Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: convergence and generalization in neural networks." *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* 2018.

Presented by: Nathan Ng, Kimia Hamidieh, and Haoran Zhang
As part of CSC 2541 Winter 2021

February 25, 2021

Warm up

Gradient Descent and Convergence to global minimum

- If cost function $C(f_\theta)$ is convex with respect to parameters θ, convergence of GD is guaranteed

Gradient Descent and Convergence to global minimum

- If cost function $C(f_\theta)$ is convex with respect to parameters θ, convergence of GD is guaranteed
- The loss function of neural networks is not convex
  - Where does the gradient descent converge?
  - Global or local minimum?

Gradient Descent and Convergence to global minimum

- If cost function $C(f_\theta)$ is convex with respect to parameters θ, convergence of GD is guaranteed
- The loss function of neural networks is not convex
  - Where does the gradient descent converge?
  - Global or local minimum?
- If the loss function is convex in *function space*,
  - Is it possible to converge to global minimum?

Gradient Descent and Convergence to global minimum

- If cost function $C(f_\theta)$ is convex with respect to parameters θ, convergence of GD is guaranteed
- The loss function of neural networks is not convex
  - Where does the gradient descent converge?
  - Global or local minimum?
- If the loss function is convex in *function space*,
  - Is it possible to converge to global minimum?
- What kind of functions are we biased towards at initialization? How do they change during training?

Neural Networks in Function Space

- Realization function of L-layer network $F^{(L)} : \mathbb{R}^P \to \mathcal{F}$, mapping parameters θ to functions f_θ in a space $\mathcal{F}$
- Inner product:

$$\langle f, g \rangle_{p^{in}} = \mathbb{E}_{x \sim p^{in}} \left[f(x)^T g(x) \right]$$

  p^{in}: distribution of training data

- Inner product defined by multi-dimensional kernel: $K : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L \times n_L}$

$$\langle f, g \rangle_K := \mathbb{E}_{x, x' \sim p^{in}} \left[f(x)^T K(x, x') g(x') \right].$$

Kernel Gradient

- The dual form $\mathcal{F}^*$: the dual space of $\mathcal{F}$ with respect to p^{in} i.e. the set of linear forms $\mu : \mathcal{F} \to \mathbb{R}$ of the form $\mu = \langle d, \cdot \rangle_{p^{in}}$ for some $d \in \mathcal{F}$.

Kernel Gradient

- The dual form
- Functional derivative of the cost C $\;C(f) = \frac{1}{2}\|f - f^*\|^2$

$$\partial_f^{in} C|_{f_\theta} = \langle f_\theta - f^*, \cdot \rangle_{p^{in}}$$

Kernel Gradient

- The dual form
- Functional derivative of the cost C
- Kernel
  $\Phi_K : \mathcal{F}^* \to \mathcal{F}$: mapping a dual element $\mu = \langle d, \cdot \rangle_{p^{in}}$ to the function f_μ such that:

$$f_\mu(x) = \Phi_K(\mu)(x) = \langle d, K(x, \cdot) \rangle_{p^{in}}$$

  Using the fact that partial application of the kernel $K_{i,\cdot}(x, \cdot)$ is a function in $\mathcal{F}$

Kernel Gradient

Kernel gradient $\nabla_K C|_{f_\theta}$ is defined as:

$$\nabla_K C|_{f_\theta} = \Phi_K\left(\partial_f^{in} C\big|_{f_\theta}\right) = \mathbb{E}_{x\sim p^{in}}\left[\left(f_\theta(x) - f^*(x)\right)^T K(\cdot, x)\right]$$

maps the functional derivative of cost to the above function.

- a generalization of GD to function spaces

$$\partial_f^{in} C|_{f_\theta} = \langle f_\theta - f^*, \cdot\rangle_{p^{in}}$$

Gradient Flow

$$
\begin{aligned}
&\partial_t f_{\theta(t)} \\
&= \partial_{\theta(t)} F(\theta(t)) \partial_t \theta(t) \\
&= -\partial_{\theta(t)} F(\theta(t)) \partial_{\theta(t)} (C \circ F)(\theta(t)) \\
&= -\partial_{\theta(t)} F(\theta(t)) \mathbb{E}_{x \sim p^{in}} \left[\left(f_{\theta(t)}(x) - f(x) \right)^T \left(\partial_{\theta(t)} F(\theta(t))(x) \right) \right] \\
&= -\mathbb{E}_{x \sim p^{in}} \left[\left(f_{\theta(t)}(x) - f(x) \right)^T \left(\partial_{\theta(t)} F(\theta(t))(\cdot) \right) \left(\partial_{\theta(t)} F(\theta(t))(x) \right) \right] \\
&= -\mathbb{E}_{x \sim p^{in}} \left[\left(f_{\theta(t)}(x) - f^*(x) \right)^T K(\cdot, x) \right] \\
&\Rightarrow K(\cdot, x) = \left(\partial_{\theta(t)} F(\theta(t))(\cdot) \right) \left(\partial_{\theta(t)} F(\theta(t))(x) \right)
\end{aligned}
$$

