ETH zürich

The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks

Aurelien Lucchi

1. Motivation

Motivation: Frequency Bias

Contributions

- Early-time learning dynamics of a two-layer fully-connected neural network can be mimicked by training a simple linear model on the inputs

Recap of Neural Tangent Kernel (NTK)

- Consider a single-output neural network $f(\boldsymbol{x};\boldsymbol{\theta})$ where $\boldsymbol{x}$ is the input and $\boldsymbol{\theta}$ is the parameters of the network.
- Around a reference network with parameters $\bar{\boldsymbol{\theta}}$, we can do a local first-order approximation:

$$f(\boldsymbol{x};\boldsymbol{\theta}) \approx f(\boldsymbol{x};\bar{\boldsymbol{\theta}}) + \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} - \bar{\boldsymbol{\theta}} \rangle.$$

Recap of Neural Tangent Kernel (NTK)

- Gradient feature map $\boldsymbol{x} \mapsto \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \bar{\boldsymbol{\theta}})$ induces a kernel $K_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{x}') := \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \bar{\boldsymbol{\theta}}), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}'; \bar{\boldsymbol{\theta}}) \rangle$ which is called the NTK
- Gradient descent training of the neural network can be viewed as kernel gradient descent on the function space with respect to the NTK.
- Use *NTK matrix* to refer to an $n \times n$ matrix that is the NTK evaluated on n datapoints.

2. Setup

Two-layer network

- Consider a two-layer fully-connected neural network with m hidden neurons defined as:

$$f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{v}) := \frac{1}{\sqrt{m}} \sum_{r=1}^{m} v_r \phi\left(\boldsymbol{w}_r^\top \boldsymbol{x}/\sqrt{d}\right) = \frac{1}{\sqrt{m}} \boldsymbol{v}^\top \phi\left(\boldsymbol{W}\boldsymbol{x}/\sqrt{d}\right),$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input, $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_m]^\top \in \mathbb{R}^{m \times d}$ is the weight matrix in the first layer, and $\boldsymbol{v} = [v_1, \dots, v_m]^\top \in \mathbb{R}^m$ is the weight vector in the second layer.

Two-layer network

- Consider a two-layer fully-connected neural network with m hidden neurons defined as:

$$f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{v}) := \frac{1}{\sqrt{m}} \sum_{r=1}^{m} v_r \phi\left(\boldsymbol{w}_r^\top \boldsymbol{x}/\sqrt{d}\right) = \frac{1}{\sqrt{m}} \boldsymbol{v}^\top \phi\left(\boldsymbol{W}\boldsymbol{x}/\sqrt{d}\right),$$

  where $\boldsymbol{x} \in \mathbb{R}^d$ is the input, $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m]^\top \in \mathbb{R}^{m \times d}$ is the weight matrix in the first layer, and $\boldsymbol{v} = [v_1, \ldots, v_m]^\top \in \mathbb{R}^m$ is the weight vector in the second layer.
- Consider the following ℓ_2 training loss:

$$L(\boldsymbol{W}, \boldsymbol{v}) := \frac{1}{2n} \sum_{i=1}^{n} \left(f(\boldsymbol{x}_i; \boldsymbol{W}, \boldsymbol{v}) - y_i\right)^2,$$

- Use *symmetric initialization* for the weights $(\boldsymbol{W}, \boldsymbol{v})$:

Gradient descent

- Let $(\boldsymbol{W}(0), \boldsymbol{v}(0))$ be a set of initial weights drawn from the symmetric initialization. Then the weights are updated according to GD:

$$\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta_1 \nabla_{\boldsymbol{W}} L\left(\boldsymbol{W}(t), \boldsymbol{v}(t)\right),$$
$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) - \eta_2 \nabla_{\boldsymbol{v}} L\left(\boldsymbol{W}(t), \boldsymbol{v}(t)\right)$$

  where η_1 and η_2 are the learning rates. Here we allow potentially different learning rates for flexibility.

