Global Convergence in Neural ODEs: Impact of Activation Functions

20252705 박근서, 20252343 임승균

Postech

December 14, 2025

Overview

1. Introduction & Motivation
2. Gradient Convergence
3. NTK Convergence
4. SPD Condition
5. Global Convergence Analysis

Introduction & Motivation

What is Neural ODE?

ResNet (Discrete, L layers):

$$\boldsymbol{h}^{\ell+1} = \boldsymbol{h}^{\ell} + \frac{1}{L}\boldsymbol{W}\phi(\boldsymbol{h}^{\ell}), \quad \ell = 0, 1, \ldots, L-1$$

Neural ODE (Continuous, $L \to \infty$):

$$\dot{\boldsymbol{h}}_t = \boldsymbol{W}\phi(\boldsymbol{h}_t), \quad t \in [0, T]$$

Pros: Continuous-depth, memory efficient, flexible time horizon
Cons: Difficult to train, no convergence guarantee

Introduction & Motivation

Research Question: ResNet vs Neural ODE

When training Neural ODEs with gradient descent,
is **global convergence** guaranteed?

ResNet: Global convergence guranteed

- NTK (Neural Tangent Kernel) theory (Jacot et al., 2018)
- In overparameterized regime, training dynamics $\approx$ kernel regression
- Key: NTK is **SPD** (Strictly Positive Definite) $\Rightarrow$ convergence

Neural ODE: Global convergence unknown

- Infinite depth $\rightarrow$ cannot use layer-by-layer induction
- Existing NTK theory does not directly apply

Contribution

1. **Gradient Convergence:** Smooth activation $\Rightarrow$ gradients are well-defined
2. **NTK Convergence:** Neural ODE's NTK converges to a deterministic kernel
3. **SPD Guarantee:** Non-polynomial activation $\Rightarrow$ NTK is SPD
4. **First global convergence guarantee for Neural ODEs!**

Introduction & Motivation

Neural ODE Definition

Model Output:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\sigma_v}{\sqrt{n}}\boldsymbol{v}^\top \phi(\boldsymbol{h}_T)$$

Hidden State Dynamics:

$$\boldsymbol{h}_0 = \frac{\sigma_u}{\sqrt{d}}\boldsymbol{U}\boldsymbol{x}, \quad \dot{\boldsymbol{h}}_t = \frac{\sigma_w}{\sqrt{n}}\boldsymbol{W}\phi(\boldsymbol{h}_t), \quad t \in [0, T]$$

Parameters:

- $\boldsymbol{\theta} = \{\boldsymbol{v}, \boldsymbol{W}, \boldsymbol{U}\}$
- $\boldsymbol{v} \in \mathbb{R}^n$: Output weights
- $\boldsymbol{W} \in \mathbb{R}^{n\times n}$: Hidden dynamics
- $\boldsymbol{U} \in \mathbb{R}^{n\times d}$: Input projection
- n: Width, T: Time horizon, ϕ: Activation function

Introduction & Motivation

Key Challenge

Problem: Is the gradient of Neural ODE well-defined?

Existing NTK Theory:

- For finite-depth networks: prove by **induction** over layers
- Neural ODE is **continuous** $\rightarrow$ induction doesn't work!

This Paper's Strategy: Approximate with finite-depth ResNet

$$f^L(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\sigma_v}{\sqrt{n}}\boldsymbol{v}^\top \phi(\boldsymbol{h}^L(\boldsymbol{x}))$$

$$\boldsymbol{h}^\ell = \boldsymbol{h}^{\ell-1} + \kappa \cdot \frac{\sigma_w}{\sqrt{n}}\boldsymbol{W}\phi(\boldsymbol{h}^{\ell-1}), \quad \kappa = \frac{T}{L}$$

As $L \rightarrow \infty$: ResNet $\rightarrow$ Neural ODE

Question: Does the ResNet gradient converge to the Neural ODE gradient?

Proposition 2

If ϕ is L_1-Lipschitz and ϕ' is L_2-Lipschitz:

$$\|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}^L - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}\| \leq \frac{C}{L}$$

Gradient Convergence

Why Smooth Activation?

