Neural Tangent Kernel

High-Dimensional Probability Analysis

Hooman Zolfaghari - Abdollah Zohrabi - Amirreza Velae

Sharif University of Technology

February 5, 2025

Emergence

- For $f_\theta : \mathbb{R}^d \to \mathbb{R}$ (NN), GD training induces:

$$\underbrace{\Theta(\theta)}_{\text{NTK}} \in \mathbb{R}^{n\times n}, \quad \Theta(\theta)_{i,j} := \nabla_\theta f_\theta(x_i)^\top \nabla_\theta f_\theta(x_j)$$

- **Dynamics** of training infinitely wide NNs $\approx$ **convex optimization** in RKHS
- **Asymptotic Property**: $\Theta(\theta^{(0)}) \to \Theta^\infty$ as width $\to \infty$

Regression Case

$$\partial_t f_t = -\Theta^{\infty}(f_t - y) \quad \text{(grad flow ODE)}$$

$$f_t = e^{-\Theta^{\infty} t} f_0 + \left(I - e^{-\Theta^{\infty} t}\right) y,$$

- Global convergence
- Linear rate for $\lambda_{\min}(\Theta^{\infty}) > 0$
- **Feature Learning Gap**: NTK regime $\neq$ real NNs (finite-width trains via $\nabla\Theta \neq 0$)

Focus Hypothesis set

- Research focused on more practical assumptions and particular settings
- Two-layer ReLU network $f_{\mathbf{W},\mathbf{a}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathrm{ReLU}(\mathbf{w}_r^\top \mathbf{x})$
- Least Squares Regression $C(W) := \frac{1}{2} \sum_{i=1}^{n} (y_i - f_{\mathbf{W},\mathbf{a}}(\mathbf{x}_i))^2$
- Resulting NTK gram matrix:

$$\mathbf{H}_{ij}^{\infty} = \mathbb{E}\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}) \left[\mathbf{x}_i^\top \mathbf{x}_j \mathbb{I} \{ \mathbf{w}^\top \mathbf{x}_i \geq 0, \mathbf{w}^\top \mathbf{x}_j \geq 0 \} \right]$$
$$= \frac{\mathbf{x}_i^\top \mathbf{x}_j \left(\pi - \arccos(\mathbf{x}_i^\top \mathbf{x}_j) \right)}{2\pi}, \quad \forall i, j \in [n].$$

- **Theorem**. If H^∞ is positive definite $\lambda_0 := \lambda_{\min}(H^\infty) > 0$, GD converges to 0 training loss w.h.p. if m is sufficiently large $\Omega(\frac{n^6}{\lambda_0^4})$.

Convergence

- Eigen-decomposition $H^\infty = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$.
- Suppose $\lambda_0 = \lambda_{\min}(H^\infty) > 0$, $\kappa = \mathcal{O}\left(\frac{\epsilon_0 \delta}{\sqrt{n}}\right)$, $\quad m = \Omega\left(\frac{n^7}{\lambda_0^4 \kappa^2 \delta^4 \epsilon^2}\right)$, $\quad \eta = \mathcal{O}\left(\frac{\lambda_0}{n^2}\right)$.
- **Theorem.** Then w.p. at least $1-\delta$ over the *random initialization*, for all $k = 0, 1, 2, \ldots$ we have

$$\|y - u(k)\|_2 = \sqrt{\sum_{i=1}^{n} \left(1 - \eta \lambda_i\right)^{2k} \left(v_i^\top y\right)^2} \pm \epsilon$$

Generalization Assumptions

- **Definition.** A distribution D over $\mathbb{R}^d \times \mathbb{R}$ is called "(λ_0, δ, n)-non-degenerate" if for n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from D, with probability at least $1-\delta$ we have

$$\lambda_{\min}\left(H^{\infty}\right) \geq \lambda_0 > 0.$$

- Fix a failure probability $\delta \in (0,1)$. Suppose our data $S = \{(x_i, y_i)\}_{i=1}^n$ are i.i.d. samples from a $(\lambda_0, \delta/3, n)$-non-degenerate distribution D, and let $\kappa = O\left(\frac{\lambda_0 \delta}{n}\right), \quad m \geq \kappa^{-2} \text{poly}\left(n, \lambda_0^{-1}, \delta^{-1}\right).$
- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,1]$ that is 1-Lipschitz in its first argument and satisfies $\ell(y,y) = 0$.