Assumptions

- The datapoints $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. samples from a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$ such that $\mathrm{Tr}[\boldsymbol{\Sigma}] = d$ and $\|\boldsymbol{\Sigma}\| = O(1)$.
- The activation function $\phi(\cdot)$ satisfies either of the followings:
  - (i) smooth activation: ϕ has bounded first and second derivatives: $|\phi'(z)| = O(1)$ and $|\phi''(z)| = O(1)$ $(\forall z \in \mathbb{R})$, or
  - (ii) piece-wise linear activation: $\phi(z) = \begin{cases} z & (z \geq 0) \\ az & (z < 0) \end{cases}$ for some $a \in \mathbb{R}, |a| = O(1)$.[1]

[1] We define $\phi'(0) = 1$ in this case.

Claim

Under previous Assumptions, the datapoints satisfy the following concentration properties:

Claim

Suppose $n \gg d$. With high probability we have $\frac{\|\boldsymbol{x}_i\|^2}{d} = 1 \pm O\left(\sqrt{\frac{\log n}{d}}\right)$ $(\forall i \in [n])$, $\frac{|\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle|}{d} = O\left(\sqrt{\frac{\log n}{d}}\right)$ $(\forall i, j \in [n], i \neq j)$, and $\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\| = \Theta(n)$.

3. Training the First Layer

Linear model

- Denote by $f_t^1 : \mathbb{R}^d \to \mathbb{R}$ the network at iteration t in this case, namely $f_t^1(\boldsymbol{x}) := f(\boldsymbol{x}; \boldsymbol{W}(t), \boldsymbol{v}(t)) = f(\boldsymbol{x}; \boldsymbol{W}(t), \boldsymbol{v}(0))$ (note that $\boldsymbol{v}(t) = \boldsymbol{v}(0)$).
- The linear model which will be proved to approximate the neural network f_t^1 in the early phase of training is $f^{\mathrm{lin1}}(\boldsymbol{x}; \boldsymbol{\beta}) := \boldsymbol{\beta}^\top \boldsymbol{\psi}_1(\boldsymbol{x})$, where

$$\boldsymbol{\psi}_1(\boldsymbol{x}) := \frac{1}{\sqrt{d}} \begin{bmatrix} \zeta \boldsymbol{x} \\ \nu \end{bmatrix}, \qquad \text{with } \zeta = \mathbb{E}[\phi'(g)]$$

$$\text{and } \nu = \mathbb{E}[g\phi'(g)] \cdot \sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}^2]/d}.$$

Main theorem for training the first layer

Main theorem for training the first layer - part 1

Let $\alpha \in (0, \frac{1}{4})$ be a fixed constant. Suppose the number of training samples n and the network width m satisfy $n \gtrsim d^{1+\alpha}$ and $m \gtrsim d^{1+\alpha}$. Suppose $\eta_1 \ll d$ and $\eta_2 = 0$. Then there exists a universal constant $c > 0$ such that with high probability, for all $0 \leq t \leq T = c \cdot \frac{d \log d}{\eta_1}$ simultaneously, the learned neural network f_t^1 and the linear model $f_t^{\mathrm{lin}1}$ at iteration t are close on average on the training data:

$$\frac{1}{n} \sum_{i=1}^{n} \left(f_t^1(\boldsymbol{x}_i) - f_t^{\mathrm{lin}1}(\boldsymbol{x}_i) \right)^2 \lesssim d^{-\Omega(\alpha)}. \tag{1}$$

Main theorem for training the first layer

Main theorem for training the first layer - part 2

Moreover, f_t^1 and $f_t^{\mathrm{lin}1}$ are also close on the underlying data distribution $\mathcal{D}$. Namely, with high probability, for all $0 \leq t \leq T$ simultaneously, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\min\{(f_t^1(\boldsymbol{x}) - f_t^{\mathrm{lin}1}(\boldsymbol{x}))^2, 1\}\right] \lesssim d^{-\Omega(\alpha)} + \sqrt{\frac{\log T}{n}}. \tag{2}$$

Remarks

- Note that this does not mean that f_t^1 and $f_t^{\mathrm{lin}1}$ are the same on the entire space $\mathbb{R}^d$ – they might still differ significantly at low-density regions of $\mathcal{D}$.
- The width requirement is mild as it only requires the width m to be larger than $d^{1+\alpha}$ for some small constant α.
- Agreement guaranteed up to iteration $T = c \cdot \frac{d \log d}{\eta_1}$ (for some constant c).
- It turns out that for well-conditioned data, after T iterations, a near optimal linear model will have been reached.
  - This means that *the neural network in the early phase approximates a linear model all the way until the linear model converges to the optimum*.

