Mean-Field Analysis of Two-Layer Neural Networks: Non-Asymptotic Rates and Generalization Bounds

Zixiang Chen[*] and Yuan Cao[†] and Quanquan Gu[‡] and Tong Zhang[§]

Abstract

A recent line of work in deep learning theory has utilized the mean-field analysis to demonstrate the global convergence of noisy (stochastic) gradient descent for training over-parameterized two-layer neural networks. However, existing results in the mean-field setting do not provide the convergence rate of neural network training, and the generalization error bound is largely missing. In this paper, we provide a mean-field analysis in a *generalized neural tangent kernel regime*, and show that noisy gradient descent with weight decay can still exhibit a "kernel-like" behavior. This implies that the training loss converges linearly up to a certain accuracy in such regime. We also establish a generalization error bound for two-layer neural networks trained by noisy gradient descent with weight decay. Our results shed light on the connection between mean field analysis and the neural tangent kernel based analysis.

1 Introduction

Deep learning has achieved tremendous practical success in a wide range of machine learning tasks (Krizhevsky et al., 2012; Hinton et al., 2012; Silver et al., 2016). However, due to the nonconvex and over-parameterized nature of modern neural networks, the success of deep learning cannot be fully explained by conventional optimization and machine learning theory.

Recently, a line of work utilized a mean-field framework to study the training of extremely wide (or even infinitely wide) neural networks (Chizat and Bach, 2018; Mei et al., 2018, 2019; Wei et al., 2019; Fang et al., 2019a,b). It has been shown that over-parameterized two-layer neural networks can be trained to a global optimizer of the training loss, despite the non-convex optimization landscape. However, most of the global convergence results proved in the line are asymptotic, and the convergence rate of the training algorithm is largely unknown, except for some specifically designed training procedure (Wei et al., 2019). Moreover, the generalization performance of neural networks trained in the mean-field regime has not been well-studied.

Compared with the mean-field analysis, another line of work studying the learning of over-parameterized neural network in the so-called "neural tangnet kernel (NTK) regime" (Jacot et al.,

[*]Department of Computer Science, University of California, Los Angeles, CA 90095, USA; e-mail: `chenzx19@cs.ucla.edu`

[†]Department of Computer Science, University of California, Los Angeles, CA 90095, USA; e-mail: `yuancao@cs.ucla.edu`

[‡]Department of Computer Science, University of California, Los Angeles, CA 90095, USA; e-mail: `qgu@cs.ucla.edu`

[§]Department of Computer Science and Mathematics, The Hong Kong University of Science and Technology, Hong Kong; e-mail: `tongzhang@tongzhang-ml.org`

2018; Du et al., 2019b; Allen-Zhu et al., 2019b; Du et al., 2019a; Zou et al., 2018; Arora et al., 2019a,b; Zou and Gu, 2019; Cao and Gu, 2019a; Ji and Telgarsky, 2020; Chen et al., 2019) is known to have its own advantages and disadvantages. On the one hand, due to the relatively simpler training dynamics, the convergence rates and generalization error bounds have been well-established in the NTK regime. On the other hand, NTK regime requires that the network weights stays very close to their initialization throughout training, which does not match the experimental observations. Moreover, due to this requirement, NTK analysis cannot handle regularizers such as weight decay, or large additive noises in the noisy gradient descent algorithm used in the mean-field analysis.

With the recent development based on the mean-field and NTK approaches, a natural question is:

Is it possible to establish a unified framework connecting the mean-field and NTK approaches?

In this paper, we give an affirmative answer to this question, and show that with an appropriate scaling, neural networks trained with noisy gradient descent and weight decay can still enjoy linear convergence rate up to certain accuracy. Moreover, we also establish generalization error bounds for the final network trained by noisy gradient descent.

We summarize the contributions of this paper as follows:

- We establish a comprehensive connection between NTK and mean-field analyses, and demonstrate that if a large scaling factor is introduced into the network, then the whole training process can be similar to the dynamics of neural tangent kernel. Our result improves existing result in Mei et al. (2019), which only shows the closeness between the two training dynamics for a limited time period $t \in [0, T]$ for some finite T. In comparison, we provide a uniform bound for all $t \in [0, +\infty)$. A direct consequence of our analysis is the linear convergence of noisy gradient descent up to certain accuracy. To the best of our knowledge, this is the first convergence rate result of noisy gradient descent for neural network training.[1]

- Our analysis demonstrates that neural network training with gradient noises and appropriate regularizers can still exhibit similar training dynamics as kernel methods, which is considered intractable in the neural tangent kernel literature, as the regularizer and gradient noises easily push the network parameters far away from the initialization. Our analysis overcomes this technical barrier by relaxing the requirement on the closeness in the parameter space to the closeness between distributions in terms of KL-divergence.

- We establish generalization bounds for the neural networks trained with noisy gradient descent with weight decay regularization under different network scalings. Our result shows that when the scaling factor is large, the infinitely wide neural network trained by noisy gradient descent with weight decay regularization can learn a class of functions that is defined based on a bounded χ^2-divergence to initialization distribution. When the scaling factor is small (of constant order), we also provide a comparable result showing that a function class defined with KL-divergence can be efficiently learnt.

[1]Although our theoretical analysis directly focus on the continuous time limit of the noisy gradient descent algorithm, our result can be easily extended to the discrete time setting by applying the approximation results established by Mei et al. (2018)

1.1 Notation

We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. For a vector $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, let $\|\mathbf{x}\|_0 = |\{x_i : x_i \neq 0, i = 1, \ldots, d\}|$ and $\|\mathbf{x}\|_\infty = \max\{|x_i| : 1 \leq i \leq d\}$ be the ℓ_0 and ℓ_∞ norms of $\mathbf{x}$ respectively. For any positive integer p, we denote the ℓ_p norm of $\mathbf{x}$ as $\|\mathbf{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{A}\|_F = (\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2)^{1/2}$ to denote its Frobenius norm. For any positive integer p, we define $\|\mathbf{A}\|_p = \sup_{\|\mathbf{v}\|_p=1} \|\mathbf{A}\mathbf{v}\|_p$ as the matrix p-norm of $\mathbf{A}$, and refer to $\|\mathbf{A}\|_2$ as the spectral norm of $\mathbf{A}$. We also define $\|\mathbf{A}\|_{\infty,\infty} = \max\{|A_{ij}| : 1 \leq i \leq m, 1 \leq j \leq n\}$. For a positive semi-definite matrix $\mathbf{A}$, we use $\lambda_{\min}(\mathbf{A})$ to denote its smallest eigenvalue.

For a positive integer n, we denote $[n] = \{1, \ldots, n\}$. For an event E, we use $\mathbb{1}\{E\}$ to denote the indicator on whether this event happens. We also use the following asymptotic notations. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C_1 such that $a_n \leq C_1 b_n$. Similarly, if there exists an absolute constant $C_2 > 0$ such that $a_n \geq C_2 b_n$, then we write $a_n = \Omega(b_n)$. We introduce $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to further hide the logarithmic terms in the Big-O and Big-Omega notations.

At last, for two distributions p and p' over $\mathbb{R}^d$, we define their 2-Wasserstein distance as

$$\mathcal{W}_2(p, p') = \Big(\inf_{\mathbf{z}\sim p, \mathbf{z}'\sim p'} \mathbb{E}_{\mathbf{z},\mathbf{z}'} \|\mathbf{z} - \mathbf{z}'\|_2^2\Big)^{1/2},$$

where the infimum is taken over all random vectors $(\mathbf{z}, \mathbf{z}') \in \mathbb{R}^d \otimes \mathbb{R}^d$ with marginal distributions p and p' respectively. We also define KullbackLeibler divergence (KL-divergence) and χ^2-divergence between p and p' as follows:

$$D_{\mathrm{KL}}(p||p') = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{p'(\mathbf{z})} d\mathbf{z}, \quad D_{\chi^2}(p||p') = \int \bigg(\frac{p(\mathbf{z})}{p'(\mathbf{z})} - 1\bigg)^2 p'(\mathbf{z}) d\mathbf{z}.$$

2 Related Work

Our analysis follows the mean-field framework adopted in the recent line of work (Chizat and Bach, 2018; Mei et al., 2018, 2019; Wei et al., 2019; Fang et al., 2019a,b). Chizat and Bach (2018) showed the convergence of gradient descent for training infinitely wide, two-layer networks under certain structural assumptions. Mei et al. (2018) proved the global convergence of noisy stochastic gradient descent and established approximation bounds between finite and infinite neural networks. Mei et al. (2019) further showed that this approximation error can be independent of the input dimension in certain cases, and proved that under certain scaling condition, the residual dynamics of noiseless gradient descent is close to the dynamics of NTK-based kernel regression within certain bounded time interval $[0, T]$. Wei et al. (2019) proved the convergence of a certain perturbed Wasserstein gradient flow, and established a generalization bound of the global minimizer of weakly regularized logistic loss. Fang et al. (2019a) extended the mean-field analysis and proposed a new concept called neural feature repopulation, which is based on a refined joint analysis on first and second layer parameters. Fang et al. (2019b) generalized this neural feature repopulation technique to deep neural networks.

Another highly relevant line of research is the study of neural network training in the “neural tangent kernel regime” (Jacot et al., 2018; Du et al., 2019b; Allen-Zhu et al., 2019b; Du et al., 2019a; Zou et al., 2018; Arora et al., 2019a,b; Zou and Gu, 2019; Cao and Gu, 2019a; Ji and

Telgarsky, 2020; Chen et al., 2019). In particular, Jacot et al. (2018) first introduced the concept of neural tangent kernel by studying the training dynamics of neural networks with square loss. Based on neural tangent kernel, Du et al. (2019b); Allen-Zhu et al. (2019b); Du et al. (2019a); Zou et al. (2018); Oymak and Soltanolkotabi (2019); Zou and Gu (2019); Su and Yang (2019); Cao et al. (2019) proved the global convergence of (stochastic) gradient descent under various settings. Chizat et al. (2019) extended the similar idea to a more general framework called "lazy training". Allen-Zhu et al. (2019a); Arora et al. (2019a); Cao and Gu (2020, 2019a); Nitanda and Suzuki (2019); Ji and Telgarsky (2020); Chen et al. (2019) established generalization bounds for over-parameterized neural networks trained by (stochastic) gradient descent.

Several recent results focused on showing that neural networks can outperform kernel methods on specific learning tasks (Wei et al., 2019; Allen-Zhu and Li, 2019; Bai and Lee, 2019; Allen-Zhu and Li, 2020). As previously mentioned, Wei et al. (2019) utilized a perturbed Wasserstein gradient flow to solve the weakly regularized optimization problem and establised a generalization bound based on normalized margin. Allen-Zhu and Li (2019) showed that three-layer ResNets can perform hierarchical learning that beats any kernel methods on certain hierarchical learning problems. Bai and Lee (2019) proposed a training procedure with randomization technique and showed that the obtained two-layer networks can outperform neural tangent kernel by a dimension factor. Allen-Zhu and Li (2020) extended the results for three-layer ResNet in Allen-Zhu and Li (2019) to multi-layer DenseNets with the quadratic activation function.

Besides the work mentioned above, this paper is also related to some other remarkable analyses on deep learning. Gunasekar et al. (2017); Soudry et al. (2018); Ji and Telgarsky (2019); Gunasekar et al. (2018a,b); Nacson et al. (2019b); Li et al. (2018b); Nacson et al. (2019a); Lyu and Li (2019) studied the implicit bias problem, and showed that when training over-parameterized models like linear networks or networks with homogeneous activation functions, (stochastic) gradient descent converges to an optimizer with specific properties. Tian (2017); Brutzkus and Globerson (2017); Li and Yuan (2017); Soltanolkotabi (2017); Du et al. (2018a,b); Zhong et al. (2017); Zhang et al. (2019); Cao and Gu (2019b) established parameter recovery guarantees for neural networks under certain assumptions on the network structure and the distribution of input data. Neyshabur et al. (2015); Bartlett et al. (2017); Neyshabur et al. (2018); Golowich et al. (2018); Arora et al. (2018); Li et al. (2018a); Wei et al. (2019) studied uniform convergence based generalization bounds for neural networks.

3 Problem Setting and Preliminaries

In this section we introduce the basic problem setting for training an infinitely wide two-layer neural network, and explain its connection to the training dynamics of standard, finitely wide neural networks.

Inspired by the study in Chizat et al. (2019); Mei et al. (2019), we introduce a scaling factor $\alpha > 0$ and study two-layer, infinitely wide neural networks of the form

$$f(p, \mathbf{x}) = \alpha \int_{\mathbb{R}^{d+1}} u h(\boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du, \tag{3.1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input, $\boldsymbol{\theta} \in \mathbb{R}^d$ and $u \in \mathbb{R}$ are the first and second layer parameters respectively, $p(\boldsymbol{\theta}, u)$ is their joint distribution, and $h(\boldsymbol{\theta}, \mathbf{x})$ is the activation function. It is easy to see that (3.1)

Algorithm 1 Noisy Gradient Descent for Training Two-layer Networks

Input: Step size η, total number of iterations T

Initialize $(\boldsymbol{\theta}_j, u_j) \sim p_0(\boldsymbol{\theta}, u)$, $j \in [m]$.

for $t = 0$ **to** $T - 1$ **do**

  Draw Gaussian noises $\zeta_{u,j} \sim N(0, \sqrt{2\eta})$, $j \in [m]$

  $u_{t+1,j} = u_{t,j} - \eta \nabla_u \widehat{Q}(\{(\boldsymbol{\theta}_{t,j}, u_{t,j})\}_{j=1}^m) - \sqrt{\lambda}\zeta_{u,j}$

  Draw Gaussian noises $\boldsymbol{\zeta}_{\boldsymbol{\theta},j} \sim N(0, \sqrt{2\eta}\mathbf{I}_d)$, $j \in [m]$

  $\boldsymbol{\theta}_{t+1,j} = \boldsymbol{\theta}_{t,j} - \eta \nabla_{\boldsymbol{\theta}} \widehat{Q}(\{(\boldsymbol{\theta}_{t,j}, u_{t,j})\}_{j=1}^m) - \sqrt{\lambda}\boldsymbol{\zeta}_{\boldsymbol{\theta},j}$

end for

is the infinite-width limit of the following neural network of finite width

$$f_m(\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m, \mathbf{x}) = \frac{\alpha}{m} \sum_{j=1}^m u_j h(\boldsymbol{\theta}_j, \mathbf{x}), \tag{3.2}$$

where m is the number of hidden nodes, $\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m$ are i.i.d. samples drawn from $p(\boldsymbol{\theta}, u)$. Note that choosing $\alpha = \sqrt{m}$ in (3.2) recovers the standard scaling in the neural tangent kernel regime (Du et al., 2019b), and setting $\alpha = 1$ in (3.1) gives the standard setting for mean-field analysis (Mei et al., 2018, 2019).

We consider training the neural network with square loss and weight decay regularization. Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be the training data set, and $\phi(y', y) = (y' - y)^2$ be the square loss function. We consider Gaussian initialization $p_0(\boldsymbol{\theta}, u) \propto \exp[-u^2/(2\sigma_u)^2 - \|\boldsymbol{\theta}\|_2^2/(2\sigma_{\boldsymbol{\theta}}^2)]$. Then for finite-width neural network (3.2), we define the training objective function as

$$\widehat{Q}(\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m) = \mathbb{E}_S[\phi(f_m(\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m, \mathbf{x}), y)] + \frac{\lambda}{m} \sum_{j=1}^m \bigg(\frac{u_j^2}{2\sigma_u^2} + \frac{\|\boldsymbol{\theta}_j\|_2^2}{2\sigma_{\boldsymbol{\theta}}^2}\bigg),$$

where $\mathbb{E}_S[\cdot]$ denotes the average over the training sample S, and $\lambda > 0$ is a regularization parameter. In order to minimize the objective function $\widehat{Q}(\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m)$ for the finite-width neural network (3.2), we consider the noisy gradient descent algorithm, which is displayed in Algorithm 1.

It has been extensively studied (Mei et al., 2018; Chizat and Bach, 2018; Mei et al., 2019; Fang et al., 2019a,b) in the mean-field regime that, the continuous-time, infinite-width limit of Algorithm 1 can be characterized by the following partial differential equation (PDE) of the distribution $p_t(\boldsymbol{\theta}, u)$[2]:

$$\frac{dp_t(\boldsymbol{\theta}, u)}{dt} = -\nabla_u[p_t(\boldsymbol{\theta}, u)g_1(t, \boldsymbol{\theta}, u)] - \nabla_{\boldsymbol{\theta}} \cdot [p_t(\boldsymbol{\theta}, u)g_2(t, \boldsymbol{\theta}, u)] + \lambda\Delta[p_t(\boldsymbol{\theta}, u)], \tag{3.3}$$

where

$$\begin{aligned} g_1(t, \boldsymbol{\theta}, u) &= -\alpha\mathbb{E}_S[\nabla_{y'}\phi(f(p_t, \mathbf{x}), y)h(\boldsymbol{\theta}, \mathbf{x})] - \lambda u/\sigma_u^2, \\ g_2(t, \boldsymbol{\theta}, u) &= -\alpha\mathbb{E}_S[\nabla_{y'}\phi(f(p_t, \mathbf{x}), y)u\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta}, \mathbf{x})] - \lambda\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2. \end{aligned}$$

Below we give an informal proposition to describe the connection between Algorithm 1 and PDE

[2]Throughout this paper, we define ∇ and Δ without subscripts as the gradient/Laplacian operators with respect to the full parameter collection $(\boldsymbol{\theta}, u)$.

(3.3). One can refer to Mei et al. (2018); Chizat and Bach (2018); Mei et al. (2019) for more details on such approximation results.

Proposition 3.1 (informal)**.** Suppose that $h(\boldsymbol{\theta}, \mathbf{x})$ is sufficiently smooth. Let $\{(\boldsymbol{\theta}_{t,j}, u_{t,j})\}_{j=1}^m$, $t \geq 0$ be output by Algorithm 1, and p_t be the solution of PDE (3.3). Then for any $t \geq 0$ and any $\mathbf{x}$, it holds that

$$\lim_{m\to\infty} \lim_{\eta\to 0} f_m(\{(\boldsymbol{\theta}_{\lfloor t/\eta \rfloor, j}, u_{\lfloor t/\eta \rfloor, j})\}_{j=1}^m, \mathbf{x}) = f(p_t, \mathbf{x}).$$

Based on Proposition 3.1, one can convert the original optimization dynamics in the parameter space to the distributional dynamics in the probability measure space. In the rest of our paper, we mainly focus on $p_t(\boldsymbol{\theta}, u)$ defined by the PDE (3.3). It is worth noting that PDE (3.3) minimizes the following energy functional

$$Q(p) = L(p) + \lambda D_{\mathrm{KL}}(p||p_0), \tag{3.4}$$

where $L(p) = \mathbb{E}_S[\phi\big((f(p, \mathbf{x}), y\big)]$ is the empirical square loss, and $D_{\mathrm{KL}}(p||p_0) = \int p \log(p/p_0) d\boldsymbol{\theta} du$ is the KL-divergence between p and p_0 (Fang et al., 2019a). The asymptotic convergence of PDE (3.3) towards the global minimum of (3.4) is recently established by Mei et al. (2018); Chizat and Bach (2018); Mei et al. (2019); Fang et al. (2019a,b).

Recall that, compared with the standard mean-field analysis, we consider the setting with an additional scaling factor α in (3.1). When α is large, we expect to build a connection to the recent results in the "neural tangent kernel regime" (Mei et al., 2019; Chizat et al., 2019), where the neural network training is similar to kernel regression using the neural tangent kernel $K(\mathbf{x}, \mathbf{x}')$ defined as $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}')$, where

$$\begin{aligned}
K_1(\mathbf{x}, \mathbf{x}') &= \int u^2 \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}') \rangle p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du, \\
K_2(\mathbf{x}, \mathbf{x}') &= \int h(\boldsymbol{\theta}, \mathbf{x}) h(\boldsymbol{\theta}, \mathbf{x}') p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du.
\end{aligned}$$

Note that the neural tangent kernel function $K(\mathbf{x}, \mathbf{x}')$ is defined based on the initialization distribution p_0. This is because the specific network scaling in the neural tangent kernel regime forces the network parameters to stay close to initialization. In our analysis, we extend the definition of neural tangent kernel function to any distribution p, and define the corresponding Gram matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ of the kernel function on the training sample S as follows:

$$\mathbf{H}(p) = \mathbf{H}_1(p) + \mathbf{H}_2(p), \tag{3.5}$$

where

$$H_1(p)_{i,j} = \mathbb{E}_p[u^2 \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i), \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_j) \rangle], \tag{3.6}$$

$$H_2(p)_{i,j} = \mathbb{E}_p[h(\boldsymbol{\theta}, \mathbf{x}_i) h(\boldsymbol{\theta}, \mathbf{x}_j)]. \tag{3.7}$$

We remark that our definition of the Gram matrix $\mathbf{H}$ is consistent with a similar definition in Mei et al. (2019).