Generalization Theorem

- **Theorem** Then w.p. at least $1-\delta$ over the random initialization *and* the training samples, the network $f_{\mathbf{W}(k),a}$ trained by GD for $k \geq \Omega\left(\frac{1}{\eta\lambda_0}\log\frac{n}{\delta}\right)$ iterations has population loss:

$$L_D\left(f_{\mathbf{W}(k),a}\right) = \mathbb{E}_{(x,y)\sim D}\left[\ell\left(f_{\mathbf{W}(k),a}(x), y\right)\right] \leq \sqrt{\frac{2y^\top\left(H^\infty\right)^{-1}y}{n}} + O\left(\sqrt{\frac{\log\left(\frac{n}{\lambda_0\delta}\right)}{n}}\right)$$

Closer Look

$$L_D(f_{\mathbf{W}(k),a}) \leq \sqrt{\frac{2\, y^\top (H^\infty)^{-1} y}{n}} + O\left(\sqrt{\frac{\log\left(\frac{n}{\lambda_0 \delta}\right)}{n}}\right)$$

- We can see that the bound depends on the Distribution $(x, y) \sim \mathscr{D}$ such that,
- $\mathbf{y}^\top (H^\infty)^{-1} \mathbf{y} \leq \|(H^\infty)^{-1}\| \|\mathbf{y}\|_2 = \frac{1}{\lambda_{\min}(H^\infty)} \|\mathbf{y}\|_2$

Motivation

- What class of functions $y = g(x)$ or distributions $(x, y) \sim \mathscr{D}$ are provably learnable ?
- This depends on definition of Learnable (PAC, Agnostic-PAC etc.)
- We chose: The bound must converge to 0 as $n \to \infty$.
- The paper mentions the case of $y = g(x)$ for some function g and gives a simple statement.
- We focus on bounding $\lambda_{\min}(H^{\infty})$.
- Then we propose a (relatively small) family of $\mathscr{D}$ that is learnable. We are yet to prove the most general class.

- **Data Scaling Assumption**:
  1. $\int \|x\|_2 \, dP_X(x) = \Theta(\sqrt{d})$
  2. $\int \|x\|_2^2 \, dP_X(x) = \Theta(d)$
  3. $\int \|x - \mathbb{E}[x]\|_2^2 \, dP_X(x) = \Omega(d)$

  These are scaling conditions on the data vector x or its centered counterpart $x - \mathbb{E}[x]$.
- **Lipschitz Concentration Assumption**: The data distribution P_X satisfies the Lipschitz concentration property. For any Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, there exists a constant $c > 0$ such that:

$$\mathbb{P}\left(\left| f(x) - \int f(x') \, dP_X(x') \right| > t \right) \leq 2e^{-ct^2 / \|f\|_{\text{Lip}}^2}$$

- **General Assumption**: This assumption includes distributions satisfying the log-Sobolev inequality or log-concave densities.

Main Theorem

Theorem (Smallest eigenvalue of limiting NTK)

Let $\{x_i\}_{i=1}^N$ be a set of i.i.d. data points from P_X, where P_X has zero mean and satisfies above assumptions. Let $K^{(L)}$ be the limiting NTK recursively defined. Then, for any even integer constant $r \geq 2$, we have with probability at least

$$1 - N e^{-\Omega(d)} - N^2 e^{-\Omega\left(d N^{-2/(r-0.5)}\right)}$$

that

$$\mathrm{LO}(d) \geq \lambda_{\min}\left(K^{(L)}\right) \geq \mu_r(\sigma)^2\, \Omega(d),$$

where $\mu_r(\sigma)$ is the r-th Hermite coefficient of the ReLU function.

Estimates from Data and Theorem Application

Content Overview:

- **Useful Estimates:** The data estimations are derived from the assumptions that we have $\|x_i\|_2^2 = \Theta(d)$ for all $i \in [N]$ with probability $1 \geq Ne^{-\Omega(d)}$.
- **Key Assumptions:**
  - **Assumption 2.1 & 2.2:** $\|x_i - x_j\|^2$ is Lipschitz continuous.
  - **Lipschitz Continuity:** $\|x_i - x_j\|^2 \leq t = dN^{-1/(r-0.5)}$, where t is the bound for $|x_i - x_j|$.