Proof sketch

- NTK matrix $\mathbf{\Theta}_1(\boldsymbol{W}) \in \mathbb{R}^{n\times n}$ at first-layer weight matrix $\boldsymbol{W}$:

$$\mathbf{\Theta}_1(\boldsymbol{W}) := (\phi'(\boldsymbol{X}\boldsymbol{W}^\top/\sqrt{d})\phi'(\boldsymbol{X}\boldsymbol{W}^\top/\sqrt{d})^\top/m) \odot (\boldsymbol{X}\boldsymbol{X}^\top/d)$$

- Kernel matrix $\mathbf{\Theta}^{\mathrm{lin1}} \in \mathbb{R}^{n\times n}$ for the linear model:

$$\mathbf{\Theta}^{\mathrm{lin1}} := \boldsymbol{\psi}_{\mathbf{1}}\boldsymbol{\psi}_{\mathbf{1}}^\top = (\zeta^2\boldsymbol{X}\boldsymbol{X}^\top + \nu^2\mathbf{1}\mathbf{1}^\top)/d.$$

Proof sketch

Proposition - Distance kernels

With high probability over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, we have $\left\|\boldsymbol{\Theta}_1(\boldsymbol{W}(0)) - \boldsymbol{\Theta}^{\mathrm{lin}1}\right\| \lesssim \frac{n}{d^{1+\alpha}}$.

Proof idea: Matrix Bernstein + entrywize Taylor expansion of $\mathbb{E}_{\boldsymbol{W}(0)} \|\boldsymbol{\Theta}_1(\boldsymbol{W}(0))\|$.

Proof sketch

Proposition - Distance kernels

With high probability over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, we have $\left\|\boldsymbol{\Theta}_1(\boldsymbol{W}(0)) - \boldsymbol{\Theta}^{\text{lin1}}\right\| \lesssim \frac{n}{d^{1+\alpha}}$.

Proof idea: Matrix Bernstein + entrywize Taylor expansion of $\mathbb{E}_{\boldsymbol{W}(0)} \|\boldsymbol{\Theta}_1(\boldsymbol{W}(0))\|$.

To finish the proof, need to carefully track:

1. the prediction difference between f_t^1 and f_t^{lin1},
2. how much the weight matrix $\boldsymbol{W}$ move away from initialization
3. how much the NTK changes.

Training the Second Layer

Second layer

- Next we consider training the second layer weights $\boldsymbol{v}$
- Denote by $f_t^2 : \mathbb{R}^d \to \mathbb{R}$ the network at iteration t in this case.

Second layer

- Next we consider training the second layer weights $\boldsymbol{v}$
- Denote by $f_t^2 : \mathbb{R}^d \to \mathbb{R}$ the network at iteration t in this case.
- Will show that training the second layer is also close to training a simple linear model $f^{\mathrm{lin2}}(\boldsymbol{x}; \boldsymbol{\gamma}) := \boldsymbol{\gamma}^\top \boldsymbol{\psi}_2(\boldsymbol{x})$ in the early phase, where:

$$\boldsymbol{\psi}_2(\boldsymbol{x}) := \begin{bmatrix} \frac{1}{\sqrt{d}}\zeta \boldsymbol{x} \\ \frac{1}{\sqrt{2d}}\nu \\ \vartheta_0 + \vartheta_1(\frac{\|\boldsymbol{x}\|}{\sqrt{d}} - 1) + \vartheta_2(\frac{\|\boldsymbol{x}\|}{\sqrt{d}} - 1)^2 \end{bmatrix} \tag{3}$$

$$\begin{cases} \vartheta_0 = \mathbb{E}[\phi(g)], \\ \vartheta_1 = \mathbb{E}[g\phi'(g)], \\ \vartheta_2 = \mathbb{E}[(\frac{1}{2}g^3 - g)\phi'(g)]. \end{cases} \tag{4}$$

Second layer

- Note that strictly speaking $f^{\mathrm{lin}2}(\boldsymbol{x};\gamma)$ is not a linear model in $\boldsymbol{x}$ because the feature map $\psi_2(\boldsymbol{x})$ contains a nonlinear feature depending on $\|\boldsymbol{x}\|$ in its last coordinate.
- Using earlier claim, proof rely on the fact that the contribution of the non-linear term is small