Backward ODE (Adjoint Equation):

$$\dot{\boldsymbol{\lambda}}_t = -\frac{\sigma_w}{\sqrt{n}} \text{diag}(\phi'(\boldsymbol{h}_t)) \boldsymbol{W}^\top \boldsymbol{\lambda}_t$$

Gradient computation requires $\phi'(\boldsymbol{h}_t)$ (derivative of activation).

ResNet Backward Pass:

$$\boldsymbol{\lambda}^{\ell-1} = \boldsymbol{\lambda}^{\ell} + \frac{T}{L} \cdot \text{diag}(\phi'(\boldsymbol{h}^{\ell-1})) \boldsymbol{W}^\top \boldsymbol{\lambda}^{\ell}$$

As $L \to \infty$, this sum becomes an integral:

$$\sum_{\ell=1}^{L} \frac{T}{L} \phi'(\boldsymbol{h}^{\ell}) \quad \longrightarrow \quad \int_0^T \phi'(\boldsymbol{h}_t)\, dt$$

NTK Convergence

Why Do We Need NTK?

Recall: Training Dynamics

$$\boldsymbol{u}^{k+1} - \boldsymbol{y} = (\boldsymbol{I} - \eta \boldsymbol{H}^k)(\boldsymbol{u}^k - \boldsymbol{y})$$

where $\boldsymbol{H}_{ij}^k = K_{\boldsymbol{\theta}^k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \nabla_\theta f(\boldsymbol{x}_i), \nabla_\theta f(\boldsymbol{x}_j) \rangle$

For Convergence:

- Need $\lambda_{\min}(\boldsymbol{H}^k) > 0$ throughout training
- In overparameterized regime ($n \to \infty$): $\boldsymbol{H}^k \approx \boldsymbol{H}^0 \approx K_\infty$
- So we need: $\lambda_{\min}(K_\infty) > 0$

Key Question

Does K_∞ even exist for Neural ODE? (infinite depth!)

NTK Convergence

Building Blocks

Step 1: Width Convergence (Proposition 4)

For fixed depth L, as width $n \to \infty$:

$$K_{\boldsymbol{\theta}}^{L} \xrightarrow{n\to\infty} K_{\infty}^{L} \quad \text{(deterministic)}$$

Step 2: Depth Convergence (Lemma 2)

For fixed width n, as depth $L \to \infty$:

$$|K_{\boldsymbol{\theta}}^{L} - K_{\boldsymbol{\theta}}| \leq \frac{C}{L} \quad \text{(uniform in } n\text{)}$$

Step 3: Moore-Osgood Theorem

If both convergences are **uniform**, the limits can be exchanged!

NTK Convergence

Theorem 2: Double Limit

$$
\begin{array}{ccc}
K_{\boldsymbol{\theta}}^{L} & \xrightarrow[\text{(Prop 4)}]{n \to \infty} & K_{\infty}^{L} \\
L \to \infty \ \Big\downarrow \ \text{(Lemma 2)} & & \Big\downarrow \ L \to \infty \\
K_{\boldsymbol{\theta}} & \xrightarrow[\text{(Thm 2)}]{n \to \infty} & K_{\infty}
\end{array}
$$

NTK Convergence

Theorem 2: Double Limit

Theorem 2

If ϕ is L_1-Lipschitz and ϕ' is L_2-Lipschitz:

$$K_{\boldsymbol{\theta}} \xrightarrow{n\to\infty} K_\infty$$

The NTK of Neural ODE converges to a deterministic kernel K_∞!

SPD Condition

Corollary 1: Statement

Corollary 1

If ϕ is Lipschitz, nonlinear, and **non-polynomial**:

$$\lambda_0 = \lambda_{\min}(K_\infty) > 0$$

Proof Outline:

1. Decompose NTK: $K_\infty = K_\infty^v + K_\infty^W + K_\infty^U$
2. Show $K_\infty^v = \sigma_v^2 \cdot \Sigma^*$ (NNGP kernel)
3. Use Hermite expansion to analyze Σ^*
4. Given Condition $\Rightarrow \Sigma^*$ is SPD $\Rightarrow K_\infty$ is SPD

SPD Condition

Step 1: NTK Decomposition

NTK Definition:

$$K_\theta(x, \bar{x}) = \langle \nabla_\theta f(x), \nabla_\theta f(\bar{x}) \rangle$$

Since $\theta = \{v, W, U\}$:

$$K_\infty = \underbrace{\left\langle \frac{\partial f}{\partial v}, \frac{\partial f}{\partial v} \right\rangle}_{K_\infty^v} + \underbrace{\left\langle \frac{\partial f}{\partial W}, \frac{\partial f}{\partial W} \right\rangle}_{K_\infty^W} + \underbrace{\left\langle \frac{\partial f}{\partial U}, \frac{\partial f}{\partial U} \right\rangle}_{K_\infty^U}$$

Each term is **positive semi-definite**, so:

$$K_\infty \geq K_\infty^v$$

Key: If K_∞^v is SPD, then K_∞ is also SPD!

