UNDERSTANDING THE IMPACT OF DIFFERENTIALLY PRIVATE TRAINING ON MEMORIZATION OF LONG-TAILED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent research shows that modern deep learning models achieve high predictive accuracy partly by memorizing individual training samples. Such memorization raises serious privacy concerns, motivating the widespread adoption of differentially private training algorithms such as DP-SGD. However, a growing body of empirical work shows that DP-SGD often leads to suboptimal generalization performance, particularly on long-tailed data that contain a large number of atypical samples. Despite these observations, a theoretical understanding of this phenomenon remains largely unexplored, and existing differential privacy analysis are difficult to extend to the nonconvex and nonsmooth neural networks commonly used in practice. In this work, we develop the first theoretical framework for analyzing DP-SGD on long-tailed data from a feature learning perspective. We show that the test error of DP-SGD-trained models on the long-tailed subpopulation is significantly larger than the test error over the entire dataset. Our analysis further characterizes the training dynamics of DP-SGD, demonstrating how gradient clipping and noise injection jointly adversely affect the model's ability to memorize informative but atypical samples. Finally, we validate our theoretical findings through extensive experiments on both synthetic and real-world datasets.

1 INTRODUCTION

Memorization, in the classical statistical regime, is traditionally viewed as a hindrance to generalization, typically mitigated through explicit regularization (Ruppert, 2004). However, modern deep learning models are often trained in highly overparameterized regimes, where empirical evidence shows that models can perfectly interpolate the training data while still generalizing well (Zhang et al., 2019; Bartlett et al., 2020; Cao et al., 2022; Kou et al., 2023). Moreover, incorporating atypical, long-tailed data during training has been shown to be crucial for attaining high predictive accuracy (Carlini et al., 2019a; Feldman, 2020).

However, this propensity for memorization raises significant privacy concerns, especially in sensitive domains such as finance (Lundervold & Lundervold, 2019; Chlap et al., 2021), healthcare (Ozbayoglu et al., 2020; Bi & Lian, 2024), and user-centric applications (Oroojlooy & Hajinezhad, 2023). Models that memorize specific training samples are vulnerable to privacy attacks, including membership inference and data extraction (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2019b). To mitigate these risks, Differential Privacy (DP) (Dwork et al., 2006), particularly through the lens of DP-SGD (Abadi et al., 2016), has become the de facto standard for providing formal privacy guarantees by limiting the influence of individual samples. While DP compromises model generalization, this degradation is particularly acute on long-tailed data. Recent empirical findings suggest that DP particularly hinders the recognition of atypical underrepresented sub-populations at test time (Carlini et al., 2019a). This phenomenon manifests as a non-uniform utility drop that exacerbates performance gaps for disadvantaged and data-complex groups (Bagdasaryan et al., 2019).

Despite these empirical observations, *previous theoretical studies only focus on the impact of general DP mechanisms on memorization (Feldman, 2020); however, a rigorous and fine-grained theoretical understanding of how DP-SGD influences memorization remains an open challenge*. Prior research on DP training has been largely confined to deriving utility bounds, often overlooking the underlying

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

training dynamics. Furthermore, these analysis frequently hinge on assumptions of convexity and smoothness, which fail to characterize the non-convex, non-smooth, and high-dimensional nature of modern large-scale neural networks (Dwork et al., 2006; Chaudhuri et al., 2011; Bassily et al., 2014; Wang & Xu, 2019; Bassily et al., 2019; Feldman et al., 2020). Recently, Ding et al. (2025) leveraged a feature learning perspective to analyze the significance of feature augmentation in the context of DP. Meanwhile, Xu & Chen (2026) investigated the disparate impact, adversarial robustness, and private fine-tuning under DP-SGD with a unified framework based on two-layer ReLU CNNs. However, neither study explicitly addresses the impact of DP on the memorization of long-tailed data.

To fill this gap, we approach the problem from a feature learning perspective, focusing on how neural networks dynamically extract task-relevant signals versus task-irrelevant components, and how such learning dynamics ultimately govern the model's generalization performance (Allen-Zhu & Li, 2020; 2022; Cao et al., 2022; Kou et al., 2023). While feature learning has been widely studied, it has seldom addressed the unique confluence of gradient clipping, noise injection, and long-tailed data distributions. To this end, we first formulate a multi-class classification task using a ReLU-activated CNN optimized via standard DP-SGD. By rigorously characterizing the impacts of gradient clipping and noise injection, we provide an in-depth analysis of training dynamics and test error to uncover why DP suppresses the memorization of long-tailed data. We conclude our contributions as follows:

- **A Novel Theoretical Framework for DP-SGD on Long-tailed Data.** To the best of our knowledge, we are the first to establish a theoretical feature learning framework for analyzing DP-SGD on long-tailed data. Leveraging a multi-patch data structure, we rigorously characterize the update dynamics of two-layer CNNs during the DP-SGD training process.
- **Theoretical Insights into Training Dynamics and Generalization.** We derive formal expressions for the training dynamics and test error under DP-SGD, providing a comparative analysis between full-set and long-tailed data distributions. Based on our fine-grained analysis, we conclude some new findings compared with previous theoretical work (Feldman, 2020). Our analysis attributes the poor performance of long-tailed data under DP-SGD to their atypicality, characterized by their dependence on weight–feature alignment (as Definition 4.7), rather than solely to low occurrence counts.
- **Empirical Validation on Synthetic and Real-world Data.** We validate our theoretical findings using two-layer CNNs and LeNet on synthetic data, MNIST, and CIFAR-10. Our experiments verify the predicted training dynamics of DP-SGD regarding the memorization process and its impact on the test error for both full data and long-tailed data.

2 Related Work

Empirical and Theoretical Studies of Memorization. Several empirical research has explored the phenomenon of memorization in neural networks. Feldman (2020) argues that memorizing labels—including those of outliers and noisy samples—is indispensable for attaining near-optimal generalization error. Furthermore, Carlini et al. (2019a) demonstrates that models must memorize anomalous samples to achieve high confidence during training. Despite these profound empirical insights, a rigorous theoretical foundation for such phenomena remains relatively elusive.

Recently, several theoretical studies attempt to model memorization from a feature-learning perspective, typically bifurcating overfitting into "benign" signal learning and "harmful" noise memorization (Cao et al., 2022; Kou et al., 2023). However, these frameworks predominantly characterize memorization as a detrimental process. This perspective stands in stark contrast to the aforementioned empirical findings, which suggest that memorization is not merely a byproduct of training but can actually be beneficial for achieving high-precision generalization performance (Carlini et al., 2019a).

Theory on Differentially Private Learning. Extensive work has focused on differential privacy, including classical results for private empirical risk minimization (Chaudhuri et al., 2011; Bassily et al., 2014; Wang & Xu, 2019) and private stochastic convex optimization (Bassily et al., 2019; Feldman et al., 2020). These studies have been further extended to various settings, such as heavy-tailed distributions (Wang et al., 2020; Hu et al., 2022) and non-Euclidean spaces (Bassily et al., 2021; Asi et al., 2021). However, the theoretical understanding of private deep learning remains relatively limited, especially from the perspective of feature learning. Recently, only a few works have investigated DP in two-layer ReLU neural networks. Ding et al. (2025) highlights the importance

of feature augmentation in DP, while Xu & Chen (2026) theoretically establishes that DP can induce disparate impacts across different subpopulations. Nevertheless, these studies are restricted to standard binary classification, where the data noise is sampled from a Gaussian distribution. In contrast, we extend the data distribution to a multi-class setting, where the noise incorporates class-dependent implicit patterns, and investigate the performance disparity of DP-SGD between all data and long-tailed data.

Privacy-Fairness Tradeoff. Recently, researchers have recognized that privacy-preserving mechanisms can inadvertently exacerbate algorithmic bias. Bagdasaryan et al. (2019) first demonstrated that DP models often exhibit a "disparate impact", where the accuracy degradation for minority classes is significantly more pronounced than for majority classes. This phenomenon was further investigated by Pujol et al. (2020), who showed that DP noise can lead to inequitable resource allocation in decision-making systems. Beyond empirical observations, Cummings et al. (2019) provided theoretical evidence that even with full distributional access, DP constraints can impede the exact achievement of fairness parity, such as the Equality of False Positives and False Negatives. However, a rigorous theoretical understanding of how DP affects fairness specifically within the context of long-tailed data distributions remains largely under-explored.

3 Problem Setup

In this section,we introduce the necessary definitions and formally describe the training of two-layer CNNs via DP-SGD under the multi-patch data distribution, laying the foundation for our subsequent theoretical analysis. We defer the detailed notation conventions to Appendix B.

Definition 3.1 (Data Generation Model)**.** Let $\mathbf{u}_1, .., \mathbf{u}_k \in \mathbb{R}^d$ be fixed vectors representing the signals contained in data points, where $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for all $i, j \in [K]$ and $i \neq j$. Then, each data point $(\mathbf{x}, y)$ with $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \in \mathbb{R}^{2d}$ and $y \in [K]$ is generated from the following distribution $\mathcal{D}$:

(1) Sample the label y following a distribution $\mathcal{K}$, whose support is $[K]$ ($K = \Theta(1)$).

(2) The noise vector $\boldsymbol{\xi}$ is generated as $\boldsymbol{\xi} = \mathbf{A}_y \boldsymbol{\zeta}$, where $\mathbf{A}_y \in \mathbb{R}^{d \times d}$ is the class-dependent matrix that determines the geometry of the noise distribution, and the coordinates of $\boldsymbol{\zeta}$ are drawn i.i.d. from $\mathcal{D}_\zeta$, a symmetric $\sigma_p = \Theta(1)$ sub-Gaussian distribution with unit variance. We assume that $\mathbf{u}_k^\top \mathbf{A}_y = 0$ for all $k \in [K]$.

(3) One of $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ is given as $\mathbf{u}_y$ representing the signal, the other is given by $\boldsymbol{\xi}$ representing noises.

This data model is widely used in feature learning literature(Cao et al., 2022; Kou et al., 2023; Xu & Chen, 2025; Ding et al., 2025). To capture the inherent heterogeneous class-dependent structures observed in real-world datasets, we adopt a class-dependent noise framework, distinguishing our approach from traditional homogeneous models. By leveraging the $\mathbf{A}_y$ eigenvalue spectrum, this setup rigorously characterizes long-tailed samples and acknowledges that their generalization hinges on memorizing specific noise patterns (Xu & Chen, 2025).

Two-layer CNN. We consider a two-layer CNN with ReLU activation, where the first layer comprises m filters (neurons) for each of the K classes, and the second-layer parameters are fixed at $1/m$ as Cao et al. (2022); Kou et al. (2023). Given an input $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, the model with weights $\mathbf{W}$ produces a K-dimensional output vector $[F_1, \ldots, F_K]^\top$, whose k-th component is defined as:

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

where $\sigma(z) = \max\{0, z\}$ is the ReLU activation function, and $\mathbf{w}_{k,r}$ denotes the weight vector of the r-th filter associated with class k.

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

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

where $\mathcal{L}(\mathbf{W}, \mathbf{x}, y) = -\log(\text{logit}_y(\mathbf{W}, \mathbf{x}))$ and $\text{logit}(\cdot)$ is the output probabilities of the network:

$$\text{logit}_y(\mathbf{W}, \mathbf{x}) = \frac{\exp(F_y(\mathbf{W}, \mathbf{x}))}{\sum_{k=1}^{K} \exp(F_k(\mathbf{W}, \mathbf{x}))}. \tag{3}$$

We study the generalization ability using the test error of on class-conditional distribution $\mathcal{D}_k$ whose probability density function is $\mathbb{P}_{\mathcal{D}_k}[(\mathbf{x}, y)] = \mathbb{P}_{\mathcal{D}}[(\mathbf{x}, y)|y = k]$, i.e., for all $k \in [K]$,

$$\mathcal{L}_{\mathcal{D}_k}(\mathbf{W}) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_k}\left[F_y(\mathbf{W}, \mathbf{x}) \neq \max_{j\in[K]}\{F_j(\mathbf{W}, \mathbf{x})\}\right]. \tag{4}$$

To preserve privacy, the learned weights $\mathbf{W}$ satisfy the definition of DP (Dwork et al., 2006):

Definition 3.2 ((ϵ, δ_{DP})-Differential privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ is (ϵ, δ_{DP})-DP if, for every pair of neighboring datasets $Z, Z' \in \mathcal{Z}$ that differ by one entry, and for any subset of outputs $S \subseteq \mathcal{R}$, the following holds: $\mathbb{P}[\mathcal{M}(Z) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(Z') \in S] + \delta_{DP}$.

DP-SGD Training algorithm. DP-SGD (Abadi et al., 2016)—which consists of gradient clipping and noise injection—is the standard training algorithm for differentially private deep learning. We train the neural network using DP-SGD with a learning rate η, i.e.,

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

$\mathcal{S}^{(t)}$ is a mini-batch of size B randomly selected at iteration t, $\mathbf{n}^{(t)}$ is the noise added for privacy protection, sampled from $\mathcal{N}(0, \sigma_n^2\mathbf{I})$, and $\text{clip}_C(\mathbf{x})$ is the gradient clipping function with a clipping threshold C on vector $\mathbf{x}$: $\text{clip}_C(\mathbf{x}) = \mathbf{x}/\max\{1, \|\mathbf{x}\|_2/C\}$. The weights of the neural network's parameters are initialized independently as $\mathbf{w}_{j,r}^{(0)} \sim \mathcal{N}(0, \sigma_0^2\mathbf{I})$ for all $j \in [K]$ and $r \in [m]$.

4 Main Results

In this section, we present our main theoretical results. We characterize the training dynamics of memorization under DP-SGD, and show that privacy protection induces suboptimal training loss. Furthermore, we analyze the resulting test error on both full-set data and long-tailed data, highlighting the disparate impact of differential privacy. We first introduce the following conditions.

Condition 4.1. *Suppose there exists a sufficiently large constant c_1 and $0 < c_2 < 1$. For certain probability parameter $\delta \in (0, 1)$, the following conditions hold:*

(a) For the neurons to learn the data patterns, for any $i, j \in [K]$, the noise patch distributions satisfy:

$$\begin{cases} \text{Tr}(\mathbf{A}_i^\top\mathbf{A}_i) \geq c_1 n\Lambda_i^{-1}\max\left\{\|\mathbf{A}_i^\top\mathbf{A}_j\|_F\log(n^2/\delta), n^{1/2}\max_{i,j\in[K]}\{\|\mathbf{A}_i^\top\mathbf{A}_j\|_F^{1/2}\}\log^{1/2}(n^2/\delta)/|\mathcal{S}_i|\right\}, \\ \|\mathbf{A}_i^\top\mathbf{A}_j\|_F/\|\mathbf{A}_i^\top\mathbf{A}_j\|_{op} \geq c_1\sqrt{\log(K/\delta)}, \quad \|\mathbf{A}_i^\top\mathbf{A}_j\|_F \geq c_2\max_{k\neq j}\{\|\mathbf{A}_i^\top\mathbf{A}_k\|_F\}. \end{cases}$$

Moreover, there exists a threshold $c' > 0$ such that $\mathbb{P}[\zeta > c'] \geq 0.4$.

(b) To ensure that the learning problem is in a sufficiently over-parameterized setting, the training dataset size n, network width m, and dimension d satisfy:

$$\begin{cases} m \geq c_1\log(n/\delta)\max_i\{(\lambda_{\max}^+(\mathbf{A}_i))^2/(\lambda_{\min}^+(\mathbf{A}_i))^2\}, \quad \min\{m, d, \text{rank}(\mathbf{A}_j)\} - 0.9m \geq Cn \\ \min_k|S_k| \geq c_1\log(Km/\delta), d \geq c_1\log(mn/\delta). \end{cases}$$

(c) We denote $\Lambda_k = \frac{C}{\|\mathbf{u}_k\|_2 + \|\mathbf{A}_k\|_F}$ as the clipping factor for class k, which quantifies the maximum change in gradient magnitude. To ensure that DP-SGD can minimize the training loss, the learning rate η, the batch size B, initialization σ_0 and the DP noise σ_n satisfy for every $k \in [K]$:

$$\begin{cases} \eta \leq \left(c_1(C + \sqrt{d}\sigma_n)(\max_{j\in[K]}\|\mathbf{u}_j\|_2 + \sqrt{1.5\,\text{Tr}(\mathbf{A}_j^\top\mathbf{A}_j)})\right)^{-1}, \\ B \geq c_2 \cdot n, \quad \sigma_0 \leq c_1^{-1}n^{-1}\phi \cdot \left(\max_{k\in[K]}\{\sqrt{\log(Km/\delta)}\|\mathbf{u}_k\|_2, \log(Km/\delta)\|\mathbf{A}_k\|_F\}\right)^{-1}, \\ c\sigma_n\sigma_p\left(\|A_k\|_F\sqrt{\log(2nmT/\delta)} + \|\mathbf{A}_k\|_2\log(2nmT/\delta)\right) \leq \Lambda_k\,\text{Tr}(\mathbf{A}_k^\top\mathbf{A}_k)/mn. \end{cases}$$

where $\phi := \min_{k_1,k_2\in[K]}\{\|\mathbf{A}_{k_1}^\top\mathbf{A}_{k_2}\|_F, \|\mathbf{u}_{k_1}\|_2^2\}$.

Similar conditions are widely made in the theoretical analysis of feature learning in neural networks (Xu & Chen, 2025; Cao et al., 2022; Kou et al., 2023; Ding et al., 2025).Compared to the standard conditions in non-private training, we impose a condition on the batch size B and relax the condition on the learning rate η, as the model cannot converge to an arbitrarily small term. We also bound $\sigma_n$to ensure that the privacy noise does not severely disrupt the training process.

Assumption 4.2 (s-non-perfect model)**.** We assume that the model is almost surely not perfect on any test example, i.e., for some constant $s > 0, \mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{x}, y\right) \geq s$, for all $(\mathbf{x}, y) \sim \mathcal{D}$.

Assumption 4.2 is relatively mild, particularly in light of the inherent stochasticity introduced by DP-SGD during the training process (Xu & Chen, 2026). Consequently, the resulting model is stochastic, making it highly improbable to attain zero cross-entropy loss on any given test sample.

Then, we first demonstrate the occurrence of memorization during training from the perspective of the training dynamics, as formalized in the theorem below.

Theorem 4.3 (noise pattern memorization)**.** *Under Condition 4.1 and Assumption 4.2, for any* $(\mathbf{x}, y)$ *in the training dataset* $\mathcal{S}$*, after* $T \geq \Omega\left((\eta\Lambda_y\|\mathbf{A}_y\|_F)^{-1} n\sqrt{m}\sigma_0\right)$ *iterations, with probability at least* $1-\delta$*, the inner product between the weight vector* $\mathbf{w}_{y,r}^{(T)}$ *and the noise vector* $\boldsymbol{\xi}$ *satisfies:*

$$\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi} \rangle \geq \Omega\left(\sum_{t=0}^{T-1} \frac{\eta\Lambda_y}{n\sqrt{m}} \left(1 - \mathit{logit}_y(\mathbf{W}^{(t)}, \mathbf{x})\right) \|\boldsymbol{\xi}\|_2^2\right) - \mathcal{O}\left(\eta T \sigma_n \sqrt{d} \|\mathbf{A}_\mathbf{y}\|_F)\right)$$

Remark 4.4. The inner product $\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi} \rangle$ quantifies the degree of alignment between the neuron weights and the heterogeneous noise, which serves as a proxy for the extent of noise pattern memorization by the model. Theorem 4.3 suggests that this memorization effect accumulates over iterations (as $(1 - \mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x})) \geq 0$). Notably, DP-SGD mitigates this undesirable process through two mechanisms: first, gradient clipping (manifested as $\Lambda_y < 1$) scales down the magnitude of updates that contribute to memorization; second, the injection of Gaussian noise (σ_n) introduces stochastic perturbations that disrupt the fine-grained alignment. Noted that, prior work only focuses on the impact of general DP mechanisms on memorization, attributing poor results to the low frequency of sub-class samples (Feldman, 2020); we provide the first new insight into the impact of DP on the alignment between model weights and inherent features.

Next, we study the terminal state of training under DP-impacted weight-feature alignment. We characterize an upper bound on the training loss of models trained with DP-SGD, highlighting the suboptimality induced by privacy preservation. We denote S_j as the set containing training data with label j in the training dataset S.

