Towards Understanding Generalization in DP-GD: A Case Study in Training Two-Layer CNNs

Zhongjie Shi[1*], **Puyu Wang**[2], **Chenyang Zhang**[3], **Yuan Cao**[3†]

[1]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, United States,
[2]Department of Computer Science, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany,
[3]Department of Statistics and Actuarial Science, School of Computing and Data Science, The University of Hong Kong, Hong Kong,
zshi332@gatech.edu, wang.puyu@cs.rptu.de, chyzhang@connect.hku.hk, yuancao@hku.hk

Abstract

Modern deep learning techniques focus on extracting intricate information from data to achieve accurate predictions. However, the training datasets may be crowdsourced and include sensitive information, such as personal contact details, financial data, and medical records. As a result, there is a growing emphasis on developing privacy-preserving training algorithms for neural networks that maintain good performance while preserving privacy. In this paper, we investigate the generalization and privacy performances of the differentially private gradient descent (DP-GD) algorithm, which is a private variant of the gradient descent (GD) by incorporating additional noise into the gradients during each iteration. Moreover, we identify a concrete learning task where DP-GD can achieve superior generalization performance compared to GD in training two-layer Huberized ReLU convolutional neural networks (CNNs). Specifically, we demonstrate that, under mild conditions, a small signal-to-noise ratio can result in GD producing training models with poor test accuracy, whereas DP-GD can yield training models with good test accuracy and privacy guarantees if the signal-to-noise ratio is not too small. This indicates that DP-GD has the potential to enhance model performance while ensuring privacy protection in certain learning tasks. Numerical simulations are further conducted to support our theoretical results.

1 Introduction

Modern deep learning (DL) algorithms are designed to extract fine-grained, high-dimensional patterns from data to achieve superior predictive performance. However, this ability to exploit intricate details can inadvertently expose sensitive information contained within the training data. In practical scenarios, datasets may include personally identifiable information such as health records, contact details, or financial transactions. Without appropriate safeguards, models trained using standard gradient descent (GD) methods are susceptible to various privacy attacks, such as membership inference or model inversion. As a result, it becomes critically important to study and understand the behavior of privacy-preserving variants of GD.

Differential Privacy (DP) (Dwork et al. 2006; Dwork, Roth et al. 2014) is a widely adopted framework for designing privacy-preserving deep learning algorithms with strong theoretical guarantees. It ensures that the output of an algorithm is minimally influenced by any single data point in the input dataset, thereby safeguarding individual privacy. Extensive research has been dedicated to developing efficient differentially private learning algorithms while maintaining strong statistical performance (Bassily et al. 2019, 2020; Bassily, Smith, and Thakurta 2014; Feldman, Koren, and Talwar 2020; Wang et al. 2022b; Chaudhuri, Monteleoni, and Sarwate 2011). In this paper, we focus on DP-GD, a commonly used private learning algorithm that introduces noise into the gradient updates to protect individual privacy. Understanding the interplay between generalization and privacy in such algorithms is crucial for advancing responsible and secure deep learning systems.

However, enforcing strong privacy guarantees often comes at the cost of model performance. Previous work has confirmed this tension between privacy and utility (Chaudhuri, Monteleoni, and Sarwate 2011; Kifer and Machanavajjhala 2011; Bassily, Guzmán, and Menart 2021; Yang et al. 2021; Carvalho et al. 2023), highlighting that increasing the level of privacy protection, typically by injecting more noise, can degrade the predictive accuracy of machine learning models, since excessive noise may obscure crucial patterns in the data, leading to underfitting or poor generalization. On the other hand, Conversely, insufficient privacy safeguards risk exposing sensitive information, eroding public trust, and violating legal obligations. This trade-off is particularly pronounced in high-stakes applications, such as healthcare diagnostics or financial forecasting, where both accuracy and data confidentiality are paramount.

Therefore, it is important to build trustworthy DL systems that can satisfy both privacy and performance requirements. Whereas achieving a satisfying balance between privacy and generalization for DP-GD is a challenging problem for general cases, we may start from the following question:

Are there specific learning tasks for which DP-GD can simultaneously provide good privacy guarantees and maintain competitive generalization performance?

[*]The work was done while the author was affiliated with The University of Hong Kong.

[†]Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

This paper provides an affirmative answer to this question by identifying a concrete binary classification task in which DP-GD not only protects privacy but also provably yields higher test accuracy than standard GD, showing that there is not always a trade-off. Our choices of learning tasks, models, theorem conditions, experiments are specifically aimed at showing the existence of scenarios where DP enhances accuracy. Specifically, we demonstrate that in the context of training two-layer Huberized ReLU CNNs under certain mild conditions, DP-GD can outperform GD in terms of test accuracy. This result reveals that privacy-preserving training does not always entail a loss in utility, and in some cases, the injected noise may even act as a form of regularization that benefits generalization. We summarize the main contributions of the paper in the following.

- We provide a refined analysis showing that when the signal-to-noise ratio is relatively low, under mild assumptions about the problem setup, model design, and hyperparameter configuration, GD can minimize the training loss to an arbitrarily small value. However, the corresponding test loss and test error remain bounded below by a constant, indicating poor generalization.
- We show that when the signal-to-noise ratio is not too small, DP-GD can, under similarly mild assumptions, achieve an arbitrarily small training loss. Furthermore, by applying the tool of early stopping, DP-GD is capable of achieving both strong generalization performance and meaningful privacy protection simultaneously.
- Comparing these two theoretical outcomes reveals that, with appropriate model architecture and careful tuning of hyperparameters, DP-GD can outperform the standard GD in terms of generalization in certain tasks. This highlights the potential of DP training algorithms not only to preserve privacy but also to enhance performance in specific scenarios, offering valuable guidance for model design and hyperparameter selection.

Notations. We use lower case letters, lower case bold face letters, and upper case bold face letters to denote scalars, vectors, and matrices respectively. For a vector $\mathbf{v} = (v_1, \cdots, v_d)^\top$, we denote by $\|\mathbf{v}\|_2 := \big(\sum_{j=1}^d v_j^2\big)^{1/2}$ its l_2 norm. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to omit the logarithmic terms.

