Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot-Guillarmod* Franck Gabriel* Clément Hongler* *EPFL

The Training of Artificial Neural Networks (ANNs) involves a priori highly non-convex optimization, yet in practice one can obtain impressive convergence and generalization.

Question How to understand mathematically the dynamics of the training of an ANN as it becomes large ?

Thanks to a new object the Neural Tangent Kernel, we can precisely describe the evolution, convergence, and generalization of ANNs of large width

Theorem In the infinite-width limit, the ANN function follows a kernel gradient descent with respect to the limiting NTK

1. ANN Training: Architecture and Optimization

- Fully connected neural net with $L+1$ layers of widths $n_0,\dots,n_L$ and nonlinearity $\sigma:\mathbb{R}\to\mathbb{R}$

$$f_\theta:\mathbb{R}^{n_0}\to\mathbb{R}^{n_L}$$

$$x=z^{(0)}(x)\to\tilde z^{(1)}(x)\to z^{(1)}(x)\to\cdots\to\tilde z^{(L)}(x)\to z^{(L)}(x)=f_\theta(x)$$

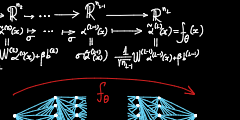
$$\tilde z^{(\ell+1)}=\frac{1}{\sqrt{n_\ell}}W^{(\ell)}z^{(\ell)}+\beta b^{(\ell)},\qquad z^{(\ell)}=\sigma(\tilde z^{(\ell)})$$

Network params $\theta=(W^{(1)},\dots,W^{(L)},b^{(1)},\dots,b^{(L)})\in\mathbb{R}^P$ ($P=\sum_{\ell=0}^{L-1}n_{\ell+1}(n_\ell+1)$); σ pointwise; β regularization

- Gaussian Initialization: $\theta_\mu\overset{iid}{\sim}\mathcal{N}(0,1)$
- Training Set: $(x_i,z_i)_{i=1,\dots,N}\in\mathbb{R}^{n_0}\times\mathbb{R}^{n_L}$ ($i\in\{1,\dots,N\}$; Training Inputs; Training Outputs)
- Training Loss: $C(\theta)=\frac{1}{N}\sum_{i=1}^{N}c(f_\theta(x_i),z_i)$

Goal: Minimize $C(\theta)$ with Gradient Descent

- Gradient Descent Step: $\theta\mapsto\theta-\gamma\nabla C(\theta)\rightsquigarrow$ Flow: $\partial_t\theta(t)=-\nabla C(\theta)$
- Inference Flow: $f_{\theta(t)}(x)\xrightarrow{t\to\infty}$? for $x\in$ {training set, test set}



2. ANN Training: Functional Formulation

Function Space

- Neural Realization Function: $F^{(L)}:\mathbb{R}^P\to\mathcal{F}$, $\theta\mapsto\left(f_\theta:\mathbb{R}^{n_0}\to\mathbb{R}^{n_L}\right)$
- Neural Tangent Kernel: $\Theta_{\theta;ij}(x,y)=\sum_{p=1}^{P}\partial_{\theta_p}f_{\theta,i}(x)\,\partial_{\theta_p}f_{\theta,j}(y)$ ($\in\mathbb{R}^{n_L\times n_L}$; $\in$ Tangent of $F^{(L)}(\mathbb{R}^P)$ at f_θ; ∇C; dF)
- Initialization: $F^{(L)}(\theta)\in\mathcal{F}$ (Non-Gaussian; C level sets; $\in$ level sets)
- Functional Cost $\mathcal{C}:\mathcal{F}\to\mathbb{R}$, $f\mapsto\frac{1}{N}\sum_{i=1}^{N}c(f(x_i),z_i)$ (Convex Nonlinear; Functional Derivative of $\mathcal{C}$ at f_θ; $\in\mathcal{F}$)
- Training Loss: $C=\mathcal{C}\circ F^{(L)}$ Chain Rule $\partial_\theta C(\theta)=D\mathcal{C}\big|_{f_\theta}\partial_\theta F^{(L)}(\theta)$
- Gradient Descent Step: $f_\theta\mapsto f_{\theta-\gamma\nabla C(\theta)}\simeq f_\theta-\gamma\sum_{p=1}^{P}\partial_{\theta_p}C(\theta)\,\partial_{\theta_p}F(\theta)$

$\rightsquigarrow$ Flow: $\partial_t f_{\theta(t)}=-\nabla_{\Theta_{\theta(t)}}\mathcal{C}\big|_{f_{\theta(t)}}=-\nabla_{\Theta_\theta}\mathcal{C}\big|_{f_\theta}$

