Neural Tangent Kernel

Convergence and Generalization of DNNs

Arthur Jacot,
Franck Gabriel, Berfin Şimşek, Francesco Spadaro, Clément Hongler

Ecole Polytechnique Fédérale de Lausanne

July 15, 2020

Neural Networks

- $L+1$ layers of n_ℓ neurons with activations $\alpha^{(\ell)}(x) \in \mathbb{R}^{n_\ell}$

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

  - Parameters $\theta = (W^{(0)}, b^{(0)}, \dots, W^{(L-1)}, b^{(L-1)})$:
    - connections weights $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell+1}}$ and bias $b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$.
  - Weights / bias balance: β.
  - Non-linearity: $\sigma : \mathbb{R} \to \mathbb{R}$.
- Network function $f_\theta(x) = \tilde{\alpha}^{(L)}(x)$.

Initialization: DNNs as Gaussian processes

- In the infinite width limit $n_1, ..., n_{L-1} \to \infty$.
- Initialize the parameters $\theta \sim \mathcal{N}(0, Id_P)$.
- The preactivations $\tilde{\alpha}_i^{(\ell)}(\cdot;\theta) : \mathbb{R}^{n_0} \to \mathbb{R}$ converge to iid Gaussian processes of covariance $\Sigma^{(\ell)}$ (Lee et al., 2018; Neal, 1996):

$$\Sigma^{(1)}(x,y) = \frac{1}{n_0} x^T y + \beta^2$$

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

- The network function $f_\theta = \tilde{\alpha}^{(L)}$ is also asymptotically Gaussian.

Training: Neural Tangent Kernel

- Training set $X = (x_1, \ldots, x_N)$ and outputs $Y_\theta = (f_\theta(x_1), \ldots, f_\theta(x_N))$.
- Convex cost $C(Y)$ defined on labels $Y \in \mathbb{R}^N$.
- Gradient descent on (non-convex) $\theta \mapsto C(Y_\theta)$

$$\partial_t \theta = -\nabla C(Y_\theta) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_\theta(x_i) \partial_{Y_i} C(Y_\theta).$$

Training: Neural Tangent Kernel

- Training set $X = (x_1, \ldots, x_N)$ and outputs $Y_\theta = (f_\theta(x_1), \ldots, f_\theta(x_N))$.
- Convex cost $C(Y)$ defined on labels $Y \in \mathbb{R}^N$.
- Gradient descent on (non-convex) $\theta \mapsto C(Y_\theta)$

$$\partial_t \theta = -\nabla C(Y_\theta) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_\theta(x_i) \partial_{Y_i} C(Y_\theta).$$

- Evolution of f_θ:

$$\partial_t f_\theta(x) = (\nabla f_\theta(x))^T \partial_t \theta = \frac{1}{N} \sum_{i=1}^{N} \underbrace{(\nabla f_\theta(x))^T \nabla f_\theta(x_i)}_{\Theta^{(L)}(x, x_i)} \partial_{Y_i} C(Y_\theta).$$

- Neural Tangent Kernel (NTK):

$$\Theta^{(L)}(x, y) := (\nabla f_\theta(x))^T \nabla f_\theta(y).$$

Asymptotics of the NTK

Theorem

As $n_1, \dots, n_{L-1} \to \infty$, there exist a fixed deterministic limiting kernel $\Theta^{(L)}_\infty$ s.t.

$$\Theta^{(L)}(t) \to \Theta^{(L)}_\infty.$$

Asymptotic dynamics:

$$f_{\theta(0)} \sim \mathcal{N}(0, \Sigma^{(L)})$$

$$\partial_t f_{\theta(t)}(x) = \frac{1}{N} \sum_{i=1}^{N} \Theta^{(L)}_\infty(x, x_i) \partial_{Y_i} C(f_\theta(X))$$

