Characterizing the Training Dynamics of Private Fine-tuning with Langevin diffusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

We show that **d**ifferentially **p**rivate **f**ull **f**ine-**t**uning (DP-FFT) can distort pre-trained backbone features based on both theoretical and empirical results. We identify the cause of the distortion as the misalignment between the pre-trained backbone and the randomly initialized linear head. We prove that a sequential fine-tuning strategy can mitigate the feature distortion: first-linear-probing-then-fine-tuning (DP-LP-FFT). A new approximation scheme allows us to derive approximate upper and lower bounds on the training loss of DP-LP and DP-FFT, in a simple but canonical setting of 2-layer neural networks with ReLU activation. Experiments on real-world datasets and architectures are consistent with our theoretical insights. Moreover, our theory suggests a trade-off of privacy budget allocation in multi-phase fine-tuning methods like DP-LP-FFT.

1 Introduction

Today, many differentially-private (DP) machine learning pipelines proceed in two phases: (1) A model is pre-trained (non-privately) on a public dataset. (2) The model is then fine-tuned on private data, using DP optimization techniques such as DP stochastic gradient descent (DP-SGD) and its variants (Hoory et al., 2021; De et al., 2022; Tang et al., 2023; Zhang et al., 2024b). Pre-training a backbone model on public data enables differentially private fine-tuning to achieve improved performance across various downstream tasks (Yu et al., 2022) and is proven to be necessary in some cases (Ganesh et al., 2023a).

Despite these advances, the effect of DP on fine-tuning training dynamics remains poorly understood. Several key questions are yet to be answered: (1) how does randomness (both of initialization and DP optimization) impact the pre-trained representations? (2) What are the convergence rates of common fine-tuning methods, such as DP **f**ull **f**ine-**t**uning (DP-FFT) and DP **l**inear **p**robing (DP-LP, where feature representations are frozen, and only the linear head is fine-tuned)? (3) Prior work suggests that combining an early stage of DP-LP with a later stage of DP-FFT yields better privacy-utility tradeoffs (Tang et al., 2023), yet there is no theoretical understanding of this phenomenon, nor is it clear how to optimally combine these fine-tuning methods.

Answering these questions theoretically requires an analysis that can capture the fine-grained optimization dynamics of DP fine-tuning. We seek a model of DP finetuning that satisfies 2 properties. (1) **Architecture-sensitivity:** The convergence dynamics must differentiate between representation learning in the backbone and learning in the linear head. The analyses of Bassily et al. (2014),Wang et al. (2022),Fang et al. (2023),Ganesh et al. (2023b) focus only on the network's dimension, failing to capture this distinction. (2) **Ability to model nonlinearities:** The model should account for the nonlinearities introduced by multi neural layers, unlike existing methods that simplify analysis by linearizing neural networks (Ye et al., 2023a; Wang et al., 2024). We propose a novel approximation of DP-SGD training dynamics based on linearizing Langevin diffusion around *the noise term*. This

Submitted to Workshop on Fine-Tuning in Machine Learning (FITML@NeurIPS 2024). Do not distribute.

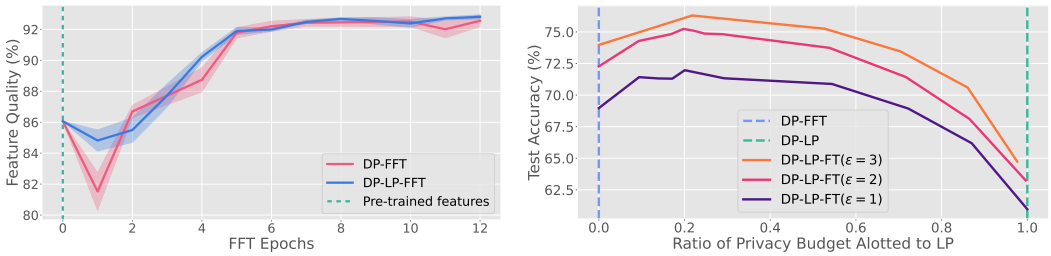


Figure 1: **Left:** Backbone feature quality evaluated by top-1 kNN accuracy on the downstream task, for ResNet-50, through public pre-training on ImageNet-1K and differentially private fine-tuning on STL-10. **Right:** Privacy budget trade-off in DP-LP-FFT, predicted in our theory, for WideResNet-16-4 on CIFAR-10 (Tang et al., 2023).

approach offers new insights into DP fine-tuning and significantly simplifies analysis by converting stochastic differential equations into ordinary differential equations (ODEs). We validate our theoretical predictions with real experiments.

Main contributions. To summarize, we make four contributions:

1. **New approximation technique:** In Section 2, we derive a first-order ODE via an asymptotic expansion of the stochastic noise in Langevin diffusion. Unlike previous methods, which linearize neural network parameters, our technique preserves the multi-layer structure of deep learning models while simplifying the analysis. This approach, commonly used in physics and control theory (Skorokhod et al., 2002), is novel in the context of private machine learning and bridges the gap between non-private neural network theory and the private regime.
2. **Understanding of feature distortion:** In Section 3, we provide a theoretical understanding of how DP fine-tuning affects feature representations. Using our approximation, we prove that, in 2-layer ReLU networks, randomly initialized linear heads distort pre-trained backbone features in the early stages of DP-FFT. Empirically Figure 1 demonstrates that feature quality evaluated on private data initially degrades during DP-FFT but later improves and surpasses pre-fine-tuning quality. Our theory also predicts that running a single epoch of DP-LP before transitioning to DP-FFT can mitigate this initial feature distortion, as shown empirically in the DP-LP-FFT curve of Figure 1 (left). This insight extends the findings of Kumar et al. (2022), who showed that LP-FFT reduces feature distortion in non-private, OOD scenarios, to in-distribution settings for both DP and non-DP cases.
3. **Theoretical convergence bounds:** In Section 4, we prove new upper and lower bounds on the training loss of DP-LP and DP-FFT, for 2-layer ReLU networks, using our approximation technique. To our knowledge, this is the first convergence analysis for DP-SGD on a non-linear neural network architecture.
4. **Mitigating feature distortion by combining fine-tuning methods:** Prior work by Tang et al. (2023) empirically showed that combining DP-LP and DP-FFT (DP-LP-FFT) can achieve better test accuracy than either method alone. In Figure 1b, we demonstrate that allocating approximately 20% of the privacy budget to DP-LP yields optimal test accuracy. In Section 5, we provide a partial theoretical explanation for this phenomenon. Specifically, our bounds suggest that DP-FFT may underperform relative to DP-LP at lower privacy budgets, while DP-LP-FFT can outperform both methods under moderate privacy budgets. These predictions are empirically verified across various architectures and benchmarks in Section 5.2.

1.1 Related Work

Similar empirical phenomena have been explored in non-private, out-of-distribution (OOD) contexts by Aghajanyan et al. (2021), Kumar et al. (2022), Trivedi et al. (2023), and Chen et al. (2024). Kumar et al. (2022) demonstrated that non-DP fine-tuning distorts pre-trained features and degrades OOD performance. Their theory, however, relies on the assumption that OOD test data exists in an orthogonal subspace to the fine-tuning training data, leaving their results unable to explain why, in many transfer learning tasks, linear-probe fine-tuning (LP-FFT) still outperforms both LP and full fine-tuning (FFT) in in-distribution (ID) settings. Our work aims to fill this research gap.

Wang et al. (2024) examined how pre-trained representations enhance DP fine-tuning through the neural collapse framework, though their focus was restricted to the final layer. Meanwhile, Tang et al. (2023) empirically observed the privacy budget trade-off for WideResNet models pre-trained on synthetic data, but without accompanying theoretical insights.

Analyses by Wang et al. (2019), Chen et al. (2020a), Ganesh et al. (2023b), and Fang et al. (2023) rely on standard convexity/non-convexity and smoothness assumptions, which abstract away the simultaneous dynamics between the backbone and linear head. Other works (Ye et al., 2023b; Wang et al., 2024) focus on linearized models, limiting their ability to capture the nuanced interactions between these components. Our explanation of representation alignment builds on the theoretical foundation laid by Min et al. (2024), which we extend to a DP context using new approximation tools.

2 Continuous modeling of differentially private fine-tuning

Notation. We denote both the deterministic and stochastic differential operators as ∂, the dot product between vectors x, y as $x^\top y$, the Euclidean norm of vector x as $\|x\|_2$, the infinity norm as $\|x\|_\infty$, the trace operator of a matrix as tr, and the ReLU activation as ϕ. For any twice differentiable function $f(x)$, we write its gradient as $\nabla_x f$ and its Hessian as $H_x f$. $\sqcup$ denotes the disjoint union. $[i] := \{1, \ldots, i\}$. We define the cosine similarity between two vectors u, v by $\cos(u, v) = \frac{u^\top v}{\|u\|_2 \|v\|_2}$. We denote r as the privacy cost estimated by Rényi divergence (Mironov, 2017).

DP-SGD Dynamics. Differential privacy (DP) is a widely-used method to evaluate privacy leakage in a database accessed through queries (Dwork & Roth, 2014). In machine learning, DP ensures that an adversary cannot confidently ascertain whether a specific training sample (or set of training samples) was in the training dataset. **D**ifferentially **P**rivate **S**tochastic **G**radient **D**escent (DP-SGD), introduced by Abadi et al. (2016), is the standard algorithm for optimizing deep neural networks while maintaining privacy guarantees.

Our fine-tuning theory is based on an analysis of DP-SGD dynamics. To study the dynamics, continuous approximations such as stochastic differential equations (e.g., Langevin diffusion) are frequently employed, though they differ from the discrete nature of real algorithms (Chourasia et al., 2021; Ye et al., 2023b). In a similar vein, Kumar et al. (2022) use gradient flow, a continuous approximation of SGD, to investigate fine-tuning dynamics in a non-private setting.

Definition 2.1 (Langevin diffusion (Ganesh et al., 2023b))**.** A Langevin diffusion is a stochastic differential equation that models the dynamics of a system under the influence of both deterministic and random forces (Lemons & Gythiel, 1997). We can define an p-dimensional Langevin diffusion to model DP-SGD as follows:

$$\partial\theta = -\nabla_\theta \mathcal{L}(\theta|f)\partial t + \sqrt{2\sigma^2}\partial Q_t, \tag{1}$$

where $\theta \in \mathbb{R}^p$ contains the neural network parameters, f is the neural network architecture, $\mathcal{L}(\cdot|f) : \mathbb{R}^p \to \mathbb{R}$ is the training loss, and $\sigma > 0$ is the noise multiplier (Abadi et al., 2016). $\{Q_t\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^m$ that models the Gaussian noise mechanism.

By Itô's lemma (Ito, 1951), the Langevin diffusion of the training loss is

$$\partial\mathcal{L} = \left[-\|\nabla_\theta \mathcal{L}(\theta|f)\|_2^2 + \sigma^2 \mathrm{tr}(H_\theta \mathcal{L})\right] \partial t + \sqrt{2\sigma^2}(\nabla_\theta \mathcal{L}(\theta|f))^\top \partial Q_t. \tag{2}$$

Ye et al. (2023b) study how the random initialization affects DP-SGD performance in linearized neural networks via Langevin diffusion. To facilitate theoretical analysis, they linearize the entire neural network using 1^{st}-order Taylor expansions at the initial parameter θ_0.

$$f(x) \approx f_{\text{lin}}(x) := f(x)\Big|_{\theta=\theta_0} + \frac{\partial f(x)}{\partial \theta}\Big|_{\theta=\theta_0} \cdot (\theta - \theta_0). \tag{3}$$

Recently, a growing body of research has employed this linearization technique to effectively explain important deep learning phenomena (Ortiz-Jimenez et al., 2021). However, linearizing the whole model removes the multi-layer interactions, making it unsuitable for our analysis.

To address this, we view the optimization trajectory of neural networks as a dynamical system, with noise in gradient updates treated as random perturbations. Applying the zeroth-order asymptotic expansion for Equation (1) at the noise multiplier σ (Freidlin et al., 2012), we approximate:

$$\partial\theta \approx \partial\tilde{\theta} = -\nabla\mathcal{L}\left(\tilde{\theta}\middle|f\right)\partial t. \tag{4}$$

This zeroth-order expansion helps circumvent the complex analysis of stochastic, non-linear equations. By substituting the approximate parameter $\tilde{\theta}$ into Equation (2), our modeling preserves the noisy behavior characteristic of DP-SGD.

3 Representation Alignment

In this section, we introduce the concept of representation alignment, present our theoretical findings, and validate them with experiments. Representation alignment refers to the process by which the classification head aligns itself with the pre-trained backbone features. During the differentially private full fine-tuning (DP-FFT) process, this alignment creates a characteristic trend in feature quality: initially, the randomly initialized linear head distorts the pre-trained features, but as it better aligns with the backbone, the distortion diminishes, and the overall quality of the backbone features improves over time.