4 Main Results

In this section we present our main results on the optimization and generalization of infinitely wide two-layer neural networks trained with noisy gradient descent. An overview of existing results in the mean-field setting and their connection to our results this paper is summarized in Table 1.

Table 1: An overview of existing results in the mean-field view.

	$\alpha = O(1)$	$\alpha \gg 1$
Optimization	Mei et al. (2018), Chizat and Bach (2018), Fang et al. (2019a)	Mei et al. (2019), Chizat et al. (2019), **This paper**
Generalization	**This paper**	**This paper**

We first introduce the following two assumptions.

Assumption 4.1. We assume that the data inputs are of bounded norm: $\|\mathbf{x}_i\|_2 \le 1$ for all $i \in [n]$.

Assumption 4.1 is a natural and mild assumption. Note that this assumption is much milder than the commonly used assumption $\|\mathbf{x}_i\|_2 = 1$ in the neural tangent kernel literature (Du et al., 2019b; Allen-Zhu et al., 2019b; Zou et al., 2019; Zou and Gu, 2019; Cao and Gu, 2019a; Ji and Telgarsky, 2020; Chen et al., 2019). We would also like to remark that all our results can be easily extended to the case when $\|\mathbf{x}_i\|_2 \le C$, $i \in [n]$ for some absolute constant C.

Assumption 4.2. We assume that the activation function has the form $h(\boldsymbol{\theta}, \mathbf{x}) = \widetilde{h}(\boldsymbol{\theta}^\top \mathbf{x})$, where $\widetilde{h}(\cdot)$ is a three-times differentiable function that satisfies the following smoothness properties:

$$
\begin{aligned}
&|\widetilde{h}(z)| \le G_1|z| + G_2, \quad |\widetilde{h}'(z)| \le G_3, \quad |\widetilde{h}''(z)| \le G_4,\\
&|\big(z\widetilde{h}'(z)\big)'| \le G_5, \quad |\widetilde{h}'''(z)| \le G_6,
\end{aligned}
$$

where $G_1, \dots, G_6$ are absolute constants.

Assumption 4.2 is by no means a strong assumption. $h(\boldsymbol{\theta}, \mathbf{x}) = \widetilde{h}(\boldsymbol{\theta}^\top \mathbf{x})$ is of the standard form in practical neural networks, and similar smoothness assumptions on $\widetilde{h}(\cdot)$ are standard in the mean field literature (Mei et al., 2018, 2019; Fang et al., 2019a,b). It is satisfied by most of the smooth activation functions including the sigmoid and hyper-tangent functions.

4.1 Optimization Results

In this section we study the optimization dynamics defined by PDE (3.3). We first introduce the following assumption.

Assumption 4.3. We assume that the Gram matrix of the neural tangent kernel is positive definite: $\lambda_{\min}(\mathbf{H}(p_0)) = \Lambda > 0$.

Assumption 4.3 is a rather weak assumption. In fact, Jacot et al. (2018) has shown that if $\|\mathbf{x}_i\|_2 = 1$ for all $i \in [n]$, Assumption 4.3 holds as long as each pair of training inputs $\mathbf{x}_1, \dots, \mathbf{x}_n$ are not parallel.

We are now ready to present our main result on the training dynamics of infinitely wide neural networks.

Theorem 4.4. Let $\Lambda = \lambda_{\min}(\mathbf{H}(p_0)) > 0$. Define $\lambda_0 = \sqrt{\Lambda/n}$. Under Assumptions 4.1, 4.2 and 4.3, there exist

$$R = \min\left\{\sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}, [\mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(1/\lambda_0))1/\lambda_0]^{-1}\right\}$$

such that if

$$\alpha \geq 8\sqrt{L(p_0)A_2^2 + \lambda A_1^2}\lambda_0^{-2}R^{-1}\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}, \tag{4.1}$$

then for all $t \in [0, +\infty)$, the following results hold:

$$L(p_t) \leq 2\exp(-2\alpha^2\lambda_0^2 t)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}$$
$$D_{\mathrm{KL}}(p_t||p_0) \leq 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4},$$

where $A_1 = 2(G_1\sigma_u^{-2} + G_3\sigma_{\boldsymbol{\theta}}^{-2})(\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2) + 2(G_2\sigma_u^{-2} + G_4)\sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2}$ and $A_2 = 2\big[((G_1 + G_3)/\sigma_u^2 + (G_3 + G_5)/\sigma_{\boldsymbol{\theta}}^2 + G_6)2\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + G_2/\sigma_u^2 + G_4\big]\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}$.

Remark 4.5. The KL-divergence bound in Theorem 5.4 increases with λ, which is counterintuitive because by (3.4), λ is the regularization parameter on $D_{\mathrm{KL}}(p_t||p_0)$. We would like to clarify that our bound is correct, and is preferable since here we are interested in the case when α is very large. A trivial bound on $D_{\mathrm{KL}}(p_t||p_0)$ that decreases in λ can be easily derived as $D_{\mathrm{KL}}(p_t||p_0) \leq \lambda^{-1}L(p_0)$ due to the fact that $Q(p_t)$ is monotonically decreasing (See Lemma A.5 in the appendix) and ϕ is non-negative.

Remark 4.6. Theorem 4.4 shows that the loss of the neural network converges linearly up to $O(\lambda^2\lambda_0^{-2}\alpha^{-2})$ accuracy, and the convergence rate essentially depends on the smallest eigenvalue of the NTK Gram matrix. This perfectly matches the results for square loss in the neural tangent kernel regime (Allen-Zhu et al., 2019b; Du et al., 2019b,a; Zou and Gu, 2019). Note that the algorithm we study is noisy gradient descent, and our objective function involves a weight decay regularizer, both of which cannot be handled by the standard technical tools used in the NTK regime. We also notice that some recent works Allen-Zhu et al. (2019a); Li et al. (2019) have studied the training of neural networks that involves noises, using proof techniques similar to the neural tangent kernel regime. However, the training algorithms that involves noises in Allen-Zhu et al. (2019a) are heavily twisted algorithms, while our training algorithm is the very standard noisy gradient descent. In comparison with Li et al. (2019), although the authors studied noisy gradient descent with certain learning rate schedule, all the theoretical results are established on a toy example, instead of a general setting. Compared with these results, our work proposes a rather intuitive and clean framework that allows large gradient noises into the neural tangent kernel type analysis.

4.2 Connection with NTK-based Analysis

In this section we study the connection between the dynamics of the PDE (3.3) and the NTK-based kernel regression. For simplicity, we assume that $y_i \in \{\pm 1\}$ for all $i \in [n]$, which implies that $L(p_0) \leq 1$.

We then have the following corollary of Theorem 4.4.

Corollary 4.7. Under the same conditions as in Theorem 4.4, for all $t \in [0, +\infty)$, we have

$$\|\mathbf{H}(p_t) - \mathbf{H}(p_0)\|_{\infty,\infty} \leq \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(\lambda_0^2\alpha))\sqrt{(A_2^2 + A_1^2\lambda)}\lambda_0^{-2}\alpha^{-1}.$$

Corollary 4.7 shows that the entries of $\mathbf{H}(p_t)$ are close to the entries of $\mathbf{H}(p_0)$. Clearly, the bounds between $\mathbf{H}(p_t)$ and $\mathbf{H}(p_0)$ in various norms can be then derived based on standard matrix perturbation results. It is also worth noting that the bound does not depend on time t, meaning that as long as α is large enough, the kernel throughout training is always close to the kernel defined at initialization, which recovers the key observation in NTK-based analysis.

We also study how close the neural network function $f(p_t, \mathbf{x})$ is to its NTK-based counterpart. We define $\mathbf{f}(t) = (f(p_t, \mathbf{x}_1), \ldots, f(p_t, \mathbf{x}_n))^\top$, where p_t is the solution of (3.3). We also define $\mathbf{f}_{\mathrm{NTK}}(t)$ to be the function value vector corresponding to the training based on NTK (Du et al., 2019b; Mei et al., 2018):

$$\frac{d[\mathbf{f}_{\mathrm{NTK}}(t) - \mathbf{y}]}{dt} = -\frac{2\alpha^2}{n}\mathbf{H}(p_0)[\mathbf{f}_{\mathrm{NTK}}(t) - \mathbf{y}],\ \mathbf{f}_{\mathrm{NTK}}(0) = \mathbf{0}.$$

The following corollary gives the bound on the distance between $\mathbf{f}(t)$ and $\mathbf{f}_{\mathrm{NTK}}(t)$.

Corollary 4.8. Under the same conditions as in Theorem 4.4, for all time $t \in [0, +\infty)$, we have

$$\frac{1}{n}\|\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)\|_2^2 \leq \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(\lambda_0^2\alpha)) \cdot (A_2^2 + \lambda A_1^2)\lambda_0^{-8}\alpha^{-2} + 2A_1^2\lambda^2\lambda_0^{-4}\alpha^{-2}.$$

where A_1 and A_2 are defined in Theorem 4.4.

Remark 4.9. Compared with Theorem 4 in Mei et al. (2019), Theorem 4.4 in this paper provides a stronger result in the following aspects: (i) Mei et al. (2019) only considers the setting without gradient noises or weight decay regularizer, while our result demonstrates that mean-field and NTK dynamics can be unified even with gradient noises and appropriate regularizer. Note that in our setting the network weights $(u_t, \boldsymbol{\theta}_t)$ is no longer close to their initialization $(u_0, \boldsymbol{\theta}_0)$[3], which, to the best of our knowledge, is beyond the scope of most existing NTK-based analyses; and (ii) Although the bound on $\|\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)\|_2$ given by Mei et al. (2019) decreases in α, it increases in t, and explodes as time goes to infinity. In comparison, our result gives a uniform bound on $\|\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)\|_2$ that does not increase in t.

4.3 Generalization Bounds for Large α

In this section, we study the generalization of the neural network trained by noisy gradient descent.

Motivated by the observation that noisy gradient descent with weight decay regularization is equivalent to minimizing the objective function with KL-divergence regularization with respect to the initialization distribution (See (3.4) and the corresponding discussion), we study the generalization bound for infinitely wide neural networks with bounded KL-divergence to the initialization distribution. In specific, for any $M > 0$, define

$$\mathcal{F}_{\mathrm{KL}}(M) = \big\{f(p, \mathbf{x}) : D_{\mathrm{KL}}(p\|p_0) \leq M\big\}. \tag{4.2}$$

The following theorem provides a generalization error bound for functions in $\mathcal{F}_{\mathrm{KL}}(M)$.

[3] This can be seen by checking Algorithm 1 or the corresponding stochastic differential equation.

Theorem 4.10. Suppose that the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from an unknown but fixed distribution $\mathcal{D}$. Let $\ell(y', y)$ be the loss function that is 1-Lipschitz in its first argument and satisfies $\ell(y, y) = 0$, $|\ell(y', y)| \leq 1$. Then under Assumptions 4.1 and 4.2, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\mathbb{E}_{\mathcal{D}}[\ell(f(\mathbf{x}), y)] \leq \mathbb{E}_S[\ell(f(\mathbf{x}), y)] + \frac{B_1 M^{1/2} \alpha}{\sqrt{n}} + B_2 M \alpha + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

for all $f \in \mathcal{F}_{\mathrm{KL}}(M)$, where

$$B_1 = \Big[4\sqrt{2G_1^2 \sigma_{\boldsymbol{\theta}}^2 d + G_2^2} + 2\sqrt{2} G_3 \sigma_u\Big] \max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} + 8 G_3 \sigma_u \max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} \cdot \sqrt{\log(n)},$$
$$B_2 = 40 G_3 \max\{\sigma_u^2, \sigma_{\boldsymbol{\theta}}^2\} + 16 G_5 \sigma_u \max\{\sigma_u^2, \sigma_{\boldsymbol{\theta}}^2\} \cdot \sqrt{\log(\sigma_u/(8\max\{\sigma_u^2, \sigma_{\boldsymbol{\theta}}^2\} M)}.$$

Theorem 4.10 provides a generalization bound for infinitely wide neural networks $f(p, \mathbf{x})$ defined in (3.1) with p close to initialization p_0. Note that when α is large, Theorem 4.4 suggests that p_t on the optimization trajectory roughly satisfies a bound $D_{\mathrm{KL}}(p_t || p_0) \leq \widetilde{O}(\lambda_0^{-4} \alpha^{-2})$ for all $t \in [0, +\infty)$. Plugging $M = \widetilde{O}(\lambda_0^{-4} \alpha^{-2})$ into the result of Theorem 4.10 gives a generalization bound as follows

$$\mathbb{E}_S[\ell(f(\mathbf{x}), y)] + \widetilde{O}(\lambda_0^{-2} n^{-1/2} + \lambda_0^{-4} \alpha^{-1}). \tag{4.3}$$

Clearly, this bound does not increase in α, which indicates that choosing a large scaling factor α will not hurt generalization. This matches with the generalization bounds derived in the neural tangent kernel regime (Arora et al., 2019a; Cao and Gu, 2020, 2019a; Ji and Telgarsky, 2020; Chen et al., 2019).

It is also worth noting that although the bound (4.3) does not increase in α, it may not have proper dependency in the sample size n, because λ_0 depends on n. However, this is natural because the KL-divergence bound $D_{\mathrm{KL}}(p_t || p_0) \leq \widetilde{O}(\lambda_0^{-4} \alpha^{-2})$ we use in the discussion above comes from Theorem 4.4, which makes no assumption on the data distribution. Suppose that the labels are simply Rademacher random variables and are independent of inputs (which is covered by Theorem 4.4), then the test error of any classier can obviously be at best a constant (Arora et al., 2019a; Cao and Gu, 2020, 2019a). In the following, we provide a corollary of Theorem 4.10, and demonstrate that under certain data distribution assumptions, the generalization error of the neural network function trained by noisy gradient descent with weight decay matches standard statistical rate.

For simplicity we consider the binary classification problem, where $y \in \{\pm 1\}$. We denote $\ell^{0\text{-}1}(y', y) = \mathbb{1}\{y' y < 0\}$.

Corollary 4.11. Suppose that the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from an unknown but fixed distribution $\mathcal{D}$, and there exists a true distribution p_{true} with $D_{\chi^2}(p_{\mathrm{true}} || p_0) < \infty$, such that

$$y = \int u h(\boldsymbol{\theta}, \mathbf{x}) p_{\mathrm{true}}(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du$$

for all $(\mathbf{x}, y) \in \mathrm{supp}(\mathcal{D})$. Let p^* be the minimizer of the energy functional (3.4), then under

Assumptions 4.1 and 4.2, provided that

$$\alpha \geq \sqrt{nD_{\chi^2}(p_{\text{true}}||p_0)} \cdot \max\{2\sqrt{\lambda}, 1\},$$

for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*, \mathbf{x}), y)] \leq 2(B_1 + B_2) \cdot \sqrt{\frac{D_{\chi^2}(p_{\text{true}}||p_0)}{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}},$$

where B_1 and B_2 are defined in Theorem 4.10.

Remark 4.12. Corollary 4.11 shows that if the target function is in the function class

$$\mathcal{F}_{\chi^2} = \bigg\{ \int uh(\boldsymbol{\theta}, \mathbf{x})p(\boldsymbol{\theta}, u)d\boldsymbol{\theta}du : D_{\chi^2}(p||p_0) < \infty \bigg\},$$

then it can be learned by over-parameterized two-layer neural networks trained with noisy gradient descent up to standard $O(1/\sqrt{n})$ accuracy, and the sample complexity depends on the χ^2-divergence between the initialization distribution and the distribution defining the target function. It is worth noting that $\mathcal{F}_{\chi^2}$ is seemingly different from the function class that can be learned by NTK, which is the NTK-induced reproducing kernel Hilbert space. We believe that studying the connection and difference between these two function classes is an interesting and important future work.

4.4 Generalization Bounds for Small α

Our generalization bounds given by Theorem 4.10 and Corollary 4.11 aim to provide a bound on the expected error for large scaling parameter α, with a focus on making the bound non-increasing in α. Although they work as a direct counterpart of the generalization results in the neural tangent kernel regime, Theorem 4.10 and Corollary 4.11 cannot cover the standard mean-field setting where $\alpha = O(1)$. In this section, we study the generalization bound of infinitely wide neural networks trained with noisy gradient descent in the setting where α is small (for example $\alpha = O(1)$). Let $\mathcal{F}_{\text{KL}}(M)$ be the function class defined in (4.2). We have the following theorem.

Theorem 4.13. Suppose that the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from an unknown but fixed distribution $\mathcal{D}$. Let $\ell(y', y)$ be the loss function that is 1-Lipschitz in its first argument and satisfies $\ell(y, y) = 0$, $|\ell(y', y)| \leq 1$. If $M \leq 1/2$, and there exists a constant G_7 such that $h(\boldsymbol{\theta}, \mathbf{x}) \leq G_7$ for all $\boldsymbol{\theta}$ and $\mathbf{x}$, then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathcal{D}}[\ell(f(\mathbf{x}), y)] \leq \mathbb{E}_S[\ell(f(\mathbf{x}), y)] + 4\alpha G_7 \sigma_u \sqrt{\frac{M}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

for all $f \in \mathcal{F}_{\text{KL}}(M)$.

Theorem 4.13 is the counterpart of Theorem 4.10 for small α. Compared with Theorem 4.10, Theorem 4.13 removes the term $O(M\alpha)$, at the expense of requiring $h(\boldsymbol{\theta}, u)$ to be bounded. We also present the following corollary of Theorem 4.13, which is a counterpart of Corollary 4.11 for small α.

Corollary 4.14. Suppose that the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from an unknown but fixed distribution $\mathcal{D}$, and there exists a true distribution p_{true} with $D_{\text{KL}}(p_{\text{true}}||p_0) < \infty$, such that

$$y = \int u h(\boldsymbol{\theta}, \mathbf{x}) p_{\text{true}}(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du$$

for all $(\mathbf{x}, y) \in \text{supp}(\mathcal{D})$. Further assume that there exists a constant G_7 such that $h(\boldsymbol{\theta}, \mathbf{x}) \leq G_7$ for all $\boldsymbol{\theta}$ and $\mathbf{x}$. Let p^* be the minimizer of the energy functional (3.4). If the regularization parameter $\lambda \leq \alpha/(4nD_{\text{KL}}(p_{\text{true}}||p_0))$, then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*, \mathbf{x}), y)] \leq 8G_7\sigma_u\sqrt{\frac{\alpha D_{\text{KL}}(p_{\text{true}}||p_0)}{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Remark 4.15. Corollary 4.14 shows that if α is small, two-layer neural networks trained by noisy gradient descent can learn the function class

$$\mathcal{F}_{\text{KL}} = \bigg\{ \int u h(\boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du : D_{\text{KL}}(p||p_0) < \infty \bigg\}.$$

In comparison, our result in Corollary 4.11 for large α shows that in that setting neural network can learn the function class $\mathcal{F}_{\chi^2}$. Since $\mathcal{F}_{\chi^2} \subsetneq \mathcal{F}_{\text{KL}}$, it seems that the regime with small α may outperform the one with large α. However, we must note that this argument is not rigorous, as Corollary 4.11 does not show that when α is large enough neural networks can *only* learn $\mathcal{F}_{\chi^2}$. We leave a more thorough study on this problem as a future work.

5 Proof of the Main Results

In this section, we present the proofs for the theorems and corollaries in Section 4.

5.1 Proof of Theorem 4.4

We first introduce the following lemma on the smallest eigenvalue of $\mathbf{H}(p)$ for distribution p close to p_0 in 2-Wasserstein distance. We remark that making the result of Lemma 5.1 hold is the major motivation of our definition of R.

