PLR-Gate: Real-Time Gradient Privacy Assessment and Gated Transmission for Secure Federated Learning

Supplementary Material

1. Preliminaries

1.1. Differential Privacy and Gaussian Mechanism

Differential Privacy (DP) [2] provides a principled framework for protecting sensitive information and has become a cornerstone of privacy-preserving deep learning. Specifically, DP formalizes the principle that the presence or absence of any single individual has only a limited impact on the distribution of the mechanism's outputs, thereby bounding what an adversary can infer about that individual from released results. The formal definition of DP is given below.

Definition 1.1 ((ε,δ)-Differential Privacy [2]) *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ε,δ)-differential privacy if, for any two adjacent datasets $D, D' \in \mathcal{D}$ differing by at most one element, and for any subset of outputs $S \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta \qquad (1)$$

Here, ε denotes the privacy budget, which quantifies the degree of potential privacy leakage risk. Smaller values of ε correspond to stronger privacy guarantees. δ is a small tolerance value that allows for a limited probability of privacy violation.

In practice, DP is enforced by injecting noise calibrated to the global sensitivity of the query (e.g., via the Laplace or Gaussian mechanisms). Specifically, in the context of DP-SGD, the Gaussian mechanism is widely adopted and is defined as follows.

Definition 1.2 (Gaussian Mechanism [3]) *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity $\Delta_2 f = \max_{D,D'} \|f(D) - f(D')\|_2$. The Gaussian Mechanism $\mathcal{M}$ adds noise scaled to $\mathcal{N}(0, \sigma^2)$ to each component of $f(D)$:*

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 I) \qquad (2)$$

For any $\varepsilon \in (0, 1]$ and $\delta > 0$, the mechanism satisfies (ε,δ)-differential privacy when:

$$\sigma \geq \frac{\Delta_2 f}{\varepsilon} \sqrt{2 \log(1.25/\delta)} \qquad (3)$$

1.2. DP-SGD

Differentially Private Stochastic Gradient Descent (DP-SGD) [1] incorporates the principles of Differential Privacy (DP) [2] into stochastic gradient descent to provide formal privacy guarantees during model training. In particular, DP-SGD bounds the influence of each training example by clipping per-sample gradients and adds calibrated Gaussian noise to the aggregated gradients before model update. This procedure rigorously enforces an (ε, δ)-DP guarantee. In practice, the noise magnitude is determined via privacy accounting methods such as the moments accountant.

2. Additional Experimental Setup

2.1. Hyperparameter Settings

MINE is trained with $N_{\text{MINE}} = 500$, $M = 500$, $K = 10$, and $\sigma_s = \|\mathbf{g}_i'\|$. Adversarial examples are generated following the procedure in [4]. Gradients are extracted from the linear layers of the MLP and the convolutional layers of the CNN models. Feature extraction is performed using Autoencoder encoders for both images and gradients, which are optimized with the Adam optimizer at a learning rate of 10^{-5}. Other MINE components employ a learning rate of 10^{-4}. During the training of the evaluated model, we configure the parameters as $\mathcal{B} = 8$, $p = 2$, $r = 0.40$ and $\eta_F = 0.01$. The privacy budget is fixed at $\delta = 10^{-5}$, and the gradient clipping threshold C is set to 1.

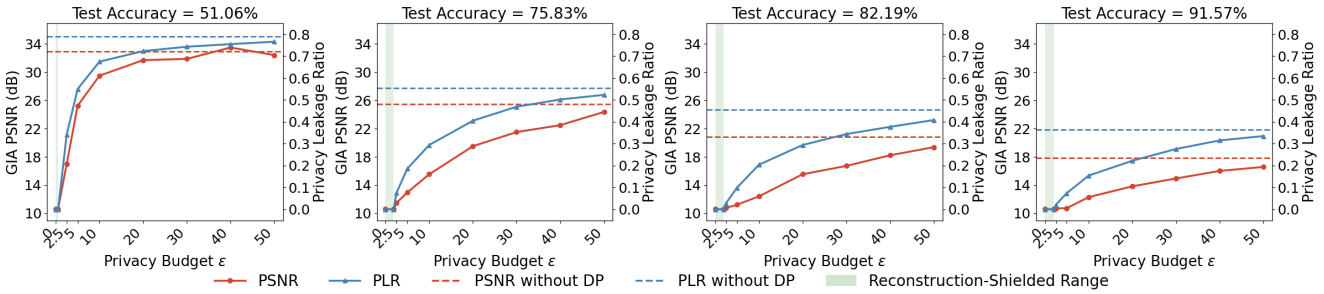
2.2. Computational Resources

All experiments are conducted on a Linux server equipped with an Intel(R) Xeon(R) Gold 5218R CPU (2.10 GHz, 251 GB DRAM) and four NVIDIA RTX A6000 GPUs (48 GB memory per GPU). All implementations are executed using PyTorch 2.0.1.