Main theorem for training the second layer

Main theorem for training the second layer

Let $\alpha \in (0, \frac{1}{4})$ be a fixed constant. Suppose $n \gtrsim d^{1+\alpha}$ and $\begin{cases} m \gtrsim d^{1+\alpha}, & \text{if } \mathbb{E}[\phi(g)] = 0 \\ m \gtrsim d^{2+\alpha}, & \text{otherwise} \end{cases}$. Suppose $\begin{cases} \eta_2 \ll d / \log n, & \text{if } \mathbb{E}[\phi(g)] = 0 \\ \eta_2 \ll 1, & \text{otherwise} \end{cases}$ and $\eta_1 = 0$. Then there exists a universal constant $c > 0$ such that with high probability, for all $0 \leq t \leq T = c \cdot \frac{d \log d}{\eta_2}$ simultaneously, s.t.

$$\frac{1}{n} \sum_{i=1}^{n} \left(f_t^2(\mathbf{x}_i) - f_t^{\text{lin}2}(\mathbf{x}_i) \right)^2 \lesssim d^{-\Omega(\alpha)}$$

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\min\{ (f_t^2(\mathbf{x}) - f_t^{\text{lin}2}(\mathbf{x}))^2, 1 \} \right] \lesssim d^{-\Omega(\alpha)}.$$

Training both layers

- Finally consider the case where both layers are trained
- NTK for training both layers is simply the sum of the first-layer NTK and the second-layer NTK
- Corresponding linear model should have its kernel being the sum of the kernels for linear models
- Proof is similar to first two theorems.

General Result

Closeness between Two Dynamics

General Idea

General idea:

- Consider an objective function of the form:

$$F(\boldsymbol{\theta}) = \frac{1}{2n} \left\| \boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{y} \right\|^2 ,$$

- Consider another linear least squares problem:

$$G(\boldsymbol{\omega}) = \frac{1}{2n} \left\| \boldsymbol{\Phi}\boldsymbol{\omega} - \boldsymbol{y} \right\|^2 ,$$

- What's happening next? We will show that the two objectives are close...

Main Objective function

- Consider an objective function of the form:

$$F(\boldsymbol{\theta}) = \frac{1}{2n} \|\boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{y}\|^2 \,,$$

where $\boldsymbol{f} : \mathbb{R}^N \mapsto \mathbb{R}^n$ is a general differentiable function, and $\boldsymbol{y} \in \mathbb{R}^n$ satisfies $\|\boldsymbol{y}\| \leq \sqrt{n}$. We denote by $\boldsymbol{J} : \mathbb{R}^N \mapsto \mathbb{R}^{n \times N}$ the Jacobian map of $\boldsymbol{f}$. Then starting from some $\boldsymbol{\theta}(0) \in \mathbb{R}^N$, the GD updates for minimizing F can be written as:

$$\begin{aligned} \boldsymbol{\theta}(t+1) &= \boldsymbol{\theta}(t) - \eta \nabla F(\boldsymbol{\theta}(t)) \\ &= \boldsymbol{\theta}(t) - \frac{1}{n} \eta \boldsymbol{J}(\boldsymbol{\theta}(t))^\top (\boldsymbol{f}(\boldsymbol{\theta}(t)) - \boldsymbol{y}). \end{aligned}$$

Linear least squares problem

- Consider another linear least squares problem:

$$G(\boldsymbol{\omega}) = \frac{1}{2n} \left\| \boldsymbol{\Phi}\boldsymbol{\omega} - \boldsymbol{y} \right\|^2,$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{n \times M}$ is a fixed matrix. Its GD dynamics started from $\boldsymbol{\omega}(0) \in \mathbb{R}^M$ can be written as:

$$\boldsymbol{\omega}(t+1) = \boldsymbol{\omega}(t) - \eta \nabla G(\boldsymbol{\omega}(t)) = \boldsymbol{\omega}(t) - \frac{1}{n}\eta \boldsymbol{\Phi}^\top (\boldsymbol{\Phi}\boldsymbol{\omega}(t) - \boldsymbol{y}).$$

Let $\boldsymbol{K} := \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$, and let

$$\boldsymbol{u}(t) := \boldsymbol{f}(\boldsymbol{\theta}(t)),$$
$$\boldsymbol{u}^{\mathrm{lin}}(t) := \boldsymbol{\Phi}\boldsymbol{\omega}(t),$$

which stand for the predictions of these two models at iteration t.