SPD Condition

Step 2: K^v_∞ and NNGP Kernel

Gradient w.r.t. output layer:

$$\frac{\partial f}{\partial v} = \frac{\sigma_v}{\sqrt{n}}\phi(h_T)$$

Therefore:

$$K^v(x, \bar{x}) = \frac{\sigma_v^2}{n}\sum_{i=1}^{n}\phi(h_T^{(i)}(x))\phi(h_T^{(i)}(\bar{x}))$$

As $n \to \infty$ (Law of Large Numbers):

$$K^v_\infty(x, \bar{x}) = \sigma_v^2 \cdot \underbrace{\mathbb{E}[\phi(h_T(x))\phi(h_T(\bar{x}))]}_{\Sigma^*(x,\bar{x})}$$

SPD Condition

Step 3: Hermite Expansion

Hermite Polynomials: $\{h_n(x)\}_{n=0}^{\infty}$ form an orthonormal basis

- $h_0(x) = 1, \quad h_1(x) = x, \quad h_2(x) = x^2 - 1, \quad \ldots$
- Orthonormal: $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[h_n(z)h_m(z)] = \delta_{nm}$

Any function can be expanded:

$$\phi(x) = \sum_{n=0}^{\infty} a_n h_n(x), \quad a_n = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(z)h_n(z)]$$

Key Property:

For $(u, \bar{u}) \sim \mathcal{N}(0, S^*)$ with correlation ρ:

$$\mathbb{E}[h_n(u)h_m(\bar{u})] = \rho^n \delta_{nm}$$

SPD Condition

Step 3: Hermite Expansion

NNGP Kernel:

$$\Sigma^*(x, \bar{x}) = \mathbb{E}[\phi(u)\phi(\bar{u})]$$

Substitute Hermite expansion:

$$\begin{aligned}
\Sigma^* &= \mathbb{E}\left[\left(\sum_{n=0}^{\infty} a_n h_n(u)\right)\left(\sum_{m=0}^{\infty} a_m h_m(\bar{u})\right)\right] \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} a_n a_m \underbrace{\mathbb{E}[h_n(u)h_m(\bar{u})]}_{\rho^n \delta_{nm}} \\
&= \sum_{n=0}^{\infty} a_n^2 \rho^n
\end{aligned}$$

SPD Condition

Step 4: Theorem 11

Theorem 11

Σ^* is SPD $\iff$ **infinitely many** $a_n \neq 0$

Proof Idea ($\Leftarrow$):

- Suppose $\Sigma^* c = 0$ for some $c \neq 0$
- Then $c^\top \Sigma^* c = \sum_{n=0}^{\infty} a_n^2 (c^\top \rho^{\circ n} c) = 0$
- Since $a_n^2 \geq 0$, we need $c^\top \rho^{\circ n} c = 0$ for all n with $a_n \neq 0$
- Infinitely many such constraints on $c \Rightarrow$ only $c = 0$ satisfies all
- Contradiction! So Σ^* is SPD.

SPD Condition

Conclusion

Non-polynomial ϕ (e.g., Softplus, Tanh, GELU):

- Cannot be written as finite sum of Hermite polynomials
- **Infinitely many** $a_n \neq 0$
- By Theorem 11: SPD **guaranteed!**

Conclusion

Non-polynomial activation $\Rightarrow \Sigma^*$ is SPD $\Rightarrow K_\infty$ is SPD

Assumption 1. Let $\{x_i, y_i\}_{i=1}^N$ be a training set. Assume the following conditions:

1. **Training set:** $x_i \in \mathbb{S}^{d-1}$ and $x_i \neq x_j$ for all $i \neq j$; moreover, $|y_i| = O(1)$.
2. **Smoothness:** The activation function ϕ and its derivative ϕ' are L_1- and L_2-Lipschitz continuous, respectively.
3. **Nonlinearity:** The activation ϕ is nonlinear and non-polynomial.