Theorem 4.5 (Training loss)**.** *Under Condition 4.1 and Assumption 4.2, for any* $(\mathbf{x}, y) \in \mathcal{S}$*, with probability at least* $1-\delta$*, the training loss satisfies:*

$$\mathcal{L}_S(\mathbf{W}^{(T)}, \mathbf{x}, y) \leq \underbrace{\exp\left(-\Omega\left(\frac{\eta T \Lambda_y |\mathcal{S}_y|}{n\sqrt{m}} \|\mathbf{u}_y\|_2^2\right)\right) \mathcal{L}(\mathbf{W}^{(0)}, \mathbf{x}, y)}_{\textit{Vanishing error}} + \underbrace{\mathcal{O}\left(\frac{n\sqrt{m} \cdot \sigma_n \sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F)}{\Lambda_y |\mathcal{S}_y| \|\mathbf{u}_y\|_2^2}\right)}_{\textit{Privacy protection error}}.$$

Remark 4.6. Theorem 4.5 characterizes the training loss under DP-SGD, which is composed of a vanishing error and a privacy protection error. While the vanishing error decays exponentially toward zero as T increases, the privacy protection error persists as an irreducible residual. This term prevents the training loss from converging to an arbitrarily small value, representing a significant departure from the "benign overfitting" phenomenon observed in standard overparameterized regimes (Cao et al., 2022; Kou et al., 2023). We also provide a corresponding lower bound of the training loss that reflects the privacy cost in Appendix D.6.

Before presenting our main theorem, we provide a formal definition of the long-tailed data distribution. Drawing inspiration from Theorem 4.3, we define the long-tailed dataset as the collection of samples whose successful fitting under standard (non-private) training hinges on the alignment between model weights and inherent features (heterogeneous noise) following Xu & Chen (2025).

Definition 4.7 (L-Long-tailed data set)**.** The L-long-tailed data distribution $\mathcal{T}_j$ for each $j \in [K]$ with model $\mathbf{W}^{(T)}$ is defined as

$$\mathbb{P}_{\mathcal{T}_j}[(\mathbf{x}, y)] = \mathbb{P}_{\mathcal{D}_j}\left[(\mathbf{x}, y) \middle| \left\langle \sum_{r \in \mathcal{R}(\boldsymbol{\xi})} \mathbf{w}_{y,r}, \boldsymbol{\xi} \right\rangle \geq L \left\| \mathbf{A}_y^\top \sum_{r \in \mathcal{R}(\boldsymbol{\xi})} \mathbf{w}_{y,r} \right\|_2\right],$$

where $\mathcal{R}(\boldsymbol{\xi}) = \{r \in [m] : \langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi} \rangle > 0\}$.

Definition 4.7 identifies atypical samples whose weight–feature alignment $\langle \sum_{r \in \mathcal{R}(\boldsymbol{\xi})} \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi} \rangle$ exceeds a threshold. Notably, we define long-tailed data using model weights obtained from non-DP (clean) training, and subsequently evaluate models trained with DP-SGD on these data points.

Definition 4.7 selects training data points whose noise patterns align better with the model weights, indicating that these samples are more effectively memorized. Due to the concentration phenomenon in high-dimensional settings, we do not define long-tailed data based on norms. Instead, we identify "long-tailed" samples using the trained model (Xu & Chen, 2025; Feldman & Zhang, 2020). An illustrative visualization is provided in Appendix A.

We define the signal-to-noise ratio as $|\mathcal{S}_k| \|\mathbf{u}_k\|_2^2 / \sqrt{|\mathcal{S}_j|} \|\mathbf{A}_k^\top \mathbf{A}_j\|_F$ and the noise correlation ratio as $\sqrt{|\mathcal{S}_k|} \|\mathbf{A}_k^\top \mathbf{A}_k\|_F / \sqrt{|\mathcal{S}_j|} \|\mathbf{A}_k^\top \mathbf{A}_j\|_F$. We present our main result in the following theorem.

Theorem 4.8 (Test error)**.** *For any $k \in [K]$, under Condition 4.1 and Assumption 4.2, there exists $T = \tilde{\mathcal{O}}(\eta^{-1} C^{-1} n \sqrt{m})$ with probability at least $1 - \delta$, the following holds:*

(1) ***All data.*** *Suppose the signal-to-noise ratio is large, i.e.,*

$$|\mathcal{S}_k| \Lambda_k \|\mathbf{u}_k\|_2^2 \geq C_1 \cdot (\max_{j \neq k} \{\sqrt{|\mathcal{S}_j|} \|\mathbf{A}_k^\top \mathbf{A}_j\|_F\} + n\sigma_n (\|\mathbf{u}_k\|_2 + \|\mathbf{A}_k\|_F)),$$

then the test error satisfies:

$$\mathcal{L}_{\mathcal{D}_k}(\mathbf{W}^{(T)}) \leq \sum_{j \neq k} \exp\left[-c_1 \cdot \left(\frac{|\mathcal{S}_k| \Lambda_k \|\mathbf{u}_k\|_2^2 - n\sigma_n \|\mathbf{u}_k\|_2 \sqrt{2\log(2/\delta)}}{\sqrt{|\mathcal{S}_j|} \|\mathbf{A}_k^\top \mathbf{A}_j\|_F + n\sigma_n \|\mathbf{A}_k\|_F} \right)^2 \right]$$

(2) ***Long-tailed data.*** *Suppose the noise correlation ratio is large, i.e.,*

$$\sqrt{|\mathcal{S}_k|} \Lambda_k \|\mathbf{A}_k^\top \mathbf{A}_k\|_F \geq C_2 \cdot (\max_{j \neq k} \{\sqrt{|\mathcal{S}_j|} \|\mathbf{A}_k^\top \mathbf{A}_j\|_F\} + n\sigma_n \sqrt{d + L^2} \|\mathbf{A}_k\|_{op} + n\sigma_n \|\mathbf{A}_k\|_F),$$

then the test error satisfies:

$$\mathcal{L}_{\mathcal{T}_k}(\mathbf{W}^{(T)}) \leq \sum_{j \neq k} \exp\left[-c_2 \cdot \left(\frac{L\sqrt{|\mathcal{S}_k|} \Lambda_k \|\mathbf{A}_k^\top \mathbf{A}_k\|_F - n\sigma_n \sqrt{d + L^2} \|\mathbf{A}_k\|_{op}}{\sqrt{|\mathcal{S}_j|} \|\mathbf{A}_k^\top \mathbf{A}_j\|_F + n\sigma_n \|\mathbf{A}_k\|_F} \right)^2 \right]$$

Here, $\Lambda_k = \frac{C}{\|\mathbf{u}_k\|_2 + \|\mathbf{A}_k\|_F}$, $C_1, C_2, C_3, c_1, c_2, c_3$ are some absolute constants.

Connections to ratios and performance. Theorem 4.8 characterizes the test performance under DP-SGD for both the general distribution and the long-tailed sub-distribution. Our results reveal two distinct generalization pathways: (i) In the high signal-to-noise ratio regime, the model achieves superior generalization by successfully learning robust signal features (Statement (1)). (ii) Conversely, when signal strength is insufficient, the model can still attain satisfactory performance on long-tailed data by memorizing atypical class-specific noise patterns, provided that the noise correlation ratio is high (i.e., the inter-class noise heterogeneity is strong) (Statement (2)).

Furthermore, the theorem quantifies how DP-SGD fundamentally hinders generalization on both data categories. Specifically, the test error upper bounds scale positively with the noise injection variance σ_n. Compared to standard (clean) training ($\sigma_n = 0$), DP-SGD necessitates significantly higher thresholds for either the signal-to-noise ratio or the noise correlation ratio to achieve the same level of generalization, highlighting the inherent utility cost of privacy-preserving optimization.

New Theoretical Findings. This theorem reveals that the performance degradation on long-tailed distributions is fundamentally rooted in the poor weight-feature alignment induced by privacy constraints. We observe that samples whose successful fitting in non-DP training hinges on the alignment between model weights and inherent features (specifically those with a large L) exhibit significantly poorer performance under DP-SGD. Distinguishing our work from Feldman (2020), which focuses on sample frequency as the source of suboptimality, we offer a more nuanced perspective (detailed explanations are provided in the following). We provide a more granular characterization by deriving an

explicit upper bound on the test error. Specifically, our bound is parameterized by the noise variance σ_n and the clipping factor C, thereby elucidating the precise mechanisms through which individual DP-SGD components impact the learning dynamics. In the non-private limit ($\Lambda_k = 1, \sigma_n = 0$), our findings are consistent with the results in Xu & Chen (2025). Furthermore, by specializing the noise to a homogeneous structure, our Statement (1) recovers the upper bound reported in the standard benign overfitting literature (Kou et al., 2023).

To further elucidate the implications of Theorem 4.8, we conduct a comparative analysis to highlight the disproportionate degradation of test accuracy on long-tailed data under privacy constraints.

Disproportionate Sensitivity to Privacy Noise. DP-SGD exerts a more pronounced negative impact on long-tailed data compared to the full-set distribution, that is $\mathcal{L}_{\mathcal{D}_k} \leq \mathcal{L}_{\mathcal{T}_k}$. By comparing the two statements in Theorem 4.8, it is evident that long-tailed generalization is more vulnerable to σ_n. Consider an initial equilibrium in clean training ($\sigma_n = 0$) where $\mathcal{L}_{\mathcal{D}_k} = \mathcal{L}_{\mathcal{T}_k}$, implying a balance between signal-to-noise ratio and noise correlation ratio ($|\mathcal{S}_k|^2\|\mathbf{u}_k\|_2^4 \approx L^2|\mathcal{S}_k|\|\mathbf{A}_k^\top \mathbf{A}_k\|_F^2$). Upon injecting privacy noise σ_n, the error bound for long-tailed data (Statement 2) increases much faster because its corresponding noise term, $n\sigma_n\sqrt{d+L^2}\|\mathbf{A}_k\|_{\text{op}}$, typically dominates the signal-side noise term $n\sigma_n\|\mathbf{u}_k\|_2\sqrt{2\log(2/\delta)}$. This follows from the fact that $\sqrt{d}\|\mathbf{A}_k\|_{\text{op}} \geq \|\mathbf{A}_k\|_F$, making the noise-memorization pathway significantly more fragile under differential privacy.

Vanished Benefits of Atypical Data. Under DP-SGD, the strategy of incorporating "longer-tailed" or more atypical data (i.e., increasing L) fails to enhance test accuracy. In standard training ($\sigma_n = 0$), atypical samples (with larger L) can actually reduce test error by providing distinct class-dependent noise patches that the model can leverage. However, our results show that in the presence of DP-SGD, particularly when $\sqrt{|\mathcal{S}_k|}\Lambda_k\|\mathbf{A}_k^\top \mathbf{A}_k\|_F \leq n\sigma_n\sqrt{d}\|\mathbf{A}_k\|_{\text{op}}$, increasing L no longer reduces the upper bound of the test error. Consequently, the model's performance remains poor even on highly atypical data, negating the potential accuracy gains from introducing such samples into the training set.

Consistency with Existing Empirical Observations. These theoretical findings echo existing empirical evidence. Prior studies have shown that rare subsets in training data are notably difficult for private models to identify during inference (Carlini et al., 2019a). Furthermore, DP-SGD has been observed to cause a disproportionate drop in accuracy for minority or vulnerable subgroups (Bagdasaryan et al., 2019). This phenomenon is fundamentally rooted in the fact that differential privacy inherently impedes the model's ability to memorize the fine-grained, atypical patterns required for classifying rare samples (Feldman, 2020).

Privacy-Utility Trade-off. To satisfy the (ϵ, δ_{DP})-DP requirement, the noise variance σ_n is set as $\mathcal{O}\left(\frac{C\sqrt{T\ln(1/\delta_{DP})}}{n\epsilon}\right)$ according to the advanced composition theorem Dwork et al. (2010). Substituting this into Theorem 4.8, we obtain the following explicit test error bound:

$$\mathcal{L}_{\mathcal{D}_k}(\mathbf{W}^{(T)}) \leq \sum_{j\neq k} \exp\left[-c_1 \cdot \left(\frac{\epsilon|\mathcal{S}_k|\Lambda_k\|\mathbf{u}_k\|_2^2 - C\sqrt{T}\|\mathbf{u}_k\|_2}{\epsilon\sqrt{|\mathcal{S}_j|}\|\mathbf{A}_k^\top \mathbf{A}_j\|_F + C\sqrt{T}\|\mathbf{A}_k\|_F}\right)^2\right]$$

This characterization yields two key insights regarding the impact of differential privacy: (i) a larger privacy budget ϵ (representing a more relaxed privacy guarantee) results in a smaller utility loss; (ii) the privacy protection error scales with the number of iterations T, suggesting that prolonged training under DP-SGD inevitably incurs a higher cost in terms of generalization performance.

5 Experiments

5.1 Experimental Setup

Datasets. For the experiments on real-world data, we utilize the MNIST and CIFAR-10 datasets, which have been widely demonstrated to have atypical long-tailed data (Carlini et al., 2019a; Feldman, 2020) and have been studied for analyzing DP-SGD (see Figure 3 in Appendix A for an illustration).

For the synthetic data experiments, each sample is generated according to the distribution specified in Definition 3.1. Specifically, we parameterize the signal vectors as $u_k = U \cdot \|u_k\|_2 \cdot z_k$ for $k \in [K]$, where z_k is the k-th standard basis vector (with the k-th entry being one and the others zero), and U is a randomly generated unitary matrix. We set the number of classes to $K = 5$ and the signal

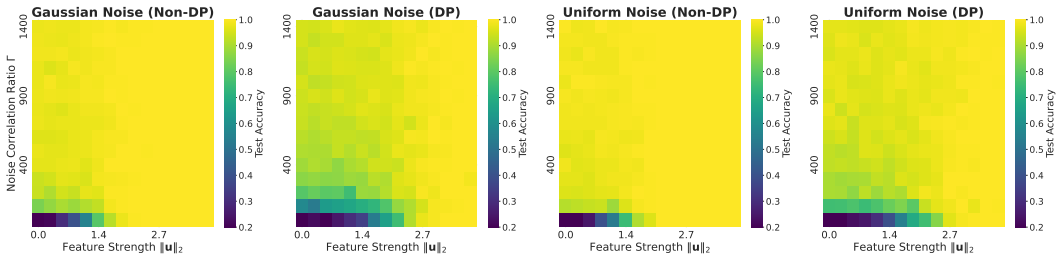
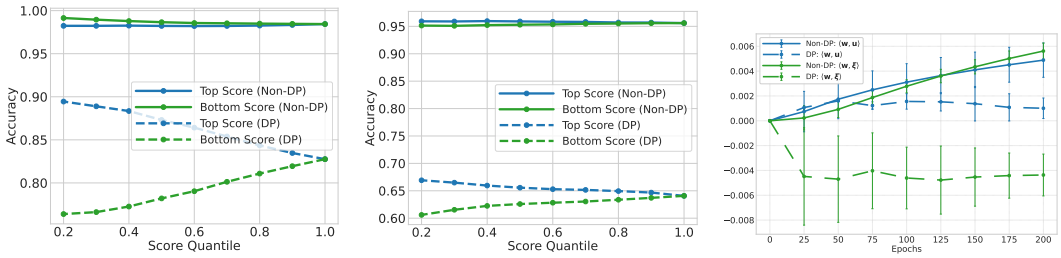


Figure 1: Test accuracy heatmap on synthetic data across feature strength and noise correlation ratio.



(a) MNIST (b) CIFAR-10 (c) Training Dynamics

Figure 2: Experimental results under DP and Non-DP: (a) and (b) show Test Accuracy across Top/Bottom Influence Score Quantiles, (c) shows the overall Training Dynamics.

dimension to $d = 1000$. For simplicity, we assume all signals have the same norm, i.e., $\|u_k\|_2$ is the same for all k. The coordinates ζ_i of the noise vector are drawn independently from either a Gaussian distribution $\mathcal{N}(0, I_d)$ or a Uniform distribution $\mathcal{U}(-\sqrt{3}, \sqrt{3})$. Regarding the matrices $\mathbf{A}_k$, we set all but one of their non-zero eigenvalues to 0.5. The remaining non-fixed eigenvalue is then tuned to vary the noise correlation ratio accordingly. Finally, we generate 500 samples for training and 500 samples for evaluation, setting $|\mathcal{S}_k| = 100$ and ensuring the noise correlation ratio $\sqrt{|\mathcal{S}_k|}\|\mathbf{A}_k^\top \mathbf{A}_k\|_F / \sqrt{|\mathcal{S}_j|}\|\mathbf{A}_k^\top \mathbf{A}_j\|_F$ is identical for all pairs $k, j \in [K], k \neq j$.

Model architectures. For synthetic data experiments, we adopt the two-layer neural network defined in equation 1, where m is set to 100 neurons. For real-world data experiments, We implement LeNet (LeCun et al., 2002) on MNIST and SmoothNets (Remerscheid et al., 2022) on CIFAR-10. Model parameters are initialized following the default initialization method of PyTorch.

Model training & evaluation. All models in our experiments are trained using DP-SGD. We set the privacy budget to $(\epsilon = 8, \delta = 1e-5)$, the gradient clipping threshold to $C = 1.0$, and the learning rate to $\eta = 0.002$. Each model is trained for 20 epochs with a batch size of $B = 256$. Finally, we report the test accuracy on both the entire dataset and the long-tailed subset. All experiments were run on a single workstation equipped with an AMD Ryzen Threadripper PRO 5955WX (16 cores, 32 threads), 128 GB of system RAM, and a NVIDIA RTX A5000 GPU.

5.2 Results Analysis

Impact of DP-SGD on Training Dynamics. We fix the noise correlation ratio $NCR = 1400$ and signal strength $\|\mathbf{u}\|_2 = 0.5$ and conduct experiments on synthetic data under both non-DP and DP settings. And we evaluate the training dynamics of $\langle \mathbf{w}, \mathbf{u} \rangle$ (signal learning) and $\langle \mathbf{w}, \boldsymbol{\xi} \rangle$ (noise pattern memorization)[1] across DP and non-DP regimes. As shown in Figure 2c, DP-SGD markedly suppresses the alignment between weights and inherent features (heterogeneous noise), thereby suppressing noise memorization, which is consistent with our theoretical analysis in Theorem 4.3 and theoretical finding. Notably, while signal learning and noise memorization exhibit comparable magnitudes in the non-DP setting, the introduction of DP causes $\langle \mathbf{w}, \boldsymbol{\xi} \rangle$ significantly lower—and even negative—relative to $\langle \mathbf{w}, \mathbf{u} \rangle$. This disparity confirms that DP disproportionately affects the memo-

[1]We denote $\langle \mathbf{w}, \mathbf{u} \rangle$ as $1/N \max_{r \in [m]} \sum_{i=1}^{N} \langle \mathbf{w}_{i,r}, \mathbf{u} \rangle$ and $\langle \mathbf{w}, \boldsymbol{\xi} \rangle$ as $1/N \max_{r \in [m]} \sum_{i=1}^{N} \langle \mathbf{w}_{i,r}, \boldsymbol{\xi}_i \rangle$

rization mechanism, thereby accounting for the utility loss observed in long-tailed data distributions where memorization is indispensable.

Impact of DP-SGD on test accuracy(synthetic data). We vary the feature strength $\|\mathbf{u}\|_2$ from 0 to 3.8 and the noise correlation ratio NCR from 0 to 1400, conducting a comparative study between DP and non-DP settings on synthetic data. Figure 1 presents the test accuracy heatmaps under different noise distributions (Gaussian and Uniform) and privacy regimes.

Our empirical results yield several key observations. First, the test accuracy increases as both the feature strength and the noise correlation ratio grow, which aligns with the test error upper bounds in Theorem 4.8. Second, it is noteworthy that even in the regime of minimal feature strength (e.g., $\|\mathbf{u}\|_2 \approx 0$), a high noise correlation ratio still results in a low test error; this suggests that noise pattern memorization is the key to achieving high test accuracy in the absence of strong signals.

Finally, a comparison between the DP and non-DP cases reveals that while models trained with DP can still achieve nearly optimal test accuracy when the feature strength is sufficiently large, their performance degrades significantly in the low-feature-strength regime, regardless of the noise correlation ratio. This indicates that DP primarily disrupts the memorization of noise patterns, which is indispensable for maintaining utility on rare, long-tailed samples that lack prominent features, corresponding to Remark 4.

