with our assumption. Consider an attack given the exact gradient of a sample. Note that the sample itself is a solution to Eq. (1). Thus, given the exact gradient of a sample, the attack can exactly recover the sample. Even if the solution is non-unique, we can still essentially assume the attack can attain the sample in the worst case.

Optimizing the gradient inversion objective may not converge to the original image. Note that the original image is always an optimal solution for the inversion loss in Eq. (1). Even though the convergence is not guaranteed, there always exists an algorithm that can converge to the original image. To our best knowledge, there is no evidence to show the attack cannot approach the worst case where the original input is recovered. Instead, empirical results have shown that the images can be recovered almost perfectly (Geiping et al., 2020). Thus, due to the sensitivity of privacy, a worst-case assumption is necessary to strictly bound possible privacy risks with arbitrary strong attacks, which is commonly imposed by the literature (Dwork, 2006; Abadi et al., 2016).

I Realistic Impact

Federated learning (FL) (McMahan et al., 2017) is a popular distributed training paradigm that benefits from the data and computation sources from multiple clients. As a principle of FL, clients will upload the local gradient based on the private data to the server for the concerns of data privacy and safety, instead of directly sharing the private data of each client. However, recent works (Geiping et al., 2020; Zhu et al., 2019) show that an attacker, who has access to the local gradient (e.g., a malicious server), can leverage the local gradient to recover the private data of the clients, which we call Deep Gradient Leakage (DGL). Even with large models like Transformer, the attacker can still successfully recover the private data given the gradient (Hatamizadeh et al., 2022).

Auditing potential privacy risks is essentially desired for privacy-sensitive applications, including but not limited to finance (Long et al., 2020), healthcare (Antunes et al., 2022; Xu et al., 2021), and clinical data (Dayan et al., 2021; Roth et al., 2020). Such privacy concerns have been discussed extensively in previous work. For instance, Zhang et al. (2023) raises the privacy concern of FL in financial crime detection while (Kaissis et al., 2021; Li et al., 2023) discusses it in the medical and healthcare applications, respectively.

To echo the demands for privacy risk auditing, we provide a fundamental tool for bounding the worst-case risks of DGL. We show multiple insights, for example, the unfairness of privacy protection using random noise defense. Thus, we expect our work can call the attention of the community to the privacy concern raised by DGL, especially the worst-case instance-level privacy risk. Moreover, We expect I^2F to be a keystone for designing more powerful defense mechanisms.