Analytical form

The linear dynamics admit a very simple analytical form.

Claim C.1

For all $t \geq 0$ we have $\boldsymbol{u}^{\mathrm{lin}}(t) - \boldsymbol{y} = \left(\boldsymbol{I} - \frac{1}{n}\eta\boldsymbol{K}\right)^{t}(\boldsymbol{u}^{\mathrm{lin}}(0) - \boldsymbol{y})$.
As a consequence, if $\eta \leq \frac{2n}{\|\boldsymbol{K}\|}$, then we have $\left\|\boldsymbol{u}^{\mathrm{lin}}(t) - \boldsymbol{y}\right\| \leq \left\|\boldsymbol{u}^{\mathrm{lin}}(0) - \boldsymbol{y}\right\|$ for all $t \geq 0$.

Analytical form - Proof idea Claim C.1

- By definition we have

$$\boldsymbol{u}^{\mathrm{lin}}(t+1) = \boldsymbol{u}^{\mathrm{lin}}(t) - \frac{1}{n}\eta \boldsymbol{K}(\boldsymbol{u}^{\mathrm{lin}}(t) - \boldsymbol{y})$$

  which implies

$$\boldsymbol{u}^{\mathrm{lin}}(t+1) - \boldsymbol{y} = \left(\boldsymbol{I} - \frac{1}{n}\eta \boldsymbol{K}\right)(\boldsymbol{u}^{\mathrm{lin}}(t) - \boldsymbol{y})$$

- Thus the first statement follows directly.
- Then the second statement can be proved by noting that $\left\|\boldsymbol{I} - \frac{1}{n}\eta \boldsymbol{K}\right\| \leq 1$ when $\eta \leq \frac{2n}{\|\boldsymbol{K}\|}$.

Assumption 1

We make the following key assumption that connects these two problems:

Key Assumption

There exist $0 < \epsilon < \|\boldsymbol{K}\|$, $R > 0$ such that for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^N$, as long as $\|\boldsymbol{\theta} - \boldsymbol{\theta}(0)\| \leq R$ and $\|\boldsymbol{\theta}' - \boldsymbol{\theta}(0)\| \leq R$, we have

$$\left\| \boldsymbol{J}(\boldsymbol{\theta}) \boldsymbol{J}(\boldsymbol{\theta}')^\top - \boldsymbol{K} \right\| \leq \epsilon.$$

We will prove this later!

Main Theorem - General result

Theorem

Suppose that the initializations are chosen so that $\boldsymbol{u}(0) = \boldsymbol{u}^{\mathrm{lin}}(0) = \mathbf{0}$, and that the learning rate satisfies $\eta \leq \frac{n}{\|\boldsymbol{K}\|}$. Suppose that Assumption 1 is satisfied with $R^2\epsilon < n$. Then there exists a universal constant $c > 0$ such that for all $0 \leq t \leq c\frac{R^2}{\eta}$:

- (closeness of predictions) $\left\|\boldsymbol{u}(t) - \boldsymbol{u}^{\mathrm{lin}}(t)\right\| \lesssim \frac{\eta t \epsilon}{\sqrt{n}}$;
- (boundedness of parameter movement) $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\| \leq R, \|\boldsymbol{\omega}(t) - \boldsymbol{\omega}(0)\| \leq R$.

Proof

We use induction to prove $\left\|\boldsymbol{u}(t) - \boldsymbol{u}^{\mathrm{lin}}(t)\right\| \lesssim \frac{\eta t \epsilon}{\sqrt{n}}$ and $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\| \leq R$.

