A Generalized Neural Tangent Kernel Analysis for Two-layer Neural Networks

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

Abstract

A recent breakthrough in deep learning theory shows that the training of over-parameterized deep neural networks can be characterized by a kernel function called *neural tangent kernel* (NTK). However, it is known that this type of results does not perfectly match the practice, as NTK-based analysis requires the network weights to stay very close to their initialization throughout training, and cannot handle regularizers or gradient noises. In this paper, we provide a generalized neural tangent kernel analysis 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. We also establish a novel generalization error bound for two-layer neural networks trained by noisy gradient descent with weight decay.

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.

A recent line of work studies the learning of over-parameterized neural networks in the so-called "neural tangent kernel (NTK) regime" (Jacot et al., 2018). It has been shown that the training of over-parameterized deep neural networks can be characterized by the training dynamics of kernel regression with the *neural tangent kernel* (NTK). Based on this, fast convergence rates can be proved for over-parameterized neural networks trained with randomly initialized (stochastic) gradient descent (Du et al., 2019b; Allen-Zhu et al., 2019b; Du et al., 2019a; Zou et al., 2019; Zou and Gu, 2019). Moreover, it has also been shown that target functions in the NTK-induced reproducing kernel Hilbert space (RKHS) can be learned by wide enough neural networks with good generalization error (Arora et al., 2019a,b; Cao and Gu, 2019).

Despite having beautiful theoretical results, the NTK-based results are known to have their limitations, for not perfectly matching the empirical observations in many aspects. Specifically, NTK-based analysis requires that the network weights stay very close to their initialization throughout

---

*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, China; e-mail: tongzhang@tongzhang-ml.org

the training. Moreover, due to this requirement, NTK-based analysis cannot handle regularizers such as weight decay, or large additive noises in the noisy gradient descent.

Given the advantages and disadvantages of the existing NTK-based results, a natural question is:

Is it possible to establish the NTK-type results under more general settings?

In this paper, we give an affirmative answer to this question by utilizing a mean-field analysis (Chizat and Bach, 2018; Mei et al., 2018, 2019; Wei et al., 2019; Fang et al., 2019a) to study neural tangent kernel. We show that with appropriate scaling, two-layer neural networks trained with noisy gradient descent and weight decay can still enjoy the nice theoretical guarantees.

We summarize the contributions of our paper as follows:

- Our analysis demonstrates that neural network training with noisy gradient 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 can 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 in the distribution space. A direct consequence of our analysis is the linear convergence of noisy gradient descent up to certain accuracy for regularized neural network training. [1]

- We establish generalization bounds for the neural networks trained with noisy gradient descent with weight decay regularization. Our result shows that the infinitely wide neural networks trained by noisy gradient descent with weight decay can learn a class of functions that are defined based on a bounded χ^2-divergence to initialization distribution. Different from standard NTK-type generalization bounds (Allen-Zhu et al., 2019a; Arora et al., 2019a; Cao and Gu, 2019), our result can handle explicit regularization. Moreover, our proof is based on an extension of the proof technique in Meir and Zhang (2003) from discrete distributions to continuous distributions, which may be of independent interest.

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 $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, and any positive integer p, we denote the ℓ_p norm of $\mathbf{x}$ as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p \right)^{1/p}$. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{m \times n}$, we denote by $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ its spectral and Frobenius norms respectively. We also define $\|\mathbf{A}\|_{\infty,\infty} = \max\{|A_{ij}| : 1 \le i \le m, 1 \le j \le n\}$. 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 $[n] = \{1, \ldots, n\}$. 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 such that $a_n \le Cb_n$. We also introduce $\widetilde{O}(\cdot)$ to hide the logarithmic terms in the Big-O notations.

At last, for two distributions p and p', we define the Kullback–Leibler divergence (KL-divergence) and χ^2-divergence between p and p' as follows:

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

---

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

2   Related Work

Our work is motivated by the recent study of neural network training in the "neural tangent kernel regime". 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, Allen-Zhu et al. (2019b); Du et al. (2019a); Zou et al. (2019) proved the global convergence of (stochastic) gradient descent under various settings. Such convergence is later studied by a line of work (Zou and Gu, 2019) with improved network width conditions in various settings. Su and Yang (2019); Cao et al. (2019) studied the convergence along different eigendirections of the NTK. Chizat et al. (2019) extended the similar idea to a more general framework called "lazy training". Liu et al. (2020) studied the optimization for over-parameterized systems of non-linear equations. Allen-Zhu et al. (2019a); Arora et al. (2019a); Cao and Gu (2020, 2019) established generalization bounds for over-parameterized neural networks trained by (stochastic) gradient descent. Li et al. (2019) studied noisy gradient descent with a certain learning rate schedule for a toy example.

Our analysis follows the mean-field framework adopted in the recent line of work (Bach, 2017; Chizat and Bach, 2018; Mei et al., 2018, 2019; Wei et al., 2019; Fang et al., 2019a,b). Bach (2017) studied the generalization performance of infinitely wide two-layer neural networks under the mean-field setting. 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,b) proposed a new concept called neural feature repopulation and extended the mean-field analysis.

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 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

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

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

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

λ = reg

α = scaling

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 $\zeta_{u,j} \sim N(0, 2\eta)$, $j \in [m]$
  $u_{t+1,j} = u_{t,j} - \eta \nabla_u \widehat{Q}(\{(\boldsymbol{\theta}_{t,j}, u_{t,j})\}_{j=1}^m) - \sqrt{\lambda}\zeta_{u,j}$
  Draw Gaussian noises $\boldsymbol{\zeta}_{\boldsymbol{\theta},j} \sim N(0, 2\eta\mathbf{I}_d)$, $j \in [m]$
  $\boldsymbol{\theta}_{t+1,j} = \boldsymbol{\theta}_{t,j} - \eta \nabla_{\boldsymbol{\theta}} \widehat{Q}(\{(\boldsymbol{\theta}_{t,j}, u_{t,j})\}_{j=1}^m) - \sqrt{\lambda}\boldsymbol{\zeta}_{\boldsymbol{\theta},j}$
end for
---

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 $p_0(\boldsymbol{\theta}, u) \propto \exp[-u^2/2 - \|\boldsymbol{\theta}\|_2^2/2]$. Then for finite-width neural network (3.2), we define the training objective function as

$$\widehat{Q}(\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m) = \mathbb{E}_S[\phi(f_m(\{(\boldsymbol{\theta}_j, u_j)\}_{j=1}^m, \mathbf{x}), y)] + \frac{\lambda}{m}\sum_{j=1}^m \left(\frac{u_j^2}{2} + \frac{\|\boldsymbol{\theta}_j\|_2^2}{2}\right), \tag{3.3}$$

where $\mathbb{E}_S[\cdot]$ denotes the empirical average over the training sample S, and $\lambda > 0$ is a regularization parameter. It is worth noting that when the network is wide enough, the neural network training is in the "interpolation regime", which gives zero training loss (first term in (3.3)). Therefore, even with a very large scaling parameter α, weight decay (second term in (3.3)) is still effective.

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) 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]:

$$\frac{dp_t(\boldsymbol{\theta}, u)}{dt} = -\nabla_u[p_t(\boldsymbol{\theta}, u)g_1(t, \boldsymbol{\theta}, u)] - \nabla_{\boldsymbol{\theta}} \cdot [p_t(\boldsymbol{\theta}, u)g_2(t, \boldsymbol{\theta}, u)] + \lambda\Delta[p_t(\boldsymbol{\theta}, u)], \tag{3.4}$$

where

$$g_1(t, \boldsymbol{\theta}, u) = -\alpha\mathbb{E}_S[\nabla_{y'}\phi(f(p_t, \mathbf{x}), y)h(\boldsymbol{\theta}, \mathbf{x})] - \lambda u,$$
$$g_2(t, \boldsymbol{\theta}, u) = -\alpha\mathbb{E}_S[\nabla_{y'}\phi(f(p_t, \mathbf{x}), y)u\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta}, \mathbf{x})] - \lambda\boldsymbol{\theta}.$$

Below we give an informal proposition to describe the connection between Algorithm 1 and the PDE (3.4). One can refer to Mei et al. (2018); Chizat and Bach (2018); Mei et al. (2019) for more details on such approximation results.