Theorem 3.

1. The parameters $\boldsymbol{\theta}^k$ stay in a neighborhood of $\boldsymbol{\theta}^0$, i.e.,

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^0\| \leq C\|X\|\sqrt{\frac{L(\boldsymbol{\theta}_0)}{\boldsymbol{\lambda}_0}},$$

2. The loss $L(\boldsymbol{\theta}^k)$ decays exponentially, i.e.,

$$L(\boldsymbol{\theta}^k) \leq \left(1 - \frac{\eta\boldsymbol{\lambda}_0}{16}\right)^k L(\boldsymbol{\theta}^0).$$

where $\lambda_0 := \lambda_{\min}(K_\infty) > 0$, and the constant $C > 0$ depends only on L_1, L_2, σ_v, σ_w, σ_u, and T.

Covergence Analysis

It is hard to show the proof of Theorem (3) in general case, so we provide the convergence analysis of Neural ODEs defined equation 1 under the gradient descent.

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\sigma_v}{\sqrt{n}}\boldsymbol{v}^\top \phi(\boldsymbol{h}_T) \tag{1}$$

Lemma 16

Lemma 16. Assume ϕ and ϕ' are L_1- and L_2-Lipschitz continuous and $\lambda_0 := \lambda_{\min}(K_{\theta^0}) > 0$. Suppose we choose the width $n = \Omega\big(\|X\|^4 \|u^0 - y\|^2 / \lambda_0^3\big)$ and the learning rate $\eta \leq 1/\|X\|^2$.
Then the parameters θ^k stay in the neighborhood of θ^0, i.e.

$$\|v^k - v^0\|,\ \|W_k - W_0\|,\ \|U^k - U^0\| \leq C\frac{\|X\|\,\|u^0 - y\|}{\lambda_0}, \qquad (2)$$

and the residual decays geometrically:

$$\|u^k - y\| \leq \left(1 - \frac{\eta\lambda_0}{8}\right)^k \|u^0 - y\|, \qquad (3)$$

where $C > 0$ depends only on $L_1, L_2, \sigma_v, \sigma_w, \sigma_u, T$.

Lemma 17

Lemma 17. Given θ, for all $t \in [0, T]$:

$$\|h_t\| \leq \|U\| \, \|x\| \, \exp\left(\frac{\sigma t}{\sqrt{n}} \|W\|\right), \tag{4}$$

$$\|\lambda_t\| \leq \frac{\|v\|}{\sqrt{n}} \, \exp\left(\frac{\sigma(T-t)}{\sqrt{n}} \|W\|\right). \tag{5}$$

Intuition: Hidden state growth controlled by integrating ODE; adjoint decays backward in time. Constants arise from σ scaling and $1/\sqrt{n}$ normalization.

Lemma 18

Lemma 18 (as in paper). For two parameter tuples θ, $\bar{\theta}$ and all $t \in [0, T]$:

$$\|h_t - \bar{h}_t\| \leq \|\theta - \bar{\theta}\| \cdot \|U\| \, \|W\| \, \exp\left(\frac{\sigma t(\|W\| + \|\bar{W}\|)}{\sqrt{n}}\right) \|x\|, \qquad (6)$$

$$\|\lambda_t - \bar{\lambda}_t\| \leq \|\theta - \bar{\theta}\| \cdot \frac{\|v\| \, \|W\|}{\sqrt{n}} \, \exp\left(\frac{\sigma(T-t)(\|W\| + \|\bar{W}\|)}{\sqrt{n}}\right). \qquad (7)$$

Intuition: Sensitivity ODEs + Grönwall give linear dependence on parameter perturbation; exponential factor from integrating Jacobians.