2 Related Work

Implicit bias of neural networks. A growing body of research has focused on the aspect of implicit bias, which refers to the intrinsic tendency of learning algorithms to favor solutions with certain underlying structures—often those considered to be "simple" or low-complexity (Neyshabur, Tomioka, and Srebro 2014; Soudry et al. 2018; Ji and Telgarsky 2019b; Wang et al. 2022a; Xie and Li 2024; Zhang, Zou, and Cao 2024). Within the context of neural networks, several studies have explored how this phenomenon manifests for GD. For instance, Lyu and Li (2019) and Ji and Telgarsky (2020) showed that when training q-homogeneous neural networks using GD, the direction of convergence aligns with a KKT point of the maximum ℓ_2-margin optimization problem. Extending this line of inquiry, Lyu et al. (2021) established a stronger convergence result under the assumption of symmetric data. Additional insights into the implicit bias of deep linear networks have been provided by Ji and Telgarsky (2019a, 2020), who demonstrated that the weight matrices in each layer eventually converge to rank-one structures. On data that is nearly orthogonal, Frei et al. (2022) proved that gradient flow in leaky ReLU networks leads to linear decision boundaries, and that the stable rank of the resulting model remains bounded by a constant. Kou, Chen, and Gu (2024) extended these results to standard gradient descent under similar data assumptions. Cao et al. (2022) investigated the GD training dynamics of two-layer polynomial ReLU CNNs, identifying specific signal-to-noise ratio thresholds that determine whether the model converges to the underlying signal or fits the noise. Building on this, Kou et al. (2023) extended the analysis to standard ReLU CNNs. More recently, Zhang et al. (2025) demonstrated that training Huberized ReLU CNNs with gradient descent enables the model to robustly learns the intrinsic dimension of the data signals. Lastly, Vardi (2023) compiled a comprehensive survey summarizing key developments and open questions in the study of implicit bias in deep learning.

Differential privacy. A large body of work has investigated the privacy and utility guarantees of differentially private gradient-based methods. The gradient perturbation mechanism, first introduced by Song, Chaudhuri, and Sarwate (2013), forms the foundation for widely studied algorithms such as DP-GD and DP-SGD. In particular, Abadi et al. (2016) proposed the first algorithm for deep learning with differential privacy. Bassily, Smith, and Thakurta (2014); Bassily et al. (2019); Bassily, Guzmán, and Menart (2021); Feldman, Koren, and Talwar (2020); Wang et al. (2021); Asi et al. (2021) demonstrated that both DP-GD and DP-SGD can achieve optimal utility bounds under different settings. Specifically, an excess risk bound of order $O(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ for non-strongly convex problems and $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ for convex problems, where n is the sample size, d is input dimension, and (ϵ, δ) are the privacy parameters. For non-convex problems, Zhang et al. (2017) established the utility bound $O(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ for DP-SGD in terms of the squared gradient norm (i.e., first-order optimality) for nonconvex smooth objectives. Wang, Chen, and Xu (2019) provided a unified analysis of DP-GD and DP-SVRG for both convex and nonconvex settings, and specifically demonstrated the utility bound $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ of DP-GD in terms of the objective gap for objectives satisfying the Polyak-Łojasiewicz (PL) condition. Very recently, Bu et al. (2023) provided the first convergence analysis of DP-GD for deep learning, using insights on the training dynamics and the neural tangent kernel (NTK). Yet, their results only showed that DP-GD with global clipping converges monotonically to zero loss without providing convergence rates.

3 Problem Setting

We consider a specific binary classification task with the use of two-layer CNNs. We present the data distribution in the following definition, where the input data com-

prises two types of components: *label-dependent signals* and *label-independent noises*. This simplified setting can already demonstrate the existence of scenarios where DP enhances accuracy. The simplifications are necessary for tractable theoretical analyses. We consider Huberized ReLU activation as it is smooth, which helps our analysis. Notably, similar setups have been considered in recent works (Li, Wei, and Ma 2019; Allen-Zhu and Li 2020, 2022; Cao et al. 2022; Kou et al. 2023), making our choice relatively standard.

Definition 1. *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be a fixed vector representing the signal contained in each data point. Each data point $(\boldsymbol{x}, y)$ with $\boldsymbol{x} = [\boldsymbol{x}^{(1)\top}, \boldsymbol{x}^{(2)\top}]^\top \in \mathbb{R}^{2d}$ and $y \in \{-1, 1\}$ is generated from the following data distribution $\mathcal{D}$:*

1. *The label y is generated as a Rademacher random variable.*
2. *A noise vector $\boldsymbol{\xi}$ is generated from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_p^2 \cdot (\boldsymbol{I} - \boldsymbol{\mu}\boldsymbol{\mu}^\top \cdot \|\boldsymbol{\mu}\|_2^{-2}))$.*
3. *One of $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ is randomly selected and then assigned as $y \cdot \boldsymbol{\mu}$, which represents the signal; the other is then given by $\boldsymbol{\xi}$, which represents noises.*
4. *The signal-to-noise ratio (SNR) is defined as SNR $= ||\boldsymbol{\mu}||_2 / \sigma_p \sqrt{d}$.*

Two-layer CNNs. We consider two-layer convolutional neural networks (CNNs)

$$f(\boldsymbol{W}, \boldsymbol{x}) = F_{+1}(\boldsymbol{W}_{+1}, \boldsymbol{x}) - F_{-1}(\boldsymbol{W}_{-1}, \boldsymbol{x}),$$

where $F_{+1}(\boldsymbol{W}_{+1}, \boldsymbol{x})$ and $F_{-1}(\boldsymbol{W}_{-1}, \boldsymbol{x})$ are defined as:

$$\begin{aligned} F_j(\boldsymbol{W}_j, \boldsymbol{x}) &= \frac{1}{m} \sum_{r=1}^{m} \left[\sigma(\langle \boldsymbol{w}_{j,r}, \boldsymbol{x}^{(1)} \rangle) + \sigma\left(\langle \boldsymbol{w}_{j,r}, \boldsymbol{x}^{(2)} \rangle \right) \right] \\ &= \frac{1}{m} \sum_{r=1}^{m} \left[\sigma(\langle \boldsymbol{w}_{j,r}, y \cdot \boldsymbol{\mu} \rangle) + \sigma\left(\langle \boldsymbol{w}_{j,r}, \boldsymbol{\xi} \rangle \right) \right] \end{aligned}$$

for $j = \{+1, -1\}$, and m is the number of convolutional filters in F_{+1} and F_{-1}. We consider the Huburized ReLU activation function $\sigma(\cdot)$ which is defined as

$$\sigma(z) = q^{-1} \kappa^{1-q} z^q \cdot \mathbf{1}_{\{z \in [0, \kappa]\}} + \left(z - \kappa + \frac{\kappa}{q} \right) \cdot \mathbf{1}_{\{z > \kappa\}}$$

where κ is the threshold between polynomial and linear functions, and $q \geq 3$, and $\mathbf{1}$ is the indicator function. We use $w_{j,r} \in \mathbb{R}^d$ to denote the weight of the r-th filter, and $\boldsymbol{W}_j$ is the collection of weights associated with F_j. We also use $\boldsymbol{W}$ to denote the collection of all weights.