Lemma 5.1. Under Assumptions 4.1, 4.2 and 4.3, for any distribution p with $\mathcal{W}_2(p, p_0) \leq R$, we have $\lambda_{\min}(\mathbf{H}(p)) \geq \Lambda/2$, where R is defined in Theorem 4.4.

Define $t^* = \inf\{t \geq 0 : \mathcal{W}_2(p_t, p_0) > R\}$, and denote $t^* = +\infty$ if the $\{t \geq 0 : \mathcal{W}_2(p_t, p_0) > R\} = \emptyset$. Then by definition, for $t \leq t^*$ we have $\mathcal{W}_2(p_t, p_0) \leq R$. The following two lemmas provide bounds on the loss function value and the KL-divergence to p_0 throughout training.

Lemma 5.2. Under Assumptions 4.1, 4.2 and 4.3, for any $t \leq t^*$, it holds that

$$\sqrt{L(p_t)} \leq \exp(-\alpha^2\lambda_0^2 t)\sqrt{L(p_0)} + A_1\lambda\alpha^{-1}\lambda_0^{-2},$$

where A_1 is defined in Theorem 4.4.

Lemma 5.3. Under Assumptions 4.1, 4.2 and 4.3, for any $t \leq t^*$,

$$D_{\text{KL}}(p_t||p_0) \leq 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4},$$

where A_1 and A_2 are defined in Theorem 4.4.

We also introduce the following Talagrand inequality (see Corollary 2.1 in Otto and Villani (2000) and Theorem 9.1.6 in Bakry et al. (2013)), which is based on the fact that p_0 is a Gaussian distribution.

Lemma 5.4 (Otto and Villani (2000))**.** The probability measure $p_0(\boldsymbol{\theta}, u) \propto \exp[-u^2/(2\sigma_u)^2 - \|\boldsymbol{\theta}\|_2^2/(2\sigma_{\boldsymbol{\theta}}^2)]$ satisfies following Talagrand inequality

$$\mathcal{W}_2(p, p_0) \leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} D_{\mathrm{KL}}(p||p_0)^{1/2}.$$

The major purpose of Lemma 5.4 is to build a connection between Lemmas 5.1, 5.2 and 5.3 based on 2-Wasserstein distance and the energy functional (3.4) which is regularized by the KL-divergence.

We are now ready to give the proof of Theorem 4.4.

Proof of Theorem 4.4. By the definition of t^*, for any $t \leq t^*$, we have

$$\begin{aligned}
\mathcal{W}_2(p_t, p_0) &\leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} D_{\mathrm{KL}}(p_t||p_0)^{1/2} \\
&\leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}\big(4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4}\big)^{1/2} \\
&\leq R/2,
\end{aligned}$$

where the first inequality is by Lemma 5.4, the second inequality is by Lemma 5.3 ,and the third inequality is due to the choice of α in (4.1).

This deduces that the set $\{t \geq 0 : \mathcal{W}_2(p_t, p_0) > R\}$ is empty and $t^* = \infty$, because otherwise $\mathcal{W}_2(p_{t^*}, p_0) = R$ by the continuity of 2-Wasserstein distance. Therefore the results of Lemmas 5.2 and 5.3 hold for all $t \in [0, +\infty)$. Squaring both sides of the result of Lemma 5.2 and applying Jensen's inequality gives

$$L(p_t) \leq 2\exp(-2\alpha^2\lambda_0^2 t)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}.$$

This completes the proof. $\square$

5.2 Proof of Corollary 4.7

To prove Corollary 4.7, we first introduce several lemmas. The following two lemmas characterize the difference between the neural tangent kernel Gram matrices defined with p_0 and some other distribution p that is close to p_0 in 2-Wasserstein distance.

Lemma 5.5. Under Assumptions 4.1 and 4.2, for any distribution p with $\mathcal{W}_2(p, p_0) \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2}$ and any $r > 0$,

$$\|\mathbf{H}_1(p) - \mathbf{H}_1(p_0)\|_{\infty,\infty} \leq G_3^2\Big[\sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2} + 2r^2G_3G_4\Big]\mathcal{W}_2(p, p_0) + 2G_3^2\mathbb{E}_{p_0}[u_0^2\,\mathbb{1}(|u_0 \geq r|)].$$

Lemma 5.6. Under Assumptions 4.1 and 4.2, for any distribution p with $\mathcal{W}_2(p, p_0) \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2}$,

$$\|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \leq \Big[4G_1G_3\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + 2G_2G_3\Big]\mathcal{W}_2(p, p_0).$$

The following lemma gives a tail bound with respect to our initialization distribution p_0, which we frequently utilize for truncation arguments.

Lemma 5.7. The initialization distribution p_0 satisfies the following tail bound:

$$\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] \leq \frac{\sigma_u^2}{2} \exp\bigg(-\frac{r^2}{4\sigma_u^2}\bigg).$$

The proof of Corollary 4.7 is presented as follows.

Proof of Corollary 4.7. Theorem 4.4 implies that

$$L(p_t) \leq 2\exp(-2\alpha^2\lambda_0^2 t)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}, \tag{5.1}$$

$$D_{\mathrm{KL}}(p_t||p_0) \leq 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4}. \tag{5.2}$$

For all $t < \infty$, we have

$$\begin{aligned}
\mathcal{W}_2(p_t, p_0) &\leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} D_{\mathrm{KL}}(p_t||p_0)^{1/2} \\
&\leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}(4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4})^{1/2} \\
&= 4\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}\sqrt{L(p_0)A_2^2 + \lambda A_1^2}\lambda_0^{-2}\alpha^{-1},
\end{aligned} \tag{5.3}$$

where the first inequality is by Lemma 5.4 and the second inequality is by (5.2). Further by the choice of α and R in theorem 4.4, we obtain that

$$\mathcal{W}_2(p_t, p_0) \leq R/2 \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}.$$

Therefore the conditions in Lemma 5.5 and Lemma 5.6 are satisfied. By Lemma 5.5, we have

$$\|\mathbf{H}_1(p_t) - \mathbf{H}_1(p_0)\|_{\infty,\infty} \leq G_3^2\Big[\sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2} + 2r^2G_3G_4\Big]\mathcal{W}_2(p, p_0) + 2G_3^2\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)]. \tag{5.4}$$

Choose $r = 2\sigma_u\sqrt{\log(\sigma_u^2\lambda_0^2\alpha/2)}$. Then by Lemma 5.7,

$$\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] \leq \alpha^{-1}\lambda_0^{-2}. \tag{5.5}$$

The definition of A_2 in theorem 4.4 implies

$$A_2 = 2\big[\big((G_1 + G_3)/\sigma_u^2 + (G_3 + G_5)/\sigma_{\boldsymbol{\theta}}^2 + G_6\big)2\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2} + G_2/\sigma_u^2 + G_4\big]\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} \geq G_3 \tag{5.6}$$

Combine (5.6) and (5.5) we have

$$2G_3^2\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] \leq 2G_3A_2\lambda_0^{-2}\alpha^{-1} \leq 2G_3\sqrt{(A_2^2 + A_1^2\lambda)}\lambda_0^{-2}\alpha^{-1}. \tag{5.7}$$

Plugging (5.7) and (5.3) into (5.4) and applying $L(p_0) \leq 1$ then gives

$$\|\mathbf{H}_1(p_t) - \mathbf{H}_1(p_0)\|_{\infty,\infty} \leq \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(\lambda_0^2\alpha))\sqrt{(A_2^2 + A_1^2\lambda)}\lambda_0^{-2}\alpha^{-1}. \tag{5.8}$$

Similarly, by Lemma 5.6, we have

$$\|\mathbf{H}_2(p_t) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \le \Big[4G_1G_3\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + 2G_2G_3\Big]\mathcal{W}_2(p, p_0), \tag{5.9}$$

Plugging (5.3) into (5.9) and applying $L(p_0) \le 1$, we obtain

$$\|\mathbf{H}_2(p_t) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \le \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}})\sqrt{(A_2^2 + A_1^2\lambda)} \cdot \alpha^{-1}\lambda_0^{-2}. \tag{5.10}$$

Since $\mathbf{H}(p_t) = \mathbf{H}_1(p_t) + \mathbf{H}_2(p_t)$, combining (5.8) and (5.10), we get

$$\begin{aligned}\|\mathbf{H}(p_t) - \mathbf{H}(p_0)\|_{\infty,\infty} &\le \|\mathbf{H}_1(p_t) - \mathbf{H}_1(p_0)\|_{\infty,\infty} + \|\mathbf{H}_2(p_t) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \\ &\le \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(\lambda_0^2\alpha))\sqrt{(A_2^2 + A_1^2\lambda)}\lambda_0^{-2}\alpha^{-1}.\end{aligned}$$

This completes the proof. □

5.3 Proof of Corollary 4.8

Here we present the proof of Corollary 4.8. The proof is based on the following lemma, which characterizes the dynamic of $\mathbf{f}(t)$ in a form that is directly comparable with the definition of $\mathbf{f}_{\mathrm{NTK}}(t)$.

Lemma 5.8. The dynamic of the residual could be written as

$$\frac{d[\mathbf{f}(t) - \mathbf{y}]}{dt} = -\frac{2\alpha^2}{n}\mathbf{H}(p_t)[\mathbf{f}(t) - \mathbf{y}] - \alpha\lambda\mathbf{I}(t),$$

where $\mathbf{I}(t)_i = \mathbb{E}_{p_t}\big[u_t h(\boldsymbol{\theta}_t, \mathbf{x}_i)/\sigma_u^2 + u_t\nabla h(\boldsymbol{\theta}_t, \mathbf{x}_i) \cdot \boldsymbol{\theta}_t/\sigma_{\boldsymbol{\theta}}^2 - u_t\Delta h(\boldsymbol{\theta}_t, \mathbf{x}_i)\big]$.

The following lemma essensially gives a bound on $\mathbf{I}(t)_i$ defined in Lemma 5.8.

Lemma 5.9. Under Assumptions 4.1 and 4.2, for all $\mathcal{W}(p, p_0) \le \sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2}$ and $\mathbf{x}$ the following inequality holds.

$$\mathbb{E}_p\big[uh(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + u\nabla h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - u\Delta h(\boldsymbol{\theta}, \mathbf{x})\big] \le A_1,$$

where A_1 is defined in Theorem 4.4.

We also have the following lemma, which states that the energy functional is monotonically decreasing during training. Note that this is not a new result, as it is to some extent an standard result, and has been discussed in Mei et al. (2018, 2019); Fang et al. (2019a).

Lemma 5.10. Let p_t be the solution of PDE (3.3). Then $Q(p_t)$ is monotonically deceasing, i.e.,

$$\frac{\partial Q(p_t)}{\partial t} \le 0. \tag{5.11}$$

The proof of Corollary 4.8 is given as follows.

Proof of Corollary 4.8. Theorem 4.4 implies that

$$
\begin{aligned}
L(p_t) &\leq 2\exp(-2\alpha^2\lambda_0^2 t)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}, \\
D_{\mathrm{KL}}(p_t||p_0) &\leq 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4}.
\end{aligned}
$$

For all $t < \infty$, we have

$$
\begin{aligned}
\mathcal{W}_2(p_t, p_0) &\leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\} D_{\mathrm{KL}}(p_t||p_0)^{1/2} \\
&\leq 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}(4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4})^{1/2} \\
&= 4\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}\sqrt{L(p_0)A_2^2 + \lambda A_1^2}\lambda_0^{-2}\alpha^{-1},
\end{aligned}
$$

where the first inequality is by Lemma 5.4, the second inequality follows by Theorem 4.4. Further by the choice of α and R in Theorem 4.4, we obtain that

$$
\mathcal{W}_2(p_t, p_0) \leq R/2 \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}.
$$

Now we can apply Corollary 4.7, which gives

$$
\|\mathbf{H}(p_t) - \mathbf{H}(p_0)\|_{\infty,\infty} \leq \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(\lambda_0^2\alpha))\sqrt{(A_2^2 + A_1^2\lambda)}\lambda_0^{-2}\alpha^{-1}.
$$

Then by standard matrix perturbation bounds, we have

$$
\|\mathbf{H}(p_t) - \mathbf{H}(p_0)\|_2 \leq \mathrm{poly}(\{G_i\}, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(\lambda_0^2\alpha))\sqrt{(A_2^2 + A_1^2\lambda)}\lambda_0^{-2}\alpha^{-1}n. \tag{5.12}
$$

By Lemma 5.8, the dynamic of $\mathbf{f}(t)$ is

$$
\frac{d[\mathbf{f}(t) - \mathbf{y}]}{dt} = -\frac{2\alpha^2}{n}\mathbf{H}(p_t)[\mathbf{f}(t) - \mathbf{y}] - \alpha\lambda\mathbf{I}(t), \tag{5.13}
$$

where $\mathbf{I}(t)_i = \mathbb{E}_{p_t}\big[u_t h(\boldsymbol{\theta}_t, \mathbf{x}_i)/\sigma_u^2 + u_t\nabla h(\boldsymbol{\theta}_t, \mathbf{x}_i)\cdot\boldsymbol{\theta}_t/\sigma_{\boldsymbol{\theta}}^2 - u_t\Delta h(\boldsymbol{\theta}_t, \mathbf{x}_i)\big]$. Combining (5.13) and the definition of $\mathbf{f}_{\mathrm{NTK}}(t)$, we get

$$
\frac{d[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)]}{dt} = -\frac{2\alpha^2}{n}[\mathbf{H}(p_t) - \mathbf{H}(p_0)][\mathbf{f}(t) - \mathbf{y}] - \frac{2\alpha^2}{n}\mathbf{H}(p_0)[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)] - \alpha\lambda\mathbf{I}(t). \tag{5.14}
$$

Denote $\epsilon(t) = \|\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)\|_2^2$. Taking inner product with $2[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)]$ on the both sides of (5.14) then gives

$$
\begin{aligned}
\frac{d\epsilon(t)}{dt} = &-\underbrace{\frac{4\alpha^2}{n}[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)]^\top[H(p_t) - H(p_0)][\mathbf{f}(t) - y]}_{I_1} - \underbrace{\frac{4\alpha^2}{n}[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)]^\top\mathbf{H}(p_0)[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)]}_{I_2} \\
&- \underbrace{2\alpha\lambda[\mathbf{f}(t) - \mathbf{f}_{\mathrm{NTK}}(t)]^\top\mathbf{I}(t)}_{I_3}.
\end{aligned} \tag{5.15}
$$

Now we bound I_1, I_2, I_3 respectively. First, for I_1 we have

$$\begin{aligned}|I_1| &\le \frac{4\alpha^2}{n}\|\mathbf{f}(t)-\mathbf{f}_{\mathrm{NTK}}(t)\|_2\|\mathbf{H}(p_t)-\mathbf{H}(p_0)\|_2\|\mathbf{f}(t)-y\|_2\\ &\le 4\alpha\lambda_0^{-2}\sqrt{A_2^2+\lambda A_1^2}\mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\sqrt{\epsilon(t)}\sqrt{nL(p_t)},\end{aligned}$$

where the first inequality is by (5.12) and the identities $\|\mathbf{f}(t)-\mathbf{f}_{\mathrm{NTK}}(t)\|_2=\sqrt{\epsilon(t)}$, $\|\mathbf{f}(t)-y\|_2=\sqrt{nL(p_t)}$. For I_2, note that by definition $\lambda_{min}(H(p_0))=\Lambda=n\lambda_0^2$. Therefore we have the following bound:

$$I_2\ge 4\alpha^2\lambda_0^2\epsilon(t).$$

Then we bound I_3,

$$|I_3|\le 2\alpha\lambda\|\mathbf{f}(t)-\mathbf{f}_{\mathrm{NTK}}(t)\|_2\|\mathbf{I}(t)\|_2\le 2\alpha\lambda\sqrt{\epsilon(t)}\sqrt{n}\|\mathbf{I}(t)\|_\infty\le 4\alpha\lambda\sqrt{\epsilon(t)}\sqrt{n}A_1,$$

where the second inequality is by inequality between 2-norm and infinity-norm and the third inequality is by Lemma 5.9 we have $\|\mathbf{I}(t)\|_\infty\le A_1$. Plugging the bounds of I_1, I_2, I_3 into (5.15) gives

$$\begin{aligned}\frac{d\epsilon(t)}{dt} &\le 4\alpha\lambda_0^{-2}\sqrt{A_2^2+\lambda A_1^2}\mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\sqrt{\epsilon(t)}\sqrt{nL(p_t)}-4\alpha^2\lambda_0^2\epsilon(t)+4\sqrt{\epsilon(t)}\alpha\lambda\sqrt{n}A_1\\ &=-4\alpha^2\lambda_0^2\sqrt{\epsilon(t)}\Big[\sqrt{\epsilon(t)}-\mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\lambda_0^{-4}\sqrt{A_2^2+\lambda A_1^2}\alpha^{-1}\sqrt{nL(p_t)}\\ &\qquad -\alpha^{-1}\lambda_0^{-2}\lambda A_1\sqrt{n}\Big].\end{aligned}\tag{5.16}$$

By Lemma 5.10, we know that Q is monotonically decreasing, which implies that

$$L(p_t)\le Q(p_t)\le Q(p_0)=L(p_0).\tag{5.17}$$

Plugging (5.17) into (5.16) gives

$$\begin{aligned}\frac{d\epsilon(t)}{dt} &\le -4\alpha^2\lambda_0^2\sqrt{\epsilon(t)}\Big[\sqrt{\epsilon(t)}-\mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\lambda_0^{-4}\sqrt{A_2^2+\lambda A_1^2}\alpha^{-1}\sqrt{nL(p_0)}\\ &\qquad -\alpha^{-1}\lambda_0^{-2}\lambda A_1\sqrt{n}\Big].\end{aligned}\tag{5.18}$$

Note that by (5.13) and the definition of $\mathbf{f}_{\mathrm{NTK}}(t)$, we have $\epsilon(0)=0$. Therefore (5.18) implies that for all time t,

$$\sqrt{\epsilon(t)}\le \mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\lambda_0^{-4}\sqrt{A_2^2+\lambda A_1^2}\alpha^{-1}\sqrt{nL(p_0)}+\alpha^{-1}\lambda_0^{-2}\lambda A_1\sqrt{n}.$$

Squaring both sides and dividing them by n, we obtain

$$\begin{aligned}\frac{1}{n}\|(\mathbf{f}(t)-\mathbf{f}_{\mathrm{NTK}}(t)^2\|_2 &\le \Big[\mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\lambda_0^{-4}\sqrt{A_2^2+\lambda A_1^2}\alpha^{-1}\sqrt{L(p_0)}+\alpha^{-1}\lambda_0^{-2}\lambda A_1\Big]^2\\ &\le \mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\lambda_0^{-8}(A_2^2+\lambda A_1^2)\alpha^{-2}L(p_0)+2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}\\ &\le \mathrm{poly}(\{G_i\},\sigma_u,\sigma_{\boldsymbol{\theta}},\log(\alpha\lambda_0^2))\lambda_0^{-8}(A_2^2+\lambda A_1^2)\alpha^{-2}+2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4},\end{aligned}$$

where the second inequality is by Jensen's inequality and the last inequality is by $L(p_0) \leq 1$. This completes the proof. $\square$

5.4 Proof of Theorem 4.10

In this subsection, we present the proof of Theorem 4.10. The first step of our proof is to convert the function class defined with the KL-divergence bound to a function class defined by Wasserstein metric. Lemma 5.4 motivates us to study the generalization bound of the function class

$$\mathcal{F}_{\mathcal{W}_2}(M') = \{f(p, \mathbf{x}) : \mathcal{W}_2(p, p_0) \leq M'\}.$$

We therefore consider the Rademacher complexity of

$$\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M')) = \mathbb{E}_{\boldsymbol{\xi}}\Bigg[\sup_{f \in \mathcal{F}_{\mathcal{W}_2}(M')} \frac{1}{n}\sum_{i=1}^n \xi_i f(\mathbf{x}_i)\Bigg],$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables. The bound on $\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M'))$ is provided in the following lemma, which is based on an "almost linear" property of $f(p, \mathbf{x})$ when $\mathcal{W}_2(p, p_0)$ is small.