Impact of DP-SGD on test accuracy(real-world data). In real-world datasets, parameters such as signal strength and noise correlation ratios are inherent and immutable. Furthermore, as the data-generating distributions of real-world datasets do not strictly align with the data structure in Definition 3.1, it is unfeasible to directly identify long-tailed samples using the criteria in Definition 4.7. To address this, we adopt a surrogate approach by identifying long-tailed samples as those that significantly enhance the heterogeneity of noise patterns, specifically through their impact on the squared Frobenius norm $\|\mathbf{A}_y^\top \mathbf{A}_y\|_F^2$. Inspired by Xu & Chen (2025), and observing that this norm is equivalent to the Frobenius norm of the covariance matrix for class y, we formalize the influence score of an image $(x, y) \in \mathcal{S}_{test,y}$ as follows:

$$\mathcal{I}(x, y) = \|\hat{\Sigma}(\mathcal{S}_{test,y})\|_F^2 - \|\hat{\Sigma}(\mathcal{S}_{test,y} \setminus \{(x, y)\})\|_F^2, \tag{6}$$

where $\hat{\Sigma}(\mathcal{S}_{test,y})$ denotes the estimated covariance of the underlying data distribution within the subset $\mathcal{S}_{test,y}$. Samples with higher influence scores are thus treated as long-tailed data, as they represent the primary sources of distributional variance and noise heterogeneity.

Figure 2a and 2b illustrates our experimental results on MNIST and CIFAR-10 under both non-DP and DP settings, where we evaluate performance on subsets partitioned by the top and bottom $X\%$ influence scores. In the non-DP regime, the test accuracy remains nearly uniform and optimal across all subsets, suggesting that the noise patterns of long-tailed samples are effectively memorized to maintain high utility. Conversely, the introduction of DP leads to a marked divergence in accuracy, where the performance on the top influence score subset (representing long-tailed data) is significantly lower than that on the bottom subset; this disparity demonstrates that DP disproportionately impairs the utility of long-tailed samples that rely on noise pattern memorization, thereby empirically validating our theoretical results in Theorem 4.8 and Remark 4.

6 Conclusion

This paper presents the first analysis of the training dynamics and generalization of two-layer neural networks trained via DP-SGD on long-tailed data. Our theoretical results demonstrate that DP suppresses the memorization of noise patterns, leading to sub-optimal training loss and a disproportionately more severe impact on the test error of long-tailed samples compared to the overall population. Empirical evaluations on both synthetic and real-world datasets validate our theoretical findings, exhibiting a high degree of consistency between the observed training dynamics, test accuracies, and our theoretical predictions.

Limitations. One limitation of this paper is the reliance on Two-layer CNNs. One future direction is to generalize our analysis to deeper model architecture.

AI USE STATEMENT

AI-based tools were used during the preparation of this manuscript for limited purposes, including language editing, proofreading, and assisting with code implementation. All mathematical derivations, theoretical results, algorithmic designs, and scientific conclusions were developed by the authors. All AI-assisted code was manually reviewed, verified, and validated by the authors, and all experimental implementations and empirical analyses were independently checked by the authors.

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.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pp. 383–392. PMLR, 2021.

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

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021.

Shuochen Bi and Yufan Lian. Advanced portfolio management in finance using deep learning and artificial intelligence techniques: Enhancing investment strategies through machine learning models. *Journal of Artificial Intelligence Research*, 4(1):233–298, 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.

Nicholas Carlini, Ulfar Erlingsson, and Nicolas Papernot. Distribution density, tails, and outliers in machine learning: Metrics and applications. *arXiv preprint arXiv:1910.13427*, 2019a.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019b.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of medical imaging and radiation oncology*, 65(5):545–563, 2021.

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.

Meng Ding, Mingxi Lei, Shaopeng Fu, Shaowei Wang, Di Wang, and Jinhui Xu. Understanding private learning from feature perspective. *arXiv preprint arXiv:2511.18006*, 2025.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st annual symposium on foundations of computer science*, pp. 51–60. IEEE, 2010.

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.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 227–236, 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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift fuer medizinische Physik*, 29(2):102–127, 2019.

Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied soft computing*, 93:106384, 2020.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 189–199, 2020.

Nicolas W. Remerscheid, Alexander Ziller, Daniel Rueckert, and Georgios Kaissis. Smoothnets: Optimizing cnn architecture design for differentially private deep learning, 2022. URL `https://arxiv.org/abs/2205.04095`.

David Ruppert. The elements of statistical learning: data mining, inference, and prediction, 2004.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1182–1189, 2019.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10081–10091. PMLR, 2020.

Ruichen Xu and Kexin Chen. Rethinking benign overfitting in two-layer neural networks. *arXiv preprint arXiv:2502.11893*, 2025.

Ruichen Xu and Kexin Chen. Differential privacy in two-layer networks: How dp-sgd harms fairness and robustness. *arXiv preprint arXiv:2603.04881*, 2026.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1524–1534. PMLR, 2019.

A Illustration of long-tailed data

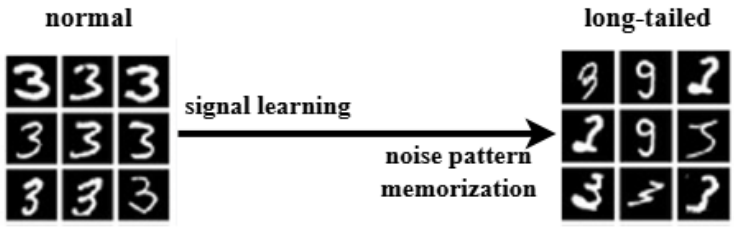


Figure 3: Illustration of long-tailed data on MNIST

Figure 3 illustrates both normal and long-tailed data from the MNIST dataset (digit 3). The samples on the left represent normal data that rely on signal learning. Conversely, the samples on the right represent atypical long-tailed data that rely on noise pattern memorization.

Remark A.1 (Justification for Definition 4.7). We think that the essence of long-tailed data is atypicality rather than mere frequency. Tail samples possess unique noise patterns $\boldsymbol{\xi}$ that do not align with the global feature $\mathbf{u}$. Definition 4.7 identifies this "intrinsic tail" by selecting samples residing in the extreme tail of the noise distribution (i.e., $\zeta' \geq L$). Given the sub-Gaussian nature of the noise, the probability of such samples occurring is exponentially small, which formally characterizes their rare property. Crucially, these samples rely fundamentally on the memorization channel to be recognized, explaining their high sensitivity to privacy-preserving training.

B Notations

We use lowercase letters, lowercase boldface letters, and uppercase boldface letters to denote scalars, vectors, and matrices, respectively. For two sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ if there exist absolute constants $C > 0$ and $N > 0$ such that $|x_n| \leq C|y_n|$ for all $n \geq N$. Similarly, we write $x_n = \Omega(y_n)$ if $y_n = O(x_n)$. We say $x_n = \Theta(y_n)$ if both $x_n = O(y_n)$ and $x_n = \Omega(y_n)$. Finally, we use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$ to denote the corresponding notations with logarithmic factors suppressed.

C Basic lemmas

In this section, we show the basic concentration inequality needed for future proofs. We follow the inequality from Xu & Chen (2025), along with several additional inequalities required for DP-SGD.

Lemma C.1 (Lemma B.1 in Xu & Chen (2025)). *Let $\mathbf{Z} \in \mathbb{R}^{m\times n}(m > n)$ be a matrix whose entries are independent and identically distributed Gaussian variables, i.e., $\mathbf{Z}_{i,j} \sim \mathcal{N}(0,1)$ for all $i \in [m], j \in [n]$. With probability at least $1-\delta$, all singular values of $\mathbf{Z}$, $\lambda_i(\mathbf{Z})$, for all $i \in [n]$ satisfies*

$$\lambda_i(\mathbf{Z}) \geq \sqrt{m} - 2\sqrt{8\log(\frac{2}{\delta}) + 8\log(9)n}. \tag{7}$$

Lemma C.1 follows from the concentration inequality of Gaussian random matrices Vershynin (2018).

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

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \geq t\sigma_0\right], \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \leq -t\sigma_0\right] \leq \exp\left(-\frac{t^2}{2}\right), \forall t \geq 0. \tag{8}$$

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 $-\sqrt{\frac{2}{\pi}}$ and a Gaussian delay positive tail. Then, by Hoeffding's inequality, we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \geq t\sigma_0\right], \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i| - \sqrt{\frac{2}{\pi}}\sigma_0 \leq -t\sigma_0\right] \leq \exp\left(-\frac{t^2}{2}\right), \forall t \geq 0. \tag{9}$$

□

Lemma C.3 (Lemma B.4 in Xu & Chen (2025)). *Suppose a zero-mean random vector $\mathbf{x} \in \mathbb{R}^d$ is generated as $\mathbf{x} = \mathbf{A}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$'s each coordinate is independent, symmetric, and σ_p sub-Gaussian with $\mathbb{E}[\zeta_j^2] = 1$, for any $j \in [d]$. Then, $\mathbf{x}$ satisfies*

$$\mathbb{P}\left[\|\mathbf{x}\|_2^2 - \mathrm{Tr}(\mathbf{A}^\top\mathbf{A}) \geq t\right] \leq 2\exp\left(-\Omega\left(\min\left\{\frac{t^2}{\|\mathbf{A}^\top\mathbf{A}\|_F^2\sigma_p^4}, \frac{t}{\|\mathbf{A}^\top\mathbf{A}\|_{op}\sigma_p^2}\right\}\right)\right). \tag{10}$$

Proof. We have

$$\|\mathbf{x}\|_2^2 = \langle\mathbf{A}\boldsymbol{\zeta}_1, \mathbf{A}\boldsymbol{\zeta}_1\rangle = \boldsymbol{\zeta}_1^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\zeta}_1 = \mathrm{Tr}\left(\mathbf{A}^\top\mathbf{A}\boldsymbol{\zeta}_1\boldsymbol{\zeta}_1^\top\right). \tag{11}$$

Then, expectation of $\|\mathbf{x}\|_2$ satisfies

$$\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] = \mathrm{Tr}(\mathbf{A}^\top\mathbf{A}). \tag{12}$$

By Bernstein's inequality, we have

$$\mathbb{P}\left[\|\mathbf{x}\|_2^2 - \mathrm{Tr}(\mathbf{A}^\top\mathbf{A}) \geq t\right] \leq 2\exp\left(-\Omega\left(\min\left\{\frac{t^2}{\|\mathbf{A}^\top\mathbf{A}\|_F^2\sigma_p^4}, \frac{t}{\|\mathbf{A}^\top\mathbf{A}\|_{\mathrm{op}}\sigma_p^2}\right\}\right)\right). \tag{13}$$

□

Lemma C.4. *Let $\mathbf{n}^{(t)} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I}_d)$ and $\mathbf{u} \in \mathbb{R}^d$ be a fixed vector. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$|\langle\mathbf{n}^{(t)}, \mathbf{u}\rangle| \leq \sigma_n\|\mathbf{u}\|_2\sqrt{2\log(2/\delta)}. \tag{14}$$

Proof. Let $Z = \langle\mathbf{n}^{(t)}, \mathbf{u}\rangle$. Since $\mathbf{n}^{(t)}$ is a Gaussian vector, Z is a zero-mean Gaussian random variable with variance $\mathbb{E}[Z^2] = \mathbf{u}^\top\mathbb{E}[\mathbf{n}^{(t)}(\mathbf{n}^{(t)})^\top]\mathbf{u} = \sigma_n^2\|\mathbf{u}\|_2^2$. Specifically, Z is $\sigma_n\|\mathbf{u}\|_2$-sub-Gaussian. Applying the standard sub-Gaussian tail bound :

$$\mathbb{P}(|Z| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma_n^2\|\mathbf{u}\|_2^2}\right). \tag{15}$$

Setting $\epsilon = \sigma_n\|\mathbf{u}\|_2\sqrt{2\log\frac{2}{\delta}}$ completes the proof. □

Lemma C.5. *Let $\mathbf{n}^{(t)} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_d)$, and let $\boldsymbol{\zeta} \in \mathbb{R}^d$ be independent of $\mathbf{n}^{(t)}$, whose coordinates are independent, zero-mean, and σ_p-sub-Gaussian. Let*

$$\boldsymbol{\xi} = A\boldsymbol{\zeta}$$

for a fixed matrix $A \in \mathbb{R}^{d\times d}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\left\langle \mathbf{n}^{(t)}, \boldsymbol{\xi} \right\rangle\right| \le C\sigma_n\sigma_p \left(\|A\|_F \sqrt{\log\left(\frac{4}{\delta}\right)} + \|A\|_{\mathrm{op}} \log\left(\frac{4}{\delta}\right) \right), \tag{16}$$

where $C > 0$ is a universal constant.

Proof. Since $\boldsymbol{\xi} = A\boldsymbol{\zeta}$, we have

$$\left\langle \mathbf{n}^{(t)}, \boldsymbol{\xi} \right\rangle = \boldsymbol{\zeta}^\top A^\top \mathbf{n}^{(t)}. \tag{17}$$

Conditioned on $\mathbf{n}^{(t)}$, the vector $A^\top \mathbf{n}^{(t)}$ is fixed. Since the coordinates of $\boldsymbol{\zeta}$ are independent, zero-mean, and σ_p-sub-Gaussian, the random variable $\boldsymbol{\zeta}^\top A^\top \mathbf{n}^{(t)}$ is $\sigma_p \|A^\top \mathbf{n}^{(t)}\|_2$-sub-Gaussian. Therefore, with conditional probability at least $1-\delta/2$,

$$\left|\boldsymbol{\zeta}^\top A^\top \mathbf{n}^{(t)}\right| \le \sigma_p \|A^\top \mathbf{n}^{(t)}\|_2 \sqrt{2\log\left(\frac{4}{\delta}\right)}. \tag{18}$$

It remains to control $\|A^\top \mathbf{n}^{(t)}\|_2$. Write $\mathbf{n}^{(t)} = \sigma_n \mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then

$$\|A^\top \mathbf{n}^{(t)}\|_2 = \sigma_n \|A^\top \mathbf{g}\|_2. \tag{19}$$

By the standard concentration inequality for the norm of a Gaussian linear transformation, with probability at least $1-\delta/2$,

$$\|A^\top \mathbf{g}\|_2 \le \|A\|_F + \|A\|_{\mathrm{op}} \sqrt{2\log\left(\frac{4}{\delta}\right)}. \tag{20}$$

Hence,

$$\|A^\top \mathbf{n}^{(t)}\|_2 \le \sigma_n \left(\|A\|_F + \|A\|_{\mathrm{op}} \sqrt{2\log\left(\frac{4}{\delta}\right)} \right). \tag{21}$$

Combining equation 18 and equation 20, and taking a union bound over the two events, we obtain

$$\begin{aligned} &\left|\left\langle \mathbf{n}^{(t)}, \boldsymbol{\xi} \right\rangle\right| \\ &\le \sigma_n\sigma_p \left(\|A\|_F + \|A\|_{\mathrm{op}} \sqrt{2\log\left(\frac{4}{\delta}\right)} \right) \sqrt{2\log\left(\frac{4}{\delta}\right)} \\ &\le C\sigma_n\sigma_p \left(\|A\|_F \sqrt{\log\left(\frac{4}{\delta}\right)} + \|A\|_{\mathrm{op}} \log\left(\frac{4}{\delta}\right) \right). \end{aligned} \tag{22}$$

This completes the proof. □

Lemma C.6 (Lemma B.6 in Xu & Chen (2025))**.** *Let $x_1, \ldots, x_m$ be m independent zero-mean Gaussian variables. Denote z_i as indicators for signs of x_i, i.e., for all $i \in [m]$,*

$$z_i = \begin{cases} 1, & x_i > 0, \\ 0, & x_i \le 0. \end{cases} \tag{23}$$

Then, we have

$$\Pr\left[\sum_{i=1}^m z_i \ge 0.4m\right] \ge 1 - \exp\left(-\frac{8}{25}m\right). \tag{24}$$

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

$$\Pr\left[m\left(\frac{1}{m}\sum_{i=1}^{m} z_i\right) \leq m\left(\frac{1}{2}-\epsilon\right)\right] \leq \exp\left(-\frac{2m^2\epsilon^2}{m(1/16)}\right). \tag{25}$$

Let $\epsilon = 0.1$, we have

$$\Pr\left[\sum_{i=1}^{m} z_i \leq 0.4m\right] \leq \exp\left(-\frac{8}{25}m\right). \tag{26}$$

Therefore, we have

$$\Pr\left[\sum_{i=1}^{m} z_i \geq 0.4m\right] \geq 1-\exp\left(-\frac{8}{25}m\right). \tag{27}$$

This completes the proof. □

Lemma C.7 (Lemma B.7 in Xu & Chen (2025)). *For any constant $t \in (0,1]$ and $x \in [-a,b], a,b > 0$, we have*

$$\log(1+t(\exp(x)-1)) \leq \Gamma(x)x, \tag{28}$$

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

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

$$\frac{\partial \log(1+t(\exp(x)-1))}{\partial t} = \frac{\exp(x)-1}{1+t(\exp(x)-1)} \geq 0. \tag{29}$$

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

$$\frac{\partial^2 \log(1+t(\exp(x)-1))}{\partial x^2} = \frac{(1-t)t\exp(x)}{[1+t(\exp(x)-1)]^2} \geq 0. \tag{30}$$

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

$$\log(1+t(\exp(x)-1)) \leq \frac{\log(1+t(\exp(-a)-1))}{-a}x, \forall x < 0. \tag{31}$$

This completes the proof. □

Lemma C.8 (Lemma B.11 in Xu & Chen (2025)). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with rank M. Suppose $\delta > 0$ and $m \geq \Omega(\log(n/\delta)(\lambda^+_{\max}(\mathbf{A}))^2/((\lambda^+_{\min}(\mathbf{A}))^2))$. With probability $1-\delta$, for any orthant $\mathcal{T} \in \mathbb{R}^m$ and vector $\mathbf{x} \in \mathbb{R}^n$ is a random sub-Gaussian vector with each coordinate follows $\mathcal{D}_\xi$, we have*

$$\mathbb{P}[\mathbf{A}\mathbf{x} \in \mathcal{T}] \leq (0.6)^M. \tag{32}$$

Proof. Using singular value decomposition for $\mathbf{A}$, for any orthant $\mathcal{T} \in \mathbb{R}^m$, we have

$$\mathbb{P}[\mathbf{A}\mathbf{x} \in \mathcal{T}] = \mathbb{P}[\mathbf{U}\mathbf{\Sigma}\mathbf{x} \in \mathcal{T}]. \tag{33}$$

Without loss of generality, we assume the orthant $\mathcal{T}$ is that with all positive entries. Then, we have

$$\mathbb{P}[\mathbf{A}\mathbf{x} \in \mathcal{T}] \leq \mathbb{P}[\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\mathbf{x} \in \mathcal{T}], \tag{34}$$

where

$$\tilde{\mathbf{U}} = \begin{bmatrix}\mathbf{1}\cdot 1/\sqrt{m} & \tilde{\mathbf{U}}_2 & \dots & \tilde{\mathbf{U}}_m\end{bmatrix}, \tag{35}$$

and

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix}\lambda^+_{\max}(\mathbf{A}) & 0 & \dots & 0 \\ 0 & \lambda^+_{\min}(\mathbf{A}) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda^+_{\min}(\mathbf{A}) \\ \dots & \dots & \dots & 0\end{bmatrix}, \tag{36}$$

Here, $\tilde{\mathbf{\Sigma}}$ is generated by replacing the non-zero singular values other than the largest one with $\lambda^+_{\min}(\mathbf{A})$. $\mathbf{1}$ is a vector with all one entries. In the following, we abbreviate $\lambda^+_{\max}(\mathbf{A})$ and $\lambda^+_{\min}(\mathbf{A})$ as $\lambda^+_{\max}$ and $\lambda^+_{\min}$ for convenience. Then, we have

$$\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\mathbf{x} = \lambda^+_{\max} \cdot \frac{1}{\sqrt{m}} \cdot \mathbf{1} \cdot x_1 + \lambda^+_{\min} \sum_{i=2}^{M} x_i \tilde{\mathbf{U}}_i. \tag{37}$$