Preliminaries and notation

- Predictions vector: $u^k = [f(x_i;\theta^k)]_{i=1}^N$ and labels y.
- Loss function: $L(\theta) := \sum_{i=1}^{N} \frac{1}{2}(f_\theta(x_i) - y_i)^2$.
- Gradient of f_θ:

$$\partial_v f_\theta(x) = \frac{\sigma_v}{\sqrt{n}}\phi(h_T)$$

$$\partial_W f_\theta(x) = \int_0^T \frac{\sigma_W}{\sqrt{n}}(\phi(h_t) \otimes \lambda_t)dt$$

$$\partial_U f_\theta(x) = \frac{\sigma_u}{\sqrt{d}}[x \otimes \lambda(0)]$$

Proof of Lemma 16

Consider the gradients of loss function $L(\theta)$

$$\frac{\partial L(\theta)}{\partial v} = \sum_{i=1}^{N} \frac{\sigma_v}{\sqrt{n}} \phi(h_T(x_i))(f_\theta(x_i) - y_i),$$

$$\frac{\partial L(\theta)}{\partial W} = \sum_{i=1}^{N} \left[\int_0^T \frac{\sigma_W}{\sqrt{n}} (\phi(h_t(x_i)) \otimes \lambda_t(x_i)) dt \right] (f_\theta(x_i) - y_i),$$

$$\frac{\partial L(\theta)}{\partial U} = \sum_{i=1}^{N} \frac{\sigma_u}{\sqrt{d}} [x_i \otimes \lambda(0)(x_i)](f_\theta(x_i) - y_i)$$

Also, the gradient descent

$$\theta^{k+1} = \theta^k - \eta \frac{\partial L(\theta^k)}{\partial \theta}$$

Proof of Lemma 16

Assume the inductive hypothesis: For all $i \leq k$,

$$\|v_i\|, \|W_i\|, \|U_i\| \leq C\sqrt{n}$$
$$\|u^i - y\| \leq (1 - \eta\alpha_0^2)^i \|u^0 - y\|$$

where $C > 0$ is a constant and $\alpha_0 := \sigma_{min}(\frac{\sigma_v}{\sqrt{n}}\Phi^0)$

Proof of Lemma 16

Closed

Without loss generality, assume $\sigma_v = 1, \sigma_w = \sigma, \sigma_u/\sqrt{d} = 1$ and $L_1 = L_2 = 1$.
Observe that

$$\|\frac{\partial f_\theta}{\partial v}\| = \|\frac{1}{\sqrt{n}}\phi(h_T)\| \leq \frac{1}{\sqrt{n}}\|U\|\|x\|e^{\sigma T\|W\|/\sqrt{n}}$$

Proof of Lemma 16

Closed

Observe that

$$\|\frac{\partial f_\theta}{\partial W}\| = \|\int_0^T \frac{\sigma_w}{\sqrt{n}}(\phi(h_t(x)) \otimes \lambda_t(x))dt\|$$
$$\leq (\sigma T)\frac{\|U\|}{\sqrt{n}}\frac{\|v\|}{\sqrt{n}}\|x\|e^{\sigma T\|W\|/\sqrt{n}}$$

Proof of Lemma 16

Closed

Observe that

$$\begin{aligned} \| \frac{\partial f_\theta}{\partial U} \| = & \ \| \frac{\sigma_u}{\sqrt{d}} [x \otimes \lambda(0)(x)] \| \\ \leq & \ \|x\| \cdot \frac{\|v\|}{\sqrt{n}} e^{\sigma T \|W\| / \sqrt{n}} \end{aligned}$$

Proof of Lemma 16

Closed

By using the inductive hypothesis, we obtain

$$\|\frac{\partial f_\theta}{\partial v}\| \leq \frac{1}{\sqrt{n}}\|U\|\|x\|e^{\sigma T\|W\|/\sqrt{n}} \leq Ce^{C\sigma T}\|x\|$$

$$\|\frac{\partial f_\theta}{\partial W}\| \leq (\sigma T)\frac{\|U\|}{\sqrt{n}}\frac{\|v\|}{\sqrt{n}}\|x\|e^{\sigma T\|W\|/\sqrt{n}} \leq (\sigma T)Ce^{C\sigma T}\|x\|$$

$$\|\frac{\partial f_\theta}{\partial U}\| \leq \ \|x\| \cdot \frac{\|v\|}{\sqrt{n}}e^{\sigma T\|W\|/\sqrt{n}} \leq Ce^{C\sigma T}\|x\|$$

