UNDERSTANDING IMPACTS OF DIFFERENTIAL PRIVACY: A UNIFIED FRAMEWORK WITH TWO-LAYER NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the growing demand for data and the increasing awareness of privacy, differentially private learning has been widely applied in various deep models. Experiments have observed several side effects of differentially private learning, including *bad learning features (performance)*, *disparate impact*, and *worse adversarial robustness* that hurt the trustworthiness of the trained models. Recent works have expected pre-training to mitigate these side effects. It is valuable to theoretically understand the impact of differential privacy on the training process. However, existing theoretical research only explained parts of the phenomena and failed to extend to non-convex and non-smooth neural networks. To fill this gap, we propose a unified framework to explain all the above phenomena by studying the feature learning process of differentially private stochastic gradient descent in two-layer ReLU convolutional neural networks. By analyzing the test loss, we find both its upper and lower bound decrease with feature-to-noise ratios (FNRs). We then show that disparate impact comes from imbalanced FNRs among different classes and subpopulation groups. Additionally, we show that the suboptimal learned features and reduced adversarial robustness are caused by the randomness of privacy-preserving noise introduced into the learned features. Moreover, we demonstrate that pre-training cannot always improve the model performance, especially with increased feature differences in the pre-training and fine-tuning datasets. Numerical results on both synthetic and real-world datasets validate our theoretical analyses.

1 INTRODUCTION

Modern deep learning models have achieved remarkable success in various applications, such as image classification (He et al., 2022) and natural language processing Vaswani et al. (2017). Many of these applications require training on datasets that contain sensitive private information. Differentially private learning, as an approach, seeks to train these models while ensuring rigorous privacy guarantees (Abadi et al., 2016). A standard differentially private learning algorithm is Differential Privacy Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016), which injects noise into network parameter updates during optimization.

While DP-SGD provides robust privacy guarantees, it introduces side effects, including

- *Bad learned features (Tramer & Boneh, 2020):* DP-SGD trained models might learn bad features that are worse than handcrafted ones.
- *Disparate impact (Bagdasaryan et al., 2019; Sanyal et al., 2022):* DP-SGD trained models achieve different accuracy on different classes and different subpopulation groups.
- *Worse adversarial robustness (Tursynbek et al., 2020):* DP-SGD trained models might be less adversarially robust than non-private models.

Moreover, recent work (De et al., 2022) has shown that pre-training improves the accuracy of DP-SGD trained models by 30% compared with training from scratch.

Understanding these phenomena is essential for deploying trustworthy, differentially private learning systems. Although several studies have explored side effects such as disparate impact (Esipova et al., 2022; Bagdasaryan et al., 2019), existing models are not applicable to ReLU neural networks due to their inherent non-convex and non-smooth nature. Moreover, a unified theoretical framework that explains all of the aforementioned phenomena is still lacking.

Thus, there remains an open problem:

How to theoretically explain the aforementioned phenomena in DP-SGD trained ReLU neural networks with a unified framework?

Facing the challenges in non-convex and non-smooth neural networks, we answer this problem in two steps. First, we derive test loss bounds of two-layer ReLU Convolutional Neural Network (CNN)s trained with DP-SGD in Section 3. Then, we extend the test loss analysis in Section 4 to explain the aforementioned phenomena.

1.1 Related Work

In this section, we discuss the related works from two perspectives. Interested readers can refer to Appendix A for a detailed discussion.

Analysis on differentially private learning side effects. After observing the side effects of differentially private learning, some researchers provided theoretical explanations. For example, Tran et al. (2021) employed Taylor expansion to understand the disparate impact from the optimization local loss landscape for twice differentiable loss functions. Sanyal et al. (2022) studied unfairness in long-tailed data distribution in an asymptotic setting where the number of training data tends to infinity. Zhang & Bu (2022) studied the adversarial robustness of private linear classifiers. However, these analyses relied on some restricted assumptions that are not applicable to neural networks due to their non-convex and non-smooth nature. In this paper, we provide a unified framework to explain all the mentioned side effects by characterizing the feature learning process of DP-SGD trained two-layer ReLU CNNs.

Feature learning in neural networks. Feature learning in neural networks explores how neural networks learn data-related patterns during training, offering insights into phenomena such as momentum (Jelassi & Li, 2022), benign overfitting (Cao et al., 2022; Kou et al., 2023), adversarial training (Allen-Zhu & Li, 2022; Li & Li, 2023), and data augmentation Zou et al. (2023). In this work, we make an initial step towards studying the generalization performance and effects in DP-SGD trained two-layer neural networks.

Although the convergence and feature learning process are analyzed in two-layer neural networks before, their approaches are not applicable to DP-SGD trained neural networks as the key properties for feature growth may not hold due to the perturbation of added noise. In this paper, we design new techniques to study the generalization performance of DP-SGD trained models.

1.2 Contributions and Main Results

We provide a unified framework to explain the side effects in DP-SGD trained two-layer ReLU CNNs. Specifically,

- We explore the feature learning process of DP-SGD trained neural networks. Considering the technical challenges posed by non-convex and non-smooth ReLU CNN and the random Differential Privacy (DP) noise, we propose a new proof technique to derive the standard and adversarial test loss bounds. The high-level idea is to use a piece-wise linear function to bound the non-linear loss function. We theoretically prove that the upper and lower bounds of test loss depend on data feature size and DP-SGD noise.
- We provide theoretical explanations for the side effects and the effectiveness of pre-training in DP-SGD trained neural networks.
  - The noise from DP-SGD highly perturbs the learned activation patterns during iterations, resulting in *bad learned features (performance)*.

- Three factors, i.e., the gradient clipping, the group or class sizes, and feature sizes, contribute to *disparate impact* in DP-SGD trained models by influencing the test loss within a given group or class.
- The DP-SGD trained models exhibit *worse adversarial robustness* because they learn non-robust, class-irrelevant features from the random DP noise.
- The fine-tuning performance on a pre-trained neural network decreases with increased feature difference. This implies that the benefits from pre-training may actually come from distribution overlaps between pre-training and fine-tuning datasets.

1.3 NOTATION

We use lowercase letters, lowercase boldface letters, and uppercase boldface letters to denote scalars, vectors, and matrices, respectively. We use $[m]$ to denote the set $\{1, \cdots, m\}$. Given two sequences $\{x_n\}$ and $\{y_n\}$, we denote $x_n = \mathcal{O}(y_n)$ if $|x_n| \leq C_1|y_n|$ for some positive constant C_1 and $x_n = \Omega(y_n)$ if $|x_n| \geq C_2|y_n|$ for some positive constant C_2. We use $x_n = \Theta(y_n)$ if both $x_n = \mathcal{O}(y_n)$ and $x_n = \Omega(y_n)$ hold. We use $\tilde{\mathcal{O}}(\cdot), \tilde{\Theta}(\cdot), \tilde{\Omega}(\cdot)$ to omit logarithmic factors in these notations. Given a set $\mathcal{T}$, we use $|\mathcal{T}|$ to denote its cardinality. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote its $\ell_p (p \geq 1)$ norm as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. The notation $(\mathbf{x}, y) \sim \mathcal{D}$ indicates that the data sample $(\mathbf{x}, y)$ is generated from a distribution $\mathcal{D}$.

2 MODEL

We consider a one-hidden layer CNN with data structured as follows. Similar data structures have been applied in (Allen-Zhu & Li, 2020; Jelassi & Li, 2022; Jelassi et al., 2022; Li & Li, 2023; Zou et al., 2023; Cao et al., 2022).[1]

Data distribution. We consider a 2-class classification problem over 2-patch inputs. Each labelled data is denoted as $(\mathbf{x}, y)$, with label $y \in \{1, 2\}$ and data vector $\mathbf{x} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) \in \mathbb{R}^{d \times 2}$. We assume that each label class y has 2-group features, i.e., the majority (common) group features $\mathbf{u}_{y,\text{Maj}}$ and the minority (rare) group features $\mathbf{u}_{y,\text{Min}}$. A sample $(\mathbf{x}, y)$ is generated from a data distribution $\mathcal{D}$ as follows.

1. The label y is randomly sampled from $\{1, 2\}$. With probability $p_c > 0$, the label is selected as $y = 1$; otherwise, it is selected as $y = 2$.
2. Each input data patch $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathbb{R}^d$ contains either feature or noise.
   - Feature patch: One data patch ($\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$) is randomly selected as the feature patch. With probability $p_f > 0.5$, this patch contains majority features $\mathbf{u}_{y,\text{Maj}}$ for $y \in \{1, 2\}$. Otherwise, this patch contains minority features $\mathbf{u}_{y,\text{Min}}$ for $y \in \{1, 2\}$.
   - Noisy patch: The remaining patch $\boldsymbol{\xi}$ is generated from a Gaussian distribution $\mathcal{N}(0, \sigma_p^2 \mathbf{H})$, where $\mathbf{H} = \mathbf{I} - \sum_{i=1}^2 \sum_{j \in \{\text{Maj,Min}\}} \mathbf{u}_{i,j} \mathbf{u}_{i,j}^\top \cdot \|\mathbf{u}_{i,j}\|_2^{-2}$.

Without loss of generality, we assume all the feature vectors are orthogonal, i.e., $\langle \mathbf{u}_{i,j}, \mathbf{u}_{i',j'} \rangle = 0$ for all $i \in \{1, 2\}, j \in \{\text{Maj}, \text{Min}\}$ when $(i, j) \neq (i', j')$. Additionally, we consider the features of majority and minority satisfy $p_f \|\mathbf{u}_{i,\text{Maj}}\|_2 > (1 - p_f) \|\mathbf{u}_{i,\text{Min}}\|_2$ for all $i \in \{1, 2\}$.

Moreover, we define distributions $\mathcal{D}_{i,j}, i \in \{1, 2\}, j \in \{\text{Maj}, \text{Min}\}$, with probability density functions given by:

$$\mathbb{P}_{\mathcal{D}_{i,j}}[(\mathbf{x}, y)] = \mathbb{P}_{\mathcal{D}}[(\mathbf{x}, y) | y = i, \text{Feature patch of } \mathbf{x} = \mathbf{u}_{i,j}] .$$

Learner model. We consider a two-layer CNN with ReLU activation as the learner model. The first layer of the CNN comprises m neurons (filters) for class 1 and m neurons for class 2. Each neuron processes the two data patches separately. The parameters of the CNN's second layer are

[1] See (Allen-Zhu & Li, 2020) for detailed experimental justifications of the data distribution.

fixed as $1/m$. Given an input vector $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, the model outputs a vector $[F_1, F_2]$ whose k^{th} element is

$$F_k(\mathbf{W}, \mathbf{x}) = \frac{1}{m} \sum_{r=1}^{m} \sum_{j=1}^{2} \sigma\left(\left\langle \mathbf{w}_{k,r}, \mathbf{x}^{(j)} \right\rangle\right), \tag{1}$$

where $\sigma(\cdot) = \max\{\cdot, 0\}$ denotes the activation function $\mathrm{ReLU}(\cdot)$. We use $\mathbf{W}$ to denote the collection of all model weights and $\mathbf{w}_{k,r}$ to denote the weight vector the r^{th} neuron associated with $F_k(\mathbf{W}, \mathbf{x})$.

Training objective. Given a training dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, we train the neural networks by minimizing the empirical risk with cross-entropy loss, i.e.,

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{W}, \mathbf{x}_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} [-\log(\mathrm{prob}_{y_i}(\mathbf{W}, \mathbf{x}_i))], \tag{2}$$

where $\mathrm{prob}(\cdot)$ represents the softmax predictions with the output of the neural network, i.e.,

$$\mathrm{prob}_{y_i}(\mathbf{W}, \mathbf{x}_i) = \frac{\exp(F_{y_i}(\mathbf{W}, \mathbf{x}_i))}{\sum_{k=1}^{2} \exp(F_k(\mathbf{W}, \mathbf{x}_i))}. \tag{3}$$

Differential privacy and training algorithm. DP (Dwork et al., 2014) (see the definition below) stands as the benchmark for quantifying privacy leakage, offering rigorous privacy guarantees.

Definition 2.1 ((α, δ)-Differential privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ is (α, δ)-DP if for every pair of neighboring datasets $Z, Z' \in \mathcal{Z}$ that differ in one entry and for any subset of output $\mathcal{S} \subseteq \mathcal{R}$, we have

$$\mathbb{P}[\mathcal{M}(Z) \in \mathcal{S}] \leq e^{\alpha} \mathbb{P}[\mathcal{M}(Z') \in \mathcal{S}] + \delta. \tag{4}$$

DP-SGD, the standard and most popular training algorithm (Abadi et al., 2016) (refer to Appendix C for details), achieves DP through an update given by

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \frac{\eta}{B} \cdot \sum_{(\mathbf{x}, y) \in \mathcal{S}^{(t)}} \mathrm{clip}_C\left(\nabla \mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{x}, y\right)\right) + \eta \cdot \mathbf{n}^{(t)}, \tag{5}$$

where η represents the learning rate and $\mathrm{clip}_C(\mathbf{x})$ is the gradient clipping function with clipping threshold C on vector $\mathbf{x}$, i.e.,

$$\mathrm{clip}_C(\mathbf{x}) = \frac{\mathbf{x}}{\max\{1, \|\mathbf{x}\|_2 / C\}}. \tag{6}$$

In (5), $\mathcal{S}^{(t)}$ represents the randomly selected mini-batch datasets with a batch size B from the training dataset $\mathcal{S}$ generated from the data distribution $\mathcal{D}$, and $\mathbf{n}^{(t)}$ is the noise used for privacy protection following $\mathcal{N}(0, \sigma_n^2 \mathbf{I})$ at iteration t. We initialize network parameters with Gaussian initialization, where all entries of $\mathbf{W}^{(0)}$ are sampled from i.i.d. Gaussian distributions $\mathcal{N}(0, \sigma_0^2 \mathbf{I})$.

3 Test Loss Analysis

In this section, we analyze the test loss of DP-SGD trained CNNs, serving as a stepping-stone for analyzing the impacts of DP-SGD in Section 4.