Here, since each coordinate of $\mathbf{x}$ is drawn from $\mathcal{D}_\xi$ and is σ_p-sub-Gaussian, a standard tail bound and a union bound imply that, with probability at least $1-\delta$,

$$|x_i| \le \sqrt{2\sigma_p^2 \log\left(\frac{2n}{\delta}\right)}, \tag{38}$$

with probability $1-\delta$. Then, we have

$$\begin{aligned}
\mathbb{P}[\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\mathbf{x} \in \mathcal{T}] &= \mathbb{P}\left[(\lambda^+_{\max} - \lambda^+_{\min}) \cdot \frac{1}{\sqrt{m}} \cdot \mathbf{1} \cdot x_1 + \lambda^+_{\min} \sum_{i=1}^{M} x_i \tilde{\mathbf{U}}_i \in \mathcal{T}\right] \\
&\le \mathbb{P}\left[\lambda^+_{\max} \cdot \frac{1}{\sqrt{m}} \cdot \mathbf{1} \cdot \sqrt{2\sigma_p^2 \log\left(\frac{2n}{\delta}\right)} + \lambda^+_{\min} \sum_{i=1}^{M} x_i \tilde{\mathbf{U}}_i \in \mathcal{T}\right] \\
&= \mathbb{P}\left[\mathbf{1} \cdot \frac{\sqrt{2\sigma_p^2 \log(2n/\delta)}\lambda^+_{\max}}{\sqrt{m}} + \lambda^+_{\min}\mathbf{x} \in \mathcal{T}\right].
\end{aligned} \tag{39}$$

As the entries of $\mathbf{x}$ are independent, we can bound each entry independently. Then, when $m \ge \Omega(\log(n/\delta)(\lambda^+_{\max})^2/((\lambda^+_{\min})^2))$, we have

$$\mathbb{P}\left[\lambda^+_{\min} x_i + \frac{\sqrt{2\sigma_p^2 \log(2n/\delta)}\lambda^+_{\max}}{\sqrt{m}} < 0\right] \ge 0.4, \tag{40}$$

by the property of $\mathcal{D}_\xi$, resulting in

$$\mathbb{P}[\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\mathbf{x} \in \mathcal{T}] \le (0.6)^M. \tag{41}$$

This completes the proof. □

Lemma C.9 (Lemma B.12 in Xu & Chen (2025)). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with rank M. Suppose $m \ge \Omega(\log(n/\delta)(\lambda^+_{\max}(\mathbf{A}))^2/((\lambda^+_{\min}(\mathbf{A}))^2))$. With probability $1-\delta$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{m} \mathbb{I}((\mathbf{A}\mathbf{x})_i \le 0) \ge 0.9m\right] \le \exp(-0.07m). \tag{42}$$

Proof. Let N_o be the number of orthants in $\mathbb{R}^m$ with at least $0.9m$ negative coordinates. For i.i.d. $\varepsilon_i \sim \text{Bernoulli}(1/2)$, Hoeffding's inequality gives

$$\frac{N_o}{2^m} = \mathbb{P}\left[\sum_{i=1}^{m} \varepsilon_i \ge 0.9m\right] \le \exp(-2m(0.4)^2) = \exp(-0.32m),$$

and hence

$$N_o \le 2^m \exp(-0.32m) = \exp\big((\log 2 - 0.32)m\big).$$

Then, by Lemma C.8, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{m} \mathbb{I}((\mathbf{A}\mathbf{x})_i \le 0) \ge 0.9m\right] &\le 0.6^M \cdot \exp((\log 2 - 0.32)m) \\
&= \exp(\log(0.6)M + (\log 2 - 0.32)m) \\
&\le \exp(-0.51M + 0.38m).
\end{aligned} \tag{43}$$

As long as $M \geq 0.9m$, we have

$$\mathbb{P}\left[\sum_{i=1}^{m} \mathbb{I}((\mathbf{A}\mathbf{x})_i \leq 0) \geq 0.9m\right] \leq \exp(-0.07m). \tag{44}$$

This completes the proof. □

Lemma C.10 (Lemma B.13 in Xu & Chen (2025)). *Suppose that* $\delta > 0$ *and* $\mathrm{Tr}(\mathbf{A}_i^\top \mathbf{A}_i) = \Omega\left(\max\left\{\left(\|\mathbf{A}_i^\top \mathbf{A}_j\|_F^2 \sigma_p^4 \log(6n/\delta)\right)^{1/2}, \|\mathbf{A}_i^\top \mathbf{A}_j\|_{op} \sigma_p^2 \log(6n/\delta)\right\}\right)$. *For all* $i, j \in [K]$, *with probability* $1-\delta$, *we have*

$$\frac{1}{2}\mathrm{Tr}(\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}) \leq \|\boldsymbol{\xi}_i\|_2^2 \leq \frac{3}{2}\mathrm{Tr}(\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}), \tag{45}$$

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_j \rangle| \leq \mathcal{O}\left(\max\left\{\left(\|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_F^2 \log\left(\frac{6n^2}{\delta}\right)\right)^{1/2}, \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_{op} \log\left(\frac{6n^2}{\delta}\right)\right\}\right). \tag{46}$$

Proof. By Lemma C.3, with probability at least $1-\delta/3n$, there exists a constant $C_1 > 0$ such that

$$\|\boldsymbol{\xi}_i\|_2^2 \leq \mathrm{Tr}(\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}) + \max\left\{\left(\|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}\|_F^2 \sigma_p^4 \frac{1}{C_1} \log\left(\frac{6n}{\delta}\right)\right)^{1/2}, \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}\|_{\mathrm{op}} \sigma_p^2 \frac{1}{C_1} \log\left(\frac{6n}{\delta}\right)\right\},$$

and

$$\|\boldsymbol{\xi}_{y_i}\|_2^2 \geq \mathrm{Tr}(\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}) - \max\left\{\left(\|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}\|_F^2 \sigma_p^4 \frac{1}{C_1} \log\left(\frac{6n}{\delta}\right)\right)^{1/2}, \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_i}\|_{\mathrm{op}} \sigma_p^2 \frac{1}{C_1} \log\left(\frac{6n}{\delta}\right)\right\}.$$

For two distinct samples, write $\boldsymbol{\xi}_{y_i} = \mathbf{A}_{y_i}\boldsymbol{\zeta}_i, \quad \boldsymbol{\xi}_{y_j} = \mathbf{A}_{y_j}\boldsymbol{\zeta}_j$, where $\boldsymbol{\zeta}_i$ and $\boldsymbol{\zeta}_j$ are independent. Then,

$$\left\langle \boldsymbol{\xi}_{y_i}, \boldsymbol{\xi}_{y_j} \right\rangle = \left\langle \mathbf{A}_{y_i}\boldsymbol{\zeta}_i, \mathbf{A}_{y_j}\boldsymbol{\zeta}_j \right\rangle = \boldsymbol{\zeta}_i^\top \mathbf{A}_{y_i}^\top \mathbf{A}_{y_j} \boldsymbol{\zeta}_j = \mathrm{Tr}\left(\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j} \boldsymbol{\zeta}_j \boldsymbol{\zeta}_i^\top\right).$$

Since the coordinates of $\boldsymbol{\zeta}_i$ and $\boldsymbol{\zeta}_j$ are independent, symmetric, and σ_p-sub-Gaussian, the bilinear Bernstein inequality implies

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{\xi}_{y_i}, \boldsymbol{\xi}_{y_j} \right\rangle\right| \geq t\right] \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sigma_p^4 \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_F^2}, \frac{t}{\sigma_p^2 \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_{\mathrm{op}}}\right\}\right)$$

for an absolute constant $c > 0$. Taking

$$t = C\max\left\{\sigma_p^2 \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_F \sqrt{\log\left(\frac{6n^2}{\delta}\right)}, \sigma_p^2 \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_{\mathrm{op}} \log\left(\frac{6n^2}{\delta}\right)\right\}$$

for a sufficiently large constant $C > 0$ gives

$$\left|\left\langle \boldsymbol{\xi}_{y_i}, \boldsymbol{\xi}_{y_j} \right\rangle\right| \leq C\max\left\{\sigma_p^2 \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_F \sqrt{\log\left(\frac{6n^2}{\delta}\right)}, \sigma_p^2 \|\mathbf{A}_{y_i}^\top \mathbf{A}_{y_j}\|_{\mathrm{op}} \log\left(\frac{6n^2}{\delta}\right)\right\}$$

with probability at least $1-\delta/(3n^2)$. A union bound over all sample pairs, together with the preceding norm bounds, completes the proof.

□

Lemma C.11 (Lemma B.14 in Xu & Chen (2026)). *Suppose that* $d = \Omega(\log(mn/\delta))$ *and* $m = \Omega(\log(1/\delta))$. *With probability at least* $1-\delta$, *for all* $r \in [m], j \in [2], i \in [n]$,

$$|\langle \mathbf{w}_{j,r}^{(0)}, \mathbf{u}_k \rangle| \leq \sqrt{2\log\left(\frac{4Km}{\delta}\right)} \|\mathbf{u}_k\|_2 \sigma_0, \tag{47}$$

$$|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \leq \mathcal{O}\left(\log\left(\frac{Km}{\delta}\right) \|\mathbf{A}_{y_i}\|_F \sigma_0\right). \tag{48}$$

Proof. We prove the first bound (47) with Hoeffding's inequality. For all $j \in [K], r \in [m], i \in [n]$, with probability $1 - \delta/(2Km)$,

$$|\langle \mathbf{w}_{j,r}^{(0)}, \mathbf{u}_k \rangle| \leq \sqrt{2 \log\left(\frac{4Km}{\delta}\right)} \|\mathbf{u}_k\|_2 \sigma_0. \tag{49}$$

Then, we prove the second bound (48) leveraging Lemma B.5. For all $j \in [K], r \in [m], i \in [n]$, with probability $1 - \delta/(2Km)$,

$$\begin{aligned} |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| &\leq \mathcal{O}\left(\max\left\{\sqrt{\log\left(\frac{Km}{\delta}\right)} \|\mathbf{A}_{y_i}\|_F, \log\left(\frac{Km}{\delta}\right) \|\mathbf{A}_{y_i}\|_{\text{op}}\right\} \sigma_0\right) \\ &\leq \mathcal{O}\left(\log\left(\frac{Km}{\delta}\right) \|\mathbf{A}_{y_i}\|_F \sigma_0\right). \end{aligned} \tag{50}$$

□

Lemma C.12 (Lemma B.15 in Xu & Chen (2025))**.** *Suppose $\delta > 0$ and $m \geq \Omega(\log(n/\delta))$. With probability at least $1 - \delta$, we have*

$$|\mathcal{S}_i^{(0)}| \geq 0.4m. \tag{51}$$

Lemma C.12 follows from Lemma C.6 and union bound.

Lemma C.13 (Lemma B.16 in Xu & Chen (2025))**.** *Suppose $\delta > 0$ and $n \geq \Omega(\log(m/\delta))$. For any neuron $\mathbf{w}_{j,r}^{(0)}, j \in [2], r \in [m]$, with probability at least $1 - \delta$, we have*

$$\sum_{i=1}^{n} \mathbb{I}(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle) \geq 0.4n. \tag{52}$$

Lemma C.13 follows from Lemma C.6 and union bound.

D Training Loss analysis

In this section, we analyze the training loss. These results are based on the high probability conclusions in Appendix C. Due to the Assumption 4.2, we consider that the training data $(\mathbf{x}, y) \in \mathcal{S}$ satisfies $1 - \text{logit}_y(\mathbf{W}, \mathbf{x}) \geq 1 - \exp(-s) = \Theta(1)$ during all training process.

D.1 Network Gradient

The mini-batch gradient on the neuron $\mathbf{w}_{k,r}$ at iteration t is

$$\begin{aligned} \nabla_{\mathbf{w}_{k,r}^{(t)}} \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) = &- \frac{1}{mB} \sum_{(\mathbf{x},y) \in \mathcal{S}^{(t)}} \left[\mathbb{I}(y = k)(1 - \text{logit}_k(\mathbf{W}^{(t)}, \mathbf{x})) \sum_{j=1}^{2} \sigma'\left(\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}^{(j)} \rangle\right) \mathbf{x}^{(j)} \right] \\ &+ \frac{1}{mB} \sum_{(\mathbf{x},y) \in \mathcal{S}^{(t)}} \left[\mathbb{I}(y \neq k) \, \text{logit}_k(\mathbf{W}^{(t)}, \mathbf{x}) \sum_{j=1}^{2} \sigma'\left(\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}^{(j)} \rangle\right) \mathbf{x}^{(j)} \right]. \end{aligned} \tag{53}$$

D.2 Bound of the Clipping Multiplier $h(C, \mathbf{x}, y)$

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

$$\left\| \nabla_{\mathbf{W}^{(t)}} \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) \right\|_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{m}} \cdot \left(\|\mathbf{u}_k\|_2 + \sqrt{Tr(A_k^T A_k)}\right)\right). \tag{54}$$

Lemma D.1 follows from Lemma C.3.

For convenience, we first define the clipping multiplier of data $(\mathbf{x}, y)$ as

$$h(C, \mathbf{x}, y) = \frac{1}{\max\left\{1, \frac{\|\nabla \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y)\|_2}{C}\right\}}. \tag{55}$$

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

By definition (55), we know that

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

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

$$h(C, \mathbf{x}, y) \geq \Omega\left(\frac{C\sqrt{m}}{\|\mathbf{u}_k\|_2 + \sqrt{Tr(A_k^T A_k)}}\right).$$

D.3 Training Loss Bounds

First, we characterize the training loss.

$$\begin{aligned}
&\mathcal{L}(\mathbf{W}^{(t+1)}, \mathbf{x}, y) - \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) \\
&= -\log\left(\mathrm{prob}_y\left(\mathbf{W}^{(t+1)}, \mathbf{x}\right)\right) + \log\left(\mathrm{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right) \\
&= \log\left(\frac{\mathrm{prob}_y(\mathbf{W}^{(t)}, \mathbf{x})}{\mathrm{prob}_y(\mathbf{W}^{(t+1)}, \mathbf{x})}\right) \\
&= \log\left(\frac{\exp\left(F_y^{(t)}(\mathbf{x})\right) / \left(\exp\left(F_y^{(t)}(\mathbf{x})\right) + \sum_{j\neq y}\exp\left(F_j^{(t)}(\mathbf{x})\right)\right)}{\exp\left(F_y^{(t+1)}(\mathbf{x})\right) / \left(\exp\left(F_y^{(t+1)}(\mathbf{x})\right) + \sum_{j\neq y}\exp\left(F_j^{(t+1)}(\mathbf{x})\right)\right)}\right) \\
&= \log\left(\frac{1 + \sum_{j\neq y}\exp\left(F_j^{(t)}(\mathbf{x}) - F_y^{(t)}(\mathbf{x})\right)\exp\left(\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right)}{1 + \sum_{j\neq y}\exp\left(F_j^{(t)}(\mathbf{x}) - F_y^{(t)}(\mathbf{x})\right)}\right) \\
&= \log\left(1 + \frac{\sum_{j\neq y}\exp\left(F_j^{(t)}(\mathbf{x}) - F_y^{(t)}(\mathbf{x})\right)\left(\exp\left(\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right) - 1\right)}{1 + \sum_{j\neq y}\exp\left(F_j^{(t)}(\mathbf{x}) - F_y^{(t)}(\mathbf{x})\right)}\right) \\
&\leq \log\left(1 + \left(1 - \mathrm{prob}_y\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right)\max_{j\neq y}\left(\exp\left(\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right) - 1\right)\right),
\end{aligned}$$

where $\Delta_y^{(t)}(\mathbf{x}) = F_y^{(t+1)}(\mathbf{x}) - F_y^{(t)}(\mathbf{x})$, $\Delta_j^{(t)}(\mathbf{x}) = F_j^{(t+1)}(\mathbf{x}) - F_j^{(t)}(\mathbf{x})$.

By Lemma C.7, we have

$$\mathcal{L}\left(\mathbf{W}^{(t+1)}, \mathbf{x}, y\right) - \mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{x}, y\right) \leq \Theta(1) \cdot \max_{j\neq y}\left(\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right). \tag{56}$$

To measure how training loss changes over iterations, we need to characterize the change of $\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$.

We first rearrange $\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$ as follows.

$$
\begin{aligned}
&\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) \\
=&\frac{1}{m}\sum_{r=1}^{m}\sum_{i=1}^{2}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{x}^{(i)}\rangle\right) - \sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{x}^{(j)}\rangle\right)\right] \\
&- \frac{1}{m}\sum_{r=1}^{m}\sum_{i=1}^{2}\left[\sigma\left(\langle \mathbf{w}_{y,r}^{(t+1)}, \mathbf{x}^{(i)}\rangle\right) - \sigma\left(\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{x}^{(i)}\rangle\right)\right] \\
=&\underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{u}_y\rangle\right) - \sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle\right)\right]}_{A} + \underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}\rangle\right) - \sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle\right)\right]}_{B} \\
&\underbrace{- \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{y,r}^{(t+1)}, \mathbf{u}_y\rangle\right) - \sigma\left(\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_y\rangle\right)\right]}_{C} \underbrace{- \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{y,r}^{(t+1)}, \boldsymbol{\xi}\rangle\right) - \sigma\left(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle\right)\right]}_{D},
\end{aligned}
\tag{57}
$$

We then bound A, B, C, D

Bound of A

$$
\begin{aligned}
A =& \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}_y}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle)h(C,\mathbf{x}_i,y_i)\,\mathrm{logit}_j(\mathbf{W}^{(t)},\mathbf{x}_i)\mathbf{u}_y \right.\right.\right. \\
&\left.\left.\left. +\eta\cdot\mathbf{n}_{j,r}^{(t)}, \mathbf{u}_y\right\rangle\right)\right] - \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle\right)\right] \\
\le& \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t)} + \eta\cdot\mathbf{n}_{j,r}^{(t)}, \mathbf{u}_y\rangle\right)\right] - \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle\right)\right] \\
\le& \frac{1}{m}\sum_{r=1}^{m}\left[\left|\langle \eta\cdot\mathbf{n}_{j,r}^{(t)}, \mathbf{u}_y\rangle\right|\right] \\
\le& \mathcal{O}(\eta\sigma_n\sqrt{d}\|\mathbf{u}_y\|_2)
\end{aligned}
$$

where the first inequality is obtained by the monotonicity of ReLU activation function; the second inequality is because ReLU function is 1-Lipschitz continuous; the last inequality is due to Lemma C.2.

Bound of B

$$
\begin{aligned}
B =& \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}^{(t)}\setminus\mathcal{S}_j^{(t)}}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)h(C,\mathbf{x}_i,y_i)\,\mathrm{logit}_j(\mathbf{W}^{(t)},\mathbf{x}_i)\boldsymbol{\xi}_i \right.\right.\right. \\
&\left.\left.\left. +\frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}_j^{(t)}}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)h(C,\mathbf{x}_i,y_i)(1-\mathrm{logit}_j(\mathbf{W}^{(t)},\mathbf{x}_i))\boldsymbol{\xi}_i + \eta\cdot\mathbf{n}_{j,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right] \\
&- \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle\right)\right] \\
\le& \frac{1}{m}\sum_{r=1}^{m}\left[\left|\left\langle \frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}^{(t)}}\boldsymbol{\xi}_i, \boldsymbol{\xi}\right\rangle\right| + \left|\left\langle \eta\mathbf{n}_{j,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right|\right] \\
\le& \mathcal{O}\left(\frac{\eta}{m\sqrt{B}}\,\mathrm{Tr}(\mathbf{A}_y^\top\mathbf{A}_y) + \eta\sqrt{d}\sigma_n\|\mathbf{A}_y\|_F\right)
\end{aligned}
$$

where the first inequality is because $\sigma'(\cdot) \geq 0, \text{logit}_j \in [0,1]$ and ReLU function is 1-Lipschitz continuous; the second inequality is because of Cauchy–Schwarz inequality, the property of 1-norm and 2-norm, and Lemma C.2 and Lemma C.10.

$$C = \frac{1}{m}\sum_{r=1}^{m}\sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)} + \frac{\eta}{mB}\cdot\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}_y^{(t)}}\sigma'\left(\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_y\rangle\right)\cdot h(C,\mathbf{x}_i,y_i)\cdot\left(1-\text{logit}_y\left(\mathbf{W}^{(t)},\mathbf{x}_i\right)\right)\cdot\mathbf{u}_y \right.$$
$$\left. +\eta\cdot\mathbf{n}_{y,r}^{(t)},\mathbf{u}_y\right\rangle\right) - \frac{1}{m}\sum_{r=1}^{m}\sigma\left(\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_y\rangle\right) \tag{58}$$

Based on Lemma C.6, we can conclude that with probability at least $1-\exp(-2m)$, the number of activated neurons at iteration t are at least $\frac{m}{4}$. Then, with probability at least $1-\exp(-\tilde{\Omega}(d))$, we have

$$C \geq \frac{1}{m}\sum_{r=1}^{m}\sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)} + \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}_y^{(t)}}\sigma'\left(\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_y\rangle\right)\cdot h(C,\mathbf{x}_i,y_i)\cdot\left(1-\text{logit}_y\left(\mathbf{W}^{(t)},\mathbf{x}_i\right)\right)\right.\right.$$
$$\left.\left.\mathbf{u}_y,\mathbf{u}_y\right\rangle\right) - \frac{1}{m}\sum_{r=1}^{m}\left|\langle\eta\cdot\mathbf{n}_{y,r}^{(t)},\mathbf{u}_y\rangle\right| - \frac{1}{m}\sum_{r=1}^{m}\sigma\left(\langle \mathbf{w}_{y,r}^{(t)},\mathbf{u}_y\rangle\right)$$
$$\geq \Omega\left(\frac{\eta C}{B\sqrt{m}(\|\mathbf{u}_y\|_2+\sqrt{\text{Tr}(\mathbf{A}_y^\top\mathbf{A}_y)})}\right)\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}_y^{(t)}}\left(1-\text{logit}_y\left(\mathbf{W}^{(t)},\mathbf{x}_i\right)\right)\|\mathbf{u}_y\|_2^2 - \mathcal{O}\left(\eta\sigma_n\sqrt{d}\|\mathbf{u}_y\|_2\right) \tag{59}$$

The second inequality is by using the bound of the clipping multiplier and Lemma C.2.

