Initialization Matters: Privacy-Utility Analysis of Overparameterized Neural Networks

Jiayuan Ye†, Zhenyu Zhu‡, Fanghui Liu†, Reza Shokri†, Volkan Cevher‡

† National University of Singapore  ‡ École Polytechnique Fédérale de Lausanne
† [first name]@comp.nus.edu.sg  ‡ [firstname].[surname]@epfl.ch

Abstract

We analytically investigate how overparameterization of models in randomized machine learning algorithms impacts the information leakage about their training data. Specifically, we prove a privacy bound for the KL divergence between model distributions on worst-case neighboring datasets, and explore its dependence on the initialization, width, and depth of fully connected neural networks. We find that this KL privacy bound is largely determined by the expected squared gradient norm relative to model parameters during training. Notably, for the special setting of linearized network, our analysis indicates that the squared gradient norm (and therefore the escalation of privacy loss) is tied directly to the per-layer variance of the initialization distribution. By using this analysis, we demonstrate that privacy bound improves with increasing depth under certain initializations (LeCun and Xavier), while degrades with increasing depth under other initializations (He and NTK). Our work reveals a complex interplay between privacy and depth that depends on the chosen initialization distribution. We further prove excess empirical risk bounds under a fixed KL privacy budget, and show that the interplay between privacy utility trade-off and depth is similarly affected by the initialization.

1 Introduction

Deep neural networks (DNNs) in the over-parameterized regime (i.e., more parameters than data) perform well in practice but the model predictions can easily leak private information about the training data under inference attacks such as membership inference attacks [42] and reconstruction attacks. [16, 6, 27] This leakage can be mathematically measured in terms of how much the algorithm’s output distribution changes if it were trained on a neighboring dataset (that only differs in one record), following the differential privacy (DP) framework [27].

To train differential private model, a typical way is randomize each gradient update in neural networks training, e.g., stochastic gradient descent (SGD), which leads to the most widely applied differentially private training algorithm in the literature – DP-SGD [1]. In each step, DP-SGD employs gradient clipping and adds calibrated Gaussian noise, and thus it comes with a differential privacy guarantee that scales with the noise multiplier (i.e., per-dimensional Gaussian noise standard deviation divided by the clipping threshold) and number of training epochs. However, this privacy bound [1] is overly general due to its independence on the network properties (e.g., width and depth) and training schemes (e.g., initializations). Accordingly, a natural question arises in the community:

How does overparameterization (e.g., increasing width and depth) of neural networks affect the (worst-case) privacy bound of the training algorithm?

We further analyze how the privacy utility trade-off is affected by overparameterization, in terms of empirical risk bounds given KL privacy budget $\varepsilon$. For the excess risk bounds, we assume the network width $m = \Omega(n)$ and data dimension $d = \Omega(n)$ are sufficiently large, and the data and network satisfy regularity assumption Assumption[22]. For NTK, He and LeCun initialization, we observe that the privacy utility trade-off improves with overparameterization (increasing depth).

Table 1: Our results for the privacy utility trade-off of training linearized network (3) via Langevin diffusion, under different width $m$, depth $L$ and initializations. We set per-layer width $m_l = d$, $m_1, \cdots, m_{L-1} = m$ and $m_L = o$. We prove privacy bound in KL divergence, and obtain excess empirical risk bounds given KL privacy budget $\varepsilon$. For the excess risk bounds, we assume the network width $m = \Omega(n)$ and data dimension $d = \Omega(n)$ are sufficiently large, and the data and network satisfy regularity assumption Assumption[22]. For NTK, He and LeCun initialization, we observe that the privacy utility trade-off improves with overparameterization (increasing depth).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LeCun [32]</td>
<td>$1/m_{l-1}$</td>
<td>$\frac{om(L-1+\frac{d}{m})}{2^{L-1+\frac{d}{m}}}$</td>
<td>$\hat{O}\left(\frac{n}{m} \cdot \frac{1}{L-1+\frac{d}{m}}\right)$</td>
<td>$\hat{O}\left(\frac{1}{n^2} + \sqrt{\frac{2\varepsilon}{md}}\right)$</td>
</tr>
<tr>
<td>He [28]</td>
<td>$2/m_{l-1}$</td>
<td>$\frac{om(L-1+\frac{d}{m})}{d}$</td>
<td>$\hat{O}\left(\frac{n}{m} \cdot \max\left(\frac{1}{L-1+\frac{d}{m}}, 2\right)\right)$</td>
<td>$\hat{O}\left(\frac{1}{n^2} + \sqrt{\frac{\max\left(\frac{1}{L-1+\frac{d}{m}}, 2\right)}{md}}\right)$</td>
</tr>
<tr>
<td>NTK [2]</td>
<td>$\frac{2}{m_l}, l &lt; L$</td>
<td>$\frac{m(L-1)}{2} + o$</td>
<td>$\hat{O}\left(\frac{n}{dm} \cdot \frac{1}{L-1+\frac{d}{m}}\right)$</td>
<td>$\hat{O}\left(\frac{1}{n^2} + \sqrt{\frac{\frac{1}{L-1+\frac{d}{m}}}{d\varepsilon}}\right)$</td>
</tr>
<tr>
<td>Xavier [23]</td>
<td>$\frac{2}{m_{l-1}+m_l}$</td>
<td>$\frac{om(L-1+\frac{d}{m})}{2^{L-1}(1+\frac{d}{m})^2}$</td>
<td>$\hat{O}\left(\frac{n}{d} \cdot \frac{(1+\frac{d}{m})(1+\frac{d}{m})}{L-1+\frac{d}{m}}\right)$</td>
<td>$\hat{O}\left(\frac{1}{n^2} + \sqrt{\frac{1}{d^2\varepsilon}}\right)$</td>
</tr>
</tbody>
</table>

To answer this question, we would need new algorithmic framework and (or) new privacy analyses. In this paper, we focus on analyzing privacy for the Langevin diffusion algorithm[4]. This is to avoid artificially setting a sensitivity constraint on the gradient update and thus making the privacy bound insensitive to the network overparameterization (as in DP-SGD analysis). Instead, we prove a KL privacy bound for Langevin diffusion that scales with the expected gradient difference between the training on any two worst-case neighboring datasets (Theorem 3.1). By proving precise upper bounds on the expected $\ell_2$-norm of this gradient difference, we obtain KL privacy bounds for fully connected neural network (Lemma 3.2 and its linearized variant (Corollary 4.2) that changes with the network width, depth and per-layer variance for the initialization distribution. We summarized the details of our KL privacy bounds in Table 1 and highlight our key observations below.

- Width always worsens privacy, under all the considered initialization distributions. Meanwhile, the interplay between network depth and privacy is much more complex and crucially depends on what initialization distribution is used and how long the training time is.
- Specifically, when the initialization distribution has small per-layer variance (such as LeCun and Xavier initialization), our KL privacy bound for training fully connected network (with a small amount of time) and for training linearized network (with finite time) decay exponentially with increasing depth, as long as the depth is large enough. To the best of our knowledge, this is the first time that an improvement of privacy bound under overparameterization is observed for randomized training algorithm.

We further analyze how the privacy utility trade-off is affected by overparameterization, in terms of excess empirical risk bound (given fixed KL privacy budget). For training linearized network using Langevin diffusion, our excess empirical risk bound scales with the lazy training distance $R$ (i.e., how close is the initialization vector to an optimal solution for the empirical risk minimization problem), as well as a gradient norm constant $B$ in Langevin diffusion. By analyzing these two terms precisely, we prove that given any fixed KL privacy budget $\varepsilon$, our risk bounds strictly improves with increasing depth and width for linearized network under LeCun and Xavier initialization. To our best knowledge, this is the first time that such a gain in privacy-utility trade-off due to overparameterization (increasing depth) is shown. Meanwhile, prior results only prove (nearly) dimension-independent privacy utility trade-off for such linear models in the literature [43][50][55]. Our improvement demonstrates the unique benefits of our new KL privacy analysis in understanding the effect of overparameterization.

1 A key difference between this paper and existing privacy utility analysis of Langevin diffusion [24] is that we analyze in the absence of gradient clipping or Lipschitz assumption on loss function. Our results also readily extend to discretized noisy GD with constant step-size (as discussed in Appendix [1]).

2 We focus on KL privacy loss because it is a more relaxed distinguishability notion than standard $(\varepsilon, \delta)$-DP, and therefore could be upper bounded even without gradient clipping. Moreover, KL divergence enables upper bound for the advantage (relative success) of various inference attacks, as studied in recent works [37][20].
1.1 Related Works

**Overparameterization in DNNs and NTK.** Theoretical demonstration on the benefit of overparameterization in DNNs occurs in global convergence [2][19], generalization [3][15]. Under proper initialization, the training dynamics of over-parameterized DNNs can be described by a kernel function, termed as neural tangent kernel (NTK) [29], which stimulates a series of analysis in DNNs. Accordingly, over-parameterization has been demonstrated to be beneficial/harmful to several topics in deep learning, e.g., robustness [14][52], covariate shift [48]. However, the relationship between overparameterization and privacy (based on the differential privacy framework) remains largely an unsolved problem, as the training dynamics typically change [13] after adding new components in the privacy-preserving learning algorithm (such as DP-SGD [1]) to enforce privacy constraints.

**Membership inference privacy risk under overparameterization.** A recent line of works [46][47] investigates how overparameterization affects the theoretical and empirical privacy in terms of membership inference advantage, and proves novel trade-off between privacy and generalization error. These are the closest works in the literature to our objective of investigating the interplay between privacy and overparameterization. However, Tan et al. [46][47] focus on proving upper bounds for an average-case privacy risk defined by the advantage (relative success) of membership inference attack on models trained from randomly sampled training dataset from a population distribution. By contrast, our KL privacy bound is heavily based on the strongest adversary model in the differential privacy definition, and holds under an arbitrary worst-case pair of neighboring datasets that only differ in one record. Our setting for model (fully connected network) is also very different from that considered in Tan et al. [46][47], thus requiring very different analysis tools.

**Differentially private learning in high dimension.** Standard results for private empirical risk minimization [3][45] and private stochastic convex optimization [10][11][4] under $\ell_1$ and $\ell_2$ constraints suggest that there is an unavoidable factor $d$ in the empirical risk and population risk that depends on the model dimension. However, for unconstrained optimization, it is possible to go across the dimension-dependency in proving risk bounds for certain class of problems (such as generalized linear model [43]). Recently, there is a growing line of works that prove dimension-independent excess risk bounds for differentially private learning, by utilizing the low-rank structure of data features [43] or gradient matrices [30][35] in training. Several follow-up works [31][12] further explore techniques to enforce the low-rank property (via random projection) and boost privacy utility trade-off. However, all the works focus on investigating a general high-dimensional problem for private learning, rather than separating the study for different network choices such as structure, width, depth and initialization. On the contrary, our study focus on the fully connected neural network and its linearized variant, which enables us to prove more precise privacy utility trade-off bounds for these particular networks under overparameterization.

2 Problem and Methodology

We consider the following standard multi-class supervised learning setting. Let $\mathcal{D} = (z_1, \ldots, z_n)$ be a finite input dataset of size $n$, where each input data record $z_i = (x_i, y_i)$ contains a $d$-dimensional input feature vector $x_i \in \mathbb{R}^d$ and a label vector $y_i \in \mathcal{Y} = \{0, 1\}^o$ on $o$ possible classes. The goal of learning is to learn a neural network output function $f_W(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by $W$ that achieves high prediction performance on the training dataset $\mathcal{D}$. Formally, we consider the empirical risk learning objective:

$$
\min_{W} \mathcal{L}(W; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \ell(f_W(x_i); y_i),
$$

where $\ell(f_W(x_i); y_i)$ is a loss function that reflects the approximation quality of model prediction $f_W(x_i)$ compared to the ground truth label $y_i$. For simplicity, throughout our analysis, we assume that the cross-entropy loss $\ell(f_W(x); y) = -\langle y, \log \text{softmax}(f_W(x)) \rangle$ is used for multi-class $o \geq 2$ network, and $\ell(f_W(x); y) = \log(1 + \exp(-y f_W(x)))$ is used for single-output network.

**Fully Connected Network.** Our investigation is based on the multi-output, fully connected, deep neural network (DNN) with ReLU activation. Denote the depth as $L$ and the width of hidden layer $l$.
We now give the definition for KL privacy, which is a more relaxed, yet closely connected privacy

We aim to understand the relation between privacy, utility and over-parameterization (depth and width) for models trained with (stochastic) gradient descent. As a first order Taylor expansion of the fully connected ReLU network at initialization parameter $W_0$, as follows.

$$
W_{lin} = W_{0,lin}(x) + \frac{\partial f_W(x)}{\partial W} \bigg|_{W=W_{0,lin}} (W - W_{lin})
$$

where $W_{lin}$ is the function output of the fully connected ReLU network at initialization $W_0$.