Step 1: proving $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\| \leq R$. We define

$$\boldsymbol{J}(\boldsymbol{\theta} \to \boldsymbol{\theta}') := \int_0^1 \boldsymbol{J}(\boldsymbol{\theta} + x(\boldsymbol{\theta}' - \boldsymbol{\theta}))dx.$$

We first prove $\|\boldsymbol{\theta}(t-1) - \boldsymbol{\theta}(0)\| \leq \frac{R}{2}$. If $t = 1$, this is trivially true. Now we assume $t \geq 2$. For each $0 \leq \tau < t-1$, by the fundamental theorem for line integrals we have

$$\begin{aligned}
\boldsymbol{u}(\tau+1) - \boldsymbol{u}(\tau) &= \boldsymbol{J}(\boldsymbol{\theta}(\tau) \to \boldsymbol{\theta}(\tau+1)) \cdot (\boldsymbol{\theta}(\tau+1) - \boldsymbol{\theta}(\tau)) \\
&= -\frac{\eta}{n}\boldsymbol{J}(\boldsymbol{\theta}(\tau) \to \boldsymbol{\theta}(\tau+1))\boldsymbol{J}(\boldsymbol{\theta}(\tau))^\top(\boldsymbol{u}(\tau) - \boldsymbol{y}).
\end{aligned}$$

Proof

Let $\boldsymbol{E}(\tau) := \boldsymbol{J}(\boldsymbol{\theta}(\tau) \to \boldsymbol{\theta}(\tau+1))\boldsymbol{J}(\boldsymbol{\theta}(\tau))^\top - \boldsymbol{K}$

$\implies$ From Assumption 1, $\|\boldsymbol{E}(\tau)\| \leq \epsilon$.

Thus

$$\begin{aligned} &\|\boldsymbol{u}(\tau+1) - \boldsymbol{y}\|^2 \\ \leq\ &\|\boldsymbol{u}(\tau) - \boldsymbol{y}\|^2 - \frac{\eta}{n}(\boldsymbol{u}(\tau) - \boldsymbol{y})^\top \boldsymbol{K}(\boldsymbol{u}(\tau) - \boldsymbol{y}) + O(\eta\epsilon). \end{aligned}$$

$$\left(\frac{\eta^2\|\boldsymbol{K}\|}{n^2} \leq \frac{\eta}{n}\right)$$

On the other hand, we have

$$\begin{aligned} &\|\boldsymbol{\theta}(\tau+1) - \boldsymbol{\theta}(\tau)\|^2 \\ \overset{GD}{=}\ &\frac{\eta^2}{n^2}\left\|\boldsymbol{J}(\boldsymbol{\theta}(\tau))^\top(\boldsymbol{u}(\tau) - \boldsymbol{y})\right\|^2 \\ \leq\ &\frac{\eta^2}{n^2}\left((\boldsymbol{u}(\tau) - \boldsymbol{y})^\top \boldsymbol{K}(\boldsymbol{u}(\tau) - \boldsymbol{y}) + O(n\epsilon)\right). \end{aligned} \tag{5}$$

Proof

Combining the above two inequalities, we obtain

$$\begin{aligned}&\|\boldsymbol{u}(\tau+1)-\boldsymbol{y}\|^2-\|\boldsymbol{u}(\tau)-\boldsymbol{y}\|^2\\ \leq& -\frac{n}{\eta}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\|^2+O(\eta\epsilon).\end{aligned}$$

Taking sum over $\tau = 0, \ldots, t-2$, we get

$$\|\boldsymbol{u}(t-1)-\boldsymbol{y}\|^2-\|\boldsymbol{u}(0)-\boldsymbol{y}\|^2 \leq -\frac{n}{\eta}\sum_{\tau=0}^{t-2}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\|^2+O(\eta t\epsilon),$$

which implies

$$\frac{n}{\eta}\sum_{\tau=0}^{t-2}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\|^2 \leq \|\boldsymbol{y}\|^2+O(\eta t\epsilon) \leq \|\boldsymbol{y}\|^2+O(R^2\epsilon)=O(n).$$

Proof

Then by the Cauchy-Schwartz inequality we have

$$\begin{aligned}
\|\boldsymbol{\theta}(t-1)-\boldsymbol{\theta}(0)\| &\leq \sum_{\tau=0}^{t-2}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\| \\
&\leq \sqrt{(t-1)\sum_{\tau=0}^{t-2}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\|^2} \\
&\leq \sqrt{t\cdot O(\eta)} \leq \sqrt{c\frac{R^2}{\eta}\cdot O(\eta)}.
\end{aligned}$$

- Choosing c sufficiently small, we can ensure $\|\boldsymbol{\theta}(t-1)-\boldsymbol{\theta}(0)\| \leq \frac{R}{2}$.