Proof of Lemma 16

Closed

We can obtain

$$
\begin{aligned}
\|v^{k+1} - v^0\| &\leq \eta \sum_{i=0}^{k} \|\frac{\partial L(\theta^i)}{\partial v}\| \\
&\leq \eta \sum_{i=0}^{k} Ce^{C\sigma T}\|X\|\|u^i - y\| \\
&\leq \eta Ce^{C\sigma T}\|X\| \sum_{i=0}^{k} (1 - \eta\alpha_0^2)^i \|u^0 - y\| \\
&\leq Ce^{C\sigma T}\|X\|\|u^0 - y\|/\alpha_0^2
\end{aligned}
$$

Proof of Lemma 16

Closed

Similarly,

$$\begin{aligned}
\|W^{k+1} - W^0\| &\leq \eta \sum_{i=0}^{k} \|\frac{\partial L(\theta^i)}{\partial W}\| \\
&\leq \eta \sum_{i=0}^{k} (\sigma T) C e^{C\sigma T} \|X\| \|u^i - y\| \\
&\leq \eta (\sigma T) C e^{C\sigma T} \|X\| \sum_{i=0}^{k} (1 - \eta \alpha_0^2)^i \|u^0 - y\| \\
&\leq (\sigma T) C e^{C\sigma T} \|X\| \|u^0 - y\| / \alpha_0^2
\end{aligned}$$

Proof of Lemma 16

Closed

Also

$$\begin{aligned}
\|U^{k+1} - U^0\| &\leq \eta \sum_{i=0}^{k} \|\frac{\partial L(\theta^i)}{\partial U}\| \\
&\leq \eta \sum_{i=0}^{k} Ce^{C\sigma T}\|X\|\|u^i - y\| \\
&\leq \eta Ce^{C\sigma T}\|X\| \sum_{i=0}^{k}(1-\eta\alpha_0^2)^i\|u^0 - y\| \\
&\leq Ce^{C\sigma T}\|X\|\|u^0 - y\|/\alpha_0^2
\end{aligned}$$

Proof of Lemma 16

Closed

If we assume $\|x\| = 1$ and $|y| = 1$,
then we need to ensure

$$Ce^{C\sigma T}\|X\|\|u^0 - y\|/\alpha_0^2 \leq C\sqrt{n}$$
$$(\sigma T)Ce^{C\sigma T}\|X\|\|u^0 - y\|/\alpha_0^2 \leq C\sqrt{n}$$

Hence,

$$\|v^{k+1}\| \leq \|v^{k+1} - v^0\| + \|v^0\| \leq C\sqrt{n}$$
$$\|W^{k+1}\| \leq \|W^{k+1} - W^0\| + \|W^0\| \leq C\sqrt{n}$$
$$\|U^{k+1}\| \leq \|U^{k+1} - U^0\| + \|U^0\| \leq C\sqrt{n}$$

Proof of Lemma 16

Consistently decreases

Observe that

$$
\begin{aligned}
u^{k+1} - y &= u^{k+1} - u^k + (u^k - y) \\
&= \left(\frac{\partial \tilde{u}}{\partial \theta}\right)^\top (\theta^{k+1} - \theta^k) + (u^k - y) \\
&= \left(\frac{\partial \tilde{u}}{\partial \theta}\right)^\top (-\eta \frac{\partial u^k}{\partial \theta})(u^k - y) + (u^k - y) \\
&= \left[I - \eta (\frac{\partial \tilde{u}}{\partial \theta})^\top (\frac{\partial u^k}{\partial \theta})\right](u^k - y) \\
&= \left[I - \eta (\frac{\partial u^k}{\partial \theta})^\top (\frac{\partial u^k}{\partial \theta})\right](u^k - y) + \eta (\frac{\partial u^k}{\partial \theta} - \frac{\partial \tilde{u}}{\partial \theta})^\top \frac{\partial u^k}{\partial \theta}(u^k -
\end{aligned}
$$

where $\tilde{u} = u(\tilde{\theta})$ and $\tilde{\theta}$ is an interpolation in between θ^k and θ^{k+1}

Proof of Lemma 16

Consistently decreases

Note that

$$\begin{aligned} \|\frac{\partial f}{\partial v} - \frac{\partial \hat{f}}{\partial v}\| = & \ \|\frac{1}{\sqrt{n}}\phi(h_T) - \frac{1}{\sqrt{n}}\phi(\bar{h}_T)\| \\ \leq & \ \frac{1}{\sqrt{n}}\|h_T - \bar{h}_T\| \\ \leq & \ \frac{C}{\sqrt{n}}\|\theta - \bar{\theta}\|e^{C\sigma T}\|x\| \end{aligned}$$