Training algorithm. The above CNN model is trained by minimizing the empirical cross-entropy loss function

$$L_S(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left[y_i \cdot f(\boldsymbol{W}, \boldsymbol{x}_i) \right],$$

where $\ell(t) = \log(1 + e^{-t})$ is the logistic loss, and $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is the training data set. Moreover, the test loss is defined as

$$L_{\mathcal{D}}(\boldsymbol{W}) := \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \ell \left[y \cdot f(\boldsymbol{W}, \boldsymbol{x}) \right],$$

and the test error is defined as

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{W}) := \mathbb{P}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left(y \cdot f(\boldsymbol{W}, \boldsymbol{x}) < 0 \right),$$

We consider DP-GD with Gaussian initialization, where each entry of $\boldsymbol{W}_{+1}$ and $\boldsymbol{W}_{-1}$ is sampled from a Gaussian distribution $\mathcal{N}(0, \sigma_0^2)$. The update rule at step t is given by

$$\boldsymbol{w}_{j,r}^{(t+1)} = \boldsymbol{w}_{j,r}^{(t)} - \eta \left(\nabla_{\boldsymbol{w}_{j,r}} L_S(\boldsymbol{W}^{(t)}) + \boldsymbol{b}_{j,r,t} \right). \quad (1)$$

where the added Gaussian noises $\boldsymbol{b}_{j,r,t} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2 \boldsymbol{I})$, and we introduce a shorthand notation $\ell_i'^{(t)} := \ell'[y_i \cdot f(\boldsymbol{W}^{(t)}, \boldsymbol{x}_i)]$.

This differs from the classical GD in that we add an additional Gaussian noise on the gradient in each iteration. Moreover, the update rule of GD is given by

$$\boldsymbol{w}_{j,r}^{(t+1)} = \boldsymbol{w}_{j,r}^{(t)} - \eta \left(\nabla_{\boldsymbol{w}_{j,r}} L_S(\boldsymbol{W}^{(t)}) \right). \quad (2)$$

4 Main Results

In this section, we present our main theoretical findings. In particular, we construct a specific binary classification task using two-layer CNNs, where the SNR satisfies the condition $\widetilde{\Omega}(n^{\frac{1}{q}}) \leq \text{SNR}^{-1} \leq \min\{\frac{\sqrt{d}}{Cm^2}, \frac{\sqrt{n}}{C}\}$. Under this setting, we show that the training loss of both GD and DP-GD can converge to an arbitrarily small value. Whereas DP-GD can outperform GD in terms of the generalization performance. We note that our SNR conditions are not intended to guide practice, but are outcomes of theoretical analyses. They define the specific regime where privacy can improve accuracy, and reasonably exclude cases with very low/high SNR, where the task is too difficult/easy and both DP-GD and GD perform similarly poorly/well. This allows us to focus on settings where DP-GD and GD can be distinguished. While the SNR conditions do not directly guide practice, our finding that privacy can sometimes enhance accuracy has practical implications by deepening our understanding of the relationship between privacy and accuracy.

4.1 Noise Memorization of GD

The theoretical analysis of GD is based on the following specific conditions, where we identify an SNR condition $\text{SNR}^{-1} \geq \widetilde{\Omega}(n^{\frac{1}{q}})$, and mild conditions on the choice of hyperparameters in the problem setting and training algorithm. We consider the learning period $0 \leq t \leq T^*$, where $T^* = \widetilde{O}\left(\frac{\kappa^{q-1} mn}{\eta \sigma_0^{q-2} (\sigma_p \sqrt{d})^q} + \frac{m^3 n}{\eta \epsilon ||\boldsymbol{\mu}||_2^2} \right)$.

Condition 1. *Suppose there exists a sufficiently large constant C, such that the following hold:*

1. *The threshold κ of the activation function is sufficiently small: $\kappa = O(1)$.*
2. *The SNR is sufficiently small: $\text{SNR}^{-1} \geq \widetilde{\Omega}(n^{\frac{1}{q}})$.*
3. *The dimension d is sufficiently large: $d \geq C m^{\frac{2q}{q-2}} n^{\frac{2q-2}{q-2}} \kappa^{-\frac{2q-2}{q-2}} (\log(\frac{mn^2}{\delta}))^2 (\log(T^*))^2$.*
4. *The training sample size n and the convolutional kernel size m of CNNs is sufficiently large: $n \geq C \log(\frac{m}{\delta})$, $m \geq C \log\left(\frac{n \widetilde{T}^*}{\delta} \right)$.*

5. *The standard deviation of Gaussian initialization* σ_0 *satisfies:* $\frac{Cn}{\sigma_p d}\sqrt{\log\left(\frac{n^2}{\delta}\right)}\log(T^*) \leq \sigma_0 \leq (C\max\{||\boldsymbol{\mu}||_2 m^{\frac{2}{q-2}} n^{\frac{1}{q-2}}\sqrt{\log(\frac{mn}{\delta})}, \sigma_p\sqrt{d}\})^{-1}\kappa^{\frac{q-1}{q-2}}$.
6. *The learning rate* η *is sufficiently small:* $\eta \leq \left(C\max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}\right)^{-1}$.

The conditions on d, n, m are to ensure that the learning problem is in a sufficiently over-parameterized setting, and similar conditions have been made in Chatterji and Long (2021); Cao et al. (2022); Frei, Chatterji, and Bartlett (2022). The condition on the SNR and the lower bound condition on σ_0 ensure that the memorization of the noises dominates the learning of the signal in GD. The upper bound on σ_0 ensures that within T^* iterations, the learning of the signal is always small and around the initialization order, even if the training loss converges. The condition imposed on η serves as a sufficient requirement to guarantee that GD can effectively minimize the training loss.

Theorem 1. *Under Condition 1, for any* $\epsilon > 0$*, denote* $T_1 = \widetilde{\Theta}\left(\frac{\kappa^{q-1}mn}{\eta\sigma_0^{q-2}(\sigma_p\sqrt{d})^q}\right)$*, and* $T_2 = T_1 + \frac{36nm^2}{\eta\sigma_p^2 d}$*. Then within* $T^* = T_1 + \widetilde{O}\left(\frac{m^3 n}{\eta\epsilon||\boldsymbol{\mu}||_2^2}\right)$ *iterations, with probability at least* $1 - 6\delta$*, we have*

1. *The training loss converges: there exists a time* $t \leq T^*$ *such that* $L_S(\boldsymbol{W}^{(t)}) \leq \epsilon$.
2. *The test loss is always large: for any* $0 \leq t \leq T^*$ *we have that* $L_{\mathcal{D}}(\boldsymbol{W}^{(t)}) \geq 0.1$.
3. *The test error is always large: suppose that* $\sigma_0 \leq \frac{C_3}{m||\boldsymbol{\mu}||_2\sqrt{d}}$ *for some small constant* C_3*. For any* $T_2 \leq t \leq T^*$*, we have that* $\mathcal{R}_{\mathcal{D}}(\boldsymbol{W}^{(t)}) \geq 0.11$.