Langevin Diffusion. In terms of optimization algorithm, we focus on the Langevin diffusion algorithm [44] with per-dimensional noise variance $\sigma^2$. Note that we aim to avoid gradient clipping while still proving KL privacy bounds. After initializing the model parameters $W_0$ at time zero, the model parameters $W_t$ at subsequent time $t$ evolves as the below stochastic differential equation.

$$
dW_t = -\nabla L(W_t; \mathcal{D}) dt + \sqrt{2\sigma^2} dB_t.
$$

Initialization Distribution. The initialization of parameters $W_0$ crucially affects the convergence of Langevin diffusion, as observed in prior literatures [30, 23, 22]. In this work, we investigate the following general class of Gaussian initialization distribution with (possibly depth-dependent) variance for the parameters in each layer. For any layer $l = 1, \cdots, L$, we have that

$$
[W']_{ij} \sim \mathcal{N}(0, \beta_i), \text{for } (i, j) \in [m_l] \times [m_{l-1}]
$$

where $\beta_1, \cdots, \beta_L > 0$ are the per-layer variance for Gaussian initialization. By choosing different variance, we recover many common initialization schemes in the literature, as summarized in Table.

2.1 Our objective and methodology

We aim to understand the relation between privacy, utility and over-parameterization (depth and width) for the Langevin diffusion algorithm (under different initialization distributions). To understand privacy, we prove a KL privacy bound for running Langevin diffusion on any two worst-case neighboring datasets. Below we first give the definition for neighboring datasets.

Definition 2.1. We denote $\mathcal{D}$ and $\mathcal{D}'$ as neighboring datasets if they are of same size and only differ in one record. For brevity, we also denote the differing records as $(x, y) \in \mathcal{D}$ and $(x', y') \in \mathcal{D}'$.

Assumption 2.2 (Bounded Data). For simplicity, we assume that all records $x$ in the data domain is bounded s.t. $\|x\|_2 \leq 1$.

We now give the definition for KL privacy, which is a more relaxed, yet closely connected privacy notion to the standard $(\varepsilon, \delta)$ differential privacy [20]. (See Appendix A.2 for more discussions.) KL privacy and relaxed variants of it are commonly used in previous literature [7, 9, 51].

Definition 2.3 (KL privacy). A randomized algorithm $A$ satisfies $\varepsilon$-KL privacy if for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, we have that the KL divergence $KL(A(\mathcal{D}) \| A(\mathcal{D}')) \leq \varepsilon$, where $A(\mathcal{D})$ denotes the algorithm’s output distribution on dataset $\mathcal{D}$.
In this paper, we prove KL privacy upper bound for \( \max_{D, D'} KL(W_{[0:T]} \| W'_{[0:T]}) \) when running Langevin diffusion on any worst-case neighboring datasets. For brevity, here (and in the remaining paper), we denote \( W_{[0:T]} \) and \( W'_{[0:T]} \) as the trajectory of model parameters throughout the Langevin diffusion processes Eq. (4) with time \( T \) on \( D \) and \( D' \) respectively.

For utility analysis, we prove upper bound for excess empirical risk given any fixed KL divergence privacy budget. We additionally make the following fair and attainable assumption on data and network regularity (which is only required for utility analysis and not needed for our privacy bound).

**Assumption 2.4** (Data and network regularity [38, Assumption 2.1]). For any training data \( x_i \in D \), it satisfies that \( \|x_i\|_2 = 1 \). Moreover, \( x_i \in D \) are i.i.d. samples from a data distribution \( P_x \) that satisfies \( \int \|x\|_2^2 dP_x(x) = 1 \). We also assume that the network only has single output.

Our ultimate goal is to precisely understand how the KL privacy bound and the excess empirical risk bound (given fixed KL privacy budget) are affected by overparameterization (i.e., increasing width and depth) under different initialization distributions.

### 3 KL Privacy for Training Fully Connected ReLU Neural Networks

In this section, we perform the composition-based KL privacy analysis for Langevin Diffusion given random Gaussian initialization distribution Eq. (5), under fully connected ReLU network. More specifically, we prove upper bound for the KL divergence between distribution of output model parameters when running Langevin diffusion on an arbitrary pair of neighboring datasets \( D \) and \( D' \).

Our first key insight is that by a Bayes rule decomposition for density function in KL divergence, it is possible to prove KL privacy under a relaxed gradient sensitivity condition (that could hold without gradient clipping).

**Theorem 3.1** (KL composition under possibly unbounded gradient difference). The KL divergence between running Langevin diffusion for DNN on neighboring datasets \( D \) and \( D' \) satisfies

\[
KL(W_{[0:T]} \| W'_{[0:T]}) = \frac{1}{2\sigma^2} \int_0^T E \left[ \|\nabla \mathcal{L}(W_t; D) - \nabla \mathcal{L}(W_t; D')\|_2^2 \right] dt .
\]

**Proof sketch.** We compute the partial derivative of KL divergence with regard to time \( t \), and then integrate it over \( t \in [0, T] \) to compute the KL divergence during training with time \( T \). For computing the limit of differentiation, we use Girsanov’s theorem to compute the KL divergence between the trajectory of Langevin diffusion processes on \( D \) and \( D' \). The complete proof is in Appendix B.1.

Theorem 3.1 is an extension of the standard additivity [49] of KL divergence (also known as chain rule [44]) for a finite sequence of distributions to continuous time processes with (possibly) unbounded drift difference. The key extension is that Theorem 3.1 does not require bounded sensitivity between the drifts of Langevin Diffusion on neighboring datasets. Instead, it only requires finite second-order moment of drift difference (in the \( \ell_2 \)-norm sense) between neighboring datasets \( D, D' \). By using this extended KL composition Theorem 3.1, we prove KL privacy bound for running Langevin diffusion algorithm (without gradient clipping) on deep neural networks, by tracking the upper bound for \( \ell_2 \) norm of the gradient difference throughout training (under mild assumptions) as follows.

**Lemma 3.2** (Drift Difference in Noisy Training). Let \( M_T \) be the subspace spanned by gradients \( \{\nabla \ell(f_{W_t}(x_i); y_i) : (x_i, y_i) \in D, t \in [0, T]\} \) on each training data record throughout Langevin diffusion \( W_t \) on \( D \). Denote \( \|W_t\|_{M_T} \) as the \( \ell_2 \) norm of the projection of the input vector onto linear space \( M_T \). Suppose that \( \exists c, \beta > 0 \) s.t. for any \( W, W' \) and \( x, y \) we have \( \|\nabla \ell(f_W(x); y) - \nabla \ell(f_{W'}(x); y)\|_2 \leq \max\{c, \beta\} \|W - W'\|_{M_T} \). Then over the randomness of Brownian motion \( B_t \) and initialization distribution \( P_W \), it satisfies that

\[
\int_0^T E \left[ \|\nabla \mathcal{L}(W_t; D) - \nabla \mathcal{L}(W_t; D')\|_2^2 \right] dt \leq 2T \cdot E \left[ \|\nabla \mathcal{L}(W_0; D) - \nabla \mathcal{L}(W_0; D')\|_2^2 \right] + \frac{2\beta^2}{n^2(2 + \beta^2)} \left\{ \left( e^{(2+\beta^2)T} - 1 \right) T + \left( \frac{\sigma^2}{2} \text{rank}(M_T) + c^2 \right) T + \frac{2e^2T}{\beta^2} \right\} .
\]
**Proof sketch.** The key is to reduce the problem of upper bounding the gradient difference at any training time $T$, to analyzing its two subcomponents: $\|\nabla \ell(f_{W_t}(x); y)) - \nabla \ell(f_{W_0}(x^\prime); y^\prime))\|_2^2 \leq 2\|\nabla \ell(f_{W_0}(x); y)) - \nabla \ell(f_{W_0}(x^\prime); y^\prime))\|_2^2 + 2\|W_t - W_0\|_{M_T}^2 + 2\varepsilon^2$, where $(x, y)$ and $(x^\prime, y^\prime)$ are the differing data between neighboring datasets $D$ and $D^\prime$. This inequality is by the Cauchy-Schwartz inequality, and by the relaxed smoothness assumption (that is explained in Remark 3.3 in details). The complete proof is in Appendix B.2.

**Remark 3.3** (Relaxed smoothness of loss function). The assumption that $\|\nabla \ell(f_{W}(x); y)) - \nabla \ell(f_{W}(x^\prime); y^\prime))\|_2 \leq \max\{c, \beta\} \|W - W^\prime\|_{M_T}$ is similar to assuming smooth loss function, but is more relaxed as it allows non-smoothness at places where the gradient is bounded. Therefore, the assumption holds under ReLU activation.

**Remark 3.4** (Gradient difference at initialization). The first term and in our upper bound linearly scales with the difference between gradients on neighboring datasets $D$ and $D^\prime$ at initialization. Under different initialization distributions, this gradient difference exhibits different dependency on the network depth and width, as we will prove theoretically in Theorem 4.1.

**Remark 3.5** (Gradient difference fluctuation during training). The second term in Lemma 3.2 bounds the change of gradient difference during training, and is proportional to the the rank of a subspace $M_T$ spanned by gradients of all training data. Intuitively, this fluctuation is because Langevin diffusion adds per-dimensional noise with variance $\sigma^2$, thus perturbing the training parameters away from the initialization at a scale of $O(\sigma \sqrt{\text{rank}(M_T)})$ in expected $\ell_2$ distance. Note that this fluctuation term does not explicitly depend on the gradient difference between datasets $D$ and $D^\prime$. However, it upper bounds the gradient difference at time $T$, due to the relaxed smoothness assumption of loss function.

**Growth of KL privacy bound with increasing training time $T$.** The first and third terms in our upper bound Lemma 3.2 grow linearly with the training time $T$, while the second term grows exponentially with regard to $T$. Consequently, for learning tasks that requires a long training time to converge, the second term will become the dominating term and the KL privacy bound suffers from exponential growth with regard to the training time. Nevertheless, observe that for $T < \frac{1}{2+\beta}$, the second component in Lemma 3.2 contains a factor $\frac{(\beta^2 + \beta)^T}{2+\beta^2} - T$ that is sublinear to $T$. That is, for small training time, the second component is smaller than the first and the third components in Lemma 3.2 that linearly scale with $T$, and thus does not dominate the privacy bound. Intuitively, this phenomenon is related to lazy training \cite{18}. In Section 5 and Figure 2, we also numerically validate that the second component does not have a high effect on the KL privacy loss in the case of small training time.

**Dependence of KL privacy bound on network overparameterization.** Under a fixed training time $T$ and noise scale $\sigma^2$, Lemma 3.2 predicts that the KL divergence upper bound in Theorem 3.1 is dependent on the gradient difference and gradient norm at initialization, and the rank of gradient subspace $\text{rank}(M_T)$ throughout training. We now discuss the how these two terms change under increasing width and depth, and whether there are possibilities to improve them under overparameterization.

1. The gradient norm at initialization crucially depend on how the per-layer variance in the Gaussian initialization distribution scales with the network width and depth. Therefore, it is possible to improve the KL privacy bound by using initialization distributions that enable smaller gradient difference at initialization, as we will theoretically show in Section 4.

2. Regarding the rank of gradient subspace $\text{rank}(M_T)$; when the gradients along the training trajectory span the whole optimization space, $\text{rank}(M)$ would equal the dimension of the learning problem. Consequently, the gradient fluctuation upper bound (and thus the KL privacy bound) worsens with increasing number of model parameters (overparameterization) in the worst-case. However, if the gradients are low-dimensional \cite{43, 50, 41} or sparse \cite{55}, it is possible that $\text{rank}(M_T)$ will be dimension-independent and thus enable better bound for gradient fluctuation (and KL privacy bound). We leave this as an interesting open problem.

**4 KL privacy bound for Linearized Network under overparameterization**

In this section, we restrict ourselves to the training of linearized networks as described in \cite{3}, and investigate the interplay between KL privacy and overparameterization (increasing width and depth).
The analysis of DNN via linearization is a commonly used technique in both theory and practice. In theory, DNN can work in the lazy training regime \(^{13}\) (also called linear regime), under which the linearized network well approximates the training dynamics for deep neural networks \(^{33}\) and has been well studied by NTK. In practice, linearized network can still achieve decent performance, which provides a good justification of linearized networks. \(^{41,39}\). We hope our analysis for linearized network serve as an initial attempt that would open a door to theoretically understanding the relationship between overparameterization and privacy.