3.1 Standard Test Loss Analysis

We first define the standard test loss of the trained model on data distribution $\mathcal{D}_{i,j}$, for any $i \in [2]$ and $j \in \{\mathrm{maj}, \mathrm{min}\}$ as follows.

$$\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{i,j}}[\mathcal{L}(\mathbf{W}, \mathbf{x}, y)]. \tag{7}$$

Our results for test loss are based on the following conditions and assumption.

Condition 3.1. Suppose that there exists a positive constants $c_1 > 0, 0 < c_2 < 1$ and a sufficiently large constants $c_3, c_4 > 0$ such that

1. The CNNs have sufficiently large number of neurons, i.e., $m \geq c_1 \cdot d$.
2. The dimension d satisfies $d \geq 50$.
3. The batch size satisfies $B \geq c_2 \cdot n$.
4. Feature size, patch noise and DP-SGD noise satisfies $\|\mathbf{u}_{i,j}\|_2 = \Theta(\sqrt{d}\sigma_p), \sigma_n \leq c_3\sigma_p$, for $i \in [2], j \in \{\text{maj}, \text{min}\}$.
5. The learning rate satisfies $\eta \leq \left(c_4(C + \sqrt{d}\sigma_n)(\max_{i,j} \|\mathbf{u}_{i,j}\|_2 + \sqrt{d}\sigma_p)\right)^{-1}$.

Condition 1 and 2 are common in theoretical analyses (e.g., (Allen-Zhu et al., 2019; Kou et al., 2023)) and also match the practical setting that modern neural networks are usually over-parameterize, i.e., have more parameters than the number of training examples. Condition 3 guarantees that batch size is proportional to the training dataset size so that stochastic gradients can take advantage of large training datasets. Condition 4 ensures that (1) the feature in not small so that the model can learn features (Kou et al., 2023); (2) the noise added by DP is in a reasonable range, i.e., privacy budgets are not too tight ($\epsilon \geq 0.5$). Condition 5 ensures that the model update is bounded.

Assumption 3.2 (s-non-perfect model)**.** We assume the model is almost surely not perfect on a test example, i.e.,

$$\mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{x}, y\right) \geq s, \ \forall (\mathbf{x}, y) \sim \mathcal{D}, \tag{8}$$

with some constant $s > 0$.

Assumption 3.2 is considered a mild assumption. Given that DP-SGD introduces randomness in the model training, the resulting trained model is randomized and is unlikely to attain a zero cross-entropy loss on a test example.

Next, we define Feature-to-Noise Ratios (FNRs) and the clipping factor to facilitate the analyses.

Definition 3.3 (Feature-to-noise ratios)**.** Feature-to-noise ratio of the class i in group j is defined as

$$\mathcal{F}_{i,j} = \frac{\|\mathbf{u}_{i,j}\|_2}{\sqrt{d}\sigma_n}, \tag{9}$$

where $\gamma_{i,j}$ denotes the expected proportion of class i group j data in the whole training dataset, i.e., $\gamma_{1,\text{maj}} = p_c p_f, \gamma_{1,\text{min}} = p_c(1-p_f), \gamma_{2,\text{maj}} = (1-p_c)p_f, \gamma_{2,\text{min}} = (1-p_c)(1-p_f)$.

Definition 3.4 (Clipping factor)**.** The clipping factor for the class i in group j is defined as

$$\Lambda_{i,j} = \frac{C}{\|\mathbf{u}_{i,j}\|_2 + \sigma_p\sqrt{d}}. \tag{10}$$

Here, the clipping factor $\Lambda_{i,j}$ quantifies the maximum change of gradient magnitude on data from class i and group j.

Based on the conditions, assumption, and definitions, we characterize upper bound test loss of DP-SGD trained models in Theorem 3.5.

Theorem 3.5. *Under Condition 3.1 and Assumption 3.2, with a probability at least* $1 - \exp(-\tilde{\Omega}(d))$*, for any* $i \in \{1, 2\}, j \in \{maj, min\}$*, the test loss of a DP-SGD trained model satisfies*

$$\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(T)}\right) \leq \bar{L}_{i,j}(\mathbf{W}^{(T)}), \tag{11}$$

where

$$\begin{aligned}\bar{L}_{i,j}(\mathbf{W}^{(T)}) =& \underbrace{\exp\left(-\Omega\left(\frac{\Lambda_{i,j}\gamma_{i,j}\|\mathbf{u}_{i,j}\|_2^2}{\sqrt{m}}T\right)\right)\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(0)}\right)}_{\text{Vanishing error}} + \underbrace{\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{mn}\gamma_{i,j}\Lambda_{i,j}}\right)}_{\text{Generalization error}} \\ &+ \underbrace{\mathcal{O}\left(\frac{\sqrt{m}}{\Lambda_{i,j}\gamma_{i,j}\mathcal{F}_{i,j}}\right)}_{\text{Privacy protection error}}.\end{aligned} \tag{12}$$

We defer the proof of Theorem 3.5 to Appendix G.2. As the randomness of noise may improve the model, the high probability test loss lower bound becomes trivial. We instead provide a lower bound of the expected test loss in the following.

Theorem 3.6. *Under Condition 3.1 and Assumption 3.2, with the number of iterations* $T \geq \Omega\left(-1/\log\left(1-\Omega\left(\eta \min_{i,j}\{\gamma_{i,j}\|\mathbf{u}_{i,j}\|_2^2\}/m\right)\right)\right)$ *and a probability at least* $1-\tilde{\mathcal{O}}(1/d)$*, for any* $i \in \{1,2\}, j \in \{maj, min\}$*, the expected test loss a DP-SGD trained model satisfies*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(T)})\right] \geq \exp\left(-\Omega\left(\frac{\gamma_{i,j}\|\mathbf{u}_{i,j}\|_2^2}{m}T\right)\right)\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(0)}) + \underbrace{\Omega\left(\frac{\sigma_p^2}{\gamma_{i,j}\mathcal{F}_{i,j}^2}\right)}_{\text{Privacy protection error}} - \mathcal{O}\left(\frac{1}{\gamma_{i,j}}\sqrt{\frac{d}{n}}\right). \tag{13}$$

Remark 3.7. Theorems 3.5 and 3.6 show that the upper and lower bound of test loss both decrease with the feature-to-noise ratio $\mathcal{F}_{i,j}$. Moreover, Theorem 3.5 illustrates the presence of non-vanishing error terms in the test loss bound, i.e., generalization error and privacy protection error. Specifically,

- Generalization error arises from data noise patch. This error decreases with $\mathcal{O}(1/\sqrt{n})$, aligning with the generalization error bounds derived in neural networks (Arora et al., 2019; Xu & Mannor, 2012).
- Privacy protection error stems from DP noise. With a fixed privacy budget ((α, δ) in Definition 2.1), the noise variance σ_n^2 increases with the number of iterations T as $\sigma_n^2 = \Theta(T)$ (Abadi et al., 2016). Consequently, the error due to privacy protection increases with the number of iterations as $\Theta(\sqrt{T})$. This rate aligns with the non-vanishing training error of DP-SGD on Lipschitz-smooth objectives (Bu et al., 2024).

Privacy-Utility Tradeoff Privacy composition (Theorem 1 in (Abadi et al., 2016)) shows that the noise variance σ_n^2 increases with $\Theta(T)$ under (α, δ)-DP. Then, the privacy protection error in (12) is in order $\mathcal{O}(\sqrt{T\log(1/\delta)}/(n\alpha))$, implying that 1) larger privacy budget leads to larger loss; 2) larger dataset has small error thanks to privacy amplification by subsampling; 3) the loss increases with the number of iterations. In addition, we demonstrate a sharp phase transition between benign and harmful privacy protection to model utility, as shown in Figure 1. Details about the simulation settings are provided in Appendix D.2.

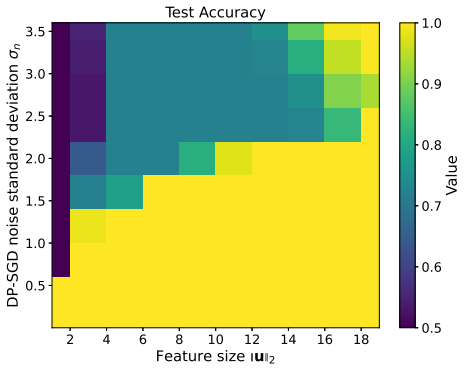


Figure 1: Illustration of the privacy-utility phase transition between benign and harmful privacy protection. The yellow region represents a benign regime where the test loss is small. The blue region represents a harmful regime where the test loss is large.

3.2 Adversarial Test Loss Analysis

In this subsection, we analyze the impact of DP-SGD on adversarial robustness as a basis for understanding the side effect of worse adversarial robustness in Section 4.

Adversarial robustness refers to a machine learning model's ability to withstand carefully manipulated input samples, commonly known as adversarial examples.

Definition 3.8 (Adversarial example)**.** For an data example $(\mathbf{x}, y)$, the corresponding adversarial example is $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\zeta}$, where $\boldsymbol{\zeta} = \arg\max_{\|\boldsymbol{\zeta}\|_p \leq \bar{\zeta}} \mathcal{L}(\mathbf{W}, \mathbf{x}+\boldsymbol{\zeta}, y)$ is the adversarial perturbation and $\bar{\zeta}$ is the perturbation radius.

We evaluate the performance of the trained model on adversarial robustness using adversarial test loss, which is defined as

$$\mathcal{L}_{\mathcal{D}}^{\text{adv}}(\mathbf{W}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\max_{\|\boldsymbol{\zeta}\|_p \leq \bar{\zeta}} \mathcal{L}(\mathbf{W}, \mathbf{x}+\boldsymbol{\zeta}, y)\right]. \tag{14}$$

We give our main result on adversarial test loss in the following theorem.

Theorem 3.9. *Under Condition 3.1 and Assumption 3.2, with probability at least* $1-\exp(-\tilde{\Omega}(d))$*, for any* $i \in \{1,2\}, j \in \{maj, min\}$*, the adversarial test loss of a DP-SGD trained model with learning rate* $\eta = \Theta(1)$ *satisfies*

$$\mathcal{L}^{\text{adv}}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(T)}\right) \leq \bar{L}_{i,j}(\mathbf{W}^{(T)}) + \underbrace{\mathcal{O}\left(\left[\frac{T}{m}C + \frac{\sqrt{Td}}{m}\sigma_n + \sqrt{d}\sigma_0\right]\bar{\zeta}d^{1-1/p}\right)}_{\text{By adversarial perturbation}}. \tag{15}$$

We defer the proof of Theorem 3.9 to Appendix G.4.

Remark 3.10. Theorem 3.9 shows that the error induced by adversarial perturbation increases at a rate of $\mathcal{O}(T)$ and contains a DP noise term $\sqrt{Td}/m\sigma_n$, which increases at a rate of $\mathcal{O}(\sqrt{T})$. Moreover, the test loss bound aligns with the excess risk bound of adversarial training (Xiao et al., 2022).

4 Understanding DP-SGD Impacts

In Sections 3, we establish the test loss bounds (Theorems 3.5 and 3.9) for a DP-SGD trained model. In this section, we extend the results to interpret the side effects and the effectiveness of pre-training. We extend the upper bound in Theorem 3.5, and we can also get a similar lower bound by extending the lower bound in Theorem 3.6 to obtain similar insights.

4.1 Interpretation of Bad Features (Performance)

Theorem 3.5 implies that the learned features are perturbed by the noise introduced during DP-SGD training. This noise prevents CNNs from learning perfect features, ultimately resulting in a non-vanishing error. Be specific, this loss increases when the feature is seriously affected by the DP noise, i.e., a smaller $\text{FNR}^{\text{DP}}_{i,j}$ (see Definition 3.3).

4.2 Interpretation of Disparate Impact

Define the distribution of class i as $\mathbb{P}_{\mathcal{D}_i}[(\mathbf{x},y)] = \mathbb{P}_{\mathcal{D}}[(\mathbf{x},y)|y=i]\,,\ i \in \{1,2\}$. We evaluate model performance on different classes by bounding the test loss on data distribution within a class.

Corollary 4.1 (Disparate impact of different classes)**.** *Under Condition 3.1 and Assumption 3.2, with probability at least* $1-\exp(-\tilde{\Omega}(d))$*, for any* $i \in \{1,2\}$*, the test loss of a DP-SGD trained model with learning rate* $\eta = \Theta(1)$ *satisfies*

$$\mathcal{L}_{\mathcal{D}_i}\left(\mathbf{W}^{(T)}\right) \leq \frac{1}{\sum_{j\in\{\text{min,maj}\}}\gamma_{i,j}} \sum_{j\in\{\text{min,maj}\}} \gamma_{i,j}\cdot \bar{L}_{i,j}\left(\mathbf{W}^{(T)}\right). \tag{16}$$

Similarly, define the distribution of group j as $\mathcal{D}_j, j \in \{\text{maj}, \text{min}\}$. We evaluate model performance across different groups (majority and minority) by the test loss on data distribution within a group.

Corollary 4.2 (Disparate impact of subpopulation groups)**.** *Under Condition 3.1 and Assumption 3.2, with probability at least* $1-\exp(-\tilde{\Omega}(d))$*, for any* $j \in \{\text{maj}, \text{min}\}$*, the test loss of a DP-SGD trained model with learning rate* $\eta = \Theta(1)$ *satisfies*

$$\mathcal{L}_{\mathcal{D}_j}\left(\mathbf{W}^{(T)}\right) \leq \frac{1}{\sum_{i=1}^{2}\gamma_{i,j}} \sum_{i=1}^{2} \gamma_{i,j}\cdot \bar{L}_{i,j}\left(\mathbf{W}^{(T)}\right). \tag{17}$$

Recall the expression of $\bar{L}$ in (12). Corollaries 4.1 uncovers three sources of disparate impact: *gradient clipping* $\Lambda_{i,j}$, *data imbalance* $\gamma_{i,j}$, and *feature disparity* $\|\mathbf{u}_{i,j}\|_2$, which correspond to three key components of FNRs and clipping factors (see Definition 3.3 and 3.4). We discuss them in detail below.