Theorem 1 shows that when the SNR is small, the training loss of GD can converge to any accuracy ϵ. However, the test loss of the trained CNN has at least a constant order. Moreover, we provide a sufficient condition on σ_0 such that the test error of the trained CNN has at least a constant order as well, when the training iteration is not too small. We note that this upper bound condition on σ_0 is only a sufficient condition. We need this condition due to the technical difficulties, and this condition has the potential to be relaxed.

Overview of proof technique At the core of our analyses is a *signal-noise decomposition* of the filters in the CNN trained by the optimization algorithm. According to the GD update rule (2), it is clear that the gradient descent iterate $\boldsymbol{w}_{j,r}^{(t)}$ is a linear combination of its random initialization $\boldsymbol{w}_{j,r}^{(0)}$, the signal vector $\boldsymbol{\mu}$ and the noise vectors in the training data $\boldsymbol{\xi}_i$, $i \in [n]$. Motivated by this observation, we introduce the following definition.

Definition 2. *Let* $\boldsymbol{w}_{j,r}^{(t)}$ *for* $j \in \{\pm 1\}$, $r \in [m]$ *be the CNN convolution filters in the* t*-th iteration of GD* (2)*. Then there exist unique coefficients* $\gamma_{j,r}^{(t)} \geq 0$ *and* $\rho_{j,r,i}^{(t)}$ *such that*

$$\boldsymbol{w}_{j,r}^{(t)} = \boldsymbol{w}_{j,r}^{(0)} + \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot j\boldsymbol{\mu} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

We further denote $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbf{1}(\rho_{j,r,i}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbf{1}(\rho_{j,r,i}^{(t)} \leq 0)$*. Then we have*

$$\boldsymbol{w}_{j,r}^{(t)} = \boldsymbol{w}_{j,r}^{(0)} + \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot j\boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i. \tag{3}$$

The normalization factors $\|\boldsymbol{\mu}\|_2^{-2}, \|\boldsymbol{\xi}_i\|_2^{-2}$ are to ensure that $\gamma_{j,r}^{(t)} \approx \langle \boldsymbol{w}_{j,r}^{(t)}, j\boldsymbol{\mu}\rangle$ tracks signal learning and $\rho_{j,r,i}^{(t)} \approx \langle \boldsymbol{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle$ tracks noise memorization. Using Definition 2, we can reduce the study of the CNN learning process to a careful assessment of the coefficients $\gamma_{j,r}^{(t)}$, $\overline{\rho}_{j,r,i}^{(t)}$, $\underline{\rho}_{j,r,i}^{(t)}$ throughout training, where $\gamma_{j,r}^{(t)}$ and $\overline{\rho}_{j,r,i}^{(t)}$ are increasing monotonically and $\underline{\rho}_{j,r,i}^{(t)}$ is decreasing monotonically. The main idea is that the memorization of noise can achieve κ and grow further to a constant level, whereas the learning of the signal is always small and around the initialization values.

Stage 1. We note that the Huberized ReLU activation function is piece-wise continuous, where the part between $[0, \kappa]$ is a polynomial of order q, and the part between $[\kappa, \infty)$ is linear, with the threshold κ. For the first stage, we show that at time $T_1 = \widetilde{\Theta}\left(\frac{\kappa^{q-1}mn}{\eta\sigma_0^{q-2}(\sigma_p\sqrt{d})^q}\right)$, there exists a neuron of noise memorization $\langle \boldsymbol{w}_{y_i,r}^{(T_1)}, \boldsymbol{\xi}_i\rangle$ that can hit the threshold κ. However, all neurons of signal learning $\langle \boldsymbol{w}_{j,r}^{(t)}, j\mu\rangle$ are around the initialization order.

Lemma 1. *Under Condition 1, we can find a time* $T_1 = \widetilde{\Theta}\left(\frac{\kappa^{q-1}mn}{\eta\sigma_0^{q-2}(\sigma_p\sqrt{d})^q}\right)$*, such that*

- $\max_r \langle \boldsymbol{w}_{y_i,r}^{(T_1)}, \boldsymbol{\xi}_i\rangle \geq \kappa$, $\max_{j,r} \overline{\rho}_{j,r,i}^{(T_1)} \geq \kappa$*, for all* $i \in [n]$.
- $\max_{j,r} \langle \boldsymbol{w}_{j,r}^{(t)}, j\mu\rangle = \widetilde{O}(\sigma_0\|\mu\|_2)$, $\max_{j,r} \gamma_{j,r}^{(T_1)} = \widetilde{O}(\sigma_0\|\mu\|_2)$*, for all* $0 \leq t \leq T_1$.
- $\max_{r,i} \left|\langle \boldsymbol{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right| = \widetilde{O}(\sigma_0\sigma_p\sqrt{d})$, $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(T_1)}| = \widetilde{O}(\sigma_0\sigma_p\sqrt{d})$*, for all* $0 \leq t \leq T_1$.

Stage 2. For the second stage, we demonstrate that the training loss will converge to the accuracy ϵ at time $T^* = \widetilde{O}\left(\frac{\kappa^{q-1}mn}{\eta\sigma_0^{q-2}(\sigma_p\sqrt{d})^q} + \frac{m^3 n}{\eta\epsilon||\boldsymbol{\mu}||_2^2}\right)$. Moreover, since $\sum_{s=T_1}^{\widetilde{T}-1}\sum_{i=1}^{n} |\ell_i'^{(t)}| \leq \sum_{s=T_1}^{\widetilde{T}-1} L_S(\boldsymbol{W}^{(s)})$, the convergence of the training error in Stage 2 indicates that the growth of the signal learning $\langle \boldsymbol{w}_{j,r}^{(t)}, j\mu\rangle$ is still small and around the initialization order.

Lemma 2. *Under Condition 1, let* $T^* = T_1 + \left\lfloor \frac{\|\boldsymbol{W}^{(T_1)} - \boldsymbol{W}^*\|_F^2}{2\eta\epsilon} \right\rfloor = T_1 + \widetilde{O}\left(\frac{m^3 n}{\eta\epsilon||\boldsymbol{\mu}||_2^2}\right)$*. Then we have*

- $\max_{j,r} \langle \boldsymbol{w}_{j,r}^{(t)}, j\boldsymbol{\mu}\rangle = \widetilde{O}(\sigma_0||\boldsymbol{\mu}||_2)$, $\max_{j,r} \gamma_{j,r}^{(t)} = \widetilde{O}(\sigma_0||\boldsymbol{\mu}||_2)$*, for all* $T_1 \leq t \leq T^*$.
- $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| = \widetilde{O}(\sigma_0\sigma_p\sqrt{d})$*, for all* $T_1 \leq t \leq T$.