3.1 Theory

Our goal is to understand (1) how does DP fine-tuning distort the pre-trained features in the backbone, and (2) under what conditions this distortion can be mitigated. We consider the simple binary classification setup from Min et al. (2024), which provides a clear and intuitive understanding of representation alignment. The results generalize to our experiments in Section 3.2. Specifically, we use a 2-layer fully-connected neural network with h hidden nodes and ReLU activation ϕ,

$$f(x) = v^\top g(x) = v^\top \phi(W^\top x) = \sum_{j=1}^{h} v_j \phi(w_j^\top x). \tag{5}$$

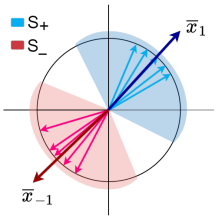


Figure 2: Visualization of Assumption 3.1.

fine-tuning on a dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$ with n inputs $x_i \in \mathbb{R}^{d_x}$, and binary labels $y_i \in \{-1, 1\}$. The objective is to minimize the training loss $\mathcal{L}(\tilde{\theta}|f) := \sum_{i=1}^n \ell(y_i, f(x_i))$, using the exponential loss $\ell(y, \hat{y}) := \exp(-y\hat{y})$. Similar results hold for logistic loss (Min et al., 2024). For simplicity, we make the following assumption.

Assumption 3.1 (Data correlation (Min et al., 2024))**.** For any pair of data $(x_i, y_i), (x_j, y_j)$, the inputs are positively/negatively correlated if the labels are the same/different.

$$\inf_{i,j\in[n]} \left[(y_1 y_2) \cdot \frac{x_1^\top x_2}{\|x_1\|_2 \|x_2\|_2} \right] := \mu > 0. \tag{6}$$

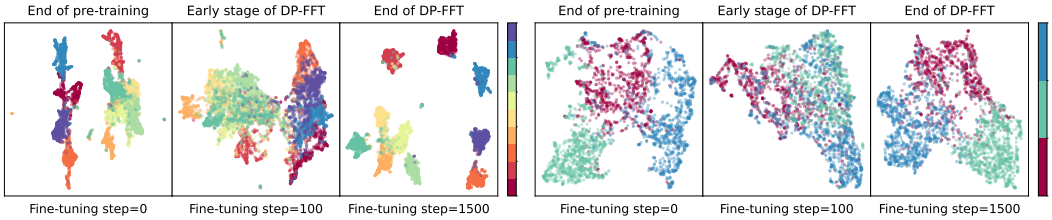
We define two cones in $\mathbb{R}^{d_x}$ that separate subspaces spanned by data points in the positive and negative classes, respectively: $S_+ = \{z \in \mathbb{R}^{d_x} : \forall i \in [n], \mathbb{I}_{x_i^\top z > 0} = \mathbb{I}_{y_i = 1}\}, S_- = \{z \in \mathbb{R}^{d_x} : \forall i \in [n], \mathbb{I}_{x_i^\top z > 0} = \mathbb{I}_{y_i = -1}\}$. Min et al. (2024) prove that $S_+ \cap S_- = \emptyset$, and $x_i \in S_{+/-}$ if $y_i = 1/-1$ (see Figure 2). We define the mean data directions of class $c \in \{-1, 1\}$ by $\bar{x}_c := \sum_{i\in[n]} x_i \cdot \mathbb{I}_{y_i = c}$.

We assume that a "clustering" behavior emerges in the pre-trained features, which allows the features to work well in transfer learning (Galanti et al., 2022). This phenomenon is well-documented in the neural collapse literature (Kothapalli, 2023), suggests that pre-trained features w_j tend to converge around the mean direction for data in class $c(j)$.

Assumption 3.2 (Collapsed neural features)**.** For each w_j in Equation (5) where $j \in [h]$ (with h denoting the dimension of the linear head), it holds that $w_j \in S_+$ or $w_j \in S_-$. We define $c(j) = 1$ if $w_j \in S_+$, and $c(j) = -1$ if $w_j \in S_-$. Thus, there is a partition $[h] = F_+ \sqcup F_-$ over the index set $[h]$, such that for each w_j,

$$\begin{cases} j \in F_+ \text{ if } w_j \in S_+, \\ j \in F_- \text{ if } w_j \in S_-. \end{cases} \tag{7}$$

Feature quality. Assumption 3.2 says that data with positive label (resp. negative) only activates the j-th neuron if $j \in F_+$ (resp. $j \in F_-$). As a result, any positive data pair $(x, y) \sim (x, y')$ activate the same set of neurons. From a contrastive learning viewpoint, it makes the representations of them



(a) STL-10 (in-distribution) (b) CIFAR-10 (out-of-distribution)

Figure 3: We pre-train (BYOL) a ResNet-50 backbone on ImageNet-1K and DP fine-tune (DP-SGD, $\epsilon = 1$) it on STL-10. We qualitatively evaluate the features in the ResNet-50 backbone by visualizing the backbone mappings (penultimate layer outputs) of data points via UMAP (McInnes et al., 2020). These results suggest that DP-FFT distorts feature quality before improving it, as predicted by Theorem 3.3.

semantically similar (Saunshi et al., 2019). Namely, when the features w_j and data inputs x_i are normalized unit vectors, the difference between representations of a positive data pair is bounded:

$$\|g(x) - g(x')\|_\infty \leq \max_{y_i = c(j) = y} \cos(w_j, x_i), \tag{8}$$

which represents the maximum cosine similarity between the features w_j and the data points.

However, FFT or DP-FFT with random initialization may reduce the feature quality.

Theorem 3.3 (Random initialization causes feature distortion)**.** *If Assumption 3.1 and Assumption 3.2 hold, and the linear head is randomly initialized by $v_0 \sim \mathcal{N}(0, \beta I_{h\times h})$, then with probability $1 - 2^{-h}$, $\exists j \in [h], \Delta t > 0$ such that during the time interval $(0, \Delta t)$, DP-FFT distorts w_j reducing its alignment with the data cluster. The cosine similarity between w_j and the data cluster mean $\bar{x}_{c(j)}$ decreases monotonically:*

$$\frac{\partial}{\partial t} \cos\left(w_j, \bar{x}_{c(j)}\right)\Big|_t < 0, \quad \forall t \in (0, \Delta t) \tag{9}$$

For a pre-trained w_j that aligns with $c(j)$-labeled data, DP-FFT (as modeled by Equation (4)) makes it deviate from $\bar{x}_{c(j)}$, the mean direction of those data. w_j is optimal when $\cos(w_j, \bar{x}_{c(j)}) = 1$. This result holds for both DP and non-DP settings and explains the potential feature distortion observed in in-distribution and non-private settings, such as those studied by Kumar et al. (2022)). The stochastic analysis of non-smooth loss, activation, cosine similarity functions is challenging without our approximation.

Next, we show that running (DP-)LP before (DP-)FFT could mitigate feature distortion.

Theorem 3.4 (DP-LP first mitigates feature distortion)**.** *Suppose Assumption 3.1 and Assumption 3.2 hold, and the linear head is randomly initialized by $v_0 \sim \mathcal{N}(0, \beta I_{h\times h})$. There exists $\Delta t > 0$ such that after running DP-LP for time Δt, switching to full fine-tuning ensures that DP-FFT does not distort the pre-trained features. Specifically, $\cos(w_j, \bar{x}_{c(j)})$ is non-decreasing for all $j \in [h]$:*

$$\frac{\partial}{\partial t} \cos\left(w_j, \bar{x}_{c(j)}\right)\Big|_t \geq 0, \quad \forall t \in (0, \Delta t) \tag{10}$$

3.2 Experiments

In this section, we show empirical evidence supporting Theorems 3.3 and 3.4.

Pre-training and Model. We pre-train Vision Transformers (ViT) and ResNet-50 backbones on ImageNet-1K using Self-Supervised Learning methods, including BYOL (Grill et al., 2020) and MoCo v2 (Chen et al., 2020b), as well as distillation methods (Touvron et al., 2021). Then we fine-tune the backbone with a linear classification head on CIFAR-10 and STL-10 using DP-SGD.

Experiment protocols. We conduct public pre-training for 100 epochs with a batch size of 256. Following this, we implement DP-SGD using the pre-trained weights and a randomly initialized linear head for 30 epochs. Each DP fine-tuning process is repeated with 5 random seeds and a batch size of 1000. We evaluate the backbone features on both the pre-training and fine-tuning datasets, measuring feature quality through top-1 kNN accuracy (Chen et al., 2023).

Private fine-tuning initially distorts features. Figure 3 qualitatively visualizes the effect of DP-FFT on feature quality with respect to the private test data. We pre-train (BYOL) a ResNet-50

backbone on ImageNet-1K and DP fine-tune (DP-SGD, $\epsilon = 1$) it on STL-10. We qualitatively assess the features of the private test data within the ResNet-50 backbone by visualizing the backbone mappings (outputs from the penultimate layer) of data points using UMAP (McInnes et al., 2020). For simplicity, we only plot 3 classes in CIFAR-10.

Figure 3 indicates that during the initial phases of DP-FFT, the randomly initialized linear head interferes with the pre-trained features in the backbone network, leading to a degradation in feature quality on both the pre-training and fine-tuning datasets. This observation validates Theorem 3.3. Concurrently, the linear head begins adapting to these pre-trained features, a process we refer to as "**representation alignment**." As this alignment progresses, the backbone starts to regain a portion of its original feature quality, which had been degraded by DP noise and shifts in data distribution.

Linear probing mitigates feature distortion. To illustrate the benefits of linear probing, we first run DP-LP for 1 epoch before transitioning to DP-FFT for the remaining epochs. In the initial steps of DP-FFT, the feature distortion is significantly weaker (Figure 1a) if we first run DP-LP. This supports the claim of Theorem 3.4. We also evaluate features on the pre-training domain (see Figure 5).

4 DP Fine-tuning Convergence Rates

Section 3 showed that DP-LP-FFT can mitigate feature distortion. A natural question is, for a fixed privacy budget, how do DP-LP and DP-FFT affect the convergence of fine-tuning loss function? We study this question under our zeroth-order approximation of Langevin diffusion (Section 4.1).

Privacy guarantees We first provide privacy guarantees of Langevin diffusion by bounding the Rényi divergence of its trajectory distributions on neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ (Mironov, 2017). Ganesh et al. (2023b) and Ye et al. (2023b) both show that the Rényi divergence linearly increases over time. We use this guarantee for all fine-tuning variants.

Theorem 4.1 (Rényi privacy guarantee (Ganesh et al., 2023b)). *Suppose we initialize a pair of neural network parameters θ, θ' by some i.i.d. distributions Θ_0, Θ_0'. We fine-tune θ, θ' respectively on neighboring datasets $\mathcal{D}, \mathcal{D}'$ via Langevin diffusion. Denote the distribution of the trajectory of θ by $\Theta_{[0,T]}$ over $[0, T]$. Similarly, denote the trajectory distribution of θ' by $\Theta'_{[0,T]}$. Then for any $\alpha \geq 1$, the Rényi divergence R_α is bounded linearly in time,*

$$r := R_\alpha(\Theta_{[0,T]} \| \Theta'_{[0,T]}) = O\left(\frac{\alpha \Delta_g T}{\sigma^2}\right) \tag{11}$$

where σ is the noise multiplier, and $\Delta_g \geq \|\nabla \mathcal{L}(\theta; \mathcal{D}) - \nabla \mathcal{L}(\theta; \mathcal{D}')\|$ is the upper bound of gradient difference between neighboring datasets. Thus, for any $\delta \in (0, 1)$, the Langevin diffusion satisfies

$$\left(\frac{\alpha \Delta_g T}{4\sigma^2} + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right) - \textit{differential privacy.} \tag{12}$$

4.1 Convergence Rates under the Zeroth-order Approximation

We follow the approximation scheme of Equation (4) to obtain the following convergence results for 2-layer ReLU neural networks. To our knowledge, these are the first convergence guarantees (approximate or not) for DP-SGD under a nonconvex, nonsmooth objective.

Theorem 4.2 (Approximate DP-LP loss convergence). *If Assumption 3.1 and Assumption 3.2 hold at $t = 0$, then we can bound the loss after running DP-LP for $t = T$:*

$$\frac{1}{\frac{1}{\mathcal{L}_c(0)} e^{-B_1 T} + \frac{A_1}{B_1}(1 - e^{-B_1 T})} \leq \mathcal{L}_c(T) \leq \frac{1}{\frac{1}{\mathcal{L}_c(0)} e^{-B_2 T} + \frac{A_2}{B_2}(1 - e^{-B_2 T})} \tag{13}$$

where $\mathcal{L}_c(t)$ denotes the training loss of data points labeled $c \in \{-1, 1\}$, $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, and

$$\begin{cases} A_1 = \sum_{w_j \in S_c} \left[\max_{y_i = c} w_j^\top x_i\right]^2 \\ B_1 = \frac{1}{2}\sigma^2 \left\{\sum_{y_i = c} \|\mathrm{relu}(W^\top x_i)\|_2^{-2}\right\}^{-1} \\ A_2 = \sum_{w_j \in S_c} \left[\min_{y_i = c} w_j^\top x_i\right]^2 \\ B_2 = \frac{1}{2}\sigma^2 \left\{\sum_{y_i = c} \|\mathrm{relu}(W^\top x_i)\|_2^{4}\right\}^{1/2} \end{cases} \tag{14}$$

are constants for DP-LP.