---

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

Proposition 3.1 (informal). Suppose that $h(\boldsymbol{\theta}, \mathbf{x})$ is sufficiently smooth, and PDE (3.4) has a unique solution p_t. Let $\{(\boldsymbol{\theta}_{t,j}, u_{t,j})\}_{j=1}^m$, $t \geq 0$ be output by Algorithm 1. 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 (3.4). It is worth noting that PDE (3.4) minimizes the following energy functional

$$Q(p) = L(p) + \lambda D_{\mathrm{KL}}(p \| p_0), \tag{3.5}$$

where $L(p) = \mathbb{E}_S[\phi((f(p, \mathbf{x}), y)]$ 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 (3.4) towards the global minimum of (3.5) is recently established (Mei et al., 2018; Chizat and Bach, 2018; Mei et al., 2019; Fang et al., 2019a).

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

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

$$K_2(\mathbf{x}, \mathbf{x}') = \int h(\boldsymbol{\theta}, \mathbf{x}) h(\boldsymbol{\theta}, \mathbf{x}') p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du.$$

the proof on grad-decent [handwritten annotation]

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 the "node-wise" ℓ_2 distance. 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:

$$\mathbf{H}(p) = \mathbf{H}_1(p) + \mathbf{H}_2(p), \tag{3.6}$$

where $\mathbf{H}_1(p)_{i,j} = \mathbb{E}_p[u^2 \langle \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i), \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_j) \rangle]$ and $\mathbf{H}_2(p)_{i,j} = \mathbb{E}_p[h(\boldsymbol{\theta}, \mathbf{x}_i) h(\boldsymbol{\theta}, \mathbf{x}_j)]$. Note 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 in Algorithm 1.

We first introduce the following two assumptions, which are required in both optimization and generalization error analyses.

Assumption 4.1. The data inputs and responses are bounded: $\|\mathbf{x}_i\|_2 \leq 1$, $|y_i| \leq 1$ 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). We would also like to remark that the bound 1 is not essential, and can be replaced by any positive constant.

Assumption 4.2. 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:

$$|\widetilde{h}(z)| \leq G_1, \quad |\widetilde{h}'(z)| \leq G_2, \quad |\widetilde{h}''(z)| \leq G_3, \quad |(z\widetilde{h}'(z))'| \leq G_4, \quad |\widetilde{h}'''(z)| \leq G_5,$$

where $G_1, \ldots, G_5$ are absolute constants, and we set $G = \max\{G_1, \ldots, G_5\}$ to simplify the bound.

$h(\boldsymbol{\theta}, \mathbf{x}) = \widetilde{h}(\boldsymbol{\theta}^\top \mathbf{x})$ 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). Assumption 4.2 is satisfied by many smooth activation functions including the sigmoid and hyper-tangent functions.

4.1 Optimization Guarantees

In order to characterize the optimization dynamics defined by PDE (3.4), we need the following additional assumption.

Assumption 4.3. The Gram matrix of the neural tangent kernel defined in (3.6) 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are not parallel.

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

Theorem 4.4. Let $\lambda_0 = \sqrt{\Lambda/n}$ and suppose that PDE (3.4) has a unique solution p_t. Under Assumptions 4.1, 4.2 and 4.3, if

$$\alpha \geq 8\sqrt{A_2^2 + \lambda A_1^2} \cdot \lambda_0^{-2} R^{-1}, \tag{4.1}$$

where $R = \min\left\{\sqrt{d+1}, [\text{poly}(G, \log(1/\lambda_0))]^{-1}\lambda_0^2\right\}$, then for all $t \in [0, +\infty)$, the following result hold:

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

where $A_1 = 2G(d+1) + 4G\sqrt{d+1}$ and $A_2 = 16G\sqrt{d+1} + 4G$.

Theorem 4.4 shows that the loss of the neural network converges linearly up to $O(\lambda^2\lambda_0^{-4}\alpha^{-2})$ accuracy, and the convergence rate depends on the smallest eigenvalue of the NTK Gram matrix. This matches the results for square loss in the neural tangent kernel regime (Du et al., 2019b). However, we would like to emphasize that the algorithm we study here is noisy gradient descent, and the objective function involves a weight decay regularizer, both of which cannot be handled by

the standard technical tools used in the NTK regime (Allen-Zhu et al., 2019b; Du et al., 2019b,a; Zou and Gu, 2019). Theorem 4.4 also shows that the KL-divergence between p_t and p_0 is bounded and decreases as α increases. This is analogous to the standard NTK results Du et al. (2019b); Allen-Zhu et al. (2019b); Zou et al. (2019) where the Euclidean distance between the parameter returned by (stochastic) gradient descent and its initialization is bounded, and decreases as the network width increases.

The results in Theorem 4.4 can also be compared with an earlier attempt by Mei et al. (2019), which uses mean-field analysis to explain NTK. While Mei et al. (2019) only reproduces the NTK-type results without regularization, our result holds for a more general setting with weight decay and noisy gradient. Another work by Tzen and Raginsky (2020) uses mean-field analysis to study the lazy training of two-layer network. They consider a very small variance in parameter initialization, which is quite different from the practice of neural network training. In contrast, our work uses standard random initialization, and exactly follows the lazy training setting with scaling factor α (Chizat et al., 2019). Moreover, Tzen and Raginsky (2020) only characterize the properties of the optimal solution without finite-time convergence result, while we characterize the whole training process with a linear convergence rate.

4.2 Generalization Bounds

Next, we study the generalization performance of the neural network obtained by minimizing the energy functional $Q(p)$. For simplicity, we consider the binary classification problem, and use the 0-1 loss $\ell^{0\text{-}1}(y', y) := \mathbb{1}\{y'y < 0\}$ to quantify the errors of the network, where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

The following theorem presents the generalization bound for neural networks trained by Algorithm 1.

Theorem 4.5. 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 $D_{\chi^2}(p_{\text{true}}||p_0) < \infty$, such that

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

for all $(\mathbf{x}, y) \in \text{supp}(\mathcal{D})$. Let p^* be the minimizer of the energy functional (3.5). Under Assumptions 4.1 and 4.2, if $\alpha \geq \sqrt{n\lambda} > 0$, then for any $\delta > 0$, with probability at least $1 - \delta$,

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

Theorem 4.5 gives the generalization bound for the global minimizer of the energy functional $Q(p)$ obtained by noisy gradient descent with weight decay. We can see that it gives a standard $1/\sqrt{n}$ error rate as long as p_{true} has a constant χ^2-divergence to p_0. Moreover, the χ^2-divergence $D_{\chi^2}(p_{\text{true}}||p_0)$ also quantifies the difficulty for a target function defined by p_{true} to be learnt. The larger $D_{\chi^2}(p_{\text{true}}||p_0)$ is, the more examples are needed to achieve the same target expected error.

Our generalization bound is different from existing NTK-based generalization results (Li and Liang, 2018; Allen-Zhu et al., 2019a; Arora et al., 2019a; Cao and Gu, 2019), which highly rely on the fact that the learned neural network weights are close to the initialization. Therefore, these generalization error bounds no longer hold with the presence of regularizer and are not applicable to

our setting. In addition, Bach (2017) studied the generalization bounds for two-layer homogeneous networks and their connection to the NTK-induced RKHS.[3] Our result based on the KL-divergence regularization is different from their setting and is not covered by their results.

5   Proof Sketch of the Main Results

In this section we present a proof sketch for Theorems 4.4 and 4.5.

5.1   Proof Sketch of Theorem 4.4

We first introduce the following definition of 2-Wasserstein distance. For two distributions p and p' over $\mathbb{R}^{d+1}$, we define

$$\mathcal{W}_2(p, p') = \left(\inf_{\gamma \in \Gamma(p, p')} \int_{\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}} \|\mathbf{z} - \mathbf{z}'\|_2^2 d\gamma(\mathbf{z}, \mathbf{z}') \right)^{1/2},$$

where $\Gamma(p, p')$ denotes the collection of all measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals p and p' on the first and second factors respectively.