Besides, for all $T_1 \leq t \leq T^*$, we have

$$\frac{1}{t-T_1+1}\sum_{s=T_1}^{t} L_S(\boldsymbol{W}^{(s)}) \leq \frac{\|\boldsymbol{W}^{(T_1)}-\boldsymbol{W}^*\|_F^2}{(2q-1)\eta(t-T_1+1)} + \frac{\epsilon}{(2q-1)}.$$

Therefore, we can find an iterate with training loss smaller than ϵ within T iterations.

Generalization analysis. To prove test loss is large, we only need to show that for a new example $(\boldsymbol{x}, y)$, the noise term $|\langle \boldsymbol{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle|$ and the signal term $|\langle \boldsymbol{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle|$ are both small. Notice that $\langle \boldsymbol{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle \sim \mathcal{N}(0, \sigma_p^2\|\boldsymbol{w}_{j,r}^{(t)}\|_2^2)$. Therefore, we can show that with high probability

$$|\langle \boldsymbol{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle| \leq \widetilde{O}\left(\sigma_0\sigma_p\sqrt{d} + \frac{mn}{\sqrt{d}}\right) = O(\kappa).$$

This, together with the first property in Lemma 2 demonstrates that $y_i f(\boldsymbol{W}^{(t)}, \boldsymbol{x}) \leq 1$, thus the test loss is large at a constant level.

To prove the lower bound of test error, we first show that at time $T_2 = T_1 + \frac{36nm^2}{\eta\sigma_p^2 d}$, the neurons have memorized the noises, i.e., $\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} \geq m$.

Lemma 3. *Under Condition 1, let $T_2 = T_1 + \frac{36nm^2}{\eta\sigma_p^2 d}$. For the time period $T_2 \leq t \leq T^*$ and all $i \in [n]$, we have*

$$\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} \geq m.$$

Denote $g(\boldsymbol{\xi}) = \sum_r \sigma(\langle \boldsymbol{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle) - \sum_r \sigma(\langle \boldsymbol{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\rangle)$, and the set $\Omega := \left\{\boldsymbol{\xi} \middle| g(\boldsymbol{\xi}) > \widetilde{O}(m\sigma_0||\boldsymbol{\mu}||_2)\right\}$. Since the signal learning of each neuron is at most $\widetilde{O}(\sigma_0||\boldsymbol{\mu}||_2)$, we can give a lower bound of the test error by

$$\begin{aligned}\mathcal{R}_{\mathcal{D}}(\boldsymbol{W}^{(t)}) &= \mathbb{P}\big(yf(\boldsymbol{W}^{(t)}, \boldsymbol{x}) < 0\big) \\ &\geq 0.5\mathbb{P}\bigg(\sum_{r=1}^{m}\sigma(\langle \boldsymbol{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle) - \sum_{r=1}^{m}\sigma(\langle \boldsymbol{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\rangle) \\ &\quad > \widetilde{O}(m\sigma_0||\boldsymbol{\mu}||_2)\bigg) \geq 0.5\mathbb{P}(\Omega).\end{aligned}$$

Finally, by utilizing Lemma 3, the symmetry property of $\boldsymbol{\xi}$ and the properties of total variance distance, we can derive that $\mathbb{P}(\Omega) \geq 0.23$, therefore get the desired results on the test error lower bound.

4.2 Signal Learning of DP-GD

The theoretical analysis of DP-GD hinges on the following specific conditions, where we identify an SNR condition $\text{SNR}^{-1} \leq \min\left\{\frac{\sqrt{d}}{Cm^2}, \frac{\sqrt{n}}{C}\right\}$, and mild conditions on the choice of hyperparameters in the problem setting and training algorithm. We consider the learning period $0 \leq t \leq \widetilde{T}^*$, where $\widetilde{T}^*$ is the maximum number of iterations.

Condition 2. *Suppose there exists a sufficiently large constant C, such that the following hold:*

1. *The threshold κ of the activation function is sufficiently small: $\kappa^2 \leq \frac{\min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}{Cm^2\max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}$.*
2. *The SNR satisfies: $\text{SNR}^{-1} \leq \min\left\{\frac{\sqrt{d}}{Cm^2}, \frac{\sqrt{n}}{C}\right\}$.*
3. *The dimension d is sufficiently large: $d \geq C\max\left\{m^4n^2, \frac{m^2n^2}{\kappa^2}\log(\frac{n^2}{\delta})(\log(T^*))^2\right\}$.*
4. *The training sample size n and the convolutional kernel size m of CNNs is sufficiently large: $n \geq C\log(\frac{m}{\delta})$, $m \geq C\log\left(\frac{n\widetilde{T}^*}{\delta}\right)$.*
5. *The standard deviation of the Gaussian initialization σ_0 satisfies: $\sigma_0 \leq \min\left\{\frac{1}{Cm\sigma_p\sqrt{d}}, \frac{\kappa}{C\max\{||\boldsymbol{\mu}||_2, \sigma_p\sqrt{d}\}\sqrt{\log\left(\frac{mn}{\delta}\right)}}\right\}$.*
6. *The learning rate η satisfies: $\frac{Cm^3\kappa^2\max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}{||\boldsymbol{\mu}||_2^2\min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} \leq \eta \leq \frac{m}{C||\boldsymbol{\mu}||_2^2}$.*
7. *The standard deviation of the Gaussian noise σ_b satisfies: $\sigma_b \leq \left(C\eta\max\left\{||\boldsymbol{\mu}||_2, \sigma_p\sqrt{d}\right\}\sqrt{\widetilde{T}^*}\log^2\left(\frac{mn\widetilde{T}^*}{\delta}\right)\right)^{-1}$.*

The condition on the SNR is to ensure that the privacy guarantee and the test loss of DP-GD are good. The upper bounds on σ_0 and η are to ensure that the test loss is good. The lower bound on η is to ensure that the value of signal learning can achieve the threshold κ at a constant order time. The condition on σ_b is to ensure that the influence of the added Gaussian noise is smaller than the learning of the signal.

The following theorem demonstrates that the training loss of DP-GD can achieve an arbitrarily small accuracy ϵ within $\widetilde{T}^* = \Theta\left(\frac{\kappa^2}{\eta^2\sigma_b^2\min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} + \frac{nm^2}{\eta\epsilon\max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}\right)$ iterations.

