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

Zixiang Chen∗ Yuan Cao∗ Quanquan Gu∗ Tong Zhang†
∗Department of Computer Science, UCLA
†Department of Computer Science and Mathematics, HKUST
{chenzx19,yuancao,qgu}@cs.ucla.edu tongzhang@tongzhang-ml.org

Abstract

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

1 Introduction

Deep learning has achieved tremendous practical success in a wide range of machine learning tasks. However, due to the nonconvex and over-parameterized nature of modern neural networks, the success of deep learning cannot be fully explained by conventional learning theory. Some recent results have established theoretical guarantees for deep learning based on two frameworks: mean-field analysis and neural tangent kernel (NTK) [1] based analysis. These two frameworks are known to have their own advantages and disadvantages.

In this work, we aim to propose a unified framework connecting the mean-field and NTK approaches. We summarize the contributions of this paper as follows:

• We establish a comprehensive connection between NTK and mean-field analyses, and demonstrate that with appropriate scaling, the whole neural network training process can be similar to the dynamics of neural tangent kernel. Our result improves existing result in Mei et al. [2], which only shows the closeness between the two dynamics for a limited time period $t \in [0, T]$ for some finite $T$. In comparison, we provide a uniform bound over $t \in [0, +\infty)$. A direct consequence of our analysis is a novel linear convergence guarantee of noisy gradient descent up to certain accuracy.

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

• We establish generalization bounds for neural networks trained with noisy gradient descent with weight decay regularization under different scalings. When the scaling factor is large, we show that infinitely wide neural networks can learn a class of functions defined based on a bounded

Preprint. Under review.
\(\chi^2\)-divergence to initialization distribution. When the scaling factor is small, we also give a comparable result showing that a function class defined with KL-divergence can be learnt.

2 Main Results

We introduce a scaling factor \(\alpha > 0\) and study two-layer, infinitely wide neural networks of the form

\[
f(p, x) = \alpha \int_{\mathbb{R}^d} u h(\theta, x)p(\theta, u) d\theta du,
\]

where \(x \in \mathbb{R}^d\) is the input, \(\theta \in \mathbb{R}^d\) and \(u \in \mathbb{R}\) are the network parameters, \(p(\theta, u)\) is their joint distribution, and \(h(\theta, x)\) is the activation function. Given a training set \(\{(x_i, y_i)\}_{i=1}^n\), we let \(E_S[\cdot]\) be the average over the training sample \(S\), \(\phi(y', y) = (y' - y)^2\) be the square loss function, and consider the training dynamic defined by the following partial differential equation, which is the continuous-time, infinite-width limit of noisy gradient descent algorithm with weight decay regularization:

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

where \(\lambda\) is the regularization parameter, and

\[
g_1(t, \theta, u) = -\alpha E_S[\nabla_y \phi(f(p_t, x), y)h(\theta, x)] - \lambda u,
\]

\[
g_2(t, \theta, u) = -\alpha E_S[\nabla_y \phi(f(p_t, x), y)u \nabla_\theta h(\theta, x)] - \lambda \theta.
\]

We consider the standard Gaussian initialization \(p_0\), define \(L(p) = E_S[\phi(f(p, x), y)]\), and let

\[
H(p) = E_p[u^2(\nabla_\theta h(\theta, x_i), \nabla_\theta h(\theta, x_j) + h(\theta, x_i)h(\theta, x_j))]
\]

be the Gram matrix of neural tangent kernel defined with parameter distribution \(p\). We have the following optimization and generalization results.

**Theorem 2.1** (informal). Suppose that \(h(\theta, x)\) satisfies certain smoothness properties. Let \(\Lambda = \lambda_{\min}(H(p_0)) > 0\), and \(\lambda_0 = \sqrt{\Lambda/n}\). If \(\alpha \geq \text{poly}(\lambda_0^{-1})\), then for all \(t \in [0, \infty)\),

\[
L(p_t) \leq 2\exp(-2\alpha^2\lambda_0^2t) L(p_0) + O(\alpha^{-2}\lambda_0^{-4}), \quad D_{\text{KL}}(p_t\|p_0) \leq O(\alpha^{-2}\lambda_0^{-4}).
\]

Moreover, define \(f(t) = (f(p_t, x_1), \ldots, f(p_t, x_n))^\top\), and let \(f_{\text{NTK}}(t)\) be its counterpart given by NTK-based kernel regression. Then

\[
\|H(p_t) - H(p_0)\|_{\infty, \infty} \leq \tilde{O}(\lambda_0^{-2}\alpha^{-1}), \quad \frac{1}{n}\|f(t) - f_{\text{NTK}}(t)\|_2^2 \leq \tilde{O}(\lambda_0^{-8}\alpha^{-2}).
\]

**Theorem 2.2.** Let \(y_0, y\) = 1 if \(y, y < 0\). Suppose that \(h(\theta, x)\) satisfies certain smoothness properties, and

\[
y = \int u h(\theta, x)p_{\text{true}}(\theta, u) d\theta du.
\]

Then \(p_t\) weakly converges to a limiting distribution \(p^*\) that satisfies:

- If \(\alpha \geq O\left(\sqrt{nD_{\chi^2}(p_{\text{true}}\|p_0)}\right)\), then with high probability,

\[
\tilde{E}_D[\ell^{0-1}(f(p^*, x), y)] \leq \tilde{O}\left(\sqrt{\frac{D_{\chi^2}(p_{\text{true}}\|p_0)}{n}}\right).
\]

- If \(\alpha = O(1)\), \(h(\theta, x)\) is bounded and \(\lambda\) is chosen appropriately small, then with high probability,

\[
\tilde{E}_D[\ell^{0-1}(f(p^*, x), y)] \leq \tilde{O}\left(\sqrt{\frac{D_{\text{KL}}(p_{\text{true}}\|p_0)}{n}}\right).
\]

References