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

Bound of D

$$
\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}_i,y_i)\in S^{(t)}\setminus S_y^{(t)}} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C,\mathbf{x}_i,y_i)\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}_i)\boldsymbol{\xi}_i \\
+ \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in S_y^{(t)}} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C,\mathbf{x}_i,y_i)(1-\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}_i))\boldsymbol{\xi}_i + \eta\cdot\mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi}\Bigg\rangle\Bigg)\Bigg] \\
- \frac{1}{m}\sum_{r=1}^{m}\Big[\sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle)\Big] \\
\geq \frac{1}{m}\sum_{r=1}^{m}\Bigg[\sigma\Bigg(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle - \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in S^{(t)}\setminus S_y^{(t)}} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C,\mathbf{x}_i,y_i)\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}_i)|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| \\
+ \frac{\eta}{mB}\sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle) h(C,\mathbf{x},y)(1-\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}))\|\boldsymbol{\xi}\|_2^2 \\
- \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in S_y^{(t)}\setminus(\mathbf{x},y)} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle)(1-h(C,\mathbf{x}_i,y_i)\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}_i))|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle|\Bigg)\Bigg] \\
- \frac{1}{m}\sum_{r=1}^{m}|\langle \eta\cdot\mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle| - \frac{1}{m}\sum_{r=1}^{m}\Big[\sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle)\Big] \\
\geq -\frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in S^{(t)}\setminus S_y^{(t)}} h(C,\mathbf{x}_i,y_i)\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}_i)|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| + \frac{2\eta}{5mB}h(C,\mathbf{x},y)(1-\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}))\|\boldsymbol{\xi}\|_2^2 \\
- \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in S_y^{(t)}\setminus(\mathbf{x},y)} h(C,\mathbf{x}_i,y_i)(1-\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}_i))|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| - \frac{1}{m}\sum_{r=1}^{m}|\langle \eta\cdot\mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle| \\
\geq \Omega\left(\frac{\eta C}{n\sqrt{m}(\|\mathbf{u}_y\|_2 + \sqrt{\mathrm{Tr}(\mathbf{A}_y^\top\mathbf{A}_y)})}\right)(1-\mathrm{logit}_y(\mathbf{W}^{(t)},\mathbf{x}))\|\boldsymbol{\xi}\|_2^2 - \mathcal{O}\left(\eta\sigma_n\sqrt{d}\|\mathbf{A}_y\|_F\right),
\end{aligned}
\tag{60}
$$

where the last inequality is by using the bound of the clipping multiplier and by the fact that the incremental term is larger than zero if a neuron is activated and by Lemma C.2.

Substituting bounds of A, B, C, D to (57), we obtain the upper bound of $\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$. With probability at least $1 - \exp(-\tilde{\Omega}(d))$,

$$
\begin{aligned}
&\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) \leq \\
&\quad \mathcal{O}(\eta\sigma_n\sqrt{d}\|\mathbf{u}_y\|_2) + \mathcal{O}\left(\frac{\eta}{m\sqrt{B}}\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) + \eta\sqrt{d}\sigma_n\|\mathbf{A}_y\|_F\right) \\
&\quad - \Omega\left(\frac{\eta C}{B\sqrt{m}(\|\mathbf{u}_y\|_2 + \sqrt{\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)})}\right) \sum_{(\mathbf{x}_i, y_i)\in\mathcal{S}_y^{(t)}} \left(1 - \operatorname{logit}_y\left(\mathbf{W}^{(t)}, \mathbf{x}_i\right)\right)\|\mathbf{u}_y\|_2^2 + \mathcal{O}\left(\eta\sigma_n\sqrt{d}\|\mathbf{u}_y\|_2\right) \\
&\quad - \Omega\left(\frac{\eta C}{B\sqrt{m}(\|\mathbf{u}_y\|_2 + \sqrt{\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)})}\right) (1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}))\|\boldsymbol{\xi}\|_2^2 + \mathcal{O}\left(\eta\sigma_n\sqrt{d}\|\mathbf{A}_y\|_F\right) \\
&\leq - \Omega\left(\frac{\eta\Lambda_y}{B\sqrt{m}}\right) \sum_{(\mathbf{x}_i, y_i)\in\mathcal{S}_y^{(t)}} \left(1 - \operatorname{logit}_y\left(\mathbf{W}^{(t)}, \mathbf{x}_i\right)\right)\|\mathbf{u}_y\|_2^2 + \mathcal{O}\left(\eta\sigma_n\sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F\right)
\end{aligned} \tag{61}
$$

Combining (56) and (61), for any $(\mathbf{x}, y) \in \mathcal{S}$ we have

$$
\begin{aligned}
\mathcal{L}(\mathbf{W}^{(t+1)}, \mathbf{x}, y) &\leq \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) - \frac{\eta\Lambda_y}{\sqrt{m}B} \sum_{(\mathbf{x}_i, y_i)\in\mathcal{S}_y^{(t)}} \left(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i)\right)\|\mathbf{u}_y\|_2^2 + \mathcal{O}\left(\eta\sigma_n\sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F\right) \\
&\overset{(a)}{=} \left(1 - \Omega\left(\frac{\eta\Lambda_y}{n\sqrt{m}}|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2\right)\right)\mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) + \mathcal{O}\left(\eta\sigma_n\sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F\right),
\end{aligned} \tag{62}
$$

This finishes the training loss analysis.

D.4 Proof of Theorem 4.3

By (5) and (53), for any $(\mathbf{x}, y) \in S$, we have:

$$
\begin{aligned}
\langle \mathbf{w}_{y,r}^{(t+1)}, \boldsymbol{\xi}\rangle - \langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle &= \Bigg\langle -\frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i)\in S^{(t)}\setminus S_y^{(t)}} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C, \mathbf{x}_i, y_i) \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i)\boldsymbol{\xi}_i \\
&\quad + \frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i)\in S_y^{(t)}} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C, \mathbf{x}_i, y_i)(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i))\boldsymbol{\xi}_i + \eta\cdot\mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi}\Bigg\rangle \\
&\geq -\frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i)\in S^{(t)}\setminus S_y^{(t)}} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C, \mathbf{x}_i, y_i) \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i)|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| \\
&\quad + \frac{\eta}{mB}\sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle) h(C, \mathbf{x}, y)(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}))\|\boldsymbol{\xi}\|_2^2 \\
&\quad - \frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i)\in S_y^{(t)}\setminus(\mathbf{x}, y)} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) h(C, \mathbf{x}_i, y_i)(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i))|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| - \frac{1}{m}\sum_{r=1}^m |\langle \eta\cdot\mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle| \\
&\geq -\frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i)\in S^{(t)}\setminus S_y^{(t)}} h(C, \mathbf{x}_i, y_i) \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i)|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| + \frac{2\eta}{5mB} h(C, \mathbf{x}, y)(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}))\|\boldsymbol{\xi}\|_2^2 \\
&\quad - \frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i)\in S_y^{(t)}\setminus(\mathbf{x}, y)} h(C, \mathbf{x}_i, y_i)(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i))|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| - \frac{1}{m}\sum_{r=1}^m |\langle \eta\cdot\mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle| \\
&\geq \Omega\left(\frac{\eta C}{n\sqrt{m}(\|\mathbf{u}_y\|_2 + \sqrt{\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)})}\right)(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}))\|\boldsymbol{\xi}\|_2^2 - \mathcal{O}\left(\eta\sigma_n\sqrt{d}\|\mathbf{A}_y\|_F\right), \\
&= \Omega\left(\frac{\eta\Lambda_y}{n\sqrt{m}}\left(1 - \operatorname{logit}_y(\mathbf{W}^{(t)}, \mathbf{x})\right)\|\boldsymbol{\xi}\|_2^2\right) - \mathcal{O}\left(\eta\sigma_n\sqrt{d}\|\mathbf{A}_y\|_F\right)
\end{aligned}
$$

where the last inequality is by Lemma C.2 and the bound of the clipping multiplier .

Then, by taking the summation, for any $(\mathbf{x}, y) \in S$, we have:

$$\begin{aligned}\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi} \rangle &\geq \Omega\left(\sum_{t=0}^{T-1} \frac{\eta\Lambda_y}{n\sqrt{m}} \left(1 - \mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x})\right) \|\boldsymbol{\xi}\|_2^2\right) - \mathcal{O}\left(\eta T \sigma_n \sqrt{d} \|\mathbf{A}_\mathbf{y}\|_F\right) + \langle \mathbf{w}_{y,r}^{(0)}, \boldsymbol{\xi} \rangle \\ &\geq \Omega\left(\sum_{t=0}^{T-1} \frac{\eta\Lambda_y}{n\sqrt{m}} \left(1 - \mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x})\right) \|\boldsymbol{\xi}\|_2^2\right) - \mathcal{O}\left(\eta T \sigma_n \sqrt{d} \|\mathbf{A}_\mathbf{y}\|_F + \log\left(\frac{Km}{\delta}\right) \|\mathbf{A}_y\|_F \sigma_0\right)\end{aligned}$$

where the last inequality is by Lemma C.11. When $T \geq \Omega\left((\eta\Lambda_y \|\mathbf{A}_y\|_F)^{-1} n\sqrt{m}\sigma_0\right)$, the last term related to initialization can be ignored.

D.5 Proof of Theorem 4.5

Proof. As (62), we have

$$\mathcal{L}(\mathbf{W}^{(t+1)}, \mathbf{x}, y) \leq \left(1 - \Theta\left(\frac{\eta\Lambda_y}{n\sqrt{m}} |\mathcal{S}_y| \|\mathbf{u}_y\|_2^2\right)\right) \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) + \mathcal{O}\left(\eta\sigma_n\sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F\right), \tag{63}$$

Combining all T iterations, we have

$$\begin{aligned}\mathcal{L}(\mathbf{W}^{(T)}, \mathbf{x}, y) \leq& \left(1 - \Omega\left(\frac{\eta\Lambda_y}{n\sqrt{m}} |\mathcal{S}_y| \|\mathbf{u}_y\|_2^2\right)\right)^T \mathcal{L}(\mathbf{W}^{(0)}, \mathbf{x}, y) \\ &+ \mathcal{O}\left(\eta\sigma_n\sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F\right) \cdot \mathcal{O}\left(\frac{n\sqrt{m}}{\eta\Lambda_y |\mathcal{S}_y| \|\mathbf{u}_y\|_2^2}\right) \\ \leq& \underbrace{\exp\left(-\Omega\left(\frac{\eta T \Lambda_y |\mathcal{S}_y|}{n\sqrt{m}} \|\mathbf{u}_y\|_2^2\right)\right) \mathcal{L}(\mathbf{W}^{(0)}, \mathbf{x}, y)}_{\text{Vanishing error}} + \underbrace{\mathcal{O}\left(\frac{n\sqrt{m} \cdot \sigma_n\sqrt{d}(\|\mathbf{u}_y\|_2 + \|\mathbf{A}_y\|_F)}{\Lambda_y |\mathcal{S}_y| \|\mathbf{u}_y\|_2^2}\right)}_{\text{Privacy protection error}}\end{aligned} \tag{64}$$

where the first inequality is obtained from the property of Geometric sequences. □

D.6 Lower bound of the training loss

Theorem D.2 (Lower bound on the training loss)**.** *Under Condition 4.1 and Assumption 4.2, with the number of iterations* $T \geq \Omega\left(-\frac{1}{\log\left(1-\Omega\left(\frac{\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{mn}\right)\right)}\right)$*, with probability at least* $1 - \delta$*, for any* $(\mathbf{x}, y) \in \mathcal{S}$*, the expected training loss satisfies*

$$\begin{aligned}\mathbb{E}\Big[\mathcal{L}(\mathbf{W}^{(T)}, \mathbf{x}, y)\Big] \geq& \left(1 - O\left(\frac{\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{mn}\right)\right)^T \mathcal{L}(\mathbf{W}^{(0)}, \mathbf{x}, y) \\ &+ \widetilde{\Omega}\left(\frac{\eta n\sigma_n^2 \left(\|\mathbf{u}_y\|_2^2 + \mathrm{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)\right)}{|\mathcal{S}_y| \|\mathbf{u}_y\|_2^2}\right) \\ &- \widetilde{O}\left(\frac{\mathrm{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) + n \max_{k\in[K]} \|\mathbf{A}_k^\top \mathbf{A}_y\|_F}{|\mathcal{S}_y| \|\mathbf{u}_y\|_2^2}\right),\end{aligned}$$

where the expectation is taken over the mini-batch sampling and the Gaussian noise injected by DP-SGD.

Remark D.3. The first term in Theorem D.2 is the vanishing error, which decays to zero as the number of iterations T increases. The second term is the privacy protection error, which provides a positive lower bound on the expected training loss under DP-SGD and increases with the privacy noise level

σ_n. Together with the upper bound in Theorem 4.5, this result shows that privacy noise prevents the training loss from converging to an arbitrarily small value. Since the proof lower bounds the loss by upper bounding the loss reduction caused by gradient updates, the clipping multiplier is relaxed using $h(C, \mathbf{x}, y) \leq 1$. Consequently, Λ_y does not explicitly appear in the lower bound, and this relaxation only makes the resulting lower bound more conservative. The last term is a memorization term arising from the model's memorization of the training noise patches. It depends on both the magnitude of the target noise pattern, characterized by $\mathrm{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)$, and the correlations between noise patterns from different classes, characterized by $\max_{k\in[K]} \|\mathbf{A}_k^\top \mathbf{A}_y\|_F$. Thus, the lower bound shows that DP prevents the training loss from converging to zero when the privacy protection error dominates the memorization term.

Proof. Fix $(\mathbf{x}, y) \in \mathcal{S}$. All expectations are taken over the mini-batch sampling and the Gaussian noise injected by DP-SGD, conditional on the training dataset and initialization. Recall that

$$\begin{aligned}
&\mathcal{L}(\mathbf{W}^{(t+1)}, \mathbf{x}, y) - \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) \\
&= \log\left(1 + \sum_{j\neq y} \mathrm{logit}_j(\mathbf{W}^{(t)}, \mathbf{x}) \left[\exp\left(\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})\right) - 1\right]\right),
\end{aligned} \tag{65}$$

where $\Delta_j^{(t)}(\mathbf{x}) = F_j(\mathbf{W}^{(t+1)}, \mathbf{x}) - F_j(\mathbf{W}^{(t)}, \mathbf{x})$. By Assumption 4.2, $1 - \mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}) \geq 1 - \exp(-s) > 0$. Therefore, a uniform lower bound on $\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x})$ for all $j \neq y$ yields a lower bound on the loss increment in equation 65.

For any $j \neq y$, we decompose

$$\begin{aligned}
\Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) = &\underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{u}_y\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle)\right]}_{A} \\
&+ \underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle)\right]}_{B} \\
&- \underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma(\langle \mathbf{w}_{y,r}^{(t+1)}, \mathbf{u}_y\rangle) - \sigma(\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_y\rangle)\right]}_{C} \\
&- \underbrace{\frac{1}{m}\sum_{r=1}^{m}\left[\sigma(\langle \mathbf{w}_{y,r}^{(t+1)}, \boldsymbol{\xi}\rangle) - \sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle)\right]}_{D}.
\end{aligned} \tag{66}$$

We next bound the four terms separately. By the orthogonality of the signal vectors and $\mathbf{u}_k^\top \mathbf{A}_y = 0$, only samples in $\mathcal{S}_y^{(t)}$ contribute to the gradient projection onto $\mathbf{u}_y$. Using the monotonicity and 1-Lipschitz continuity of ReLU, $0 \leq h(C, \mathbf{x}_i, y_i) \leq 1$, and $0 \leq \mathrm{logit}_j \leq 1$, we obtain

$$\begin{aligned}
A \geq &-\frac{\eta}{mB}\sum_{(\mathbf{x}_i, y_i)\in\mathcal{S}_y^{(t)}} h(C, \mathbf{x}_i, y_i)\,\mathrm{logit}_j(\mathbf{W}^{(t)}, \mathbf{x}_i)\|\mathbf{u}_y\|_2^2 \\
&+ \frac{\eta}{m}\sum_{r=1}^{m}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle)\langle \mathbf{n}_{j,r}^{(t)}, \mathbf{u}_y\rangle \\
\geq &-\frac{\eta|\mathcal{S}_y^{(t)}|}{mB}\|\mathbf{u}_y\|_2^2 + \frac{\eta}{m}\sum_{r=1}^{m}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y\rangle)\langle \mathbf{n}_{j,r}^{(t)}, \mathbf{u}_y\rangle.
\end{aligned} \tag{67}$$

Similarly, the gradient projection onto $\boldsymbol{\xi}$ gives

$$\begin{aligned} B \geq & -\frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{(t)}} h(C, \mathbf{x}_i, y_i) \left| \mathbb{I}(y_i = j) - \mathrm{logit}_j(\mathbf{W}^{(t)}, \mathbf{x}_i) \right| |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \\ & + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle) \langle \mathbf{n}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle \\ \geq & -\frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{(t)}} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle) \langle \mathbf{n}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle. \end{aligned} \tag{68}$$

For the correct class, the signal contribution satisfies

$$\begin{aligned} C \leq & \frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_y^{(t)}} h(C, \mathbf{x}_i, y_i) \left(1 - \mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i)\right) \|\mathbf{u}_y\|_2^2 \\ & + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_y \rangle) \langle \mathbf{n}_{y,r}^{(t)}, \mathbf{u}_y \rangle \\ \leq & \frac{\eta |\mathcal{S}_y^{(t)}|}{mB} \|\mathbf{u}_y\|_2^2 + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_y \rangle) \langle \mathbf{n}_{y,r}^{(t)}, \mathbf{u}_y \rangle. \end{aligned} \tag{69}$$

Finally, the correct-class noise contribution satisfies

$$\begin{aligned} D \leq & \frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{(t)}} h(C, \mathbf{x}_i, y_i) \left| \mathbb{I}(y_i = y) - \mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}_i) \right| |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \\ & + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi} \rangle) \langle \mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi} \rangle \\ \leq & \frac{\eta}{mB} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{(t)}} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi} \rangle) \langle \mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi} \rangle. \end{aligned} \tag{70}$$

In particular, when $(\mathbf{x}_i, y_i) = (\mathbf{x}, y)$, the corresponding inner product equals $\|\boldsymbol{\xi}\|_2^2$, which captures the decrease in training loss caused by memorizing the target noise patch.