To derive a composition-based KL privacy bound for training linearized network, we apply Theorem \(^{33}\) which requires an upper bound for the norm of gradient difference between the training processes on neighboring datasets \(D\) and \(D'\) at any time \(t\). Note that the empirical risk function for training linearized models enjoys convexity, and therefore requires a relatively short amount of training time for convergence. Therefore intuitively, the gradient difference between neighboring datasets does not change a lot during training, thus allowing us to prove tighter upper bound for the gradient difference norm for linearized networks (than Lemma \(^{32}\)).

In the following theorem, we prove that for linearized network, the gradient difference throughout training has a uniform upper bound that only depends on the network width, depth and initialization.

**Theorem 4.1** (Gradient Difference throughout training linearized network). Under Assumption \(^{2.2}\) taking over the randomness of the random initialization and the Brownian motion, for any \(t \in [0, T]\), running Langevin diffusion on linearized network Eq. \(^{3}\) satisfies that

\[
\mathbb{E} \left[ \left\| \nabla \mathcal{L}(W^{lin}_t; D) - \mathcal{L}(W^{lin}_t; D') \right\|^2 \right] \leq \frac{4B}{n^2},
\]

where \(n\) is the training dataset size, and \(B\) is a constant that only depends on the network width, depth and initialization distribution as follows.

\[
B := o \left( \prod_{l=1}^{L-1} \frac{\beta_l m_l}{2} \right) \sum_{l=1}^{L} \frac{\beta_l}{\beta_l^m},
\]

where \(m\) is the number of output classes, \(\{m_l\}_{l=1}^{L}\) are the per-layer network widths, and \(\{\beta_l\}_{l=1}^{L}\) are the variances of Gaussian initialization at each layer.

Theorem \(^{4.1}\) provides an precise analytical upper bound for the gradient difference during training linearized network, by tracking the gradient distribution under fully connected feed-forward ReLU network with Gaussian weight matrices. The full proof is in Appendix C.1 and is heavily based on similar techniques for computing the gradient distribution in the NTK literature \(^{2,52}\). By plugging Eq. \(^{7}\) into Theorem \(^{3.1}\) we have the following KL privacy bound for training linearized network.

**Corollary 4.2** (KL privacy bound for training linearized network). Under Assumption \(^{2.2}\) and neural networks \(^{3}\) initialized by Gaussian distribution with per-layer variance \(\{\beta_l\}_{l=1}^{L}\), running Langevin diffusion for linearized network with time \(T\) on any neighboring datasets satisfies that

\[
KL(W^{lin}_{[0:T]} \| W^{lin}_{[0:T]}) \leq \frac{2BT}{n^2\sigma^2},
\]

where \(B\) is the constant that specifies the gradient norm upper bound, given by \(^{8}\).

**Overparameterization affects privacy differently under different initialization.** Corollary \(^{4.2}\) and Theorem \(^{4.1}\) proves that the effect of network overparameterization on KL privacy bound crucially relies on how the per-layer Gaussian initialization variance \(\beta_l\) is scaled with the per-layer network width \(m_l\) and depth \(L\). We summarize our KL privacy bound for linearized network under different width, depth and initialization schemes in Table \(^{1}\) and elaborate the comparison below.

1. **LeCun initialization** uses small, width-independent variance for initializing the first layer \(\beta_1 = \frac{1}{m}\) (where \(d\) is the number of input features), and width-dependent variance \(\beta_2 = \cdots = \beta_L = \frac{1}{m}\) for initializing all the subsequent layers. Therefore, the second term \(\sum_{l=1}^{L} \frac{\beta_l}{\beta_l^m}\) in the constant \(B\) \(^{8}\) increases linearly with the width \(m\) and depth \(L\). However, due to \(\frac{m_l}{2} < 1\) for all \(l = 2, \cdots, L\), the first product term \(\prod_{l=1}^{L} \frac{\beta_l m_l}{2}\) in constant \(B\) decays with the increasing depth. Therefore, by combining the two terms, we prove that the KL privacy bound worsens with increasing width, but improves with increasing depth (as long as the depth is large enough). Similarly, under Xavier
We now investigate the properties of numerically estimated KL divergence upper bound under different initialization distributions. In Figure 1, we show the growth of the KL privacy loss estimate (mean \( \mu \) and expected squared gradient norm during training. Therefore, we could numerically estimate the KL privacy loss by empirically averaging the squared gradient norm in multiple runs of the noisy GD algorithm. We consider the training datasets \( D \) as the subset of CIFAR-10 dataset that contains the ‘car’ and ‘plane’ labels. For neighboring dataset, we consider \( D' \) over all possible dataset constructed by randomly removing one training record from \( D \), or by adding one random record from test set with class ‘car’ or ‘plane’ to \( D \). (Note that this is different from the general neighboring notion we consider for the analysis in the paper, but our privacy bound Theorem 3.1 still holds under this add-or-remove-one neighboring notion).

We run noisy gradient descent with constant step-size 0.01 for 50 epochs on both datasets. We compute the mean and standard deviation of squared norm of gradient difference across 6 runs, and report the corresponding KL privacy loss computed by applying Theorem 3.1.

Numerical evidence for the growth of KL privacy bound with regard to training time. We evaluate over a specific setting of fully connected network with width 1024 and depth 10 under different initialization distributions. In Figure 1 we show the growth of the KL privacy loss estimate (mean in solid lines and standard deviation in shaded area) over the training process (as the number of epochs grows). We observe that the KL privacy loss grows at a close to linear rate at the beginning of training (< 10 epochs). Specifically, the KL privacy loss under LeCun and Xavier initialization distribution is close to zero at the beginning of training (< 10 epochs). This is due to the small per-layer variance for model parameters in LeCun and Xavier initialization, which contributes to small gradient norm at the beginning of training (i.e., the first term in the KL privacy bound Lemma 3.2).

However, as the number of epochs grow to > 10, the KL privacy loss grows faster than linear accumulation, reflecting that the exponential dependence of the second term in the KL privacy bound Lemma 3.2 on training time \( T \) is reasonable in this instance of practical training.

Numerical evidence for the relation between KL privacy bound and network overparameterization. We now investigate the properties of numerically estimated KL divergence upper bound under overparameterization, i.e., increasing network depth and width. We observe in Figure 2c that the numerically estimated KL divergence upper bound accumulates with increasing width and time.
Figure 2: Numerically estimated KL privacy loss during noisy gradient descent on fully connected ReLU network with increasing width and depth under different initialization. We observe that KL privacy loss grows with width under all evaluated initialization in Figure 2c. In terms of depth, at the beginning of training (20 epochs), KL privacy loss worsens (increases) with depth under He initialization, but first worsens (increases) with depth (≤ 8) and then improves (decreases) with depth (≥ 8) under Xavier and LeCun initializations. At later phases of the training (50 epochs), KL privacy worsens (increases) with depth under all evaluated initializations. This is consistent with Lemma 3.2 and suggests that by choosing appropriate initialization distributions (Xavier and LeCun) and reducing the number of total training epochs, it is possible to obtain KL privacy loss that improves (decreases) with increasing depth for large depth.

under all evaluated initialization distributions. This is consistent with empirical observations in the literature [47] that larger (i.e., wider) models suffer from higher privacy loss, given a fixed network depth. The relationship between KL privacy and network depth, however, is more complicated and highly depend on the initialization distributions and amount of training time. As we observe in Figure 2a and Figure 2b, only when the training time is small (20 epochs) and when the initialization distribution is LeCun or Xavier, it is possible to observe a numerical KL privacy loss that improves (decreases) with depth as long as the depth is large enough (> 8). These results suggest that the smaller per-layer variance in the LeCun and Xavier initialization distribution (compared to He initialization) contributes to smaller gradient norm at initialization, thus contributing to improved dependency of KL privacy loss on increasing depth. This is consistent with our discussion after Lemma 3.2 on the dependency of KL privacy bound for DNN on increasing width and depth, and validates our bound Theorem 4.1 for gradient difference norm at initialization. To this end, we validate that the choice of initialization distribution affects the dependency of KL privacy loss on increasing width and depth.

6 Utility guarantees for Training Linearized Network

Our privacy analysis suggests that training linearized network under certain initialization schemes (such as LeCun initialization) enable significantly better privacy bounds under overparameterization by increasing depth. In this section, we further prove utility bounds for Langevin diffusion under initialization schemes and investigate the effect of overparameterization on the privacy utility trade-off. In other words, we aim to understand whether there is any utility degradation for training linearized networks when using the more privacy-preserving initialization schemes.

Convergence of training linearized network. We now prove convergence of excess empirical risk in training linearized network via Langevin diffusion. This is a well-studied problem in the literature for noisy gradient descent. We extend the convergence theorem to continuous-time Langevin diffusion below and investigate factors that affect the convergence under overparameterization.

Lemma 6.1 (Extension of [40, Theorem 2] and [43, Theorem 3.1]). Let $L_{lin}^{(0)}(W; D)$ be the empirical risk function of linearized newtork Eq. (3) expanded at initialization vector $W_{lin}^{0}$. Let $W_{*}^{0}$ be an $\alpha$-near-optimal solution for the ERM problem such that $L_{lin}^{(0)}(W_{*}^{0}; D) = \min W L_{lin}^{(0)}(W; D) \leq \alpha$. Let $D = x_1, \cdots, x_n$ be an arbitrary training dataset of size $n$, and denote $M_{0} = (\nabla f_{W_{lin}^{0}}(x_1), \cdots, \nabla f_{W_{lin}^{0}}(x_n))^{T}$ as the NTK feature matrix at initialization. Then running
Langevin diffusion \[^4\] on \( L_0^{\text{lin}}(W) \) with time \( T \) and initialization vector \( W_0^{\text{lin}} \) satisfies
\[
\mathbb{E}[\ell_0^{\text{lin}}(\bar{W}_T)] - \min_W L_0^{\text{lin}}(W; \mathcal{D}) \leq \alpha + \frac{R}{2T} + \frac{1}{2} \sigma^2 \text{rank}(M_0)
\]
where the expectation is over Brownian motion \( B_T \) in Langevin diffusion Eq. (4). \( \bar{W}_T^{\text{lin}} = \frac{1}{T} \int W_t^{\text{lin}} dt \) is the average of all iterates, and \( R = \| W_0^{\text{lin}} - W^*_0 \|^2_{M_0} \) is the gap between initialization parameters \( W_0^{\text{lin}} \) and solution \( W^*_0 \).

**Remark 6.2.** The excess empirical risk bound Lemma 6.1 is smaller if data is low-rank, e.g., image data, then rank \( (M_0) \) is small. This is consistent with the prior dimension-independent private learning literature [30, 31, 35] and shows the benefit of low-dimensional gradients on private learning.

See Appendix D.1 for the full proof. Lemma 6.1 highlights that the excess empirical risk scales with the gap \( R \) between initialization and optima (which we refer as the lazy training distance), the rank of the gradient subspace rank \( (M_0) \), and the constant \( B \) that specifies upper bound for expected gradient norm during training. Specifically, the smaller the lazy training distance \( R \), the better the excess risk bound for Langevin diffusion given fixed training time \( T \) and noise variance \( \sigma^2 \). We have discussed how overparameterization affects the gradient norm constant \( B \) and the gradient subspace rank \( (M_0) \) in Section 4. Therefore, we only still need to investigate how the lazy training distance \( R \) changes with the network width, depth, and initialization, as follows.

**Lazy training distance \( R \) decreases with increasing depth.** It is widely observed in the literature [18, 53, 59] that under appropriate choices of initializations, gradient descent on fully connected neural network falls under a lazy training regime. That is, with high probability, there exists a (nearly) optimal solution for the ERM problem that is close to the initialization parameters in terms of \( \ell_2 \) norm. Moreover, this lazy training distance \( R \) is closely related to the smallest eigenvalue of the NTK matrix. In the following proposition, we compute a near-optimal solution via the pseudo inverse distance. Moreover, this lazy training distance \( R \) is closely related to the smallest eigenvalue of the NTK matrix [38].

**Lemma 6.3 (Bounding lazy training distance via smallest eigenvalue of the NTK matrix).** Suppose that the data and network regularity Assumption 2.2 holds. Let \( \xi_l \) be an auxiliary variable that takes value 1 is \( m_l = \tilde{O}(n) \), and zero otherwise with \( m_l \) being width for layer \( l \) and \( n \) being number of training data. Let \( L_0^{\text{lin}}(W) \) be the empirical risk function defined below Eq. (3), for linearized network expanded at initialization vector \( W_0^{\text{lin}} \). Then for any initialization vector \( W_0^{\text{lin}} \), there exists a corresponding near optimal solution \( W_0^{\text{lin}} \), such that \( L_0^{\text{lin}}(W_0^{\text{lin}}) \leq \frac{1}{\pi^*}, \text{rank}(M_0) = n \) and
\[
R \leq \tilde{O} \left( \max \left\{ \frac{1}{\beta_L \left( \prod_{l=1}^{L-1} \frac{m_l}{\beta_L} \right)^{\frac{1}{2}}} \right\} \frac{n}{d \sum_{l=1}^{L} \xi_l \beta_l^{-1}} \right),
\]
with high probability over random initialization Eq. (5), where \( \tilde{O} \) ignores logarithmic factors with regard to \( n, m, L, \) and tail probability \( \delta \).