Gradient clipping. Theorem 3.5 shows that the test loss depends on the clipping factor $\Lambda_{i,j}$. This implies that a class or group with a larger gradient norm will experience more aggressive clipping, leading to poorer feature learning performance.

Data imbalance. As the privacy protection error terms in Theorems 3.5 and 3.6 decrease with the data proportion $\gamma_{i,j}$, a group or a class with more data enjoys better performance and vice versa. This raises a risk of worse model performance on the skewed data sources[2] and the long-tailed distributed applications (Feldman & Zhang, 2020).[3]

Feature disparity. The feature-to-noise ratio $\mathcal{F}_{i,j}$ (see Definition 3.3) depends on the feature sizes of the data $\mathbf{u}_{i,j}$. In real-world applications, data from different classes or groups may have significantly different feature sizes, resulting in divergent model performances.

Remark 4.3. Recent papers, e.g., (Esipova et al., 2022) have pointed out that the disparate impact mainly arises from gradient clipping. However, we show that even without gradient clipping, different groups still exhibit different performances due to data imbalance and the intrinsic differences in feature sizes.

4.3 Interpretation of Adversarial Robustness

As shown in Theorem 3.9, DP-SGD leads to a high adversarial test loss. We interpret worse adversarial robustness from two perspectives.

Feature learning. As pointed out in Allen-Zhu & Li (2020), an adversarially robust model typically removes the non-robust class-irrelevant noises and learns robust features. However, DP-SGD injects much noise during the training process, making the neural networks inevitably learn non-robust class-irrelevant noises.

Network parameter growth. Due to the noise added by DP-SGD, the network parameters have a consistently growing norm with the number of iterations. As adversarial perturbation $\boldsymbol{\zeta}$ attacks the model by changing the neurons' activated inner products, i.e.,

$$F_k(\mathbf{W}, \mathbf{x}+\boldsymbol{\zeta}) = \frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left[\sigma\left(\left\langle \mathbf{w}_{k,r}, \mathbf{x}^{(j)}\right\rangle + \underbrace{\left\langle \mathbf{w}_{k,r}, \boldsymbol{\zeta}^{(j)}\right\rangle}_{\text{By adversarial perturbation}}\right)\right], \tag{18}$$

higher network parameter norms result in increased vulnerability to adversarial attacks.

4.4 Pre-training and fine-tuning

As demonstrated in Berrada et al. (2023), pre-training on public datasets can significantly reduce the accuracy drop and mitigate the side effects caused by DP-SGD. With pre-training, the neural network utilizes good pre-trained features and the loss at initialization $\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(0)})$ is relatively small. As a result, fine-tuning only needs a small number of iterations, leading to smaller privacy protection errors and thus, mild side effects. In Tramèr et al. (2022), the authors pointed out that many of the existing pre-training and fine-tuning datasets satisfy the private fine-tuning data distribution is essentially a subset of the public data distribution (e.g. pre-training on ImageNet and fine-tuning on CIFAR-10). So, in this subsection, we explore how the distribution shifts affect the private fine-tuning performance.

Consider a two-layer CNN model (defined in Section 2) that is trained on a simplified data distribution $\mathcal{D}_{\text{pt}}$ that each class only has one feature with the same magnitude, i.e., $\mathbf{u}_1$ for class 1, and $\mathbf{u}_2$ for class 2

[2]Take ImageNet (Deng et al., 2009) as an example. More than 45% of ImageNet data comes from the United States, corresponding to only 4% of the world's population; In contrast, China and India contribute just 3% of ImageNet data (Zou & Schiebinger, 2018). Thus, DP-SGD trained models on ImageNet may perform poorly on tasks concerning China and India.

[3]For example, in the SUN dataset (Xiao et al., 2010), the number of examples in each class displays a long-tailed structure (Feldman, 2020).

and $\|\mathbf{u}_1\|_2 = \|\mathbf{u}_2\|_2$, with SGD and fine-tuned on similar data distribution $\mathcal{D}_{\text{ft}}$ (data is first generated from $\mathcal{D}_{\text{pt}}$ and then rotated with an angle θ) (Details are discussed in Appendix B). Based on the results in Kou et al. (2023), a ReLU CNN trained with gradient descent learns the feature in a constant order. Therefore, we assume that the pre-trained trained model is $\tilde{\mathbf{w}}_{j,r} = C_1 \cdot \mathbf{u}_j + C_3 \cdot \mathcal{N}(0, \sigma_p \mathbf{H})$ for simplicity. Then we can characterize the fine-tuning test loss as follows (We define Λ_i in a similar manner that $\Lambda_i = C/\|\mathbf{u}_i\|_2 + \sigma_p\sqrt{d}$).

Proposition 4.4. *Suppose, then the fine-tuned test loss satisfies*

$$\mathcal{L}_{\mathcal{D}_{\text{ft}}}(\mathbf{W}^{(T)}) \leq \exp\left(-\Omega\left(\frac{\Lambda_i \|\mathbf{u}_i\|_2^2}{\sqrt{m}} T\right)\right) \cdot \tilde{L} + \mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{mn}\Lambda_i}\right) + \mathcal{O}\left(\frac{\sqrt{md}\sigma_n}{\Lambda_i \|\mathbf{u}_i\|_2}\right), \tag{19}$$

where

$$\begin{aligned} \tilde{L} = &-\frac{1}{2}\ln\left(\frac{\exp(C_1 \cos\theta \|\mathbf{u}_1\|_2^2)}{\exp(C_1 \cos\theta \|\mathbf{u}_1\|_2^2) + \exp(C_1 \sin\theta \|\mathbf{u}_1\|_2^2 + C_3\sigma_p^2)}\right) \\ &-\frac{1}{2}\ln\left(\frac{\exp(C_1 \cos\theta \|\mathbf{u}_2\|_2^2)}{\exp(C_1 \cos\theta \|\mathbf{u}_2\|_2^2) + \exp(C_3\sigma_p^2)}\right). \end{aligned} \tag{20}$$

Remark 4.5. A worth noting fact is that $\tilde{L}$ increases with θ, meaning that the fine-tuning test loss decreases with increased feature difference, i.e., θ. Even worse, when $\tilde{L} > \mathcal{L}_{\mathcal{D}_2}(\mathbf{W}^{(0)})$, the pre-training can lead to worse performance than training from scratch. Hence, as indicated in Tramèr et al. (2022), the remarkable good performance of pre-training on private learning may come from the distribution overlap between pre-training and fine-tuning datasets.

5 Experiments

In this section, we use synthetic data and real-world datasets MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) to verify our theory.

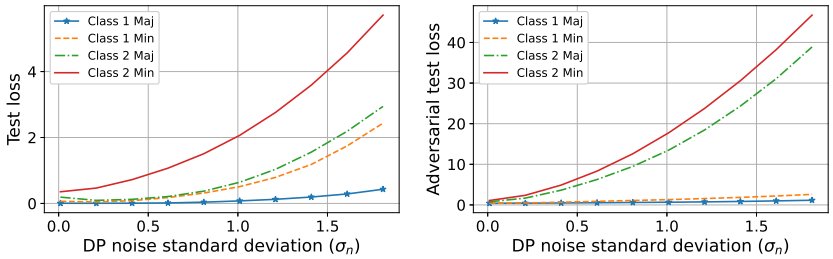
5.1 Synthetic datasets

Synthetic data generation. We generate a synthetic dataset following the data distribution described in Section 2. Specifically, we set the sizes of both training and test datasets to 450. We set the feature vector length in each patch to 100. The feature vector sizes and dataset sizes of the majority and minority groups of classes 0 and 1 are specified as $\|\mathbf{u}_{1,\text{maj}}\|_2 = 4, \gamma_{1,\text{maj}} = 44\%, \|\mathbf{u}_{1,\text{min}}\|_2 = 2, \gamma_{1,\text{min}} = 22\%, \|\mathbf{u}_{2,\text{maj}}\|_2 = 1.5, \gamma_{2,\text{maj}} = 22\%$ and $\|\mathbf{u}_{2,\text{min}}\|_2 = 0.5, \gamma_{2,\text{min}} = 11\%$. The difference in feature vector sizes and dataset sizes characterize feature disparity and data imbalance. In addition, we set the standard deviation of the noise patch to $\sigma_p = 0.2$.

We train a two-layer CNN with ReLU activation function and cross-entropy loss (see Section 2). We set the number of neurons as 64, i.e., $m = 32$. We utilize the default initialization in PyTorch. We train the CNN with DP-SGD with batch size $B = 128$ for 20 epochs.

Test loss. We experiment on the model's test loss of groups with different features in Figure 2 (a). We observe that the test loss generally increases with the DP noise standard deviation, aligning with the findings in Theorem 3.5 where the upper bound of test loss depends on the corresponding FNR. Furthermore, the class and the group with a larger feature size incur smaller test losses. It is also worth noting that the gaps among the groups become more significant with increasing noise standard deviation.

Adversarial robustness. In Figure 2(b), we assess the adversarial robustness by attacking the model with the projected gradient descent method (generate the adversarial examples by maximizing the loss with projected gradient descent) with $\bar{\zeta} = 0.02$. We observe that the DP-SGD trained model experiences a degradation in adversarial robustness on certain groups/classes. This aligns with the results from Theorem 3.5 that the upper bound of the model's adversarial test loss increases with DP noise standard deviation.



(a) Test loss versus DP noise standard deviation σ_n.

(b) Adversarial test loss attacked with the Projected Gradient Descent Method.

Figure 2: Model standard test loss and adversarial test loss.

5.2 Real-world datasets

Setup. For MNIST and CIFAR-10, we train LeNet and a CNN following the architecture in (Tramer & Boneh, 2020) with DP-SGD. We fix the privacy budget as $\alpha = 3, \delta = 10^{-5}$. We fix the gradient clipping threshold as $C = 0.1$ (Tramer & Boneh, 2020). We fix the batch size as 256 and try various learning rates. We use the DP-SGD implementation in Opacus (Yousefpour et al., 2021). We obtain the adversarial example through the projected gradient descent method with $\bar{\zeta} = 4/255$.

Impact of feature size. In real images, it is hard to distinguish which part of the images represents the features. We simply see the object as features and regard the background as noise. To emulate the image backgrounds, we apply zero-padding to the periphery of the input images and the resize the image back to the original size (as shown in Figure 3 in the appendix). The table below shows both the test accuracy and adversarial test accuracy with different padding ratios. As indicated in the theory, the model accuracy decreases with the padding ratios (feature sizes).

Padding ratio	0%	13%	26%	38%	50%	62%	75%
MNIST	97%	97%	97%	96%	95%	94%	86%
CIFAR-10	58%	56%	54%	52%	51%	48%	46%
MNIST (adversarial)	95%	93%	89%	77%	50%	20%	1%
CIFAR-10 (adversarial)	3%	3%	2%	2%	1%	0%	0%

Pre-training vs. fine-tuning. We maintain the training dataset unrotated. We rotate the test images and split them into fine-tuning training dataset and test dataset. We additionally use ResNet-18 (He et al., 2016) for CIFAR-10. The table below shows the fine-tuning test performance under different rotation angles. As indicated in the theory, the model accuracy decreases with the rotation angle.

Rotation angle	0°	22.5°	45°	67.5°	90°
MNIST	99%	97%	94%	94%	94%
CIFAR-10 (CNN Tramer & Boneh (2020))	70%	59%	51%	49%	51%
CIFAR-10 (ResNet-18)	91%	66%	43%	37%	44%

6 Conclusions and Future Works

In this paper, we investigate the side effects of DP-SGD in two-layer ReLU CNNs. We show that the side effects of DP-SGD trained CNNs depend on data's feature, data noise, and privacy-preserving noise. Our results uncover three sources of disparate impact: *gradient clipping*, *data imbalance*, and *feature disparity*. In addition, we show that the privacy-preserving noise introduces randomness into the learned features, leading to bad learned features and worse adversarial robustness. Moreover, we demonstrate that the fine-tuning performance with pre-trained models decreases with increased feature differences in the pre-training and fine-tuning datasets. Numerical results on both synthetic and real-world datasets validate our theoretical analyses.

7 Reproducibility Statement

We put the source code in the Supplementary Material. All the experimental results can be reproduced with the source code.

References

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*, pp. 308–318, 2016.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, 2019.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32:15479–15488, 2019.

Leonard Berrada, Soham De, Judy Hanwen Shen, Jamie Hayes, Robert Stanforth, David Stutz, Pushmeet Kohli, Samuel L Smith, and Borja Balle. Unlocking accuracy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.

Zhiqi Bu, Hua Wang, Zongyu Dai, and Qi Long. On the convergence and calibration of deep learning with differential privacy. *Transactions on machine learning research*, 2023, 2023.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pp. 309–315, 2019.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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.

Maria S Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C Cresswell. Disparate impact in differential privacy from gradient misalignment. *arXiv preprint arXiv:2206.07737*, 2022.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, pp. 6944–6959. PMLR, 2022.

Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2521–2529. PMLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pp. 9965–10040. PMLR, 2022.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pp. 17615–17659. PMLR, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Binghui Li and Yuanzhi Li. Why clean generalization and robust overfitting both happen in adversarial training. *arXiv preprint arXiv:2306.01271*, 2023.

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, 2023.

Lucas Rosenblatt, Julia Stoyanovich, and Christopher Musco. A simple and practical method for reducing the disparate impact of differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning? In *Uncertainty in Artificial Intelligence*, pp. 1738–1748. PMLR, 2022.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *Forty-first International Conference on Machine Learning*, 2022.

Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34:27555–27565, 2021.

Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets. Robustness threats of differential privacy. *arXiv preprint arXiv:2012.07828*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:6000–6010, 2017.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459, 2022.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86:391–423, 2012.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

Hanlin Zhang, Xuechen Li, Prithviraj Sen, Salim Roukos, and Tatsunori Hashimoto. A closer look at the calibration of differentially private learners. *arXiv preprint arXiv:2210.08248*, 2022.