Asymptotics of the NTK

Theorem

As $n_1, \ldots, n_{L-1} \to \infty$, there exist a fixed deterministic limiting kernel $\Theta_\infty^{(L)}$ s.t.

$$\Theta^{(L)}(t) \to \Theta_\infty^{(L)}.$$

Asymptotic dynamics:

$$f_{\theta(0)} \sim \mathcal{N}(0, \Sigma^{(L)})$$

$$\partial_t f_{\theta(t)}(x) = \frac{1}{N} \sum_{i=1}^{N} \Theta_\infty^{(L)}(x, x_i) \partial_{Y_i} C(f_\theta(X))$$

Guarantee of convergence: NTK Gram matrix $\Theta_\infty^{(L)}(X, X)$

$$\partial_t C(f_\theta(X)) = -\left(\nabla C\right)^T \Theta_\infty^{(L)}(X, X) \nabla C \leq -\lambda_0 \left\|\nabla C\right\|^2.$$

Asymptotics of the NTK

1. First proof Jacot et al., 2018: sequential limit $n_1 \to \infty, \ldots, n_{L-1} \to \infty$.
2. Simultaneous limit ($n_1 = n_{L-1} = w \to \infty$), finite width bounds Arora et al., 2019; Lee et al., 2019

$$\left|\Theta^{(L)}(0) - \Theta^{(L)}_{\infty}\right| = O(w^{-\frac{1}{2}})$$
$$\left|\Theta^{(L)}(0) - \Theta^{(L)}(t)\right| = O(w^{-\frac{1}{2}}).$$

3. Tight rates Huang and Yau, 2019

$$\left|\Theta^{(L)}(0) - \Theta^{(L)}(t)\right| = O(w^{-1}).$$

MSE Loss

MSE loss $C(Y) = \frac{1}{N} \| Y - Y^* \|^2$ for some true labels Y^*.

1. Linear ODE on the training set

$$\partial_t Y_{\theta(t)} = \frac{2}{N} \Theta_\infty^{(L)}(X, X) \left(Y^* - Y_{\theta(t)} \right).$$

2. Solution: $f_{\theta(t)}$ is Gaussian for all t with mean

$$\mathbb{E}\left[f_\theta(x) \right] = \Theta_\infty^{(L)}(x, X) \left(\Theta_\infty^{(L)}(X, X) \right)^{-1} \left(I_N - e^{-\frac{2t}{N} \Theta_\infty^{(L)}(X,X)} \right) Y^*.$$

MSE Loss

MSE loss $C(Y) = \frac{1}{N} \| Y - Y^* \|^2$ for some true labels Y^*.

1. Linear ODE on the training set

$$\partial_t Y_{\theta(t)} = \frac{2}{N} \Theta_\infty^{(L)}(X, X) \left(Y^* - Y_{\theta(t)}\right).$$

2. Solution: $f_{\theta(t)}$ is Gaussian for all t with mean

$$\mathbb{E}\left[f_\theta(x)\right] = \Theta_\infty^{(L)}(x, X) \left(\Theta_\infty^{(L)}(X, X)\right)^{-1} \left(I_N - e^{-\frac{2t}{N}\Theta_\infty^{(L)}(X,X)}\right) Y^*.$$

1. As $t \to \infty$ the mean converges to the ridgless kernel predictor w.r.t. the NTK

$$\Theta_\infty^{(L)}(x, X) \left(\Theta_\infty^{(L)}(X, X)\right)^{-1} Y^*.$$

"Wide DNNs perform NTK Kernel Ridge Regression"