The full proof is in Appendix D.2. We refer to \( R \) as the approximate lazy training distance because \( W_0^{\text{lin}} \) is only a near optimal solution for the ERM problem. By using Lemma 6.3, we provide a summary of bounds for \( R \) under different initializations in Table 1. We observe that the lazy training distance \( R \) decreases with increasing width and depth under LeCun, He and NTK initializations, while under Xavier initialization \( R \) only decreases with increasing depth.

**Privacy & Excess empirical risk tradeoff for Langevin diffusion on linearized network.** We now use the approximate lazy training distance \( R \) to prove empirical risk bound and combine it with our KL privacy bound Section 3 to show the privacy utility trade-off under overparameterization.

**Corollary 6.4 (Privacy utility tradeoff for training linearized network).** Assume that the data and network regularity Assumption 2.4 holds. Let \( B = \left( \prod_{l=1}^{L-1} \frac{m_l}{\beta_l} \right) \sum_{l=1}^{L-1} \frac{m_l}{\beta_l} \) be the gradient norm constant proved in Eq. (8), and let \( R \leq \tilde{O} \left( \max \left\{ \frac{1}{\beta_L \left( \prod_{l=1}^{L-1} \frac{m_l}{\beta_L} \right)^{\frac{1}{2}}} \right\} \frac{n}{d \sum_{l=1}^{L} \xi_l \beta_l^{-1}} \right) \) be the lazy training distance bound proved in Lemma 6.3. Then by setting \( \sigma^2 = \frac{2B}{\epsilon n^2} \) and \( T = \sqrt{\frac{cnR}{2B}} \), we have
that releasing all iterates of Langevin diffusion with time \( T \) satisfies \( \varepsilon \)-KL privacy, and has empirical excess risk upper bounded by

\[
\mathbb{E}[\mathcal{L}_0^{lin}(\bar{W}_T^{lin})] - \min_W \mathcal{L}_0^{lin}(W; \mathcal{D}) \leq \hat{O} \left( \frac{1}{n^2} + \sqrt{\frac{BR}{\varepsilon n}} \right)
\]  

with high probability over random initialization Eq. (5), where the expectation is over Brownian motion \( B_T \) in Langevin diffusion Eq. (4), and \( \hat{O} \) ignores logarithmic factors with regard to width \( m \), depth \( L \), number of training data \( n \) and tail probability \( \delta \).

Specifically, under large data dimension \( d = \tilde{\Omega}(n) \) and hidden-layer width \( m = \tilde{\Omega}(n) \), we have

\[
\mathbb{E}[\mathcal{L}_0^{lin}(\bar{W}_T^{lin})] - \min_W \mathcal{L}_0^{lin}(W; \mathcal{D}) \leq \hat{O} \left( \frac{1}{n^2} + \sqrt{\frac{1}{2^{L-1}d_\varepsilon}} \max\{1, \beta_L \prod_{l=1}^{L-1} \beta_l m_l\} \right)
\]  

A summary of our excess empirical risk bounds under different initializations is in Table 1.

See Appendix D.3 for the full proof. Corollary 6.4 proves that the excess empirical risk worsens in the presence of a stronger privacy constraint, i.e., a small privacy budget \( \varepsilon \), thus contributing to a trade-off between privacy and utility. However, the excess empirical risk also scales with the approximate lazy training distance \( R \) and the gradient norm constant \( B \). These constants depend on network width, depth and initialization distributions, and therefore we prove privacy-utility trade-offs for training linearized network that changes with overparameterization, as summarized in Table 1.

We would like to highlight that our privacy utility trade-off bound under LeCun and Xavier initialization strictly improves with increasing depth as long as the data and network satisfy regularity Assumption 2.4 and the hidden-layer width is large enough. To our best knowledge, this is the first time that a strictly improving privacy utility trade-off under overparameterization is shown in literature. This shows the benefits of precisely bounding the gradient norm in our KL privacy analysis for overparameterized models.

### 7 Conclusion

We prove new KL privacy bound for training fully connected ReLu network (and its linearized variant) using the Langevin diffusion algorithm, and investigate how privacy is affected by the network width, depth and initialization. Our results suggest that there is a complex interplay between privacy and overparameterization (width and depth) that crucially relies on what initialization distribution is used and how much the gradient fluctuates during training. To this end, we show that for training a linearized variant of fully connected network with finite time, it is possible to prove a KL privacy bound that improves with depth, as long as the initialization distribution is set appropriately (such as LeCun and Xavier). We also study the excess empirical and population risk bounds for linearized network, and prove that the privacy-utility trade-off similarly improves as depth increases under LeCun and Xavier initialization. This shows the gain of our new privacy analysis for capturing the effect of overparameterization. We leave it as an important open problem as to whether our privacy utility trade-off results for linearized network could be generalized to deep neural networks.

### References


## Contents

1 Introduction
   1.1 Related Works ................................................. 3

2 Problem and Methodology
   2.1 Our objective and methodology .................................. 4

3 KL Privacy for Training Fully Connected ReLU Neural Networks ................................................. 5

4 KL privacy bound for Linearized Network under overparameterization ................................................. 6

5 Numerical validation of our KL privacy bounds ................................................. 8

6 Utility guarantees for Training Linearized Network ................................................. 9

7 Conclusion ................................................. 11

A Symbols and definitions
   A.1 Additional notations ................................................. 15
   A.2 Relation between KL privacy Definition 2.3 to differential privacy ................................................. 16

B Deferred proofs for Section 3 ................................................. 16
   B.1 Deferred proofs for Theorem 3.1 ................................................. 16
   B.2 Deferred proofs for Lemma 3.2 ................................................. 18

C Deferred proofs for Section 4 ................................................. 19
   C.1 Bounding the gradient norm at initialization ................................................. 19
   C.2 Deferred proof for Theorem 4.1 ................................................. 22

D Deferred proofs for Section 6 ................................................. 22
   D.1 Deferred proof for Lemma 6.1 on convergence of training linearized network ................................................. 22
   D.2 Deferred proof for Lemma 6.3 ................................................. 23
   D.3 Deferred proof for Corollary 6.4 ................................................. 26

E Discussion on extending our results to Noisy GD with constant step-size ................................................. 27

A Symbols and definitions

A.1 Additional notations ................................................. 15

Vecorization: Vec(·) denotes the transformation that takes an input matrix \( A = (a_{ij})_{i \in [r], j \in [c]} \in \mathbb{R}^{r \times c} \) (with \( r \) rows and \( c \) columns) and outputs a \( rc \)-dimensional column vector: \( \text{Vec}(A) = (a_{1,1}, \ldots, a_{r,1}, a_{1,2}, \ldots, a_{r,2}, \ldots, a_{1,c}, \ldots, a_{r,c})^T \).

Softmax function: \( \text{softmax}(y) = \frac{e^{y[j]}}{\sum_{j=1}^o e^{y[j]}} \) where \( o \) is the number of output classes.

\( o \): number of output classes for the neural network.
A.2 Relation between KL privacy Definition 2.3 to differential privacy

KL privacy is a more relaxed, yet closely connected privacy notion to \((\varepsilon, \delta)\) differential privacy [20].

1. KL privacy and differential privacy are both worst-case privacy notions over all possible neighboring datasets, by requiring bounded distinguishability between the algorithm’s output distributions on neighboring datasets in statistical divergence. The difference is that KL privacy requires bounded KL divergence, while \((\varepsilon, \delta)\)-differential privacy is equivalent to bounded Hockey-stick divergence [5].

2. KL privacy and differential privacy are both definitions based on the privacy loss random variable \(\log \frac{P(A(D))}{P(A(D'))} \sim A(D)\) (following the definition in [11, Equation 1]). KL privacy implies that the privacy loss random variable has a bounded first order moment, while differential privacy requires a high probability argument that the privacy loss random variable is bounded by \(\varepsilon\) with probability \(1 - \delta\). Therefore, KL privacy is generally a more relaxed notion than differential privacy.

3. Translation to each other: For \(\varepsilon = 0\), KL privacy (bounded first-order moment of privacy loss random variable) implies \((0, \delta)\)-differential privacy with \(\delta = \sqrt{\frac{2}{\varepsilon}}\) by Pinsker inequality. Higher order moments of the privacy loss random variable suffice to prove \((\varepsilon, \delta)\)-differential privacy for \(\varepsilon > 0\). Note that \((\varepsilon, \delta)\)-DP with \(\delta > 0\) does not necessarily imply KL privacy, as the privacy loss random variable may be large at the tail event with \(\delta\) probability.

4. Due to the connection to the privacy loss random variable (which is closely connected to the likelihood ratio test for membership hypothesis testing), both differential privacy and KL privacy incur upper bound on the performance curve of inference attacks, such as the membership inference and attribute inference [17, 26], as we discuss in Footnote 2.

B Deferred proofs for Section 3

B.1 Deferred proofs for Theorem 3.1

To prove the new composition theorem, we will use the Girsanov’s Theorem. Here we follow the presentation of [17, Theorem 6].

Theorem B.1 (Implication of Girsanov’s theorem [17, Theorem 6]). Let \((X_t)_{t \in [0, \eta]}\) and \((\tilde{X}_t)_{t \in [0, \eta]}\) be two continuous-time processes over \(\mathbb{R}^r\). Let \(P_T\) be the probability measure that corresponds to the trajectory of \((X_t)_{t \in [0, \eta]}\), and let \(Q_T\) be the probability measure that corresponds to the trajectory of \((\tilde{X}_t)_{t \in [0, \eta]}\). Suppose that the process \((X_t)_{t \in [0, \eta]}\) follows

\[dX_t = b_t dt + \sigma dB_t,\]

where \((B_t)_{t \in [0, T]}\) is a standard Brownian motion over \(P_T\), and the process \((\tilde{X}_t)_{t \in [0, \eta]}\) follows

\[d\tilde{X}_t = b_t dt + \sigma d\tilde{B}_t,\]

where \((\tilde{B}_t)_{t \in [0, T]}\) is a standard Brownian motion over \(Q_T\) with \(d\tilde{B}_t = dB_t + \frac{1}{2} \left( b_t - \tilde{b}_t \right) \). Assume that \(\sigma\) is a \(r \times r\) symmetric positive definite matrix. Then, provided that Novikov’s condition holds,

\[\mathbb{E}_{Q_T} \exp \left( \frac{1}{2} \int_0^\eta \| \sigma^{-1} (b_t - \tilde{b}_t) \|^2 dt \right) < \infty, \tag{13}\]

we have that

\[\frac{dP_T}{dQ_T} = \exp \left( \int_0^\eta \sigma^{-1} (b_t - \tilde{b}_t) dB_t - \frac{1}{2} \int_0^\eta \| \sigma^{-1} (b_t - \tilde{b}_t) \|^2 dt \right).\]

We are now ready to apply Girsanov’s theorem to prove the following new KL privacy composition theorem for Langevin diffusion processes on neighboring datasets \(D\) and \(D'\).
With time $t$, and $W_t$ denotes the model parameter after running Langevin diffusion on dataset $D$ with time $t$. By definition of partial derivative, we have that

$$\frac{\partial KL(W_{[0:t]}||W_{[0:t]})}{\partial t} = \lim_{\eta \to 0} \frac{KL(W_{[0:t+\eta]}||W'_{[0:t+\eta]}) - KL(W_{[0:t]}||W'_{[0:t]})}{\eta}$$

Now we compute the term $KL(W_{[0:t+\eta]}||W'_{[0:t+\eta]})$ as follows.

$$KL(W_{[0:t+\eta]}||W'_{[0:t+\eta]}) = E_{w_{[0:t+\eta]} \sim p_{[0:t+\eta]}} \left[ \log \left( \frac{p_{[t:t+\eta]}(w_{[0:t]})}{p'_{[t:t+\eta]}(w_{[0:t]})} \right) \right]$$

where $p_{[t:t+\eta]}(w_{[0:t]})$ is the conditional distribution of model parameter $W_{[t:t+\eta]}$ during running Langevin diffusion on dataset $D$, given that $W_{[0:t]} = w_{[0:t]}$. Similarly, $p'_{[t:t+\eta]}(w_{[0:t]})$ is the conditional distribution during running Langevin diffusion on dataset $D'$.