Key Result:

- **Theorem 3.1 Outcome:**

$$\|x_i - x_j\|_2^2 = \Theta(d) \quad \forall i \in [N], \quad |x_i - x_j|^r \leq dN^{-1/(r-0.5)} \quad \forall i \neq j.$$

The equation holds with the same probability as stated in the theorem.

Matrix Analysis

Lemma 3.1 Application:

Define Gram matrix kernel as:

$$H \triangleq K(L) = \sum_{l=1}^{L} G(l) \circ G(l+1) \circ G(l+2) \circ \cdots \circ G(L)$$

Eigenvalue Bound:

$$\lambda_{\min}(K(L)) \geq \sum_{l=1}^{L} \lambda_{\min}(G(l))$$

Matrix Eigenvalue Estimates

Final Eigenvalue Bound:

$$\lambda_{\min}(G(2)) = \lambda_{\min}(D\,\mathbb{E}\left[\sigma(X^T w)\sigma(X^T w)^T\right] D)$$

where $D = \mathrm{diag}(\|x_i\|_2^2)$.

$$\lambda_{\min}(G(2)) \geq \mu(\sigma)\lambda_{\min}(D(X^*)^T(X^*)^T D)$$

$$\lambda_{\min}(G(2)) \geq \lambda_{\min}\left(\sum_{i\in[N]} \|x_i\|_2^2 (X^* X^T)\right)$$

At last, by Gershgorin circle theorem we have:

$$\lambda_{\min}\left((X^* r)(X^r)^T\right) \geq \min_{i\in[N]} \|x_i\|_2^{2r} - (N-1)\max_{i\neq j} \left|\langle x_i, x_j\rangle\right|^r \geq \Omega(d)$$

Similar works

- Data scaling assumption. The data distribution P_X satisfies the following properties:
  1. $\int \|x\|_2 \, dP_X(x) = \Theta(\sqrt{d})$.
  2. $\int \|x\|_2^2 \, dP_X(x) = \Theta(d)$.
  3. $\int \left\| x - \int x' \, dP_X(x') \right\|_2^2 dP_X(x) = \Omega(d)$.
- These are just scaling conditions on the data vector x or its centered counterpart $x - \mathbb{E}x$. We remark that the data can have any scaling, but in this paper we fix it to be of order d for convenience. We further assume the following condition on the data distribution.

- Lipschitz concentration assumption. The data distribution P_X satisfies the Lipschitz concentration property. Namely, for every Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, there exists an absolute constant $c > 0$ such that, for all $t > 0$,

$$\mathbb{P}\left(|f(x) - \int f(x')\, dP_X(x')| \;>\; t\right) \;\leq\; 2\, e^{-c\, t^2 / \|f\|_{\mathrm{Lip}}^2}.$$

- In general, this assumption covers the whole family of distributions that satisfy the log-Sobolev inequality with a dimension-independent constant (or distributions with log-concave densities).

Theorem (Smallest eigenvalue of limiting NTK)

Let $\{x_i\}_{i=1}^N$ be a set of i.i.d. data points from P_X, where P_X has zero mean and satisfies Assumptions 2.1 and 2.2. Let $K^{(L)}$ be the limiting NTK recursively defined in (9). Then, for any even integer constant $r \geq 2$, we have with probability at least

$$1 - N e^{-\Omega(d)} - N^2 e^{-\Omega\left(d N^{-2/(r-0.5)}\right)}$$

that

$$\mathrm{LO}(d) \geq \lambda_{\min}\left(K^{(L)}\right) \geq \mu_r(\sigma)^2 \, \Omega(d),$$

where $\mu_r(\sigma)$ is the r-th Hermite coefficient of the ReLU function given by (8).

Gershgorin Circle Theorem

Statement: Let $A = [a_{ij}]$ be an $n \times n$ matrix. The eigenvalues of A lie within the union of disks D_i in the complex plane, centered at a_{ii} with radius $\sum_{j \neq i} |a_{ij}|$:

$$D_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

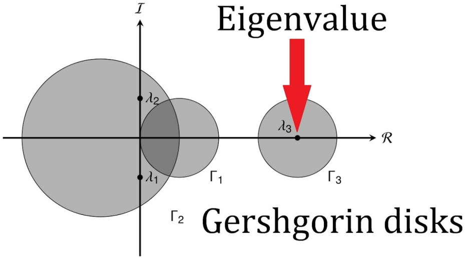

Figure 1: Gershgorin Circle Theorem

- Most papers and our main reference assume $\|x_i\| = 1$ and $|y_i| \leq 1$, for simplicity.
- The diagonal $H^{\infty}_{ii} = \frac{1}{2}$.
- Denote $\rho = \max_{i,j\neq i} |x_i^{\top} x_j|$. Considering $ft(t) = \frac{t(\pi - \arccos(t))}{2\pi}$, we get:

$$H^{\infty}_{i,j\neq i} \leq \frac{\rho(\pi - \arccos(\rho))}{2\pi} \leq \frac{1}{2}$$

- We find the **Gershgorin circle** theorem and get

$$\lambda_{\min}(H^{\infty}) \geq \frac{1}{2} - (n-1)\frac{\rho(\pi - \arccos(\rho))}{2\pi}$$

Examples

- Thus, the bound depends on the maximum "correlation" between two distinct i.i.d x_i in a sample of size n. This will depend on the distribution of x on S^{d-1}.
- The problem is now bounding ρ such that:
- $x_i \sim \mathrm{Unif}(S^{d-1})$:

$$\max_{1\le i<j\le n} \left|\langle X_i, X_j\rangle\right| \le C\sqrt{\frac{\log(\frac{n}{\delta})}{d}}$$

- x_i are isotropic, mean-zero, sub-Gaussian vectors:

$$\max_{1\le i<j\le n} |\langle X_i, X_j\rangle| \le C\sqrt{\frac{\log(\frac{n}{\delta})}{d}} \max_{1\le i\le n} \|X_i\| .$$

Learnable Distributions

- $\mathbf{y}$ has sub-Gaussian Coordinates and $\mathbb{E}[y_i^2] = 2C^2$
- So w.h.p. $\|y\|_2 \leq C\sqrt{n}$
- We need $\lambda_0 \|y\|_2 \leq Cn$, so we want:

$$\lambda_0 \leq C\sqrt{n} \implies \frac{1}{2} - (n-1)\frac{\rho(\pi - \arccos(\rho))}{2\pi} \geq \frac{1}{C\sqrt{n}}$$

- For $0 \leq \rho \leq 1$ gives a computable bound. approximately $O(\frac{1}{n})$.
- So we need distributions on x_i such that $\sup_{i,j} |\langle x_i, x_j \rangle| = O(\frac{1}{n})$ for n i.i.d samples w.h.p.

Weyl's Inequality

Weyl's Inequality provides bounds on the eigenvalues of the sum of two Hermitian matrices.

Statement: Let A and B be $n \times n$ Hermitian matrices with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ for A and $\mu_1 \leq \cdots \leq \mu_n$ for B. Then the eigenvalues ν_i of the matrix $A + B$ satisfy:

$$\lambda_i + \mu_j \leq \nu_{i+j-1} \leq \lambda_i + \mu_j \quad \text{for all} \quad 1 \leq i, j \leq n.$$

Define:

$$A = H - \mathbb{E}[H]$$

- Note that it can be shown that:

$$\mathbb{E}[H] = (\frac{1}{2} - s_d)\mathbf{1} + s_d J$$

  and hence $\lambda_{\min}(\mathbb{E}[H]) = \frac{1}{2} - s_d$.
- Now we can take adventage of Weyl's inequality to bound the smallest eigenvalue of H^∞ via:

$$\lambda_{\min}(H^\infty) \geq \lambda_{\min}(A) + \lambda_{\min}(\mathbb{E}[H]) = \lambda_{\min}(A) + \frac{1}{2} - s_d$$

Main Observations

- Let $\mathbb{P}\left[|H - \mathbb{E}[H]|_{op} \geq t\right] \leq \delta_t$. Then, with probability at least $1-\delta_t$, we have:

$$\lambda_{\min}(H^{\infty}) \geq -t + \frac{1}{2} - s_d$$

or equivalently:

$$\lambda_{\min}(H^{\infty}) \geq max\left\{-t + \frac{1}{2} - s_d, 0\right\}$$

- We know that $|H - \mathbb{E}[H]|_{op}$ grows linearly with N. Hence we can make a simple observation that one should have $O(d)$ samples to have a non-zero lower bound on the smallest eigenvalue of H^{∞}.

Future Directions & Other Ideas