Yuan Zhang and Zhiqi Bu. Differentially private optimizers can learn adversarially robust models. *arXiv preprint arXiv:2211.08942*, 2022.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. *arXiv preprint arXiv:2303.08433*, 2023.

James Zou and Londa Schiebinger. Design ai so that it's fair. *Nature*, 559(7714):324–326, 2018.

A Additional Related Work

Side effects in differentially private learning. Side effects of DP have been widely studied in deep learning literature. *Disparate impact* was initially observed in classification tasks Bagdasaryan et al. (2019) and generative tasks (Ganev et al., 2022). Then, researchers have proposed several methods, such as a regularization approach (Tran et al., 2021), re-weighting, and stratification methods (Esipova et al., 2022; Rosenblatt et al., 2024) to mitigate the disparate impacts. A recent paper (Berrada et al., 2023) showed that the DP models pre-trained with large datasets and fine-tuned with large batch size can have marginal disparate effects. We discuss the implication of this work to pre-training in Appendix **??**.

The interplay between DP and *fairness* is also a widely studied topic. Cummings et al. (2019) showed that exact fairness is not compatible with DP under the PAC learning setting. Sanyal et al. (2022) showed that it is not possible to build accurate learning algorithms that are both private and fair when data follows a specific kind of long-tailed distribution. Mangold et al. (2023) bounded the difference in fairness levels between private and non-private models under the assumption that the confidence margin is Lipschitz-continuous.

Some other side effects have also been studied. Tramer & Boneh (2020) showed that Differentially Private Learning (DPL) may perform worse after *bad feature learning* compared with learning handcraft features. Tursynbek et al. (2020) studied *adversarial robustness* in DPL and showed that models trained by DPL may be more vulnerable compared with non-private models. Zhang & Bu (2022) studied the adversarial robustness of private linear classifiers and showed differentially privately fine-tuned pre-trained models may be robust under certain parameter settings. In addition, Zhang et al. (2022) studied calibration of DPL and observed *miscalibration* across a wide range of vision and language tasks. Bu et al. (2023) studied DPL with neural tangent kernel (NTK) and demonstrated that a large clipping threshold may benefit the calibration of DPL.

In these works, researchers have studied the convergence of DPL and explored explanations for the aforementioned side effects. However, these analyses relied on some restricted assumptions that are not applicable to neural networks because (1) differentially private neural networks training is not in the NTK regime as the noise keeps the network parameter far away from the initialization during training; (2) training loss of ReLU neural networks is non-convex and non-smooth, contradicting the assumptions in most analyses. In this work, we aim to overcome these challenges and explain the side effects of DP-SGD on a two-layer ReLU CNN.

B Details about pre-training and fine-tuning data distributions

To illustrate the impact of pre-training, we simplify the data distribution as follows. We consider the following pre-training and fine-tuning data distributions. We control their difference by a parameter θ.

Pre-training data distribution. We consider a 2-class classification problem over 2-patch inputs. Each labelled data is denoted as $(\mathbf{x}, y)$, with label $y \in \{1, 2\}$ and data vector $\mathbf{x} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) \in \mathbb{R}^{d\times 2}$. A sample $(\mathbf{x}, y)$ is generated from a data distribution $\mathcal{D}$ as follows.

1. The label y is randomly sampled from $\{1, 2\}$. With probability $1/2$, the label is selected as $y = 1$; otherwise, it is selected as $y = 2$.
2. Each input data patch $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathbb{R}^d$ contains either feature or noise.
   - Feature patch: One data patch ($\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$) is randomly selected as the feature patch. This patch contains a feature $\mathbf{u}_y$ for $y \in \{1, 2\}$.
   - Noisy patch: The remaining patch $\boldsymbol{\xi}$ is generated from a Gaussian distribution $\mathcal{N}(0, \sigma_p^2 \mathbf{H})$, where $\mathbf{H} = \mathbf{I} - \sum_{i=1}^{2} \mathbf{u}_i \mathbf{u}_i^\top \cdot \|\mathbf{u}_i\|_2^{-2}$.

Fine-tuning data distribution. We consider a 2-class classification problem over 2-patch inputs. Each labelled data is denoted as $(\mathbf{x}, y)$, with label $y \in \{1, 2\}$ and data vector $\mathbf{x} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) \in \mathbb{R}^{d\times 2}$. A sample $(\mathbf{x}, y)$ is generated from a data distribution $\mathcal{D}$ as follows.

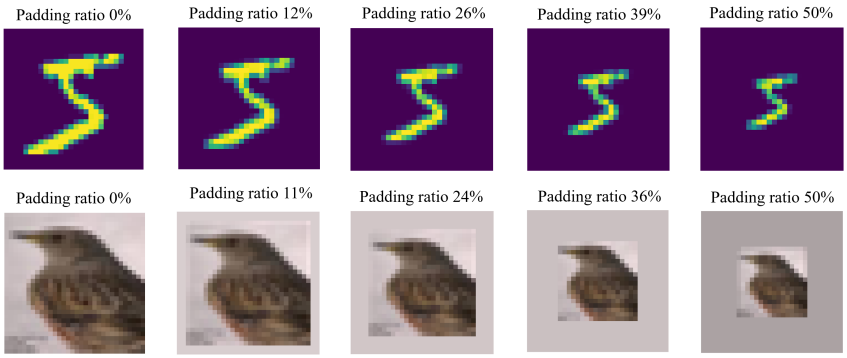


Figure 3: Examples on using image padding to control feature sizes. For digits and objects, we pad the images with their background color.

1. The label y is randomly sampled from $\{1, 2\}$. With probability $1/2$, the label is selected as $y = 1$; otherwise, it is selected as $y = 2$.
2. Each input data patch $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathbb{R}^d$ contains either feature or noise.
   - Feature patch: One data patch ($\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$) is randomly selected as the feature patch. For $y = 1$, this patch contains a feature $\mathbf{u}_1' = \cos\theta\mathbf{u}_1 + \sin\theta\mathbf{u}_2$. For $y = 2$, this this patch contains a feature $\mathbf{u}_2' = \cos\theta\mathbf{u}_2 - \sin\theta\mathbf{u}_1$.
   - Noisy patch: The remaining patch $\boldsymbol{\xi}$ is generated from a Gaussian distribution $\mathcal{N}(0, \sigma_p^2\mathbf{H})$, where $\mathbf{H} = \mathbf{I} - \sum_{i=1}^{2} \mathbf{u}_i'(\mathbf{u}_i')^\top \cdot \|\mathbf{u}_i'\|_2^{-2}$.

C DP-SGD Algorithm

Input: training set $\mathcal{S}$, learning rate η, DP noise standard deviation σ_n, batch size B
initialize $\mathbf{W}^{(0)}$ randomly
for each round $t = 1, 2, \cdots, T$ **do**
  Take a random training subset $\mathcal{S}^{(t)}$ uniformly from $\mathcal{S}$ with probability $\frac{B}{n}$
  Compute $\mathbf{g}^{(t)} = \frac{1}{B}\sum_{(\mathbf{x},y)\in\mathcal{S}^{(t)}} \nabla\mathcal{L}\left(\mathbf{W}^{(t-1)}, \mathbf{x}, y\right) + \mathcal{N}\left(0, \sigma_n^2\mathbf{I}\right)$
  $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \eta\mathbf{g}^{(t)}$
end for

Algorithm 1: DP-SGD

D Details about experiments

In this section, we introduce the experimental details.

D.1 Visualization of padding images

In Figure 3, we present examples of padding images.

D.2 Experimental details of Figure 1

Because we mainly characterize the privacy-utility tradeoff, we simply set 2 classes with equal feature sizes and equal dataset sizes 100. We vary the feature size from 0 to 21 and the DP-SGD noise standard deviation σ_n from 0 to 4. We set the noise patch standard deviation σ_p as 0.02. We set the clipping threshold as $C = 2$. We set the number of neurons as $m = 32$.

E SOME DATASETS AND NEURAL NETWORK ARCHITECTURES OBSERVING SIDE EFFECTS

Bagdasaryan et al. (2019) has identified disparate effects in datasets such as MNIST, CelebA, and Twitter posts using ResNet and LSTM networks. Unfairness has been observed in CelebA and CIFAR-10 datasets using ResNet networks (Sanyal et al., 2022). Zhang et al. (2022) observed miscalibration in QNLI, QQP, SST-2 with fine-tuned RoBERTa-base.

F OVERVIEW OF CHALLENGES AND PROOF SKETCH

In this section, we outline the main challenges in studying feature learning of DP-SGD on CNNs and the key proof techniques employed to overcome the challenges.

F.1 CHALLENGE 1: NON-SMOOTHNESS OF THE RELU ACTIVATION FUNCTION

The first challenge arises from the non-smoothness of the ReLU activation function. Some existing papers (e.g., (Girgis et al., 2021)) analyzed DP-SGD with Lipschitz-smoothness-based approaches. However, this kind of approach is not applicable to ReLU neural networks.

In a two-layer CNNs, we can track the neurons' feature learning process through their gradients. For any $i \in \{1, 2\}, r \in [m]$, the gradient on $\mathbf{w}_{i,r}^{(t)}$ can be decomposed as

$$\nabla_{\mathbf{w}_{i,r}^{(t)}} \mathcal{L}_{\mathcal{S}}(\mathbf{W}^{(t)}) = \underbrace{\sum_{j \in \{\text{maj,min}\}} (\mu_{i,j}\mathbf{u}_{i,j} - \mu_{3-i,j}\mathbf{u}_{3-i,j})}_{\text{Data features}} + \underbrace{\sum_{k=1}^{n} \boldsymbol{\xi}_k (\bar{\rho}_{i,j,k}\mathbb{I}(y_k = i) - \underline{\rho}_{i,j,k}\mathbb{I}(y_k \neq i))}_{\text{Data noise}},$$

where $\mu_{i,j}, \mu_{3-i,j}, \bar{\rho}_{i,j,k}, \underline{\rho}_{i,j,k} \geq 0$ are constants. The neurons tend to learn both class-relevant data features and data noise of the targeted class while unlearning others.

Some existing approaches (e.g., (Cao et al., 2022)) bound the feature learning process by characterizing the leading neurons that learn the most features. However, this approach only works for ReLU^q ($q > 2$) activation functions, where the leading neurons dominate other neurons during training. As the ReLU function is piece-wise linear, these approaches fail in ReLU neural networks.

To overcome this challenge, we study the feature learning process by analyzing the dynamics of the model outputs $F_i(\mathbf{W}, \mathbf{x}), i \in [2]$ defined in Section 2

F.2 CHALLENGE 2: RANDOMNESS FROM DP-SGD

The second challenge stems from the randomness introduced by DP-SGD. The learning process is significantly perturbed due to the random noise in DP-SGD. Kou et al. (2023) attempted to bound the feature learning process based on the monotonicity of the weights of feature vectors. However, due to the randomness from DP-SGD, the weights are not consistently increasing.

To address this challenge, we track the increments of the model outputs for any data point $(\mathbf{x}, y) \sim \mathcal{D}_{i,j}$ instead. The key proposition is presented as follows.

Proposition F.1. *For any $(\mathbf{x}, y) \sim \mathcal{D}_{i,j}, i \in [2], j \in \{maj, min\}$, with probability at least $1 - \exp(-\tilde{\Omega}(d))$,*

- *The increment of model output for the targeted class y satisfies*

$$\begin{aligned}\Delta_y^{(t)}(\mathbf{x}) =& F_y\left(\mathbf{W}^{(t+1)}, \mathbf{x}\right) - F_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \\ \geq& \Omega\left(\frac{\eta}{\sqrt{m}} \cdot \gamma_{i,j} \cdot \Lambda_{i,j} \cdot \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{i,j}}\left[1 - \text{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right]\right) \cdot \|\mathbf{u}_{i,j}\|_2^2 \\ & - \mathcal{O}\left(\frac{\eta}{m}\|\mathbf{u}_{i,j}\|_2^2 + \eta\sigma_n\sqrt{d}\|\mathbf{u}_{i,j}\|_2 + \frac{\eta}{m\sqrt{n}}d\sigma_p^2 + \eta d\sigma_n\sigma_p\right).\end{aligned} \tag{21}$$

- *The increment of model output for the other class* $3-y$ *satisfies*

$$\begin{aligned}\Delta_{3-y}^{(t)}(\mathbf{x}) =& F_{3-y}\left(\mathbf{W}^{(t+1)}, \mathbf{x}\right) - F_{3-y}\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \\ \leq& \mathcal{O}\left(\eta\sigma_n\sqrt{d}\left\|\mathbf{u}_{i,j}\right\|_2 + \frac{\eta}{m\sqrt{n}}d\sigma_p^2 + \eta d\sigma_n\sigma_p\right).\end{aligned} \tag{22}$$

Proposition F.1 shows that the model output on the data $(\mathbf{x}, y)$ with respect to the targeted class, $F_y(\mathbf{W}, \mathbf{x})$, tends to increase over iterations (see term 1 in the RHS of (21)). However, due to the noise perturbation introduced by DP-SGD, the model increment $\Delta_y^{(t)}$ becomes smaller and cannot always be positive (because of the second negative term in RHS of (21)). In addition, the model output on the data $(\mathbf{x}, y)$ with respect to the other class may increase due to the randomness from batches and DP-SGD. The results of Proposition F.3 allow us to track the test loss increments, as shown in the following subsection.

F.3 Challenge 3: Non-Linearity of Cross-Entropy and Softmax Functions

Due to the non-linearity of cross-entropy and softmax functions, the model output increment bounds in Proposition F.1 cannot be directly applied to bounding test loss.

To tackle this challenge, we bound the non-linear functions with a piece-wise linear function, as stated in Lemma F.2.

Lemma F.2. *Under Assumption 3.2, we have*

$$\mathcal{L}\left(\mathbf{W}^{(t+1)}, \mathbf{x}, y\right) - \mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{x}, y\right) \leq c_1 \cdot \sigma\left(\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right) - c_2 \cdot \sigma\left(\Delta_y^{(t)}(\mathbf{x}) - \Delta_{3-y}^{(t)}(\mathbf{x})\right),$$

for some constants $c_1, c_2 > 0$.