Kernel Ridge Regression

- Random inputs $x \sim \mathcal{D}$ in a compact domain Ω.
- Labels $Y_i^* = f^*(x_i) + \epsilon e_i$ for $e_i \sim \mathcal{N}(0, 1)$.
- For a kernel K and ridge $\lambda > 0$, the KRR predictor is

$$\hat{f}_\lambda(x) = K(x, X)\left(K(X, X) + \lambda I_N\right)^{-1} Y^*$$

Kernel Ridge Regression

- Random inputs $x \sim \mathcal{D}$ in a compact domain Ω.
- Labels $Y_i^* = f^*(x_i) + \epsilon e_i$ for $e_i \sim \mathcal{N}(0,1)$.
- For a kernel K and ridge $\lambda > 0$, the KRR predictor is

$$\hat{f}_\lambda(x) = K(x,X)\left(K(X,X) + \lambda I_N\right)^{-1} Y^*$$

- Risk $R(\hat{f}_\lambda) = \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\hat{f}_\lambda(x) - f^*(x)\right)^2\right] + \epsilon^2 = \left\|\hat{f}_\lambda - f^*\right\|_{\mathcal{D}}^2 + \epsilon^2$.
- Empirical Risk $\hat{R}(\hat{f}_\lambda) = \frac{1}{N}\left\|\hat{Y}_\lambda - Y^*\right\|^2$.

Objects of interest

- Random Sampling operator $\mathcal{O}(f) = (f(x_1), ..., f(x_N))^T$ from $\mathcal{F}$ to $\mathbb{R}^N$.
- Noiseless predictor $\epsilon = 0$ (for $K : \mathcal{F}^* \to \mathcal{F}$ and $\mathcal{O}^T : \mathbb{R}^N \to \mathcal{F}^*$):

$$\hat{f}_\lambda = \frac{1}{N} K \mathcal{O}^T \left(\frac{1}{N} \mathcal{O} K \mathcal{O}^T + \lambda I_N \right)^{-1} \mathcal{O} f^*$$

Objects of interest

- Random Sampling operator $\mathcal{O}(f) = (f(x_1), ..., f(x_N))^T$ from $\mathcal{F}$ to $\mathbb{R}^N$.
- Noiseless predictor $\epsilon = 0$ (for $K : \mathcal{F}^* \to \mathcal{F}$ and $\mathcal{O}^T : \mathbb{R}^N \to \mathcal{F}^*$):

$$
\begin{aligned}
\hat{f}_\lambda &= \frac{1}{N} K \mathcal{O}^T \left(\frac{1}{N} \mathcal{O} K \mathcal{O}^T + \lambda I_N \right)^{-1} \mathcal{O} f^* \\
&= \frac{1}{N} K \mathcal{O}^T \mathcal{O} \left(\frac{1}{N} K \mathcal{O}^T \mathcal{O} + \lambda I_{\mathcal{F}} \right)^{-1} f^*
\end{aligned}
$$

Objects of interest

- Random Sampling operator $\mathcal{O}(f) = (f(x_1), ..., f(x_N))^T$ from $\mathcal{F}$ to $\mathbb{R}^N$.
- Noiseless predictor $\epsilon = 0$ (for $K : \mathcal{F}^* \to \mathcal{F}$ and $\mathcal{O}^T : \mathbb{R}^N \to \mathcal{F}^*$):

$$\begin{aligned}
\hat{f}_\lambda &= \frac{1}{N} K \mathcal{O}^T \left(\frac{1}{N} \mathcal{O} K \mathcal{O}^T + \lambda I_N \right)^{-1} \mathcal{O} f^* \\
&= \frac{1}{N} K \mathcal{O}^T \mathcal{O} \left(\frac{1}{N} K \mathcal{O}^T \mathcal{O} + \lambda I_{\mathcal{F}} \right)^{-1} f^* \\
&\underset{N\to\infty}{\to} \underbrace{T_K \left(T_K + \lambda I_{\mathcal{F}}\right)^{-1}}_{\tilde{A}_\lambda} f^*
\end{aligned}$$

for the *integral operator* $(T_K f)(x) = \mathbb{E}_{w\sim\mathcal{D}}\left[K(x,w)f(w)\right]$.