Neural Tangent Kernel

Main Idea

$$\partial_t f_\theta(t) = -\mathbb{E}_{x \sim p^{in}} \left[\left(f_{\theta(t)}(x) - f^*(x) \right)^T K(\cdot, x) \right]$$

If the kernel remains constant, we have a linear differential equation with solution:

$$f_t = f^* + e^{-t\Pi}(f_0 - f^*)$$

where Π is a map of : $f \mapsto \Phi_K \left(\langle f, \cdot \rangle_{p^{in}} \right)$

Main Idea

$$\partial_t f_\theta(t) = -\mathbb{E}_{x\sim p^{in}}\left[\left(f_{\theta(t)}(x) - f^*(x)\right)^T K(\cdot, x)\right]$$

During training, the network function f_θ evolves along the (negative) kernel gradient

$$\partial_t f_{\theta(t)} = -\nabla_{\Theta^{(L)}} C|_{f_{\theta(t)}}$$

with respect to the *neural tangent kernel* (NTK)

$$\Theta^{(L)}(\theta)\left(x, x'\right) = \sum_{p=1}^{P}\left(\partial_{\theta_p} F^{(L)}(\theta)(x)\right)^T \left(\partial_{\theta_p} F^{(L)}(\theta)\left(x'\right)\right)$$

$$\Rightarrow \Theta^{(L)}(\theta) = \sum_{p=1}^{P} \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta)$$

Neural Tangent Kernel

$$\Theta^{(L)}(\theta) = \sum_{p=1}^{P} \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta)$$

- Depends on the parameters $\Rightarrow$ random at initialization, time-dependent
- By Theorem 1. and 2. at infinite width limit:
  - Converges to a deterministic limit at initialization
  - Fixed during training

Notation

- Network function $f_\theta(x) := \tilde{\alpha}^{(L)}(x;\theta)$, where

$$\alpha^{(0)}(x;\theta) = x$$
$$\tilde{\alpha}^{(\ell+1)}(x;\theta) = \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(x;\theta) + \beta b^{(\ell)}$$
$$\alpha^{(\ell)}(x;\theta) = \sigma(\tilde{\alpha}^{(\ell)}(x;\theta))$$

Initialization

- In the infinite width limit $n_1, ..., n_{L-1} \to \infty$
- Initialize the parameters $\theta \sim \mathcal{N}(0, Id_n)$
- The output functions (pre-activations) $f_{\theta,k}$, for $k = 1, ..., n_L$, is a Gaussian processes of covariance $\Sigma^{(L)}$

$$\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^T x' + \beta^2$$

$$\Sigma^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}\left(0, \Sigma^{(L)}\right)}[\sigma(f(x))\sigma(f(x'))] + \beta^2$$

Initialization

An L-layer neural network at **initialization**, and when $n_1, \ldots, n_{L-1} \to \infty$, then the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel $\Theta^{(L)} \to \Theta_\infty^{(L)} \otimes Id_n$

$$\begin{aligned}
\Theta_\infty^{(1)}(x, x') &= \Sigma^{(1)}(x, x') \\
\Theta_\infty^{(L+1)}(x, x') &= \Theta_\infty^{(L)}(x, x')\dot{\Sigma}^{(L+1)}(x, x') + \Sigma^{(L+1)}(x, x'),
\end{aligned}$$

Proof by induction

Training

Given a training direction $t \mapsto d_t \in F$, the parameters θ_p are trained following the differential equation:

$$\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}, d_t \right\rangle_{p^{in}}$$

If $\int_0^T \|d_t\|_{p^{in}} dt$ is bounded for any training time T, and $n_1, \ldots, n_{L-1} \to \infty$ then for any $t \in [0, T]$,

$$\Theta^{(L)} \to \Theta_\infty^{(L)} \otimes Id_{n_L}$$

Dynamics of Gradient Descent in Function Space

$$\partial_t f_t = \Phi_K \left(\langle f^* - f, \cdot \rangle_{p^{in}} \right) \text{ where, } K = \Theta_\infty^{(L)} \otimes Id_{n_L}$$

Solution:

$$f_t = f^* + e^{-t\Pi}(f_0 - f^*)$$

- Convergence to global minimum
  if Π is positive definite, as $t \to \infty$, $f_t \to f^*$ and then $C(f_t)$ converges to global minimum

Dynamics of Gradient Descent in Function Space

$$\partial_t f_t = \Phi_K \left(\langle f^* - f, \cdot \rangle_{p^{in}} \right) \text{ where, } K = \Theta_\infty^{(L)} \otimes Id_{n_L}$$

Solution:

$$f_t = f^* + e^{-t\Pi}(f_0 - f^*)$$

- Convergence to global minimum
- Motivation for early stopping
  avoid fitting the eigenfunctions of $f^* - f_0$ with lower eigenvalues