Divergence of Neural Tangent Kernel in Classification Problems

Zixiong Yu [1,2] Songtao Tian [1] Guhan Chen [1]

[1]Tsinghua University [2]Huawei Technologies Ltd.

TSINGHUA UNIVERSITY

Background: Problem Setup

- Notations:
  - For sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ if $|x_n| \le C|y_n|$ for some constant $C > 0$ when n is large.
  - $i \in [n]$ denotes $i \in \{1, 2, \ldots, n\}$.
  - Sample matrix $X = (x_1^T, \ldots, x_n^T)^T$, labels $Y = (y_1, \ldots, y_n)^T$.
  - $f(X) = (f(x_1), \ldots, f(x_n))^T$ applies f entry-wise.
- Binary Classification:

$$(x, y) \in \mathcal{X} \times \{0, 1\}, \quad \mathcal{X} \subset \mathbb{R}^d \text{ is compact},$$

with unknown distribution, no separability assumptions.
- Neural Networks:
  - Fully Connected Network (FCN):

$$\begin{aligned} &\alpha^{(0)}(x) = x, \\ &\alpha^{(l)}(x) = \sqrt{\frac{2}{m_l}}\,\sigma\big(W^{(l)}\alpha^{(l-1)}(x) + b^{(l)}\big), \quad l = 1, \ldots, L, \\ &f(x;\theta) = W^{(L+1)}\alpha^{(L)}(x), \end{aligned}$$

where $\sigma(x) = \max(0, x)$, $W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$, $b^{(l)} \in \mathbb{R}^{m_l}$, etc.
  - Residual Network (ResNet):

$$\begin{aligned} &\alpha^{(0)}(x) = \sqrt{\frac{1}{m_0}}(Ax + b), \\ &\widetilde{\alpha}^{(l)}(x) = \sqrt{\frac{2}{m_l}}\,\sigma\big(W^{(l)}\alpha^{(l-1)}(x) + b^{(l)}\big), \\ &\alpha^{(l)}(x) = \alpha^{(l-1)}(x) \;+\; \alpha\sqrt{\frac{1}{m_l}}\big(V^{(l)}\widetilde{\alpha}^{(l)}(x) + d^{(l)}\big), \\ &f(x;\theta) = W^{(L+1)}\alpha^{(L)}(x), \end{aligned}$$

with $m_0 = \cdots = m_L$, and $A \in \mathbb{R}^{m_0 \times d}$, etc.
- Initialization: FCN: $W^{(l)}, b^{(l)} \sim \mathcal{N}(0, 1)$. ResNet: $W^{(l)}, V^{(l)} \sim \mathcal{N}(0, 1)$, $b^{(l)} = d^{(l)} = 0$.
- Loss Function: Cross-Entropy:

$$\begin{aligned} &\mathcal{L}(\theta) = -\sum_{i=1}^{n}\Big[y_i \ln o_i + (1 - y_i)\ln(1 - o_i)\Big], \\ &o_i = \frac{1}{1 + \exp\big(-f(x_i;\theta)\big)}, \quad \ell(x) = \ln\big(1 + e^{-x}\big). \end{aligned}$$

Residual $u_i = |o_i - y_i|$.

Neural Tangent Kernel (NTK)

- NNK/Empirical NTK:

$$K_t(x, x') = \big\langle \nabla_\theta f(x;\theta_t), \nabla_\theta f(x';\theta_t)\big\rangle,$$

$$\frac{d}{dt}f(x;\theta_t) \;=\; \sum_{i=1}^{n} K_t(x, x_i)\,\big(2y_i - 1\big)\,u_i.$$

- NTK Regime : Previous literature shows in regression problem: As $m \to \infty$, $K_t(x, x') \to K(x, x')$ with high probability.
  - FCN NTK (defined recursively):

$$\Theta^{(l+1)}(x, x') = \Theta^{(l)}(x, x')\,\dot{\Sigma}^{(l+1)}(x, x') + \Sigma^{(l+1)}(x, x'),$$

so $K^{\mathrm{FC}}(x, x') = \Theta^{(L+1)}(x, x')$.
  - ResNet NTK (defined on the explicit formula of homogeneous $r^{(L)}$):

$$K^{\mathrm{Res}}(x, x') = \|\tilde{x}\|\; r^{(L)}\big(\phi(x), \phi(x')\big)\; \|\tilde{x}'\|.$$

- Positive Definiteness: Proposition: $K(x, x')$ is strictly positive definite ($\lambda_{\min}(K) > 0$).