Theorem 2. *Under Condition 2, for any $\epsilon > 0$, denote $\widetilde{T}_1 = \Theta\left(\frac{\kappa^2}{\eta^2\sigma_b^2\min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}\right)$, if we choose $\widetilde{T}^* = \widetilde{T}_1 + \Theta\left(\frac{nm^2}{\eta\epsilon\max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}\right)$, with probability at least $1 - 6\delta$, we have $L_S(\boldsymbol{W}^{(\widetilde{T}^*)}) \leq \epsilon$.*

In the following theorem, we indicate that with a proper choice of σ_b to control the level of the injected noise in DP-GD, and with the tool of early stopping at around $\widetilde{T}_2 = \Theta(\frac{m}{\eta||\boldsymbol{\mu}||_2^2})$ iterations, we can achieve a good test error and a good privacy guarantee simultaneously for DP-GD.

Theorem 3. *Under Condition 2, by choosing $\sigma_b = \Theta\left(\sqrt{\frac{||\boldsymbol{\mu}||_2^2}{\eta m^3\max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}}\right)$, and $\widetilde{T}_2 = \Theta\left(\frac{m}{\eta||\boldsymbol{\mu}||_2^2}\right)$. With probability at least $1 - 6\delta$, we have $\mathcal{R}_{\mathcal{D}}(\boldsymbol{W}^{(\widetilde{T}_2)}) \leq 0.01$. Moreover, the DP-GD with $\widetilde{T}_2$ iterations satisfies $\left(\frac{C_4m^3\max^2\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}{n^2||\boldsymbol{\mu}||_2^4}\log\frac{2}{\delta}, \delta\right)$-DP for some positive constant C_4.*

This theorem highlights the generalization advantage of DP-GD over standard GD. The key insight is that the in-

jected Gaussian noise in DP-GD facilitates the signal learning component in surpassing the threshold κ of the Huberized ReLU when $t \geq \widetilde{T}_1$. Once this threshold is crossed, the derivative $\sigma'(z)$ becomes 1, allowing the signal to grow more effectively. In contrast, the signal learning component remains at initialization order, and $\sigma'(z)$ remains very small, limiting the signal learning process in standard GD.

Overview of proof technique We utilize the same signal-noise decomposition method as in Definition 2 as the core method for our analysis, the different point is that we need the additional added Gaussian noise terms $\eta \sum_{k=0}^{t-1} \boldsymbol{b}_{j,r,k}$ in DP-GD.

Definition 3. *Let $\boldsymbol{w}_{j,r}^{(t)}$ for $j \in \{+1, -1\}$, $r \in [m]$ be the convolution filters of the CNN at the t-th iteration of noisy SGD. Then there exist unique coefficients $\gamma_{j,r}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ such that*

$$\boldsymbol{w}_{j,r}^{(t)} = \boldsymbol{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot ||\boldsymbol{\mu}||_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot ||\boldsymbol{\xi}_i||_2^{-2} \cdot \boldsymbol{\xi}_i - \eta \sum_{k=0}^{t-1} \boldsymbol{b}_{j,r,k}. \quad (4)$$

By further denoting $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$, and $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$, we have

$$\boldsymbol{w}_{j,r}^{(t)} = \boldsymbol{w}_{j,r}^{(0)} + j\gamma_{j,r}^{(t)} \cdot ||\boldsymbol{\mu}||_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \cdot ||\boldsymbol{\xi}_i||_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot ||\boldsymbol{\xi}_i||_2^{-2} \cdot \boldsymbol{\xi}_i - \eta \sum_{k=0}^{t-1} \boldsymbol{b}_{j,r,k}. \quad (5)$$

Training loss. We briefly introduce the proof of the convergence of training loss for DP-GD. The key idea is to analyze the growth of the signal and noises together during the training process, which is denoted as $\lambda_i^{(t)} = \frac{1}{m} \sum_{r=1}^{m} \left(\gamma_{y_i,r}^{(t)} + \overline{\rho}_{y_i,r,i}^{(t)} \right)$. It is different from the analysis of GD since we cannot identify whether the memorization of noise or the learning of signal dominates during the training process, due to the perturbations caused by the added Gaussian noise in the gradient. However, the added Gaussian noises $\eta \sum_{k=0}^{t-1} \boldsymbol{b}_{j,r,k}$ in DP-GD can help at least one neuron in $\langle \boldsymbol{w}_{j,r}^{(t)}, j\boldsymbol{\mu} \rangle$ or $\langle \boldsymbol{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ to achieve the threshold κ of the Huberized ReLU activation function, for each $j \in \{\pm 1\}$ and $i \in [n]$, when the iteration $t \geq \widetilde{T}_1$ is large enough. This property makes it different from the analysis of GD, since it ensures that at least one neuron in signal learning to hit κ, and therefore help the signal learning term to grow outside the scope of the initialization order.

Specifically, for the time $t \geq \widetilde{T}_1 = \Theta\left(\frac{\kappa^2}{\eta^2 \sigma_b^2 \min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} \right)$, for each $j \in \{\pm 1\}$ and $i \in [n]$, we have

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

Therefore, according to the update rule of $\gamma_{j,r}^{(t)}$, $\overline{\rho}_{j,r,i}^{(t)}$ based on Definition 3, we have

$$\lambda_i^{(t+1)} \geq \lambda_i^{(t)} + \frac{\eta \max\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}{12nm^2} e^{-\lambda_i^{(t)}},$$

based on this iteration inequality, we can prove that $L_S(\boldsymbol{W}^{(\widetilde{T}^*)}) \leq \epsilon$ within $\widetilde{T}^*$ iterations.

Test error. To prove the upper bound of the test error, we need the following lemma, which demonstrates that the learning of signal $\frac{1}{m} \sum_{r=1}^{m} \gamma_{j,r}^{(t)}$ can be as large as $\Theta(\frac{1}{m})$ after $\widetilde{T}_2$ iterations.

Lemma 4. *Under Condition 2, denote $c_1 = \frac{3\eta \max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}{nm}$, $\widetilde{T}_1 = \Theta\left(\frac{\kappa^2}{\eta^2 \sigma_b^2 \min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} \right)$, and $\widetilde{T}_2 = \widetilde{T}_1 + \Theta\left(e^{c_1} (\widetilde{T}_1 + \frac{1}{c_1}) \right)$. Then, for any $\widetilde{T}_2 \leq t \leq \widetilde{T}^*$, we have*

$$\frac{1}{m} \sum_{r=1}^{m} \gamma_{j,r}^{(t)} = \Omega\left(\frac{n||\boldsymbol{\mu}||_2^2}{e^{c_1} m \max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} \right), \quad (6)$$

This lemma is important for the generalization analysis of DP-GD, since it tells us when $t \geq \widetilde{T}_2$, the signal learning can escape the initialization order.

Next, the following lemma shows that if σ_b and t are chosen to satisfy a condition, we can derive a high probability upper bound of the test error at time t. The condition is chosen to ensure that the signal learning term can dominate the influence of added Gaussian noise to achieve the right classification.