Lemma F.2 allows us to apply the model output increment bounds in Proposition F.1 to bound the test loss, as shown in Proposition F.3.

Proposition F.3. *Under Condition 3.1 and Assumption 3.2, with probability at least* $1-\exp\left(-\tilde{\Omega}(d)\right)$, *for any* $i \in [2], j \in \{maj, min\}$, *we have*

$$\begin{aligned}&\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t+1)}\right) - \mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right) \leq -\Omega\left(\frac{\eta}{\sqrt{m}} \cdot \gamma_{i,j} \cdot \Lambda_{i,j} \cdot \left\|\mathbf{u}_{i,j}\right\|_2^2\right) \cdot \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{i,j}}\left[1 - \text{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right] \\ &\underbrace{+ \mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2 + \eta\sigma_n\sqrt{d}\left\|\mathbf{u}_{i,j}\right\|_2 + \frac{\eta}{m\sqrt{n}}d\sigma_p^2 + \eta d\sigma_n\sigma_p\right)}_{:=\phi}.\end{aligned}$$

With the fact that under Assumption 3.2,

$$1 - \text{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right) = \Theta(1) \cdot \mathcal{L}\left(\left(\mathbf{W}^{(t)}\right), \mathbf{x}, y\right) \tag{23}$$

holds, Proposition F.3 can be applied to establish the following test loss bound

$$\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t+1)}\right) \leq \left(1 - \Omega\left(\frac{\eta}{\sqrt{m}} \cdot \gamma_{i,j} \cdot \Lambda_{i,j} \cdot \left\|\mathbf{u}_{i,j}\right\|_2^2\right)\right) \cdot \mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right) + \phi. \tag{24}$$

Recursively applying (24) over T iterations yields Theorem 3.5.

G Proof

G.1 Preliminaries

Lemma G.1 (Gaussian distribution tail bound)**.** *A variable* x *following* $\mathcal{N}(0, \sigma_0^2)$ *satisfies*

$$\mathbb{P}[x \geq t\sigma_0], \mathbb{P}[x \leq -t\sigma_0] \leq \exp\left(-\frac{t^2}{2}\right), \forall t \geq 0. \tag{25}$$

Lemma G.2 (Chi-squared distribution tail bound)**.** *A variable* $\mathbf{x} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ *satisfies With probability at least* $1 - \exp(-td/10)$*, we have*

$$\mathbb{P}\left[\|\mathbf{x}\|_2 \geq \sigma_0 \sqrt{2(t+1)d}\right] \leq \exp(-(t+1)d/10), \forall t \geq 0. \tag{26}$$

Lemma G.3 (Half-normal distribution concentration bound)**.** *Suppose* $x_1, x_2, \cdots, x_n \sim \mathcal{N}(0, \sigma_0^2)$*. Then,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \geq t\sigma_0\right], \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \leq -t\sigma_0\right] \leq \exp\left(-\frac{t^2}{2}\right), \forall t \geq 0. \tag{27}$$

Proof. First, half-normal variables $|x_i|, \forall i \in [n]$ are sub-Gaussian as a half-normal variable has a negative tail bounded by $-\sqrt{2/\pi}$ and a Gaussian delay positive tail. Then, by Hoeffding's inequality, we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \geq t\sigma_0\right], \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \leq -t\sigma_0\right] \leq \exp\left(-\frac{t^2}{2}\right), \forall t \geq 0. \tag{28}$$

□

Lemma G.4. *Let* $x_1, \cdots, x_m$ *be* m *independent zero-mean Gaussian variables. Denote* z_i *as indicators for signs of* x_i*, i.e., for all* $i \in [m]$*,*

$$z_i = \begin{cases} 1, & x_i > 0, \\ 0, & x_i \leq 0. \end{cases} \tag{29}$$

Then, we have

$$\mathbb{P}\left[\sum_{i=1}^{m} z_i \geq \frac{m}{4}\right] \geq 1 - \exp\left(-2m\right). \tag{30}$$

Proof. Because $z_i, i \in [m]$ are bounded in $[0, 1]$, $z_i, i \in [m]$ are sub-Gaussian variables. By Hoeffding's inequality, we have

$$\mathbb{P}\left[m \cdot \left(\frac{1}{m}\sum_{i=1}^{m} z_i\right) \leq m \cdot \left(\frac{1}{2} - \epsilon\right)\right] \leq \exp\left(\frac{2m^2\epsilon^2}{m(1/16)}\right). \tag{31}$$

Let $\epsilon = 1/4$, we have

$$\mathbb{P}\left[\sum_{i=1}^{m} z_i \leq \frac{m}{4}\right] \leq \exp\left(-2m\right). \tag{32}$$

Therefore, we have

$$\mathbb{P}\left[\sum_{i=1}^{m} z_i \geq \frac{m}{4}\right] \geq 1 - \exp\left(-2m\right). \tag{33}$$

This completes the proof. □

Lemma G.5. *Let* x_1 *be a Gaussian variable following* $\mathcal{N}(0, \sigma_1)$ *and* x_2 *be a Gaussian variable following* $\mathcal{N}(0, \sigma_2)$*. Then, with probability at least* $1 - \exp(t_1^2/2) - \exp(t_2^2/2)$*, we have*

$$\langle x_1, x_2 \rangle \leq t_1 t_2 \sigma_1 \sigma_2. \tag{34}$$

Lemma G.5 can be proved by using Lemma G.1.

Lemma G.6. *For* N *Independent and Identically Distributed (IID) random variables* $x_1, \cdots, x_N \in [0, 1]$ *with expectation* μ*, we have*

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} x_i - \mu \leq \sqrt{\frac{t}{N}}\right] \geq 1 - \exp\left(-2t\right) \tag{35}$$

with $t > 0$*.*

Lemma G.6 can be proved by Hoeffding's inequality.

Lemma G.7. *For any constant $t \in (0, 1]$ and $x \in [-a, b], a, b > 0$, we have*

$$\log(1 + t \cdot (\exp(x) - 1)) \leq \Gamma(x)x, \tag{36}$$

where $\Gamma(x) = \mathbb{I}(x \geq 0) + [\log(1 + t \cdot (\exp(-a) - 1))/ - a] \cdot \mathbb{I}(x < 0)$.

Proof. First, considering $x \geq 0$, we have

$$\frac{\partial \log(1 + t \cdot (\exp(x) - 1))}{\partial t} = \frac{\exp(x) - 1}{1 + t \cdot (\exp(x) - 1)} \geq 0. \tag{37}$$

Thus, $\log(1 + t \cdot (\exp(x) - 1)) \leq x, \forall x > 0$. Second, considering $x < 0$, we have

$$\frac{\partial^2 \log(1 + tv(\exp(x) - 1))}{\partial x^2} = \frac{(1 - t)t \exp(x)}{[1 + t(\exp(x) - 1)]^2} \geq 0. \tag{38}$$

So $\log(1 + t(\exp(x) - 1))$ is a convex function of x. We can conclude that

$$\log(1 + t(\exp(x) - 1)) \leq \frac{\log(1 + t(\exp(-a) - 1))}{-a} x, \forall x < 0. \tag{39}$$

This completes the proof. □

Lemma G.8. *For $x \in [x_0, 1]$ and $x_0 > 0$, we have*

$$1 - x \geq \frac{1 - x_0}{-\log(x_0)} \cdot (-\log(x)). \tag{40}$$

Lemma G.8 can be proved by applying the convexity of $-\log(x)$.

Lemma G.9. *For a geometric sequence defined as $z_{t+1} = \beta z_t$ for a constant $\beta < 1$, we have*

$$\sum_{t=1}^{T} z_t = \frac{1 - \beta^T}{1 - \beta} \cdot z_1. \tag{41}$$

Lemma G.9 is obtained from the property of Geometric sequences.

Lemma G.10. *For $x \leq x_0$, we have*

$$\frac{1}{\exp(x) + 1} \geq \frac{1}{\exp(\bar{x}_0) + 1} - \frac{\exp(\bar{x}_0)}{(\exp(\bar{x}_0) + 1)^2} \cdot (x - \bar{x}_0), \tag{42}$$

where $\bar{x}_0 = |x_0|$.

Lemma G.10 can be prove by the monotonicity and convexity of function $f(x) = 1/\exp(x) + 1$ with $x > 0$.

Lemma G.11. *In each iteration t, with probability at least $1 - \exp(-\Omega(d))$, for any $(\mathbf{x}, y) \in \mathcal{D}_{i,j}$, we have*

$$\left\| \nabla_{\mathbf{W}^{(t)}} \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) \right\|_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{m}} \cdot \left(\|\mathbf{u}_{i,j}\|_2 + \sigma_p \sqrt{d} \right) \right). \tag{43}$$

Lemma G.11 follows from Lemma G.2.

Lemma G.12. *For a variable $x \in [a, b] (a < 0, b > 0)$, the function $f(x) = \log(1 + x)$ satisfies*

$$f(x) \geq \frac{\log(1 + b)}{b} x \cdot \mathbb{I}(x \geq 0) + \frac{\log(1 + a)}{-a} x \cdot \mathbb{I}(x < 0). \tag{44}$$

Lemma G.12 can be proved by the monotonicity and concavity of the $\log(\cdot)$ function.

Lemma G.13. *For any $(\mathbf{x}, y) \sim \mathcal{D}$, With probability at least $1 - 1/d$, we have*

$$\frac{\sigma_p^2 d}{2} \leq \|\boldsymbol{\xi}\|_2 \leq \frac{3\sigma_p^2 d}{2} \tag{45}$$

Proof. By Bernstein's inequality, with probability $1 - 1/d$, we have

$$| \|\boldsymbol{\xi}\|_2 - \sigma_p^2 (d - 2)| = \mathcal{O}(\sigma_p^2 \sqrt{d \log(2d)}). \tag{46}$$

As $d \geq 50$, we have

$$\frac{\sigma_p^2 d}{2} \leq \|\boldsymbol{\xi}\|_2 \leq \frac{3\sigma_p^2 d}{2}, \tag{47}$$

with probability $1 - 1/d$. □

G.2 Proof of Theorem 3.5

In this subsection, we will prove Theorem 3.5. For convenience, we first define the clipping multiplier of data $(\mathbf{x}, y)$ as

$$h(C, \mathbf{x}, y) = \frac{1}{\max\left\{1, \left\|\nabla \mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{x}, y\right)\right\|_2 / C\right\}}. \tag{48}$$

Then, we compute the gradient of the neural networks and prove a bound for it.

G.2.1 Network Gradient

The stochastic gradient on $\mathbf{w}_{q,r}, q \in \{1, 2\}$ at iteration t is

$$\begin{aligned}
&\nabla_{\mathbf{w}_{q,r}^{(t)}} \mathcal{L}_{\mathcal{S}}(\mathbf{W}^{(t)}) \\
=& -\frac{1}{mB} \cdot \sum_{(\mathbf{x},y)\in\mathcal{S}^{(t)}} \left[\mathbb{I}\left(y = q\right) \cdot \left(1 - \text{prob}_q(\mathbf{W}^{(t)}, \mathbf{x})\right) \cdot \sum_{j=1}^{2} \sigma'\left(\left\langle \mathbf{w}_{q,r}^{(t)}, \mathbf{x}^{(j)} \right\rangle\right) \cdot \mathbf{x}^{(j)}\right] \\
&+ \frac{1}{mB} \cdot \sum_{(\mathbf{x},y)\in\mathcal{S}^{(t)}} \left[\mathbb{I}\left(y \neq q\right) \cdot \text{prob}_q(\mathbf{W}^{(t)}, \mathbf{x}) \cdot \sum_{j=1}^{2} \sigma'\left(\left\langle \mathbf{w}_{q,r}^{(t)}, \mathbf{x}^{(j)} \right\rangle\right) \cdot \mathbf{x}^{(j)}\right].
\end{aligned} \tag{49}$$

G.2.2 Bound of the Clipping Multiplier $h(C, \mathbf{x}, y)$

By definition (48), we know that

$$h(C, \mathbf{x}, y) \leq 1. \tag{50}$$

In addition, from Lemma G.11, we know that with probability at least $1 - \exp(\Omega(d))$,

$$h(C, \mathbf{x}, y) \geq \Omega\left(\frac{C\sqrt{m}}{\|\mathbf{u}_{i,j}\|_2 + \sigma_p\sqrt{d}}\right). \tag{51}$$

G.2.3 Loss Increment

For any data $(\mathbf{x}, y) \sim \mathcal{D}_{i,j}, i \in \{1, 2\}, j \in \{\text{Maj}, \text{Min}\}$, with some rearrangement, we can express the increment of the loss as

$$\begin{aligned}
&\mathcal{L}(\mathbf{W}^{(t+1)}, \mathbf{x}, y) - \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) = -\log\left(\text{prob}_y\left(\mathbf{W}^{(t+1)}, \mathbf{x}\right)\right) + \log\left(\text{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right) \\
=& \log\left(1 + \left(1 - \text{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right)\left(\exp\left(\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right) - 1\right)\right),
\end{aligned} \tag{52}$$

where $\Delta_y^{(t)}(\mathbf{x}) = F_y^{(t+1)}(\mathbf{x}) - F_y^{(t)}(\mathbf{x}), \Delta_{3-y}^{(t)}(\mathbf{x}) = F_{3-y}^{(t+1)}(\mathbf{x}) - F_{3-y}^{(t)}(\mathbf{x})$ represent the model output increments at iteration t. As we can see in (52), one key factor that control the loss increment is $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$. We then bound the term it as follows. We first decompose $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$ as following.

$$\begin{aligned}
&\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) \\
=& \underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t+1)}, \mathbf{u}_{i,j}\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\right\rangle\right)\right]}_{A} + \underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t+1)}, \boldsymbol{\xi}\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right]}_{B} \\
&\underbrace{-\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{y,r}^{(t+1)}, \mathbf{u}_{i,j}\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_{i,j}\right\rangle\right)\right]}_{C} \underbrace{-\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{y,r}^{(t+1)}, \boldsymbol{\xi}\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right]}_{D},
\end{aligned} \tag{53}$$

where $\boldsymbol{\xi}$ is the noise patch of a data sample $(\mathbf{x}, y)$ generated from $\mathcal{D}_{i,j}$. We then upper bound A, B and lower bound C, D to find the upper bound of $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x})$.