We also introduce the perturbation region $\mathcal{B}(p_0, R) := \{p | \mathcal{W}_2(p, p_0) \leq R\}$ based on the Wasserstein distance to the initialization, where R defined in Theorem 4.4 gives the perturbation radius. We would like to highlight that compared with standard NTK-based analyses Allen-Zhu et al. (2019a); Arora et al. (2019a); Cao and Gu (2019) which are based on a perturbation region around initial weight parameter, our proof is based upon the 2-Wasserstein neighborhood around p_0. Such an extension is essential to handle weight decay and gradient noises and is one of our key technical contributions.

The proof of Theorem 4.5 can be divided into the following three steps.

Step 1: Landscape properties when p_t is close to p_0. We first consider the situation when the distribution p_t is close to initial distribution p_0.

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.

Lemma 5.1 shows that when p_t is close to p_0 in 2-Wasserstein distance, the Gram matrix at p_t is strictly positive definite. This further implies nice landscape properties around p_t, which enables our analysis in the next step.

Step 2: Loss and regularization bounds when p_t is close to p_0. With the results in Step 1, we establish loss and regularization bounds when p_t stays in $\mathcal{B}(p_0, R)$ for some time period $[0, t^*]$, with $t^* = \inf\{t \geq 0 : \mathcal{W}_2(p_t, p_0) > R\}$. We have $t^* = +\infty$ if the $\{t \geq 0 : \mathcal{W}_2(p_t, p_0) > R\} = \emptyset$. The following lemma shows that the loss function decreases linearly in the time perioud $[0, t^*]$.

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

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

where A_1 is defined in Theorem 4.4.

---

[3]Although it is not named "neural tangent kernel", the kernel function studied in (Bach, 2017) is essentially NTK.

Besides the bound on $L(p_t)$, we also have a bound on the KL-divergence between p_t and p_0, as is given in the following lemma.

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

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

where A_1 and A_2 are defined in Theorem 4.4.

Here we would like to remark that the bound in Lemma 5.3 does not increase with time t, which is an important feature of our result. This is achieved by jointly considering two types of bounds on the KL-divergence between p_t and p_0: the first type of bound is on the time derivative of $D_{\mathrm{KL}}(p_t, p_0)$ based on the training dynamics described by (3.4), and the second type of bound is a direct KL-divergence bound based on the monotonicity of the energy functional $Q(p_t)$. The detailed proof of this lemma is deferred to the appendix.

Step 3: Large scaling factor α ensures distribution closeness throughout training. When α is sufficiently large, p_t will not escape from the perturbation region. To show this, we utilize 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 in our setting p_0 is a Gaussian distribution.

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

$$\mathcal{W}_2(p, p_0) \leq 2\sqrt{D_{\mathrm{KL}}(p || p_0)}.$$

The main purpose of Lemma 5.4 is to build a connection between the 2-Wasserstein ball around p_0 and the KL-divergence ball.

We are now ready to finalize the proof. Note that given our results in Step 2, it suffices to show that $t^* = +\infty$, which is proved based on a reduction to absurdity.

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

$$\mathcal{W}_2(p_t, p_0) \leq 2D_{\mathrm{KL}}(p_t || p_0)^{1/2} \leq 2\big(4A_2^2 \alpha^{-2} \lambda_0^{-4} + 4A_1^2 \lambda \alpha^{-2} \lambda_0^{-4}\big)^{1/2} \leq R/2,$$

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 inequality in Lemma 5.2 and applying Jensen's inequality gives

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

This completes the proof. $\square$

5.2 Proof Sketch of Theorem 4.5

For any $M > 0$, we consider the following class of infinitely wide neural network functions characterized by the KL-divergence to initialization

$$\mathcal{F}_{\text{KL}}(M) = \big\{ f(p, \mathbf{x}) : D_{\text{KL}}(p\|p_0) \leq M \big\}. \tag{5.1}$$

Our proof consists of the following two steps.

Step 1: A KL-divergence based Rademacher complexity bound. Motivated by the KL-divergence regularization in the energy functional $Q(p)$, we first derive a Rademacher complexity bound for the function class $\mathcal{F}_{\text{KL}}(M)$, which is given as the following lemma.

Lemma 5.5. Suppose that $|h(\boldsymbol{\theta}, \mathbf{x})| \leq G$ for all $\boldsymbol{\theta}$ and $\mathbf{x}$, and $M \leq 1/2$. Then

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

Lemma 5.5 is another key technical contribution of our paper. Different from previous NTK-based generalization error analysis that utilizes the approximation of neural network functions with linear models (Lee et al., 2019; Cao and Gu, 2019), we use mean-field analysis to directly bound the Rademacher complexity of neural network function class. At the core of the proof for Lemma 5.5 is an extension of Meir and Zhang (2003) for discrete distributions to continuous distributions, which is of independent interest.

Step 2: Expected 0-1 loss bound over $\mathcal{F}_{\text{KL}}(M)$. We bound the expected 0-1 loss by the square root of the empirical square loss function and the Rademacher complexity.

Lemma 5.6. For any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds uniformly over all $f \in \mathcal{F}_{\text{KL}}(M)$:

$$\mathbb{E}_{\mathcal{D}}[\ell^{\text{0-1}}(f(\mathbf{x}), y)] \leq \sqrt{\mathbb{E}_S[|f(\mathbf{x}) - y|^2]} + 4\mathfrak{R}_n(\mathcal{F}_{\text{KL}}(M)) + 6\sqrt{\frac{\log(2/\delta)}{2n}}.$$

The bound in Lemma 5.6 utilizes the property of square loss instead of margin-based arguments (Bartlett et al., 2017). In this way, we are able to obtain a tighter bound, as our setting uses square loss as the objective.

We now finalize the proof by deriving the loss and regularization bounds at p^* and plugging them into Lemma 5.6.

Proof of Theorem 4.5. Let $\overline{D} = D_{\chi^2}(p_{\text{true}}\|p_0) < \infty$. Define

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

Then 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. The training loss of $\widehat{p}$ can be calculated as follows:

$$L(\widehat{p}) = \mathbb{E}_S\left[\alpha \int uh(\theta, \mathbf{x})\widehat{p}(u, \theta) du d\theta - y\right]^2 = \mathbb{E}_S\left(0 + \alpha \cdot \frac{1}{\alpha}y - y\right)^2 = 0. \tag{5.2}$$

10

Moreover, by the fact that KL-divergence is upper bounded by the χ^2-divergence, we have

$$D_{\mathrm{KL}}(\widehat{p}\|p_0) \leq D_{\chi^2}(\widehat{p}\|p_0) = \int \left[\frac{\alpha-1}{\alpha} + \frac{p_{\mathrm{true}}(\boldsymbol{\theta}, u)}{\alpha p_0(\boldsymbol{\theta}, u)} - 1 \right]^2 p_0(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du = \alpha^{-2}\overline{D}, \qquad (5.3)$$

where the last equation is by the definition of χ^2-divergence. Now we have

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

where the first inequality follows by the optimality of p^*, and we plug (5.2), (5.3) into the definition of the energy function $Q(p)$ in (3.5) to obtain the second inequality. Applying the definition of $Q(p)$ again gives the following two bounds:

$$L(p^*) = \mathbb{E}_S[|f(p^*, \mathbf{x}) - y|^2] \leq \alpha^{-2}\lambda\overline{D}, \qquad (5.4)$$
$$D_{\mathrm{KL}}(p^*\|p_0) \leq \alpha^{-2}\overline{D}. \qquad (5.5)$$

By (5.5), we have $f(p^*, \mathbf{x}) \in \mathcal{F}_{\mathrm{KL}}(\alpha^{-2}\widehat{D})$. Therefore, applying Lemma 5.6 with $M = \alpha^{-2}\widehat{D}$ gives

$$\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(p^*, \mathbf{x}), y)] \leq \sqrt{\mathbb{E}_S[|f(p^*, \mathbf{x}) - y|^2]} + 4\mathfrak{R}_n(\mathcal{F}_{\mathrm{KL}}(\alpha^{-2}\widehat{D})) + 6\sqrt{\frac{\log(2/\delta)}{2n}}$$

$$\leq \sqrt{\alpha^{-2}\lambda\overline{D}} + 8G\alpha\sqrt{\frac{\alpha^{-2}\widehat{D}}{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}}$$

$$\leq (8G+1)\sqrt{\frac{\widehat{D}}{n}} + 6\sqrt{\frac{\log(2/\delta)}{2n}},$$

where the second inequality follows from (5.4) and Lemma 5.5, and the third inequality follows from the assumption that $\alpha \geq \sqrt{n\lambda}$. This finishes the proof. $\qquad \square$

6   Conclusion

In this paper, we demonstrate that the neural tangent kernel based regression can characterize neural network training dynamics in a general setting where weight decay and gradient noises are implemented. This leads to the linear convergence of noisy gradient descent up to certain accuracy. Compared with existing analysis in the neural tangent kernel regime, our work points out an important observation that as long as the distribution of parameters stays close to the initialization, it does not matter whether the parameters themselves are close to their initial values. We also establish a novel generalization bound for the neural network trained by noisy gradient descent with weight decay regularization.