Lemma 5.11. For any $M' > 0$,

$$\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M')) \leq \frac{B_1 M' \alpha}{\sqrt{n}} + B_2 M'^2 \alpha,$$

where

$$\widehat{B}_1 = \sqrt{2G_1^2\sigma_{\boldsymbol{\theta}}^2 d + G_2^2} + G_3\sigma_u/\sqrt{2} + 2G_3\sigma_u\sqrt{\log(n)},\ \ \widehat{B}_2 = 5G_3 + 2G_4\sigma_u\sqrt{\log(\sigma_u/(2M'^2))}.$$

Proof of Theorem 4.10. By Lemma 5.4, we have

$$\mathcal{F}_{\mathrm{KL}}(M) \subseteq \mathcal{F}_{\mathcal{W}_2}(2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}M^{1/2}).$$

Applying Lemma 5.11 with $M' = 2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}M^{1/2}$ gives

$$\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M')) \leq \frac{\widehat{B}_1 M' \alpha}{\sqrt{n}} + \widehat{B}_2 M'^2 \alpha = \frac{B_1 M^{1/2}\alpha}{2\sqrt{n}} + B_2 M\alpha/2,$$

Now by the standard properties of Rademacher complexity (Bartlett and Mendelson, 2002; Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014), we have

$$\mathbb{E}_{\mathcal{D}}[\ell(f(\mathbf{x}), y)] \leq \mathbb{E}_S[\ell(f(\mathbf{x}), y)] + \frac{B_1 M^{1/2}\alpha}{\sqrt{n}} + B_2 M\alpha + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

for all $f \in \mathcal{F}(M) \subseteq \mathcal{F}_{\mathcal{W}_2}(2\max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}M^{1/2})$. This completes the proof. $\square$

5.5 Proof of Corollary 4.11

Proof of Corollary 4.11. Throughout the proof, we denote $\overline{D} = D_{\chi^2}(p_{\text{true}}||p_0) < \infty$ to simplify the notation. Define

$$\widehat{p} = \frac{\alpha - 1}{\alpha} \cdot p_0 + \frac{1}{\alpha} \cdot p_{\text{true}}.$$

Then obviously we have $\int \widehat{p}(\theta, u) du d\theta = 1$, $\widehat{p}(\theta, u) \geq 0$, meaning that $\widehat{p}$ is a well-defined density function. Moreover, the training loss of $\widehat{p}$ can be calculated as follows:

$$L(\widehat{p}) = \mathbb{E}_S \bigg[\alpha \int u h(\theta, \mathbf{x}) \widehat{p}(u, \theta) du d\theta - y \bigg]^2 = \mathbb{E}_S \bigg(0 + \alpha \cdot \frac{1}{\alpha} y - y \bigg)^2 = 0. \tag{5.19}$$

Similarly, we can calculate the χ^2-divergence between $\widehat{p}$ and p_0:

$$D_{\chi^2}(\widehat{p}||p_0) = \int \bigg[\frac{\widehat{p}(\boldsymbol{\theta}, u)}{p_0(\boldsymbol{\theta}, u)} - 1 \bigg]^2 p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du = \int \bigg[\frac{\alpha - 1}{\alpha} + \frac{p_{\text{true}}(\boldsymbol{\theta}, u)}{\alpha p_0(\boldsymbol{\theta}, u)} - 1 \bigg]^2 p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du = \alpha^{-2} \overline{D},$$

where we remind the readers that we define $\overline{D} = D_{\chi^2}(p_{\text{true}}||p_0)$ to shorten the notation. Now by the fact that KL-divergence is upper bounded by the χ^2-divergence, we have

$$D_{\text{KL}}(\widehat{p}||p_0) \leq D_{\chi^2}(\widehat{p}||p_0) = \alpha^{-2} \overline{D}. \tag{5.20}$$

Invoking the definition of the energy function $Q(p)$ in (3.4) gives

$$Q(p^*) \leq Q(\widehat{p}) = L(\widehat{p}) + \lambda D_{\text{KL}}(\widehat{p}||p_0) \leq \alpha^{-2} \lambda \overline{D},$$

where the first inequality follows by the optimality of p^*, and we plug in (5.19) and (5.20) to obtain the second inequality. Applying the definition of $Q(p)$ again gives the following two bounds:

$$L(p^*) = \frac{1}{n} \sum_{i=1}^{n} [f(p^*, \mathbf{x}_i) - y_i]^2 \leq \alpha^{-2} \lambda \overline{D}, \tag{5.21}$$

$$D_{\text{KL}}(p^*||p_0) \leq \alpha^{-2} \overline{D}. \tag{5.22}$$

Now we introduce the following ramp loss function, which is frequently used in the analysis of generalization bounds (Bartlett et al., 2017; Li et al., 2018a).

$$\ell_{\text{ramp}}(y', y) = \begin{cases} 0 & \text{if } y'y \geq 1/2, \\ -2y'y + 1, & \text{if } 0 \leq y'y < 1/2, \\ 1, & \text{if } y'y < 0. \end{cases} \tag{5.23}$$

Then by definition, we see that $\ell_{\text{ramp}}(y', y)$ is 2-Lipschitz in the first argument and satisfies $\ell(y, y) = 0$, $|\ell(y', y)| \leq 1$, since $y \in \{\pm 1\}$. Moreover, we also have

$$\ell^{0\text{-}1}(y', y) \leq \ell_{\text{ramp}}(y', y)$$

for all y' and y. Taking expectations over $\mathcal{D}$ we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*,\mathbf{x}),y)] &\leq 2\mathbb{E}_{\mathcal{D}}[\ell_{\mathrm{ramp}}(f(p^*,\mathbf{x}),y)/2] \\
&\leq \mathbb{E}_S[\ell_{\mathrm{ramp}}(f(\mathbf{x}),y)] + \frac{2B_1\overline{D}^{1/2}}{\sqrt{n}} + 2B_2\overline{D}\alpha^{-1} + 6\sqrt{\frac{\log(2/\delta)}{2n}}, \qquad (5.24)
\end{aligned}
$$

where the second inequality follows by (5.22) and the application of Theorem 4.10 to $\ell_{\mathrm{ramp}}(f(p^*,\mathbf{x}),y)/2$ with $M=\alpha^{-2}\overline{D}$. We now proceed to bound the empirical ramp loss utilizing (5.21). By (5.21), for any $i\in[n]$, we have

$$
|f(p^*,\mathbf{x}_i)-y_i|^2 \leq n\cdot\frac{1}{n}\sum_{i=1}^n[f(p^*,\mathbf{x}_i)-y_i]^2 \leq n\cdot\alpha^{-2}\lambda\overline{D} \leq 1/4,
$$

where the first inequality follows by simply upper bounding $|f(p^*,\mathbf{x}_i)-y_i|^2$ with the sum over all $i\in[n]$, and the second inequality follows by the assumption that $\alpha \geq 2\sqrt{n\lambda D_{\chi^2}(p_{\mathrm{true}}||p_0)}$. Since $y_i\in\{\pm1\}$ for all $i\in[n]$, we see that $f(p^*,\mathbf{x}_i)\cdot y_i \geq 1/2$ for all $i\in[n]$. Therefore by the definition of ramp loss we have

$$
\mathbb{E}_S[\ell_{\mathrm{ramp}}(f(\mathbf{x}),y)] = 0,
$$

and therefore by (5.24) we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*,\mathbf{x}),y)] &\leq \frac{2B_1\overline{D}^{1/2}}{\sqrt{n}} + 2B_2\overline{D}\alpha^{-1} + 6\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \frac{2B_1\overline{D}^{1/2}}{\sqrt{n}} + 2B_2\overline{D}\cdot\frac{1}{\sqrt{n\overline{D}}} + 6\sqrt{\frac{\log(2/\delta)}{2n}} \\
&= 2(B_1+B_2)\cdot\frac{\sqrt{\overline{D}}}{\sqrt{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}},
\end{aligned}
$$

where the second inequality follows by the assumption that $\alpha \geq \sqrt{nD_{\chi^2}(p_{\mathrm{true}}||p_0)}$. This completes the proof. $\square$

5.6 Proof of Theorem 4.13

We introduce the following bound on the Rademacher complexity $\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M))$. Note that applying Lemma 5.4 and Lemma 5.11 can also lead to a bound on $\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M))$, but here we propose a different bound which is more suitable for the case when α is small.

Lemma 5.12. Suppose that $|h(\boldsymbol{\theta},\mathbf{x})| \leq G_7$ for all $\boldsymbol{\theta}$ and $\mathbf{x}$, and $M\leq 1/2$. Then

$$
\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M)) \leq 2\alpha G_7\sigma_u\sqrt{\frac{M}{n}}.
$$

We give the following proof for Theorem 4.13, which is rather straight-forward given Lemma 5.12.

Proof of Theorem 4.13. By Lemma 5.12, we have

$$\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M)) \le 2\alpha G_7 \sigma_u \sqrt{\frac{M}{n}}.$$

Now by the standard properties of Rademacher complexity (Bartlett and Mendelson, 2002; Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014), we have

$$\mathbb{E}_{\mathcal{D}}[\ell(f(\mathbf{x}), y)] \le \mathbb{E}_S[\ell(f(\mathbf{x}), y)] + 4\alpha G_7 \sigma_u \sqrt{\frac{M}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

for all $f \in \mathcal{F}(M)$. This completes the proof. □

5.7 Proof of Corollary 4.14

The proof of Corollary 4.14 is similar to the proof of Corollary 4.11.

Proof of Corollary 4.14. Throughout the proof, we denote $\widehat{D} = D_{\mathrm{KL}}(p_{\mathrm{true}}||p_0) < \infty$ to simplify the notation. Set $\widehat{p} = (\alpha - 1)p_0/\alpha + p_{\mathrm{true}}/\alpha$. Then with the exact same proof as the proof of Corollary 4.11 (see (5.19)), we have

$$L(\widehat{p}) = 0. \tag{5.25}$$

Moreover, by the convexity of KL-divergence, we have

$$D_{\mathrm{KL}}(\widehat{p}||p_0) \le \frac{\alpha - 1}{\alpha} \cdot D_{\mathrm{KL}}(p_0||p_0) + \frac{1}{\alpha} \cdot D_{\mathrm{KL}}(p_{\mathrm{true}}||p_0) = \frac{\alpha - 1}{\alpha} \cdot 0 + \frac{1}{\alpha}\widehat{D} = \frac{1}{\alpha}\widehat{D}. \tag{5.26}$$

Therefore we have

$$Q(p^*) \le Q(\widehat{p}) = L(\widehat{p}) + \lambda D_{\mathrm{KL}}(\widehat{p}||p_0) = \alpha^{-1}\lambda\widehat{D},$$

where the first inequality is due to the optimality of p^*, and the first equality follows by the definition of $Q(p)$ in (3.4) and the definition of $\widehat{p}$. Applying the definition of $Q(p)$ again gives the following two bounds:

$$L(p^*) = \frac{1}{n}\sum_{i=1}^{n}[f(p^*, \mathbf{x}_i) - y_i]^2 \le \alpha^{-1}\lambda\widehat{D}, \tag{5.27}$$

$$D_{\mathrm{KL}}(p^*||p_0) \le \alpha^{-1}\widehat{D}. \tag{5.28}$$

Consider the same ramp loss function $\ell_{\mathrm{ramp}}(y', y)$ defined in (5.23) in the proof of Corollary 4.11. Then again, $\ell_{\mathrm{ramp}}(y', y)$ is 2-Lipschitz in the first argument, $\ell(y, y) = 0$, $|\ell(y', y)| \le 1$, and

$$\ell^{0\text{-}1}(y', y) \le \ell_{\mathrm{ramp}}(y', y)$$

for all y' and y. Taking expectations over $\mathcal{D}$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*,\mathbf{x}),y)] &\leq 2\mathbb{E}_{\mathcal{D}}[\ell_{\mathrm{ramp}}(f(p^*,\mathbf{x}),y)/2] \\ &\leq \mathbb{E}_S[\ell_{\mathrm{ramp}}(f(\mathbf{x}),y)] + 8G_7\sigma_u\sqrt{\frac{\alpha\widehat{D}}{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}} \end{aligned} \tag{5.29}$$

where the second inequality follows by (5.28) and the application of Theorem 4.13 to $\ell_{\mathrm{ramp}}(f(p^*,\mathbf{x}),y)/2$ with $M = \alpha^{-1}\widehat{D}$. Similar to the proof of Corollary 4.11, we aim to utilize (5.27) to establish an upper bound for $\mathbb{E}_S[\ell_{\mathrm{ramp}}(f(p^*,\mathbf{x}),y)]$. By (5.27), we have

$$|f(p^*,\mathbf{x}_i) - y_i|^2 \leq n\cdot L(p^*) \leq n\cdot\alpha^{-1}\lambda\widehat{D} \leq 1/4$$

for all $i\in[n]$, where we use the assumption that $\lambda \leq \alpha/(4nD_{\mathrm{KL}}(p_{\mathrm{true}}||p_0))$ to derive the second inequality. Therefore by $y_i\in\{\pm1\}$, $i\in[n]$ we see that $f(p^*,\mathbf{x}_i)\cdot y_i \geq 1/2$ for all $i\in[n]$. Therefore by the definition of ramp loss we have

$$\mathbb{E}_S[\ell_{\mathrm{ramp}}(f(\mathbf{x}),y)] = 0.$$

Plugging this result into (5.29) then yields

$$\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*,\mathbf{x}),y)] \leq 8G_7\sigma_u\sqrt{\frac{\alpha\widehat{D}}{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}},$$

which completes the proof. □

6 Conclusions and Future Work

In this paper we establish a connection between NTK and mean-field analyses, and demonstrate that if a large scaling factor is introduced into the network function, the whole training dynamic is similar to the dynamics of neural tangent kernel. This also leads to the linear convergence of noisy gradient descent up to certain accuracy. Compared with standard analysis in the neural tangent kernel regime, our work points out an important observation that as long as the distribution of parameters stay close to the initialization, it does not matter whether the parameters themselves are close to their initial values. We also establish generalization bounds for the neural networks trained with noisy gradient descent with weight decay regularization under different network scalings.

One interesting future direction is to extend our results to multi-layer networks, where the approach proposed by Fang et al. (2019b) might be leveraged. Due to the popularity of the non-smooth activation functions like ReLU, relaxing our assumption on the smoothness of the activation function can be an important problem to study. Further investigation on the function classes $\mathcal{F}_{\chi^2}$, $\mathcal{F}_{\mathrm{KL}}$ and their relation to the reproducing kernel Hilbert space defined by the neural tangent kernel is also a future direction of vital importance.

A Proof of Lemmas in Section 5

In this section we provide the proofs of lemmas we use in Section 5 for the proof of our main results. We first introduce the following notations:

$$\widehat{g}_1(t, \boldsymbol{\theta}, u) = -\alpha \mathbb{E}_S[\nabla_f \phi(f(p_t, \mathbf{x}), y) h(\boldsymbol{\theta}, \mathbf{x})], \tag{A.1}$$
$$\widehat{g}_2(t, \boldsymbol{\theta}, u) = -\alpha \mathbb{E}_S[\nabla_f \phi(f(p_t, \mathbf{x}), y) u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})]. \tag{A.2}$$

A.1 Proof of Lemma 5.1

Here we give the proof of Lemma 5.1. We first introduce the following lemma which summarizes some basic properties of the activation function $h(\boldsymbol{\theta}, u)$.

Lemma A.1. Under Assumptions 4.1 and 4.2, for all $\mathbf{x}$ and $\boldsymbol{\theta}$, it holds that $|h(\boldsymbol{\theta}, \mathbf{x})| \leq G_1 \|\boldsymbol{\theta}\|_2 + G_2$, $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \leq G_3$, $|\Delta h(\boldsymbol{\theta}, \mathbf{x})| \leq G_4$, $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_1, x) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_2, x)\|_2 \leq G_4 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$, $\|\nabla_{\boldsymbol{\theta}}\big(\nabla_{\boldsymbol{\theta}} h(\theta, \mathbf{x}) \cdot \boldsymbol{\theta}\big)\|_2 \leq G_5$, $\|\nabla_{\boldsymbol{\theta}} \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \leq G_6$.

We are now ready to provide the proof of Lemma 5.1.

Proof of Lemma 5.1. Here we first give the definition of R in Theorem 4.4 with specific polynomial dependencies.

$$\begin{aligned}
R &= \min\Big\{\sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}, [\mathrm{poly}(\{G_i\}_{i=1}^7, \sigma_u, \sigma_{\boldsymbol{\theta}}, \log(n/\Lambda))n/\Lambda]^{-1}\Big\} \\
&\leq \min\Big\{\sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}, \\
&\quad \Big(8G_3^2\sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2} + 64G_3G_4\log(8\Lambda^{-1}nG_3^2\sigma_u^2) + 16G_1G_3\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + 8G_2G_3\Big)^{-1} n^{-1}\Lambda\Big\}.
\end{aligned}$$

Note that the definition of R, the results for Lemmas 5.5 and 5.6 hold for all p with $\mathcal{W}_2(p, p_0) \leq R$. Now by Lemma 5.5, for any p with $\mathcal{W}_2(p, p_0) \leq R$ and any $r > 0$,

$$\|\mathbf{H}_1(p) - \mathbf{H}_1(p_0)\|_{\infty,\infty} \leq G_3^2 R\sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2} + 2r^2 G_3 G_4 R + 2G_3^2 \mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)]. \tag{A.3}$$

Choose $r = 2\sigma_u\sqrt{\log(8\Lambda^{-1}nG_3^2\sigma_u^2)}$, then by Lemma 5.7 we have

$$\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] \leq \frac{\Lambda}{16nG_3^2}. \tag{A.4}$$

Moreover, by the definition of R, we have

$$R \leq \Big(8G_3^2\sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2} + 16G_3G_4r^2\Big)^{-1} n^{-1}\Lambda. \tag{A.5}$$

Plugging the bounds on $\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)]$ and R given by (A.4) and (A.5) into (A.3) gives

$$\begin{aligned}
\|\mathbf{H}_1(p) - \mathbf{H}_1(p_0)\|_{\infty,\infty} &\leq G_3^2 R\sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2} + 2r^2 G_3 G_4 R + G_3^2 \mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \\
&\geq r|)] \leq \frac{\Lambda}{8n} + \frac{\Lambda}{8n} \\
&= \frac{\Lambda}{4n}.
\end{aligned} \tag{A.6}$$

By Lemma 5.6, for any distribution p with $\mathcal{W}_2(p, p_0) \le R$,

$$\|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \le \Big[4G_1G_3\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + 2G_2G_3\Big]R. \tag{A.7}$$

The definition of R also leads to the following bound:

$$R \le \Big(16G_1G_3\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + 8G_2G_3\Big)^{-1} n^{-1}\Lambda. \tag{A.8}$$

Therefore we can plug the bound (A.8) into (A.7), which gives

$$\|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \le \frac{\Lambda}{4n}. \tag{A.9}$$

Combining (A.6) and (A.9) further gives

$$\|\mathbf{H}(p) - \mathbf{H}(p_0)\|_{\infty,\infty} \le \|\mathbf{H}_1(p) - \mathbf{H}_1(p_0)\|_{\infty,\infty} + \|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \le \frac{\Lambda}{2n}.$$

Then by standard matrix perturbation bounds, we have $\lambda_{\min}(\mathbf{H}(p)) \ge \lambda_{\min}(\mathbf{H}(p_0)) - \|\mathbf{H}(p) - \mathbf{H}(p_0)\|_2 \ge \lambda_{\min}(\mathbf{H}(p_0)) - n\|\mathbf{H}(p) - \mathbf{H}(p_0)\|_{\infty,\infty} \ge \Lambda/2$, which finishes the proof. □

A.2 Proof of Lemma 5.2

Here we give the proof of Lemma 5.2. The following lemma summarizes some basic calculation on the training dynamics. Here we remind the readers that the definitions of $\widehat{g}_1(t, \boldsymbol{\theta}, u)$ and $\widehat{g}_2(t, \boldsymbol{\theta}, u)$ are given in (A.1) and (A.2) respectively.