Therefore, by using the Markov property of the Langevin diffusion process and the definition of KL divergence in Eq. (16), we have that

$$KL(W_{[0:t+\eta]}||W'_{[0:t+\eta]}) = E_{w_{[0:t+\eta]} \sim p_t} \left[ KL(p_{[t:t+\eta]}(\cdot || w_t)) \right] + KL(p_t, p'_{t})$$

Now to compute the term $KL(p_{[t:t+\eta]}(\cdot || w_t))$, we only need to apply the implication Theorem [1] of Girsanov’s theorem. Specifically, we apply Theorem [1] to the following two Langevin diffusion processes $(W_{t+s}|t \in [0, \eta])$ and $(W'_{t+s}|t \in [0, \eta])$, conditioning on the observation $W_{t|t} = W'_{t|t} = w_t$ at time $t$.

$$dW_{t+s} = -\nabla L(W_{t+s}|t; D) dt + \sqrt{2\sigma^2} dB_s$$

$$dW'_{t+s} = -\nabla L(W'_{t+s}|t; D') dt + \sqrt{2\sigma^2} dB_s$$

Note that when $\eta$ is small enough, we have that the Novikov’s condition in Eq. (15) holds because the exponent inside integration $\frac{1}{2} \int_0^\eta \|\sigma_s^{-1}(b_s - \tilde{b}_s)\|^2 d\sigma$ scales linearly with $\eta$ that can be arbitrarily small. Therefore, by applying Girsanov’s theorem, we have that

$$KL(p_{[t:t+\eta]}(\cdot || w_t)) = E \left[ \int_0^\eta \sigma_s^{-1}(b_s - \tilde{b}_s) dB_s - \frac{1}{2} \int_0^\eta \|\sigma_s^{-1}(b_s - \tilde{b}_s)\|^2 d\sigma \right]$$

where $b_s - \tilde{b}_s = -\nabla L(W_{t+s}|t; D) + \nabla L(W_{t+s}|t; D')$. By $dB_s = dB_s + \frac{1}{\sigma}(b_s - \tilde{b}_s)$ and Itô integration with regard to $W_{[t:t+\eta]}$, we have that

$$KL(p_{[t:t+\eta]}(\cdot || w_t)) = \frac{E \left[ \int_0^\eta \|\nabla L(W_{t+s}|t; D) - \nabla L(W_{t+s}|t; D')\|^2 d\sigma \right]}{2\sigma^2} + KL(p_t, p'_{t})$$

By plugging Eq. (15) into Eq. (17), we have that

$$KL(p_{[t:t+\eta]}, p'_{[t:t+\eta]}) = \frac{E \left[ \int_0^\eta \|\nabla L(W_{t+s}|t; D) - \nabla L(W_{t+s}|t; D')\|^2 d\sigma \right]}{2\sigma^2} + KL(p_t, p'_{t})$$
By plugging Eq. (19) into Eq. (15), and by exchanging the order of expectation and integration, we have that
\[
\frac{\partial K(p_t, p_t')}{\partial t} = \frac{1}{2\sigma^2} \lim_{\eta \to 0} \int_0^T \mathbb{E}_{p_t+s} \left[ \| \nabla \mathcal{L}(W_{t+s}; D) - \nabla \mathcal{L}(W_{t+s}; D') \|^2_2 \right] \, ds
\]
\[
= \frac{1}{2\sigma^2} \mathbb{E}_{p_t} \left[ \| \nabla \mathcal{L}(W_t; D) - \nabla \mathcal{L}(W_t; D') \|^2_2 \right]
\]
(20)
Integrating Eq. (20) on \( t \in [0, T] \) finishes the proof.

\[\Box\]

### B.2 Deferred proofs for Lemma 3.2

**Lemma 3.2** Let \( M_T \) be the subspace spanned by gradients \( \{ \nabla \ell(f_{W_t}(x; y)) : (x, y) \in D, t \in [0, T] \} \) on each training data record throughout Langevin diffusion \( \{ W_t \}_{t \in [0, T]} \). Denote \( \| \cdot \|_{M_T} \) as the \( \ell_2 \) norm of the projection of the input vector onto linear space \( M_T \). Suppose that \( 2c, \beta > 0 \) s.t. for any \( W, W' \) and \( x, y \) we have \( \| \nabla \ell(f_{W}(x; y)) - \nabla \ell(f_{W'}(x; y)) \|_2 \leq \max \{ c, \beta \} \| W - W' \|_{M_T} \).

Then over the randomness of Brownian motion \( B_t \) and initialization distribution \( \mathcal{N} \), it satisfies that
\[
\int_0^T \mathbb{E} \left[ \| \nabla \mathcal{L}(W_t; D) - \nabla \mathcal{L}(W_t; D') \|^2_2 \right] \, dt \leq 2T \cdot \mathbb{E} \left[ \| \nabla \mathcal{L}(W_0; D) - \nabla \mathcal{L}(W_0; D') \|^2_2 \right]
\]
\[
+ \frac{2(2 + \beta^2)}{n^2(2 + \beta^2)} \left( e \left( 2 + \beta^2 \right) - 1 - T \right) \cdot \left( \mathbb{E} \left[ \| \nabla \mathcal{L}(W_0; D) \|_2^2 \right] + \mathbb{E} \left[ \| \nabla \mathcal{L}(W_0; D') \|_2^2 \right] \right) + \frac{2c^2 T}{n^2} \cdot \mathbb{I}
\]

**Proof.** By definition of the neighboring datasets \( D \) and \( D' \), and the definition of empirical risk Eq. (1), we have that for any \( W \), it satisfies that
\[
\| \nabla \mathcal{L}(W; D) - \nabla \mathcal{L}(W; D') \|^2_2 = \frac{1}{n^2} \| \nabla \ell(f_{W}(x; y)) - \nabla \ell(f_{W'}(x; y)) \|^2_2,
\]
(21)
where \( (x, y) \) and \( (x', y') \) are the differing records between neighboring datasets \( D \) and \( D' \). By the assumption that \( \| \nabla \ell(f_{W}(x; y)) - \nabla \ell(f_{W'}(x; y)) \|_2 \leq \max \{ c, \beta \} \| W - W' \|_{M_T} \), and by the Cauchy-Schwarz inequality, we further have that for any \( W \) and \( W_t \), it satisfies that
\[
\| \nabla \ell(f_{W_t}(x; y)) - \nabla \ell(f_{W_t}(x'; y')) \|^2_2 \leq 2 \| \nabla \ell(f_{W_0}(x; y)) - \nabla \ell(f_{W_0}(x'; y')) \|^2_2
\]
\[
+ 2\beta^2 \| W_t - W_0 \|^2_{M_T} + 2c^2.
\]
(22)
The first term \( \| \nabla \ell(f_{W_t}(x; y)) - \nabla \ell(f_{W_0}(x; y')) \|^2_2 \) is constant during training (as it only depends on the initialization). Therefore, we only need to bound the second term \( \| W_t - W_0 \|^2_{M_T} \). For brevity, we denote the function \( d(W) = \| W - W_0 \|^2_{M_T} \). For clarity, we further denote \( p_t \) as the distribution of model parameters after running Langevin diffusion on dataset \( D \) with time \( t \), and similarly denote \( p_t' \) as the distribution of model parameters after running Langevin diffusion on dataset \( D' \) with time \( t \). Then by definition we have that
\[
\frac{\partial}{\partial t} \mathbb{E}_{p_t}[d(W)] = \lim_{\eta \to 0} \frac{\mathbb{E}_{p_{t+\eta}}[d(W)] - \mathbb{E}_{p_t}[d(W)]}{\eta}.
\]
(23)
Denote \( \Gamma_s \) as the following random operator on model parameters \( \theta \).
\[
\Gamma_s(W) = \theta - s \nabla \mathcal{L}(W; D) + \sqrt{2\sigma^2 s}Z
\]
where \( Z \sim \mathcal{N}(0, \mathbb{I}) \). We first claim that the following equation holds.
\[
\lim_{\eta \to 0} \frac{\mathbb{E}_{p_{t+\eta}}[d(W)] - \mathbb{E}_{p_t}[d(\Gamma_\eta(W))]}{\eta} = 0
\]
(24)
This is by using Euler-Maruyama discretization method to approximate the solution \( W_t \) of SDE Eq. (4). More specifically, the approximation error \( \mathbb{E}_{p_{t+\eta}}[d(W)] - \mathbb{E}_{p_t}[d(\Gamma_\eta(W))] \) is of size \( O(\eta^2 r^2) \) for small \( \eta \), where \( r \) is the dimension of \( W \).
Therefore, by plugging Eq. (24) into Eq. (23), we have that
\[
\frac{\partial}{\partial t} \mathbb{E}_{p_t}[d(W)] = \lim_{\eta \to 0} \frac{\mathbb{E}_{p_t}[d(\Gamma_\eta(W))] - \mathbb{E}_{p_t}[d(W)]}{\eta}
\]
Recall that $\nabla^2 d(W)$ exists almost everywhere with regard to $W \sim p_t$. Therefore we could approximate the term $\mathbb{E}_{p_t}[d(\Gamma_\eta(W)); D, D')$ via its second-order Taylor expansion at $W$ as follows.
\[
\frac{\partial}{\partial t} \mathbb{E}_{p_t}[d(W)] = \lim_{\eta \to 0} \frac{\mathbb{E}_{p_t}[(\nabla^2 d(W), -\nabla \nabla L(W); D) + \sqrt{2\sigma^2}\eta Z] + \sigma^2 Z^T \nabla^2 d(W) Z + o(\eta)]}{\eta}
\]
By plugging the assumption into Eq. (22) and Eq. (21), followed by integration ver time $t$ we have that
\[
\frac{\partial}{\partial t} \mathbb{E}_{p_t}[\|W - W_0\|^2_{M_T}] \leq 2\mathbb{E}_{p_t}[(\|W - W_0, \nabla L(W); D)] + \sigma^2 \text{rank}(M_T)
\]
By solving the above ordinary differential inequality on $t \in [0, T]$, we have that
\[
\mathbb{E}_{p_t}[\|W - W_0\|^2_{M_T}] \leq \frac{(2 + \beta^2)\mathbb{E}_{p_t}[\|W - W_0\|^2_{M_T}] + \mathbb{E}[\|\nabla L(W_0; D)\|^2] + \sigma^2 \text{rank}(M_T) + c^2}{2 + \beta^2} - T
\]
By plugging Eq. (30) into Eq. (22) and Eq. (21), followed by integration ver time $t \in [0, T]$, we have that
\[
\begin{align*}
\int_0^T \mathbb{E}_{p_t} \left[\|\nabla L(W; D) - \nabla L(W; D')\|^2\right] dt &\leq 2T \cdot \mathbb{E}_{p_0} \left[\|\nabla L(W_0; D) - \nabla L(W; D')\|^2\right] \\
&\quad + \frac{2\beta^2}{n^2(2 + \beta^2)} \left(\frac{(2 + \beta^2)T - 1}{2 + \beta^2} - T\right) \cdot \left(\mathbb{E}_{p_0} \left[\|\nabla L(W; D)\|^2\right] + \sigma^2 \text{rank}(M_T) + c^2\right) + \frac{2\epsilon^2 T}{n^2}.
\end{align*}
\]
This suffice to prove the statement.

\[\square\]

C \hspace{1cm} Deferred proofs for Section 4

C.1 Bounding the gradient norm at initialization

To bound the moment of $\ell_2$ norm of the gradient $\frac{\partial f(x)}{\partial W}$ of network output function, we need the following (extended) lemmas from Zhu et al. [52].

Lemma C.1 ([52 Lemma 1]). Let $w \sim \mathcal{N}(0, \sigma^2 I_n)$, then for two fixed non-zero vectors $h_1, h_2 \in \mathbb{R}^n$ whose correlation is unknown, define two random variables $X = (w^T h_1 | w^T h_2 > 0)^2$ and $Y = s(w^T h_1)^2$, where $s \sim Ber(1, 1/2)$ follows a Bernoulli distribution with 1 trial and $\frac{1}{2}$ success rate, and $s$ and $w$ are independent random variables. Then $X$ and $Y$ have the same distribution.
Lemma C.2 (Extension of [52, Lemma 2]). Given a fixed non-zero matrix $H_1 \in \mathbb{R}^{p \times r}$ and a fixed non-zero vector $h_2 \in \mathbb{R}^p$ and let $W \in \mathbb{R}^{q \times p}$ be a random matrix with i.i.d. entries $W_{ij} \sim \mathcal{N}(0, \beta)$ and a matrix (or vector) $V = \phi'(Wh_2)WH_1 \in \mathbb{R}^{q \times r}$, then, we have $\mathbb{E}[\|V\|_F^2] = \frac{q \beta}{2}$.