When we take $n = h = 2, y_1 = -y_2, w_1 = x_1 = -w_2 = -x_2$, the upper and lower bounds are equal and we achieve a tight bound on the DP-LP loss.

Theorem 4.3 (Approximate DP-FFT loss convergence)**.** *For simplicity, we assume that $\|x_i\|_2 = R$ for all $i \in [n]$. If Assumption 3.1 and Assumption 3.2 hold, and we consider a balanced initialization $\|W\|_F^2 = \|v_0\|_2^2$ (Min et al., 2023) at $t = 0$, then*

(i) we lower bound the loss after running DP-FFT for $T > 0$:

$$\mathcal{L}_c(T) \geq \frac{1}{\frac{1}{\mathcal{L}_c(0)} e^{(1-\exp(\lambda_c T))A_l C_l/\lambda_c} + \frac{B_l}{C_l}\left[1 - e^{(1-\exp(\lambda_c T))A_l C_l/\lambda_c}\right]} \tag{15}$$

where we define $A_l = \|W_0\|_F^2, B_l = 2R^2, C_l = \frac{R^2\sigma^2(1+\mu^2)}{2}$ and $\lambda_c = 2R\mathcal{L}_c(0)$.

(ii) we upper bound the loss after running DP-FFT for $T > 0$:

$$\mathcal{L}_c(T) \leq \frac{1}{\frac{B_u}{C_u}(1 - e^{-A_c C_u T}) + \frac{1}{\mathcal{L}_c(0)} e^{-A_c C_u T}} \tag{16}$$

where we define $A_c = \sum_{w_j \in S_c}\left[v_{j,t=0}^2 + \|w_j\|_2^2\right], B_u = R^2\mu^2$ and $C_u = \frac{1}{2}R^2\sigma^2$.

5 Budget Allocation between DP-LP and DP-FFT

Consider the DP-LP-FFT fine-tuning strategy, which first applies DP-LP for some portion r of the privacy budget (i.e. for some number of training iterations), then uses the remaining privacy budget for DP-FFT. In this section, we ask: given a fixed privacy budget, how should we allocate it across DP-LP and DP-FFT? Our results, both theoretical and empirical, suggest that at low total privacy budget, one should allocate more of the total privacy budget to DP-LP.

5.1 Results under Zeroth-order Approximation

We first show how to allocate privacy budget to avoid the feature distortion analyzed in Section 3, under our zeroth-order approximation.

Theorem 5.1 (Estimated privacy budget allocated to DP-LP)**.** *If Assumption 3.1 and Assumption 3.2 hold at $t = 0$, then for any $\rho \in (0, 1)$, with probability $(1-\rho)^h$, we can avoid feature distortion by spending*

$$r \propto \sigma^4\sqrt{\ln(2/\rho)} \tag{17}$$

amount r of privacy budget on DP-LP, where σ is the noise multiplier. That is, we ensure that $\forall j \in [h]$, and any $t > 0$ after DP-LP,

$$\frac{\partial}{\partial t}\cos\left(w_j, \bar{x}_{c(j)}\right)\Big|_t \geq 0 \tag{18}$$

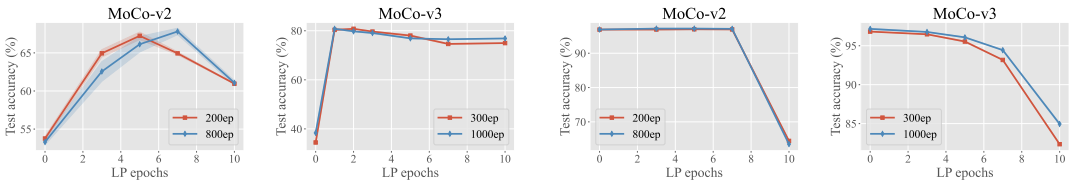
According to Theorem 5.1, we should spend greater ratio of privacy budget on DP-LP if the total privacy budget is smaller.

5.2 Experiments

To illustrate the privacy budget trade-off, we empirically evaluate the benefits of DP-LP-FFT on real data and architectures.

DP-LP-FFT outperforms other fine-tuning methods: Pre-training on synthetic data. We follow the setup in Tang et al. (2023) and generate utility curves for $\epsilon = 1, 2, 3$ (Figure 1b). We pre-train WideResNet with synthetic images generated from StyleGAN-oriented (Baradad et al., 2021) , and fine-tune it with DP-SGD on CIFAR-10. The x-axis sweeps the fraction of privacy budget allocated to DP-LP, and the remaining budget is used for DP-FFT. We find that at various privacy levels, DP-LP-FFT gives a clear advantage over either DP-FFT or DP-LP alone.

Figure 1b presents a different trend from our theoretical prediction, where we expect the optimal budget ratio for DP-LP to increase as the privacy noise grows. A possible intuitive explanation is



(a) Private utility curves ($\sigma = 0.3$) (b) Non-private utility curves

Figure 4: Utility curves for pretraining on ImageNet-1K and fine-tuning on CIFAR-10 over ResNet-50, with pretrained features from MoCo-v2 and MoCo-v3 (Chen et al., 2020b; Chen* et al., 2021). We compare the performance from pre-trained weights of different pre-training epochs (200/800 epochs for MoCo-v2, 300/1k epochs for MoCo-v3). The x-axis sweeps the number of LP epochs from 0 to 10; the remaining epochs (out of 10) use FFT.

that, in the Figure 1b experiments, the pre-training data is synthetic, making it 'distant' from the CIFAR-10 fine-tuning data distribution. This divergence may violate our assumption that the pre-trained weights w_j are well-aligned with the fine-tuning data x_i.

DP-LP-FFT outperforms other fine-tuning methods: Pre-training on ImageNet-1K. Figure 4 illustrates the utility curves on ResNet-50 for $\sigma = 0, 0.3$. To demonstrate utility curves for DP-LP-FFT, we vary the number of epochs of linear probing from $e_{LP} = 0$ to $e_{LP} = 10$; all remaining epochs (out of 10 total) are allocated to full fine-tuning, i.e., $e_{FFT} = 10 - e_{LP}$. Note that full fine-tuning corresponds to $e_{LP} = 0$ (the leftmost point of our subplots), and linear probing corresponds to $e_{LP} = 10$. We observe that for non-private optimization (Figure 4b), full fine-tuning achieves the highest test accuracy. However, for DP-SGD (Figure 4a), linear probing outperforms full fine-tuning, and DP-LP-FFT outperforms both DP-LP and DP-FFT.

Model	ResNet_{18}			MobileNet_{v3}			$\text{Transformer}_{\texttt{DeiT}}$		
ϵ	∞	1.29	0.57	∞	1.29	0.57	∞	1.29	0.26
LP	$68.54_{0.02}$	$67.90_{0.12}$	$66.60_{0.04}$	$71.12_{0.31}$	$69.54_{0.08}$	$67.32_{0.03}$	$95.74_{0.04}$	$93.61_{0.08}$	$94.21_{0.08}$
LP-FFT	$72.66_{0.12}$	$68.65_{0.08}$	$59.79_{1.03}$	$71.30_{0.11}$	$71.18_{0.06}$	$66.94_{0.08}$	$96.82_{0.08}$	$93.66_{0.15}$	$93.62_{0.05}$
FFT	$73.69_{0.03}$	$59.79_{1.03}$	$53.82_{0.37}$	$77.02_{0.31}$	$63.06_{0.05}$	$45.12_{0.07}$	$96.17_{0.08}$	$90.31_{0.53}$	$84.19_{0.82}$

Table 1: Test accuracies of DP-LP, DP-LP-FFT, and DP-FFT on various architectures.

Comparing DP fine-tuning methods. As suggested by Theorem 5.1, as the noise scale σ increases, the best fine-tuning strategy changes from DP-FFT (small σ, low privacy regime) to DP-LP-FFT, to DP-LP (large σ, high privacy regime). To qualitatively test this prediction, we sweep over different noise scales σ and fix other hyperparameters in each benchmark and model architecture. We sort the rows by the number of parameters of each model and the noise scale in an ascending order. For each experiment setting, we report average test accuracies with standard errors. As expected, among the three fine-tuning methods (Table 1), DP-FFT almost always does the best under small noise scales (including the non-private setting where $\sigma = 0$), DP-LP-FFT does the best under moderate noise scales, and DP-LP does the best under large noise scales. The close non-DP (ϵ) performance of FFT and LP-FFT on transformer architectures is consistent with previous observations in Kumar et al. (2022, Table 1). We also provide results with LoRA (see Table 2).

6 Conclusion and Discussion

We characterize the training dynamics of DP fine-tuning under a simplified theoretic setup (2-layer neural networks, separable datasets with -1/1 labels) using a Langevin diffusion-based approximation of DP-SGD, with an asymptotic expansion of random perturbations in dynamical systems as an approximation for Langevin diffusion. Our theory identifies and explains the phenomenon of representation distortion and alignment during DP fine-tuning, which we confirm empirically. Our work takes a step towards understanding how different private fine-tuning strategies can be mixed to improve performance, which could be useful for designing or mixing other strategies, such as memory-efficient zeroth-order optimization with differential privacy (Zhang et al., 2024a).

Limitations and open questions There are several open questions we cannot cover in this work, such as generalizing our results to multi-layer neural networks with our approximation technique, the effect of other loss functions on the fine-tuning dynamics, and loss lower bounds for DP-LP/FFT without the zeroth-order approximation. Moreover, it is unclear how to apply our theory to other fine-tuning methods like LoRA (Hu et al., 2022), as well as generative models for which neural collapse does not happen. Understanding whether the zeroth-order approximation can facilitate analysis in these settings is an interesting and important question for future work.

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*, CCS '16, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OQ08SN70M1V.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/arora18a.html.

Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In *Advances in Neural Information Processing Systems*, 2021.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pp. 464–473, USA, 2014. IEEE Computer Society. ISBN 9781479965175. doi: 10.1109/FOCS.2014.56. URL https://doi.org/10.1109/FOCS.2014.56.

Annie S Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. Project and probe: Sample-efficient adaptation by interpolating orthogonal features. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=f6CBQYxXvr.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: a geometric perspective. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Yubei Chen, Zeyu Yun, Yi Ma, Bruno Olshausen, and Yann LeCun. Minimalistic unsupervised representation learning with the sparse manifold transform. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nN_nBVKAhhD.

Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14771–14781. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7c6c1a7bfde175bed616b39247ccace1-Paper.pdf.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/coates11a.html.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking High-Accuracy Differentially Private Image Classification through Scale. *arXiv preprint arXiv:2204.13650*, 2022.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/fe131d7f5a6b38b23cc967316c13dae2-Paper.pdf.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Huang Fang, Xiaoyun Li, Chenglin Fan, and Ping Li. Improved convergence of differential private SGD with gradient clipping. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FRLswckPXQ5.

M.I. Freidlin, J. Szücs, and A.D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer, 2012. ISBN 9783642258473. URL http://books.google.de/books?id=p8LFMILAiMEC.

Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkpoTaxA-.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SwIp410B6aQ.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pp. 10611–10627. PMLR, 2023a.

Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Universality of langevin diffusion for private optimization, with applications to sampling from rashomon sets. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1730–1773. PMLR, 12–15 Jul 2023b. URL https://proceedings.mlr.press/v195/ganesh23a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Shlomo Hoory, Amir Feder, Avichai Tendler, Alon Cohen, Sofia Erell, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and Yossi Matias. Learning and evaluating a differentially private pre-trained language model. In Oluwaseyi Feyisetan, Sepideh Ghanavati, Shervin Malmasi, and Patricia Thaine (eds.), *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pp. 21–29, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.privatenlp-1.3. URL https://aclanthology.org/2021.privatenlp-1.3.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kiyosi Ito. On stochastic differential equations. *Mem. Amer. Math. Soc.*, 1951(4):51, 1951. ISSN 0065-9266.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QTXocpAP9p.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Don S. Lemons and Anthony Gythiel. Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]. *American Journal of Physics*, 65(11):1079–1081, 11 1997. ISSN 0002-9505. doi: 10.1119/1.18725. URL https://doi.org/10.1119/1.18725.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Hancheng Min, Rene Vidal, and Enrique Mallada. On the convergence of gradient flow on multi-layer linear models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24850–24887. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/min23d.html.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer reLU networks with small initialization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QibPzdVrRu.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

Guillermo Ortiz-Jimenez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. What can linearized neural networks actually say about generalization? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8998–9010. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4b5deb9a14d66ab0acc3b8a2360cde7c-Paper.pdf.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Anatoli V. Skorokhod, Frank C. Hoppensteadt, and Habib Salehi. *Random Perturbation Methods with Applications in Science and Engineering*. Applied mathematical sciences (Springer-Verlag New York Inc.) ; v. 150. Springer, 2002. ISBN 0387954279. doi: 10.1115/1.1579453.