Main Results

Divergence of Network. At standard NTK initialization, the NNK converges in probability as width grows *independently of the loss function*. In regression, it remains convergent throughout training, forming the basis of NTK theory. However, under cross-entropy loss for classification, this "NTK regime" no longer holds.

We denote by

$$\widetilde{\lambda}_0(t) \;=\; \lambda_{\min}\big(K_t(X, X)\big).$$

Theorem (Divergence of Network)

Fix the training samples $\{(x_i, y_i)\}_{i \in [n]}$. For both fully-connected and residual networks in classification with cross-entropy loss, if $\widetilde{\lambda}_0(t)$ stays above a positive constant C during training, then

$$\lim_{t \to \infty} |f_t(x_i)| = +\infty \quad \textit{for all } i \in [n].$$

Interpretation. This theorem shows that if the NNK matrix remains uniformly strictly positive definite ($\widetilde{\lambda}_0(t) \ge C > 0$), then the network outputs $|f_t(x_i)|$ at the samples will diverge to infinity. Equivalently, some parameters must deviate drastically from their initial values as training progresses.

Corollary (Failure of the NTK regime)

Under Theorem 1's conditions, for any initial parameter θ_0,

$$\limsup_{t \to \infty} \|\theta_t - \theta_0\|_\infty \;=\; \infty.$$

Note that the network width m does not prevent this divergence. The key premise is $\widetilde{\lambda}_0(t) \ge C > 0$, which would be satisfied if NNK uniformly converged to NTK — precisely the assumption contradicted in our next theorem.

Divergence of NTK. In regression cases, if m is large enough, the NNK matrix converges uniformly to a fixed NTK, enabling kernel methods for generalization. In classification, **NTK does not remain fixed**: it evolves over time, signaling the breakdown of classical NTK theory for classification.

Theorem (Divergence of NTK)

Let λ_0 be the minimum eigenvalue of the (FCN or ResNet) NTK at initialization, e.g. $\lambda_0 = \lambda_{\min}(K^{\mathrm{FC}}(X, X))$. Under cross-entropy training, there exist $x, x' \in \mathcal{X}$ such that

$$\sup_{t \ge 0}\Big|K_t^{\mathrm{FC}}(x, x') - K^{\mathrm{FC}}(x, x')\Big| \;\ge\; \frac{\lambda_0}{2n^2},$$

and similarly for K^{Res}.

Significance. Unlike regression, where one can make $\sup_{t \ge 0}|K_t(x, x') - K(x, x')| \to 0$ by large width, here a persistent gap $\frac{\lambda_0}{2n^2}$ remains. This means the evolving NNK $\neq$ fixed NTK, thus we cannot directly approximate a trained network by a static kernel. New methods are hence required for analyzing classification tasks.

Experiments

Synthetic Data: Divergence of the Fully Connected Network Function The fully connected network has three hidden layers, with an input dimension of $d = 2$ and a width of $m = 2000$. For the training set, six input vectors are spaced on the unit sphere $\mathbb{S}^d$, with interleaved labels to eliminate the influence of data separability. The network is trained for 10,000 epochs with a learning rate of 0.1. The output values of the network at the six training points are plotted during training.