Lemma 5. *Under Condition 2, let $\widetilde{T}_2 \leq t \leq \widetilde{T}^*$ and satisfies that*

$$\eta \sigma_b ||\boldsymbol{\mu}||_2 \sqrt{t} \leq \frac{n||\boldsymbol{\mu}||_2^2}{Cm \max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}$$

for some large constant C, we have

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{W}^{(t)}) \leq \exp\left(-C_2 \left(\frac{n||\boldsymbol{\mu}||_2^2}{e^{c_1} m \max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} \cdot \frac{1}{\sigma_0 \sigma_p \sqrt{d} + \frac{\sigma_p m}{||\boldsymbol{\mu}||_2} + \frac{mn}{\sqrt{d}} + \eta \sigma_b \sigma_p \sqrt{d\widetilde{T}_2}} \right)^2 \right),$$

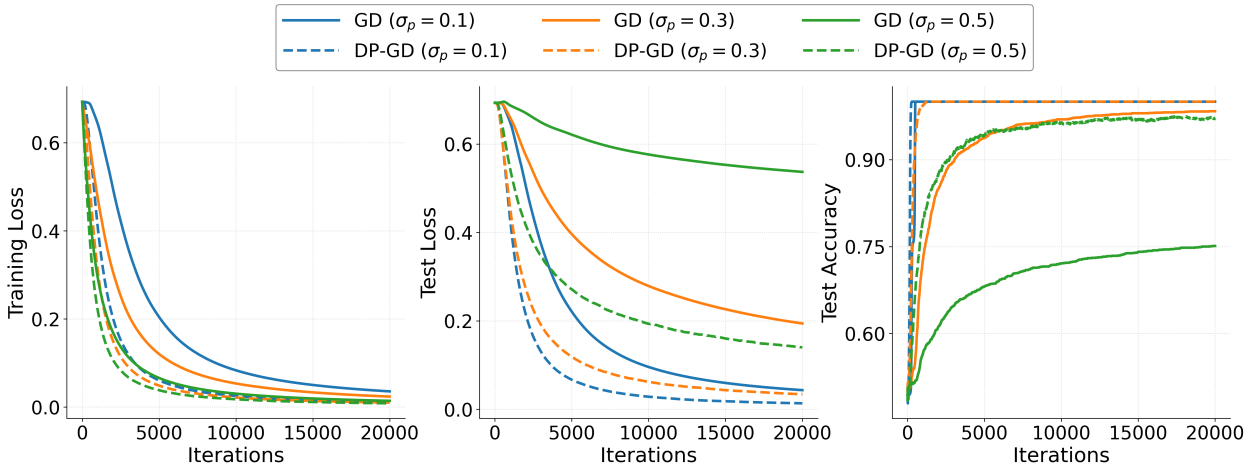
where C_2 is a positive constant.

A condition for good test error and DP. The following lemma gives a DP guarantee for DP-GD at time T.

Lemma 6 (Privacy guarantee)**.** *The DP-GD algorithm with T iterations satisfies $(\frac{T\lambda(2||\boldsymbol{\mu}||_2^2 + 3\sigma_p^2 d)}{\sigma_b^2 n^2 m} + \frac{\log(2/\delta)}{\lambda - 1}, \delta)$-DP for any $\lambda > 1$.*

Based on Lemma 5 and Lemma 6, we can select a condition to ensure that DP-GD can achieve a good test error and a good DP guarantee at the same time. We choose η large enough depending on σ_b to ensure that

$$\widetilde{T}_1 = \Theta\left(\frac{\kappa^2}{\eta^2 \sigma_b^2 \min\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}} \right) = O(1).$$



(a) Training Loss (b) Test Loss (c) Test Accuracy

Figure 1: Training loss, test loss, and test accuracy of two-layer CNNs trained with GD and DP-GD. Results are shown for three noise levels ($\sigma_p \in \{0.1, 0.3, 0.5\}$) with fixed signal strength ($\|\boldsymbol{\mu}\|_2 = 1$).

We also choose η to be not that large, so that $c_1 = O(1)$. Under these two conditions on η, we have

$$\widetilde{T}_2 = \widetilde{T}_1 + \Theta\Big(e^{c_1}(\widetilde{T}_1 + \frac{1}{c_1})\Big) = \Theta\Big(\frac{nm}{\eta \max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}\Big).$$

Furthermore, we choose σ_b to be the largest condition such that it satisfies

$$\eta\sigma_b\|\boldsymbol{\mu}\|_2\sqrt{\widetilde{T}_2} \leq \frac{n||\boldsymbol{\mu}||_2^2}{Cm\max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}},$$
$$\eta\sigma_b\sigma_p\sqrt{d\widetilde{T}_2} \leq \frac{n||\boldsymbol{\mu}||_2^2}{Cm\max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}},$$

so that the DP guarantee is as good as possible. Such a choice of σ_b further provides us with a detailed condition on the lower bound of η. Furthermore, according to Lemma 5, we need to choose σ_0, σ_p, d to satisfy the conditions to ensure that

$$\sigma_0\sigma_p\sqrt{d} + \frac{\sigma_p m}{||\boldsymbol{\mu}||_2} + \frac{mn}{\sqrt{d}} \leq \frac{n||\boldsymbol{\mu}||_2^2}{Cm\max\{n||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}.$$

Finally, according to Lemma 6, DP-GD with $\widetilde{T}_2$ iterations satisfies $\Big(\frac{C_4 m^3 \max^2\{||\boldsymbol{\mu}||_2^2, \sigma_p^2 d\}}{n^2||\boldsymbol{\mu}||_2^4}\log\frac{2}{\delta}, \delta\Big)$-DP for some positive constant C_4. Therefore, we need the SNR condition that $\frac{\sigma_p\sqrt{d}}{||\boldsymbol{\mu}||_2} \leq \frac{\sqrt{n}}{C}$ to ensure that the DP guarantee is good.

5 Numerical Experiments

In this section, we empirically validate our theoretical analysis by demonstrating that a two-layer CNN trained with DP-GD exhibits stronger noise robustness than one trained with GD. Following Definition 1, we generate three datasets with fixed signal scale $\|\boldsymbol{\mu}\|_2 = 1$ and noise levels $\sigma_p \in \{0.1, 0.3, 0.5\}$. Unless otherwise stated, we set input dimension $d = 2000$, number of convolutional kernels $m = 100$, training sample size $n = 1000$, DP-GD noise parameter $\sigma_b = 0.01$, and learning rate $\eta = 0.1$. Although these settings are milder than the theoretical requirements in Condition 2, they still capture the over-parameterized regime and support the generality of our theory. The code is available at https://github.com/ZhongjieSHI/Paper-Codes.