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. We denote $\mathbf{f}(t) = (f(p_t, \mathbf{x}_1), \ldots, f(p_t, \mathbf{x}_n))^\top$. Moreover,

we define

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

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

A.1 Proof of Lemma 5.1

Here we give the proof of Lemma 5.1. The following lemma 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 $|h(\boldsymbol{\theta}, \mathbf{x})| \leq G$, $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \leq G$, $|\Delta h(\boldsymbol{\theta}, \mathbf{x})| \leq G$, $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_1, x) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_2, x)\|_2 \leq G\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$, $\|\nabla_{\boldsymbol{\theta}}(\nabla_{\boldsymbol{\theta}} h(\theta, \mathbf{x}) \cdot \boldsymbol{\theta})\|_2 \leq G$, $\|\nabla_{\boldsymbol{\theta}} \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \leq G$.

We also give the following two lemmas to characterize the difference between the Gram matrices defined with p_0 and some other distribution p that is close to p_0 in 2-Wasserstein distance.

Lemma A.2. Under Assumptions 4.1 and 4.2, for any distribution p with $\mathcal{W}_2(p, p_0) \leq \sqrt{d+1}$ and any $r > 0$,

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

Lemma A.3. Under Assumptions 4.1 and 4.2, for any distribution p with $\mathcal{W}_2(p, p_0) \leq \sqrt{d+1}$,

$$\|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \leq 2G^2 \mathcal{W}_2(p, p_0).$$

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

Lemma A.4. The initialization distribution p_0 satisfies the following tail bound:

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

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.

$$R = \min\left\{\sqrt{d+1}, [\text{poly}(G, \log(n/\Lambda))n/\Lambda]^{-1}\right\}$$

$$\leq \min\left\{\sqrt{d+1}, \left(8G^2\sqrt{8d+10} + 64G^2 \log(8\Lambda^{-1}nG^2)\right)^{-1} n^{-1}\Lambda\right\}.$$

Note that the definition of R, the results for Lemmas A.2 and A.3 hold for all p with $\mathcal{W}_2(p, p_0) \leq R$. Now by Lemma A.2, for any p with $\mathcal{W}_2(p, p_0) \leq R$ and any $r > 0$,

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

Choose $r = 2\sqrt{\log(8\Lambda^{-1}nG^2)}$, then by Lemma A.4 we have

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

Moreover, by the definition of R, we have

$$R \leq \left(8G^2\sqrt{8d+10} + 16G^2r^2\right)^{-1}n^{-1}\Lambda. \tag{A.5}$$

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

$$\|\mathbf{H}_1(p) - \mathbf{H}_1(p_0)\|_{\infty,\infty} \leq G^2 R\sqrt{8d+10} + 2r^2 G^2 R + G^2 \mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)] \tag{A.6}$$

$$\leq \frac{\Lambda}{8n} + \frac{\Lambda}{8n} \tag{A.7}$$

$$= \frac{\Lambda}{4n}. \tag{A.8}$$

By Lemma A.3, for any distribution p with $\mathcal{W}_2(p, p_0) \leq R$,

$$\|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \leq 2G^2 R. \tag{A.9}$$

The definition of R also leads to the following bound:

$$R \leq (8G^2)^{-1}n^{-1}\Lambda. \tag{A.10}$$

Therefore we can plug the bound (A.10) into (A.9), which gives

$$\|\mathbf{H}_2(p) - \mathbf{H}_2(p_0)\|_{\infty,\infty} \leq \frac{\Lambda}{4n}. \tag{A.11}$$

Combining (A.6) and (A.11) further gives

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

Then by standard matrix perturbation bounds, we have $\lambda_{\min}(\mathbf{H}(p)) \geq \lambda_{\min}(\mathbf{H}(p_0)) - \|\mathbf{H}(p) - \mathbf{H}(p_0)\|_2 \geq \lambda_{\min}(\mathbf{H}(p_0)) - n\|\mathbf{H}(p) - \mathbf{H}(p_0)\|_{\infty,\infty} \geq \Lambda/2$, 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.5. Let p_t be the solution of PDE (3.4). Then the following identity holds.

$$\frac{\partial L(p_t)}{\partial t} = -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta}du - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta}du$$

$$+ \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u + \widehat{g}_2 \cdot \boldsymbol{\theta} - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2]d\boldsymbol{\theta}du. \tag{A.12}$$

Lemma A.5 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.6 below essentially serves as a bound on the first two terms on the right-hand side of (A.12).

Lemma A.6. Under Assumptions 4.1, 4.2 and 4.3, let λ_0 be defined in Theorem 4.4. Then for $t \leq t^*$, it holds that

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

Lemma A.7. Under Assumptions 4.1 and 4.2, let A_1 be defined in Theorem 4.4. Then for $t \leq t^*$, it holds that

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

We now present the proof of Lemma 5.2, which is based on the calculations in Lemmas A.5, A.6 and A.7 as well as the application of Gronwall's inequality.

Proof of Lemma 5.2. By Lemma A.5, we have

$$\frac{\partial L(p_t)}{\partial t} = -\underbrace{\left[\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta}du + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta}d\boldsymbol{\theta}du\right]}_{I_1}$$

$$+ \lambda\underbrace{\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u + \widehat{g}_2 \cdot \boldsymbol{\theta} - \nabla \cdot \widehat{g}_1 - \nabla \cdot \widehat{g}_2]d\boldsymbol{\theta}du}_{I_2}., \tag{A.13}$$

For I_1, we have

$$I_1 = 4\alpha^2 \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\big[|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)h(\boldsymbol{\theta}, \mathbf{x})]|^2 \tag{A.14}$$

$$+ \|\mathbb{E}_S[(f(p_t, \mathbf{x}) - y)u\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta}, \mathbf{x})]\|_2^2\big]d\boldsymbol{\theta}du \tag{A.15}$$

$$\geq 2\alpha^2\lambda_0^2 L(p_t), \tag{A.16}$$

where the equation 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.6. For I_2, we directly apply Lemma A.7 and obtain

$$I_2 \leq 2A_1\alpha\lambda\sqrt{L(p_t)}. \tag{A.17}$$

Plugging the bounds (A.16) and (A.17) into (A.13) yields

$$\frac{\partial L(p_t)}{\partial t} \leq -2\alpha^2\lambda_0^2 L(p_t) + 2A_1\alpha\lambda\sqrt{L(p_t)}. \tag{A.18}$$

14

Now denote $V(t) = \sqrt{L(p_t)} - A_1\lambda\alpha^{-1}\lambda_0^{-2}$. Then (A.18) 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 $V(0) = \sqrt{L(p_0)} - A_1\lambda\alpha^{-1}\lambda_0^{-2} \leq \sqrt{L(p_0)} \leq 1$, we have

$$\sqrt{L(p_t)} \leq \exp(-\alpha^2\lambda_0^2 t) + A_1\lambda\alpha^{-1}\lambda_0^{-2}. \tag{A.19}$$

This completes the proof. □

A.3   Proof of Lemma 5.3

In this subsection we present the proof of Lemma 5.3.

Lemma A.8. Under Assumptions 4.1, 4.2 and 4.3, let λ_0 be defined in Theorem 4.4. Then for $t \leq t^*$ the following inequality holds

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

If $\lambda \neq 0$, the KL distance bound given by Lemma A.8 depends on t, we can give a tighter bound by the monotonically deceasing property of $Q(p_t)$ given by 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 A.9. Let p_t be the solution of PDE (3.4). Then $Q(p_t)$ is monotonically deceasing, i.e.,

$$\frac{\partial Q(p_t)}{\partial t} \leq 0. \tag{A.20}$$

Proof of Lemma 5.3. Notice that for $\lambda = 0$, Lemma A.8 directly implies the conclusion. So in the rest of the proof we consider the situation where $\lambda > 0$. Denote $t_0 = A_1^{-1}\alpha^{-1}\lambda^{-1}$, we consider two cases $t_0 \geq t_*$ and $t_0 < t_*$ respectively.