Proof

Then by the Cauchy-Schwartz inequality we have

$$
\begin{aligned}
\|\boldsymbol{\theta}(t-1)-\boldsymbol{\theta}(0)\| &\leq \sum_{\tau=0}^{t-2}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\| \\
&\leq \sqrt{(t-1)\sum_{\tau=0}^{t-2}\|\boldsymbol{\theta}(\tau+1)-\boldsymbol{\theta}(\tau)\|^2} \\
&\leq \sqrt{t\cdot O(\eta)} \leq \sqrt{c\frac{R^2}{\eta}\cdot O(\eta)}.
\end{aligned}
$$

- Choosing c sufficiently small, we can ensure $\|\boldsymbol{\theta}(t-1)-\boldsymbol{\theta}(0)\| \leq \frac{R}{2}$.
- Using the exact same method, can prove $\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(t-1)\| \leq \frac{R}{2}$

Proof

Step 2: proving $\left\|\boldsymbol{u}(t)-\boldsymbol{u}^{\mathrm{lin}}(t)\right\| \lesssim \frac{\eta t \epsilon}{\sqrt{n}}$.

- Same as before we have

$$\boldsymbol{u}(t)-\boldsymbol{y}=\left(\boldsymbol{I}-\frac{\eta}{n}\boldsymbol{K}\right)(\boldsymbol{u}(t-1)-\boldsymbol{y})-\frac{\eta}{n}\boldsymbol{E}(t-1)(\boldsymbol{u}(t-1)-\boldsymbol{y}),$$

  where $\boldsymbol{E}(t-1)=\boldsymbol{J}(\boldsymbol{\theta}(t-1),\boldsymbol{\theta}(t))\boldsymbol{J}(\boldsymbol{\theta}(t-1))^{\top}-\boldsymbol{K}$.

- Since $\|\boldsymbol{\theta}(t-1)-\boldsymbol{\theta}(0)\| \leq R$ and $\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\| \leq R$, we know from Assumption 1 that $\|\boldsymbol{E}(t-1)\| \leq \epsilon$. Moreover, from Claim C.1 we know

$$\boldsymbol{u}^{\mathrm{lin}}(t)-\boldsymbol{y}=\left(\boldsymbol{I}-\frac{\eta}{n}\boldsymbol{K}\right)(\boldsymbol{u}^{\mathrm{lin}}(t-1)-\boldsymbol{y}).$$

- Combine...

Proof Main Theorem

First Layer

Assumption 1

The next lemma verifies Assumption 1 for training the first layer.

Lemma to prove Assumption 1

Let $R = \sqrt{d \log d}$. With high probability over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, for all $\boldsymbol{W}, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{m \times d}$ such that $\|\boldsymbol{W} - \boldsymbol{W}(0)\|_F \leq R$ and $\left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}(0)\right\|_F \leq R$, we have

$$\left\|\boldsymbol{J}_1(\boldsymbol{W}, \boldsymbol{v})\boldsymbol{J}_1(\widetilde{\boldsymbol{W}}, \boldsymbol{v})^\top - \boldsymbol{\Theta}^{\mathrm{lin1}}\right\| \lesssim \frac{n}{d^{1+\frac{\alpha}{7}}}.$$

We can then apply the previous result to prove the main theorem.

Assumption 1

The next lemma verifies Assumption 1 for training the first layer.

Lemma to prove Assumption 1

Let $R = \sqrt{d \log d}$. With high probability over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, for all $\boldsymbol{W}, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{m \times d}$ such that $\|\boldsymbol{W} - \boldsymbol{W}(0)\|_F \leq R$ and $\left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}(0)\right\|_F \leq R$, we have

$$\left\| \boldsymbol{J}_1(\boldsymbol{W}, \boldsymbol{v}) \boldsymbol{J}_1(\widetilde{\boldsymbol{W}}, \boldsymbol{v})^\top - \boldsymbol{\Theta}^{\mathrm{lin1}} \right\| \lesssim \frac{n}{d^{1+\frac{\alpha}{7}}}.$$

We can then apply the previous result to prove the main theorem.