Lemma A.2. Let p_t be the solution of PDE (3.3). Then the following identity holds.

$$\begin{aligned}\frac{\partial L(p_t)}{\partial t} &= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta} du - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta} d\boldsymbol{\theta} du \\ &\quad + \lambda\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du. \end{aligned} \tag{A.10}$$

Lemma A.2 decomposes the time derivative of $L(p_t)$ into several terms. The following two lemmas further provides bounds on these terms. Note that by the definition in (A.1) and (A.2), Lemma A.3 below essentially serves as a bound on the first two terms on the right-hand side of (A.10).

Lemma A.3. Under Assumptions 4.1, 4.2 and 4.3, let λ_0 be defined in Theorem 4.4. Then for $t \le t^*$, it holds that

$$\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\big[|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)h(\boldsymbol{\theta}, \mathbf{x})]|^2 + \|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})]\|_2^2\big] d\boldsymbol{\theta} du \ge \frac{\lambda_0^2}{2} L(p_t).$$

Lemma A.4. Under Assumptions 4.1 and 4.2, let A_1 be defined in Theorem 4.4. Then for $t \le t^*$,it holds that

$$\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du \le 2\alpha A_1\sqrt{L(p_t)}.$$

We now present the proof of Lemma 5.2, which is based on the calculations in Lemmas A.2, A.3 and A.4 as well as the application of Gronwall's inequality.

Proof of Lemma 5.2. By Lemma A.2, we have

$$\frac{\partial L(p_t)}{\partial t} = -\underbrace{\left[\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta} du + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta} d\boldsymbol{\theta} du\right]}_{I_1}$$
$$+ \underbrace{\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla \cdot \widehat{g}_1 - \nabla \cdot \widehat{g}_2] d\boldsymbol{\theta} du}_{I_2} ., \tag{A.11}$$

For I_1, we have

$$\begin{aligned} I_1 &= 4\alpha^2 \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\big[|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)h(\boldsymbol{\theta}, \mathbf{x})]|^2 + \|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})]\|_2^2\big] d\boldsymbol{\theta} du \\ &\geq 2\alpha^2\lambda_0^2 L(p_t), \end{aligned} \tag{A.12}$$

where the equality follows by the definitions of $\widehat{g}_1(t, \boldsymbol{\theta}, u)$, $\widehat{g}_2(t, \boldsymbol{\theta}, u)$ in (A.1), (A.1), and the inequality follows by Lemma A.3. For I_2, we directly apply Lemma A.4 and obtain

$$I_2 \leq 2A_1\alpha\lambda\sqrt{L(p_t)}. \tag{A.13}$$

Plugging the bounds (A.12) and (A.13) into (A.11) yields

$$\frac{\partial L(p_t)}{\partial t} \leq -2\alpha^2\lambda_0^2 L(p_t) + 2A_1\alpha\lambda\sqrt{L(p_t)}. \tag{A.14}$$

Now denote $V(t) = \sqrt{L(p_t)} - A_1\lambda\alpha^{-1}\lambda_0^{-2}$. Then (A.14) implies that[4]

$$\frac{\partial V(t)}{\partial t} \leq -\alpha^2\lambda_0^2 V(t).$$

By Gronwall's inequality we further get

$$V(t) \leq \exp(-\alpha^2\lambda_0^2 t)V(0).$$

By $V(0) = \sqrt{L(p_0)} - A_1\lambda\alpha^{-1}\lambda_0^{-2} \leq \sqrt{L(p_0)}$, we have

$$\sqrt{L(p_t)} \leq \exp(-\alpha^2\lambda_0^2 t)\sqrt{L(p_0)} + A_1\lambda\alpha^{-1}\lambda_0^{-2}. \tag{A.15}$$

This completes the proof. □

A.3 Proof of Lemma 5.3

In this subsection we present the proof of Lemma 5.3.

[4]The derivation we present here works as long as $L(p_t) \neq 0$. A more thorough but complicated analysis can deal with the case when $L(p_t) = 0$ for some t. However for simplicity we omit the more complicated proof, since loss equaling to zero is a trivial case for a learning problem.

Lemma A.5. Under Assumptions 4.1, 4.2 and 4.3, let λ_0 be defined in Theorem 4.4. Then for $t \le t^*$ the following inequality holds

$$D_{\mathrm{KL}}(p_t||p_0) \le 2A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 2A_2^2A_1^2\lambda^2\lambda_0^{-4}t^2.$$

If $\lambda \neq 0$, the KL distance bound given by Lemma A.5 depends on t, we can give a tighter bound by the monotonically deceasing property of $Q(p_t)$ given by Lemma 5.10.

Proof of Lemma 5.3. Notice that for $\lambda = 0$, Lemma A.5 directly implies the conclusion. So in the rest of the proof we consider the situation where $\lambda > 0$. Denote $t_0 = A_1^{-1}\alpha^{-1}\lambda^{-1}\sqrt{L(p_0)}$, we consider two cases $t_0 \ge t_*$ and $t_0 < t_*$ respectively.

If $t_0 \ge t_*$, then for $t \le t^*$ we have $t \le t_0$

$$\begin{aligned}
D_{\mathrm{KL}}(p_t||p_0) &\le 2A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 2A_2^2A_1^2\lambda^2\lambda_0^{-4}t^2 \\
&\le 2A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 2A_2^2A_1^2\lambda^2\lambda_0^{-4}t_0^2 \\
&= 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) \\
&\le 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4},
\end{aligned}$$

where the first inequality is by Lemma A.5 and the second inequality is by $t \le t_0$.

If $t_0 < t_*$, then for $t \le t_0$, we also have

$$D_{\mathrm{KL}}(p_t||p_0) \le 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) \le 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4}.$$

For $t_0 < t \le t_*$, consider $Q(p_t) = L(p_t) + \lambda D_{\mathrm{KL}}(p_t||p_0)$. The monotonically deceasing property of $Q(p_t)$ in Lemma 5.10 implies that,

$$D_{\mathrm{KL}}(p_t||p_0) \le \lambda^{-1}Q(p_t) \le \lambda^{-1}Q(p_{t_0}). \tag{A.16}$$

Now we bound $Q(p_{t_0})$. We first bound $L(p_{t_0})$. Squaring both sides of the result of Lemma 5.2 and applying Jensen's inequality now gives

$$L(p_t) \le 2\exp(-2\alpha^2\lambda_0^2 t)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}. \tag{A.17}$$

Plugging $t_0 = A_1^{-1}\alpha^{-1}\lambda^{-1}\sqrt{L(p_0)}$ into (A.17) gives

$$\begin{aligned}
L(p_{t_0}) &\le 2\exp(-2\alpha^2\lambda_0^2 t_0)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4} \\
&= 2\exp\big(-2A_1^{-1}\lambda^{-1}\alpha\lambda_0^2\sqrt{L(p_0)}\big)L(p_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4} \\
&\le 4A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4},
\end{aligned} \tag{A.18}$$

where the last inequality is by $\exp(-2z) = [\exp(-z)]^2 \le [1/z]^2$ for any $z > 0$. We then bound $D_{\mathrm{KL}}(p_{t_0}||p_0)$. By Lemma A.5, we have

$$D_{\mathrm{KL}}(p_{t_0}||p_0) \le 2A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 2A_2^2A_1^2\lambda^2\lambda_0^{-4}t_0^2 = 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0). \tag{A.19}$$

Plugging (A.18) and (A.19) into (A.16) gives

$$D_{\mathrm{KL}}(p_t||p_0) \le \lambda^{-1}Q(p_{t_0}) = \lambda^{-1}L(p_{t_0}) + D_{\mathrm{KL}}(p_{t_0}||p_0) \le 4A_2^2\alpha^{-2}\lambda_0^{-4}L(p_0) + 4A_1^2\lambda\alpha^{-2}\lambda_0^{-4}.$$

This completes the proof. $\square$

A.4 Proof of Lemma 5.5

The following lemma bounds the second moment of a distribution p that is close to p_0 in 2-Wasserstein distance.

Lemma A.6. For $\mathcal{W}_2(p, p_0) \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}$, the following bound holds:

$$\mathbb{E}_p(\|\boldsymbol{\theta}\|_2^2 + u^2) \leq 4\sigma_{\boldsymbol{\theta}}^2 \cdot d + 4\sigma_u^2$$

The following lemma is a reformulation of Lemma C.8 in Xu et al. (2018). For completeness, we provide its proof in Appendix B.

Lemma A.7. For $\mathcal{W}_2(p, p_0) \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}$, let $g(u, \boldsymbol{\theta}) : \mathbb{R}^{d+1} \to \mathbb{R}$ be a C^1 function such that

$$\sqrt{\nabla_u g(u, \boldsymbol{\theta})^2 + \|\nabla_{\boldsymbol{\theta}} g(u, \boldsymbol{\theta})\|^2} \leq C_1 \sqrt{u^2 + \|\boldsymbol{\theta}\|_2^2} + C_2, \forall \mathbf{x} \in \mathbb{R}^{d'}$$

for some constants $C_1, C_2 \geq 0$. Then

$$\big|\mathbb{E}_p[g(u, \boldsymbol{\theta})] - \mathbb{E}_{p_0}[g(u_0, \boldsymbol{\theta}_0)]\big| \leq \Big(2C_1 \sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2} + C_2\Big) \mathcal{W}_2(p, p_0).$$

Proof of Lemma 5.5. Let π^* be the optimal coupling of $\mathcal{W}_2(p, p_0)$. Then we have

$$\begin{aligned}
\big|H_1(p)_{i,j} - H_1(p_0)_{i,j}\big| &= \big|\mathbb{E}_{\pi^*}[u^2 \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_j)] - \mathbb{E}_{\pi^*}[u_0^2 \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_j)]\big| \\
&\leq \underbrace{\big|\mathbb{E}_{\pi^*}[(u^2 - u_0^2) \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j)]\big|}_{I_1} \\
&\quad + \underbrace{\Big|\mathbb{E}_{\pi^*}\big[u_0^2 \big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big)\big]\Big|}_{I_2}.
\end{aligned} \tag{A.20}$$

We first bound I_1 as follows.

$$\begin{aligned}
I_1 &\leq G_3^2 \mathbb{E}_{\pi^*}[|u^2 - u_0^2|] \\
&\leq G_3^2 \sqrt{\mathbb{E}_{\pi^*}[(u - u_0)^2]} \sqrt{\mathbb{E}_{\pi^*}[(u + u_0)^2]} \\
&\leq G_3^2 \mathcal{W}_2(p, p_0) \sqrt{2\mathbb{E}_p[u^2] + 2\mathbb{E}_{p_0}[u_0^2]} \\
&\leq G_3^2 \mathcal{W}_2(p, p_0) \sqrt{8\sigma_{\boldsymbol{\theta}}^2 \cdot d + 10\sigma_u^2},
\end{aligned} \tag{A.21}$$

where the first inequality is by $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\|_2 \leq G_3$ in Lemma A.1, the second inequality is by Cauchy-Schwarz inequality, the third inequality is by Jensen's inequality and the last inequality is by Lemma A.6. Next, We bound I_2 in (A.20). For any given $r > 0$ we have

$$
\begin{aligned}
I_2 &\le \mathbb{E}_{\pi^*}\big[u_0^2\,\mathbb{1}(|u_0<r|)\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\quad + \mathbb{E}_{\pi^*}\big[u_0^2\,\mathbb{1}(|u_0\ge r|)\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\le r^2\mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\quad + 2G_3^2\mathbb{E}_{\pi^*}[u_0^2\,\mathbb{1}(|u_0\ge r|)],
\end{aligned}
\tag{A.22}
$$

where the second inequality is by $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\|_2 \le G_3$ Lemma A.1. We further bound the first term on the right-hand side of (A.22),

$$
\begin{aligned}
&\mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i)\cdot\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\quad\le \mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\cdot\big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big)\big|\big] \\
&\qquad + \mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\cdot\big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i)\big)\big|\big] \\
&\quad\le 2G_3G_4\mathcal{W}_2(p, p_0),
\end{aligned}
\tag{A.23}
$$

where the last inequality is by $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \le G_3$ and $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x})\|_2 \le G_4\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2$ in Lemma A.1. Plugging (A.23) into (A.22) yields

$$
I_2 \le 2r^2G_3G_4\mathcal{W}_2(p, p_0) + 2G_3^2\mathbb{E}_{\pi^*}[u_0^2\,\mathbb{1}(|u_0\ge r|)]. \tag{A.24}
$$

Further plugging (A.21) and (A.24) into (A.20), we obtain

$$
\big|H_1(p)_{i,j} - H_1(p_0)_{i,j}\big| \le G_3^2\mathcal{W}_2(p, p_0)\sqrt{8\sigma_{\boldsymbol{\theta}}^2\cdot d + 10\sigma_u^2} + 2r^2G_3G_4\mathcal{W}_2(p, p_0) + 2G_3^2\mathbb{E}_{p_0}[u_0^2\,\mathbb{1}(|u_0\ge r|)].
$$

This finishes the proof. $\square$

A.5 Proof of Lemma 5.6

Here we provide the proof of Lemma 5.6, which is essentially based on a direct application of Lemma A.1 and the definition of 2-Wasserstein distance.

Proof of Lemma 5.6. Denote $\widehat{H}_{i,j}(\boldsymbol{\theta}, u) = h(\boldsymbol{\theta}, \mathbf{x}_i)h(\boldsymbol{\theta}, \mathbf{x}_j)$, then we have $H_2(p)_{i,j} = \mathbb{E}_p[\widehat{H}_{i,j}(\boldsymbol{\theta}, u)]$. Calculating the gradient of $\widehat{H}_{i,j}(\boldsymbol{\theta}, u)$, we have

$$
\nabla_u\widehat{H}_{i,j}(\boldsymbol{\theta}, u) = 0, \qquad \|\nabla_{\boldsymbol{\theta}}\widehat{H}_{i,j}(\boldsymbol{\theta}, u)\|_2 \le 2\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i)\|_2|h(\boldsymbol{\theta}, \mathbf{x}_j)| \le 2G_3(G_1\|\boldsymbol{\theta}\|_2 + G_2),
$$

where the second inequality is by Lemma A.1 . Applying Lemma A.7 gives

$$
|H_2(p)_{i,j} - H_2(p_0)_{i,j}| \le \Big[4G_1G_3\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2\cdot d} + 2G_2G_3\Big]\mathcal{W}_2(p, p_0).
$$

This finializes our proof. $\square$

A.6 Proof of Lemma 5.7

Lemma 5.7 gives a tail bound on p_0, which is essentially a basic property of Gaussian distribution. For completeness we present the detailed proof as follows.

Proof of Lemma 5.7. By the definition of p_0 we have

$$\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] = \frac{2}{\sqrt{2\pi\sigma_u^2}} \int_r^\infty u_0^2 \exp(-u_0^2/2\sigma_u^2) du_0 = \frac{2\sigma_u^2}{\sqrt{\pi}} \int_{r^2/2\sigma_u^2}^\infty t^{1/2} \exp(-t) dt$$

Now by the fact that $4z/\pi \leq \exp(z), \forall z \in \mathbb{R}$, we have

$$\mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] \leq \sigma_u^2 \int_{r^2/2\sigma_u^2}^\infty \exp(-t/2) dt = \frac{\sigma_u^2}{2} \exp\bigg(-\frac{r^2}{4\sigma_u^2}\bigg),$$

which finalizes our proof. □

A.7 Proof of Lemma 5.8

We first introduce some notations on the first variations. For $i \in [n]$, $\frac{\partial \mathbf{f}(t)_i}{\partial p_t}$, $\frac{\partial L(p_t)}{\partial p_t}$, $\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}$ and $\frac{\partial Q(p_t)}{\partial p_t}$ are defined as follows.

$$\frac{\partial \mathbf{f}(t)_i}{\partial p_t} := \alpha u h(\boldsymbol{\theta}, \mathbf{x}_i), \tag{A.25}$$

$$\frac{\partial L(p_t)}{\partial p_t} := \mathbb{E}_S\big[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big) \cdot \alpha u h(\boldsymbol{\theta}, \mathbf{x})\big], \tag{A.26}$$

$$\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} := \log(p_t/p_0) + 1, \tag{A.27}$$

$$\frac{\partial Q(p_t)}{\partial p_t} := \frac{\partial L(p_t)}{\partial p_t} + \lambda \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} = \mathbb{E}_S\big[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big) \cdot \alpha u h(\boldsymbol{\theta}, \mathbf{x}) + \lambda \log(p_t/p_0) + \lambda\big]. \tag{A.28}$$

The following lemma summarizes some direct calculations on the relation between these first variations defined above and the time derivatives of $\mathbf{f}(t)_i$, $L(p_t)$, $D_{\mathrm{KL}}(p_t||p_0)$ and $Q(p_t)$. Note that these results are well-known results in literature, but for completeness we present the detailed calculations in Appendix B.8.

Lemma A.8. Let $\frac{\partial \mathbf{f}(t)_i}{\partial p_t}$, $\frac{\partial L(p_t)}{\partial p_t}$, $\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}$, $\frac{\partial Q(p_t)}{\partial p_t}$ be the first variations defined in (A.25), (A.26), (A.27) and (A.28). Then

$$\begin{aligned}
\frac{\partial[\mathbf{f}(t)_i - y_i]}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial \mathbf{f}(t)_i}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du, \\
\frac{\partial L(p_t)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du, \\
\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du, \\
\frac{\partial Q(p_t)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial Q(p_t)}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du.
\end{aligned}$$

The following lemma summarizes the calculation of the gradients of the first variations defined in (A.26), (A.27) and (A.28).

Lemma A.9. Let $\frac{\partial L(p_t)}{\partial p_t}$, $\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}$ and $\frac{\partial Q(p_t)}{\partial p_t}$ be the first variations defined in (A.26), (A.27) and (A.28). Then their gradients with respect to u and $\boldsymbol{\theta}$ are given as follows:

$$
\begin{aligned}
&\nabla_u \frac{\partial L(p_t)}{\partial p_t} = -\widehat{g}_1(t, \boldsymbol{\theta}, u), \nabla_{\boldsymbol{\theta}} \frac{\partial L(p_t)}{\partial p_t} = -\widehat{g}_2(t, \boldsymbol{\theta}, u),\\
&\nabla_u \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} = u/\sigma_u^2 + \nabla_{\boldsymbol{\theta}} \log(p_t), \nabla_{\boldsymbol{\theta}} \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} = \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t),\\
&\nabla \frac{\partial Q(p_t)}{\partial p_t} = \nabla \frac{\partial L(p_t)}{\partial p_t} + \lambda \nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}.
\end{aligned}
$$

Moreover, the PDE (3.3) can be written as

$$
\frac{dp_t}{dt} = \nabla \cdot \bigg[p_t(\boldsymbol{\theta}, u) \nabla \frac{\partial Q(p_t)}{\partial p_t} \bigg].
$$

Proof of Lemma 5.8. The first variation of $\mathbf{f}(t)_i$ is $\frac{\partial \mathbf{f}(t)_i}{\partial p_t} = \alpha u h(\boldsymbol{\theta}, \mathbf{x}_i)$. So we have that

$$
\begin{aligned}
\frac{\partial [\mathbf{f}(t)_i - y_i]}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial \mathbf{f}(t)_i}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du\\
&= \alpha \int_{\mathbb{R}^{d+1}} u h(\boldsymbol{\theta}, \mathbf{x}_i) \nabla \cdot \bigg[p_t(\boldsymbol{\theta}, u) \nabla \frac{\partial Q(p_t)}{\partial p_t} \bigg] d\boldsymbol{\theta} du\\
&= -\alpha \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \nabla [u h(\boldsymbol{\theta}, \mathbf{x}_i)] \cdot \nabla \frac{\partial Q(p_t)}{\partial p_t} d\boldsymbol{\theta} du\\
&= \underbrace{-\alpha \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \nabla [u h(\boldsymbol{\theta}, \mathbf{x}_i)] \cdot \nabla \frac{\partial L(p_t)}{\partial p_t} d\boldsymbol{\theta} du}_{J_1}\\
&\qquad - \alpha \lambda \underbrace{\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \nabla [u h(\boldsymbol{\theta}, \mathbf{x}_i)] \cdot \nabla \frac{\partial D_{KL}(p_t||p_0)}{\partial p_t} d\boldsymbol{\theta} du}_{J_2}, \qquad \text{(A.29)}
\end{aligned}
$$

where the first equation is by the property of first variation, the second equation is by Lemma A.9, the third equation is by integrate by part and the last equation is by Lemma A.9. For J_1, by definition we have

$$
\begin{aligned}
J_1 &= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) h(\boldsymbol{\theta}, \mathbf{x}_i) \widehat{g}_1(t, \boldsymbol{\theta}, u) d\boldsymbol{\theta} du - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \widehat{g}_2(t, \boldsymbol{\theta}, u) d\boldsymbol{\theta} du\\
&= 2\alpha \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) h(\boldsymbol{\theta}, \mathbf{x}_i) \mathbb{E}_S[(f(p_t, \mathbf{x}) - y) h(\boldsymbol{\theta}, \mathbf{x})] d\boldsymbol{\theta} du\\
&\qquad + 2\alpha \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \mathbb{E}_S[(f(p_t, \mathbf{x}) - y) u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})] d\boldsymbol{\theta} du\\
&= \frac{2\alpha}{n} \sum_{j=1}^n H_{i,j}(p)(f(p_t, \mathbf{x}_j) - y_j),
\end{aligned}
$$

where the first equation is by Lemma A.9, the second equation is by the definition of $\widehat{g}_1$ and $\widehat{g}_2$.

