Supplementary Material

Bochao Liu[1,2], Pengju Wang[1,2], and Shiming Ge[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

1 Procedure of DP-SAD

The procedure of our DP-SAD is shown in Alg. 1. It can be described as the following four steps:

- train a teacher model ϵ_ψ without protection.
- initialize the student model ϵ_θ and the discriminator ϵ_ϕ.
- randomly sample a batch of data : $\{x_i\}_{i=1}^{B}$.
- calculate the loss function with Eq. (12) and update the student and the discriminator with Eq. (17).

We run the last two steps until the termination condition is reached.

Algorithm 1: DP-SAD

Require: Private data $\mathcal{D}$, time step T, batch size B, training iterations N, learning rates γ and γ_d, the teacher ϵ_ψ, the student ϵ_θ and the discriminator ϵ_ϕ.

Train a teacher model with private data $\mathcal{D}$ without protection.
Initialize θ_0 and ϕ_0 with Xavier.
for $k < N$ **do**
  sample a batch of data from $\mathcal{D}$: $\{x_i\}_{i=1}^{B}$.
  Compute the final loss $\mathcal{L}$ with Eq. (12) and get the differentially private gradients $\bar{g}$ with Eq. (17).
  Update the student with $\theta_{k+1} = \theta_k - \gamma \cdot \bar{g}$
  Compute the loss $\mathcal{L}_r = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{adv}^{i,r}$ and update the discriminator with $\phi_{k+1} = \phi_k - \gamma_d \cdot \partial \mathcal{L}_r / \partial \phi$.
end for
return θ_N

2 Convergence Analysis

We assume that the function to be optimized is $\mathcal{L}(\theta)$, where θ is the parameter of the student model. We follow the standard assumptions same as [1]:

$$\begin{aligned}
&(1)\ ||\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')||_2 \leq \tau_1||\theta - \theta'||_2; \\
&(2)\ \mathcal{L}(\theta) \geq \mathcal{L}(\theta') + \nabla\mathcal{L}(\theta')^T(\theta - \theta') + \frac{\tau_2}{2}||\theta - \theta'||_2^2; \\
&(3)\ \nabla\mathcal{L}(\theta)^T\mathbb{E}_t[\bar{g}(\theta;t)] \geq \mu||\nabla\mathcal{L}(\theta)||_2^2,
\end{aligned} \tag{1}$$

where $\nabla\mathcal{L}(\theta)$ is the true gradient (Eq (13) in the main text), $\bar{g}(\theta;t)$ is the gradient we used to update the student (Eq (15) in the main text), $\mathbb{E}[\cdot]$ is the symbol for mean calculation, $\mathbb{V}[\cdot]$ is the symbol for variance calculation and $\tau_1, \tau_2, \mu, \mu_e, \mu_v, c$ are non-negative constants. According to assumption (1), we have:

Lemma 1. *For any two weights θ and θ', the difference of the objective function $\mathcal{L}(\theta) - \mathcal{L}(\theta')$ is limited by the distance between the weights.*

$$\mathcal{L}(\theta) \leq \mathcal{L}(\theta') + \nabla\mathcal{L}(\theta')^T(\theta - \theta') + \frac{\tau_1}{2}||\theta - \theta'||_2^2. \tag{2}$$

Proof. The Taylor expansion of the objective function $\mathcal{L}(\theta)$ can be expressed as:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta') + \nabla\mathcal{L}(\theta')^T(\theta - \theta') + \frac{1}{2}(\theta - \theta')^T\nabla^2\mathcal{L}(\vartheta)(\theta - \theta'), \tag{3}$$

where ϑ is any point between θ and θ'. According to assumption (1), we know the Hessian matrix satisfies:

$$\nabla^2\mathcal{L}(\theta) \leq \tau_1. \tag{4}$$

Combining Eq. (3) and Eq. (4), we get Lemma. 1.

Based on the assumption (2), we have:

Lemma 2. *For any weight θ, the distance between $\mathcal{L}(\theta)$ and the minimum value $\mathcal{L}(\theta^*)$ is limited by $\nabla\mathcal{L}(\theta)$ as follows*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \leq \frac{1}{2\tau_2}||\nabla\mathcal{L}(\theta)||_2^2. \tag{5}$$

Proof. We regard the right side of the inequality as a quadratic function on θ. When $\theta = \theta' - \frac{1}{\tau_2}\nabla\mathcal{L}(\theta')$, it takes the minimum value $\mathcal{L}(\theta') - \frac{1}{2\tau_2}||\nabla\mathcal{L}(\theta')||_2^2$. Substituting it into assumption (2) and letting $\theta = \theta^*$, we can get Lemma 2.