- Mercer's Theorem:
  - T_K has eigenvalues d_k and eigenfunctions $f^{(k)}$.
  - T_K is trace class $\sum_{k=1}^{\infty} d_k < \infty$.

Expected Predictor

Theorem (Jacot et al., 2020)

For $\lambda > 0$ we have

$$\mathbb{E}\left[\hat{f}_\lambda(x)\right] \approx \tilde{A}_\vartheta f^* = T_K\left(T_K + \vartheta I_{\mathcal{F}}\right)^{-1} f^*$$

where the Signal Capture Threshold $\vartheta(\lambda, N, T_K)$ is the unique positive solution of

$$\vartheta = \lambda + \frac{\vartheta}{N}\mathrm{Tr}\left[T_K(T_K + \vartheta I_{\mathcal{F}})^{-1}\right].$$

Expected Predictor

Theorem (Jacot et al., 2020)

For $\lambda > 0$ we have

$$\mathbb{E}\left[\hat{f}_\lambda(x)\right] \approx \tilde{A}_\vartheta f^* = T_K\left(T_K + \vartheta I_{\mathcal{F}}\right)^{-1} f^*$$

where the Signal Capture Threshold $\vartheta(\lambda, N, T_K)$ is the unique positive solution of

$$\vartheta = \lambda + \frac{\vartheta}{N}\mathrm{Tr}\left[T_K(T_K + \vartheta I_{\mathcal{F}})^{-1}\right].$$

For $f^* = \sum_k b_k f^{(k)}$ we have $\mathbb{E}\left[\hat{f}_\lambda(x)\right] \approx \sum_k \frac{d_k}{d_k+\vartheta} b_k f^{(k)}$:

- When $d_k \gg \vartheta$, $\frac{d_k}{d_k+\vartheta} \simeq 1 \implies$ signal is captured.
- When $d_k \ll \vartheta$,$\frac{d_k}{d_k+\vartheta} \simeq 0 \implies$ signal is lost.

Expected Risks

Theorem

$$
\begin{aligned}
R\left(\mathbb{E}\left[\hat{f}_\lambda\right]\right) &\approx \left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta)f^*\right\|_{\mathcal{D}}^2 + \epsilon^2 \\
\mathbb{E}\left[R\left(\hat{f}_\lambda\right)\right] &\approx \partial_\lambda \vartheta \left(\left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta)f^*\right\|_{\mathcal{D}}^2 + \epsilon^2\right).
\end{aligned}
$$

For $f^* = \sum_k b_k f^{(k)}$, $\left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta)f^*\right\|_{\mathcal{D}}^2 = \sum_k \frac{\vartheta^2}{(d_k+\vartheta)^2} b_k^2$.

Expected Risks

Theorem

$$R\left(\mathbb{E}\left[\hat{f}_\lambda\right]\right) \approx \left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta) f^*\right\|_{\mathcal{D}}^2 + \epsilon^2$$

$$\mathbb{E}\left[R\left(\hat{f}_\lambda\right)\right] \approx \partial_\lambda \vartheta \left(\left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta) f^*\right\|_{\mathcal{D}}^2 + \epsilon^2\right).$$

For $f^* = \sum_k b_k f^{(k)}$, $\left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta) f^*\right\|_{\mathcal{D}}^2 = \sum_k \frac{\vartheta^2}{(d_k+\vartheta)^2} b_k^2$.

Theorem

$$\mathbb{E}\left[\hat{R}\left(\hat{f}_\lambda\right)\right] \approx \partial_\lambda \vartheta \frac{\lambda^2}{\vartheta^2} \left(\left\|(I_{\mathcal{F}} - \tilde{A}_\vartheta) f^*\right\|_{\mathcal{D}}^2 + \epsilon^2\right).$$

$\Longrightarrow$ relation $R\left(\hat{f}_\lambda\right) \approx \frac{\vartheta^2}{\lambda^2} \hat{R}\left(\hat{f}_\lambda\right)$.