Proof of Lemma 16

Consistently decreases

Similarly,

$$
\begin{aligned}
\|\frac{\partial f}{\partial W} - \frac{\partial \hat{f}}{\partial W}\| &\leq \frac{\sigma}{\sqrt{n}} \| \int_0^T \phi(h_t) \otimes \lambda_t - \phi(\bar{h}_t) \otimes \bar{\lambda}_t dt \| \\
&\leq \frac{\sigma}{\sqrt{n}} \int_0^T \Big(\|h_t - \bar{h}_t\| \|\lambda_t\| + \|\bar{h}_t\| \|\lambda_t - \bar{\lambda}_t\| \Big) dt \\
&\leq C \frac{\sigma}{\sqrt{n}} \int_0^T \|\theta - \bar{\theta}\| e^{C\sigma t} \|x\| \cdot e^{C\sigma(T-t)} dt \\
&\leq (\sigma T) \frac{C}{\sqrt{n}} \|\theta - \bar{\theta}\| e^{C\sigma T} \|x\|.
\end{aligned}
$$

and $\|\frac{\partial f}{\partial U} - \frac{\partial \bar{f}}{\partial U}\| \leq \|x\| \|\lambda_0 - \bar{\lambda}_0\| \leq \frac{C}{\sqrt{n}} \|\theta - \bar{\theta}\| e^{C\sigma T} \|x\|.$

Proof of Lemma 16

Consistently decreases

Hence, we have

$$\|\frac{\partial f}{\partial \theta} - \frac{\partial \bar{f}}{\partial \theta}\| = \|\frac{\partial f}{\partial v} - \frac{\partial \bar{f}}{\partial v}\| + \|\frac{\partial f}{\partial W} - \frac{\partial \bar{f}}{\partial W}\| + \|\frac{\partial f}{\partial U} - \frac{\partial \bar{f}}{\partial U}\|$$
$$\leq (\sigma T)\frac{C}{\sqrt{n}}\|\theta - \bar{\theta}\|e^{C\sigma T}\|x\|.$$

Then

$$\|\frac{\partial u^k}{\partial \theta} - \frac{\partial \tilde{u}}{\partial \theta}\| \leq (\sigma T)\frac{C}{\sqrt{n}}\|\theta^k - \tilde{\theta}\|e^{C\sigma T}\|X\|$$
$$\leq (\sigma T)\frac{C}{\sqrt{n}}\|\theta^k - \theta^{k+1}\|e^{C\sigma T}\|X\|$$

where we can use the fact $\tilde{\theta} = \alpha\theta^k + (1-\alpha)\theta^{k+1}$ for some $\alpha \in [0, 1]$.

Proof of Lemma 16

Consistently decreases

Observe that

$$\begin{aligned}\|\theta^{k+1} - \theta^k\| &= \eta \|\frac{\partial L(\theta^k)}{\partial \theta}\| = \eta \|\left(\frac{\partial u^k}{\partial \theta}\right)^\top (u^k - y)\| \\ &\leq \eta(\sigma T) C e^{C\sigma T} \|X\| \|u^k - y\|.\end{aligned}$$

Proof of Lemma 16

Consistently decreases

Hence, we obtain

$$\|\frac{\partial u^k}{\partial \theta} - \frac{\partial \tilde{u}}{\partial \theta}\| \leq \eta(\sigma T)^2 \frac{C}{\sqrt{n}} e^{C\sigma T} \|X\|^2 \|u^k - y\|$$

Proof of Lemma 16

Consistently decreases

using the assumption $\sqrt{n} \geq C(\sigma T)^2 e^{C\sigma T}\|X\|^2\|u^0 - y\|/\alpha_0^3$,