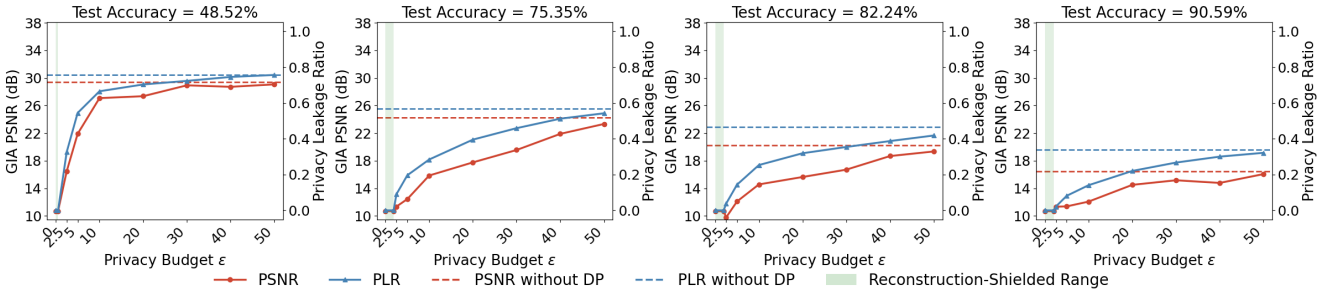
3. Additional Ablation Studies

3.1. Impact of Evaluated Model State on Gradient Privacy Risk under Varying Privacy Budgets

This section examines how the trained model state affects gradient privacy risk under different privacy budgets ε. Experiments are conducted on the CelebA dataset using both MLP and AlexNet as representative architectures. As illustrated in Fig.1, the two models exhibit consistent behavioral patterns. In the early training stages, a modest increase in ε from an initially small value leads to a sharp rise in the leakage of DP-AGs with respect to OTIs. This leakage quickly saturates as ε approaches approximately 10, beyond which GIAs already achieve high-fidelity reconstructions, and further relaxation of ε produces negligible increases in gradient



(a) MLP



(b) AlexNet

Figure 1. Trends of reconstructed image PSNR and PLR with varying privacy budgets ε under different model states, on the CelebA dataset. Higher test accuracy corresponds to later training epochs, indicating a more converged model.

privacy risk. In contrast, as the model converges, the growth of DP-AG privacy risk progressively slows with increasing ε and eventually plateaus at a lower level. At this stage, the progressively convergent model state inherently suppresses reconstruction fidelity, even in the absence of differentially private noise, thereby attenuating the sensitivity of gradient privacy risk to variations in the privacy budget. Importantly, the convergence thresholds observed in both the PSNR of reconstructions and the PLR closely align with those obtained under non-private conditions (i.e., $\varepsilon = \infty$). These thresholds characterize the upper bound of the gradient privacy risk achievable for a given model state and batch size.

Moreover, when ε falls within an extremely small range (approaching 0), reconstruction performance reaches its lowest saturation point and no longer deteriorates with further decreases in ε. Meanwhile, the PLR remains near zero without any discernible upward or downward trend. This phenomenon arises because the excessive noise introduced under such stringent DP constraints effectively prevents DP-AGs from revealing OTI information. In this regime, the gradients are already fully protected by DP, and adding larger-magnitude noise yields no further privacy gains. We define this interval as the reconstruction-shielded range, in which sufficient noise perturbation completely suppresses the recovery of OTI-related private information. Notably, models closer to convergence exhibit a broader reconstruction-shielded range, as their internal representations inherently weaken the dependence between DP-AGs and OTIs. Consequently, a smaller magnitude of noise (i.e., a larger ε) is sufficient for DP-AGs to begin leaking private information about OTIs.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 1

[2] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006. 1

[3] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014. 1

[4] Jingyang Zhang, Yiran Chen, and Hai Li. Privacy leakage of adversarial training models in federated learning systems. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 108–114, 2022. 1