Here, we prove that $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x})$ is bounded.

Lemma G.14. *For any $(\mathbf{x}, y) \sim \mathcal{D}$, with probability at least $1 - \exp -\tilde{\Omega}(d)$, we have*

$$|\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x})| \leq \mathcal{O}\left(\eta(C + \sqrt{d}\sigma_n)(\max_{i,j} \|\mathbf{u}_{i,j}\|_2 + \sqrt{d}\sigma_p)\right). \tag{54}$$

Proof. With Lemma G.2, we have

$$\begin{aligned} |\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x})| \leq & 2\eta\left(C + \left\|\mathbf{n}^{(t)}\right\|_2\right)\left(\max_{i,j} \|\mathbf{u}_{i,j}\|_2 + \|\boldsymbol{\xi}\|_2\right) \\ \leq & \mathcal{O}\left(\eta(C + \sqrt{d}\sigma_n)(\max_{i,j} \|\mathbf{u}_{i,j}\|_2 + \sqrt{d}\sigma_p)\right), \end{aligned} \tag{55}$$

with probability at least $1 - \exp(-\tilde{\Omega}(d))$. By the learning rate condition in Condition 3.1, we can conclude that $|\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x})|$ is upper bounded by a constant. □

For the term A, with probability at least $1 - \exp(-\tilde{\Omega}(d))$, we have the following inequality,

$$\begin{aligned} A = & \frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\Big\langle \mathbf{w}_{3-y,r}^{(t)} - \frac{\eta}{mB}\cdot \sum_{(\mathbf{x}_k, y_k)\in\mathcal{S}_{i,j}^{(t)}} \sigma'\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\cdot h(C, \mathbf{x}_k, y_k)\cdot \mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)}, \mathbf{x}_k\Big)\cdot \mathbf{u}_{i,j} \\ & + \ \eta\cdot \mathbf{n}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\Bigg] - \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\right\rangle\right)\right] \\ \overset{(a)}{\leq} & \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t)} + \eta\mathbf{n}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\right\rangle\right)\right] - \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\right\rangle\right)\right] \\ \overset{(b)}{\leq} & \frac{1}{m}\sum_{r=1}^{m}\left[\left|\left\langle \eta\mathbf{n}^{(t)}, \mathbf{u}_{i,j}\right\rangle\right|\right] \\ \overset{(c)}{\leq} & \mathcal{O}(\eta\sigma_n\sqrt{d}\,\|\mathbf{u}_{i,j}\|_2), \end{aligned} \tag{56}$$

where (a) is obtained by the monotonicity of ReLU activation function; (b) is because ReLU function is 1-Lipschitz continuous; (c) is due to Lemma G.3.

For the term B, we have that with probability at least $1-\exp(-\tilde{\Omega}(d))$,

$$
\begin{aligned}
B=&\frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\Big\langle \mathbf{w}_{3-y,r}^{(t)}-\frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_y^{(t)}}\sigma'\left(\Big\langle \mathbf{w}_{3-y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\right)\cdot h(C,\mathbf{x}_k,y_k)\cdot \mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\cdot\boldsymbol{\xi}_k\\
&+\frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{3-y}^{(t)}}\sigma'\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\Big)\cdot h(C,\mathbf{x},y)\cdot\Big(1-\mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\cdot\boldsymbol{\xi}_k+\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\Bigg)\Bigg]\\
&-\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\Big\langle \mathbf{w}_{3-y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\right)\right]\\
\overset{(a)}{\leq}&\frac{1}{m}\sum_{r=1}^{m}\left[\left|\left\langle\frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}^{(t)}}\boldsymbol{\xi}_k,\boldsymbol{\xi}\right\rangle\right|+\left|\left\langle\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\right\rangle\right|\right]\\
\overset{(b)}{\leq}&\mathcal{O}\left(\frac{\eta}{m\sqrt{B}}\sqrt{d}\sigma_p+\eta\sqrt{d}\sigma_n\right)\cdot\mathcal{O}\left(\sqrt{d}\sigma_p\right)\\
=&\mathcal{O}\left(\frac{\eta}{m\sqrt{B}}d\sigma_p^2+\eta d\sigma_n\sigma_p\right),
\end{aligned}
\tag{57}
$$

where (a) is because $\sigma'(\cdot)\geq 0$, $\mathrm{prob}_y, \mathrm{prob}_{3-y}\in[0,1]$ and ReLU function is 1-Lipschitz continuous; (b) is because of Cauchy–Schwarz inequality, the property of 1-norm and 2-norm, and Lemma G.3.

For the term C, we have

$$
\begin{aligned}
C=&\frac{1}{m}\sum_{r=1}^{m}\sigma\Bigg(\Big\langle \mathbf{w}_{y,r}^{(t)}+\frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\sigma'\Big(\Big\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\Big)\cdot h(C,\mathbf{x}_k,y_k)\cdot\Big(1-\mathrm{prob}_y\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\cdot\mathbf{u}_{i,j}\\
&+\eta\cdot\mathbf{n}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\Bigg)-\frac{1}{m}\sum_{r=1}^{m}\sigma\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\right)
\end{aligned}
\tag{58}
$$

Based on Lemma G.4, we can conclude that with probability at least $1-\exp(-2m)$, the number of activated neurons at iteration t are at least $m/4$. Then, with probability at least $1-\exp(-\tilde{\Omega}(d))$, we have

$$
\begin{aligned}
C\geq&\frac{1}{m}\sum_{r=1}^{m}\sigma\Bigg(\Big\langle \mathbf{w}_{y,r}^{(t)}+\frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\sigma'\Big(\Big\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\Big)\cdot h(C,\mathbf{x}_k,y_k)\cdot\Big(1-\mathrm{prob}_y\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\\
&\mathbf{u}_{i,j},\mathbf{u}_{i,j}\rangle)-\frac{1}{m}\sum_{r=1}^{m}\left|\Big\langle\eta\cdot\mathbf{n}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\right|-\frac{1}{m}\sum_{r=1}^{m}\sigma\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\right)\\
\geq&\Omega\left(\frac{\eta C}{B\sqrt{m}(\|\mathbf{u}_{i,j}\|_2+\sigma_p\sqrt{d})}\right)\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\Big(1-\mathrm{prob}_y\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\|\mathbf{u}_{i,j}\|_2^2-\mathcal{O}\Big(\eta\sigma_n\sqrt{d}\,\|\mathbf{u}_{i,j}\|_2\Big),
\end{aligned}
\tag{59}
$$

The second inequality is by using the bound of the clipping multiplier.

Therefore, by Lemmas G.3 and G.6, with probability at least $1-\exp(-\Omega(d))-\exp(-\Omega(n))$, we have

$$
\begin{aligned}
C\geq&\Omega\left(\frac{\eta\gamma_{i,j}C\,\|\mathbf{u}_{i,j}\|_2^2}{\sqrt{m}(\|\mathbf{u}_{i,j}\|_2+\sigma_p\sqrt{d})}\right)\mathbb{E}_{(\mathbf{x}_k,y_k)\sim\mathcal{D}_{i,j}}\left[1-\mathrm{prob}_{y_k}\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\right]-\mathcal{O}\left(\frac{\eta}{m}\|\mathbf{u}_{i,j}\|_2^2\right)\\
&-\mathcal{O}\left(\eta\sigma_n\sqrt{d}\,\|\mathbf{u}_{i,j}\|_2\right).
\end{aligned}
\tag{60}
$$

In the following, we prove the bound of D. Similar to the proof of bound of B, we have that with probability at least $1-\exp\left(-\tilde{\Omega}(d)\right)-\exp\left(-\Omega(n)\right)$, we have

$$
\begin{aligned}
D =& \frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\Big\langle \mathbf{w}_{y,r}^{(t)} - \frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{3-y}^{(t)}}\sigma'\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\right)h(C,\mathbf{x}_k,y_k)\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\boldsymbol{\xi}_k \\
&+ \frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{y}^{(t)}}\sigma'\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\right)h(C,\mathbf{x}_k,y_k)\left(1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\right)\boldsymbol{\xi}_k + \mathbf{n}_{y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\Bigg)\Bigg] \\
&- \frac{1}{m}\sum_{r=1}^{m}\sigma\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\right) \\
\geq& -\frac{1}{m}\sum_{r=1}^{m}\left[\left|\left\langle \frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}^{(t)}}\boldsymbol{\xi}_k,\boldsymbol{\xi}\right\rangle\right| + \left|\left\langle \eta\mathbf{n}_{y,r}^{(t)},\boldsymbol{\xi}\right\rangle\right|\right] \\
\geq& -\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2 + \eta d\sigma_n\sigma_p\right).
\end{aligned}
\tag{61}
$$

Substituting bounds of A, B, C, D to (53), we obtain the upper bound of $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$. With probability at least $1-\exp\left(-\tilde{\Omega}(d)\right)$,

$$
\begin{aligned}
&\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) \\
\leq& \underbrace{-\Omega\left(\frac{\eta\gamma_{i,j}C\left\|\mathbf{u}_{i,j}\right\|_2^2}{\sqrt{m}(\left\|\mathbf{u}_{i,j}\right\|_2+\sigma_p\sqrt{d})}\right)\cdot\left[\mathbb{E}_{(\mathbf{x}_k,y_k)\sim\mathcal{D}_{i,j}}\left[1-\mathrm{prob}_{y_k}\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\right]\right]}_{\Phi_1} \\
&\underbrace{+\mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)+\mathcal{O}\left(\eta\sigma_n\sqrt{d}\left\|\mathbf{u}_{i,j}\right\|_2\right)+\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2+\eta d\sigma_n\sigma_p\right)}_{\Phi_2}.
\end{aligned}
\tag{62}
$$

Armed with the loss increment bound (62), we prove the test loss bound of each data $(\mathbf{x},y)\sim\mathcal{D}$ in the next sub section.

G.2.4 Test Loss Bound

Under Assumption 3.2, for any $(\mathbf{x},y)\sim\mathcal{D}$, we have

$$
1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}\right)\geq 1-\exp(-s). \tag{63}
$$

By (52) and Lemma G.7, with probability at least $1-\exp(-\tilde{\Omega}(d))$, we can upper bound the loss increment by a piece-wise linear function,

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t+1)}\right)-\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right) \\
=&\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{i,j}}\left[\log\left(1+\left(1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}\right)\right)\cdot\left(\exp\left(\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})\right)-1\right)\right)\right] \\
\overset{(a)}{\leq}& -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{i,j}}\left[\Gamma\left(\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})\right)\cdot\Phi_1\right]+\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{i,j}}\left[\Gamma\left(\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})\right)\cdot\Phi 2\right]
\end{aligned}
\tag{64}
$$

where (a) is by Lemma G.7 and Lemma (62). Then, substituting (62) to the above inequality yields

$$\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t+1)}\right)-\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right)\overset{(a)}{\leq}-\Omega\left(\frac{\eta\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{\sqrt{m}}\right)\cdot\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right)+\mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)$$
$$+\mathcal{O}\left(\eta\sigma_n\sqrt{d}\left\|\mathbf{u}_{i,j}\right\|_2\right)+\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2+\eta d\sigma_n\sigma_p\right), \tag{65}$$

where (a) is obtain by Lemma G.8. $\underline{\Gamma}=-\log(1+(1-\exp(-s))\cdot(\exp(-a)-1))/a, \gamma=-\exp(-s)/\ln(1-\exp(-s))$ and a is the lower bound of $\Delta_{3-y}^{(t)}-\Delta_y^{(t)}$ (By Lemma G.14, $\Delta_{3-y}^{(t)}-\Delta_y^{(t)}$ is lower bounded by a constant). Therefore, we have

$$\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t+1)}\right)\leq\left(1-\Omega\left(\frac{\eta\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{\sqrt{m}}\right)\right)\cdot\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right)+\mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)$$
$$+\mathcal{O}\left(\eta\sigma_n\sqrt{d}\left\|\mathbf{u}_{i,j}\right\|_2\right)+\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2+\eta d\sigma_n\sigma_p\right). \tag{66}$$

Then, combining all T iterations and using Lemma G.9, we have

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(T)}\right)\\
&\leq\left(1-\Omega\left(\frac{\eta\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{\sqrt{m}}\right)\right)^T\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(0)}\right)+\left(\mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)+\mathcal{O}\left(\eta\sigma_n\sqrt{d}\left\|\mathbf{u}_{i,j}\right\|_2\right)\right.\\
&\left.+\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2+\eta d\sigma_n\sigma_p\right)\right)\cdot\mathcal{O}\left(\frac{\sqrt{m}}{\eta\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}\right)\\
&\leq\exp\left(-\Omega\left(\frac{T\eta\gamma_{i,j}\Lambda_{i,j}}{\sqrt{m}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)\right)\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(0)}\right)+\mathcal{O}\left(\sqrt{\frac{d}{mn}}\frac{1}{\gamma_{i,j}\Lambda_{i,j}}\right)\\
&+\mathcal{O}\left(\frac{\sqrt{m}\sigma_n\sqrt{d}}{\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2}+\frac{1}{\sqrt{mn}}\frac{d\sigma_p^2}{\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}+\frac{\sqrt{m}d\sigma_n\sigma_p}{\gamma_{i,j}\Lambda_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}\right).
\end{aligned} \tag{67}$$

Setting the parameters with Condition 3.1 yields the conclusion. This completes the proof.