Proof. According to the definition of $V = \phi'(Wh_2)WH_1 \in \mathbb{R}^{q \times r}$, we have:

$$\|V\|_F^2 = \sum_{i=1}^{q} \sum_{j=1}^{r} \left( D_{i,j}(Wh[i], H_1^{[j]}) \right)^2,$$

where $D_{i,j} = 1_{\{(Wh[i], H_1^{[j]}) \geq 0\}}$, $Wh[i]$ is the $i$-th row of $W$, and $H_1^{[j]}$ is the $j$-th column vector of $H_1$. Therefore, by Lemma C.1 with i.i.d. Bernoulli random variable $\rho_1, \cdots, \rho_q \sim \text{Ber}(1, 1/2)$, we have

$$\|V\|_F^2 = \sum_{i=1}^{q} \sum_{j=1}^{r} \rho_i \|H_1^{[j]}\|_2^2 \tilde{w}_{ij}^2.$$

where $\tilde{w}_{ij} = \langle Wh[i], H_1^{[j]} \rangle / (\sqrt{\beta\|H_1^{[j]}\|_2^2})$. By the fact that $Wh[i]$ has i.i.d. Gaussian entries, for any fixed $j$, we have that $\tilde{w}_{ij} \sim \mathcal{N}(0, 1), i = 1, \cdots, q$ independently. Therefore, we have

$$\mathbb{E}[\|V\|_F^2] = \sum_{i=1}^{q} \sum_{j=1}^{r} \mathbb{E}[\rho_i \|H_1^{[j]}\|_2^2 \mathbb{E}[\tilde{w}_{ij}^2] = \frac{q \beta}{2} \mathbb{E}[\|H_1\|_F^2].$$

Now, we are ready to prove output gradient expectation at random initialization as follows.

Lemma C.3 (Output Gradient Expectation Bound at Random Initialization). Fix any data record $x$, then over the randomness of the initialization distributions for $W_1, \cdots, W_L$, i.e., $W_l \sim \mathcal{N}(0, \beta I)$ for $l = 1, \cdots, L - 1$, it satisfies that

$$\mathbb{E}_W \left[ \|\partial f(x) / \partial \text{Vec}(W_l)\|_F^2 \right] = \|x\|_2^2 \sigma \left( \prod_{i=1}^{L-1} \beta_i m_i / 2 \right) \sum_{i=1}^{L} \beta_i / \beta_i.$$

Proof. We use $\text{Vec}(W_l)$ to denote the concatenation of all row vector of the parameter matrix $W_l$. By chain rule, for $l = 1, \cdots, L - 1$, we have that

$$\frac{\partial f(x)}{\partial \text{Vec}(W_l)} = \frac{\partial h_L(x)}{\partial h_{L-1}(x)} \left( \prod_{i=1}^{L-1} \frac{\partial h_{L-1}(x)}{\partial h_{L-1-i}(x)} \right) \frac{\partial h_{L-1}}{\text{Vec}(W_l)}$$

$$= W_L \left( \prod_{i=1}^{L-1} \sigma_{L-1} W_{L-i} \right) \sigma_L' \left( \begin{array}{c} h_{L-1}^T \cdots 0 \\ \vdots \\ 0 \cdots h_{L-1}^T \end{array} \right)_{m \times m \times m_{L-1}}.$$  

Similarly, for the $L$-th layer, we have that

$$\frac{\partial f(x)}{\partial \text{Vec}(W_L)} = \left( \begin{array}{ccc} h_{L-1}^T & 0 & \cdots \\ \vdots & \vdots & \vdots \\ 0 & h_{L-1}^T \end{array} \right)_{o \times o \times m_{L-1}}.$$  

By properties of ReLU activation $\phi$, we have $\phi'_L - 1 = \text{diag}[\text{sgn}(W_{L-1}h_{L-1-i})]$, where $\text{sgn}(x) = \{1, x > 0; 0, x \leq 0\}$ operates coordinate-wise with regard to the input matrix. Therefore, we have that for $l = 1, \cdots, L - 1$

$$\frac{\partial f(x)}{\partial \text{Vec}(W_l)} = W_L \left( \prod_{i=1}^{L-1} \text{diag}[\text{sgn}(W_{L-1}h_{L-1-i})] W_{L-i} \right) \cdot \text{diag}[\text{sgn}(W_l h_{L-1})] \left( \begin{array}{ccc} h_{L-1}^T & 0 & \cdots \\ \vdots & \vdots & \vdots \\ 0 & h_{L-1}^T \end{array} \right)_{m \times m \times m_{L-1}}.$$  

20
For notational simplicity, we introduce the notation of $t_{l'}^t$ for $l = 1, \cdots, L - 1$ and $l \leq l' < L$ as follows.

$$t_{l'}^t := \left( \prod_{i=L-l'}^{L-1} \text{diag}(\text{sgn}(W_{L-i} h_{L-1-i})) W_{L-i} \right) \cdot \text{diag}(\text{sgn}(W_{l} h_{l-1})) \begin{pmatrix} h_{l-1}^t & 0 & \cdots \\ \vdots & \vdots & \vdots \\ 0 & 0 & h_{l-1}^t \end{pmatrix}_{m_l \times m_l m_{l-1}}.$$

Then by definition, we have that

$$E_W \left[ \frac{\partial f(x)}{\partial \text{Vec}(W_l)} \right] = E_W \left[ W_L t_{L-1}^{L-1} \right] = E_W \left[ \frac{\|W_L t_{L-1}^{L-1}\|_F^2}{\|t_{L-1}^{L-1}\|_F^2} \cdot \frac{\|t_{L-1}^{L-1}\|_F^2}{\|t_{L-1}^{L-1}\|_F^2} \cdots \frac{\|t_{1}^{L-1}\|_F^2}{\|t_{1}^{L-1}\|_F^2} \cdot \|t_{1}^{L-1}\|_F^2 \right]$$

$$= E_{W_1, \ldots, W_l} \left[ \|t_{1}^{L-1}\|_F^2 \right] E_{W_{l+1}} \left[ \|W_L e_1\|_F^2 \right] = \beta_{L, o}.$$

By Lemma C.2 for any $l = 1, \ldots, L - 2$ and $l \leq l' \leq L - 2$, we have that

$$E_{W_{l+1}} \left[ \frac{\|t_{l+1}^{L-1}\|_F^2}{\|t_{l}^{L-1}\|_F^2} \right] = \frac{\beta_{l+1}}{2} m_{l+1}.$$

We now bound the last term $E_{W_1, \ldots, W_l} \left[ \|t_{1}^{L-1}\|_F^2 \right]$ for $l = 1, \ldots, L - 1$. By definition, we have that

$$E_{W_1, \ldots, W_l} \left[ \|t_{1}^{L-1}\|_F^2 \right] = E_{W_1, \ldots, W_l} \left[ \sum_{i=1}^{m_l} |W_i|_{h_{l-1} \leq 0} \cdot \|h_{l-1}\|_2 \right] = \frac{m_l}{2} E_{W_1, \ldots, W_l} \left[ \|h_{l-1}\|_2 \right].$$

To bound the $E_{W_1, \ldots, W_l} \left[ \|h_{l-1}\|_2 \right]$, note that by Lemma C.2 we prove that for any $l = 1, \ldots, L - 1$

$$E_{W_1, \ldots, W_l} \left[ \frac{\|h_{l-1}(x)\|_2^2}{\|h_{l-2}(x)\|_2^2} \right] = \frac{\beta_{l-1}}{2} m_{l-1}.$$

Therefore, for any $l = 1, \ldots, L$, we have that

$$E_{W_1, \ldots, W_{l-1}} \left[ \|h_{l-1}(x)\|_2 \right] = E_{W_1, \ldots, W_{l-1}} \left[ \frac{\|h_{l-1}(x)\|_2}{\|h_{l-2}(x)\|_2} \cdot \frac{\|h_{l-2}(x)\|_2}{\|h_{l-2}(x)\|_2} \right] \cdot \|x\|_2$$

$$= \left( \prod_{i=1}^{l-1} \frac{\beta_i}{2} m_i \right) \|x\|_2.$$

By plugging (43) into (40), we have that

$$E_{W_1, \ldots, W_l} \left[ \|t_{1}^{L} \|_F^2 \right] = \frac{m_l}{2} \left( \prod_{i=1}^{l-1} \frac{\beta_i}{2} m_i \right) \|x\|_2.$$

By combining (38), (39) and (44), we have for any $l = 1, \ldots, L - 1$

$$E_W \left[ \frac{\partial f(x)}{\partial \text{Vec}(W_l)} \right] = \frac{m_l}{2} \left( \prod_{i=1}^{L-1} \frac{\beta_i}{2} m_i \right) \left( \prod_{i=l+1}^{L-1} \frac{\beta_i}{2} m_i \right) \cdot \beta_{L, o} \cdot \|x\|_2^2 = \frac{\beta_L}{\beta_L} \|x\|_2^2 \left( \prod_{i=1}^{L-1} \frac{\beta_i}{2} m_i \right) \cdot \frac{L-1}{2} \frac{\beta_i}{2} m_i.$$
On the other hand, by plugging Eq. (43) (under $\ell = L$) into Eq. (37), we have that

$$
E_W \left[ \| \frac{\partial f(x)}{\partial \text{Vec}(W_L)} \|_2^2 \right] = o \left( \prod_{i=1}^{L-1} \beta_i \frac{m_i}{2} \right) \|x\|_2^2
$$

Therefore,

$$
E_W \left[ \| \frac{\partial f(x)}{\partial \text{Vec}(W)} \|_2^2 \right] = \sum_{l=1}^L \| \frac{\partial f(x)}{\partial W_l} \|_2^2 = \|x\|_2^2 o \left( \prod_{i=1}^{L-1} \frac{\beta_i m_i}{2} \right) \sum_{l=1}^L \frac{\beta_L}{\beta_l},
$$

which suffices to prove Eq. (34).

C.2 Deferred proof for Theorem 4.1

Finally, we prove that the gradient difference between two training datasets under linearized network is bounded by a constant throughout training (which only depends on the network width, depth and initialization distribution).

**Theorem 4.1** Under Assumption 2.2, taking over the randomness of the random initialization and the Brownian motion, for any $t \in [0, T]$, running Langevin diffusion on linearized network Eq. (3) satisfies that

$$
E \left[ \| \nabla L(W_{t, W}; D) - L(W_{t, W}; D') \|_2^2 \right] \leq \frac{4B}{n^2},
$$

where $n$ is the training dataset size, and $B$ is a constant that only depends on the network width, depth and initialization distribution as follows.

$$
B := o \left( \prod_{i=1}^{L-1} \frac{\beta_i m_i}{2} \right) \sum_{l=1}^L \frac{\beta_L}{\beta_l},
$$

where $o$ is the number of output classes, $\{m_i\}_{i=1}^L$ are the per-layer network widths, and $\{\beta_i\}_{i=1}^L$ are the variances of Gaussian initialization at each layer.

**Proof.** Denote $W$ as the initialization parameters and denote $W_{t, W}$ as the parameters for linearized network after training time $t$. Then the gradient difference under linearized network and cross-entropy loss function is as follows.

$$
\| \nabla L(W_t; D) - \nabla L(W_t; D') \|_2 = \| \frac{\nabla f_W(x)}{n} (\text{softmax}(f_W(x)) - y) - \frac{\nabla f_W(x')}{n} (\text{softmax}(f_W(x')) - y') \|_2
\leq \frac{2}{n^2} \left( \| \nabla f_W(x) \|_2^2 + \| \nabla f_W(x') \|_2^2 \right).
$$

Plugging Lemma C.3 into the above equation with data Assumption 2.2 suffice to prove the result.

D Deferred proofs for Section 6

D.1 Deferred proof for Lemma 6.1 on convergence of training linearized network

In this section, we prove empirical risk bound for the average of all iterates of Langevin diffusion, building on standard results [40, Theorem 2] and [43, Theorem 3.1] for the (discrete time) stochastic gradient descent algorithm.