Proof of Lemma: relies on two elements:

- Proposition on Distance kernels stated earlier
- Bound on Jacobian perturbation

Proposition - Distance kernels

Proposition - Distance kernels

With high probability over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, we have $\left\|\boldsymbol{\Theta}_1(\boldsymbol{W}(0)) - \boldsymbol{\Theta}^{\mathrm{lin1}}\right\| \lesssim \frac{n}{d^{1+\alpha}}$.

To prove this proposition we will prove $\boldsymbol{\Theta}_1(\boldsymbol{W}(0))$ is close to its expectation $\boldsymbol{\Theta}_1^*$, and then prove $\boldsymbol{\Theta}_1^*$ is close to $\boldsymbol{\Theta}^{\mathrm{lin1}}$. We do these steps in the next two propositions.

Proof Proposition - Distance kernels

First-layer NTK - Concentration

With high probability over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, we have

$$\|\boldsymbol{\Theta}_1(\boldsymbol{W}(0)) - \boldsymbol{\Theta}_1^*\| \leq \frac{n}{d^{1+\alpha}}.$$

Proof idea: Matrix Bernstein.

First-layer NTK - Approximation

With high probability over the training data $\boldsymbol{X}$, we have

$$\left\|\boldsymbol{\Theta}_1^* - \boldsymbol{\Theta}^{\mathrm{lin1}}\right\| \lesssim \frac{n}{d^{1+\alpha}}.$$

Proof idea: Entrywize Taylor expansion of $\mathbb{E}_{\boldsymbol{W}(0)} \|\boldsymbol{\Theta}_1(\boldsymbol{W}(0))\| +$ concentration bounds.

Proof Assumption 1

Jacobian perturbation for the first layer

If ϕ is a smooth activation, then w.h.p. over the training data $\boldsymbol{X}$, we have

$$\left\| \boldsymbol{J}_1(\boldsymbol{W}, \boldsymbol{v}) - \boldsymbol{J}_1(\widetilde{\boldsymbol{W}}, \boldsymbol{v}) \right\| \lesssim \sqrt{\frac{n}{md}} \left\| \boldsymbol{W} - \widetilde{\boldsymbol{W}} \right\|_F, \quad \forall \boldsymbol{W}, \widetilde{\boldsymbol{W}} \tag{6}$$

If ϕ is a piece-wise linear activation, then w.h.p. over the random initialization $\boldsymbol{W}(0)$ and the training data $\boldsymbol{X}$, we have

$$\| \boldsymbol{J}_1(\boldsymbol{W}, \boldsymbol{v}) - \boldsymbol{J}_1(\boldsymbol{W}(0), \boldsymbol{v}) \| \lesssim \sqrt{\frac{n}{d}} \left(\frac{\| \boldsymbol{W} - \boldsymbol{W}(0) \|^{1/3}}{m^{1/6}} + \left(\frac{\log n}{m} \right)^{1/4} \right) \tag{7}$$

Proof Assumption 1

$$
\begin{aligned}
&\left\| \boldsymbol{J}_1(\boldsymbol{W}, \boldsymbol{v}) \boldsymbol{J}_1(\widetilde{\boldsymbol{W}}, \boldsymbol{v})^\top - \boldsymbol{\Theta}^{\mathrm{lin1}} \right\| \\
&\leq \left\| \boldsymbol{J}_1(\boldsymbol{W}, \boldsymbol{v}) \boldsymbol{J}_1(\widetilde{\boldsymbol{W}}, \boldsymbol{v})^\top - \boldsymbol{J}_1(\boldsymbol{W}(0), \boldsymbol{v}) \boldsymbol{J}_1(\boldsymbol{W}(0), \boldsymbol{v})^\top \right\| \\
&+ \left\| \boldsymbol{J}_1(\boldsymbol{W}(0), \boldsymbol{v}) \boldsymbol{J}_1(\boldsymbol{W}(0), \boldsymbol{v})^\top - \boldsymbol{\Theta}^{\mathrm{lin1}} \right\| \\
&\leq \frac{n}{d^{1+\frac{\alpha}{7}}},
\end{aligned}
$$

where the last inequality uses the two previous lemma.

Things I didn't cover that are in the paper

- Proof for second layer: similar to first one
- Experiments: two-layer neural network with erf activation and width 256 on synthetic data generated
- Extensions to Multi-Layer and Convolutional Neural Networks

The end

The end