G.3 Proof of Theorem 3.6

Proof. Recall that in (52), we have

$$\begin{aligned}
&\mathcal{L}(\mathbf{W}^{(t+1)},\mathbf{x},y)-\mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y)\\
&=\log\left(1+\left(1-\text{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}\right)\right)\left(\exp\left(\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})\right)-1\right)\right)\\
&\overset{(a)}{\geq}c_0^{(t)}\cdot\left(1-\text{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}\right)\right)\left(\exp\left(\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})\right)-1\right)\\
&\overset{(b)}{=}\Omega\left(\exp\left(\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})\right)-1\right),
\end{aligned} \tag{68}$$

where $c_0^{(t)}>0$ for any $t\in\{0\}\cup[T-1]$ are constants. Here (a) is obtained from Lemma G.12 Then, we bound $\Delta_{3-y}^{(t)}(\mathbf{x})-\Delta_y^{(t)}(\mathbf{x})$; (b) is by $1-\exp(-s)\leq 1-\text{prob}_y(\mathbf{W},\mathbf{x})\leq 1$ with Assumption

3.2. Next, we will prove a lower bound of $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$. Recall that in (53), we have

$$
\begin{aligned}
&\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) \\
=&\underbrace{\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t+1)}, \mathbf{u}_{i,j}\Big\rangle\Big) - \sigma\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\Big]}_{A} + \underbrace{\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t+1)}, \boldsymbol{\xi}\Big\rangle\Big) - \sigma\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \boldsymbol{\xi}\Big\rangle\Big)\Big]}_{B} \\
&\underbrace{-\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\Big(\Big\langle \mathbf{w}_{y,r}^{(t+1)}, \mathbf{u}_{i,j}\Big\rangle\Big) - \sigma\Big(\Big\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\Big]}_{C} \underbrace{-\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\Big(\Big\langle \mathbf{w}_{y,r}^{(t+1)}, \boldsymbol{\xi}\Big\rangle\Big) - \sigma\Big(\Big\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\Big\rangle\Big)\Big]}_{D},
\end{aligned}
\tag{69}
$$

We then find the lower bounds of A, B and upper bounds of C, D to obtain the lower bound of $\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$.

For the term A, with probability at least $1 - \exp\Big(-\tilde{\Omega}(d)\Big)$, we have

$$
\begin{aligned}
A=&\frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\Big\langle \mathbf{w}_{3-y,r}^{(t)} - \frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\sigma'\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\cdot h(C,\mathbf{x}_k,y_k)\cdot \mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\cdot\mathbf{u}_{i,j} \\
&+\ \eta\cdot\mathbf{n}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Bigg)\Bigg] - \frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\Big] \\
\overset{(a)}{\geq}& -\frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\sigma'\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\Big\rangle\Big)\cdot h(C,\mathbf{x}_k,y_k)\cdot \mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\cdot\|\mathbf{u}_{i,j}\|_2^2 \\
&+\frac{1}{m}\sum_{r=1}^{m}\langle\eta\cdot\mathbf{n}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\rangle, \\
\overset{(b)}{\geq}& -\frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\Big(1-\mathrm{prob}_y\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\cdot\|\mathbf{u}_{i,j}\|_2^2 + \langle\eta\cdot\mathbf{n}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\rangle \\
\overset{(c)}{\geq}& -\frac{\eta\gamma_{i,j}}{m}\mathbb{E}_{(\mathbf{x}_k,y_k)\sim\mathcal{D}_{i,j}}\Big(1-\mathrm{prob}_y\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\cdot\|\mathbf{u}_{i,j}\|_2^2 - \frac{\eta}{m}\sqrt{\frac{d}{B}}\,\|\mathbf{u}_{i,j}\|_2^2 + \frac{\eta}{m}\sum_{r=1}^{m}\langle\mathbf{n}_{3-y,r}^{(t)}, \mathbf{u}_{i,j}\rangle,
\end{aligned}
\tag{70}
$$

where (a) is due to the Condition 3.1. As $\|\mathbf{u}_{i,j}\|_2 = \Theta(\sqrt{d}\sigma_p)$ and $\sigma_p = \Omega(\sigma_n)$, we have that $\|\mathbf{u}_{i,j}\|_2 = \Omega(\sqrt{d}\sigma_n)$. (b) is by the fact that $\sigma'(\cdot), h(C,\mathbf{x}_k,y_k) \leq 1$; (c) is by Lemma G.6.

For the term B, with probability at least $1-\exp\left(-\tilde{\Omega}(d)\right)$, we have

$$
\begin{aligned}
B =& \frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\Bigg\langle \mathbf{w}_{3-y,r}^{(t)} - \frac{\eta}{mB}\cdot \sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_y^{(t)}} \sigma'\left(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \boldsymbol{\xi}_k\Big\rangle\right)\cdot h(C,\mathbf{x}_k,y_k)\cdot \mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\cdot\boldsymbol{\xi}_k \\
&+\frac{\eta}{mB}\cdot \sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{3-y}^{(t)}} \sigma'\Big(\Big\langle \mathbf{w}_{3-y,r}^{(t)}, \boldsymbol{\xi}_k\Big\rangle\Big)\cdot h(C,\mathbf{x},y)\cdot\Big(1-\mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\cdot\boldsymbol{\xi}_k+\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\Bigg)\Bigg] \\
&-\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\Big\langle \mathbf{w}_{3-y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\right)\right] \\
\geq& -\frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_y^{(t)}} \sigma'\left(\Big\langle \mathbf{w}_{3-y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\right)\cdot h(C,\mathbf{x}_k,y_k)\cdot \mathrm{prob}_{3-y}\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\cdot|\langle\boldsymbol{\xi}_k,\boldsymbol{\xi}\rangle|+\langle\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\rangle \\
\geq& -\mathcal{O}\left(\frac{\eta}{m\sqrt{B}}d\sigma_p^2\right)+\frac{1}{m}\sum_{r=1}^{m}\eta\langle\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\rangle,
\end{aligned}
\tag{71}
$$

For the term C, with probability at least $1-\exp\left(-\tilde{\Omega}(d)\right)$, we have

$$
\begin{aligned}
C =& \frac{1}{m}\sum_{r=1}^{m}\sigma\Bigg(\Bigg\langle \mathbf{w}_{y,r}^{(t)}+\frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{i,j}^{(t)}}\sigma'\Big(\Big\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\Big)\cdot h(C,\mathbf{x}_k,y_k)\cdot\Big(1-\mathrm{prob}_y\Big(\mathbf{W}^{(t)},\mathbf{x}_k\Big)\Big)\cdot\mathbf{u}_{i,j} \\
&+\eta\cdot\mathbf{n}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\Big)-\frac{1}{m}\sum_{r=1}^{m}\sigma\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_{i,j}\Big\rangle\right) \\
\leq& 0,
\end{aligned}
\tag{72}
$$

where the inequality is by Condition 3.1, which implies $\|\mathbf{u}_{i,j}\|_2 = \Omega(\sqrt{d}\sigma_n)$.

For the term D, with probability at least $1-\exp\left(-\tilde{\Omega}(d)\right)$, we have

$$
\begin{aligned}
D =& \frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\Bigg\langle \mathbf{w}_{y,r}^{(t)}-\frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_{3-y}^{(t)}}\sigma'\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\right)h(C,\mathbf{x}_k,y_k)\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\boldsymbol{\xi}_k \\
&+\frac{\eta}{mB}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{S}_y^{(t)}}\sigma'\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}_k\Big\rangle\right)h(C,\mathbf{x}_k,y_k)\left(1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\right)\boldsymbol{\xi}_k+\mathbf{n}_{y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\Bigg)\Bigg] \\
&-\frac{1}{m}\sum_{r=1}^{m}\sigma\left(\Big\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}\Big\rangle\right) \\
\leq& \mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2\right)+\frac{1}{m}\sum_{r=1}^{m}\eta\langle\mathbf{n}_{y,r}^{(t)},\boldsymbol{\xi}\rangle,
\end{aligned}
\tag{73}
$$

where the inequality is by Condition 3.1 that $\sigma_n = \mathcal{O}(\sigma_p)$, $h(C,\mathbf{x}_k,y_k)<1$.

Combining the bounds together, we have

$$\begin{aligned}
&\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x}) \geq -\frac{\eta\gamma_{i,j}}{m}\mathbb{E}_{(\mathbf{x}_k,y_k)\sim\mathcal{D}_{i,j}}\left(1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\right)\cdot\|\mathbf{u}_{i,j}\|_2^2 \\
&+\frac{1}{m}\sum_{r=1}^{m}\langle\eta\cdot\mathbf{n}_{3-y,r}^{(t)},\mathbf{u}_{i,j}\rangle - \mathcal{O}\left(\frac{\eta}{m\sqrt{B}}d\sigma_p^2\right) + \frac{1}{m}\sum_{r=1}^{m}\langle\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\rangle - \mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2\right) \\
&-\frac{1}{m}\sum_{r=1}^{m}\eta\langle\mathbf{n}_{y,t}^{(t)},\boldsymbol{\xi}\rangle - \mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\|\mathbf{u}_{i,j}\|_2^2\right),
\end{aligned} \tag{74}$$

with probability at least $1-\exp\left(-\tilde{\Omega}(d)\right)$. Substituting (74) into (68), we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\mathbf{W}^{(t+1)},\mathbf{x},y)] - \mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y) \\
\geq&\mathbb{E}\left[\Omega\left(\exp\left(\Delta_{3-y}^{(t)}(\mathbf{x}) - \Delta_{y}^{(t)}(\mathbf{x})\right)-1\right)\right] \\
\geq&\Omega\left(\mathbb{E}\left[\exp\left(-\frac{\eta\gamma_{i,j}}{m}\mathbb{E}_{(\mathbf{x}_k,y_k)\sim\mathcal{D}_{i,j}}\left(1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\right)\cdot\|\mathbf{u}_{i,j}\|_2^2 + \frac{1}{m}\sum_{r=1}^{m}\langle\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\rangle\right.\right.\right. \\
&\left.\left.\left.+\frac{1}{m}\sum_{r=1}^{m}\langle\eta\cdot\mathbf{n}_{3-y,r}^{(t)},\mathbf{u}_{i,j}\rangle - \frac{1}{m}\sum_{r=1}^{m}\eta\langle\mathbf{n}_{y,t}^{(t)},\boldsymbol{\xi}\rangle - \mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2 + \frac{\eta}{m}\sqrt{\frac{d}{n}}\|\mathbf{u}_{i,j}\|_2^2\right)\right)-1\right]\right).
\end{aligned} \tag{75}$$

Here with a probability at least $1-1/d$, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1}{m}\sum_{r=1}^{m}\langle\eta\mathbf{n}_{3-y,r}^{(t)},\mathbf{u}_{i,j}\rangle + \frac{1}{m}\sum_{r=1}^{m}\langle\eta\mathbf{n}_{3-y,r}^{(t)},\boldsymbol{\xi}\rangle - \frac{1}{m}\sum_{r=1}^{m}\eta\langle\mathbf{n}_{y,t}^{(t)},\boldsymbol{\xi}\rangle\right)\right] \\
=&\exp\left(\eta^2\frac{\|\mathbf{u}_{i,j}\|_2^2\sigma_n^2 + 2\|\boldsymbol{\xi}\|_2^2\sigma_n^2}{2m}\right) \\
=&\exp\left(\Theta\left(\eta^2\frac{\|\mathbf{u}_{i,j}\|_2^2\sigma_n^2 + \sigma_p^2 d\sigma_n^2}{2m}\right)\right),
\end{aligned} \tag{76}$$

where the least equality is by Lemma G.13. With a probability at least $1-\tilde{\mathcal{O}}(1/d)$, we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\mathbf{W}^{(t+1)},\mathbf{x},y)] - \mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y) \\
\geq&\Omega\left(-\frac{\eta\gamma_{i,j}}{m}\mathbb{E}_{(\mathbf{x}_k,y_k)\sim\mathcal{D}_{i,j}}\left(1-\mathrm{prob}_y\left(\mathbf{W}^{(t)},\mathbf{x}_k\right)\cdot\|\mathbf{u}_{i,j}\|_2^2\right.\right. \\
&\left.\left.+\Omega\left(\frac{\eta^2\sigma_n^2\|\mathbf{u}_{i,j}\|_2^2}{2m}\right) + \Omega\left(\frac{\eta^2 d\sigma_n^2\sigma_p^2}{2m}\right) - \mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2\right) - \mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\|\mathbf{u}_{i,j}\|_2^2\right)\right)\right), \\
\geq&-\mathcal{O}\left(\frac{\eta\gamma_{i,j}}{m}\|\mathbf{u}_{i,j}\|_2^2\right)\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t)}\right) - \mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\|\mathbf{u}_{i,j}\|_2^2\right) \\
&+\Omega\left(\frac{\eta^2\sigma_n^2\|\mathbf{u}_{i,j}\|_2^2}{2m}\right) + \Omega\left(\frac{\eta^2 d\sigma_n^2\sigma_p^2}{2m}\right) - \mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2\right).
\end{aligned} \tag{77}$$

where the second equality is by Lemma G.8 (the $(1-\mathrm{prob}_y(\mathbf{W}^{(t)},\mathbf{x}))$ is almost surely lower bounded). Then, with a probability at least $1-\tilde{\mathcal{O}}(1/d)$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{n}^{(t)}}[\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(t+1)})] \\
\geq&\left(1-\mathcal{O}\left(\frac{\eta\gamma_{i,j}\|\mathbf{u}_{i,j}\|_2^2}{m}\right)\right)\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(t)}) - \mathcal{O}\left(\frac{\eta}{m}\sqrt{\frac{d}{n}}\|\mathbf{u}_{i,j}\|_2^2\right) \\
&+\Omega\left(\frac{\eta^2\sigma_n^2\|\mathbf{u}_{i,j}\|_2^2}{2m}\right) + \Omega\left(\frac{\eta^2 d\sigma_n^2\sigma_p^2}{2m}\right) - \mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2\right).
\end{aligned} \tag{78}$$