Xinyu Tang, Ashwinee Panda, Vikash Sehwag, and Prateek Mittal. Differentially private image classification by learning priors from random processes. *CoRR*, abs/2306.06076, 2023. URL https://doi.org/10.48550/arXiv.2306.06076.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on*

Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Puja Trivedi, Danai Koutra, and Jayaraman J. Thiagarajan. A closer look at model adaptation using feature distortion and simplicity bias. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wkg_b4-IwTZ.

Chendi Wang, Yuqing Zhu, Weijie J Su, and Yu-Xiang Wang. Neural collapse meets differential privacy: Curious behaviors of NoisyGD with near-perfect representation learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52334–52360. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24cu.html.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6526–6535. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/wang19c.html.

Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private sgd with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56:306–336, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.09.001. URL https://www.sciencedirect.com/science/article/pii/S1063520321000841.

Jiayuan Ye, Zhenyu Zhu, Fanghui Liu, Reza Shokri, and Volkan Cevher. Initialization matters: Privacy-utility analysis of overparameterized neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=IKvxmnHjkL.

Jiayuan Ye, Zhenyu Zhu, Fanghui Liu, Reza Shokri, and Volkan Cevher. Initialization matters: Privacy-utility analysis of overparameterized neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5419–5446. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1165af8b913fb836c6280b42d6e0084f-Paper-Conference.pdf.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference 2016*, York, France, January 2016. British Machine Vision Association. doi: 10.5244/C.30.87. URL https://enpc.hal.science/hal-01832503.

Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12135–12144, 2022. doi: 10.1109/CVPR52688.2022.01183.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. DPZero: private fine-tuning of language models without backpropagation. In *Forty-first International Conference on Machine Learning*, 2024a.

Xinwei Zhang, Zhiqi Bu, Steven Wu, and Mingyi Hong. Differentially private SGD without clipping bias: An error-feedback approach. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=uFbWHyTlPn.

A Additional experiment results

In this section, we provide more experiment results and detailed configurations.

Evaluations back in the pre-training distribution (Figure 5). We also evaluate the feature quality on ImageNet1-K, the pre-training dataset. The representation alignment for the pre-training domain is different: once a proper alignment is achieved, the backbone gradually recovers a portion of its original feature quality, which had been compromised due to DP noise and distribution-shift.

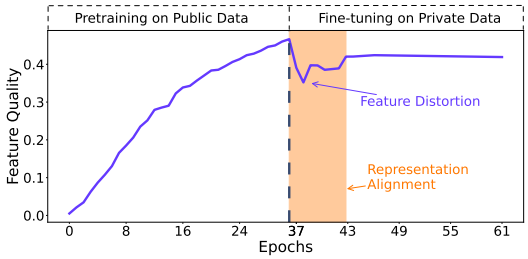


Figure 5: Backbone feature quality evaluated by average top-1 kNN accuracy on the pre-training dataset, for ResNet-50, through public pre-training on ImageNet-1K and differentially private fine-tuning on STL-10.

More experiments on parameter-efficient fine-tuning (PEFT) methods. We conduct experiments with another fine-tuning trick: differentially private LoRA (Hu et al., 2022). We run experiments on the Mini-DeiT-Ti architecture, where we use LoRA instead of full fine-tuning. In these experiments (Table 2), our batch size is 1000, and our LoRA rank is set to 8. We observe the same trend as what we saw for full fine-tuning; namely, as we increase the noise scale (i.e., as we reduce epsilon, giving a stronger privacy guarantee), it becomes more beneficial to use LP-LoRA or even just LP.

$\text{Transformer}_{\texttt{DeiT}}$					
ϵ	∞	12.28	1.29	0.57	0.26
LP	$95.81_{0.05}$	$95.55_{0.05}$	$94.80_{0.06}$	$94.21_{0.08}$	$92.48_{0.27}$
LP-LoRA	$96.2_{0.05}$	$95.90_{0.03}$	$94.81_{0.08}$	$94.18_{0.05}$	$91.99_{0.19}$
LoRA	$96.26_{0.05}$	$95.50_{0.06}$	$94.76_{0.08}$	$93.05_{0.09}$	$91.28_{0.43}$

Table 2: Test accuracies of LP, LP-LoRA, LoRA on $\text{Transformer}_{\texttt{DeiT}}$.

Experiment setup in Table 1. We use batch size 1000 and and sweep over a range of learning rates $\{9, 5, 1, 0.5, 0.2, 0.15, 0.1, 0.05, 0.025\}$.

Summary of experiment configurations. We run experiments on five deep learning models and four transfer learning benchmarks to verify if our theoretical prediction, the existence of concave utility curves, generalizes to deep neural networks and real datasets. Each experimental setting comprises: (1) a model architecture, (2) a (larger) dataset for public pretraining, and (3) a (smaller) dataset as the private data for fine-tuning. The benchmarks we use are:

- ImageNet-1K→CIFAR-10. ImageNet-1K is a large-scale dataset. We initialize pretrained features of ResNet-50 from MoCo-v2 Chen et al. (2020b) and MoCo-v3 Chen* et al. (2021), trained on ImageNet-1K Russakovsky et al. (2015) without privacy. We then privately fine-tune the ResNet-50 on CIFAR-10.
- ImageNet-1K→STL-10. We pretrain a DeiT model on ImageNet then pretrain a Mini-DeiT-Ti model with weight distillation from the DeiT model Touvron et al. (2021); Zhang et al. (2022). After that, we privately fine-tune the Mini-DeiT-Ti model on STL-10 Coates et al. (2011) for 20 epochs.
- CIFAR-10→STL-10. We pretrain the feature extractor on CIFAR-10 Krizhevsky (2009) using stochastic gradient descent without privacy mechanisms. Then we finetune the pre-trained features and a randomly initialized linear head on STL-10. This benchmark has been studied in the context of domain adaptation French et al. (2018); Kumar et al. (2022).

The training subset of STL-10 only contains 500 images. To align with the small scale fine-tuning data, we run the experiments with smaller and data-efficient models: MobileNet-v3 and ResNet-18.

- RandP→CIFAR-10. To reproduce the results of Tang et al. (2023) and verify the general existence of concave utility curves, we also consider a slightly non-standard pretraining protocol. We pretrain a wide residual network (WRN) Zagoruyko & Komodakis (2016) on synthetic images generated by random diffusion processes. We follow the settings in Tang et al. (2023).

We employ early stopping, and select the optimal learning rate based on the accuracy of the in-distribution validation.

B Technical results

Lemma B.1 (Holder's inequality for sums)**.** *For a sequence $x = [x_i]_{i=1}^n$ of positive real numbers and $p > 0$, define $\|x\|_p := (\sum_{i=1}^n x_i^p)^{1/p}$. Then for any pair of positive real numbers $p > 0, q > 0$ with $\frac{1}{p} + \frac{1}{q} = 1$, and any pair of sequence of positive real numbers x and y,*

$$\|xy\|_1 \le \|x\|_p \|y\|_q$$

Lemma B.2 (Reverse Holder's inequality for sums)**.** *For a sequence $x = [x_i]_{i=1}^n$ of positive real numbers and $p > 0$, define $\|x\|_p := (\sum_{i=1}^n x_i^p)^{1/p}$. Then for any pair of positive real numbers $p > 0, q > 0$ with $\frac{1}{p} - \frac{1}{q} = 1$, and any pair of sequence of positive real numbers x and y,*

$$\|xy\|_1 \ge \|x\|_p \|y\|_{-q}$$

Lemma B.3 (Reverse QM-AM inequality for sums)**.** *For a sequence $x = [x_i]_{i=1}^n$ of positive real numbers,*

$$\left(\sum_{i=1}^n x_i\right)^2 \ge \sum_{i=1}^n x_i^2$$

Lemma B.4 (μ-coherent data conic hull (Min et al., 2024, Lemma 5))**.** *Define a conic hull $K := \mathcal{CH}(\{y_i x_i : i \in [n]\}) = \{\sum_{i=1}^n a_i y_i x_i : \forall a_i \ge 0, i \in [n]\}$. If Assumption 3.1 holds, i.e. the dataset is separable, then K is μ-coherent:*

$$\forall z_1, z_2 \in K \backslash \{0\}, \quad \cos(z_1, z_2) \ge \mu$$

Corollary B.5 (Orthogonally separable $\implies$ linearly separable (Min et al., 2024))**.** *If Assumption 3.1 holds, then $\exists \gamma > 0$ and $z \in \mathbb{S}^{D-1}$ such that*

$$\forall i \in [n], \quad y_i \langle z, x_i \rangle \ge \gamma$$

Proof of Corollary B.5. We prove the existence statement by picking a valid pair of z, γ. Take $z := \frac{y_1 x_1}{\|x_1\|_2}$. Then $\forall i \in [n]$,

$$\begin{aligned} y_i \langle z, x_i \rangle =& \|x_i\|_2 \cos(y_1 x_1, y_i x_i) \\ & //\text{by Lemma B.4} \\ \ge& \|x_i\|_2 \mu \\ \ge& \mu \cdot \min_{i \in [n]} \|x_i\|_2 \end{aligned}$$

Therefore $\gamma = \mu \cdot \min_{i \in [n]} \|x_i\|_2$. □

C Appendix: Representation alignment

C.1 Theory

The Langevin diffusion of w_j on a n-sized data cluster ($j \in [h]$) is

$$\dot{w_j} = \sum_{i=1}^n y_i \exp(-y_i f(x_i; W, v)) v_j \text{relu}'(w_j^\top x_i) x_i + \sigma \partial Q_t, \tag{19}$$

where Q_t is a vector containing D independent 1-dimensional Brownian motion.

The Langevin diffusion of v on a n-sized data cluster is

$$\dot{v} = \sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W, v))\text{relu}(W^\top x_i) + \sigma \partial Q_t,$$

where Q_t is a vector containing h independent 1-dimensional Brownian motion.

We rewrite the Langevin diffusion by asymptotic expansion (Freidlin et al., 2012, Equation 2.1, Chapter 2.2),

$$\begin{cases} v_j \approx v_j^{(0)} + \sigma v_j^{(1)} + \cdots \\ w_j \approx w_j^{(0)} + \sigma w_j^{(1)} + \cdots, \end{cases} \tag{20}$$

i.e. we expand the Langevin diffusion as a linear combination of the original gradient flow and a linear stochastic diffusion.

$$\begin{cases} \dot{v}_j^{(0)} = \sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\text{relu}((w_j^{(0)})^\top x_i) \\ \dot{w}_j^{(0)} = \sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \text{relu}'((w_j^{(0)})^\top x_i) x_i. \end{cases} \tag{21}$$

Lemma C.1 (Zeroth order invariance of locally linearized LD)**.** *If we rewrite the Langevin diffusion by asymptotic expansion (Freidlin et al., 2012, Equation 2.1, Chapter 2.2),*

$$\begin{cases} v_j \approx v_j^{(0)} + \sigma v_j^{(1)} \\ w_j \approx w_j^{(0)} + \sigma w_j^{(1)}. \end{cases}$$

then the layer invariance still holds for zeroth order approximation

$$\frac{d}{dt}[(v_j^{(0)})^2 - \|w_j^{(0)}\|_2^2] = 0. \tag{22}$$

This result is similar to the imbalance matrix in gradient flow (Arora et al., 2018; Du et al., 2018; Min et al., 2023).

We are ready to prove Theorem 3.3.