Moreover, for J_2, by Lemma A.9 we obtain,

$$
\begin{aligned}
J_2 &= \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) h(\boldsymbol{\theta}, \mathbf{x}_i)[u/\sigma_u^2 + \nabla_u \log(p_t)] d\boldsymbol{\theta} du + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) u \nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{x}_i) \cdot [\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)] d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[u h(\boldsymbol{\theta}, \mathbf{x}_i)/\sigma_u^2 + u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2] d\boldsymbol{\theta} du \\
&\quad + \int_{\mathbb{R}^{d+1}} [h(\boldsymbol{\theta}, \mathbf{x}_i) \nabla_u p_t(t, \boldsymbol{\theta}, u) + u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \nabla_{\boldsymbol{\theta}} p_t(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[u h(\boldsymbol{\theta}, \mathbf{x}_i)/\sigma_u^2 + u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - u \Delta h(\boldsymbol{\theta}, \mathbf{x}_i)] d\boldsymbol{\theta} du.
\end{aligned}
$$

Plugging the calculations of J_1 and J_2 above into (A.29) completes the proof. □

A.8 Proof of Lemma 5.9

Here we give the proof of Lemma 5.9.

Proof of Lemma 5.9. The proof is based on the smoothness properties of $h(\boldsymbol{\theta}, \mathbf{x})$ given in Lemma A.1. We have

$$
\begin{aligned}
&\mathbb{E}_p\big[\big(u h(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + u \nabla h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - u \Delta h(\boldsymbol{\theta}, \mathbf{x})\big)\big] \\
&\quad \le \mathbb{E}_p\big[|u|(G_1\|\boldsymbol{\theta}\|_2 + G_2)/\sigma_u^2 + G_3|u|\|\boldsymbol{\theta}\|_2/\sigma_{\boldsymbol{\theta}}^2 + G_4|u|\big] \\
&\quad = (G_1/\sigma_u^2 + G_3/\sigma_{\boldsymbol{\theta}}^2)\mathbb{E}_p[|u|\|\boldsymbol{\theta}\|_2] + (G_2/\sigma_u^2 + G_4)\mathbb{E}_p[|u|] \\
&\quad \le (G_1/\sigma_u^2 + G_3/\sigma_{\boldsymbol{\theta}}^2)\mathbb{E}_p\bigg[\frac{u^2 + \|\boldsymbol{\theta}\|_2^2}{2}\bigg] + (G_2/\sigma_u^2 + G_4)\sqrt{\mathbb{E}_p[u^2]},
\end{aligned}
$$

where the first inequality is by $|h(\boldsymbol{\theta}, \mathbf{x})| \le G_1\|\boldsymbol{\theta}\|_2 + G_2$, $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \le G_3$ and $|\Delta h(\boldsymbol{\theta}, \mathbf{x})| \le G_4$ in Lemma A.1, the second inequality is by Young's inequality and Cauchy-Schwartz inequality. Now by $\mathcal{W}(p, p_0) \le \sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2}$ and Lemma A.6, we have

$$
\begin{aligned}
&\mathbb{E}_p\Big[\big(u h(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + u \nabla h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - u \Delta h(\boldsymbol{\theta}, \mathbf{x})\big)\Big] \\
&\quad \le (G_1/\sigma_u^2 + G_3/\sigma_{\boldsymbol{\theta}}^2) 2(\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2) + (G_2/\sigma_u^2 + G_4) 2\sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2} \\
&\quad = A_1.
\end{aligned}
$$

This completes proof. □

A.9 Proof of Lemma 5.10

Proof of Lemma 5.10. By Lemma A.8, we get

$$
\begin{aligned}
\frac{\partial Q(p_t)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial Q(p_t)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial Q(p_t)}{\partial p_t} \nabla \cdot \bigg[p_t(\boldsymbol{\theta}, u)\nabla\frac{\partial Q(p_t)}{\partial p_t}\bigg] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\bigg\|\nabla\frac{\partial Q(p_t)}{\partial p_t}\bigg\|_2^2 d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\widehat{g}_2 - \lambda\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \lambda\nabla_{\boldsymbol{\theta}}\log(p_t)\|_2^2 - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|\widehat{g}_1 - \lambda u/\sigma_u^2 - \lambda\nabla_u\log(p_t)|^2 \\
&\le 0,
\end{aligned}
$$

where the third inequality is by applying integration by parts and the fourth equality is by Lemma A.9. $\square$

A.10 Proof of Lemma 5.11

The following lemma shows the when p is close to p_0 in 2-Wasserstein distance, $f(p, \mathbf{x})$ is almost an expectation of a linear function in $(\boldsymbol{\theta}, u)$. It is the counterpart of Lemma 4.1 in Cao and Gu (2019a) for smooth activation functions in the mean-field view.

Lemma A.10. Under Assumptions 4.1 and 4.2, let $(\boldsymbol{\theta}, u)$, $(\boldsymbol{\theta}', u')$, and $(\boldsymbol{\theta}_0, u_0)$ be the parameters following distributions $p(\boldsymbol{\theta}, u)$, $p'(\boldsymbol{\theta}', u')$ and $p_0(\boldsymbol{\theta}_0, u_0)$ respectively. For any $M' > 0$, if $\mathcal{W}_2(p, p_0) \le M'$, then for any coupling π between p and p', it holds that

$$
\begin{aligned}
&\mathbb{E}_\pi\big|u'h(\boldsymbol{\theta}', \mathbf{x}) - uh(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x})(u' - u) - u\langle\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big| \\
&\quad \le \big[G_3 + 2G_4\sigma_u\sqrt{\log(\sigma_u^2/2M'^2)}\big]\mathbb{E}_\pi[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2] + 4G_3M'\sqrt{\mathbb{E}_\pi[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2]}.
\end{aligned}
$$

Proof of Lemma 5.11. We use standard notation convention and denote by $(\boldsymbol{\theta}, u)$ and $(\boldsymbol{\theta}_0, u_0)$ the random variables following distributions $p(\boldsymbol{\theta}, u)$ and $p_0(\boldsymbol{\theta}_0, u_0)$ respectively. For any p, let $\pi^* = \pi^*(p, p_0)$ be the optimal coupling of $\mathcal{W}_2(p, p_0)$. Then we have

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M'))) &= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}}\Bigg\{\sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_p\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i uh(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg]\Bigg\} \\
&= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}}\Bigg\{\sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i uh(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg]\Bigg\} \\
&= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}}\Bigg\{\sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i [uh(\boldsymbol{\theta}, \mathbf{x}_i) - u_0h(\boldsymbol{\theta}_0, \mathbf{x}_i)]\Bigg]\Bigg\},
\end{aligned}
$$

where the third equation follows by the fact that $p_0(\boldsymbol{\theta}_0, u_0)$ is a Gaussian density with mean zero and $\mathbb{E}_{p_0}[u_0h(\boldsymbol{\theta}_0, \mathbf{x})] = 0$ for all $\mathbf{x}$. Let r is a thresholding parameter whose value will be chosen later in the proof. We further expand and upper-bound the right-hand side above into several terms:

$$
\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M'))) \le I_1 + I_2 + I_3 + I_4, \tag{A.30}
$$

where

$$
\begin{aligned}
I_1 &= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}} \Bigg\{ \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i h(\boldsymbol{\theta}_0,\mathbf{x}_i)(u-u_0)\Bigg]\Bigg\},\\
I_2 &= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}} \Bigg\{ \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i u_0\, \mathbb{1}\{u_0\le r\}\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0,\mathbf{x}_i), \boldsymbol{\theta}-\boldsymbol{\theta}_0\rangle\Bigg]\Bigg\},\\
I_3 &= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}} \Bigg\{ \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i u_0\, \mathbb{1}\{u_0> r\}\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0,\mathbf{x}_i), \boldsymbol{\theta}-\boldsymbol{\theta}_0\rangle\Bigg]\Bigg\},\\
I_4 &= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}} \Bigg\{ \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i [u h(\boldsymbol{\theta},\mathbf{x}_i) - u_0 h(\boldsymbol{\theta}_0,\mathbf{x}_i) - h(\boldsymbol{\theta}_0,\mathbf{x}_i)(u-u_0)\\
&\qquad\qquad - u_0\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0,\mathbf{x}_i), \boldsymbol{\theta}-\boldsymbol{\theta}_0\rangle]\Bigg]\Bigg\}.
\end{aligned}
$$

Since $\mathcal{W}_2(p_0,p_0)=0\le M'$, applying Lemma A.10 with π being the optimal coupling in $\mathcal{W}_2(p,p_0)$ gives

$$
I_4 \le M'^2\alpha\big[5G_3 + 2G_4\sigma_u\sqrt{\log(\sigma_u/(2M'^2))}\big], \tag{A.31}
$$

For I_1, By Cauchy-Schwarz inequality we have

$$
\begin{aligned}
\mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i h(\boldsymbol{\theta}_0,\mathbf{x}_i)(u-u_0)\Bigg] &\le \sqrt{\mathbb{E}_{\pi^*}[(u-u_0)^2]}\cdot\sqrt{\mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i h(\boldsymbol{\theta}_0,\mathbf{x}_i)\Bigg]^2}\\
&\le \mathcal{W}_2(p,p_0)\cdot\sqrt{\mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i h(\boldsymbol{\theta}_0,\mathbf{x}_i)\Bigg]^2}.
\end{aligned}
$$

Therefore

$$
I_1 \le M'\alpha\cdot\mathbb{E}_{\boldsymbol{\xi}}\Bigg\{\sqrt{\mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i h(\boldsymbol{\theta}_0,\mathbf{x}_i)\Bigg]^2}\Bigg\} \le M'\alpha\cdot\sqrt{\mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^n \xi_i h(\boldsymbol{\theta}_0,\mathbf{x}_i)\Bigg]^2},
$$

where the second inequality follows by Jensen's inequality. Directly calculating the expectation with respect to $\boldsymbol{\xi}$ and applying Lemma A.1 gives

$$
I_1 \le \frac{M'\alpha}{\sqrt{n}}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\pi^*}[h^2(\boldsymbol{\theta}_0,\mathbf{x}_i)]} \le \frac{M'\alpha}{\sqrt{n}}\cdot\sqrt{\mathbb{E}_{\pi^*}[(G_1\|\boldsymbol{\theta}_0\|_2+G_2)^2]} \le \frac{\sqrt{2}M'\alpha}{\sqrt{n}}\cdot\sqrt{G_1^2\sigma_{\boldsymbol{\theta}}^2 d + G_2^2}. \tag{A.32}
$$

Similarly, for I_2 we have

$$
\begin{aligned}
&\mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^{n}\xi_i u_0\,\mathbb{1}\{u_0 \le r\}\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i), \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle\Bigg] \\
&\qquad = \mathbb{E}_{\pi^*}\Bigg[u_0\,\mathbb{1}\{u_0 \le r\} \cdot \bigg\langle \frac{1}{n}\sum_{i=1}^{n}\xi_i \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i), \boldsymbol{\theta} - \boldsymbol{\theta}_0\bigg\rangle\Bigg] \\
&\qquad \le \mathbb{E}_{\pi^*}\Bigg[|u_0|\,\mathbb{1}\{u_0 \le r\}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \cdot \bigg\|\frac{1}{n}\sum_{i=1}^{n}\xi_i \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\bigg\|_2\Bigg] \\
&\qquad \le \sqrt{\mathbb{E}_{\pi^*}[u_0^2\,\mathbb{1}\{u_0 \le r\}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2]} \cdot \sqrt{\mathbb{E}_{\pi^*}\Bigg[\bigg\|\frac{1}{n}\sum_{i=1}^{n}\xi_i \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\bigg\|_2^2\Bigg]} \\
&\qquad \le r\mathcal{W}_2(p, p_0) \cdot \sqrt{\mathbb{E}_{\pi^*}\Bigg[\bigg\|\frac{1}{n}\sum_{i=1}^{n}\xi_i \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\bigg\|_2^2\Bigg]},
\end{aligned}
$$

where we apply Cauchy-Schwarz inequality to obtain the second inequality. Therefore by Jensen's inequality we have

$$
\begin{aligned}
I_2 &\le rM'\alpha \cdot \mathbb{E}_{\boldsymbol{\xi}}\Bigg\{\sqrt{\mathbb{E}_{\pi^*}\Bigg[\bigg\|\frac{1}{n}\sum_{i=1}^{n}\xi_i \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\bigg\|_2^2\Bigg]}\Bigg\} \\
&\le rM'\alpha \cdot \sqrt{\mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{\pi^*}\Bigg[\bigg\|\frac{1}{n}\sum_{i=1}^{n}\xi_i \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\bigg\|_2^2\Bigg]} \\
&= rM'\alpha \cdot \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}_{\pi^*}[\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\|_2^2]} \\
&\le \frac{G_3 rM'\alpha}{\sqrt{n}}.
\end{aligned}
\tag{A.33}
$$

For I_3, we have

$$
\begin{aligned}
I_3 &= \alpha \cdot \mathbb{E}_{\boldsymbol{\xi}}\Bigg\{\sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^{n}\xi_i u_0\,\mathbb{1}\{u_0 > r\}\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i), \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle\Bigg]\Bigg\} \\
&\le \alpha \cdot \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}\Bigg[\frac{1}{n}\sum_{i=1}^{n}|u_0|\,\mathbb{1}\{|u_0| > r\}\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i)\|_2\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2\Bigg] \\
&\le G_3\alpha \cdot \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \mathbb{E}_{\pi^*}[|u_0|\,\mathbb{1}\{|u_0| > r\}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2] \\
&\le G_3\alpha \cdot \sup_{p:\mathcal{W}_2(p,p_0)\le M'} \sqrt{\mathbb{E}_{\pi^*}[u_0^2\,\mathbb{1}\{|u_0| > r\}]} \cdot \sqrt{\mathbb{E}_{\pi^*}[\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2]} \\
&\le G_3 M'\alpha \cdot \sqrt{\mathbb{E}_{\pi^*}[u_0^2\,\mathbb{1}\{|u_0| > r\}]} \\
&\le \frac{G_3 M'\alpha\sigma_u}{\sqrt{2}} \cdot \exp\bigg(-\frac{r^2}{8\sigma_u^2}\bigg),
\end{aligned}
\tag{A.34}
$$

where the third inequality follows by Cauchy-Schwarz inequality, and the last inequality follows by Lemma 5.7. Summing up the bounds for I_1, I_2, I_3, I_4 in (A.31), (A.32), (A.33), (A.34) into (A.30), we have

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M'))) \leq \frac{\sqrt{2}M'\alpha}{\sqrt{n}} \cdot \sqrt{G_1^2\sigma_{\boldsymbol{\theta}}^2 d + G_2^2} + \frac{G_3 r M'\alpha}{\sqrt{n}} + \frac{G_3 M'\alpha\sigma_u}{\sqrt{2}} \cdot \exp\bigg(-\frac{r^2}{8\sigma_u^2}\bigg) \\
+ M'^2\alpha\big[5G_3 + 2G_4\sigma_u\sqrt{\log(\sigma_u/(2M'^2))}\big].
\end{aligned}
$$

Setting $r = 2\sigma_u\sqrt{\log(n)}$, we have

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{F}_{\mathcal{W}_2}(M'))) \leq \frac{M'\alpha}{\sqrt{n}} \cdot \Big[\sqrt{2G_1^2\sigma_{\boldsymbol{\theta}}^2 d + G_2^2} + 2G_3\sigma_u\sqrt{\log(n)} + G_3\sigma_u/\sqrt{2}\Big] \\
+ M'^2\alpha\big[5G_3 + 2G_4\sigma_u\sqrt{\log(\sigma_u/(2M'^2))}\big].
\end{aligned}
$$

This completes the proof. □

A.11 Proof of Lemma 5.12

Proof of Lemma 5.12. Our proof is inspired by the Rademacher complexity bound for discrete distributions given by Meir and Zhang (2003). Let γ be a parameter whose value will be determined later in the proof. We have

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M)) &= \frac{\alpha}{\gamma} \cdot \mathbb{E}_{\boldsymbol{\xi}}\Bigg[\sup_{p: D_{\mathrm{KL}}(p||p_0) \leq M} \int_{\mathbb{R}^{d+1}} \frac{\gamma}{n}\sum_{i=1}^n \xi_i u h(\boldsymbol{\theta}, \mathbf{x}_i) p(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du\Bigg] \\
&\leq \frac{\alpha}{\gamma} \cdot \Bigg\{M + \mathbb{E}_{\boldsymbol{\xi}} \log\Bigg[\int \exp\Bigg(\frac{\gamma}{n}\sum_{i=1}^n \xi_i u h(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg) p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du\Bigg]\Bigg\} \\
&\leq \frac{\alpha}{\gamma} \cdot \Bigg\{M + \log\Bigg[\int \mathbb{E}_{\boldsymbol{\xi}} \exp\Bigg(\frac{\gamma}{n}\sum_{i=1}^n \xi_i u h(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg) p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du\Bigg]\Bigg\},
\end{aligned}
$$

where the first inequality follows by the Donsker-Varadhan representation of KL-divergence (Donsker and Varadhan, 1983), and the second inequality follows by Jensen's inequality. Note that $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables. By standard tail bound we have

$$
\mathbb{E}_{\boldsymbol{\xi}} \exp\Bigg[\frac{\gamma}{n}\sum_{i=1}^n \xi_i u h(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg] \leq \exp\Bigg[\frac{\gamma^2}{2n^2}\sum_{i=1}^n u^2 h^2(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg].
$$

Therefore

$$
\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M))) \leq \frac{\alpha}{\gamma} \cdot \Bigg\{M + \log\Bigg[\int \exp\Bigg(\frac{\gamma^2}{2n^2}\sum_{i=1}^n u^2 h^2(\boldsymbol{\theta}, \mathbf{x}_i)\Bigg) p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du\Bigg]\Bigg\}.
$$

Now by the assumption that $h(\boldsymbol{\theta}, \mathbf{x}) \le G_7$, we have

$$\begin{aligned}\int \exp\left(\frac{\gamma^2}{2n^2}\sum_{i=1}^{n} u^2 h^2(\boldsymbol{\theta}, \mathbf{x}_i)\right) p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du &\le \int \exp\left(\frac{\gamma^2 G_7^2}{2n} u^2\right) p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du \\ &= \frac{1}{\sqrt{2\pi}\sigma_u} \cdot \sqrt{\frac{2\pi}{\sigma_u^{-2} - \gamma^2 G_7^2 n^{-1}}} \\ &= \sqrt{\frac{1}{1 - \sigma_u^2 \gamma^2 G_7^2 n^{-1}}}.\end{aligned}$$

Therefore we have

$$\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M)) \le \frac{\alpha}{\gamma} \cdot \left[M + \log\left(\sqrt{\frac{1}{1 - \sigma_u^2 \gamma^2 G_7^2 n^{-1}}}\right)\right].$$

Setting $\gamma = \sigma_u^{-1} B^{-1} \sqrt{Mn}$ and applying the inequality $\log(1 - z) \ge -2z$ for $z \in [0, 1/2]$ gives

$$\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(M)) \le \frac{G_7 \sigma_u \alpha}{\sqrt{Mn}} \cdot \left[M + \log\left(\sqrt{\frac{1}{1 - M}}\right)\right] \le 2\alpha G_7 \sigma_u \sqrt{\frac{M}{n}}.$$

This completes the proof. □

B Proof of Lemmas in Appendix A

In this section we provide the proof of technical lemmas we use in Appendix A.