If $t_0 \geq t_*$, then for $t \leq t^*$ we have $t \leq t_0$

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

where the first inequality is by Lemma A.8 and the second inequality is by $t \leq t_0$.

---

[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.

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

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

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 A.9 implies that,

$$D_{\mathrm{KL}}(p_t||p_0) \le \lambda^{-1}Q(p_t) \le \lambda^{-1}Q(p_{t_0}). \tag{A.21}$$

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

$$L(p_t) \le 2\exp(-2\alpha^2\lambda_0^2 t) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}. \tag{A.22}$$

Plugging $t_0 = A_1^{-1}\alpha^{-1}\lambda^{-1}$ into (A.22) gives

$$\begin{aligned}
L(p_{t_0}) &\le 2\exp(-2\alpha^2\lambda_0^2 t_0) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4} \\
&= 2\exp\left(-2A_1^{-1}\lambda^{-1}\alpha\lambda_0^2\right) + 2A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4} \\
&\le 4A_1^2\lambda^2\alpha^{-2}\lambda_0^{-4}, \tag{A.23}
\end{aligned}$$

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.8, we have

$$D_{\mathrm{KL}}(p_{t_0}||p_0) \le 2A_2^2\alpha^{-2}\lambda_0^{-4} + 2A_2^2A_1^2\lambda^2\lambda_0^{-4}t_0^2 = 4A_2^2\alpha^{-2}\lambda_0^{-4}. \tag{A.24}$$

Plugging (A.23) and (A.24) into (A.21) gives

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

This completes the proof. $\qquad\square$

A.4   Proof of Lemma 5.5

Proof of Lemma 5.5. 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

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

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

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

Therefore

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

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

$$\int\exp\left(\frac{\gamma^2}{2n^2}\sum_{i=1}^{n}u^2 h^2(\boldsymbol{\theta}, \mathbf{x}_i)\right)p_0(\boldsymbol{\theta}, u)d\boldsymbol{\theta}du \leq \int\exp\left(\frac{\gamma^2 G^2}{2n}u^2\right)p_0(\boldsymbol{\theta}, u)d\boldsymbol{\theta}du$$

$$= \frac{1}{\sqrt{2\pi}}\cdot\sqrt{\frac{2\pi}{1-\gamma^2 G^2 n^{-1}}}$$

$$= \sqrt{\frac{1}{1-\gamma^2 G^2 n^{-1}}}.$$

Therefore we have

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

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

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

This completes the proof. $\qquad\square$

A.5   Proof of Lemma 5.6

Proof of Lemma 5.6. We first introduce the following ramp loss function, which is frequently used in the analysis of generalization bounds (Bartlett et al., 2017; Li et al., 2018) for binary classification problems.

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

Then by definition, we see that $\ell_{\mathrm{ramp}}(y', y)$ is 2-Lipschitz in the first argument, $\ell_{\mathrm{ramp}}(y, y) = 0$, $|\ell_{\mathrm{ramp}}(y', y)| \leq 1$, and

$$\ell^{0\text{-}1}(y', y) \leq \ell_{\mathrm{ramp}}(y', y) \leq |y' - y| \qquad (\text{A.25})$$

for all $y' \in \mathbb{R}$ and $y \in \{\pm 1\}$. By the Lipschitz and boundedness properties of the ramp loss, we apply the standard properties of Rademacher complexity (Bartlett and Mendelson, 2002; Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014) and obtain that with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathcal{D}}[\ell_{\text{ramp}}(f(\mathbf{x}), y)/2] \leq \mathbb{E}_S[\ell_{\text{ramp}}(f(\mathbf{x}), y)/2] + 2\mathfrak{R}_n(\mathcal{F}_{\text{KL}}(M)) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

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

$$\mathbb{E}_{\mathcal{D}}[\ell^{0\text{-}1}(f(\mathbf{x}), y)] \leq 2\mathbb{E}_{\mathcal{D}}[\ell_{\text{ramp}}(f(\mathbf{x}), y)/2]$$

$$\leq \mathbb{E}_S[\ell_{\text{ramp}}(f(\mathbf{x}), y)] + 4\mathfrak{R}_n(\mathcal{F}_{\text{KL}}(M)) + 6\sqrt{\frac{\log(2/\delta)}{2n}}$$

$$\leq \mathbb{E}_S[|f(\mathbf{x}) - y|] + 4\mathfrak{R}_n(\mathcal{F}_{\text{KL}}(M)) + 6\sqrt{\frac{\log(2/\delta)}{2n}}$$

$$\leq \sqrt{\mathbb{E}_S[|f(\mathbf{x}) - y|^2]} + 4\mathfrak{R}_n(\mathcal{F}_{\text{KL}}(M)) + 6\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Here the first and third inequalities follow by the first and second parts of the inequality in (A.25) respectively, and the last inequality uses Jensen's inequality. This completes the proof. $\quad\square$

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 $\|\mathbf{x}\|_2 \leq 1$.

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

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

By $|\widetilde{h}(z)| \leq G$ in Assumption 4.2 and $\|\mathbf{x}\|_2 \leq 1$ in Assumption 4.1, we have

$$|h(\boldsymbol{\theta}, \mathbf{x})| \leq G,$$

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

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

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

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

Moreover, based on the same assumptions we also have

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

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

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

Finally, by $|\widetilde{h}'''(z)| \leq G$ and $\|\mathbf{x}\|_2 \leq 1$, we have

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

This completes the proof. $\qquad\square$

B.2 Proof of Lemma A.2

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

Lemma B.1. For $\mathcal{W}_2(p, p_0) \leq \sqrt{d+1}$, the following bound holds:

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

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 B.2. For $\mathcal{W}_2(p, p_0) \leq \sqrt{d+1}$, 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

$$
\big|\mathbb{E}_p[g(u, \boldsymbol{\theta})] - \mathbb{E}_{p_0}[g(u_0, \boldsymbol{\theta}_0)]\big| \leq \Big(2C_1\sqrt{d+1} + C_2\Big)\mathcal{W}_2(p, p_0).
$$

Proof of Lemma A.2. Let π^* be the optimal coupling of $\mathcal{W}_2(p, p_0)$. Then we have

$$
\begin{aligned}
\big|\mathbf{H}_1(p)_{i,j} - \mathbf{H}_1(p_0)_{i,j}\big| &= \big|\mathbb{E}_{\pi^*}[u^2\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_j)] - \mathbb{E}_{\pi^*}[u_0^2\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}_j)]\big| \\
&\leq \underbrace{\big|\mathbb{E}_{\pi^*}[(u^2 - u_0^2)\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j)]\big|}_{I_1} \\
&\quad + \underbrace{\big|\mathbb{E}_{\pi^*}\big[u_0^2\big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big)\big]\big|}_{I_2}.
\end{aligned}
\tag{B.1}
$$

We first bound I_1 as follows.