The results are summarized in Figure 1. Figure 1(a) shows that both GD and DP-GD achieve fast training loss decay across all noise levels, consistent with Theorem 1 and Theorem 2. However, their generalization performance differs substantially as noise increases. As shown in Figures 1(b) and 1(c): (i) for $\sigma_p = 0.1$, both methods achieve near-zero test loss and almost 100% accuracy; (ii) for $\sigma_p = 0.3$, DP-GD attains lower test loss (by about 0.2) and slightly higher accuracy (100% vs. 95%); (iii) for $\sigma_p = 0.5$, DP-GD maintains strong generalization (test loss < 0.2, 95% accuracy), while GD largely fails, reaching only 75% accuracy. These results confirm that DP-GD significantly improves generalization performance under high noise, in line with our theoretical predictions.

6 Conclusion and Future Work

This paper presents a theoretical study showing that, in certain binary classification tasks using two-layer Huberized ReLU CNNs, under mild conditions on the problem setting, model architecture, and hyperparameters in the training algorithm, the training loss of both GD and DP-GD can converge to an arbitrarily small value. Nonetheless, by using the tool of early stopping, DP-GD can achieve better generalization performance than GD under appropriate signal-to-noise conditions, and ensure good privacy guarantees at the same time, highlighting its potential in achieving a trustworthy deep learning scheme in certain learning tasks. An important future work direction is to derive a more refined analysis of the privacy guarantee by utilizing the detailed properties during the training dynamics. It is also interesting to generalize the results in this paper to other algorithms and learning tasks.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. Yuan Cao is partially supported by NSFC 12301657 and Hong Kong RGC-ECS 27308624. The work of Puyu Wang is partially supported by the Alexander von Humboldt Foundation.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

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

Allen-Zhu, Z.; and Li, Y. 2022. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd annual symposium on foundations of computer science (FOCS)*, 977–988. IEEE.

Asi, H.; Feldman, V.; Koren, T.; and Talwar, K. 2021. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, 393–403. PMLR.

Bassily, R.; Feldman, V.; Guzmán, C.; and Talwar, K. 2020. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, 4381–4391.

Bassily, R.; Feldman, V.; Talwar, K.; and Guha Thakurta, A. 2019. Private stochastic convex optimization with optimal rates. In *Advances in neural information processing systems*, volume 32.

Bassily, R.; Guzmán, C.; and Menart, M. 2021. Differentially private stochastic optimization: New results in convex and non-convex settings. In *Advances in Neural Information Processing Systems*, volume 34, 9317–9329.

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

Bu, Z.; Wang, H.; Dai, Z.; and Long, Q. 2023. On the convergence and calibration of deep learning with differential privacy. *Transactions on machine learning research*, 2023: https–openreview.

Cao, Y.; Chen, Z.; Belkin, M.; and Gu, Q. 2022. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35: 25237–25250.

Carvalho, T.; Moniz, N.; Faria, P.; and Antunes, L. 2023. Towards a data privacy-predictive performance trade-off. *Expert Systems with Applications*, 223: 119785.

Chatterji, N. S.; and Long, P. M. 2021. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129): 1–30.

Chaudhuri, K.; Monteleoni, C.; and Sarwate, A. D. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

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

Frei, S.; Chatterji, N. S.; and Bartlett, P. 2022. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, 2668–2703. PMLR.

Frei, S.; Vardi, G.; Bartlett, P.; Srebro, N.; and Hu, W. 2022. Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data. In *The Eleventh International Conference on Learning Representations*.

Ji, Z.; and Telgarsky, M. 2019a. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*.

Ji, Z.; and Telgarsky, M. 2019b. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, 1772–1798. PMLR.

Ji, Z.; and Telgarsky, M. 2020. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33: 17176–17186.

Kifer, D.; and Machanavajjhala, A. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 193–204.

Kou, Y.; Chen, Z.; Chen, Y.; and Gu, Q. 2023. Benign overfitting in two-layer relu convolutional neural networks. In *International conference on machine learning*, 17615–17659. PMLR.

Kou, Y.; Chen, Z.; and Gu, Q. 2024. Implicit Bias of Gradient Descent for Two-layer ReLU and Leaky ReLU Networks on Nearly-orthogonal Data. *Advances in Neural Information Processing Systems*, 36.

Li, Y.; Wei, C.; and Ma, T. 2019. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, 11669–11680.

Lyu, K.; and Li, J. 2019. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. In *7th International Conference on Learning Representations, ICLR 2019*.

Lyu, K.; Li, Z.; Wang, R.; and Arora, S. 2021. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34: 12978–12991.

Neyshabur, B.; Tomioka, R.; and Srebro, N. 2014. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Song, S.; Chaudhuri, K.; and Sarwate, A. D. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, 245–248. IEEE.

Soudry, D.; Hoffer, E.; Nacson, M. S.; Gunasekar, S.; and Srebro, N. 2018. The Implicit Bias of Gradient Descent on Separable Data. *J. Mach. Learn. Res.*, 19: 70:1–70:57.

Vardi, G. 2023. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93.

Wang, B.; Meng, Q.; Zhang, H.; Sun, R.; Chen, W.; Ma, Z.-M.; and Liu, T.-Y. 2022a. Does Momentum Change the Implicit Regularization on Separable Data? In *Advances in Neural Information Processing Systems*, volume 35, 26764–26776. Curran Associates, Inc.

Wang, D.; Chen, C.; and Xu, J. 2019. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, 6526–6535. PMLR.

Wang, P.; Lei, Y.; Ying, Y.; and Zhang, H. 2022b. Differentially private SGD with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56: 306–336.

Wang, P.; Yang, Z.; Lei, Y.; Ying, Y.; and Zhang, H. 2021. Differentially private empirical risk minimization for AUC maximization. *Neurocomputing*, 461: 419–437.

Xie, S.; and Li, Z. 2024. Implicit Bias of AdamW: ℓ_∞-Norm Constrained Optimization. In *International Conference on Machine Learning*, 54488–54510. PMLR.

Yang, Z.; Lei, Y.; Wang, P.; Yang, T.; and Ying, Y. 2021. Simple stochastic and online gradient descent algorithms for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 34, 20160–20171.

Zhang, C.; Gao, P.; Zou, D.; and Cao, Y. 2025. Gradient Descent Robustly Learns the Intrinsic Dimension of Data in Training Convolutional Neural Networks. *arXiv preprint arXiv:2504.08628*.

Zhang, C.; Zou, D.; and Cao, Y. 2024. The Implicit Bias of Adam on Separable Data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhang, J.; Zheng, K.; Mou, W.; and Wang, L. 2017. Efficient private ERM for smooth objectives. *arXiv preprint arXiv:1703.09947*.