B.1 Proof of Lemma A.1

Here we provide the proof of Lemma A.1, which is essentially based on direct calculations on the activation function and the assumption that $\|\mathbf{x}\|_2 \le 1$.

Proof of Lemma A.1. By $h(\boldsymbol{\theta}, \mathbf{x}) = \widetilde{h}(\boldsymbol{\theta}^\top \mathbf{x})$, we have the following identities.

$$\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) = \widetilde{h}'(\boldsymbol{\theta}^\top \mathbf{x})\mathbf{x}, \quad \Delta h(\boldsymbol{\theta}^\top \mathbf{x}) = \sum_{i=1} \widetilde{h}''(\boldsymbol{\theta}^\top \mathbf{x}) x_i^2 = \widetilde{h}''(\boldsymbol{\theta}^\top \mathbf{x})\|\mathbf{x}\|_2^2, \quad \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} = \widetilde{h}'(\boldsymbol{\theta}^\top \mathbf{x})\boldsymbol{\theta}^\top \mathbf{x}.$$

By $|\widetilde{h}(z)| \le G_1|z| + G_2$ in Assumption 4.2 and $\|\mathbf{x}\|_2 \le 1$ in Assumption 4.1, we have

$$|h(\boldsymbol{\theta}, \mathbf{x})| \le G_1|\boldsymbol{\theta}^\top \mathbf{x}| + G_2 \le G_1\|\boldsymbol{\theta}\|_2 + G_2,$$

which gives the first bound. The other results can be derived similarly, which we present as follows. By $|\widetilde{h}'(z)| \le G_3$ and $\|\mathbf{x}\|_2 \le 1$, we have

$$\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 = \|\widetilde{h}'(\boldsymbol{\theta}^\top \mathbf{x})\mathbf{x}\|_2 \le G_3,$$

which gives the second bound. By $|\widetilde{h}''(z)| \le G_4$ and $\|\mathbf{x}\|_2 \le 1$, we have

$$|\Delta h(\boldsymbol{\theta}, \mathbf{x})| = |\widetilde{h}''(\boldsymbol{\theta}^\top \mathbf{x})\|\mathbf{x}\|_2^2| \le G_4.$$

Moreover, based on the same assumptions we also have

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_1, \mathbf{x}) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_2, \mathbf{x})\|_2 &= \|\widetilde{h}'(\boldsymbol{\theta}_1^\top \mathbf{x})\mathbf{x} - \widetilde{h}'(\boldsymbol{\theta}_2^\top \mathbf{x})\mathbf{x}\|_2 \\
&\le |\widetilde{h}'(\boldsymbol{\theta}_1^\top \mathbf{x}) - \widetilde{h}'(\boldsymbol{\theta}_2^\top \mathbf{x})| \\
&\le G_4 |\boldsymbol{\theta}_1^\top \mathbf{x} - \boldsymbol{\theta}_2^\top \mathbf{x}| \\
&\le G_4 \|\boldsymbol{\theta}_1^\top - \boldsymbol{\theta}_2^\top\|_2.
\end{aligned}$$

Therefore the third and fourth bounds hold. Applying the bound $|\big(z\widetilde{h}'(z)\big)'| \le G_5$ and $\|\mathbf{x}\|_2 \le 1$ gives the fifth bound:

$$\|\nabla_{\boldsymbol{\theta}}\big(\nabla_{\boldsymbol{\theta}} h(\theta, \mathbf{x}) \cdot \boldsymbol{\theta}\big)\|_2 = \|\nabla_{\boldsymbol{\theta}}\big(\widetilde{h}'(\boldsymbol{\theta}^\top \mathbf{x})\boldsymbol{\theta}^\top \mathbf{x}\big)\|_2 = \|\mathbf{x}\|_2 \big|\big(z\widetilde{h}'(z)\big)'|_{z=\boldsymbol{\theta}^\top \mathbf{x}}\big| \le G_5.$$

Finlaly, by $|\widetilde{h}'''(z)| \le G_6$ and $\|\mathbf{x}\|_2 \le 1$, we have

$$\|\nabla_{\boldsymbol{\theta}} \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 = \|\nabla_{\boldsymbol{\theta}} \widetilde{h}''(\boldsymbol{\theta}^\top \mathbf{x})\|_2 \|\mathbf{x}\|_2^2 \le |\widetilde{h}'''(\boldsymbol{\theta}^\top \mathbf{x})| \|\mathbf{x}\|_2^3 \le G_6.$$

This completes the proof. $\square$

B.2 Proof of Lemma A.2

Proof of Lemma A.2. By Lemma A.8, we have the following chain rule

$$\begin{aligned}
\frac{\partial L(p_t)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t} \nabla \cdot \bigg[p_t(\boldsymbol{\theta}, u) \nabla \frac{\partial Q(p_t)}{\partial p_t}\bigg] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial L(p_t)}{\partial p_t}\bigg] \cdot \bigg[\nabla \frac{\partial Q(p_t)}{\partial p_t}\bigg] d\boldsymbol{\theta} du \\
&= \underbrace{-\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial L(p_t)}{\partial p_t}\bigg] \cdot \bigg[\nabla \frac{\partial L(p_t)}{\partial p_t}\bigg] d\boldsymbol{\theta} du}_{I_1} \\
&\quad \underbrace{- \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial L(p_t)}{\partial p_t}\bigg] \cdot \bigg[\nabla \frac{\partial D(p_t||p_0)}{\partial p_t}\bigg] d\boldsymbol{\theta} du}_{I_2}, \qquad \text{(B.1)}
\end{aligned}$$

where the second and last equality is by Lemma A.9, the third inequality is by apply integration by parts. We now proceed to calculate I_1 and I_2 based on the calculations of derivatives in Lemma A.9. For I_1, we have

$$I_1 = \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta} du + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) |\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta} d\boldsymbol{\theta} du. \qquad \text{(B.2)}$$

Similarly, for I_2, we have

$$
\begin{aligned}
I_2 &= \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[-\widehat{g}_1(t, \boldsymbol{\theta}, u)] \cdot [u/\sigma_u^2 + \nabla_u \log(p_t)] d\boldsymbol{\theta} du \\
&\quad + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[-\widehat{g}_2(t, \boldsymbol{\theta}, u)] \cdot [\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u/\sigma_u^2 + \widehat{g}_2(t, \boldsymbol{\theta}, u)\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2] d\boldsymbol{\theta} du \\
&\quad - \int_{\mathbb{R}^{d+1}} [\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot \nabla_u p_t(t, \boldsymbol{\theta}, u) + \widehat{g}_2(t, \boldsymbol{\theta}, u) \cdot \nabla_{\boldsymbol{\theta}} p_t(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u/\sigma_u^2 + \widehat{g}_2(t, \boldsymbol{\theta}, u)\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2] d\boldsymbol{\theta} du \\
&\quad + \int_{\mathbb{R}^{d+1}} p_t(t, \boldsymbol{\theta}, u)[\nabla_u \cdot \widehat{g}_1(t, \boldsymbol{\theta}, u) + \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du, \qquad \text{(B.3)}
\end{aligned}
$$

where the second equality is by $p_t \nabla \log(p_t) = \nabla p_t$ and the third equality is by applying integration by parts. Plugging (B.2) and (B.3) into (B.1), we get

$$
\begin{aligned}
\frac{\partial L(p_t)}{\partial t} &= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta} du - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta} d\boldsymbol{\theta} du \\
&\quad + \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du.
\end{aligned}
$$

This completes the proof. □

B.3 Proof of Lemma A.3

Here we prove Lemma A.3, which is based on its connection to the Gram matrix of neural tangent kernel.

Proof of Lemma A.3. We first remind the readers of the definitions of the Gram matrices in (3.5), (3.6) and (3.7). Let $\mathbf{b}(p_t) = (f(p_t, \mathbf{x}_1) - y_1, \ldots, f(p_t, \mathbf{x}_n) - y_n)^\top \in \mathbb{R}^n$. Then by the definitions of $\mathbf{H}_1(p_t)$ and $\mathbf{H}_2(p_t)$ in (3.6) and (3.7), we have

$$
\begin{aligned}
\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\big[|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)h(\boldsymbol{\theta}, \mathbf{x})]|^2 d\boldsymbol{\theta} du &= \frac{1}{n^2}\mathbf{b}(p_t)^\top \mathbf{H}_1(p_t)\mathbf{b}(p_t), \\
\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\big[\|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})]\|_2^2\big] d\boldsymbol{\theta} du &= \frac{1}{n^2}\mathbf{b}(p_t)^\top \mathbf{H}_2(p_t)\mathbf{b}(p_t).
\end{aligned}
$$

Therefore by (3.5) we have

$$
\begin{aligned}
&\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\big[|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)h(\boldsymbol{\theta}, \mathbf{x})]|^2 + \|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})]\|_2^2\big] d\boldsymbol{\theta} du \\
&\quad = \frac{1}{n^2}\mathbf{b}(p_t)^\top \mathbf{H}(p_t)\mathbf{b}(p_t). \qquad \text{(B.4)}
\end{aligned}
$$

By the definition of t^*, for $t \le t^*$ we have $\mathcal{W}_2(p_t, p_0) \le R$, and therefore applying Lemma 5.1 gives

$$\frac{1}{n^2}\mathbf{b}(p_t)^\top \mathbf{H}(p_t)\mathbf{b}(p_t) \ge \frac{\Lambda\|\mathbf{b}(p_t)\|_2^2}{2n^2} = \frac{\lambda_0^2}{2}L(p_t), \tag{B.5}$$

where the equality follows by the definition of $\mathbf{b}(p_t)$. Plugging (B.5) into (B.4) completes the proof. $\square$

B.4 Proof of Lemma A.4

The proof of Lemma A.4 is based on direct applications of Lemma A.1. We present the proof as follows.

Proof of Lemma A.4. We have the following identities

$$\begin{aligned}
\widehat{g}_1(t,\boldsymbol{\theta},u) &= -\mathbb{E}_S[\nabla_f \phi(f(p_t,\mathbf{x}),y)\alpha h(\boldsymbol{\theta},\mathbf{x})],\\
\widehat{g}_2(t,\boldsymbol{\theta},u) &= -\mathbb{E}_S[\nabla_f \phi(f(p_t,\mathbf{x}),y)\alpha u \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta},\mathbf{x})],\\
\nabla_u \cdot \widehat{g}_1(t,\boldsymbol{\theta},u) &= 0,\\
\nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2(t,\boldsymbol{\theta},u) &= -\mathbb{E}_S[\nabla_f \phi(f(p_t,\mathbf{x}),y)\alpha u \Delta h(\boldsymbol{\theta},\mathbf{x})].
\end{aligned}$$

Base on these inequalities we can derive

$$\begin{aligned}
&\bigg| \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta},u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du \bigg|\\
&\quad = \bigg| \alpha \mathbb{E}_S\Big[\nabla_f \phi(f(p_t,\mathbf{x}),y)\mathbb{E}_{p_t}\Big[\big(u_t h(\boldsymbol{\theta}_t,\mathbf{x}_i)/\sigma_u^2 + u_t \nabla h(\boldsymbol{\theta}_t,\mathbf{x}_i)\cdot \boldsymbol{\theta}_t/\sigma_{\boldsymbol{\theta}}^2 - u_t \Delta h(\boldsymbol{\theta}_t,\mathbf{x}_i)\big)\Big]\Big]\bigg|\\
&\quad \le 2\alpha A_1 \mathbb{E}_S[|f(p_t,\mathbf{x}) - y|]\\
&\quad \le 2\alpha A_1 \sqrt{L(p_t)},
\end{aligned}$$

where the first inequality is by Lemma 5.9 and the last inequality is by Jensen's inequality. $\square$

B.5 Proof of Lemma A.5

The following lemma summarizes the calculation on the time derivative of $D_{\mathrm{KL}}(p_t||p_0)$.

Lemma B.1. Let p_t be the solution of PDE (3.3). Then the following identity holds.

$$\begin{aligned}
\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} &= -\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta},u)\|\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)\|_2^2 - \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta},u)|u/\sigma_u^2 + \nabla_u \log(p_t)|^2\\
&\quad + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta},u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du.
\end{aligned}$$

In the calculation given by Lemma B.1, we can see that the (potentially) positive term in $\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t}$ naturally coincides with the corresponding term in $\frac{\partial L(p_t)}{\partial t}$ given by Lemma A.2, and a bound of it has already been given in Lemma A.4. However, for the analysis of the KL-divergence term, we present the following new bound, which eventually leads to a sharper result.

Lemma B.2. Under Assumptions 4.1 and 4.2, for $\mathcal{W}_2(p, p_0) \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 \cdot d + \sigma_u^2}$, it holds that

$$\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du \leq 2\alpha A_2 \sqrt{L(p_t)} \sqrt{D_{\mathrm{KL}}(p_t||p_0)},$$

where A_2 is defined in theorem 4.4.

Proof of Lemma A.5. By Lemma B.1,

$$\begin{aligned}
\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} &= -\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \|\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)\|_2^2 - \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) |u/\sigma_u^2 + \nabla_u \log(p_t)|^2 \\
&\quad + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du \\
&\leq 2A_2\alpha \sqrt{L(p_t)} \sqrt{D_{\mathrm{KL}}(p_t||p_0)}, \qquad \text{(B.6)}
\end{aligned}$$

where the inequality is by Lemma B.2. Notice that $\sqrt{D_{\mathrm{KL}}(p_0||p_0)} = 0$, $\sqrt{D_{\mathrm{KL}}(p_t||p_0)}$ is differentiable at $\sqrt{D_{\mathrm{KL}}(p_t||p_0)} \neq 0$ and from (B.6) the derivative

$$\frac{\partial \sqrt{D_{\mathrm{KL}}(p_t||p_0)}}{\partial t} = \frac{\partial D_{\mathrm{KL}}(p_t||p_0))}{\partial t} \frac{1}{2\sqrt{D_{\mathrm{KL}}(p_t||p_0)}} \leq A_2 \alpha \sqrt{L(p_t)},$$

which implies

$$\begin{aligned}
\sqrt{D_{\mathrm{KL}}(p_t||p_0)} &\leq \int_0^t A_2 \alpha \sqrt{L(p_s)} ds \\
&\leq A_2 \alpha \int_0^t \exp(-\alpha^2 \lambda_0^2 s) \sqrt{L(p_0)} + A_1 \lambda \alpha^{-1} \lambda_0^{-2} ds \\
&\leq A_2 \alpha^{-1} \lambda_0^{-2} \sqrt{L(p_0)} + A_2 A_1 \lambda \lambda_0^{-2} t,
\end{aligned}$$

where the second inequality holds due to Lemma 5.2. Squaring both sides and applying Jensen's inequality now gives

$$D_{\mathrm{KL}}(p_t||p_0) \leq 2A_2^2 \alpha^{-2} \lambda_0^{-4} L(p_0) + 2A_2^2 A_1^2 \lambda^2 \lambda_0^{-4} t^2.$$

This completes the proof. □

B.6 Proof of Lemma A.6

Proof of Lemma A.6. Let $\pi^*(p_0, p)$ be the coupling that achieves the 2-Wasserstein distance between p_0 and p. Then by definition,

$$\begin{aligned}
\mathbb{E}_{\pi^*}(\|\boldsymbol{\theta}\|_2^2 + u^2) &\leq \mathbb{E}_{\pi^*}(2\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 + 2\|\boldsymbol{\theta}_0\|_2^2 + 2(u - u_0)^2 + 2u_0^2) \\
&\leq 2R^2 + 2\sigma_{\boldsymbol{\theta}}^2 \cdot d + 2\sigma_u^2 \\
&\leq 4\sigma_{\boldsymbol{\theta}}^2 \cdot d + 4\sigma_u^2,
\end{aligned}$$

where the last inequality is by the assumption that $\mathcal{W}_2(p, p_0) \leq \sqrt{\sigma_{\boldsymbol{\theta}}^2 d + \sigma_u^2}$. This finishes the proof. □

B.7 Proof of Lemma A.7

Proof of Lemma A.7. By Lemma C.8 in Xu et al. (2018), we have that

$$\Big|\mathbb{E}_p[g(u,\boldsymbol{\theta})] - \mathbb{E}_{p_0}[g(u_0,\boldsymbol{\theta}_0)]\Big| \le (C_1\sigma + C_2)\mathcal{W}_2(p,p_0),$$

where $\sigma^2 = \max\{\mathbb{E}_p[u^2 + \boldsymbol{\theta}^2], \mathbb{E}_{p_0}[u_0^2 + \boldsymbol{\theta}_0^2]\}$. Then by Lemma A.6, we get $\sigma \le 2\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d}$. Substituting the upper bound of σ into the above inequality completes the proof. $\square$

B.8 Proof of Lemma A.8

Proof of Lemma A.8. By chain rule and the definition of $\mathbf{f}(t)$, we have

$$\begin{aligned}
\frac{\partial[\mathbf{f}(t)_i - y_i]}{\partial t} &= \frac{d}{dt}\int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta},\mathbf{x}) p_t(\boldsymbol{\theta},u) d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta},\mathbf{x}) \frac{dp_t}{dt}(\boldsymbol{\theta},u) d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial \mathbf{f}(t)_i}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du,
\end{aligned}$$

where the last equality follows by the definition of the first variation $\frac{\partial L(p_t)}{\partial p_t}$. This proves the first identity. Now we bound the second identity,

$$\begin{aligned}
\frac{\partial L(p_t)}{\partial t} &= \mathbb{E}_S\bigg[\nabla_{y'}\phi\big(f(p_t,\mathbf{x}),y\big)\frac{d}{dt}f(p_t,\mathbf{x})\bigg] \\
&= \mathbb{E}_S\bigg[\nabla_{y'}\phi\big(f(p_t,\mathbf{x}),y\big)\frac{d}{dt}\int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta},\mathbf{x}) p_t(\boldsymbol{\theta},u) d\boldsymbol{\theta} du\bigg] \\
&= \mathbb{E}_S\bigg[\nabla_{y'}\phi\big(f(p_t,\mathbf{x}),y\big)\int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta},\mathbf{x}) \frac{dp_t(\boldsymbol{\theta},u)}{dt} d\boldsymbol{\theta} du\bigg] \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du,
\end{aligned}$$

where the last equality follows by the definition of the first variation $\frac{\partial L(p_t)}{\partial p_t}$. This proves the second identity. Similarly, for $\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t}$, we have

$$\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} = \frac{d}{dt}\int p_t \log(p_t/p_0) d\boldsymbol{\theta} du = \int \frac{dp_t}{dt}\log(p_t/p_0) + \frac{dp_t}{dt} d\boldsymbol{\theta} du = \int_{\mathbb{R}^{d+1}} \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du.$$

Notice that $Q(p_t) = L(p_t) + \lambda D_{\mathrm{KL}}(p_t||p_0)$, so we have

$$\begin{aligned}
\frac{\partial Q(p_t)}{\partial t} &= \frac{\partial L(p_t)}{\partial t} + \lambda\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du + \lambda\int_{\mathbb{R}^{d+1}} \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial Q(p_t)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du,
\end{aligned}$$

where the last equality is by the definition $\frac{\partial Q(p_t)}{\partial p_t} = \frac{\partial L(p_t)}{\partial p_t} + \lambda \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial pt}$. This completes the proof. □

B.9 Proof of Lemma A.9

Proof of Lemma A.9. By Lemma A.8,we have

$$
\begin{aligned}
\nabla_u \frac{\partial L}{\partial p_t} &= \nabla_u \mathbb{E}_S\big[\nabla_{y'}\phi\big(f(p_t,\mathbf{x}),y\big)\alpha u h(\boldsymbol{\theta},\mathbf{x})\big] = -\widehat{g}_1(t,\boldsymbol{\theta},u),\\
\nabla_{\boldsymbol{\theta}} \frac{\partial L}{\partial p_t} &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_S\big[\nabla_{y'}\phi\big(f(p_t,\mathbf{x}),y\big)\alpha u h(\boldsymbol{\theta},\mathbf{x})\big] = -\widehat{g}_2(t,\boldsymbol{\theta},u),\\
\nabla_u \frac{\partial D_{kL}(p_t||p_0)}{\partial p_t} &= \nabla_u(\log(p_t/p_0)+1) = u/\sigma_u^2 + \nabla_u \log(p_t),\\
\nabla_{\boldsymbol{\theta}} \frac{\partial D_{kL}(p_t||p_0)}{\partial p_t} &= \nabla_{\boldsymbol{\theta}}(\log(p_t/p_0)+1) = \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t).
\end{aligned}
$$