**Lemma 6.1** (Extension of [40, Theorem 2] and [43, Theorem 3.1]) Let $L_0^{lin}(W; D)$ be the empirical risk function of linearized network Eq. (3) expanded at initialization vector $W_0^{lin}$. Let $W_0^*$ be an $\alpha$-near-optimal solution for the ERM problem such that $L_0^{lin}(W_0^*; D) - \min_W L_0^{lin}(W; D) \leq \alpha$. Let $D = x_1, \cdots, x_n$ be an arbitrary training dataset of size $n$, and denote $M_0 =$
We now rewrite where

\[ \text{(50)} \]

\[ \text{(49)} \]

Therefore by plugging (50) into (49), we have that

\[ \text{(5)} \]

\[ \text{(4)} \]

\[ \text{Proof.} \]

Our proofs are heavily based on the idea in [43] Theorem 3.1] to work only in the parameter space spanned by the input feature vectors. And our proof serves as an extension of their bound to the continuous-time Langevin diffusion algorithm. We begin by using convexity of the empirical loss function \( L_{0}^{\text{lin}}(W; D) \) for linearized network to prove the following standard results

\[ \mathcal{L}_{0}^{\text{lin}}(W_{T}^{\text{lin}}; D) - \mathcal{L}_{0}^{\text{lin}}(W_{0}^{\text{lin}}; D) \leq (W_{T}^{\text{lin}} - W_{0}^{\text{lin}}, \nabla L_{0}^{\text{lin}}(W_{t}^{\text{lin}}; D)) \]  

(47)

Denote \( M_{0} = (\nabla f_{W_{0}^{\text{lin}}}(x_{1}) \cdots \nabla f_{W_{0}^{\text{lin}}}(x_{n})) \). By computing the gradient under cross entropy loss and linearized network, we have \( \nabla L_{0}^{\text{lin}}(W_{T}^{\text{lin}}; D) \) lies in the column space of \( M_{0} \). Denote \( \Pi M_{0} \) as the projection operator to the column space of \( M_{0} \), then (47) can be rewritten as

\[ \mathcal{L}_{0}^{\text{lin}}(W_{T}^{\text{lin}}; D) - \mathcal{L}_{0}^{\text{lin}}(W_{0}^{\text{lin}}; D) \leq (\Pi M_{0}(W_{T}^{\text{lin}} - W_{0}^{\text{lin}}), \nabla L_{0}^{\text{lin}}(W_{T}^{\text{lin}}; D)). \]  

(48)

By taking expectation over the randomness of Brownian motion in Langevin diffusion, we have

\[ \mathbb{E}[\mathcal{L}_{0}^{\text{lin}}(W_{T}^{\text{lin}})] - \mathcal{L}_{0}^{\text{lin}}(W_{0}^{\text{lin}}; D) \leq \frac{1}{T} \int_{0}^{T} \mathbb{E}[(\Pi M_{0}(W_{t}^{\text{lin}} - W_{0}^{\text{lin}}), \nabla L_{0}^{\text{lin}}(W_{t}^{\text{lin}}; D))] dt \]  

(49)

We now rewrite \( \mathbb{E}[(\Pi M_{0}(W_{t}^{\text{lin}} - W_{0}^{\text{lin}}), \nabla L_{0}^{\text{lin}}(W_{t}^{\text{lin}}; D))] \) by computing

\[ \frac{\partial}{\partial t} \mathbb{E}[\|W_{t}^{\text{lin}} - W_{0}^{\text{lin}}\|^{2}_{M_{0}}] \]  

where \( \|W_{t}^{\text{lin}} - W_{0}^{\text{lin}}\|^{2}_{M_{0}} = \Pi M_{0}(W_{t}^{\text{lin}} - W_{0}^{\text{lin}})^{\top} \Pi M_{0}(W_{t}^{\text{lin}} - W_{0}^{\text{lin}})^{\top} \). By applying (25) with \( p_{t} \) being the distribution for \( W_{t}^{\text{lin}} \) in Langevin diffusion for linearized network starting from point initialization \( W_{0}^{\text{lin}} \), and with function \( d(W) = \|W - W_{0}^{\text{lin}}\|^{2}_{M_{0}} \), we have that

\[ \frac{\partial}{\partial t} \mathbb{E}[\|W_{t}^{\text{lin}} - W_{0}^{\text{lin}}\|^{2}_{M_{0}}] \leq -2 \mathbb{E}[\|W_{t}^{\text{lin}} - W_{0}^{\text{lin}}\|^{2}_{M_{0}}] + \sigma^{2} \text{rank}(M_{0}). \]  

(50)

Therefore by plugging (50) into (49), we have that

\[ \mathbb{E}[\mathcal{L}_{0}^{\text{lin}}(W; D) - \mathcal{L}_{0}^{\text{lin}}(W_{0}^{\text{lin}}; D)] \leq -\frac{1}{2T} \int_{0}^{T} \frac{\partial}{\partial t} \mathbb{E}[\|W_{t}^{\text{lin}} - W_{0}^{\text{lin}}\|^{2}_{M_{0}}] dt + \frac{1}{2} \sigma^{2} \text{rank}(M_{0}) \]  

(51)

\[ \leq \frac{1}{2T}\|W_{t}^{\text{lin}} - W_{0}^{\text{lin}}\|^{2}_{M_{0}} + \frac{1}{2} \sigma^{2} \text{rank}(M_{0}) \]  

(52)

\[ \square \]

### D.2 Deferred proof for Lemma 6.3

To bound lazy training distance of training linearized network, we would need the following auxiliary Lemma about high probability upper bound for the final layer output of linearized network at initialization.

**Lemma D.1.** Fix any data record \( x \), then with high probability \( 1 - \delta \) over random initialization Eq. (5) of model weight matrices \( W^{1}, \ldots, W^{L} \) for layer 1, \( \cdots, L \), i.e., \( W_{l} \sim \mathcal{N}(0, \beta_{l}I) \), it satisfies that

\[ \|f_{W}(x)\|_{2} \leq \|x\|_{2} \bar{O}\left(\beta_{L} \prod_{i=1}^{L-1} \beta_{i}m_{i}\right) \]  

(53)

where \( \bar{O} \) ignores logarithmic terms with regard to width \( m \), depth \( L \) and tail probability \( \delta \).
Proof. To bound the term \( \| f_{W_i} \|_2 \), by definition Eq. (2), for any \( x \), we have that
\[
\| f_{W_i} (x) \|_2^2 = \| h_L (x) \|_2^2 \cdots \| h_1 (x) \|_2^2 \| h_0 (x) \|_2^2 / \| x \|_2^2 \quad \text{(54)}
\]
We now bound the terms in the right-hand-side of Eq. (54) one by one.

Regarding the first term \( \| h_l (x) \|_2^2 / \| h_{l-1} (x) \|_2^2 \) in Eq. (54), observe that by the network output definition Eq. (2), we have that
\[
\| h_l (x) \|_2^2 = \| W^l h_{l-1} (x) \|_2^2 \leq \beta_l \| h_{l-1} (x) \|_2^2 \| \hat{w} \|^2 \quad \text{(55)}
\]
where \( \hat{w} \sim \mathcal{N}(0, 1) \) and the last equality is by rotational invariance of Gaussian distribution used for initializing \( L \)-th layer weight matrix \( W^L \in \mathbb{R}^{1 \times m} \). Therefore, by tail probability expression for standard Gaussian random variable, we have that with high probability \( 1 - \delta / L \) over random initialization of \( W^L \in \mathbb{R}^{1 \times m} \), it satisfies that
\[
\frac{\| h_l (x) \|_2^2}{\| h_{l-1} (x) \|_2^2} \leq 2 \beta_l \log \frac{L}{\delta} \quad \text{(56)}
\]
Regarding the terms \( \| h_i (x) \|_2^2 / \| h_{i-1} (x) \|_2^2 \) in Eq. (54) for layer \( l = 1, \ldots, L - 1 \), by setting \( H_1 = h_2 = h_{l-1} \) in Eq. (33), we immediately prove that over random initialization of weight matrix \( W^l \in \mathbb{R}^{m_i \times m_{i-1}} \), it satisfies that
\[
\frac{\| h_i (x) \|_2^2}{\| h_{i-1} (x) \|_2^2} = \sum_{i=1}^{m_i} \rho_i \hat{w}_i^2 \quad \text{(57)}
\]
where \( \rho_1, \ldots, \rho_{m_i} \) i.i.d. \( \text{Ber}(1, 1/2) \) and \( \hat{w}_1, \ldots, \hat{w}_{m_i} \) i.i.d. \( \mathcal{N}(0, 1) \) and \( \rho_i \) and \( \hat{w}_i \) are independent.

By tail probability expression for Gaussian random variable \( \hat{w}_i \), we have that \( P(\hat{w}_i^2 \geq t) \leq e^{-t/2} \) for any \( t > 0 \). By union bound over \( i = 1, \ldots, m_i \), we prove that for any layer \( l = 1, \ldots, L - 1 \), with high probability \( 1 - \delta / 2L \) over random initialization of weight matrix \( W^l \in \mathbb{R}^{m_i \times m_{i-1}} \), it satisfies that
\[
\max_i \hat{w}_i^2 \leq 2 \log \frac{2m_i L}{\delta} \quad \text{(58)}
\]
Moreover, by applying Hoeffding’s inequality to i.i.d. Bernoulli r.v.s \( \rho_1, \ldots, \rho_{m_i} \), we prove with high probability \( 1 - \delta / 2L \), it satisfies that \( \sum_{i=1}^{m_i} \rho_i \leq m_i (1 + \log(2L/\delta)) \). By combining it with Eq. (58) via union bound, and plugging the result into Eq. (60), we prove for any \( l = 1, \ldots, L - 1 \), it satisfies with high probability \( 1 - \delta / 4 \) over random initialization of weight matrix \( W^l \in \mathbb{R}^{m_i \times m_{i-1}} \) that
\[
\frac{\| h_i (x) \|_2^2}{\| h_{i-1} (x) \|_2^2} = m_i \beta_l \log \frac{2m_i L}{\delta} \cdot (1 + \log(2L/\delta)) \quad \text{(59)}
\]
By using union bound over Eq. (59) for layer \( l = 1, \ldots, L - 1 \) and Eq. (56), we have that with high probability \( 1 - \delta \) over random initialization Eq. (5), it satisfies that
\[
\frac{\| h_L (x) \|_2^2}{\| h_0 (x) \|_2^2} \leq \hat{O} \left( \prod_{i=1}^{L-1} \beta_i m_i \right) \quad \text{(60)}
\]
where \( \hat{O} \) ignores logarithmic factors with regard to \( m, L \) and \( \delta \).

By plugging Eq. (60) into Eq. (54), we prove that with high probability \( 1 - \delta \) over initialization Eq. (5), the following bound holds.
\[
\| f_W (x) \|_2 \leq \| x \|_2^2 \hat{O} \left( \beta_L \left( \prod_{i=1}^{L-1} \beta_i m_i \right) \right), \quad \text{(61)}
\]
where \( \hat{O} \) ignores logarithmic terms with regard to width \( m \), depth \( L \) and tail probability \( \delta \).
Lemma 6.3 (Bounding lazy training distance via smallest eigenvalue of the NTK matrix) Suppose that the data and network regularity Assumption 2.2 holds. Let $\xi_l$ be an auxiliary variable that takes value 1 is $m_l = \Omega(n)$, and zero otherwise with $m_l$ being width for layer $l$ and $n$ being number of training data. Let $\mathcal{L}_{\theta_{lin}}^0(W)$ be the empirical risk function defined below Eq. (1) for linearized network expanded at initialization vector $W_{0}^{lin}$. Then for any initialization vector $W_{0}^{lin}$, there exists a corresponding near optimal solution $W_{0}^{lin}$, such that $\mathcal{L}_{\theta_{lin}}^0(W_{0}^{lin}) \leq \frac{1}{n^2}$, rank($M_0$) = $n$ and

$$R \leq \tilde{O}\left(\max\left\{\frac{1}{\beta_L} \left(\prod_{l=1}^{L-1} \beta_l^{-1} m_l\right), 1\right\}\right) \frac{n}{d \sum_{l=1}^{L} \xi_l \beta_l^{-1}}.$$  
(62)

with high probability over random initialization Eq. (5), where $\tilde{O}$ ignores logarithmic factors with regard to $n$, $m$, $L$, and tail probability $\delta$.