Combining equation 67–equation 70, we obtain

$$\begin{aligned} \Delta_j^{(t)}(\mathbf{x}) - \Delta_y^{(t)}(\mathbf{x}) \geq & -\frac{2\eta}{mB} \left(|\mathcal{S}_y^{(t)}| \|\mathbf{u}_y\|_2^2 + \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{(t)}} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \right) \\ & + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_y \rangle) \langle \mathbf{n}_{j,r}^{(t)}, \mathbf{u}_y \rangle + \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle) \langle \mathbf{n}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle \\ & - \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \mathbf{u}_y \rangle) \langle \mathbf{n}_{y,r}^{(t)}, \mathbf{u}_y \rangle - \frac{\eta}{m} \sum_{r=1}^{m} \sigma'(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi} \rangle) \langle \mathbf{n}_{y,r}^{(t)}, \boldsymbol{\xi} \rangle. \end{aligned} \tag{71}$$

Conditional on $\mathbf{W}^{(t)}$ and $\mathcal{S}^{(t)}$, the last four terms in equation 71 form a centered Gaussian random variable. By Lemma C.9, a constant fraction of the neurons are activated on both $\mathbf{u}_y$ and $\boldsymbol{\xi}$. Moreover, $\mathbf{u}_y^\top \boldsymbol{\xi} = 0$, and the Gaussian perturbations corresponding to different classes and neurons are

independent. Hence,

$$\begin{aligned}\mathrm{Var}\Bigg[\frac{\eta}{m}\sum_{r=1}^{m}\Big(&\sigma'(\langle\mathbf{w}_{j,r}^{(t)},\mathbf{u}_y\rangle)\langle\mathbf{n}_{j,r}^{(t)},\mathbf{u}_y\rangle+\sigma'(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\rangle)\langle\mathbf{n}_{j,r}^{(t)},\boldsymbol{\xi}\rangle\\&-\sigma'(\langle\mathbf{w}_{y,r}^{(t)},\mathbf{u}_y\rangle)\langle\mathbf{n}_{y,r}^{(t)},\mathbf{u}_y\rangle-\sigma'(\langle\mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}\rangle)\langle\mathbf{n}_{y,r}^{(t)},\boldsymbol{\xi}\rangle\Big)\,\Big|\,\mathbf{W}^{(t)},\mathcal{S}^{(t)}\Bigg]\\&=\Omega\left(\frac{\eta^2\sigma_n^2}{m}\left(\|\mathbf{u}_y\|_2^2+\|\boldsymbol{\xi}\|_2^2\right)\right).\end{aligned}\tag{72}$$

Using the moment-generating function of a centered Gaussian variable, we have

$$\mathbb{E}[\exp(Z)\mid\mathbf{W}^{(t)},\mathcal{S}^{(t)}]=\exp\left(\frac{\mathrm{Var}(Z\mid\mathbf{W}^{(t)},\mathcal{S}^{(t)})}{2}\right).\tag{73}$$

Applying equation 73, $\exp(z)-1\geq z$, and equation 71 yields

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}(\mathbf{W}^{(t+1)},\mathbf{x},y)\,\middle|\,\mathbf{W}^{(t)},\mathcal{S}^{(t)}\right]\\&\geq\mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y)-\frac{2\eta}{mB}\left(|\mathcal{S}_y^{(t)}|\|\mathbf{u}_y\|_2^2+\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}^{(t)}}|\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}\rangle|\right)\\&\quad+\Omega\left(\frac{\eta^2\sigma_n^2}{m}\left(\|\mathbf{u}_y\|_2^2+\|\boldsymbol{\xi}\|_2^2\right)\right).\end{aligned}\tag{74}$$

Since every sample is included in a uniformly sampled mini-batch with probability B/n, taking expectation over $\mathcal{S}^{(t)}$ gives

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\mathbf{W}^{(t+1)},\mathbf{x},y)\mid\mathbf{W}^{(t)}\right]\geq\mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y)&-\frac{2\eta|\mathcal{S}_y|}{mn}\|\mathbf{u}_y\|_2^2\\&-\frac{2\eta}{mn}\sum_{i=1}^{n}|\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}\rangle|+\Omega\left(\frac{\eta^2\sigma_n^2}{m}\left(\|\mathbf{u}_y\|_2^2+\|\boldsymbol{\xi}\|_2^2\right)\right).\end{aligned}\tag{75}$$

By Assumption 4.2,

$$-\frac{2\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{mn}\geq-\frac{2\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{smn}\mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y).\tag{76}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\mathbf{W}^{(t+1)},\mathbf{x},y)\right]\geq&\left(1-\frac{2\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{smn}\right)\mathbb{E}\left[\mathcal{L}(\mathbf{W}^{(t)},\mathbf{x},y)\right]\\&+\Omega\left(\frac{\eta^2\sigma_n^2}{m}\left(\|\mathbf{u}_y\|_2^2+\|\boldsymbol{\xi}\|_2^2\right)\right)-\frac{2\eta}{mn}\sum_{i=1}^{n}|\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}\rangle|.\end{aligned}\tag{77}$$

Combining all T iterations yields

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\mathbf{W}^{(T)},\mathbf{x},y)\right]\geq&\left(1-\frac{2\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{smn}\right)^T\mathcal{L}(\mathbf{W}^{(0)},\mathbf{x},y)\\&+\left[1-\left(1-\frac{2\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{smn}\right)^T\right]\\&\cdot\left[\Omega\left(\frac{s\eta n\sigma_n^2\left(\|\mathbf{u}_y\|_2^2+\|\boldsymbol{\xi}\|_2^2\right)}{|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}\right)-\frac{s\sum_{i=1}^{n}|\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}\rangle|}{|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}\right].\end{aligned}\tag{78}$$

For

$$T \geq \Omega\left(-\frac{1}{\log\left(1-\Omega\left(\frac{\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{mn}\right)\right)}\right), \tag{79}$$

the geometric accumulation factor in equation 78 is bounded below by a positive constant. On the event of Lemma C.10,

$$\frac{1}{2}\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) \leq \|\boldsymbol{\xi}\|_2^2 \leq \frac{3}{2}\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y). \tag{80}$$

Moreover,

$$\sum_{i=1}^{n} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| \leq \widetilde{O}\left(\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) + n \max_{k\in[K]} \|\mathbf{A}_k^\top \mathbf{A}_y\|_F\right). \tag{81}$$

Substituting equation 80 and equation 81 into equation 78, and absorbing the constant s into the asymptotic notation, gives

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\mathbf{W}^{(T)}, \mathbf{x}, y)\right] \geq\ & \left(1 - O\left(\frac{\eta|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}{mn}\right)\right)^T \mathcal{L}(\mathbf{W}^{(0)}, \mathbf{x}, y) \\ & + \widetilde{\Omega}\left(\frac{\eta n \sigma_n^2 \left(\|\mathbf{u}_y\|_2^2 + \operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)\right)}{|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}\right) \\ & - \widetilde{O}\left(\frac{\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) + n\max_{k\in[K]} \|\mathbf{A}_k^\top \mathbf{A}_y\|_F}{|\mathcal{S}_y|\|\mathbf{u}_y\|_2^2}\right).\end{aligned} \tag{82}$$

A union bound over the data-concentration and activation events completes the proof. □

E Test Error Analysis

In this section, we analyze the test error. Similar to the proof in Appendix D, the results are based on the high probability conclusions in Appendix C. In order to characterize the test loss, we first prove the following key lemmas.

E.1 Key Lemmas

Lemma E.1. *Define*

$$S_i^{(t)} = \left\{r \in [m] : \left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle > 0\right\}, \tag{83}$$

for all $(\mathbf{x}_i, y_i) \in S$. Under Condition 4.1 and Assumption 4.2, for any $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$ and $t \in [T]$, we have

$$\frac{1 - \operatorname{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i)}{1 - \operatorname{logit}_{y_j}(\mathbf{W}^{(t)}, \mathbf{x}_j)} \leq \kappa, \tag{84}$$

for a constant $\kappa > 1$, and

$$S_i^{(t)} \subseteq S_i^{(t+1)}. \tag{85}$$

Proof. We first prove the first statement. By Assumption 4.2, for every $(\mathbf{x}_i, y_i) \in S$ and $t \in [T]$, we have

$$-\log \operatorname{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i) \geq s. \tag{86}$$

Therefore,

$$\operatorname{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i) \leq \exp(-s), \tag{87}$$

which implies

$$1 - \exp(-s) \leq 1 - \operatorname{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i) \leq 1. \tag{88}$$

Hence, for any $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in S$,

$$\frac{1 - \operatorname{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i)}{1 - \operatorname{logit}_{y_j}(\mathbf{W}^{(t)}, \mathbf{x}_j)} \leq \frac{1}{1 - \exp(-s)} =: \kappa. \tag{89}$$

We next prove the second statement. For any $(\mathbf{x}_i, y_i) \in S$ and $r \in S_i^{(t)}$, by the DP-SGD update rule, we have

$$\begin{aligned}
&\left\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \right\rangle - \left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \\
&= \frac{\eta}{mn} h(C, \mathbf{x}_i, y_i) \left(1 - \mathrm{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i)\right) \sigma'\left(\left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle\right) \|\boldsymbol{\xi}_i\|_2^2 \\
&\quad - \frac{\eta}{mn} \sum_{i' \neq i} h(C, \mathbf{x}_{i'}, y_{i'}) \,\mathrm{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_{i'}) \sigma'\left(\left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i'} \right\rangle\right) \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&\quad + \eta \left\langle \mathbf{n}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle .
\end{aligned} \tag{90}$$

Since $r \in S_i^{(t)}$, we have

$$\sigma'\left(\left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle\right) = 1. \tag{91}$$

Moreover, by Assumption 4.2 and the lower bound of the clipping factor in Condition 4.1,

$$\begin{aligned}
&h(C, \mathbf{x}_i, y_i) \left(1 - \mathrm{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i)\right) \|\boldsymbol{\xi}_i\|_2^2 \\
&\geq \Lambda_{y_i}(1 - \exp(-s)) \|\boldsymbol{\xi}_i\|_2^2.
\end{aligned} \tag{92}$$

By the noise geometry condition in Condition 4.1, the cross-sample term satisfies

$$\begin{aligned}
&\left| \sum_{i' \neq i} h(C, \mathbf{x}_{i'}, y_{i'}) \,\mathrm{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_{i'}) \sigma'\left(\left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i'} \right\rangle\right) \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \right| \\
&\leq \frac{1}{4} \Lambda_{y_i}(1 - \exp(-s)) \|\boldsymbol{\xi}_i\|_2^2.
\end{aligned} \tag{93}$$

In addition, by Lemma C.5 with the failure probability set to $\delta/(nmT)$ and a union bound over all $i \in [n]$, $r \in [m]$, and $t \in [T]$, with probability at least $1 - \delta$, we have

$$\begin{aligned}
\left| \left\langle \mathbf{n}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \right| \leq C\sigma_n\sigma_p \Bigg(&\|A_{y_i}\|_F \sqrt{\log\left(\frac{2nmT}{\delta}\right)} \\
&+ \|A_{y_i}\|_2 \log\left(\frac{2nmT}{\delta}\right) \Bigg).
\end{aligned} \tag{94}$$

By the DP-noise condition in Condition 4.1,

$$\sigma_n \sqrt{2\log\left(\frac{2KmnT}{\delta}\right)} \leq \frac{\Lambda_{y_i}(1 - \exp(-s))}{4mn} \|\boldsymbol{\xi}_i\|_2, \tag{95}$$

and hence

$$\left| \left\langle \mathbf{n}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \right| \leq \frac{\Lambda_{y_i}(1 - \exp(-s))}{4mn} \|\boldsymbol{\xi}_i\|_2^2. \tag{96}$$

Combining equation 90, equation 92, equation 93, and equation 96, we obtain

$$\begin{aligned}
&\left\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \right\rangle - \left\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \\
&\geq \frac{\eta}{2mn} \Lambda_{y_i}(1 - \exp(-s)) \|\boldsymbol{\xi}_i\|_2^2 \geq 0.
\end{aligned} \tag{97}$$

Thus,

$$\left\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \right\rangle > 0 \tag{98}$$

for every $r \in S_i^{(t)}$, which implies

$$S_i^{(t)} \subseteq S_i^{(t+1)}. \tag{99}$$

This completes the proof. □

Lemma E.2. *Under Condition 4.1, for any $j \in [K]$, $l \in [K] \setminus \{j\}$, $(\mathbf{x}_q, y_q) \in S_j$, $(\mathbf{x}_a, y_a) \in S_l$, and $r \in S_q^{(0)}$, we have*

$$\begin{aligned}
&\sum_{t'=0}^{t-1} h(C, \mathbf{x}_q, y_q) \left(1 - \mathrm{logit}_{y_q}(\mathbf{W}^{(t')}, \mathbf{x}_q)\right) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_q \right\rangle\right) \\
&\geq \Omega\left(\frac{1}{n} \sum_{t'=0}^{t-1} h(C, \mathbf{x}_a, y_a) \,\mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x}_a) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right) \frac{\mathrm{Tr}(\mathbf{A}_l^\top \mathbf{A}_l)}{\max_{l_1, l_2 \in [K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}}\right) \\
&\quad - O\Bigg(\frac{m\sigma_n \sigma_p \sqrt{t}}{\max_{l_1, l_2 \in [K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}} \Bigg[\|\mathbf{A}_l\|_F \sqrt{\log\left(\frac{2Kmn T}{\delta}\right)} \\
&\qquad\qquad + \|\mathbf{A}_l\|_2 \log\left(\frac{2Kmn T}{\delta}\right)\Bigg]\Bigg). \qquad (100)
\end{aligned}$$

Proof. By the update rule of DP-SGD, for $(\mathbf{x}_a, y_a) \in S_l$ with $l \neq j$, we have

$$\begin{aligned}
&\left\langle \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_a \right\rangle \\
&\leq \frac{\eta}{mn} \sum_{(\mathbf{x}_i, y_i) \in S \setminus \{(\mathbf{x}_a, y_a)\}} h(C, \mathbf{x}_i, y_i) \left(1 - \mathrm{logit}_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i)\right) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle\right) |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_a \rangle| \\
&\quad - \frac{\eta}{mn} h(C, \mathbf{x}_a, y_a) \,\mathrm{logit}_j(\mathbf{W}^{(t)}, \mathbf{x}_a) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_a \right\rangle\right) \|\boldsymbol{\xi}_a\|_2^2 + \eta \left\langle \mathbf{n}_{j,r}^{(t)}, \boldsymbol{\xi}_a \right\rangle. \qquad (101)
\end{aligned}$$

By Lemma C.5, with probability at least $1 - \delta$, for all distinct training samples,

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_a \rangle| \leq O\left(\max_{l_1, l_2 \in [K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}\right). \qquad (102)$$

Therefore, summing equation 101 from $t' = 0$ to t, we obtain

$$\begin{aligned}
&\left\langle \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_a \right\rangle \\
&\quad + \frac{\eta}{mn} \sum_{t'=0}^{t} h(C, \mathbf{x}_a, y_a) \,\mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x}_a) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right) \|\boldsymbol{\xi}_a\|_2^2 \\
&\leq O\Bigg(\frac{\eta}{mn} \sum_{(\mathbf{x}, y) \in S \setminus \{(\mathbf{x}_a, y_a)\}} \sum_{t'=0}^{t} h(C, \mathbf{x}, y) \left(1 - \mathrm{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})\right) \\
&\qquad\qquad \cdot \max_{l_1, l_2 \in [K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}\Bigg) + \eta \left|\sum_{t'=0}^{t} \left\langle \mathbf{n}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right|. \qquad (103)
\end{aligned}$$

We next bound the accumulated Gaussian perturbation. Since $\mathbf{n}_{j,r}^{(t')} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ are independent across iterations,

$$\sum_{t'=0}^{t} \mathbf{n}_{j,r}^{(t')} \sim \mathcal{N}\left(\mathbf{0}, (t+1)\sigma_n^2 \mathbf{I}\right). \qquad (104)$$

Applying Lemma C.10 with σ_n replaced by $\sqrt{t+1}\sigma_n$, and taking a union bound over all $j \in [K]$, $r \in [m]$, training samples, and $t \in [T]$, with probability at least $1 - \delta$,

$$\begin{aligned}
\left|\sum_{t'=0}^{t} \left\langle \mathbf{n}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right| \leq O\Bigg(\sigma_n \sigma_p \sqrt{t+1} \Bigg[&\|\mathbf{A}_l\|_F \sqrt{\log\left(\frac{2Kmn T}{\delta}\right)} \\
&+ \|\mathbf{A}_l\|_2 \log\left(\frac{2Kmn T}{\delta}\right)\Bigg]\Bigg). \qquad (105)
\end{aligned}$$

On the other hand, by the ReLU activation and the update rule, we have

$$\begin{aligned}
&\left\langle \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_a \right\rangle \\
\geq\ & -O\Bigg(\frac{\eta}{mn} \sum_{(\mathbf{x},y)\in S\setminus\{(\mathbf{x}_a,y_a)\}} \sum_{t'=0}^{t} h(C,\mathbf{x},y) \left(1 - \mathrm{logit}_y(\mathbf{W}^{(t')},\mathbf{x})\right) \\
&\qquad \cdot \max_{l_1,l_2\in[K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)} \Bigg) \\
&- \frac{\eta}{mn} h(C,\mathbf{x}_a,y_a)\|\boldsymbol{\xi}_a\|_2^2 - \eta \left| \sum_{t'=0}^{t} \left\langle \mathbf{n}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle \right|. \qquad (106)
\end{aligned}$$

Combining equation 103, equation 105, and equation 106, and using the learning-rate condition in Condition 4.1, we obtain

$$\begin{aligned}
&\frac{\eta}{mn} \sum_{t'=0}^{t-1} h(C,\mathbf{x}_a,y_a)\,\mathrm{logit}_j(\mathbf{W}^{(t')},\mathbf{x}_a)\sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right) \|\boldsymbol{\xi}_a\|_2^2 \\
=\ & O\Bigg(\frac{\eta}{mn} \sum_{(\mathbf{x},y)\in S\setminus\{(\mathbf{x}_a,y_a)\}} \sum_{t'=0}^{t-1} h(C,\mathbf{x},y) \left(1 - \mathrm{logit}_y(\mathbf{W}^{(t')},\mathbf{x})\right) \\
&\qquad \cdot \max_{l_1,l_2\in[K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)} \Bigg) \\
&+ O\left(\eta\sigma_n\sigma_p\sqrt{t} \left[\|\mathbf{A}_l\|_F \sqrt{\log\left(\frac{2Kmn T}{\delta}\right)} + \|\mathbf{A}_l\|_2 \log\left(\frac{2Kmn T}{\delta}\right) \right] \right). \qquad (107)
\end{aligned}$$

By Lemma E.1, for $r \in S_q^{(0)}$, we have $r \in S_q^{(t')}$ for every $t' \leq t$. Moreover, the first statement of Lemma E.1 gives a uniform constant bound on the ratio of the residual probabilities. Therefore,

$$\begin{aligned}
&\sum_{t'=0}^{t-1} h(C,\mathbf{x}_q,y_q) \left(1 - \mathrm{logit}_{y_q}(\mathbf{W}^{(t')},\mathbf{x}_q)\right) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_q \right\rangle\right) \\
\geq\ & \Omega\Bigg(\frac{1}{n} \sum_{t'=0}^{t-1} h(C,\mathbf{x}_a,y_a)\,\mathrm{logit}_j(\mathbf{W}^{(t')},\mathbf{x}_a)\sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right) \frac{\|\boldsymbol{\xi}_a\|_2^2}{\max_{l_1,l_2\in[K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}} \Bigg) \\
&- O\Bigg(\frac{m\sigma_n\sigma_p\sqrt{t}}{\max_{l_1,l_2\in[K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}} \Bigg[\|\mathbf{A}_l\|_F \sqrt{\log\left(\frac{2Kmn T}{\delta}\right)} \\
&\qquad + \|\mathbf{A}_l\|_2 \log\left(\frac{2Kmn T}{\delta}\right) \Bigg] \Bigg). \qquad (108)
\end{aligned}$$

Finally, by Lemma C.10,

$$\|\boldsymbol{\xi}_a\|_2^2 \geq \frac{1}{2}\,\mathrm{Tr}(\mathbf{A}_l^\top \mathbf{A}_l). \qquad (109)$$

Substituting this bound into equation 108 yields

$$\begin{aligned}
&\sum_{t'=0}^{t-1} h(C, \mathbf{x}_q, y_q) \left(1 - \mathrm{logit}_{y_q}(\mathbf{W}^{(t')}, \mathbf{x}_q)\right) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_q \right\rangle\right) \\
&\geq \Omega\left(\frac{1}{n}\sum_{t'=0}^{t-1} h(C, \mathbf{x}_a, y_a) \,\mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x}_a) \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_a \right\rangle\right) \frac{\mathrm{Tr}(\mathbf{A}_l^\top \mathbf{A}_l)}{\max_{l_1,l_2\in[K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}}\right) \\
&\quad - O\left(\frac{m\sigma_n\sigma_p\sqrt{t}}{\max_{l_1,l_2\in[K]} \|\mathbf{A}_{l_1}^\top \mathbf{A}_{l_2}\|_F \sqrt{\log(3n^2/\delta)}} \left[\|\mathbf{A}_l\|_F \sqrt{\log\left(\frac{2Kmn T}{\delta}\right)} + \|\mathbf{A}_l\|_2 \log\left(\frac{2Kmn T}{\delta}\right)\right]\right).
\end{aligned} \tag{110}$$

This completes the proof. □

Lemma E.3. *Under Condition 4.1, for any $j \in [K]$, with probability at least $1-\delta$, we have*

$$\left\|(\mathbf{W}_j^{(T)})^\top \mathbf{A}_j\right\|_F \leq O\left(\sqrt{m}\log(T) + \eta\sigma_n\sqrt{mT}\left[\|\mathbf{A}_j\|_F + \|\mathbf{A}_j\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)}\right]\right). \tag{111}$$

Proof. Due to Assumption 4.2, for any $t \in [0, T]$ and $(\mathbf{x}, y) \in S$, we have

$$\mathrm{logit}_y(\mathbf{W}^{(t)}, \mathbf{x}) \leq \kappa, \tag{112}$$

where $\kappa < 1$ is a constant to be specified later.