We consider the update at step k as $\theta_{k+1} = \theta_k - \gamma \cdot \bar{g}(\theta_k;t)$. Based on Lemma. 1, we have:

$$\mathcal{L}(\theta_{k+1}) \leq \mathcal{L}(\theta_k) - \gamma\nabla\mathcal{L}(\theta_k)^T\bar{g}(\theta_k;t) + \frac{\tau_1}{2}\gamma^2||g(\theta_k;t)||_2^2. \tag{6}$$

Taking the expectations on both sides gives:

$$\mathbb{E}[\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k)] \leq -\gamma \nabla \mathcal{L}(\theta_k)^T \mathbb{E}[\bar{g}(\theta_k, t)] + \frac{\tau_1}{2}\gamma^2 \mathbb{E}[||\bar{g}(\theta_k; t)||_2^2]. \quad (7)$$

Since $\bar{g} = CLIP(g, C) = g/max(1, \frac{||g||_2}{C}) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$, combining with Cauchy Schwartz inequality yields:

$$\mathbb{E}[||\bar{g}(\theta; t)||_2^2] \leq 2C^2 + 2\sigma^2 C^2 d, \quad (8)$$

where $d = ||z||^2$, $z \sim \mathcal{N}(0, \mathbf{I})$. Substituting Eq. (8) into Eq. 7 and combining it with assumption (3) yields:

$$\mathbb{E}[\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k)] \leq -\gamma\mu ||\nabla \mathcal{L}(\theta)||_2^2 + \frac{\tau_1}{2}\gamma^2(2C^2 + 2\sigma^2 C^2 d). \quad (9)$$

Combined with Lemma. 2, we have:

$$\mathbb{E}[\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k)] \leq -2\tau_2 \gamma\mu \mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)] + \frac{\tau_1}{2}\gamma^2(2C^2 + 2\sigma^2 C^2 d). \quad (10)$$

We define $d_k = \mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)$ and after the transformation we have:

$$d_{k+1} - \frac{\tau_1 \gamma C^2(1 + \sigma^2 d)}{2\tau_2 \mu} \leq (1 - 2\tau_2 \gamma\mu)(d_k - \frac{\tau_1 \gamma C^2(1 + \sigma^2 d)}{2\tau_2 \mu}). \quad (11)$$

Our DP-SAD converges when we guarantee that $0 < 2\tau_2\gamma\mu < 1$ and the error from the minimum $\mathcal{L}(\theta^*)$ is $\frac{\tau_1 \gamma C^2(1+\sigma^2 d)}{2\tau_2 \mu}$.

3 Experimental Details

Datasets. MNIST and FashionMNIST are both 10-class datasets containing 60,000 training images and 10,000 testing images. Each image is 28×28 grayscale image. CelebA is a face attribute dataset, which contains 202,599 color images of celebrity faces. We use the official preprocessed version with the face alignment and resize the images to 64×64×3. We create CelebA-H and CelebA-G based on it. CelebA-H is a classification dataset with hair color (black/blonde/brown) as the label and CelebA-G is a classification dataset with gender as the label.

Baselines. DP-GAN is to directly apply the DPSGD training strategy to the training process of WGAN. Because WGAN itself satisfies the Lipschitz condition, the effect from gradient clipping is eliminated. PATE-GAN, DP-MERF, GS-WGAN, P3GM, G-PATE and DataLens are all based on PATE framework with different teacher aggregation strategies. All of the above baselines achieve differential privacy based on the Gaussian mechanism. DPGEN achieves differential privacy based on randomized response mechanism. PSG incorporates the downstream task into training to improve its data quality, but it requires repeated training for different downstream tasks. Both DP-DM and DP-LDM are implemented by applying DPSGD directly to diffusion models for differentially

private generative modeling. We get the experimental results by running official codes or from original papers.