Proof of Theorem 3.3. The explicit expression of the cosine value is

$$\cos(w_j, \bar{x}_{c(j)}) = \frac{w_j^\top \bar{x}_{c(j)}}{\|w_j\|_2 \|\bar{x}_{c(j)}\|_2} \tag{23}$$

Without loss of generality, let $\|\bar{x}_{c(j)}\|_2 = 1$. To show that the cosine value decreases with high probability, we only need to prove that the derivative of $\frac{(w_j^\top \bar{x}_{c(j)})^2}{\|w_j\|_2^2}$ is negative at $t = 0$ with high probability. The explicit derivative expression is

$$\begin{aligned} \frac{\partial}{\partial t}\cos(w_j, \bar{x}_{c(j)}) =& \frac{2(w_j^\top \bar{x}_{c(j)})}{\|w_j\|_2^2}\left[\|w_j\|_2^2 \bar{x}_{c(j)}^\top \frac{\partial w_j}{\partial t} - \bar{x}_{c(j)}^\top w_j w_j^\top \frac{\partial w_j}{\partial t}\right] && (24) \\ =& \frac{2(w_j^\top \bar{x}_{c(j)})}{\|w_j\|_2^2}\left[\|w_j\|_2^2 \bar{x}_{c(j)} - (\bar{x}_{c(j)}^\top w_j) w_j\right]^\top \frac{\partial w_j}{\partial t} && (25) \\ & //\text{by Assumption 3.2} && (26) \\ \text{sign}\left(\frac{\partial}{\partial t}\cos(w_j, \bar{x}_{c(j)})\right) =& \text{sign}\left(\left[\|w_j\|_2^2 \bar{x}_{c(j)} - (\bar{x}_{c(j)}^\top w_j) w_j\right]^\top \frac{\partial w_j}{\partial t}\right) && (27) \\ =& \text{sign}\left(v_j(\|w_j\|_2^2 - (\bar{x}_{c(j)}^\top w_j)^2)\right) && (28) \\ =& \text{sign}(v_j) && (29) \end{aligned}$$

Since we initialize $v \sim \mathcal{N}(0, \beta I_{h\times h})$, with probability $1 - 2^{-h}$, there exists j such that $v_j < 0$ at $t = 0 \implies \frac{\partial}{\partial t}\cos(w_j, \bar{x}_{c(j)}) < 0$ at $t = 0$. By the continuity of the approximated Langevin diffusion, there exists $\Delta t > 0$ such that for any $t \in (0, \Delta t)$,

$$\frac{\partial}{\partial t}\cos(w_j, \bar{x}_{c(j)}) < 0. \tag{30}$$

□

Proof of Theorem 3.4. In the proof of Theorem 3.3, we show that for $w_j \in S_c, c \in \{-1, 1\}$,

$$\operatorname{sign}\left(\frac{\partial}{\partial t}\cos(w_j, \bar{x}_{c(j)})\right) = \operatorname{sign}(v_j) \cdot \operatorname{sign}(c) \tag{31}$$

To mitigate the feature distortion after some time index Δt, we only need $c \cdot v_j > 0$. For DP-LP, every $\frac{\partial}{\partial t} v_j$ increases/decreases if $c = 1/-1$. Therefore, for any initialization, there exists Δt such that $\operatorname{sign}(v_j) = \operatorname{sign}(c)$ after time index Δt. If we switch to DP-FFT after Δt, $\frac{\partial}{\partial t}\cos(w_j, \bar{x}_{c(j)}) > 0$ for any $j \in [h]$. Thus $\cos(w_j, \bar{x}_{c(j)})$ is non-decreasing in DP-FFT. □

D Approximate convergence of DP-LP and DP-FFT

D.1 Approximate DP-LP convergence

We add some extra notations for the following proofs:

- Positive data subset $\mathcal{I}_+ := \{i \in [n] : y_i > 0\}$
- Negative data subset $\mathcal{I}_- := \{i \in [n] : y_i < 0\}$
- Positive head cluster $\mathcal{V}_+(t) := \{j \in [h] : \operatorname{sign}(v_j(t)) > 0\}$
- Negative head cluster $\mathcal{V}_-(t) := \{j \in [h] : \operatorname{sign}(v_j(t)) < 0\}$
- Index function $\mathscr{I} : \mathbb{R}^D \to \{\mathcal{I}_+, \mathcal{I}_-\}$ maps feature vector to its cluster

$$\mathscr{I}(w) = \begin{cases} \mathcal{I}_+ & w \in S_+ \\ \mathcal{I}_- & w \in S_- \\ \emptyset & \text{otherwise} \end{cases}$$

We first derive the upper bound for approximate DP-LP.

Upper bound proof of Theorem 4.2. We construct a lower bound of the drift terms in the zeroth order approximation

$$\|\nabla_v \mathcal{L}^{(0)}\|_2^2 = \sum_{j=1}^{h}\left(\sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\operatorname{relu}((w_j^{(0)})^\top x_i)\right)^2 \tag{32}$$

$$= \sum_{j=1}^{h}\left(\sum_{i\in\mathscr{I}(w_j^{(0)})} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\operatorname{relu}((w_j^{(0)})^\top x_i)\right)^2 \tag{33}$$

$$\geq \sum_{j=1}^{h}\left[\min_{i\in\mathscr{I}(w_j^{(0)})} \operatorname{relu}((w_j^{(0)})^\top x_i)\right]^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\right)^2 \tag{34}$$

$$= \sum_{j=1}^{h}\left[\min_{i\in\mathscr{I}(w_j^{(0)})} \operatorname{relu}((w_j^{(0)})^\top x_i)\right]^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\right)^2 \tag{35}$$

$$= \sum_{j\in\mathcal{V}_+}\left[\min_{i\in\mathcal{I}_+} \operatorname{relu}((w_j^{(0)})^\top x_i)\right]^2 (\mathcal{L}_+^{(0)})^2 + \sum_{j\in\mathcal{V}_-}\left[\min_{i\in\mathcal{I}_-} \operatorname{relu}((w_j^{(0)})^\top x_i)\right]^2 (\mathcal{L}_+^{(0)})^2 \tag{36}$$

$$\geq \min\left\{\sum_{j\in\mathcal{V}_+}\left[\min_{i\in\mathcal{I}_+} \operatorname{relu}((w_j^{(0)})^\top x_i)\right]^2, \sum_{j\in\mathcal{V}_-}\left[\min_{i\in\mathcal{I}_-} \operatorname{relu}((w_j^{(0)})^\top x_i)\right]^2\right\}\left[(\mathcal{L}_+^{(0)})^2 + (\mathcal{L}_-^{(0)})^2\right] \tag{37}$$

$$\geq \frac{1}{2}\min\left\{\sum_{j\in\mathcal{V}_+}\left[\min_{i\in\mathcal{I}_+}\mathrm{relu}((w_j^{(0)})^\top x_i)\right]^2, \sum_{j\in\mathcal{V}_-}\left[\min_{i\in\mathcal{I}_-}\mathrm{relu}((w_j^{(0)})^\top x_i)\right]^2\right\}\left[\mathcal{L}_+^{(0)}+\mathcal{L}_-^{(0)}\right]^2 \tag{38}$$

$$= \frac{1}{2}\min\left\{\sum_{j\in\mathcal{V}_+}\left[\min_{i\in\mathcal{I}_+}\mathrm{relu}((w_j^{(0)})^\top x_i)\right]^2, \sum_{j\in\mathcal{V}_-}\left[\min_{i\in\mathcal{I}_-}\mathrm{relu}((w_j^{(0)})^\top x_i)\right]^2\right\}(\mathcal{L}^{(0)})^2 \tag{39}$$

We construct an upper bound of the diffusion terms in the zeroth order approximation

$$\begin{aligned}
&\frac{1}{2}\sigma^2\sum_{i=1}^n \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\\
=&\frac{1}{2}\sigma^2\sum_{i=1}^n \left\{\ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\right\}\cdot\left\{\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right\}\\
&//\text{by Lemma B.1}\\
\leq&\frac{1}{2}\sigma^2\left\{\sum_{i=1}^n \ell^2(y_i, f(x_i; W^{(0)}, v^{(0)}))\right\}^{1/2}\cdot\left\{\sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^4\right\}^{1/2}\\
&//\text{by Lemma B.3}\\
\leq&\frac{1}{2}\sigma^2\left\{\sum_{i=1}^n \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\right\}\cdot\left\{\sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^4\right\}^{1/2}\\
=&\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\cdot\left\{\sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^4\right\}^{1/2}
\end{aligned}$$

Then we have an upper bound

$$\mathcal{L}^{(0)}(T) \leq \frac{1}{\frac{1}{\mathcal{L}^{(0)}(0)}e^{-BT}+\frac{A}{B}(1-e^{-BT})}$$

where constants A, B are defined as

$$\begin{cases} A = \frac{1}{2}\min\left\{\sum_{j\in\mathcal{V}_+}\left[\min_{i\in\mathcal{I}_+}\mathrm{relu}((w_j^{(0)})^\top x_i)\right]^2, \sum_{j\in\mathcal{V}_-}\left[\min_{i\in\mathcal{I}_-}\mathrm{relu}((w_j^{(0)})^\top x_i)\right]^2\right\} \\ B = \frac{1}{2}\sigma^2\left\{\sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^4\right\}^{1/2} \end{cases}$$

□

We give the lower bound of approxiamte DP-LP below. We first give a loose lower bound as a warm-up. Then we improve the techniques and provide a tight lower bound.

Loose lower bound proof of Theorem 4.2. We rewrite the Langevin diffusion by asymptotic expansion (Freidlin et al., 2012, Equation 2.1, Chapter 2.2)

$$\begin{aligned}
\dot{\mathcal{L}}^{(0)} =& -\|\nabla_v\mathcal{L}^{(0)}\|_2^2 + \frac{1}{2}\sigma^2\sum_{i=1}^n y_i^2\ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\\
=& -\|\nabla_v\mathcal{L}^{(0)}\|_2^2 + \frac{1}{2}\sigma^2\sum_{i=1}^n \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\\
\geq& -\|\nabla_v\mathcal{L}^{(0)}\|_2^2 + \left(\min_{i\in\mathcal{V}_+^{(0)}}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\sum_{i\in\mathcal{V}_+^{(0)}}\ell(y_i, f(x_i; W^{(0)}, v^{(0)}))
\end{aligned}$$

$$
\begin{aligned}
&+ \left(\min_{i\in\mathcal{V}_-^{(0)}} \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\sum_{i\in\mathcal{V}_-^{(0)}}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\\
=&-\|\nabla_v\mathcal{L}^{(0)}\|_2^2+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\sum_{i\in[n]}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\\
=&-\|\nabla_v\mathcal{L}^{(0)}\|_2^2+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\\
=&-\sum_{j=1}^h\left(\sum_{i=1}^n y_i\exp(-y_i f(x_i;W^{(0)},v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\\
&//\text{by trapping}\\
=&-\sum_{j\in\mathcal{V}_+^{(0)}}\left(\sum_{i\in\mathcal{I}_+}\exp(-f(x_i;W^{(0)},v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2\\
&-\sum_{j\in\mathcal{V}_-^{(0)}}\left(\sum_{i\in\mathcal{I}_-}\exp(f(x_i;W^{(0)},v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2\\
&+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\\
\geq&-\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)\sum_{j\in\mathcal{V}_+^{(0)}}\left(\sum_{i\in\mathcal{I}_+}\exp(-f(x_i;W^{(0)},v^{(0)}))\right)^2\\
&-\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)\sum_{j\in\mathcal{V}_-^{(0)}}\left(\sum_{i\in\mathcal{I}_-}\exp(f(x_i;W^{(0)},v^{(0)}))\right)^2\\
&+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\\
&//a^2+b^2\leq(a+b)^2\text{ when }a>0,b>0\\
\geq&-\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)\sum_{j\in[h]}\left(\sum_{i\in[n]}\exp(-f(x_i;W^{(0)},v^{(0)}))\right)^2\\
&+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\\
\geq&-h\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)\left(\sum_{i\in[n]}\exp(-f(x_i;W^{(0)},v^{(0)}))\right)^2+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}\\
\geq&-h\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)(\mathcal{L}^{(0)})^2+\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)\cdot\frac{1}{2}\sigma^2\mathcal{L}^{(0)}
\end{aligned}
$$

In linear probing, the coefficients $h\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)$ and $\frac{1}{2}\sigma^2\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)$ are constants. We replace them with dummy notation A and B. We solve the first-order nonlinear ODE by turning it into a first-order linear ODE.

$$
\begin{aligned}
\dot{\mathcal{L}}^{(0)}\geq&-A(\mathcal{L}^{(0)})^2+B\mathcal{L}^{(0)}\\
\frac{1}{(\mathcal{L}^{(0)})^2}\dot{\mathcal{L}}^{(0)}\geq&-A+B\frac{1}{\mathcal{L}^{(0)}}
\end{aligned}
$$

$$-\frac{d}{dt}\left(\frac{1}{\mathcal{L}^{(0)}}\right) \geq -A + B\frac{1}{\mathcal{L}^{(0)}}$$
$$\mathcal{L}^{(0)}(T) \geq \frac{1}{\frac{1}{\mathcal{L}^{(0)}(0)}e^{-BT} + \frac{A}{B}(1-e^{-BT})}$$

□

Remark D.1 (On the qualitative properties of loose DP-LP lower bound). If we take the limit to initial point, then the lower bound degenerate to the initial loss value.

$$\lim_{t\to 0} \frac{1}{\frac{1}{\mathcal{L}^{(0)}(0)}e^{-BT} + \frac{A}{B}(1-e^{-BT})} = \mathcal{L}^{(0)}(t=0) = \mathcal{L}(t=0) \tag{40}$$

If we take the limit to infinite time,

$$\lim_{t\to\infty} \frac{1}{\frac{1}{\mathcal{L}^{(0)}(0)}e^{-BT} + \frac{A}{B}(1-e^{-BT})} = \frac{B}{A} = \frac{\frac{1}{2}\sigma^2\left(\min_{i\in[n]}\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\right)}{h\left(\max_{j\in[h],i\in[n]}(\mathrm{relu}((w_j^{(0)})^\top x_i))^2\right)} \tag{41}$$

the following interpretation holds:

1. For larger noise $\sigma\uparrow$, the lower bound is higher, i.e. worse performance.
2. For bad alignment between pretrained features $W^(0)$ and data points, both the denominator and the numerator could shrink. It is not obvious how the lower bound changes.