Combining all the iterations, with a probability at least $1-\tilde{\mathcal{O}}\left(1/d\right)$ we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{n}^{(0)},\cdots,\mathbf{n}^{(T-1)}}[\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(T)})] \\
&\geq \left(1-\mathcal{O}\left(\frac{\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{m}\right)\right)^T \mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(0)}) + \frac{1-\left(1-\mathcal{O}\left(\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2/m\right)\right)^T}{\mathcal{O}\left(\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2/m\right)} \\
&\quad \left[\Omega\left(\frac{\eta^2\sigma_n^2\left\|\mathbf{u}_{i,j}\right\|_2^2}{m}\right)+\Omega\left(\frac{\eta^2 d\sigma_n^2\sigma_p^2}{m}\right)-\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2+\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)\right].
\end{aligned} \tag{79}$$

With the number of iterations $T \geq \Omega\left(-1/\log\left(1-\Omega\left(\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2/m\right)\right)\right)$ and a probability at least $1-\tilde{\mathcal{O}}\left(1/d\right)$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{n}^{(0)},\cdots,\mathbf{n}^{(T-1)}}[\mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(T)})] \\
&\geq \left(1-\mathcal{O}\left(\frac{\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{m}\right)\right)^T \mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(0)}) + \Omega\left(\frac{m}{\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}\right) \\
&\quad \left[\Omega\left(\frac{\eta^2\sigma_n^2\left\|\mathbf{u}_{i,j}\right\|_2^2}{m}\right)+\Omega\left(\frac{\eta^2 d\sigma_n^2\sigma_p^2}{m}\right)-\mathcal{O}\left(\frac{\eta}{m\sqrt{n}}d\sigma_p^2+\frac{\eta}{m}\sqrt{\frac{d}{n}}\left\|\mathbf{u}_{i,j}\right\|_2^2\right)\right] \\
&\geq \left(1-\mathcal{O}\left(\frac{\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{m}\right)\right)^T \mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(0)}) \\
&\quad + \left[\Omega\left(\frac{\eta\sigma_n^2}{\gamma_{i,j}}\right)+\Omega\left(\frac{\eta d\sigma_n^2\sigma_p^2}{\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}\right)-\mathcal{O}\left(\frac{d\sigma_p^2}{\sqrt{n}\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}+\frac{1}{\gamma_{i,j}}\sqrt{\frac{d}{n}}\right)\right] \\
&\geq \left(1-\mathcal{O}\left(\frac{\eta\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}{m}\right)\right)^T \mathcal{L}_{\mathcal{D}_{i,j}}(\mathbf{W}^{(0)}) + \Omega\left(\frac{\eta d\sigma_n^2\sigma_p^2}{\gamma_{i,j}\left\|\mathbf{u}_{i,j}\right\|_2^2}\right)-\mathcal{O}\left(\frac{1}{\gamma_{i,j}}\sqrt{\frac{d}{n}}\right).
\end{aligned} \tag{80}$$

This completes the proof. □

G.4 Proof of Theorem 3.9

Proof. Based on (52), for any $(\mathbf{x},y)\sim\mathcal{D}_{i,j}$, we have

$$\begin{aligned}
&\mathcal{L}\left(\mathbf{W}^{(t+1)},\mathbf{x}+\boldsymbol{\zeta}^{(t+1)}(\mathbf{x}),y\right)-\mathcal{L}\left(\mathbf{W}^{(t+1)},\mathbf{x},y\right) \\
&=\log\left(1+\left(1-\text{prob}_y\left(\mathbf{W}^{(t+1)},\mathbf{x}\right)\right)\cdot\left(\exp\left(\tilde{\Delta}_{3-y}^{(t+1)}(\mathbf{x})-\tilde{\Delta}_y^{(t+1)}(\mathbf{x})\right)-1\right)\right),
\end{aligned} \tag{81}$$

where

$$\boldsymbol{\zeta}^{(t+1)}(\mathbf{x})=\arg\max_{\|\boldsymbol{\zeta}\|_p\leq\zeta}\mathcal{L}\left(\mathbf{W}^{(t+1)},\mathbf{x}+\boldsymbol{\zeta},y\right), \tag{82}$$

and

$$\begin{aligned}
&\tilde{\Delta}_{3-y}^{(t+1)}(\mathbf{x})-\tilde{\Delta}_y^{(t+1)}(\mathbf{x}) \\
&=\underbrace{\frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left[\sigma\left(\left\langle\mathbf{w}_{3-y,r}^{(t+1)},\mathbf{x}^{(j)}+\boldsymbol{\zeta}^{(t+1)}(\mathbf{x})^{(j)}\right\rangle\right)-\sigma\left(\left\langle\mathbf{w}_{3-y,r}^{(t+1)},\mathbf{x}^{(j)}\right\rangle\right)\right]}_{E} \\
&\quad\underbrace{-\frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left[\sigma\left(\left\langle\mathbf{w}_{y,r}^{(t+1)},\mathbf{x}^{(j)}+\boldsymbol{\zeta}^{(t+1)}(\mathbf{x})^{(j)}\right\rangle\right)-\sigma\left(\left\langle\mathbf{w}_{y,r}^{(t+1)},\mathbf{x}^{(j)}\right\rangle\right)\right]}_{F}.
\end{aligned} \tag{83}$$

Then, we bound E and F.

For the term E, with probability at least $1-\exp\left(-\Omega(d)\right)$, we have

$$\begin{aligned}
E \overset{(a)}{\leq} & \frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left|\left\langle \mathbf{w}_{3-y,r}^{(t+1)}, \boldsymbol{\zeta}^{(t+1)}\left(\mathbf{x}\right)^{(j)}\right\rangle\right| \\
\overset{(b)}{\leq} & \frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left\|\mathbf{w}_{3-y,r}^{(t+1)}\right\|_2\left\|\boldsymbol{\zeta}^{(t+1)}\left(\mathbf{x}\right)^{(j)}\right\|_2 \\
\overset{(c)}{\leq} & \frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left\|\mathbf{w}_{3-y,r}^{(t+1)}\right\|_2\left\|\boldsymbol{\zeta}^{(t+1)}\left(\mathbf{x}\right)^{(j)}\right\|_p d^{1-1/p} \\
\leq & \frac{2}{m}\sum_{r=1}^{m}\left(\sum_{t'=0}^{t}\left\|\mathbf{w}_{3-y,r}^{(t'+1)}-\mathbf{w}_{3-y,r}^{(t')}\right\|_2+\left\|\mathbf{w}_{3-y,r}^{(0)}\right\|_2\right)\bar{\zeta}d^{1-1/p} \\
\overset{(d)}{\leq} & \mathcal{O}\left(\left[t\frac{\eta}{m}C+\frac{\eta}{m}\sqrt{td}\sigma_n+\sqrt{d}\sigma_0\right]\bar{\zeta}d^{1-1/p}\right),
\end{aligned} \tag{84}$$

where (a) is because $\mathrm{ReLU}(\cdot)$ is 1-Lipschitz continuous; (b) is due to the Cauchy-Schwarz inequality; (c) is because of the Hölder's inequality; (d) is due to Lemma G.2.

Similarly, for the term F, with probability at least $1-\exp\left(-\Omega(d)\right)$, we have we have

$$\begin{aligned}
F \geq & -\frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left|\left\langle \mathbf{w}_{y,r}^{(t+1)}, \boldsymbol{\zeta}^{(t+1)}\left(\mathbf{x}\right)^{(j)}\right\rangle\right| \\
\geq & -\frac{1}{m}\sum_{r=1}^{m}\sum_{j=1}^{2}\left\|\mathbf{w}_{y,r}^{(t+1)}\right\|_2\left\|\boldsymbol{\zeta}^{(t+1)}\left(\mathbf{x}\right)^{(j)}\right\|_2 \\
\geq & -\frac{2}{m}\sum_{r=1}^{m}\left(\sum_{t'=0}^{t}\left\|\mathbf{w}_{y,r}^{(t'+1)}-\mathbf{w}_{y,r}^{(t')}\right\|_2+\left\|\mathbf{w}_{y,r}^{(0)}\right\|_2\right)\bar{\zeta}d^{1-1/p} \\
\geq & -\Omega\left(\left[t\frac{\eta}{m}C+\frac{\eta}{m}\sqrt{td}\sigma_n+\sqrt{d}\sigma_0\right]\bar{\zeta}d^{1-1/p}\right).
\end{aligned} \tag{85}$$

Combing with the bounds of E, F, with probability at least $1-\exp(-\tilde{\Omega}(d))$, we have

$$\tilde{\Delta}_{3-y}^{(t+1)}\left(\mathbf{x}\right)-\tilde{\Delta}_{y}^{(t+1)}\left(\mathbf{x}\right)\leq\mathcal{O}\left(\left[t\frac{\eta}{m}C+\frac{\eta}{m}\sqrt{td}\sigma_n+\sqrt{d}\sigma_0\right]\bar{\zeta}d^{1-1/p}\right). \tag{86}$$

Then, with probability at least $1-\exp(-\tilde{\Omega}(d))$, we have

$$\begin{aligned}
& \mathcal{L}_{\mathcal{D}_{i,j}}^{\text{adv}}\left(\mathbf{W}^{(t+1)}\right)-\mathcal{L}_{\mathcal{D}_{i,j}}\left(\mathbf{W}^{(t+1)}\right) \\
\leq & \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{i,j}}\left[\Gamma\left(\tilde{\Delta}_{3-y}^{(t+1)}\left(\mathbf{x}\right)-\tilde{\Delta}_{y}^{(t+1)}\left(\mathbf{x}\right)\right)\left(\tilde{\Delta}_{3-y}^{(t+1)}\left(\mathbf{x}\right)-\tilde{\Delta}_{y}^{(t+1)}\left(\mathbf{x}\right)\right)\right] \\
\leq & \mathcal{O}\left(\left[t\frac{\eta}{m}C+\frac{\eta}{m}\sqrt{td}\sigma_n+\sqrt{d}\sigma_0\right]\bar{\zeta}d^{1-1/p}\right).
\end{aligned} \tag{87}$$

Combining with Theorem 3.5 and setting parameters with Condition 3.1, with probability at least $1-\exp(-\tilde{\Omega}(d))$, we have

$$\mathcal{L}_{\mathcal{D}_{i,j}}^{\text{adv}}\left(\mathbf{W}^{(T)}\right)\leq\bar{L}_{i,j}+\mathcal{O}\left(\left[\frac{T}{m}C+\frac{\sqrt{Td}}{m}\sigma_n+\sqrt{d}\sigma_0\right]\bar{\zeta}d^{1-1/p}\right). \tag{88}$$

This completes the proof. □

H PROOF OF PROPOSITION 4.4

Proof. We have

$$\sigma(\langle \tilde{\mathbf{w}}_{1,r}, \mathbf{u}_1' \rangle) = C_1 \cos\theta \left\|\mathbf{u}_1'\right\|_2 \left\|\mathbf{u}_1\right\|_2, \tag{89}$$

$$0 \leq \sigma(\langle \tilde{\mathbf{w}}_{1,r}, \boldsymbol{\xi} \rangle) \leq C_3 \sigma_p^2, \tag{90}$$

$$\sigma(\langle \tilde{\mathbf{w}}_{2,r}, \mathbf{u}_1' \rangle) = C_1 \sin\theta \left\|\mathbf{u}_1'\right\|_2 \left\|\mathbf{u}_2\right\|_2, \tag{91}$$

$$0 \leq \sigma(\langle \tilde{\mathbf{w}}_{2,r}, \boldsymbol{\xi} \rangle) \leq C_3 \sigma_p^2, \tag{92}$$

$$\sigma(\langle \tilde{\mathbf{w}}_{1,r}, \mathbf{u}_2' \rangle) = 0, \tag{93}$$

$$\sigma(\langle \tilde{\mathbf{w}}_{2,r}, \mathbf{u}_2' \rangle) = C_1 \cos\theta \left\|\mathbf{u}_2\right\|_2 \left\|\mathbf{u}_2'\right\|_2. \tag{94}$$

Using the above inequalities (equalities), we have

$$\begin{aligned} \mathcal{L}_{\mathcal{D}_2}(\tilde{\mathbf{W}}) \leq & -\frac{1}{2} \ln \left(\frac{\exp(C_1 \cos\theta \left\|\mathbf{u}_1'\right\|_2 \left\|\mathbf{u}_1\right\|_2)}{\exp(C_1 \cos\theta \left\|\mathbf{u}_1'\right\|_2 \left\|\mathbf{u}_1\right\|_2) + \exp(C_1 \sin\theta \left\|\mathbf{u}_1'\right\|_2 \left\|\mathbf{u}_2\right\|_2 + C_3 \sigma_p^2)} \right) \\ & -\frac{1}{2} \ln \left(\frac{\exp(C_1 \cos\theta \left\|\mathbf{u}_2'\right\|_2 \left\|\mathbf{u}_2\right\|_2)}{\exp(C_1 \cos\theta \left\|\mathbf{u}_2'\right\|_2 \left\|\mathbf{u}_2\right\|_2) + \exp(C_3 \sigma_p^2)} \right). \end{aligned} \tag{95}$$

Based on Theorem 3.5 and $\left\|\mathbf{u}_1\right\|_2 = \left\|\mathbf{u}_1\right\|_2 = \left\|\mathbf{u}_1'\right\|_2 = \left\|\mathbf{u}_2'\right\|_2$, we have

$$\mathcal{L}_{\mathcal{D}_{\mathrm{ft}}}(\mathbf{W}^{(T)}) \leq \exp\left(-\Omega\left(\frac{\Lambda_i \left\|\mathbf{u}_i\right\|_2^2}{\sqrt{m}} T \right) \right) \cdot \tilde{L} + \mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{mn}\Lambda_i} \right) + \mathcal{O}\left(\frac{\sqrt{md}\sigma_n}{\Lambda_i \left\|\mathbf{u}_i\right\|_2} \right), \tag{96}$$

This completes the proof. □

I EXPERIMENTS COMPUTE RESOURCES

The simulations are conducted on a commodity machine with Intel® Core i7-9700 CPU with a NVidia® 3090Ti GPU.

J BROADER IMPACTS

Our work studies side effects in DP-SGD trained models. One important side effect is unfairness. Our work identifies data imbalance as one source of unfairness, indicating that collecting balanced data is significant for maintaining fairness. Our work also shows the potential of design algorithms to adjust group weights to improve model fairness.