Implementations. We set the norm bound C to 10^{-6}. We set the training epoch to 100 for all models and compute the σ by RDP. We set the trade-off weight λ and time step T to 1 and 500, respectively. We set the initial values of both γ and γ_d to 10^{-4}, and employ a "CosineAnnealingLR" to adjust them. We set ω in Eq.(5) to 1.8, batch size to 128, β_{start} to 1e-4, β_{end} to 0.028, σ=1.9 for ε=1 and σ=0.6 for ε=10. A simplified version of the model structures for different resolutions is shown in Fig. 1

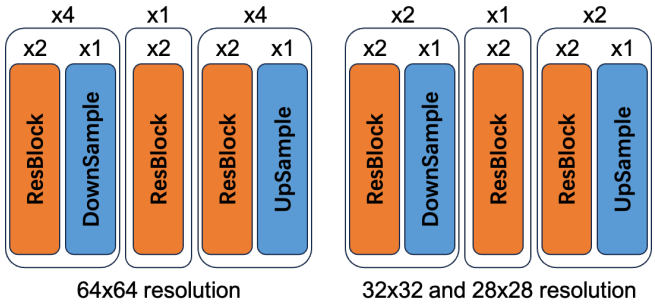

Fig. 1: Model structure for different resolutions.

4 Disscusion about different loss functions

To investigate the impact of the loss function, we compare the classification accuracy on MNIST with different loss functions. The results are shown in Tab. 1. The adversarial loss (ADV) forces the student manifold to conform to the teacher manifold, which accelerates the convergence process. The teacher MSE loss (MSE-T) can also accelerate the convergence process. When the privacy budget (ε) is determined, faster convergence leads to a model with better performance.

Table 1: Classification accuracy comparisons on MNIST with different loss functions.

MSE-T	MSE	ADV	e=20	e=50	e=100
✓			0.5811	0.7452	0.9499
✓	✓		0.6058	0.7617	0.9512
✓	✓	✓	**0.6949**	**0.8272**	**0.9761**

5 Limitations

There are still some limitations to DP-SAD: i) because we need to set a large time step T to dilute the effect of DP noise, which leads to low efficiency in sampling the generated images. Unlike GANs, which require only a single inference step to generate the final image, our method necessitates T inference iterations to produce the final image; ii) our method does not employ the architecture of latent diffusion models; instead, it integrates with a VAE. The architecture of latent diffusion models is more advantageous for generating high-resolution images; iii) inevitably, the model might inadvertently assimilate the biases present within the dataset. In future research, efforts could be directed toward mitigating the acquisition of these biases by incorporating specific prompts during the model's training phase.

6 Extened Visualization Results

We further visualize the generated results on three datasets, including MNIST, FMNIST and CelebA, under different privacy budget. The results are shown in Fig. 2, Fig. 3, Fig. 4 and Fig. 5. We find that there is not much difference visually between the results for $\varepsilon = 1$ and $\varepsilon = 10$, which demonstrates the potential of our method to generate higher-resolution images with stronger privacy protection.

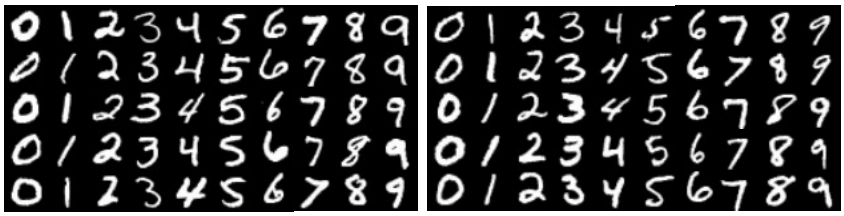

Fig. 2: Visualization results of MNIST with 28×28 resolution under $\varepsilon = 1$ (left) and $\varepsilon = 10$ (right).

References

1. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM pp. 223–311 (2018)

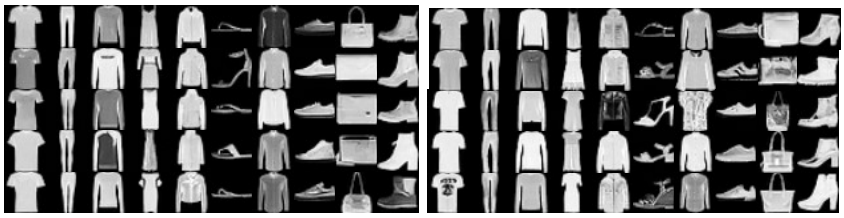

Fig. 3: Visualization results of FMNIST with 28×28 resolution under $\varepsilon = 1$ (left) and $\varepsilon = 10$ (right).

Fig. 4: Visualization results of CelebA with 64×64 resolution under $\varepsilon = 1$ (left) and $\varepsilon = 10$ (right).

Fig. 5: Visualization of CelebA at 128×128 resolution.