We first bound $\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\rangle$ for a training sample $(\mathbf{x}, y) \in S$. By the definition of softmax and equation 112, we have

$$\begin{aligned}
&(1-\kappa)\exp\left(\sigma\left(\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\right\rangle\right)\right) \\
&\leq (1-\kappa)\exp\left(\sigma\left(\left\langle \mathbf{w}_{y,r}^{(T)}, \mathbf{u}_y\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\right\rangle\right)\right) \\
&\leq \kappa \sum_{k\neq y} \exp\left(\sigma\left(\left\langle \mathbf{w}_{k,r}^{(T)}, \mathbf{u}_y\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{k,r}^{(T)}, \boldsymbol{\xi}\right\rangle\right)\right).
\end{aligned} \tag{113}$$

For every $k \in [K]$ and $r \in [m]$, the DP-SGD iterate satisfies

$$\begin{aligned}
\mathbf{w}_{k,r}^{(T)} = \mathbf{w}_{k,r}^{(0)} &- \frac{\eta}{mn}\sum_{t'=0}^{T-1}\sum_{(\mathbf{x}_i,y_i)\in S} h(C, \mathbf{x}_i, y_i)\nabla_{\mathbf{w}_{k,r}} \ell(\mathbf{W}^{(t')}, \mathbf{x}_i, y_i) \\
&+ \eta\sum_{t'=0}^{T-1} \mathbf{n}_{k,r}^{(t')}.
\end{aligned} \tag{114}$$

Since $0 < h(C, \mathbf{x}_i, y_i) \leq 1$, clipping does not increase the magnitude of each gradient contribution. Therefore, following the same deterministic gradient argument as in the non-private case, together with Lemmas C.10 and C.11 and Condition 4.1, we have

$$\begin{aligned}
&\sigma\left(\left\langle \mathbf{w}_{k,r}^{(T)}, \mathbf{u}_y\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{k,r}^{(T)}, \boldsymbol{\xi}\right\rangle\right) \\
&\leq \sqrt{2\log\left(\frac{4Km}{\delta}\right)}\|\mathbf{u}_y\|_2\sigma_0 + O\left(\log\left(\frac{Km}{\delta}\right)\|\mathbf{A}_y\|_F\sigma_0\right) + \frac{\eta}{mn}\|\boldsymbol{\xi}\|_2^2 \\
&\quad + \eta\left|\left\langle \sum_{t'=0}^{T-1}\mathbf{n}_{k,r}^{(t')}, \mathbf{u}_y\right\rangle\right| + \eta\left|\left\langle \sum_{t'=0}^{T-1}\mathbf{n}_{k,r}^{(t')}, \boldsymbol{\xi}\right\rangle\right|.
\end{aligned} \tag{115}$$

We next control the two Gaussian perturbation terms. Since the perturbations are independent across iterations,

$$\sum_{t'=0}^{T-1} \mathbf{n}_{k,r}^{(t')} \sim \mathcal{N}\left(\mathbf{0}, T\sigma_n^2 \mathbf{I}\right). \tag{116}$$

Hence, by Lemma C.4 and a union bound over $k \in [K]$ and $r \in [m]$,

$$\left|\left\langle \sum_{t'=0}^{T-1} \mathbf{n}_{k,r}^{(t')}, \mathbf{u}_y \right\rangle\right| \le \sigma_n \sqrt{T} \|\mathbf{u}_y\|_2 \sqrt{2\log\left(\frac{2Km}{\delta}\right)}. \tag{117}$$

Moreover, applying Lemma C.5 with σ_n replaced by $\sigma_n\sqrt{T}$ gives

$$\left|\left\langle \sum_{t'=0}^{T-1} \mathbf{n}_{k,r}^{(t')}, \boldsymbol{\xi} \right\rangle\right| \le O\left(\sigma_n \sigma_p \sqrt{T}\left[\|\mathbf{A}_y\|_F \sqrt{\log\left(\frac{2Km}{\delta}\right)} + \|\mathbf{A}_y\|_2 \log\left(\frac{2Km}{\delta}\right)\right]\right). \tag{118}$$

Substituting equation 115–equation 118 into equation 113, and using Lemma C.10 and Condition 4.1, we obtain

$$\begin{aligned}
&(1-\kappa)\exp\left(\sigma\left(\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\right\rangle\right)\right) \\
&\le 1.1\kappa K \exp\Bigg(\frac{3\eta}{2mn}\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) \\
&\qquad + O\Bigg(\eta\sigma_n\sqrt{T}\Bigg[\|\mathbf{u}_y\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_F\sqrt{\log\left(\frac{2Km}{\delta}\right)} \\
&\qquad\qquad + \sigma_p\|\mathbf{A}_y\|_2\log\left(\frac{2Km}{\delta}\right)\Bigg]\Bigg)\Bigg).
\end{aligned} \tag{119}$$

Taking logarithms yields

$$\begin{aligned}
\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\right\rangle \le\ & \log\left(\frac{\kappa}{1-\kappa}K\right) + \frac{3\eta}{2mn}\operatorname{Tr}(\mathbf{A}_y^\top \mathbf{A}_y) \\
&+ O\Bigg(\eta\sigma_n\sqrt{T}\Bigg[\|\mathbf{u}_y\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_F\sqrt{\log\left(\frac{2Km}{\delta}\right)} \\
&\qquad + \sigma_p\|\mathbf{A}_y\|_2\log\left(\frac{2Km}{\delta}\right)\Bigg]\Bigg).
\end{aligned} \tag{120}$$

Taking

$$\kappa = \frac{T}{1.1K + T}, \tag{121}$$

and using Condition 4.1, we obtain

$$\begin{aligned}
\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\right\rangle \le O\Bigg(&\log(T) + \eta\sigma_n\sqrt{T}\Bigg[\|\mathbf{u}_y\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_F\sqrt{\log\left(\frac{2Km}{\delta}\right)} \\
&+ \sigma_p\|\mathbf{A}_y\|_2\log\left(\frac{2Km}{\delta}\right)\Bigg]\Bigg).
\end{aligned} \tag{122}$$

We next translate the scalar bound into a bound on $\|(\mathbf{w}_{y,r}^{(T)})^\top \mathbf{A}_y\|_2$. Let $\boldsymbol{\xi}' = \mathbf{A}_y\boldsymbol{\zeta}'$ be an independent sample from $\mathcal{D}_y$. Conditioned on $\mathbf{w}_{y,r}^{(T)}$, we have

$$\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}'\right\rangle = \left\langle \mathbf{A}_y^\top \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\zeta}'\right\rangle. \tag{123}$$

By the anti-concentration property in Condition 4.1, with constant probability,

$$\left|\left\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}' \right\rangle\right| \geq c \left\|\mathbf{A}_y^\top \mathbf{w}_{y,r}^{(T)}\right\|_2. \tag{124}$$

Combining equation 122 and equation 124 yields

$$\left\|\mathbf{A}_y^\top \mathbf{w}_{y,r}^{(T)}\right\|_2 \leq O\Bigg(\log(T) + \eta\sigma_n\sqrt{T}\Bigg[\|\mathbf{u}_y\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_F\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_2 \log\left(\frac{2Km}{\delta}\right)\Bigg]\Bigg). \tag{125}$$

Finally,

$$\left\|(\mathbf{W}_y^{(T)})^\top \mathbf{A}_y\right\|_F^2 = \sum_{r=1}^{m} \left\|\mathbf{A}_y^\top \mathbf{w}_{y,r}^{(T)}\right\|_2^2. \tag{126}$$

Therefore,

$$\left\|(\mathbf{W}_y^{(T)})^\top \mathbf{A}_y\right\|_F \leq O\Bigg(\sqrt{m}\log(T) + \eta\sigma_n\sqrt{mT}\Bigg[\|\mathbf{u}_y\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_F\sqrt{\log\left(\frac{2Km}{\delta}\right)} + \sigma_p\|\mathbf{A}_y\|_2 \log\left(\frac{2Km}{\delta}\right)\Bigg]\Bigg). \tag{127}$$

Since $\|\cdot\|_2 \leq \|\cdot\|_F$, the desired result follows. □

Lemma E.4. *Suppose that Condition 4.1 holds. For any $j \in [K]$, let $\boldsymbol{\xi} = \mathbf{A}_j\boldsymbol{\zeta}$ be an independent test noise vector, where $\boldsymbol{\zeta} \sim \mathcal{D}_\zeta$. Then, with probability at least $1-\delta$,*

$$\sum_{r=1}^{m} \mathbb{I}\left(\left\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi} \right\rangle > 0\right) \geq 0.1m. \tag{128}$$

Proof. For class $j \in [K]$, define

$$\mathbf{W}_j^{(T)} = \begin{bmatrix} (\mathbf{w}_{j,1}^{(T)})^\top \\ \vdots \\ (\mathbf{w}_{j,m}^{(T)})^\top \end{bmatrix} \in \mathbb{R}^{m\times d}. \tag{129}$$

For each iteration t, similarly define the Gaussian noise matrix

$$\mathbf{N}_j^{(t)} = \begin{bmatrix} (\mathbf{n}_{j,1}^{(t)})^\top \\ \vdots \\ (\mathbf{n}_{j,m}^{(t)})^\top \end{bmatrix} \in \mathbb{R}^{m\times d}. \tag{130}$$

Recall that the DP-SGD update is

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \frac{\eta}{B}\sum_{(\mathbf{x}_i,y_i)\in S^{(t)}} h(C,\mathbf{x}_i,y_i)\nabla L(\mathbf{W}^{(t)},\mathbf{x}_i,y_i) + \eta\mathbf{N}^{(t)}. \tag{131}$$

For the neuron $\mathbf{w}_{j,r}^{(t)}$, the contribution of a training sample $(\mathbf{x}_i, y_i)$ to the update is

$$\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)} = \frac{\eta}{mB}\sum_{(\mathbf{x}_i,y_i)\in S^{(t)}} h(C,\mathbf{x}_i,y_i)\Big[\mathbb{I}(y_i=j)\left(1-\mathrm{logit}_j(\mathbf{W}^{(t)},\mathbf{x}_i)\right) - \mathbb{I}(y_i\neq j)\,\mathrm{logit}_j(\mathbf{W}^{(t)},\mathbf{x}_i)\Big] \cdot \left[\sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}_{y_i}\right\rangle\right)\mathbf{u}_{y_i} + \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\right\rangle\right)\boldsymbol{\xi}_i\right] + \eta\mathbf{n}_{j,r}^{(t)}. \tag{132}$$

By the orthogonality assumption $\mathbf{u}_k^\top \mathbf{A}_j = \mathbf{0}$ for all $j, k \in [K]$, multiplying equation 132 by $\mathbf{A}_j$ eliminates all signal components. Summing over $t = 0, \ldots, T-1$, we obtain

$$\mathbf{W}_j^{(T)} \mathbf{A}_j = \mathbf{D}_1 + \mathbf{D}_2, \tag{133}$$

where

$$\boxed{\mathbf{D}_1 := \left(\mathbf{W}_j^{(0)} + \eta \sum_{t=0}^{T-1} \mathbf{N}_j^{(t)} \right) \mathbf{A}_j} \tag{134}$$

is the Gaussian component, and

$$\boxed{\mathbf{D}_2 := \begin{bmatrix} \beta_{j,1,1} & \cdots & \beta_{j,1,n} \\ \vdots & \ddots & \vdots \\ \beta_{j,m,1} & \cdots & \beta_{j,m,n} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_1^\top \\ \vdots \\ \boldsymbol{\xi}_n^\top \end{bmatrix} \mathbf{A}_j} \tag{135}$$

is the training-dependent component. Here, for every $r \in [m]$ and $i \in [n]$,

$$\begin{aligned} \beta_{j,r,i} := \frac{\eta}{mB} \sum_{t=0}^{T-1} & \mathbb{I}\left((\mathbf{x}_i, y_i) \in S^{(t)} \right) h(C, \mathbf{x}_i, y_i) \\ & \cdot \left[\mathbb{I}(y_i = j) \left(1 - \mathrm{logit}_j(\mathbf{W}^{(t)}, \mathbf{x}_i) \right) - \mathbb{I}(y_i \neq j) \, \mathrm{logit}_j(\mathbf{W}^{(t)}, \mathbf{x}_i) \right] \\ & \cdot \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \right). \end{aligned} \tag{136}$$

We first bound the rank of $\mathbf{D}_2$. Since the coefficient matrix in equation 135 has size $m \times n$, we have

$$\mathrm{rank}(\mathbf{D}_2) \leq \min\{m, n, d, \mathrm{rank}(\mathbf{A}_j)\} \leq n. \tag{137}$$

Notice that this rank bound holds regardless of the dependence of $\beta_{j,r,i}$ on the entire DP-SGD trajectory.

We next analyze $\mathbf{D}_1$. The initialization satisfies

$$\mathbf{w}_{j,r}^{(0)} \sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2 \mathbf{I}_d \right), \tag{138}$$

while the injected noises $\mathbf{n}_{j,r}^{(t)}$ are mutually independent and satisfy

$$\mathbf{n}_{j,r}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \sigma_n^2 \mathbf{I}_d \right). \tag{139}$$

Therefore,

$$\mathbf{w}_{j,r}^{(0)} + \eta \sum_{t=0}^{T-1} \mathbf{n}_{j,r}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \left(\sigma_0^2 + \eta^2 T \sigma_n^2 \right) \mathbf{I}_d \right). \tag{140}$$

Moreover, these vectors are independent across $r \in [m]$. Consequently,

$$\mathbf{W}_j^{(0)} + \eta \sum_{t=0}^{T-1} \mathbf{N}_j^{(t)} = \sqrt{\sigma_0^2 + \eta^2 T \sigma_n^2} \, \mathbf{Z}_j, \tag{141}$$

where $\mathbf{Z}_j \in \mathbb{R}^{m \times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Hence,

$$\mathbf{D}_1 = \sqrt{\sigma_0^2 + \eta^2 T \sigma_n^2} \, \mathbf{Z}_j \mathbf{A}_j. \tag{142}$$

Let

$$\mathbf{A}_j = \mathbf{U}_j \boldsymbol{\Sigma}_j \mathbf{V}_j^\top \tag{143}$$

be the singular value decomposition of $\mathbf{A}_j$. Since Gaussian distributions are rotationally invariant, $\mathbf{Z}_j \mathbf{U}_j$ is again a Gaussian matrix with i.i.d. standard normal entries on the $\mathrm{rank}(\mathbf{A}_j)$-dimensional column space. Therefore,

$$\mathrm{rank}(\mathbf{D}_1) = \min\{m, d, \mathrm{rank}(\mathbf{A}_j)\} \tag{144}$$

almost surely.

Combining this with equation 137, we obtain

$$\begin{aligned}\operatorname{rank}\left(\mathbf{W}_j^{(T)}\mathbf{A}_j\right) &= \operatorname{rank}(\mathbf{D}_1+\mathbf{D}_2)\\ &\geq \operatorname{rank}(\mathbf{D}_1)-\operatorname{rank}(\mathbf{D}_2)\\ &\geq \min\{m,d,\operatorname{rank}(\mathbf{A}_j)\}-n. \end{aligned} \tag{145}$$

By Condition 4.1,

$$\min\{m,d,\operatorname{rank}(\mathbf{A}_j)\}-n \geq 0.9m. \tag{146}$$

We next control the smallest relevant nonzero singular value. By Lemma C.1 and the singular value decomposition of $\mathbf{A}_j$, with probability at least $1-\delta$,

$$\begin{aligned}&\lambda_{\min\{m,d,\operatorname{rank}(\mathbf{A}_j)\}}(\mathbf{D}_1)\\ &\quad\geq c\sqrt{m}\sqrt{\sigma_0^2+\eta^2T\sigma_n^2}\,\lambda_{\min}^{+}(\mathbf{A}_j),\end{aligned} \tag{147}$$

where $\lambda_{\min}^{+}(\mathbf{A}_j)$ denotes the smallest nonzero singular value of $\mathbf{A}_j$.

Since $\operatorname{rank}(\mathbf{D}_2)\leq n$, the singular-value inequality for a rank-n perturbation implies

$$\begin{aligned}&\lambda_{\min\{m,d,\operatorname{rank}(\mathbf{A}_j)\}-n}\left(\mathbf{W}_j^{(T)}\mathbf{A}_j\right)\\ &\quad\geq \lambda_{\min\{m,d,\operatorname{rank}(\mathbf{A}_j)\}}(\mathbf{D}_1)\\ &\quad\geq c\sqrt{m}\sqrt{\sigma_0^2+\eta^2T\sigma_n^2}\,\lambda_{\min}^{+}(\mathbf{A}_j).\end{aligned} \tag{148}$$

On the other hand, by Lemma E.3,

$$\begin{aligned}\lambda_1\left(\mathbf{W}_j^{(T)}\mathbf{A}_j\right) &\leq \left\|\mathbf{W}_j^{(T)}\mathbf{A}_j\right\|_F\\ &\leq O\Bigg(\sqrt{m}\log(T)+\eta\sigma_n\sqrt{mT}\Bigg[\|\mathbf{u}_j\|_2\sqrt{\log\left(\frac{2Km}{\delta}\right)}\\ &\qquad+\sigma_p\|\mathbf{A}_j\|_F\sqrt{\log\left(\frac{2Km}{\delta}\right)}+\sigma_p\|\mathbf{A}_j\|_2\log\left(\frac{2Km}{\delta}\right)\Bigg]\Bigg).\end{aligned} \tag{149}$$

Therefore, if the over-parameterization condition in Condition 4.1 satisfies

$$m\geq c_1\log\left(\frac{d}{\delta}\right)\frac{\left[\log(T)+\eta\sigma_n\sqrt{T}\left(\|\mathbf{u}_j\|_2\sqrt{\log(2Km/\delta)}+\sigma_p(\|\mathbf{A}_j\|_F+\|\mathbf{A}_j\|_2)\sqrt{\log(2Km/\delta)}\right)\right]^2}{(\sigma_0^2+\eta^2T\sigma_n^2)\left(\lambda_{\min}^{+}(\mathbf{A}_j)\right)^2}, \tag{150}$$

then the singular-value condition required by Lemma C.9 is satisfied.

Now condition on the entire training trajectory and the resulting matrix $\mathbf{W}_j^{(T)}$. The test vector $\boldsymbol{\zeta}\sim\mathcal{D}_\zeta$ is independent of the training trajectory, and

$$\mathbf{W}_j^{(T)}\boldsymbol{\xi}=\mathbf{W}_j^{(T)}\mathbf{A}_j\boldsymbol{\zeta}. \tag{151}$$

By equation 146, the matrix $\mathbf{W}_j^{(T)}\mathbf{A}_j$ has rank at least $0.9m$, and by equation 148, equation 149, and equation 150, its nonzero singular values satisfy the condition of Lemma C.9.