Kernel Alignement Risk Estimator

Proposition

$$\vartheta \approx \frac{1}{\frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}K(X,X)+\lambda I_N\right)^{-1}\right]}.$$

Kernel Alignement Risk Estimator

Proposition

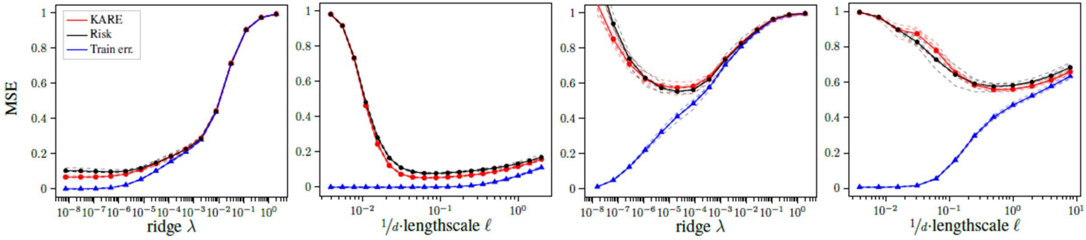
$$\vartheta \approx \frac{1}{\frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}K(X,X)+\lambda I_N\right)^{-1}\right]}.$$

Kernel Alignement Risk Estimator (KARE)

$$R\left(\hat{f}_\lambda\right) \approx \frac{\frac{1}{N}\left(Y^*\right)^T\left(\frac{1}{N}K(X,X)+\lambda I_N\right)^{-2}Y^*}{\left(\frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}K(X,X)+\lambda I_N\right)^{-1}\right]\right)^2}.$$

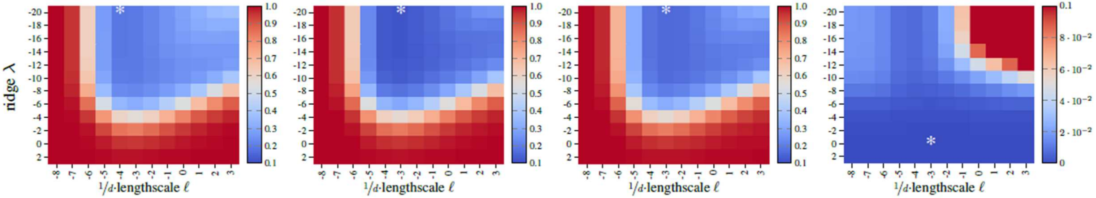
Bias term is approximated by $\frac{\frac{1}{N}(Y^*)^T\left(\frac{1}{N}K(X,X)+\lambda I_N\right)^{-2}Y^*}{\frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}K(X,X)+\lambda I_N\right)^{-2}\right]}$.

Kernel Alignement Risk Estimator



(a) MNIST, $\ell = d$ (b) MNIST, $\lambda = 10^{-5}$ (c) Higgs, $\ell = d$ (d) Higgs, $\lambda = 10^{-4}$

Kernel Alignement Risk Estimator



(a) Risk (b) KARE Predictions (c) Cross Val. Predictions (d) Log-likelihood Estim.

Conclusion

1. Wide networks perform Kernel Ridge Regression w.r.t. the NTK.
2. Convergence is guaranteed whenever the NTK is positive definite.
3. Generalization for a general Kernel:
    1. The SCT describes which components are learned.
    2. The test loss can be predicted from the training data using the KARE.

Bibliography I

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*.

Huang, J. and Yau, H.-T. (2019). Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*.

Jacot, A., Şimşek, B., Spadaro, F., Hongler, C., and Gabriel, F. (2020). Kernel alignment risk estimator: Risk prediction from training data.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589. Curran Associates, Inc.

Bibliography II

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8572–8583.

Lee, J. H., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2018). Deep Neural Networks as Gaussian Processes. *ICLR*.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.