Kernel Gradient: $\nabla^{\Theta_\theta}_{f_\theta}\mathcal{C}(f_\theta)=\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_L}\Theta_{ij}(x_i,x_k)\frac{\partial}{\partial f_{\theta,j}(x_i)}\mathcal{C}\big|_{f_\theta}\right)_{k=1,\dots,N}$

3. Infinite-Width Limit: $n_1,\dots,n_{L-1}\to\infty$

- $\left(\Theta_{\theta(0)}\right)_{ij;x,y}\to\mathcal{Z}$ centered Gaussian Process with $Cov(\mathcal{Z}_{ij}(x),\mathcal{Z}_{i'j'}(y))=\Sigma(x,y)\delta_{ij}$ (Neal, 1996 [2]; Explicitly Computable)
- $H_{\theta(0)}\to H^{\infty}\otimes\mathrm{Id}_{n_L}$, $\left(H^{\infty}\otimes\mathrm{Id}_{n_L}\right)(x,y)=\left(H^{\infty}(x,y)\right)\delta_{ij}$

} At Initialization

- $\left|\Theta_{\theta(w)}-\Theta_{\theta(0)}\right|\xrightarrow[n_1,\dots,n_{L-1}\to\infty]{\text{pointwise}}0\rightsquigarrow\partial_t f_{\theta(t)}=-\nabla_{\Theta^{\infty}\otimes\mathrm{Id}_{n_L}}\mathcal{C}\big|_{f_{\theta(t)}}$

} During Training

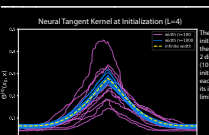
4 Consequences as $n_1,\dots,n_{L-1}\to\infty$

Optimal Emergent Learning

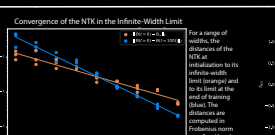
- Individually, the weights evolve less and less: $|\theta_t(0)-\theta(0)|\xrightarrow{n_1,\dots,n_{L-1}\to\infty}0$ … yet learning occurs at each layer
- The influence of (x_i,z_i) on the prediction $f_\theta(x)$ for a test point x can be understood using the NTK (Training data)
- Convergence to global min of $\mathcal{C}$ is guaranteed if Θ_∞ is positive-definite (Proven for $\|x_i\|=\frac{1}{n_0}\|z_i\|=1$)

Least-Squares Regression

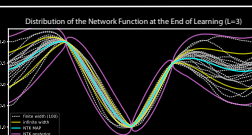
- $f_{\theta(0)}$ is Gaussian for any $t\geq 0$
- $\mathbb{E}[f_{\theta(t)}]\xrightarrow{n_1,\dots,n_{L-1}\to\infty}$ Bayesian Max A Posteriori (wrt. $\mathcal{N}(0,\Theta^{\infty}\otimes\mathrm{Id})$ prior) $=$ Kernel Ridge Regression (with $\lambda\to 0$: regularization)
- Learning governed by NTK PCA: convergence rate along $v_i\propto\lambda_i$ (PCA direction; PCA eigenvalue)



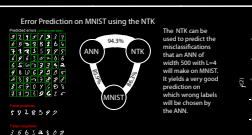
The NTK at initialization on the unit circle for 2 different widths (10 random initializations each), along with its infinite-width limit.



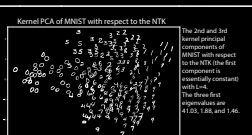
For a range of widths, the distances of the NTK at initialization to its infinite-width limit (orange) and to its limit at the end of training (blue). The distances are computed in Frobenius norm on a fixed batch.



Distribution of f_∞ at the end of training for an L^2-cost: 20 trials for an ANN of width 100, compared to the common mean (blue) and 80% confidence intervals of the infinite-width Gaussian distribution (green) and the posterior for NTK prior (pink).



The NTK can be used to predict the misclassifications that an ANN of width 500 with $L=4$ will make on MNIST. It yields a very good prediction on which wrong labels will be chosen by the ANN.



The 2nd and 3rd kernel principal components of MNIST with respect to the NTK (the first component is essentially constant) with $L=4$. The three first eigenvalues are 41.03, 1.88, and 1.46.

References

[1] A. Jacot, F. Gabriel, C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks, NeurIPS 2018, arXiv:1806.07572

[2] R.M. Neal. Bayesian Learning for Neural Networks, 1996

Acknowledgments

ERC CG CRITICAL (Gabriel)
ERC SG CONSTAMIS (Hongler)
Blavatnik Family Foundation
Latsis Family Foundation