In the following result, we modify the proof, replace the $\min(\cdot)$, and provide a tighter bound.

Tight lower bound proof of Theorem 4.2. This is an alternative construction of a lower bound for drift terms in the zeroth order approximation

$$\begin{aligned}
\|\nabla_v \mathcal{L}^{(0)}\|_2^2 =& \sum_{j=1}^h \left(\sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
=& \sum_{j\in\mathcal{V}_+^{(0)}} \left(\sum_{i\in\mathcal{I}_+} \exp(-f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
&+ \sum_{j\in\mathcal{V}_-^{(0)}} \left(\sum_{i\in\mathcal{I}_-} \exp(f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
&//\text{by Lemma B.3} \\
\leq& \left(\sum_{j\in\mathcal{V}_+^{(0)}}\sum_{i\in\mathcal{I}_+} \exp(-f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
&+ \left(\sum_{j\in\mathcal{V}_-^{(0)}}\sum_{i\in\mathcal{I}_-} \exp(f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
\leq& \left(\sum_{j\in[h]}\sum_{i\in[n]} \exp(-f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
=& \left(\sum_{i\in[n]}\sum_{j\in[h]} \exp(-f(x_i; W^{(0)}, v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right] \exp(-f(x_i; W^{(0)}, v^{(0)})) \right)^2 \\
&\quad //\text{by Lemma B.1} \\
&\leq \left(\sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right]^2 \right) \left(\sum_{i\in[n]} \exp(-f(x_i; W^{(0)}, v^{(0)}))^2 \right) \\
&\quad //\text{by Lemma B.3} \\
&\leq \left(\sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right]^2 \right) \left(\sum_{i\in[n]} \exp(-f(x_i; W^{(0)}, v^{(0)})) \right)^2 \\
&\leq \left(\sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right]^2 \right) (\mathcal{L}^{(0)})^2
\end{aligned}
$$

We replace the A constant by $\sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right]^2$. This is an alternative construction of a lower bound for diffusion-resulted terms in the zeroth order approximation

$$
\begin{aligned}
&\frac{1}{2}\sigma^2 \sum_{i=1}^n \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2 \\
=&\frac{1}{2}\sigma^2 \sum_{i=1}^n \left\{ \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \right\} \cdot \left\{ \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2 \right\} \\
&\quad //\text{by Lemma B.2} \\
\geq&\frac{1}{2}\sigma^2 \left\{ \sum_{i=1}^n \ell^{1/2}(y_i, f(x_i; W^{(0)}, v^{(0)})) \right\}^2 \cdot \left\{ \sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^{-2} \right\}^{-1} \\
&\quad //\text{by Lemma B.3} \\
\geq&\frac{1}{2}\sigma^2 \left\{ \sum_{i=1}^n \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \right\} \cdot \left\{ \sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^{-2} \right\}^{-1} \\
\geq&\frac{1}{2}\sigma^2 \mathcal{L}^{(0)} \cdot \left\{ \sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^{-2} \right\}^{-1}
\end{aligned}
$$

We replace the B constant by $\left\{ \sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^{-2} \right\}^{-1}$ in the previous proof of loose lower bound of Theorem 4.2. Similarly,

$$
\mathcal{L}^{(0)}(T) \geq \frac{1}{\frac{1}{\mathcal{L}^{(0)}(0)} e^{-BT} + \frac{A}{B}(1 - e^{-BT})}
$$

where $A = \sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right]^2, B = \frac{1}{2}\sigma^2 \left\{ \sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^{-2} \right\}^{-1}$. The limit of this lower bound is

$$
\lim_{t\to\infty} \frac{1}{\frac{1}{\mathcal{L}^{(0)}(0)} e^{-BT} + \frac{A}{B}(1 - e^{-BT})} = \frac{B}{A} = \frac{1}{2}\sigma^2 \left\{ \sum_{i=1}^n \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^{-2} \right\}^{-1} \left\{ \sum_{i\in[n]} \left[\max_{j\in[h]} \mathrm{relu}((w_j^{(0)})^\top x_i) \right]^2 \right\}^{-1}
$$

□

Example D.2 (On the downstream alignment of pretrained features (Theorem 4.2))**.** Here we provide an example on how the pretrained feature space affects the linear probing lower bound in Theorem 4.2 in the **overparametrized** regime. Consider one data point x_+ and two pretrained features $w_{+,1}, w_{+,2}$ with $\|x_+\|_2 = \|w_{+,1}\|_2 = \|w_{+,2}\|_2 = 1, \cos(x_+, w_{+,2}) = \frac{1}{3}\pi$.

1. If we get lucky such that $w_{+,1} = x_+$, then the limit is $\frac{B}{A} = \frac{15}{24}\sigma^2$.

2. If the $w_{+,1}$ is not so good for the downstream task such that $\cos(x_+, w_{+,1}) = \frac{1}{6}\pi$, then the limit becomes $\frac{B}{A} = \frac{16}{24}\sigma^2$.

Since $\frac{16}{24} > \frac{15}{24}$, we can tell that when the pretrained features do not align well with the downstream task, the lower bound gets higher, i.e. worse performance.

D.2 Approximate DP-FT convergence

Analysis of DP-FFT loss diffusion. In the following 0^{th}-order approximation of loss Langevin diffusion, denote the drift term by W-gradient as T_1, the drift term by v-gradient as T_2, the diffusion term by W-hessian as T_3, the diffusion term by v-hessian as T_4.

$$\dot{\mathcal{L}}^{(0)} = -\underbrace{\left\|\nabla_W \mathcal{L}^{(0)}\right\|_F^2}_{T_1} - \underbrace{\left\|\nabla_v \mathcal{L}^{(0)}\right\|_2^2}_{T_2} \tag{42}$$

$$+ \frac{1}{2}\sigma^2 \sum_{i=1}^n y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \left(\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2 + \sum_{j=1}^h (v_j^{(0)})^2 [\mathrm{relu}'((w_j^{(0)})^\top x_i)]^2 \|x_i\|_2^2 \right) \tag{43}$$

$$= -\sum_{j=1}^h \left(\sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \mathrm{relu}((w_j^{(0)})^\top x_i) \right)^2 \tag{44}$$

$$- \sum_{j=1}^h \left\| \sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} x_i \right\|_2^2 \tag{45}$$

$$+ \frac{1}{2}\sigma^2 \sum_{i=1}^n y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \left(\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2 + \sum_{j=1}^h (v_j^{(0)})^2 \mathbb{1}^2_{(w_j^{(0)})^\top x_i > 0} \|x_i\|_2^2 \right) \tag{46}$$

$$= \underbrace{-\sum_{j=1}^h \left(\sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \mathrm{relu}((w_j^{(0)})^\top x_i) \right)^2}_{T_2} \tag{47}$$

$$\underbrace{- \sum_{j=1}^h \left\| \sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} x_i \right\|_2^2}_{T_1} \tag{48}$$

$$\underbrace{+ \frac{1}{2}\sigma^2 \sum_{i=1}^n y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2}_{T_4} \tag{49}$$

$$\underbrace{+ \frac{1}{2}\sigma^2 \sum_{i=1}^n y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \sum_{j=1}^h (v_j^{(0)})^2 \mathbb{1}^2_{(w_j^{(0)})^\top x_i > 0} \|x_i\|_2^2}_{T_3} \tag{50}$$

Upper bound proof of Theorem 4.3. **1. Upper bounds for T_1, T_3.** For T_1, the key idea is $\|x\|_2^2 \geq \langle x, z\rangle^2$ for any unit vector z.

$$T_1 = -\sum_{j=1}^h \left\| \sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} x_i \right\|_2^2$$

$$
\begin{aligned}
&//\text{since } \forall x \in \mathbb{R}^D, z \in \mathbb{S}^{D-1}, \|x\|_2^2 \geq \langle x, z\rangle^2 \\
\leq & -\sum_{j=1}^{h}\left\langle \sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i>0} x_i, z\right\rangle^2 \\
= & -\sum_{j=1}^{h}\left(\sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i>0} \langle x_i, z\rangle\right)^2 \\
= & -\sum_{j=1}^{h}(v_j^{(0)})^2\left(\sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\mathbb{1}_{(w_j^{(0)})^\top x_i>0} \langle x_i, z\rangle\right)^2 \\
&//\text{pick } z = \frac{y_1 x_1}{\|x_1\|_2}, \text{ by Corollary B.5} \\
\leq & -\gamma^2\sum_{j=1}^{h}(v_j^{(0)})^2\left(\sum_{i=1}^{n} \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\right)^2 \\
= & -\gamma^2\sum_{j=1}^{h}(v_j^{(0)})^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\right)^2 \\
= & -\gamma^2\sum_{j=1}^{h}(v_j^{(0)})^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\right)^2
\end{aligned}
$$

For T_3, we align its form with T_1.

$$
\begin{aligned}
T_3 = & \frac{1}{2}\sigma^2\sum_{i=1}^{n} y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\sum_{j=1}^{h}(v_j^{(0)})^2 \mathbb{1}^2_{(w_j^{(0)})^\top x_i>0}\|x_i\|_2^2 \\
& //\text{since } \forall i \in [n], |y_i| = 1 \\
= & \frac{1}{2}\sigma^2\sum_{i=1}^{n} \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))\sum_{j=1}^{h}(v_j^{(0)})^2 \mathbb{1}_{(w_j^{(0)})^\top x_i>0}\|x_i\|_2^2 \\
= & \frac{1}{2}\sigma^2\sum_{j=1}^{h}(v_j^{(0)})^2\sum_{i=1}^{n} \|x_i\|_2^2 \mathbb{1}_{(w_j^{(0)})^\top x_i>0}\ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
\leq & \frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\sum_{j=1}^{h}(v_j^{(0)})^2\sum_{i=1}^{n} \mathbb{1}_{(w_j^{(0)})^\top x_i>0}\ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
= & \frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\sum_{j=1}^{h}(v_j^{(0)})^2\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))
\end{aligned}
$$

2. Upper bounds of T_2, T_4. For T_2, we use linear separability.

$$
\begin{aligned}
T_2 = & -\sum_{j=1}^{h}\left(\sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\text{relu}((w_j^{(0)})^\top x_i)\right)^2 \\
& //\text{by Corollary B.5} \\
\leq & -\sum_{j=1}^{h}\left(\sum_{i\in[n]} \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\gamma\|w_j^{(0)}\|_2\right)^2 \\
= & -\gamma^2\sum_{j=1}^{h}\|w_j^{(0)}\|_2^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \exp(-y_i f(x_i; W^{(0)}, v^{(0)}))\right)^2
\end{aligned}
$$

$$= -\gamma^2 \sum_{j=1}^{h} \|w_j^{(0)}\|_2^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \right)^2$$

For T_4, we align its form with T_3.

$$\begin{aligned}
T_4 =& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2 \\
& //\text{since } \forall i \in [n], |y_i| = 1 \\
=& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2 \\
=& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \sum_{j \in [h]} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \langle w_j^{(0)}, x_i \rangle^2 \\
\leq& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \sum_{j \in [h]} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \|w_j^{(0)}\|_2^2 \|x_i\|_2^2 \\
\leq& \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \sum_{j=1}^{h} \|w_j^{(0)}\|_2^2 \sum_{i \in [n]} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
=& \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \sum_{j=1}^{h} \|w_j^{(0)}\|_2^2 \sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))
\end{aligned}$$

3. Combine upper bounds of T_1, T_2, T_3, T_4.

$$\begin{aligned}
\dot{\mathcal{L}}^{(0)} =& T_1 + T_2 + T_3 + T_4 \\
\leq& -\gamma^2 \sum_{j=1}^{h} \left[(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2 \right] \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \right)^2 \\
& + \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \sum_{j=1}^{h} \left[(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2 \right] \sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
& //\text{abbr. } \ell_i := \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
=& -\gamma^2 \sum_{j=1}^{h} \left[(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2 \right] \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell_i \right)^2 \\
& + \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \sum_{j=1}^{h} \left[(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2 \right] \sum_{i \in \mathscr{I}(w_j^{(0)})} \ell_i \\
=& \sum_{j=1}^{h} \left[(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2 \right] \left\{ -\gamma^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell_i \right)^2 + \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell_i \right) \right\}
\end{aligned}$$

$\because (v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2 \geq (v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2$

$\therefore$ When the drift term (negative) still dominates the dynamics, we take $t = 0$ for $(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2$.