$$\begin{aligned}
\|\frac{\partial u^k}{\partial \theta} - \frac{\partial u^0}{\partial \theta}\| &\leq (\sigma T)\frac{C}{\sqrt{n}}\|\theta^k - \theta^0\| e^{C\sigma T}\|X\| \\
&\leq (\sigma T)\frac{C}{\sqrt{n}} e^{C\sigma T}\|X\| \sum_{i=0}^{k-1}\|\theta^{i+1} - \theta^i\| \\
&\leq \eta(\sigma T)^2\frac{C}{\sqrt{n}} e^{C\sigma T}\|X\|^2 \sum_{i=0}^{k-1}\|u^i - y\| \\
&\leq \eta(\sigma T)^2\frac{C}{\sqrt{n}} e^{C\sigma T}\|X\|^2 \sum_{i=0}^{k-1}(1 - \eta\alpha_0^2)\|u^0 - y\| \\
&\leq \eta(\sigma T)^2\frac{C}{\sqrt{n}} e^{C\sigma T}\|X\|^2\|u^0 - y\|/\alpha_0^2 \leq \alpha_0/2
\end{aligned}$$

Proof of Lemma 16

Consistently decreases

It follows from Weyl's inequality that

$$\sigma_{min}\left(\frac{\partial u^k}{\partial \theta}\right) \geq \sigma_{min}\left(\frac{\partial u^0}{\partial \theta}\right) - \|\frac{\partial u^k}{\partial \theta} - \frac{\partial u^0}{\partial \theta}\| \geq \alpha_0/2$$

and so

$$\lambda_{min}\left[\left(\frac{\partial u^k}{\partial \theta}\right)^\top \left(\frac{\partial u^k}{\partial \theta}\right)\right] \geq \alpha_0^2/4$$

Proof of Lemma 16

Consistently decreases

Therefore, we obtain

$$\begin{aligned}
\|u^{k+1} - y\| &\le [1 - \eta\alpha_0^2/4]\|u^k - y\| + \eta^2(\sigma T)^3 \frac{C}{\sqrt{n}} e^{C\sigma T}\|X\|^3 \|u^k - y\|^2 \\
&\le \left[1 - \eta\alpha_0^2/4 + \eta^2(\sigma T)^3 \frac{C}{\sqrt{n}} e^{C\sigma T}\|X\|^3\|u^0 - y\|\right]\|u^k - y\| \\
&= \left[1 - \eta\left(\alpha_0^2/4 - \eta(\sigma T)^3 \frac{C}{\sqrt{n}} e^{C\sigma T}\|X\|^3\|u^0 - y\|\right)\right]\|u^k - y\| \\
&\le [1 - \eta\alpha_0^2/8]\|u^k - y\|,
\end{aligned}$$

where we assume $\sqrt{n} \ge 8C(\sigma T)^3 e^{C\sigma T}\|X\|^3\|u^0 - y\|/\alpha_0^2$

Proof of Lemma 16

Therefore, we show that

$$\|v^{k+1} - v^0\|, \ \|W^{k+1} - W_0\|, \ \|U^{k+1} - U^0\| \leq C\frac{\|X\| \, \|u^0 - y\|}{\lambda_0},$$
$$\|u^{k+1} - y\| \leq \left(1 - \frac{\eta\lambda_0}{8}\right)^k \|u^0 - y\|,$$

By induction, we prove the Lemma 16.

Conclusion

Assumption 1. Let $\{x_i, y_i\}_{i=1}^N$ be a training set. Assume the following conditions:

1. **Training set:** $x_i \in \mathbb{S}^{d-1}$ and $x_i \neq x_j$ for all $i \neq j$; moreover, $|y_i| = O(1)$.
2. **Smoothness:** The activation function ϕ and its derivative ϕ' are L_1- and L_2-Lipschitz continuous, respectively.
3. **Nonlinearity:** The activation ϕ is nonlinear and non-polynomial.

Conclusion

Theorem 3.

1. The parameters $\boldsymbol{\theta}^k$ stay in a neighborhood of $\boldsymbol{\theta}^0$, i.e.,

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^0\| \leq C\|X\|\sqrt{\frac{L(\boldsymbol{\theta}_0)}{\boldsymbol{\lambda}_0}},$$

2. The loss $L(\boldsymbol{\theta}^k)$ decays exponentially, i.e.,

$$L(\boldsymbol{\theta}^k) \leq \left(1 - \frac{\eta\boldsymbol{\lambda}_0}{16}\right)^k L(\boldsymbol{\theta}^0).$$

where $\lambda_0 := \lambda_{\min}(K_\infty) > 0$, and the constant $C > 0$ depends only on L_1, L_2, σ_v, σ_w, σ_u, and T.