This proves the first four identities. For the last one, by the definition

$$
\nabla \frac{\partial Q(p_t)}{\partial p_t} = \nabla \frac{\partial L(p_t)}{\partial p_t} + \lambda \nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t},
$$

we have

$$
\begin{aligned}
\nabla \cdot \bigg[p_t(\boldsymbol{\theta},u)\nabla \frac{\partial Q(p_t)}{\partial p_t}\bigg] &= \nabla \cdot \bigg[p_t(\boldsymbol{\theta},u)\nabla \frac{\partial L}{\partial p_t}\bigg] + \lambda \nabla \cdot \bigg[p_t(\boldsymbol{\theta},u)\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t}\bigg]\\
&= -\nabla_u \cdot [p_t(\boldsymbol{\theta},u)\widehat{g}_1] - \nabla_{\boldsymbol{\theta}} \cdot [p_t(\boldsymbol{\theta},u)\widehat{g}_2] + \lambda \nabla_u \cdot [p_t(\boldsymbol{\theta},u)u/\sigma_u^2]\\
&\quad + \lambda \nabla_{\boldsymbol{\theta}} \cdot [p_t(\boldsymbol{\theta},u)\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2] + \lambda \nabla \cdot [p_t \nabla \log(p_t)]\\
&= -\nabla_u \cdot [p_t(\boldsymbol{\theta},u)g_1(t,\boldsymbol{\theta},u)] - \nabla_{\boldsymbol{\theta}} \cdot [p_t(\boldsymbol{\theta},u)g_2(t,\boldsymbol{\theta},u)] + \lambda \Delta[p_t(\boldsymbol{\theta},u)]\\
&= \frac{dp_t}{dt},
\end{aligned}
$$

where the third equality is by the definition $g_1(t,\boldsymbol{\theta},u) = \widehat{g}_1(t,\boldsymbol{\theta},u) - \lambda u/\sigma_u^2$, $g_2(t,\boldsymbol{\theta},u) = \widehat{g}_2(t,\boldsymbol{\theta},u) - \lambda\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2$ and $p_t \nabla \log(p_t) = \nabla p_t$. □

B.10 Proof of Lemma A.10

Proof of Lemma A.10. Consider $(\boldsymbol{\theta},u)$, $(\boldsymbol{\theta}',u')$ and $(\boldsymbol{\theta}_0,u_0)$ following distributions $p(\boldsymbol{\theta},u)$, $p'(\boldsymbol{\theta}',u')$ and $p(\boldsymbol{\theta}_0,u_0)$ respectively. First, based on Lemma A.1 on the smoothness of h, we can derive the following two bounds on the first-order approximation of $h(\boldsymbol{\theta}',\mathbf{x})$:

$$
\begin{aligned}
|h(\boldsymbol{\theta}',\mathbf{x}) - h(\boldsymbol{\theta},\mathbf{x}) - \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta},\mathbf{x}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle| &\le |h(\boldsymbol{\theta}',\mathbf{x}) - h(\boldsymbol{\theta},\mathbf{x})| + |\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta},\mathbf{x}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle|\\
&\le 2G_3\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|_2, \qquad (\mathrm{B.7})\\
|h(\boldsymbol{\theta}',\mathbf{x}) - h(\boldsymbol{\theta},\mathbf{x}) - \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta},\mathbf{x}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle| &= |\langle \nabla_{\boldsymbol{\theta}} h(\widetilde{\boldsymbol{\theta}},\mathbf{x}) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta},\mathbf{x}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle|\\
&\le G_4\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|_2^2, \qquad (\mathrm{B.8})
\end{aligned}
$$

where $\widetilde{\boldsymbol{\theta}}$ in (B.8) is a point on the line segment connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, which also lead to the inequality in (B.8). Therefore we have

$$
\begin{aligned}
&\big|u'h(\boldsymbol{\theta}', \mathbf{x}) - uh(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x})(u' - u) - u\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big| \\
&\qquad = \big|(u' - u)\big(h(\boldsymbol{\theta}', \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x})\big) + u\big(h(\boldsymbol{\theta}', \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x}) - \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big)\big| \\
&\qquad \le G_3|u' - u|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 + \big|u\big(h(\boldsymbol{\theta}', \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x}) - \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big)\big| \\
&\qquad \le G_3|u' - u|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 + 2G_3|u - u_0|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 + \big|u_0\big(h(\boldsymbol{\theta}', \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x}) - \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big)\big| \\
&\qquad \le G_3|u' - u|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 + 2G_3|u - u_0|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 + rG_4\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + 2G_3|u_0\,\mathbb{1}(u_0 \ge r)|\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2,
\end{aligned}
$$

where we apply the basic properties of $h(\boldsymbol{\theta}, \mathbf{x})$ in Lemma A.1 as well as (B.7) and (B.7) to derive the three inequalities, and in the last inequality we also introduce a thresholding parameter r whose value will be determined later. Taking expectation then gives

$$
\begin{aligned}
&\mathbb{E}_{\pi(p', p)}\big|u'h(\boldsymbol{\theta}', \mathbf{x}) - uh(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x})(u' - u) - u\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big| \\
&\qquad = \mathbb{E}_{\pi^*(p', p_0), \pi(p', p)}\big|u'h(\boldsymbol{\theta}', \mathbf{x}) - uh(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x})(u' - u) - u\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big| \\
&\qquad \le (G_3 + rG_4)\mathbb{E}_{\pi}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2] + 2G_3M'\sqrt{\mathbb{E}_{\pi}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2]} \\
&\qquad\qquad + 2G_3\sqrt{\mathbb{E}_{p'}[u_0^2\,\mathbb{1}(|u_0| \ge r)]}\sqrt{\mathbb{E}_{\pi}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2]},
\end{aligned}
$$

where the inequality follows by Cauchy-Schwarz inequality and the assumption that $\mathcal{W}_2(p, p_0) \le M'$. Now setting $r = 2\sigma_u\sqrt{\log(\sigma_u^2/2M'^2)}$ and applying Lemma 5.7 gives

$$
\mathbb{E}_{p_0}[u_0^2\,\mathbb{1}(|u_0 \ge r|)] \le \frac{\sigma_u^2}{2}\exp\bigg(-\frac{r^2}{4\sigma_u^2}\bigg) \le M'^2,
$$

and therefore we obtain

$$
\begin{aligned}
&\mathbb{E}_{\pi(p', p)}\big|u'h(\boldsymbol{\theta}', \mathbf{x}) - uh(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta}, \mathbf{x})(u' - u) - u\langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big| \\
&\qquad \le \big[G_3 + 2G_4\sigma_u\sqrt{\log(\sigma_u^2/2M'^2)}\big]\mathbb{E}_{\pi}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2] + 4G_3M'\sqrt{\mathbb{E}_{\pi}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + (u - u')^2]},
\end{aligned}
$$

This completes the proof. □

C Proof of Lemmas in Appendix B

C.1 Proof of Lemma B.1

Proof of Lemma B.1. By Lemma A.8, we have

$$
\begin{aligned}
\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \nabla \cdot \bigg[p_t(\boldsymbol{\theta}, u) \nabla \frac{\partial Q(p_t)}{\partial p_t} \bigg] d\boldsymbol{\theta} du \\
&= - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] \cdot \bigg[\nabla \frac{\partial Q(p_t)}{\partial p_t} \bigg] d\boldsymbol{\theta} du \\
&= -\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] \cdot \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] d\boldsymbol{\theta} du \\
&\quad - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] \cdot \bigg[\nabla \frac{\partial L(p_t)}{\partial p_t} \bigg] d\boldsymbol{\theta} du, \qquad \text{(C.1)}
\end{aligned}
$$

where the second and last equality is by Lemma A.9, the third inequality is by apply integration by parts multiple times. We further calculate by Lemma A.9,

$$
\begin{aligned}
&\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] \cdot \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] d\boldsymbol{\theta} du \\
&\quad = \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \|\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)\|^2 + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) |u/\sigma_u^2 + \nabla_u \log(p_t)|_2^2. \qquad \text{(C.2)}
\end{aligned}
$$

Moreover, for the second term on the right-hand side of (C.1) we have

$$
\begin{aligned}
&\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \bigg[\nabla \frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial p_t} \bigg] \cdot \bigg[\nabla \frac{\partial L(p_t)}{\partial p_t} \bigg] d\boldsymbol{\theta} du \\
&\quad = \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) [-\widehat{g}_1(t, \boldsymbol{\theta}, u)] \cdot [u/\sigma_u^2 + \nabla_u \log(p_t)] d\boldsymbol{\theta} du \\
&\qquad + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) [-\widehat{g}_2(t, \boldsymbol{\theta}, u)] \cdot [\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)] d\boldsymbol{\theta} du \\
&\quad = - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) [\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u/\sigma_u^2 + \widehat{g}_2(t, \boldsymbol{\theta}, u) \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2] d\boldsymbol{\theta} du \\
&\qquad - \int_{\mathbb{R}^{d+1}} [\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot \nabla_u p_t(t, \boldsymbol{\theta}, u) + \widehat{g}_2(t, \boldsymbol{\theta}, u) \cdot \nabla_{\boldsymbol{\theta}} p_t(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du \\
&\quad = - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) [\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u/\sigma_u^2 + \widehat{g}_2(t, \boldsymbol{\theta}, u) \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2] d\boldsymbol{\theta} du \\
&\qquad + \int_{\mathbb{R}^{d+1}} p_t(t, \boldsymbol{\theta}, u) [\nabla_u \cdot \widehat{g}_1(t, \boldsymbol{\theta}, u) + \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du, \qquad \text{(C.3)}
\end{aligned}
$$

where the second equality is by $p_t \nabla \log(p_t) = \nabla p_t$ and the third equality is by applying integration by parts. Then plugging (C.2) and (C.3) into (C.1), we get

$$\begin{aligned}\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} &= -\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 + \nabla_{\boldsymbol{\theta}} \log(p_t)\|_2^2 - \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|u/\sigma_u^2 + \nabla_u \log(p_t)|^2 \\ &\quad + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du.\end{aligned}$$

This completes the proof. $\square$

C.2 Proof of Lemma B.2

Proof of Lemma B.2. We remind the readers the definitions of $\widehat{g}_1$ and $\widehat{g}_2$ in (A.1) and (A.1). We have

$$\begin{aligned}&\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du \\ &\quad = 2\alpha \mathbb{E}_S\bigg[(f(p_t, \mathbf{x}) - y) \int_{\mathbb{R}^{d+1}} \big(uh(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - u\Delta h(\boldsymbol{\theta}, \mathbf{x})\big) p_t(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du\bigg].\end{aligned}$$

Denote $I(\boldsymbol{\theta}, u, \mathbf{x}) = uh(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - u\Delta h(\boldsymbol{\theta}, \mathbf{x})$, then we have

$$|\nabla_u I(\boldsymbol{\theta}, u, \mathbf{x})| = |h(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \Delta h(\boldsymbol{\theta}, \mathbf{x})| \le (G_1/\sigma_u^2 + G_3/\sigma_{\boldsymbol{\theta}}^2)\|\boldsymbol{\theta}\|_2 + G_2/\sigma_u^2 + G_4, \tag{C.4}$$

where the inequality holds by Lemma A.1. Similarly, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}} I(\boldsymbol{\theta}, u, \mathbf{x})\|_2 &= \|u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})/\sigma_u^2 + u\nabla_{\boldsymbol{\theta}}\big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta}\big)/\sigma_{\boldsymbol{\theta}}^2 - u\nabla_{\boldsymbol{\theta}} \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\big)\|_2 \\ &\le (G_3/\sigma_u^2 + G_5/\sigma_{\boldsymbol{\theta}}^2 + G_6)|u|.\end{aligned} \tag{C.5}$$

Therefore, combining the bounds in (C.4) and (C.5) yields

$$\sqrt{\nabla_u I(\boldsymbol{\theta}, u, \mathbf{x})^2 + \|\nabla_{\boldsymbol{\theta}} I(\boldsymbol{\theta}, u, \mathbf{x})\|_2^2} \le (G_1/\sigma_u^2 + G_3/\sigma_{\boldsymbol{\theta}}^2 + G_3/\sigma_u^2 + G_5/\sigma_{\boldsymbol{\theta}}^2 + G_6)\sqrt{u^2 + \|\boldsymbol{\theta}\|_2^2} + G_2/\sigma_u^2 + G_4.$$

By Lemma A.7, we have that

$$\begin{aligned}\mathbb{E}_{p_t}[I(\boldsymbol{\theta}_t, u_t, \mathbf{x})] - \mathbb{E}_{p_0}[I(\boldsymbol{\theta}_0, u_0, \mathbf{x})] &\le \Big[\big((G_1 + G_3)/\sigma_u^2 + (G_3 + G_5)/\sigma_{\boldsymbol{\theta}}^2 + G_6\big) 2\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d} + G_2/\sigma_u^2 + G_4\Big] \\ &\quad \cdot \mathcal{W}(p_0, p_t) \\ &\le A_2 \sqrt{D_{\mathrm{KL}}(p_t||p_0)},\end{aligned}$$

where the last inequality is by Lemma 5.4 and $A_2 = 2\big[\big((G_1{+}G_3)/\sigma_u^2{+}(G_3{+}G_5)/\sigma_{\boldsymbol{\theta}}^2{+}G_6\big)2\sqrt{\sigma_u^2 + \sigma_{\boldsymbol{\theta}}^2 \cdot d}{+}G_2/\sigma_u^2 + G_4\big] \max\{\sigma_u, \sigma_{\boldsymbol{\theta}}\}$. By $\mathbb{E}_{p_0}[I(\boldsymbol{\theta}_0, u_0, \mathbf{x})] = \mathbb{E}_{p_0}[u_0]\mathbb{E}_{p_0}[h(\boldsymbol{\theta}_0, \mathbf{x})/\sigma_u^2 + \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}) \cdot \boldsymbol{\theta}_0/\sigma_{\boldsymbol{\theta}}^2 - \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x})] = 0$, we further have

$$\mathbb{E}_{p_t}[I(\boldsymbol{\theta}_t, u_t, \mathbf{x})] \le A_2 \sqrt{D_{\mathrm{KL}}(p_t||p_0)}. \tag{C.6}$$

Then we have

$$\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u/\sigma_u^2 + \widehat{g}_2 \cdot \boldsymbol{\theta}/\sigma_{\boldsymbol{\theta}}^2 - \nabla \cdot \widehat{g}_1 - \nabla \cdot \widehat{g}_2] d\boldsymbol{\theta} du$$
$$= 2\alpha \mathbb{E}_S\big[(\mathbf{f}(t) - y)\mathbb{E}_{p_t}[I(\boldsymbol{\theta}_t, u_t, \mathbf{x})]\big]$$
$$\leq 2\alpha A_2 \sqrt{D_{\mathrm{KL}}(p_t || p_0)} \sqrt{L(p_t)},$$

where the last inequality is by (C.6) and Cauchy-Schwarz inequality. This completes the proof. □

References

Allen-Zhu, Z. and Li, Y. (2019). What can ResNet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems.*

Allen-Zhu, Z. and Li, Y. (2020). Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413* .

Allen-Zhu, Z., Li, Y. and Liang, Y. (2019a). Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems.*

Allen-Zhu, Z., Li, Y. and Song, Z. (2019b). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning.*

Arora, S., Du, S., Hu, W., Li, Z. and Wang, R. (2019a). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning.*

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. and Wang, R. (2019b). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems.*

Arora, S., Ge, R., Neyshabur, B. and Zhang, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning.*

Bai, Y. and Lee, J. D. (2019). Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619* .

Bakry, D., Gentil, I. and Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators*, vol. 348. Springer Science & Business Media.

Bartlett, P. L., Foster, D. J. and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems.*

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3** 463–482.

Brutzkus, A. and Globerson, A. (2017). Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org.

Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X. and Gu, Q. (2019). Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198* .

Cao, Y. and Gu, Q. (2019a). Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*.

Cao, Y. and Gu, Q. (2019b). Tight sample complexity of learning one-hidden-layer convolutional neural networks. In *Advances in Neural Information Processing Systems*.

Cao, Y. and Gu, Q. (2020). Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Chen, Z., Cao, Y., Zou, D. and Gu, Q. (2019). How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360* .

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*.

Chizat, L., Oyallon, E. and Bach, F. (2019). On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*.

Donsker, M. D. and Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics* **36** 183–212.

Du, S., Lee, J., Li, H., Wang, L. and Zhai, X. (2019a). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*.

Du, S. S., Lee, J. D. and Tian, Y. (2018a). When is a convolutional filter easy to learn? In *International Conference on Learning Representations*.

Du, S. S., Lee, J. D., Tian, Y., Singh, A. and Poczos, B. (2018b). Gradient descent learns one-hidden-layer CNN: Dont be afraid of spurious local minima. In *International Conference on Machine Learning*.

Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2019b). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

Fang, C., Dong, H. and Zhang, T. (2019a). Over parameterized two-level neural networks can learn near optimal feature representations. *arXiv preprint arXiv:1910.11508* .

Fang, C., Gu, Y., Zhang, W. and Zhang, T. (2019b). Convex formulation of overparameterized deep neural networks. *arXiv preprint arXiv:1911.07626* .

Golowich, N., Rakhlin, A. and Shamir, O. (2018). Size-independent sample complexity of neural networks. In *Conference On Learning Theory*.

Gunasekar, S., Lee, J., Soudry, D. and Srebro, N. (2018a). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*.

Gunasekar, S., Lee, J. D., Soudry, D. and Srebro, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B. and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* **29** 82–97.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*.

Ji, Z. and Telgarsky, M. (2019). The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*.

Ji, Z. and Telgarsky, M. (2020). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Li, X., Lu, J., Wang, Z., Haupt, J. and Zhao, T. (2018a). On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond. *arXiv preprint arXiv:1806.05159* .

Li, Y., Ma, T. and Zhang, H. (2018b). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*.

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

Li, Y. and Yuan, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*.

Lyu, K. and Li, J. (2019). Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890* .

Mei, S., Misiakiewicz, T. and Montanari, A. (2019). Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*.

Mei, S., Montanari, A. and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences* **115** E7665–E7671.

Meir, R. and Zhang, T. (2003). Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research* **4** 839–860.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Nacson, M. S., Gunasekar, S., Lee, J. D., Srebro, N. and Soudry, D. (2019a). Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *arXiv preprint arXiv:1905.07325* .

Nacson, M. S., Srebro, N. and Soudry, D. (2019b). Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*.

Neyshabur, B., Bhojanapalli, S., McAllester, D. and Srebro, N. (2018). A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representation*.

Neyshabur, B., Tomioka, R. and Srebro, N. (2015). Norm-based capacity control in neural networks. In *Conference on Learning Theory*.

Nitanda, A. and Suzuki, T. (2019). Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870* .

Otto, F. and Villani, C. (2000). Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis* **173** 361–400.

Oymak, S. and Soltanolkotabi, M. (2019). Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674* .

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* **529** 484–489.

Soltanolkotabi, M. (2017). Learning ReLUs via gradient descent. In *Advances in Neural Information Processing Systems*.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* **19** 2822–2878.

Su, L. and Yang, P. (2019). On learning over-parameterized neural networks: A functional approximation prospective. In *Advances in Neural Information Processing Systems*.

Tian, Y. (2017). An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560* .

Wei, C., Lee, J. D., Liu, Q. and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. *Advances in Neural Information Processing Systems* .

Xu, P., Chen, J., Zou, D. and Gu, Q. (2018). Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*.

Zhang, X., Yu, Y., Wang, L. and Gu, Q. (2019). Learning one-hidden-layer ReLU networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L. and Dhillon, I. S. (2017). Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning*.

Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888* .

Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2019). Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning* .

Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*.