$$\dot{\mathcal{L}}^{(0)} \leq \sum_{j=1}^{h} \left[(v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2 \right] \left\{ -\gamma^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell_i \right)^2 + \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell_i \right) \right\}$$

4. Decompose loss by trapping. If the trapping condition holds, we can decompose the loss $\mathcal{L}^{(0)} = \mathcal{L}_+^{(0)} + \mathcal{L}_-^{(0)}$, where $\mathcal{L}_*^{(0)}$ is only controlled by w_j if $w_j^{(0)} \in \mathcal{S}_*$ ($* \in \{+,-\}$).

$$\begin{aligned}
\dot{\mathcal{L}}_*^{(0)} \leq & \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \left[(v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2\right] \left\{-\gamma^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right)^2 + \frac{1}{2}\sigma^2 \left(\max_{i\in[n]} \|x_i\|_2^2\right) \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right)\right\} \\
\leq & \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \left[(v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2\right] \left\{-\gamma^2 \left(\mathcal{L}_*^{(0)}\right)^2 + \frac{1}{2}\sigma^2 \left(\max_{i\in[n]} \|x_i\|_2^2\right) \mathcal{L}_*^{(0)}\right\}
\end{aligned}$$

Let $u = 1/\mathcal{L}_*^{(0)}, A = \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \left[(v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2\right], B = \gamma^2, C = \frac{1}{2}\sigma^2 \left(\max_{i\in[n]} \|x_i\|_2^2\right)$. Then

$$\begin{aligned}
-\frac{du}{dt} \leq & -AB + ACu \\
AB\exp(ACt) \leq & \frac{d}{dt}(ue^{ACt}) \\
\frac{B}{C}(\exp(ACt) - 1) \leq & ue^{ACt} - u_0 \\
\frac{B}{C}(\exp(ACt) - 1) + u_0 \leq & ue^{ACt} \\
\frac{B}{C}(1 - \exp(-ACt)) + u_0 e^{-ACt} \leq & u \\
\mathcal{L}_*^{(0)} \leq & \frac{1}{\frac{B}{C}(1 - e^{-ACt}) + \frac{1}{\mathcal{L}_{t=0,*}^{(0)}} e^{-ACt}}
\end{aligned}$$

The time limit of the upper bound is

$$\lim_{t\to\infty} \mathcal{L}_*^{(0)} \leq \frac{C}{B} = \frac{\sigma^2}{2\gamma^2} \left(\max_{i\in[n]} \|x_i\|_2^2\right) = \frac{1}{2} \frac{\max_{i\in[n]} \|x_i\|_2^2}{\min_{i\in[n]} \|x_i\|_2^2} \sigma^2 \frac{1}{\mu^2}$$

5. Combine clustered losses.

$$\begin{aligned}
\mathcal{L}^{(0)} = & \mathcal{L}_-^{(0)} + \mathcal{L}_+^{(0)} \\
\leq & \frac{1}{\frac{B}{C}(1 - e^{-A_+Ct}) + \frac{1}{\mathcal{L}_{t=0,+}^{(0)}} e^{-A_+Ct}} + \frac{1}{\frac{B}{C}(1 - e^{-A_-Ct}) + \frac{1}{\mathcal{L}_{t=0,-}^{(0)}} e^{-A_-Ct}}
\end{aligned}$$

□

Lower bound (type I) proof of Theorem 4.3. **1. Upper bounds for T_1, T_3.** For T_1, the key idea is $\|x\|_2^2 \geq \langle x, z\rangle^2$ for any unit vector z.

$$\begin{aligned}
T_1 = & -\sum_{j=1}^h \left\| \sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} x_i \right\|_2^2 \\
& //\text{since } \forall x \in \mathbb{R}^D, z \in \mathbb{S}^{D-1}, \|x\|_2^2 \geq \langle x, z\rangle^2 \\
\leq & -\sum_{j=1}^h \left\langle \sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} x_i, z \right\rangle^2 \\
= & -\sum_{j=1}^h \left(\sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \langle x_i, z\rangle \right)^2 \\
= & -\sum_{j=1}^h (v_j^{(0)})^2 \left(\sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \langle x_i, z\rangle \right)^2
\end{aligned}$$

$$
\begin{aligned}
&//\text{pick } z = \frac{y_1 x_1}{\|x_1\|_2}, \text{ by Corollary B.5} \\
\leq& -\gamma^2 \sum_{j=1}^{h} (v_j^{(0)})^2 \left(\sum_{i=1}^{n} \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \right)^2 \\
=& -\gamma^2 \sum_{j=1}^{h} (v_j^{(0)})^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \right)^2 \\
=& -\gamma^2 \sum_{j=1}^{h} (v_j^{(0)})^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \right)^2
\end{aligned}
$$

For T_3, we align its form with T_1.

$$
\begin{aligned}
T_3 =& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \sum_{j=1}^{h} (v_j^{(0)})^2 \mathbb{1}^2_{(w_j^{(0)})^\top x_i > 0} \|x_i\|_2^2 \\
&//\text{since } \forall i \in [n], |y_i| = 1 \\
=& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \sum_{j=1}^{h} (v_j^{(0)})^2 \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \|x_i\|_2^2 \\
=& \frac{1}{2}\sigma^2 \sum_{j=1}^{h} (v_j^{(0)})^2 \sum_{i=1}^{n} \|x_i\|_2^2 \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
\leq& \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \sum_{j=1}^{h} (v_j^{(0)})^2 \sum_{i=1}^{n} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \\
=& \frac{1}{2}\sigma^2 \left(\max_{i \in [n]} \|x_i\|_2^2 \right) \sum_{j=1}^{h} (v_j^{(0)})^2 \sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)}))
\end{aligned}
$$

2. Upper bounds of T_2, T_4. For T_2, we use linear separability.

$$
\begin{aligned}
T_2 =& -\sum_{j=1}^{h} \left(\sum_{i=1}^{n} y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \text{relu}((w_j^{(0)})^\top x_i) \right)^2 \\
&//\text{by Corollary B.5} \\
\leq& -\sum_{j=1}^{h} \left(\sum_{i \in [n]} \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} \gamma \|w_j^{(0)}\|_2 \right)^2 \\
=& -\gamma^2 \sum_{j=1}^{h} \|w_j^{(0)}\|_2^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \right)^2 \\
=& -\gamma^2 \sum_{j=1}^{h} \|w_j^{(0)}\|_2^2 \left(\sum_{i \in \mathscr{I}(w_j^{(0)})} \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \right)^2
\end{aligned}
$$

For T_4, we align its form with T_3.

$$
\begin{aligned}
T_4 =& \frac{1}{2}\sigma^2 \sum_{i=1}^{n} y_i^2 \ell(y_i, f(x_i; W^{(0)}, v^{(0)})) \|\text{relu}((W^{(0)})^\top x_i)\|_2^2 \\
&//\text{since } \forall i \in [n], |y_i| = 1
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{2}\sigma^2\sum_{i=1}^{n}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\\
=&\frac{1}{2}\sigma^2\sum_{i=1}^{n}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\sum_{j\in[h]}\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\langle w_j^{(0)},x_i\rangle^2\\
\le&\frac{1}{2}\sigma^2\sum_{i=1}^{n}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\sum_{j\in[h]}\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\|w_j^{(0)}\|_2^2\|x_i\|_2^2\\
\le&\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\sum_{j=1}^{h}\|w_j^{(0)}\|_2^2\sum_{i\in[n]}\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\\
=&\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\sum_{j=1}^{h}\|w_j^{(0)}\|_2^2\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))
\end{aligned}
$$

3. Combine upper bounds of T_1,T_2,T_3,T_4.

$$
\begin{aligned}
\dot{\mathcal{L}}^{(0)}=&T_1+T_2+T_3+T_4\\
\le&-\gamma^2\sum_{j=1}^{h}\left[(v_j^{(0)})^2+\|w_j^{(0)}\|_2^2\right]\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\right)^2\\
&+\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\sum_{j=1}^{h}\left[(v_j^{(0)})^2+\|w_j^{(0)}\|_2^2\right]\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\\
&//\text{abbr. }\ell_i:=\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\\
=&-\gamma^2\sum_{j=1}^{h}\left[(v_j^{(0)})^2+\|w_j^{(0)}\|_2^2\right]\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)^2\\
&+\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\sum_{j=1}^{h}\left[(v_j^{(0)})^2+\|w_j^{(0)}\|_2^2\right]\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\\
=&\sum_{j=1}^{h}\left[(v_j^{(0)})^2+\|w_j^{(0)}\|_2^2\right]\left\{-\gamma^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)^2+\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)\right\}
\end{aligned}
$$

$\because (v_j^{(0)})^2+\|w_j^{(0)}\|_2^2\ge(v_{j,t=0}^{(0)})^2+\|w_{j,t=0}^{(0)}\|_2^2$

$\therefore$ When the drift term (negative) still dominates the dynamics, we take $t=0$ for $(v_j^{(0)})^2+\|w_j^{(0)}\|_2^2$.

$$
\dot{\mathcal{L}}^{(0)}\le\sum_{j=1}^{h}\left[(v_{j,t=0}^{(0)})^2+\|w_{j,t=0}^{(0)}\|_2^2\right]\left\{-\gamma^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)^2+\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)\right\}
$$

4. Decompose loss by trapping. If the trapping condition holds, we can decompose the loss $\mathcal{L}^{(0)}=\mathcal{L}_+^{(0)}+\mathcal{L}_-^{(0)}$, where $\mathcal{L}_*^{(0)}$ is only controlled by w_j if $w_j^{(0)}\in\mathcal{S}_*$ ($*\in\{+,-\}$).

$$
\dot{\mathcal{L}}_*^{(0)}\le\sum_{j\in[h],w_j^{(0)}\in\mathcal{S}_*}\left[(v_{j,t=0}^{(0)})^2+\|w_{j,t=0}^{(0)}\|_2^2\right]\left\{-\gamma^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)^2+\frac{1}{2}\sigma^2\left(\max_{i\in[n]}\|x_i\|_2^2\right)\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)\right\}
$$

$$\leq \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \left[(v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2\right] \left\{-\gamma^2 \left(\mathcal{L}_*^{(0)}\right)^2 + \frac{1}{2}\sigma^2 \left(\max_{i\in[n]} \|x_i\|_2^2\right) \mathcal{L}_*^{(0)}\right\}$$

Let $u = 1/\mathcal{L}_*^{(0)}, A = \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \left[(v_{j,t=0}^{(0)})^2 + \|w_{j,t=0}^{(0)}\|_2^2\right], B = \gamma^2, C = \frac{1}{2}\sigma^2 \left(\max_{i\in[n]} \|x_i\|_2^2\right)$. Then

$$\begin{aligned}
-\frac{du}{dt} \leq& -AB + ACu \\
AB\exp(ACt) \leq& \frac{d}{dt}(ue^{ACt}) \\
\frac{B}{C}(\exp(ACt) - 1) \leq& ue^{ACt} - u_0 \\
\frac{B}{C}(\exp(ACt) - 1) + u_0 \leq& ue^{ACt} \\
\frac{B}{C}(1 - \exp(-ACt)) + u_0 e^{-ACt} \leq& u \\
\mathcal{L}_*^{(0)} \leq& \frac{1}{\frac{B}{C}(1 - e^{-ACt}) + \frac{1}{\mathcal{L}_{t=0,*}^{(0)}} e^{-ACt}}
\end{aligned}$$

The time limit of the upper bound is

$$\lim_{t\to\infty} \mathcal{L}_*^{(0)} \leq \frac{C}{B} = \frac{\sigma^2}{2\gamma^2} \left(\max_{i\in[n]} \|x_i\|_2^2\right) = \frac{1}{2} \frac{\max_{i\in[n]} \|x_i\|_2^2}{\min_{i\in[n]} \|x_i\|_2^2} \sigma^2 \frac{1}{\mu^2}$$

5. Combine clustered losses.

$$\begin{aligned}
\mathcal{L}^{(0)} =& \mathcal{L}_-^{(0)} + \mathcal{L}_+^{(0)} \\
\leq& \frac{1}{\frac{B}{C}(1 - e^{-A_+Ct}) + \frac{1}{\mathcal{L}_{t=0,+}^{(0)}} e^{-A_+Ct}} + \frac{1}{\frac{B}{C}(1 - e^{-A_-Ct}) + \frac{1}{\mathcal{L}_{t=0,-}^{(0)}} e^{-A_-Ct}}
\end{aligned}$$

□

Lower bound (type III) proof of Theorem 4.3. **1. Lower bounds for T_1, T_3.** For T_1, we use $\left(\max_{k\in[n]} \|x_k\|_2^2\right)$.