Therefore, Lemma C.9 gives

$$\Pr_{\boldsymbol{\zeta}}\left[\sum_{r=1}^{m}\mathbb{I}\left(\left(\mathbf{W}_j^{(T)}\mathbf{A}_j\boldsymbol{\zeta}\right)_r\leq 0\right)\geq 0.9m\right]\leq\exp(-0.07m). \tag{152}$$

Since

$$\left(\mathbf{W}_j^{(T)}\mathbf{A}_j\boldsymbol{\zeta}\right)_r = \left\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\right\rangle, \tag{153}$$

we conclude that, with probability at least $1-\delta$,

$$\sum_{r=1}^{m}\mathbb{I}\left(\left\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\right\rangle > 0\right) \geq 0.1m. \tag{154}$$

This completes the proof. □

E.2 Proof of Statement 1 in Theorem 4.8

In this part, we prove Statement 1 in Theorem 4.8. For data samples following $(\mathbf{x}, y) \sim \mathcal{D}$, the test error satisfies

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{W}^{(t)}) &= \mathbb{P}\left[\arg\max_k F_k(\mathbf{W}^{(T)}, \mathbf{x}, y) \neq y\right] \\ &\leq \sum_{j\neq y}\mathbb{P}\left[F_y(\mathbf{W}^{(T)}, \mathbf{x}, y) \leq F_j(\mathbf{W}^{(T)}, \mathbf{x}, y)\right] \\ &= \sum_{j\neq y}\mathbb{P}\left[\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(T)}, \mathbf{u}_y\rangle) + \sigma(\langle \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi}\rangle) \leq \sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{j,r}^{(T)}, \mathbf{u}_y\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\rangle)\right]. \end{aligned} \tag{155}$$

For the features, we bound the loss for any $(\mathbf{x}, y) \sim \mathcal{D}$ through

$$\mathcal{L}(\mathbf{W}^{(T)}, \mathbf{x}, y) \leq \sum_{j\neq y}\mathbb{P}\left[\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(T)}, \mathbf{u}_y\rangle) \leq \sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{j,r}^{(T)}, \mathbf{u}_y\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\rangle)\right]. \tag{156}$$

Then, we can bound the model outputs as

$$\begin{aligned} &\sum_{r=1}^{m}\sigma\left(\langle \mathbf{w}_{y,r}^{(T)}, \mathbf{u}_y\rangle\right) \\ &= \sum_{r=1}^{m}\sigma\left(\left\langle \mathbf{w}_{y,r}^{(0)} + \frac{\eta}{B}\sum_{t'=1}^{T}\sum_{(\mathbf{x}_l, y_l)\in\mathcal{S}_y^{(t)}}(1-\mathrm{logit}_y(\mathbf{W}^{(t')}, \mathbf{x}_l))h(C, x_l, y_l)\sigma'(\langle \mathbf{w}_{y,r}^{(t')}, \mathbf{u}_y\rangle)\mathbf{u}_y + \eta\cdot\sum_{t'=1}^{T}n_{y,r}^{(t')}, \mathbf{u}_y\right\rangle\right) \\ &\geq \frac{3m}{10}\frac{\eta}{n}\sum_{t'=1}^{T}\sum_{(\mathbf{x}_l, y_l)\in\mathcal{S}_y}(1-\mathrm{logit}_y(\mathbf{W}^{(t')}, \mathbf{x}_l))\|\mathbf{u}_y\|_2^2\cdot\left(\frac{C\sqrt{m}}{\|\mathbf{u}_y\|_2 + \sqrt{Tr(A_y^T A_y)}}\right) - \eta T\sigma_n\|\mathbf{u}_y\|_2\sqrt{2\log(2/\delta)}, \end{aligned} \tag{157}$$

where the last inequality is by Lemma C.4 and the bound of the clipping multiplier. And

$$\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\rangle) \leq \sum_{r=1}^{m}|\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\rangle| = \sum_{r=1}^{m}\|\mathbf{A}_k^\top \mathbf{w}_{j,r}^{(T)}\|_2|\zeta'|, \tag{158}$$

where ζ' is a sub-Gaussian variable. Suppose that $\delta > 0$ and $\|\mathbf{A}_y^\top\mathbf{A}_j\|_F/\|\mathbf{A}_y^\top\mathbf{A}_j\|_{\mathrm{op}} \geq \Theta(\sqrt{\log(K^2/\delta)})$. For the term $\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\|_2$, with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\left\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\right\|_2 \\
&= \Bigg\| \mathbf{A}_y^\top \Bigg(\mathbf{w}_{j,r}^{(0)} + \frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_j} (1 - \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x})) h(C,x,y) \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi} \rangle) \boldsymbol{\xi} \\
&\quad - \frac{\eta}{n} \sum_{(\mathbf{x},y)\in\mathcal{S}\setminus\mathcal{S}_j} \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x}) h(C,x,y) \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi} \rangle) \boldsymbol{\xi} + \eta \cdot \sum_{t'=0}^{T-1} n_{j,r}^{(t)} \Bigg) \Bigg\|_2 \\
&\le \left\| \mathbf{A}_y^\top \mathbf{A}_j \left(\frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_j} (1 - \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x})) \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi} \rangle) \boldsymbol{\zeta} \right) \right\|_2 \qquad (159) \\
&\quad + \left\| \sum_{k\neq j} \mathbf{A}_y^\top \mathbf{A}_k \left(\frac{\eta}{n} \sum_{(\mathbf{x},y)\in\mathcal{S}_k} \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x}) \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi} \rangle) \boldsymbol{\zeta} \right) \right\|_2 + \eta \cdot T\sigma_n \sqrt{Tr(A_y^\top A_y)} \\
&\overset{(a)}{=} \Theta \left(\left\| \mathbf{A}_y^\top \left(\frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_j} (1 - \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x})) \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi} \rangle) \boldsymbol{\xi} \right) \right\|_2 \right) + \eta \cdot T\sigma_n \sqrt{Tr(A_y^\top A_y)} \\
&\overset{(b)}{=} \Theta \left(\frac{\eta}{n} \|\mathbf{A}_y^\top \mathbf{A}_j\|_F \cdot \sqrt{\sum_{(\mathbf{x},y)\in\mathcal{S}_j} \left(\sum_{t'=0}^{T-1} (1 - \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x})) \right)^2} \right) + \eta \cdot T\sigma_n \sqrt{\mathrm{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)},
\end{aligned}
$$

where the inequality is by Lemma E.2 and Condition 4.1, (a) is based on Lemma E.2, and (b) is based on the concentration of random vectors (Theorem 6.2.6 in Vershynin (2018)). Substituting (157) and (158) into (156), for any $(\mathbf{x}, y) \sim \mathcal{D}$ we have

$$
\begin{aligned}
&\mathcal{L}(\mathbf{W}^{(T)}, \mathbf{x}, y) \\
&\le \sum_{j\neq y} \mathbb{P} \left[\frac{m^{3/2}}{5} \frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_y} (1 - \mathrm{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})) \Lambda_y \|\mathbf{u}_y\|_2^2 - \eta T\sigma_n \|\mathbf{u}_y\|_2 \sqrt{2\log(2/\delta)} \le \sum_{r=1}^{m} |\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi} \rangle| + |\langle \mathbf{w}_{j,r}^{(T)}, \mathbf{u}_y \rangle| \right] \\
&= \sum_{j\neq y} \mathbb{P} \left[\frac{m^{3/2}}{5} \frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_y} (1 - \mathrm{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})) \Lambda_y \|\mathbf{u}_y\|_2^2 - \eta T\sigma_n \|\mathbf{u}_y\|_2 \sqrt{2\log(2/\delta)} \le \sum_{r=1}^{m} \left\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\right\|_2 |\zeta'| + |\langle \mathbf{w}_{j,r}^{(0)}, \mathbf{u}_y \rangle| \right] \\
&\le \sum_{j\neq y} \mathbb{P} \Bigg[\sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_y} (1 - \mathrm{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})) \Lambda_y \|\mathbf{u}_y\|_2^2 - \eta T\sigma_n \|\mathbf{u}_y\|_2 \sqrt{2\log(2/\delta)} \\
&\quad \le c \left(\|\mathbf{A}_y^\top \mathbf{A}_j\|_F \cdot \sqrt{\sum_{(\mathbf{x},y)\in\mathcal{S}_j} \left(\sum_{t'=0}^{T-1} (1 - \mathrm{logit}_j(\mathbf{W}^{(t')}, \mathbf{x})) \right)^2} + \eta \cdot T\sigma_n \sqrt{\mathrm{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)} \right) |\zeta'| + \mathcal{O}\left(\frac{n}{\eta} \sqrt{\log\left(\frac{4Km}{\delta}\right)} \|\mathbf{u}_y\|_2 \sigma_0 \right) \Bigg] \\
&\le \sum_{j\neq y} \exp \left[-c_1 \cdot \left(\frac{|\mathcal{S}_y| \Lambda_y \|\mathbf{u}_y\|_2^2 - \sigma_n \|\mathbf{u}_y\|_2 \sqrt{2\log(2/\delta)}}{\sqrt{|\mathcal{S}_j|} \|\mathbf{A}_y^\top \mathbf{A}_j\|_F + \sigma_n \sqrt{\mathrm{Tr}(\mathbf{A}_y^\top \mathbf{A}_y)}} \right)^2 \right],
\end{aligned}
$$

where $c, c_1 > 0$ are some constants and the last inequality is obtained from Hoeffding's inequality.

E.3 Proof of Statement 2 in Theorem 4.8

Lemma E.5. *for $(\mathbf{x}, y)$ sampled from L-long-tailed data distribution T_i defined in Definition 4.7, each $i \in [K]$, the following bound holds with probability at least $1 - \delta$:*

$$
\|\boldsymbol{\xi}\|_2 \le O\left(\sqrt{L^2 + d} \|\mathbf{A}_y\|_2 \right)
$$

Proof. For $(\mathbf{x}, y) \sim T_i$, $\left\langle \sum_{r\in\mathcal{R}(\boldsymbol{\xi})} \mathbf{w}_{y,r}^{(T)}, \boldsymbol{\xi} \right\rangle \ge L \|\mathbf{A}_y^\top \sum_{r\in\mathcal{R}(\boldsymbol{\xi})} \mathbf{w}_{y,r}^{(T)}\|_2$

Let $u = \frac{\mathbf{A}^\top \mathbf{w}_{y,r}^{(T)}}{\|\mathbf{A}^\top \mathbf{w}_{y,r}^{(T)}\|_2}$ be the unit direction vector associated with the screening condition. By the linearity of the inner product, the condition $(\mathbf{x}, y) \sim \mathcal{T}_i$ is equivalent to: $u^\top \boldsymbol{\xi} \ge L$

Due to the rotational invariance of $\mathcal{N}(0, I_d)$, we can decompose $\boldsymbol{\xi}$ as $\boldsymbol{\xi} = zu + \boldsymbol{\xi}_\perp$, where $z = u^\top \boldsymbol{\xi} \in \mathbb{R}$ is a scalar and $\boldsymbol{\xi}_\perp = (I - uu^\top)\boldsymbol{\xi} \in \mathbb{R}^d$ is the component orthogonal to u. Here, $z \sim \mathcal{N}(0,1)$ and $\boldsymbol{\xi}_\perp \sim \mathcal{N}(0, I - uu^\top)$ are independent. The squared norm is $\|\boldsymbol{\xi}\|_2^2 = z^2 + \|\boldsymbol{\xi}_\perp\|_2^2$.

The term $\|\boldsymbol{\xi}_\perp\|_2^2$ follows a Chi-squared distribution with $d-1$ degrees of freedom, i.e., $\|\boldsymbol{\xi}_\perp\|_2^2 \sim \chi^2_{d-1}$. By the Laurent-Massart concentration inequality, with probability at least $1 - \delta/2$:

$$\|\boldsymbol{\xi}_\perp\|_2^2 \leq d - 1 + 2\sqrt{(d-1)\ln(2/\delta)} + 2\ln(2/\delta). \tag{160}$$

Given the condition $z > L$, we examine the tail of the truncated Gaussian. For any $s > 0$:

$$\mathbb{P}(z > L + s \mid z > L) = \frac{1 - \Phi(L+s)}{1 - \Phi(L)} \leq \exp\left(-Ls - \frac{s^2}{2}\right). \tag{161}$$

By setting this probability to $\delta/2$, we find that with probability at least $1 - \delta/2$, $z \leq L + \frac{\ln(2/\delta)}{L}$. Squaring this term yields $z^2 \leq L^2 + 2\ln(2/\delta) + o(1)$.

Applying a union bound over the events, we obtain:

$$\|\boldsymbol{\xi}\|_2^2 \leq d + L^2 + 2\sqrt{d\ln(2/\delta)} + 4\ln(2/\delta). \tag{162}$$

Taking the square root and using the property $\|\mathbf{A}_y\boldsymbol{\xi}\|_2 \leq \|\mathbf{A}_y\|_{\text{op}}\|\boldsymbol{\xi}\|_2$ completes the proof. □

Proof of Statement 2: Furthermore, for long-tailed data distribution, by Lemma E.2 and union bound, we have

$$\mathcal{L}(\mathbf{W}^{(t)}, \mathbf{x}, y) \leq \sum_{j \neq y} \mathbb{P}\left[\sum_{r=1}^m \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) \leq \sum_{r=1}^m \sigma\left(\left\langle \mathbf{w}_{j,r}^{(T)}, \mathbf{u}_y\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right]. \tag{163}$$

First, we consider the model behavior under non-DP training, and select long-tailed data by Definition 4.7. Then, we bound the gap between DP and non-DP setting.

Suppose $\delta > 0$, by Condition 4.1, we have $\|\mathbf{A}_y^\top \mathbf{A}_y\|_F / \|\mathbf{A}_y^\top \mathbf{A}_y\|_{\text{op}} \geq \Theta(\sqrt{\log(K^2/\delta)})$, for any $y \in [K]$. With probability at least $1 - \delta$, for all $y \in [K]$, under **non-DP** setting, we have

$$\begin{aligned}
&\left\|\mathbf{A}_y^\top \mathbf{w}_{y,r,clean}^{(T)}\right\|_2 \\
&= \Bigg\| \mathbf{A}_y^\top \Bigg(\mathbf{w}_{y,r,clean}^{(0)} + \frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_y} (1 - \text{logit}_y(\mathbf{W}^{(t')}, \mathbf{x}))\sigma'(\langle \mathbf{w}_{y,r,clean}^{(t')}, \boldsymbol{\xi}\rangle)\boldsymbol{\xi} \\
&\quad - \frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}\setminus\mathcal{S}_y} \text{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})\sigma'(\langle \mathbf{w}_{y,r,clean}^{(t')}, \boldsymbol{\xi}\rangle)\boldsymbol{\xi} \Bigg) \Bigg\|_2 \\
&\geq \left\| \mathbf{A}_y^\top \mathbf{A}_y \left(\frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_y} (1 - \text{logit}_y(\mathbf{W}^{(t')}, \mathbf{x}))\sigma'(\langle \mathbf{w}_{y,r,clean}^{(t')}, \boldsymbol{\xi}\rangle)\boldsymbol{\zeta} \right) \right\|_2 \\
&\quad - \left\| \sum_{j\neq y} \mathbf{A}_y^\top \mathbf{A}_j \left(\frac{\eta}{n} \sum_{t'=0}^{T-1} \sum_{(\mathbf{x},y)\in\mathcal{S}_j} \text{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})\sigma'(\langle \mathbf{w}_{y,r,clean}^{(t')}, \boldsymbol{\xi}\rangle)\boldsymbol{\zeta} \right) \right\|_2 \\
&\overset{(a)}{=} \Omega\left(\frac{\eta}{n} \|\mathbf{A}_y^\top \mathbf{A}_y\|_F \cdot \sqrt{\sum_{(\mathbf{x},y)\in\mathcal{S}_y} \left(\sum_{t'=0}^{T-1} (1 - \text{logit}_y(\mathbf{W}^{(t')}, \mathbf{x})) \right)^2} \right),
\end{aligned} \tag{164}$$

where (a) is obtained by the condition $\|\mathbf{A}_y^\top \mathbf{A}_y\|_F = \Omega(\|\mathbf{A}_y^\top \mathbf{A}_j\|_F)$ for all $j, y \in [K]$ and $j \neq y, K = \Theta(1)$.

Then, we consider DP setting. Similar to the proof of (159), we also have

$$\left\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\right\|_2 \leq \Theta\left(\frac{\eta}{n}\|\mathbf{A}_y^\top \mathbf{A}_j\|_F \cdot \sqrt{\sum_{(\mathbf{x},y)\in\mathcal{S}_j}\left(\sum_{t'=0}^{T-1}(1-\mathrm{logit}_j(\mathbf{W}^{(t')},\mathbf{x}))\right)^2} + \eta\cdot T\sigma_n\sqrt{Tr(A_y^\top A_y)}\right), \tag{165}$$

Denote $\bar{\mathcal{S}}_j(\boldsymbol{\xi}) = \{r\in[m] : \langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\xi}\rangle > 0\}$. We have

$$\left\|\sum_{r\in\bar{\mathcal{S}}_j(\boldsymbol{\xi})}\mathbf{A}_y^\top \mathbf{w}_{y,r}^{(T)}\right\|_2 \geq \Theta\left(\frac{\eta m}{n}\|\mathbf{A}_y^\top \mathbf{A}_y\|_F \cdot \sqrt{\sum_{(\mathbf{x},y)\in\mathcal{S}_y}\left(\sum_{t'=0}^{T-1}(1-\mathrm{logit}_y(\mathbf{W}^{(t')},\mathbf{x}))\right)^2}\right), \tag{166}$$

because Lemma E.4 holds.

Finally, we have:

$$\begin{aligned}
&\mathcal{L}(\mathbf{W}^{(T)},\mathbf{x},y) \leq \sum_{j\neq y}\mathbb{P}\left[\frac{1}{m}\sum_{r=1}^m \sigma(\langle \mathbf{w}_{y,r}^{(T)},\boldsymbol{\xi}\rangle) \leq \frac{1}{m}\sum_{r=1}^m \sigma(\langle \mathbf{w}_{j,r}^{(T)},\mathbf{u}_y\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(T)},\boldsymbol{\xi}\rangle)\right]\\
&= \sum_{j\neq y}\mathbb{P}\left[\left|\sum_{r\in\bar{\mathcal{S}}_y(\boldsymbol{\xi})}\langle \mathbf{w}_{y,r}^{(T)},\boldsymbol{\xi}\rangle\right| \leq \sum_{r=1}^m\left|\langle \mathbf{w}_{j,r}^{(T)},\boldsymbol{\xi}\rangle\right| + \left|\langle \mathbf{w}_{j,r}^{(T)},\mathbf{u}_y\rangle\right|\right]\\
&\leq \sum_{j\neq y}\mathbb{P}\left[L\cdot\left\|\sum_{r\in\bar{\mathcal{S}}_y(\boldsymbol{\xi})}\mathbf{A}_y^\top \mathbf{w}_{y,r,clean}^{(T)}\right\|_2 h(C,\mathbf{x},y) - \eta T\sigma_n\sqrt{L^2+d}\|\mathbf{A}_y\|_{op} - \sum_{r=1}^m\left\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\right\|_2\right.\\
&\qquad\left.\leq \sum_{r=1}^m|\langle \mathbf{w}_{j,r}^{(T)},\boldsymbol{\xi}\rangle| - \sum_{r=1}^m\mathbb{E}|\langle \mathbf{w}_{j,r}^{(T)},\boldsymbol{\xi}\rangle| + |\langle \mathbf{w}_{j,r}^{(0)},\mathbf{u}_y\rangle|\right]\\
&\leq \sum_{j\neq y}\exp\left(-c'\frac{\left(\begin{array}{c} L\left\|\sum_{r\in\bar{\mathcal{S}}_y(\boldsymbol{\xi})}\mathbf{A}_y^\top \mathbf{w}_{y,r,\mathrm{clean}}^{(T)}\right\|_2 h(C,\mathbf{x},y) - \eta T\sigma_n\sqrt{L^2+d}\,\|\mathbf{A}_y\|_{\mathrm{op}}\\ -\sum_{r=1}^m\left\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\right\|_2 - \mathcal{O}\left(\sqrt{\log(4Km/\delta)}\,\|\mathbf{u}_y\|_2\sigma_0\right)\end{array}\right)^2}{\sum_{r=1}^m\left\|\mathbf{A}_y^\top \mathbf{w}_{j,r}^{(T)}\right\|_2^2}\right)\\
&\leq \sum_{j\neq y}\exp\left(-c_2\cdot\frac{(L\cdot\sqrt{|\mathcal{S}_y|}\Lambda_y\|\mathbf{A}_y^\top \mathbf{A}_y\|_F - n\sigma_n\sqrt{L^2+d}\|\mathbf{A}_y\|_{op})^2}{|\mathcal{S}_j|\|\mathbf{A}_y^\top \mathbf{A}_j\|_F^2}\right),
\end{aligned} \tag{167}$$

where $c'>0$ is a constant, the first inequality is by the condition for long-tailed data that $|\langle \mathbf{w}_{y,r}^{(T)},\boldsymbol{\xi}\rangle| \geq L\left\|\mathbf{A}_y^\top \mathbf{w}_{y,r}^{(T)}\right\|_2$ and the gap between DP and non-DP setting, the second inequality is by Hoeffding's inequality and the last inequality is by (164) and (165).