$$
\begin{aligned}
I_1 &\leq G^2 \mathbb{E}_{\pi^*}[|u^2 - u_0^2|] \\
&\leq G^2 \sqrt{\mathbb{E}_{\pi^*}[(u - u_0)^2]} \sqrt{\mathbb{E}_{\pi^*}[(u + u_0)^2]} \\
&\leq G^2 \mathcal{W}_2(p, p_0) \sqrt{2\mathbb{E}_p[u^2] + 2\mathbb{E}_{p_0}[u_0^2]} \\
&\leq G^2 \mathcal{W}_2(p, p_0) \sqrt{8d + 10},
\end{aligned}
\tag{B.2}
$$

where the first inequality is by $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\|_2 \leq G$ 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 B.1. Next, We bound I_2 in (B.1). For any given $r > 0$ we have

$$
\begin{aligned}
I_2 &\leq \mathbb{E}_{\pi^*}\big[u_0^2 \mathbb{1}(|u_0 < r|)\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\quad + \mathbb{E}_{\pi^*}\big[u_0^2 \mathbb{1}(|u_0 \geq r|)\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\leq r^2 \mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] + 2G^2 \mathbb{E}_{\pi^*}[u_0^2 \mathbb{1}(|u_0 \geq r|)], \quad (\text{B.3})
\end{aligned}
$$

where the second inequality is by $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i)\|_2 \leq G$ Lemma A.1. We further bound the first term on the right-hand side of (B.3),

$$
\begin{aligned}
\mathbb{E}_{\pi^*}\big[&\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i) \cdot \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big|\big] \\
&\leq \mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) \cdot \big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_j) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j)\big)\big|\big] \\
&\quad + \mathbb{E}_{\pi^*}\big[\big|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_j) \cdot \big(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, x_i) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, x_i)\big)\big|\big] \\
&\leq 2G^2 \mathcal{W}_2(p, p_0),
\end{aligned}
\tag{B.4}
$$

where the last inequality is by $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \leq G$ and $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x})\|_2 \leq G\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2$ in Lemma A.1. Plugging (B.4) into (B.3) yields

$$
I_2 \leq 2r^2 G^2 \mathcal{W}_2(p, p_0) + 2G^2 \mathbb{E}_{\pi^*}[u_0^2 \mathbb{1}(|u_0 \geq r|)].
\tag{B.5}
$$

Further plugging (B.2) and (B.5) into (B.1), we obtain

$$
\begin{aligned}
\big|\mathbf{H}_1(p)_{i,j} - \mathbf{H}_1(p_0)_{i,j}\big| &\leq G^2 \mathcal{W}_2(p, p_0)\sqrt{8d + 10} + 2r^2 G^2 \mathcal{W}_2(p, p_0) \\
&\quad + 2G^2 \mathbb{E}_{p_0}[u_0^2 \mathbb{1}(|u_0 \geq r|)].
\end{aligned}
$$

This finishes the proof. $\square$

B.3   Proof of Lemma A.3

Here we provide the proof of Lemma A.3, which is essentially based on a direct application of Lemma A.1 and the definition of 2-Wasserstein distance.

Proof of Lemma A.3. Denote $\widehat{H}_{i,j}(\boldsymbol{\theta}, u) = h(\boldsymbol{\theta}, \mathbf{x}_i) h(\boldsymbol{\theta}, \mathbf{x}_j)$, then we have $\mathbf{H}_2(p)_{i,j} = \mathbb{E}_p[\widehat{H}_{i,j}(\boldsymbol{\theta}, u)]$. Calculating the gradient of $\widehat{H}_{i,j}(\boldsymbol{\theta}, u)$, we have

$$
\nabla_u \widehat{H}_{i,j}(\boldsymbol{\theta}, u) = 0, \qquad \|\nabla_{\boldsymbol{\theta}} \widehat{H}_{i,j}(\boldsymbol{\theta}, u)\|_2 \leq 2\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}_i)\|_2 |h(\boldsymbol{\theta}, \mathbf{x}_j)| \leq 2G^2,
$$

where the second inequality is by Lemma A.1 . Applying Lemma B.2 gives

$$|\mathbf{H}_2(p)_{i,j} - \mathbf{H}_2(p_0)_{i,j}| \leq 2G^2 \mathcal{W}_2(p, p_0).$$

This finializes our proof. □

B.4 Proof of Lemma A.4

Lemma A.4 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 A.4. By the definition of p_0 we have

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

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

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

which finalizes our proof. □

B.5 Proof of Lemma A.5

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.

$$\frac{\partial \mathbf{f}(t)_i}{\partial p_t} := \alpha u h(\boldsymbol{\theta}, \mathbf{x}_i), \tag{B.6}$$

$$\frac{\partial L(p_t)}{\partial p_t} := \mathbb{E}_S\big[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big) \cdot \alpha u h(\boldsymbol{\theta}, \mathbf{x})\big], \tag{B.7}$$

$$\frac{\partial D_{\mathrm{KL}}(p_t || p_0)}{\partial p_t} := \log(p_t/p_0) + 1, \tag{B.8}$$

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

$$= \mathbb{E}_S\big[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big) \cdot \alpha u h(\boldsymbol{\theta}, \mathbf{x}) + \lambda \log(p_t/p_0) + \lambda\big]. \tag{B.9}$$

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 C.3.

Lemma B.3. 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 (B.6), (B.7),

(B.8) and (B.9). 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 (B.7), (B.8) and (B.9).

Lemma B.4. 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 (B.7), (B.8) and (B.9). Then their gradients with respect to u and $\boldsymbol{\theta}$ are given as follows:

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

Moreover, the PDE (3.4) can be written as

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

Proof of Lemma A.5. By Lemma B.3, we have the following chain rule

$$
\begin{aligned}
\frac{\partial L(p_t)}{\partial t} &= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t} \frac{dp_t}{dt} d\boldsymbol{\theta} du \\
&= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t} \nabla \cdot \left[p_t(\boldsymbol{\theta}, u) \nabla \frac{\partial Q(p_t)}{\partial p_t} \right] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \left[\nabla \frac{\partial L(p_t)}{\partial p_t} \right] \cdot \left[\nabla \frac{\partial Q(p_t)}{\partial p_t} \right] d\boldsymbol{\theta} du \\
&= \underbrace{-\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \left[\nabla \frac{\partial L(p_t)}{\partial p_t} \right] \cdot \left[\nabla \frac{\partial L(p_t)}{\partial p_t} \right] d\boldsymbol{\theta} du}_{I_1} \\
&\quad \underbrace{- \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \left[\nabla \frac{\partial L(p_t)}{\partial p_t} \right] \cdot \left[\nabla \frac{\partial D(p_t||p_0)}{\partial p_t} \right] d\boldsymbol{\theta} du}_{I_2}, \quad \text{(B.10)}
\end{aligned}
$$

where the second and last equation is by Lemma B.4, 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

22

. For I_1, we have

$$I_1 = \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \|\widehat{g}_1(t, \boldsymbol{\theta}, u)\|_2^2 d\boldsymbol{\theta} du + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) |\widehat{g}_2(t, \boldsymbol{\theta}, u)|^2 d\boldsymbol{\theta} d\boldsymbol{\theta} du. \tag{B.11}$$

Similarly, for I_2, we have

$$\begin{aligned}
I_2 &= \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[-\widehat{g}_1(t, \boldsymbol{\theta}, u)] \cdot [u + \nabla_u \log(p_t)] d\boldsymbol{\theta} du \\
&\quad + \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[-\widehat{g}_2(t, \boldsymbol{\theta}, u)] \cdot [\boldsymbol{\theta} + \nabla_{\boldsymbol{\theta}} \log(p_t)] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u + \widehat{g}_2(t, \boldsymbol{\theta}, u)\boldsymbol{\theta}] d\boldsymbol{\theta} du \\
&\quad - \int_{\mathbb{R}^{d+1}} [\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot \nabla_u p_t(t, \boldsymbol{\theta}, u) + \widehat{g}_2(t, \boldsymbol{\theta}, u) \cdot \nabla_{\boldsymbol{\theta}} p_t(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du \\
&= -\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u + \widehat{g}_2(t, \boldsymbol{\theta}, u)\boldsymbol{\theta}] d\boldsymbol{\theta} du \\
&\quad + \int_{\mathbb{R}^{d+1}} p_t(t, \boldsymbol{\theta}, u)[\nabla_u \cdot \widehat{g}_1(t, \boldsymbol{\theta}, u) + \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du, \tag{B.12}
\end{aligned}$$

where the second equation is by $p_t \nabla \log(p_t) = \nabla p_t$ and the third equation is by applying integration by parts. Plugging (B.11) and (B.12) into (B.10), we get

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

This completes the proof. $\qquad\square$

B.6 Proof of Lemma A.6

Here we prove Lemma A.6, which is based on its connection to the Gram matrix of neural tangent kernel.