$$\begin{aligned}
T_1 =& -\sum_{j=1}^h \left\| \sum_{i=1}^n y_i \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) v_j^{(0)} \mathbb{1}_{(w_j^{(0)})^\top x_i > 0} x_i \right\|_2^2 \\
& //\text{abbr. } \ell_i := \exp(-y_i f(x_i; W^{(0)}, v^{(0)})) \\
=& -\sum_{j=1}^h \left\| \sum_{i\in\mathscr{I}(w_j^{(0)})} y_i \ell_i v_j^{(0)} x_i \right\|_2^2 \\
=& -\sum_{j=1}^h \left\| \sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i v_j^{(0)} x_i \right\|_2^2 \\
=& -\sum_{j\in[h]} (v_j^{(0)})^2 \left\| \sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i x_i \right\|_2^2 \\
\geq& -\sum_{j\in[h]} (v_j^{(0)})^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i \|x_i\|_2 \right)^2
\end{aligned}$$

$$\geq -\left(\max_{k\in[n]}\|x_k\|_2^2\right)\sum_{j\in[h]}(v_j^{(0)})^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)^2$$

For T_3, we align its form with T_1.

$$\begin{aligned}
T_3 =&\frac{1}{2}\sigma^2\sum_{i=1}^n y_i^2\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\sum_{j=1}^h(v_j^{(0)})^2\mathbb{1}^2_{(w_j^{(0)})^\top x_i>0}\|x_i\|_2^2\\
=&\frac{1}{2}\sigma^2\sum_{i=1}^n \ell_i\sum_{j=1}^h(v_j^{(0)})^2\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\|x_i\|_2^2\\
=&\frac{1}{2}\sigma^2\sum_{j\in[h]}(v_j^{(0)})^2\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\|x_i\|_2^2\\
\geq&\frac{1}{2}\sigma^2\left(\min_{k\in[n]}\|x_k\|_2^2\right)\sum_{j\in[h]}(v_j^{(0)})^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)
\end{aligned}$$

2. Lower bounds for T_2, T_4. For T_2, we use $\langle x,y\rangle\leq\|x\|_2\|y\|_2$.

$$\begin{aligned}
T_2 =&-\sum_{j=1}^h\left(\sum_{i=1}^n y_i\exp(-y_i f(x_i;W^{(0)},v^{(0)}))\mathrm{relu}((w_j^{(0)})^\top x_i)\right)^2\\
=&-\sum_{j=1}^h\left(\sum_{i\in\mathscr{I}(w_j^{(0)})} y_i\exp(-y_i f(x_i;W^{(0)},v^{(0)}))(w_j^{(0)})^\top x_i\right)^2\\
=&-\sum_{j\in[h]}\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\langle w_j^{(0)},x_i\rangle\right)^2\\
\geq&-\sum_{j\in[h]}\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\|w_j^{(0)}\|_2\|x_i\|_2\right)^2\\
\geq&-\left(\max_{k\in[n]}\|x_k\|_2^2\right)\sum_{j\in[h]}\|w_j^{(0)}\|_2^2\left(\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\right)^2
\end{aligned}$$

For T_4, we align its form with T_2.

$$\begin{aligned}
T_4 =&\frac{1}{2}\sigma^2\sum_{i=1}^n y_i^2\ell(y_i,f(x_i;W^{(0)},v^{(0)}))\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\\
&//\text{since }\forall i\in[n],|y_i|=1\\
=&\frac{1}{2}\sigma^2\sum_{i=1}^n \ell(y_i,f(x_i;W^{(0)},v^{(0)}))\|\mathrm{relu}((W^{(0)})^\top x_i)\|_2^2\\
=&\frac{1}{2}\sigma^2\sum_{i=1}^n \ell(y_i,f(x_i;W^{(0)},v^{(0)}))\sum_{j\in[h]}\mathbb{1}_{(w_j^{(0)})^\top x_i>0}\langle w_j^{(0)},x_i\rangle^2\\
=&\frac{1}{2}\sigma^2\sum_{j\in[h]}\sum_{i\in\mathscr{I}(w_j^{(0)})}\ell_i\langle w_j^{(0)},x_i\rangle^2\\
&//\text{by Lemma B.4}
\end{aligned}$$

$$
\begin{aligned}
\geq & \frac{1}{2}\sigma^2 \sum_{j\in[h]} \sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i \mu^2 \|w_j^{(0)}\|_2^2 \|x_i\|_2^2 \\
= & \frac{1}{2}\sigma^2\mu^2 \sum_{j\in[h]} \|w_j^{(0)}\|_2^2 \sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i \|x_i\|_2^2 \\
\geq & \frac{1}{2}\sigma^2\mu^2 \left(\min_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h]} \|w_j^{(0)}\|_2^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right)
\end{aligned}
$$

3. Combine lower bounds of T_1, T_2, T_3, T_4.

$$
\begin{aligned}
\dot{\mathcal{L}}^{(0)} = & T_1 + T_2 + T_3 + T_4 \\
\geq & -\left(\max_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h]} \left[(v_j^{(0)})^2 + \|w_j^{(0)}\|_2^2\right] \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right)^2 \\
& + \frac{1}{2}\sigma^2 \left(\min_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h]} \left[(v_j^{(0)})^2 + \mu^2\|w_j^{(0)}\|_2^2\right] \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right) \\
& //\text{by balancedness, } \|w_j^{(0)}\|_2^2 = (v_j^{(0)})^2 \\
\geq & -2\left(\max_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h]} \|w_j^{(0)}\|_2^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right)^2 + \frac{\sigma^2(1+\mu^2)}{2} \left(\min_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h]} \|w_j^{(0)}\|_2^2 \left(\sum_{i\in\mathscr{I}(w_j^{(0)})} \ell_i\right)
\end{aligned}
$$

4. Decompose loss by trapping. If the trapping condition holds, we can decompose the loss $\mathcal{L}^{(0)} = \mathcal{L}_+^{(0)} + \mathcal{L}_-^{(0)}$, where $\mathcal{L}_*^{(0)}$ is only controlled by w_j if $w_j^{(0)} \in \mathcal{S}_*$ ($* \in \{+,-\}$).

$$
\begin{aligned}
\dot{\mathcal{L}}_*^{(0)} \geq & -2\left(\max_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \|w_j^{(0)}\|_2^2 (\mathcal{L}_*^{(0)})^2 + \frac{\sigma^2(1+\mu^2)}{2} \left(\min_{k\in[n]} \|x_k\|_2^2\right) \sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \|w_j^{(0)}\|_2^2 \mathcal{L}_*^{(0)} \\
= & \left\{\sum_{j\in[h], w_j^{(0)}\in\mathcal{S}_*} \|w_j^{(0)}\|_2^2\right\} \cdot \left\{-2\left(\max_{k\in[n]} \|x_k\|_2^2\right)(\mathcal{L}_*^{(0)})^2 + \frac{\sigma^2(1+\mu^2)}{2}\left(\min_{k\in[n]} \|x_k\|_2^2\right)\mathcal{L}_*^{(0)}\right\}
\end{aligned}
$$

The time limit of the loss lower bound is

$$
\lim_{t\to\infty} \mathcal{L}_*^{(0)} \geq \frac{1}{2} \frac{\min_{k\in[n]} \|x_k\|_2^2}{\max_{k\in[n]} \|x_k\|_2^2} \sigma^2 \frac{1+\mu^2}{2}
$$

By the previous lower bound proof,

$$
\|W^{(0)}\|_F^2 \leq \|W_0^{(0)}\|_F^2 e^{2(\max_{k\in[n]} \|x_i\|_2)\mathcal{L}_0^{(0)} t}
$$

Let $u = \frac{1}{\mathcal{L}_*^{(0)}}, A = \|W_0^{(0)}\|_F^2, \lambda_2 = 2(\max_{k\in[n]} \|x_i\|_2)\mathcal{L}_0^{(0)}, B = 2\max_{k\in[n]} \|x_k\|_2^2, C = \frac{\sigma^2(1+\mu^2)}{2} \min_{k\in[n]} \|x_k\|_2^2$. Then consider integrating factor $\exp(AC/\lambda_2 \exp(\lambda_2 t))$.

$$
\begin{aligned}
-\frac{d}{dt}u \geq & Ae^{\lambda_2 t}(-B + Cu) \\
ABe^{\lambda_2 t} \geq & ACe^{\lambda_2 t}u + \frac{d}{dt}u \\
ABe^{\lambda_2 t} \exp(AC/\lambda_2 \exp(\lambda_2 t)) \geq & AC \exp(AC/\lambda_2 \exp(\lambda_2 t)) e^{\lambda_2 t} u + \exp(AC/\lambda_2 \exp(\lambda_2 t)) \frac{d}{dt}u \\
\frac{B}{C}\frac{d}{dt}[\exp(AC/\lambda_2 \exp(\lambda_2 t))] \geq & \frac{d}{dt}(u \cdot \exp(AC/\lambda_2 \exp(\lambda_2 t)))
\end{aligned}
$$

$$\frac{B}{C}[\exp(AC/\lambda_2\exp(\lambda_2 t)) - \exp(AC/\lambda_2)] \geq u \cdot \exp(AC/\lambda_2\exp(\lambda_2 t)) - u_0 \cdot \exp(AC/\lambda_2)$$

$$\frac{B}{C}[1 - \exp(AC/\lambda_2(1-\exp(\lambda_2 t)))] \geq u - u_0 \cdot \exp(AC/\lambda_2(1-\exp(\lambda_2 t)))$$

$$\mathcal{L}_*^{(0)} \geq \frac{1}{\frac{1}{\mathcal{L}_{*,t=0}^{(0)}} e^{AC/\lambda_2(1-\exp(\lambda_2 t))} + \frac{B}{C}\left[1 - e^{AC/\lambda_2(1-\exp(\lambda_2 t))}\right]}$$

5. Combine clustered losses.

$$\mathcal{L}^{(0)} = \mathcal{L}_-^{(0)} + \mathcal{L}_+^{(0)}$$

$$\geq \frac{1}{\frac{1}{\mathcal{L}_{+,t=0}^{(0)}} e^{AC/\lambda_2(1-\exp(\lambda_2 t))} + \frac{B}{C}\left[1 - e^{AC/\lambda_2(1-\exp(\lambda_2 t))}\right]} + \frac{1}{\frac{1}{\mathcal{L}_{-,t=0}^{(0)}} e^{AC/\lambda_2(1-\exp(\lambda_2 t))} + \frac{B}{C}\left[1 - e^{AC/\lambda_2(1-\exp(\lambda_2 t))}\right]}$$

□

D.3 Privacy budget allocation

Proof of Theorem 5.1. For any $j \in [h]$, with probability $1-\rho$, its initial absolute value is bounded by

$$|v_j| \leq \sqrt{2\beta^2 \ln(2/\rho)} \tag{51}$$

Then with probability $(1-\rho)^h$, the maximum worse initial value is bounded by

$$\max_{j\in[h]}(c_j \cdot v_j) \leq \sqrt{\beta^2 \ln(2/\rho)} \tag{52}$$

where we define c_j by $w_j \in S_{c_j}$. The approximate DP-LP dynamics is

$$\dot{v}_j = \sum_{i=1}^{n} y_i \ell_i \mathrm{relu}(w_j^\top x_i) \tag{53}$$

Say $w_j \in S_c$ for some $c \in \{-1, 1\}$, then during DP-LP, when $\mathrm{sign}(v_j(T)) = \mathrm{sign}(v_j(0))$,

$$|v_j(T) - v_j(0)| = \int_0^T \sum_{y_i=c} \ell_i \mathrm{relu}(w_j^\top x_i) dt \tag{54}$$

$$\geq \min_{y_i=c} |\mathrm{relu}(w_j^\top x_i)| \int_0^T \mathcal{L}_c(t) dt \tag{55}$$

$$//\text{by Theorem 4.2} \tag{56}$$

$$\geq \min_{y_i=c} \mathrm{relu}(w_j^\top x_i) \frac{\frac{1}{2}\sigma^2 \left\{\sum_{y_i=c} \|\mathrm{relu}(W^\top x_i)\|_2^{-2}\right\}^{-1}}{\sum_{w_j\in S_c}\left[\max_{y_i=c} w_j^\top x_i\right]^2} \tag{57}$$

$$= \frac{1}{2}\sigma^2 \frac{\min_{y_i=c}\mathrm{relu}(w_j^\top x_i)}{\sum_{w_j\in S_c}\left[\max_{y_i=c} w_j^\top x_i\right]^2} \left\{\sum_{y_i=c} \|\mathrm{relu}(W^\top x_i)\|_2^{-2}\right\}^{-1} \tag{58}$$

$$= \frac{1}{2}\sigma^2 Q \tag{59}$$

where we define a constant Q to describe the pre-training quality. If the pre-trained features are better, Q becomes larger. To mitigate the feature distortion, we need $c \cdot v_j > 0$, then the necessary DP-LP run-time is

$$\Delta t \propto \frac{\sigma^2}{Q}\sqrt{\beta^2 \ln(2/\rho)} \propto \frac{\sigma^2}{Q}\sqrt{\ln(2/\rho)} \tag{60}$$

where we ignore β as it is typically pre-determined in real implementations (e.g. the Linear layers in PyTorch). □