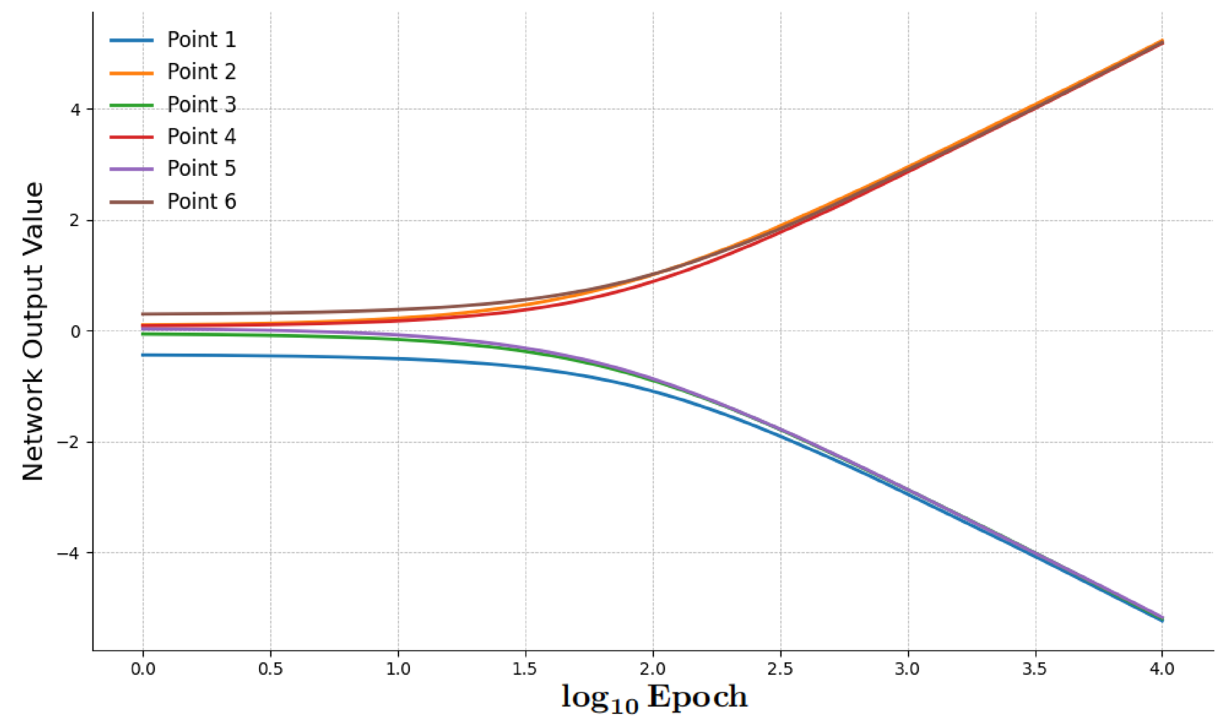

Figure: The output values of the network at the six training points over the training process. Despite starting from interleaved labels, the network function diverges at all six points.

Real Data: MNIST Experiment A four-layer fully connected neural network with a width of $m = 500$ was trained on the MNIST dataset for 100,000 epochs with a learning rate of 0.5. The NTK values for three selected points were computed, showing disconvergence during training.

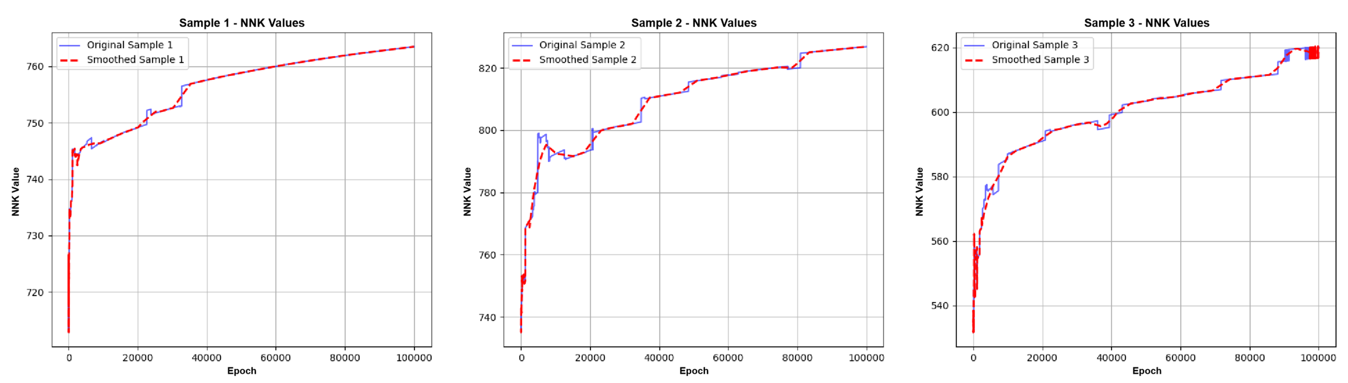

Figure: The NTK values on the MNIST dataset for three selected points. The blue lines represent the original NTK values, while the red dashed lines show the smoothed values.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

[2] Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. *Journal of Machine Learning Research*, 25(82):1–47, 2024.

Corresponding author: Guhan Chen (chen-gh23@mails.tsinghua.edu.cn)