Proof of Lemma A.6. We first remind the readers of the definitions of the Gram matrices in (3.6). 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), we have

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

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

Therefore by (3.6) we have

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

$$= \frac{1}{n^2} \mathbf{b}(p_t)^\top \mathbf{H}(p_t) \mathbf{b}(p_t). \tag{B.13}$$

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

$$\frac{1}{n^2} \mathbf{b}(p_t)^\top \mathbf{H}(p_t) \mathbf{b}(p_t) \geq \frac{\Lambda \|\mathbf{b}(p_t)\|_2^2}{2n^2} = \frac{\lambda_0^2}{2} L(p_t), \tag{B.14}$$

where the equation follows by the definition of $\mathbf{b}(p_t)$. Plugging (B.14) into (B.13) completes the proof. $\square$

B.7 Proof of Lemma A.7

Lemma B.5. Under Assumptions 4.1 and 4.2, for all $\mathcal{W}(p, p_0) \leq \sqrt{d+1}$ and $\mathbf{x}$ the following inequality holds.

$$\big| \mathbb{E}_p \big[uh(\boldsymbol{\theta}, \mathbf{x}) + u\nabla h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} - u\Delta h(\boldsymbol{\theta}, \mathbf{x}) \big] \big| \leq A_1,$$

where A_1 is defined in Theorem 4.4.

The proof of Lemma A.7 is based on direct applications of Lemma B.5. We present the proof as follows.

Proof of Lemma A.7. We have the following identities:

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

Base on these identities we can derive

$$\left| \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u + \widehat{g}_2 \cdot \boldsymbol{\theta} - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du \right|$$

$$= \left| \alpha \mathbb{E}_S \left[\nabla_f \phi(f(p_t, \mathbf{x}), y) \mathbb{E}_{p_t} \left[\big(u_t h(\boldsymbol{\theta}_t, \mathbf{x}_i) + u_t \nabla h(\boldsymbol{\theta}_t, \mathbf{x}_i) \cdot \boldsymbol{\theta}_t - u_t \Delta h(\boldsymbol{\theta}_t, \mathbf{x}_i) \big) \right] \right] \right|$$

$$\leq 2\alpha A_1 \mathbb{E}_S[|f(p_t, \mathbf{x}) - y|]$$

$$\leq 2\alpha A_1 \sqrt{L(p_t)},$$

where the first inequality is by Lemma B.5, the second inequality is by Jensen's inequality. $\square$

B.8 Proof of Lemma A.8

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

Lemma B.6. Let p_t be the solution of PDE (3.4). Then the following identity holds.

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

In the calculation given by Lemma B.6, 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.5, and a bound of it has already been given in Lemma A.7. However, for the analysis of the KL-divergence term, we present the following new bound, which eventually leads to a sharper result.

Lemma B.7. Under Assumptions 4.1 and 4.2, let A_2 be defined in Theorem 4.4. Then for $t \le t^*$, it holds that

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

Proof of Lemma A.8. By Lemma B.6,

$$\frac{\partial D_{\mathrm{KL}}(p_t||p_0)}{\partial t} = -\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\boldsymbol{\theta} + \nabla_{\boldsymbol{\theta}} \log(p_t)\|_2^2 - \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|u + \nabla_u \log(p_t)|^2$$
$$+ \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u + \widehat{g}_2 \cdot \boldsymbol{\theta} - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2]d\boldsymbol{\theta}du$$
$$\le 2A_2\alpha\sqrt{L(p_t)}\sqrt{D_{\mathrm{KL}}(p_t||p_0)}, \tag{B.15}$$

where the inequality is by Lemma B.7. Notice that $\sqrt{D_{\mathrm{KL}}(p_0||p_0)} = 0$, $\sqrt{D_{\mathrm{KL}}(p_t||p_0)}$ is differentiable at $\sqrt{D_{\mathrm{KL}}(p_t||p_0)} \ne 0$ and from (B.15) the derivative

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

which implies

$$\sqrt{D_{\mathrm{KL}}(p_t||p_0)} \le \int_0^t A_2\alpha\sqrt{L(p_s)}ds$$
$$\le A_2\alpha \int_0^t \exp(-\alpha^2\lambda_0^2 s) + A_1\lambda\alpha^{-1}\lambda_0^{-2}ds$$
$$\le A_2\alpha^{-1}\lambda_0^{-2} + A_2A_1\lambda\lambda_0^{-2}t,$$

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

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

This completes the proof. □

B.9 Proof of Lemma A.9

Proof of Lemma A.9. By Lemma B.3, we get

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

where the third equation is by applying integration by parts and the fourth equation is by Lemma B.4. $\qquad\square$

C Proof of Auxiliary Lemmas in Appendix B

C.1 Proof of Lemma B.1

Proof of Lemma B.1. Let $\pi^*(p_0, p)$ be the coupling that achieves the 2-Wasserstein distance between p_0 and p. Then by definition,

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

where the last inequality is by the assumption that $\mathcal{W}_2(p, p_0) \leq \sqrt{d+1}$. This finishes the proof. $\quad\square$

C.2 Proof of Lemma B.2

Proof of Lemma B.2. By Lemma C.8 in Xu et al. (2018), we have that

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

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 B.1, we get $\sigma \leq 2\sqrt{d+1}$. Substituting the upper bound of σ into the above inequality completes the proof. $\quad\square$

C.3 Proof of Lemma B.3

Proof of Lemma B.3. By chain rule and the definition of $\mathbf{f}(t)$, we have

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

$$= \int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta}, \mathbf{x}_i)\frac{dp_t}{dt}(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du$$

$$= \int_{\mathbb{R}^{d+1}} \frac{\partial \mathbf{f}(t)_i}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du,$$

where the last equation follows by the definition of the first variation $\frac{\partial \mathbf{f}(t)_i}{\partial p_t}$. This proves the first identity. Now we bound the second identity,

$$\frac{\partial L(p_t)}{\partial t} = \mathbb{E}_S\left[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big)\frac{d}{dt}f(p_t, \mathbf{x})\right]$$

$$= \mathbb{E}_S\left[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big)\frac{d}{dt}\int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta}, \mathbf{x}) p_t(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du\right]$$

$$= \mathbb{E}_S\left[\nabla_{y'}\phi\big(f(p_t, \mathbf{x}), y\big)\int_{\mathbb{R}^{d+1}} \alpha u h(\boldsymbol{\theta}, \mathbf{x})\frac{dp_t(\boldsymbol{\theta}, u)}{dt} d\boldsymbol{\theta} du\right]$$

$$= \int_{\mathbb{R}^{d+1}} \frac{\partial L(p_t)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du,$$

where the last equation 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

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

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

$$= \int_{\mathbb{R}^{d+1}} \frac{\partial Q(p_t)}{\partial p_t}\frac{dp_t}{dt} d\boldsymbol{\theta} du,$$

where the last equation 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. $\square$

C.4 Proof of Lemma B.4

Proof of Lemma B.4. By Lemma B.3,we have

$$\nabla_u \frac{\partial L}{\partial p_t} = \nabla_u \mathbb{E}_S \big[\nabla_{y'} \phi \big(f(p_t, \mathbf{x}), y \big) \alpha u h(\boldsymbol{\theta}, \mathbf{x}) \big] = -\widehat{g}_1(t, \boldsymbol{\theta}, u),$$

$$\nabla_{\boldsymbol{\theta}} \frac{\partial L}{\partial p_t} = \nabla_{\boldsymbol{\theta}} \mathbb{E}_S \big[\nabla_{y'} \phi \big(f(p_t, \mathbf{x}), y \big) \alpha u h(\boldsymbol{\theta}, \mathbf{x}) \big] = -\widehat{g}_2(t, \boldsymbol{\theta}, u),$$

$$\nabla_u \frac{\partial D_{kL}(p_t||p_0)}{\partial p_t} = \nabla_u (\log(p_t/p_0) + 1) = u + \nabla_u \log(p_t),$$

$$\nabla_{\boldsymbol{\theta}} \frac{\partial D_{kL}(p_t||p_0)}{\partial p_t} = \nabla_{\boldsymbol{\theta}} (\log(p_t/p_0) + 1) = \boldsymbol{\theta} + \nabla_{\boldsymbol{\theta}} \log(p_t).$$

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

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

where the third equation is by the definition $g_1(t, \boldsymbol{\theta}, u) = \widehat{g}_1(t, \boldsymbol{\theta}, u) - \lambda u$, $g_2(t, \boldsymbol{\theta}, u) = \widehat{g}_2(t, \boldsymbol{\theta}, u) - \lambda \boldsymbol{\theta}$ and $p_t \nabla \log(p_t) = \nabla p_t$. $\qquad \square$

C.5 Proof of Lemma B.5

Here we give the proof of Lemma B.5.