Proof. Given arbitrary initialization parameters $W_{0}^{lin}$, we first construct an solution $W_{0}^{lin}$ that is nearly optimal for the ERM problem over $\mathcal{L}_{\theta_{lin}}^0(W)$. Specifically, let $W_{0}^{lin}$ have the following expression.

$$W_{0}^{lin} = \begin{pmatrix} \nabla f_{W_{0}^{lin}}(x_1)^T \\ \vdots \\ \nabla f_{W_{0}^{lin}}(x_n)^T \end{pmatrix}$$

where $M_0 = \begin{pmatrix} d \prod_{l=1}^{L-1} m_l \cdot \left(\prod_{l=1}^{L} \beta_l\right) \cdot \left(\sum_{l=1}^{L} \xi_l \beta_l^{-1}\right) \end{pmatrix}$ is the NTK feature matrix at initialization and $\dagger$ denotes the pseudo-inverse. When the data regularity assumption Assumption 2.4 holds, by applying existing bounds for the smallest eigenvalue of the NTK matrix $M_0 M_0^T$ in [38, Theorem 4.1], with high probability over random initialization it satisfies that

$$O\left(\left(\prod_{l=1}^{L-1} m_l \cdot \left(\prod_{l=1}^{L} \beta_l\right) \cdot \left(\sum_{l=1}^{L} \xi_l \beta_l^{-1}\right)\right) \geq \lambda_{min}(M_0 M_0^T) \right) \geq \Omega\left(\left(\prod_{l=1}^{L-1} m_l \cdot \left(\prod_{l=1}^{L} \beta_l\right) \cdot \left(\sum_{l=1}^{L} \xi_l \beta_l^{-1}\right)\right) \right),$$

(64)

where $\xi_l$ is an auxiliary variable that takes value one if $m_l = \Omega(n)$, and zero otherwise, with $m_l$ being the width of layer $l$ and $n$ being the number of training data.

Eq. (64) implies that rank($M_0$) = $n$ with high probability, and therefore $M_0^\dagger = M_0^T (M_0 M_0^T)^{-1}$

$$\begin{pmatrix} \nabla f_{W_{0}^{lin}}(x_1) \\ \vdots \\ \nabla f_{W_{0}^{lin}}(x_n) \end{pmatrix} = \begin{pmatrix} 2 \ln n \cdot y_1 \\ \vdots \\ 2 \ln n \cdot y_n \end{pmatrix}$$

with high probability. By plugging it into the cross-entropy loss for the single-output network defined below Eq. (1), we have that the solution $W_{0}^{lin}$ satisfies

$$\mathcal{L}_{\theta_{lin}}^0(W_{0}^{lin}) = \log(1 + \exp(-2 \ln n)) < \frac{1}{n^2}. $$

(65)
We now only need to prove that the solution $W_{0}^\perp$ is close to the initialization parameters $W_{0}^{lin}$ in $\ell_2$ norm with high probability. By applying the holder inequality on (63), we have that
\[
R = \|W_{0}^\perp - W_{0}^{lin}\|_2 \leq \|M_0\|_2^2 \left\| \begin{pmatrix} 2 \ln n \cdot y_1 - f_{W_{0}^{lin}}(x_1) \\ \vdots \\ 2 \ln n \cdot y_n - f_{W_{0}^{lin}}(x_n) \end{pmatrix} \right\|_2^2
\]
(66)
\[
\leq \frac{1}{\lambda_{\min}(M_0M_0^\top)} \left\| \begin{pmatrix} 2 \ln n \cdot y_1 - f_{W_{0}^{lin}}(x_1) \\ \vdots \\ 2 \ln n \cdot y_n - f_{W_{0}^{lin}}(x_n) \end{pmatrix} \right\|_2^2
\]
(67)
We now prove upper bounds for the two terms on the right hand side separately. For the first term, by using Eq. (64), we have that with high probability
\[
\frac{1}{\lambda_{\min}(M_0M_0^\top)} \leq O \left( \frac{d \prod_{l=1}^{L-1} m_l}{\prod_{l=1}^{L} \beta_l} \cdot \left( \sum_{l=1}^{L} \xi_l \beta_l^{-1} \right) \right)
\]
(68)
where $\xi_l$ is an auxiliary variable that takes value one if $m_l = \tilde{O}(n)$, and zero otherwise, with $m_l$ being the width of layer $l$ and $n$ being the number of training data.

For the second term, by Cauchy-Schwarz inequality, we have that
\[
\left\| \begin{pmatrix} 2 \ln n \cdot y_1 - f_{W_{0}^{lin}}(x_1) \\ \vdots \\ 2 \ln n \cdot y_n - f_{W_{0}^{lin}}(x_n) \end{pmatrix} \right\|_2^2 \leq 2 \cdot (2 \ln n)^2 \sum_{i=1}^{n} y_i^2 + 2 \sum_{i=1}^{n} \|f_{W_{0}^{lin}}(x_i)\|_2^2
\]
(69)
By using Lemma D.1 under data Assumption 2.4, we have $\|f_{W_{0}^{lin}}(x)\|_2 \leq \tilde{O} \left( \beta_L \left( \prod_{i=1}^{L-1} \beta_i m_i \right) \right)$. By plugging this result into Eq. (69), we have that with high probability over random initializaition Eq. (5), it satisfies that
\[
\left\| \begin{pmatrix} 2 \ln n \cdot y_1 - f_{W_{0}^{lin}}(x_1) \\ \vdots \\ 2 \ln n \cdot y_n - f_{W_{0}^{lin}}(x_n) \end{pmatrix} \right\|_2^2 = \tilde{O} \left( n + n \beta_L \left( \prod_{i=1}^{L-1} \beta_i m_i \right) \right)
\]
(70)
where $\tilde{O}$ ignores logarithmic factors with regard to $n$, $m$, $L$, and tail probability $\delta$. Therefore, by combining Eq. (68) and Eq. (70) with union bound, and by plugging the result into Eq. (67), we have that with high probability over random initialization Eq. (5)
\[
R \leq \tilde{O} \left( \frac{n + n \beta_L \left( \prod_{i=1}^{L-1} \beta_i m_i \right)}{d \beta_L \prod_{l=1}^{L-1} \beta_l m_l \cdot \left( \sum_{l=1}^{L} \xi_l \beta_l^{-1} \right)} \right)
\]
\[
\leq \tilde{O} \left( \max \left\{ \frac{n}{\beta_L \left( \prod_{i=1}^{L-1} \beta_i m_i \right)}, 1 \right\} \cdot \frac{1}{d \sum_{l=1}^{L} \xi_l \beta_l^{-1}} \right)
\]
where $\xi_l$ is an auxiliary variable that takes value one if $m_l = \tilde{O}(n)$, and zero otherwise, with $m_l$ being the width of layer $l$ and $n$ being the number of training data, and where $\tilde{O}$ ignores logarithmic factors with regard to $n$, $m$, $L$, and tail probability $\delta$.

D.3 Deferred proof for Corollary 6.4
A summary of our excess empirical risk bounds under different initializations is in Table 1. 

**Theorem E.1** (KL composition for noisy GD under possibly unbounded gradient difference) Let the iterative update in noisy GD algorithm be defined by: $W_{(k+1)} = W_{(k)} - \eta \nabla \mathcal{L}(W_{(k)}; D) + \sqrt{2\eta\sigma^2}Z_k$, where $Z_k \sim \mathcal{N}(0, I)$. Then the KL divergence between running noisy GD for DNN $(2)$ on neighboring datasets $D$ and $D'$ satisfies

$$KL(W_{(1:)}, W'_{(1:)}) = \frac{1}{2\sigma^2} \sum_{k=0}^{K-1} \eta \cdot \mathbb{E} \left[ \left\| \nabla \mathcal{L}(W_{(k)}; D) - \nabla \mathcal{L}(W_{(k)}; D') \right\|_2^2 \right].$$

**Corollary 6.4** (Privacy utility trade-off for linearized network) Assume that the data and network regularity Assumption 2.4 holds. Let $B = \left( \prod_{i=1}^{L-1} \frac{\beta_i m_i}{2} \right) \sum_{i=1}^{L} \frac{\beta_i}{\beta_i + m_i}$ be the gradient norm constant proved in Eq. (8), and let $R \leq \hat{O} \left( \max \left\{ \frac{1}{\beta_L (\prod_{i=1}^{L-1} \beta_i m_i) \left( \frac{1}{m_i} \right)^{L-1} \prod_{i=1}^{L-1} \beta_i m_i} \right\} \right)$ be the lazy training distance bound proved in Lemma 6.3. Then by setting $\sigma^2 = \frac{2BT}{\varepsilon n^2}$ and $T = \sqrt{\frac{2\varepsilon n R}{2B}}$, we have that releasing all iterates of Langevin diffusion with time $T$ satisfies $\varepsilon$-KL privacy, and has empirical excess risk upper bounded by

$$\mathbb{E}[\mathcal{L}_0^{lin}(\tilde{W}_T^{lin})] - \min_W \mathcal{L}_0^{lin}(W; D) \leq \hat{O} \left( \frac{1}{n^2} + \sqrt{\frac{1}{2L-1 \varepsilon n^4 \omega} \max\{1, \beta L \prod_{i=1}^{L-1} \beta_i m_i\}} \right).$$

Specifically, under large data dimension $d = \Omega(n)$ and hidden-layer width $m = \Omega(n)$, we have

$$\mathbb{E}[\mathcal{L}_0^{lin}(\tilde{W}_T^{lin})] - \min_W \mathcal{L}_0^{lin}(W; D) \leq \hat{O} \left( \frac{1}{n^2} + \sqrt{\frac{1}{2L-1 \varepsilon n^4 \omega} \max\{1, \beta L \prod_{i=1}^{L-1} \beta_i m_i\}} \right).$$

A summary of our excess empirical risk bounds under different initializations is in Table 1.

**Proof.** By setting $W_0 = W_0^{\frac{1}{2\eta}}$ in Lemma 6.1 with $W_0^{\frac{1}{2\eta}}$ constructed as Lemma 6.3, we have with high probability over random initialization Eq. (5), we have

$$\mathbb{E}[\mathcal{L}_0^{lin}(\tilde{W}_T^{lin})] - \min_W \mathcal{L}_0^{lin}(W; D) \leq \hat{O} \left( \frac{1}{n^2} + \frac{R}{2T} + \frac{\sigma^2 n}{2} \right).$$

where $R \leq \hat{O} \left( \max \left\{ \frac{1}{\beta_L (\prod_{i=1}^{L-1} \beta_i m_i) \left( \frac{1}{m_i} \right)^{L-1} \prod_{i=1}^{L-1} \beta_i m_i} \right\} \right)$ by Lemma 6.3.

Meanwhile, to ensure $\varepsilon$-KL privacy, by Corollary 4.2, we only need to set $\sigma^2 = \frac{2BT}{\varepsilon n^2}$ where $B = \left( \prod_{i=1}^{L-1} \frac{\beta_i m_i}{2} \right) \sum_{i=1}^{L} \frac{\beta_i}{\beta_i + m_i}$ by plugging the single output assumption $o = 1$ into Eq. (8). By plugging $\sigma^2 = \frac{2BT}{\varepsilon n^2}$ into (73), we prove that

$$\mathbb{E}[\mathcal{L}_0^{lin}(\tilde{W}_T^{lin})] - \min_W \mathcal{L}_0^{lin}(W; D) \leq \frac{1}{n^2} + \frac{R}{2T} + \frac{BT}{\varepsilon n}.$$
Therefore, by plugging Eq. (78) into Eq. (77), we have that

\[ KL(p_{(k)}) = 0 \]

Now we expand the term \( KL(p_{(1,k+1)}, p'_{(1,k+1)}) \) by the Bayes rule as follows.

\[
KL(p_{(1,k+1)}, p'_{(1,k+1)})
\]

\[
= \mathbb{E}_{p_{(1,k+1)}}(w_{(1,k+1)}) \left[ \log \left( \frac{p_{(k+1)|((1:k))}^W(W_{(k+1)}|W_{(1:k)})p_{(1:k)}(W_{(1:k)})}{p'_{(k+1)|((1:k))}^W(W_{(k+1)}|W_{(1:k)})p'_{(1:k)}(W_{(1:k)})} \right) \right] + KL(p_{(1:k)}, p'_{(1:k)})
\]

Observe that conditioned on fixed model parameters \( W_{(1:k)} \) at iteration \( 1, \cdots, k \), the distributions \( p_{(k+1)|((1:k))}, p'_{(k+1)|((1:k))} \) are Gaussian with per-dimensional variance \( \sigma^2 \). Therefore, by computing the KL divergence between two multivariate Gaussians, we have that

\[
KL(p_{(k+1)|((1:k))}, p'_{(k+1)|((1:k))}) = \frac{1}{2\sigma^2} \cdot \mathbb{E} \left\| \nabla L(W_{(k)}; D) - \nabla L(W_{(k)}; D') \right\|^2_2
\]

Therefore, by plugging Eq. (78) into Eq. (77), we have that

\[
KL(p_{(1,k+1)}, p'_{(1,k+1)}) = \frac{\eta}{2\sigma^2} \mathbb{E} \left\| \nabla L(W_{(k)}; D) - \nabla L(W_{(k)}; D') \right\|^2_2 + KL(p_{(1:k)}, p'_{(1:k)})
\]

By summing (79) over \( k = 0, \cdots, K-1 \) and observing that \( KL(p_{(0)}, p'_{(0)}) = 0 \) (as the initialization distribution is the same between noisy GD on \( D \) and \( D' \)), we finish the proof for Eq. (75). □