Proof of Lemma B.5. The proof is based on the smoothness properties of $h(\boldsymbol{\theta}, \mathbf{x})$ given in Lemma A.1. We have

$$\begin{aligned}
\big| \mathbb{E}_p \big[\big(u h(\boldsymbol{\theta}, \mathbf{x}) + u \nabla h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} - u \Delta h(\boldsymbol{\theta}, \mathbf{x}) \big) \big] \big| \\
\leq \mathbb{E}_p \big[|u| G + G |u| \|\boldsymbol{\theta}\|_2 + G |u| \big] \\
= G \mathbb{E}_p [|u| \|\boldsymbol{\theta}\|_2] + 2 G \mathbb{E}_p [|u|] \\
\leq G \mathbb{E}_p \left[\frac{u^2 + \|\boldsymbol{\theta}\|_2^2}{2} \right] + 2 G \sqrt{\mathbb{E}_p [u^2]},
\end{aligned}$$

where the first inequality is by $|h(\boldsymbol{\theta}, \mathbf{x})| \leq G$, $\|\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x})\|_2 \leq G$ and $|\Delta h(\boldsymbol{\theta}, \mathbf{x})| \leq G$ in Lemma A.1, the second inequality is by Young's inequality and Cauchy-Schwartz inequality. Now by

$\mathcal{W}(p, p_0) \leq \sqrt{d+1}$ and Lemma B.1, we have

$$\begin{aligned}
\Big|\mathbb{E}_p\Big[\big(uh(\boldsymbol{\theta}, \mathbf{x}) + u\nabla h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} - u\Delta h(\boldsymbol{\theta}, \mathbf{x})\big)\Big]\Big| & \\
\leq 2G(d+1) + 4G\sqrt{d+1} & \\
= A_1. &
\end{aligned}$$

This completes proof. $\qquad\qquad\square$

C.6   Proof of Lemma B.6

Proof of Lemma B.6. By Lemma B.3, we have

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

where the second and last equations are by Lemma B.4, the third inequality is by applying integration by parts multiple times. We further calculate by Lemma B.4,

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

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

$$
\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u) \left[\nabla \frac{\partial D_{\mathrm{KL}}(p_t \| p_0)}{\partial p_t} \right] \cdot \left[\nabla \frac{\partial L(p_t)}{\partial p_t} \right] d\boldsymbol{\theta} du
$$

$$
= \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[-\widehat{g}_1(t, \boldsymbol{\theta}, u)] \cdot [u + \nabla_u \log(p_t)] d\boldsymbol{\theta} du
$$

$$
+ \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[-\widehat{g}_2(t, \boldsymbol{\theta}, u)] \cdot [\boldsymbol{\theta} + \nabla_{\boldsymbol{\theta}} \log(p_t)] d\boldsymbol{\theta} du
$$

$$
= - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u + \widehat{g}_2(t, \boldsymbol{\theta}, u)\boldsymbol{\theta}] d\boldsymbol{\theta} du
$$

$$
- \int_{\mathbb{R}^{d+1}} [\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot \nabla_u p_t(t, \boldsymbol{\theta}, u) + \widehat{g}_2(t, \boldsymbol{\theta}, u) \cdot \nabla_{\boldsymbol{\theta}} p_t(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du
$$

$$
= - \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1(t, \boldsymbol{\theta}, u) \cdot u + \widehat{g}_2(t, \boldsymbol{\theta}, u)\boldsymbol{\theta}] d\boldsymbol{\theta} du
$$

$$
+ \int_{\mathbb{R}^{d+1}} p_t(t, \boldsymbol{\theta}, u)[\nabla_u \cdot \widehat{g}_1(t, \boldsymbol{\theta}, u) + \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2(t, \boldsymbol{\theta}, u)] d\boldsymbol{\theta} du, \tag{C.3}
$$

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

$$
\frac{\partial D_{\mathrm{KL}}(p_t \| p_0)}{\partial t} = -\lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)\|\boldsymbol{\theta} + \nabla_{\boldsymbol{\theta}} \log(p_t)\|_2^2 - \lambda \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)|u + \nabla_u \log(p_t)|^2
$$

$$
+ \int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u + \widehat{g}_2 \cdot \boldsymbol{\theta} - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du.
$$

This completes the proof. $\qquad\square$

C.7 Proof of Lemma B.7

Proof of Lemma B.7. We remind the readers the definitions of $\widehat{g}_1$ and $\widehat{g}_2$ in (A.1) and (A.1). We have

$$
\int_{\mathbb{R}^{d+1}} p_t(\boldsymbol{\theta}, u)[\widehat{g}_1 \cdot u + \widehat{g}_2 \cdot \boldsymbol{\theta} - \nabla_u \cdot \widehat{g}_1 - \nabla_{\boldsymbol{\theta}} \cdot \widehat{g}_2] d\boldsymbol{\theta} du
$$

$$
= 2\alpha \mathbb{E}_S \left[(f(p_t, \mathbf{x}) - y) \int_{\mathbb{R}^{d+1}} \left(uh(\boldsymbol{\theta}, \mathbf{x}) + u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} - u\Delta h(\boldsymbol{\theta}, \mathbf{x}) \right) p_t(\boldsymbol{\theta}, u) d\boldsymbol{\theta} du \right].
$$

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

$$
|\nabla_u I(\boldsymbol{\theta}, u, \mathbf{x})| = |h(\boldsymbol{\theta}, \mathbf{x}) + \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} - \Delta h(\boldsymbol{\theta}, \mathbf{x})| \le G\|\boldsymbol{\theta}\|_2 + 2G, \tag{C.4}
$$

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

$$
\|\nabla_{\boldsymbol{\theta}} I(\boldsymbol{\theta}, u, \mathbf{x})\|_2 = \|u\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) + u\nabla_{\boldsymbol{\theta}} \left(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}) \cdot \boldsymbol{\theta} \right) - u\nabla_{\boldsymbol{\theta}} \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{x}))\|_2
$$

$$
\le 3G|u|. \tag{C.5}
$$

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

$$\sqrt{\nabla_u I(\boldsymbol{\theta}, u, \mathbf{x})^2 + \|\nabla_{\boldsymbol{\theta}} I(\boldsymbol{\theta}, u, \mathbf{x})\|_2^2} \le 4G\sqrt{u^2 + \|\boldsymbol{\theta}\|_2^2} + 2G.$$

By Lemma B.2, we have that

$$\mathbb{E}_{p_t}[I(\boldsymbol{\theta}_t, u_t, \mathbf{x})] - \mathbb{E}_{p_0}[I(\boldsymbol{\theta}_0, u_0, \mathbf{x})] \le \left[8G\sqrt{d+1} + 2G\right]\mathcal{W}(p_0, p_t)$$
$$\le A_2\sqrt{D_{\mathrm{KL}}(p_t||p_0)},$$

where the last inequality is by Lemma 5.4 and $A_2 = 16G\sqrt{d+1} + 4G$. By $\mathbb{E}_{p_0}[I(\boldsymbol{\theta}_0, u_0, \mathbf{x})] = \mathbb{E}_{p_0}[u_0]\mathbb{E}_{p_0}[h(\boldsymbol{\theta}_0, \mathbf{x}) + \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x}) \cdot \boldsymbol{\theta}_0 - \Delta_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_0, \mathbf{x})] = 0$, 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

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

where the last inequality is by (C.6) and Cauchy-Schwarz inequality. This completes the proof. $\qquad\square$

References

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*.

BACH, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research* **18** 629–681.

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.

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. (2019). Generalization bounds of stochastic gradient descent for wide and deep 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*.

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., 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* .

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*.

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

LEE, J., XIAO, L., SCHOENHOLZ, S. S., BAHRI, Y., SOHL-DICKSTEIN, J. and PENNINGTON, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*.

LI, X., LU, J., WANG, Z., HAUPT, J. and ZHAO, T. (2018). On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond. *arXiv preprint arXiv:1806.05159* .

LI, Y. and LIANG, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*.

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*.

LIU, C., ZHU, L. and BELKIN, M. (2020). Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307* .

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.

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.

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.

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

TZEN, B. and RAGINSKY, M. (2020). A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled mckean-vlasov dynamics. *arXiv preprint arXiv:2002.01987* .

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*.

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*.