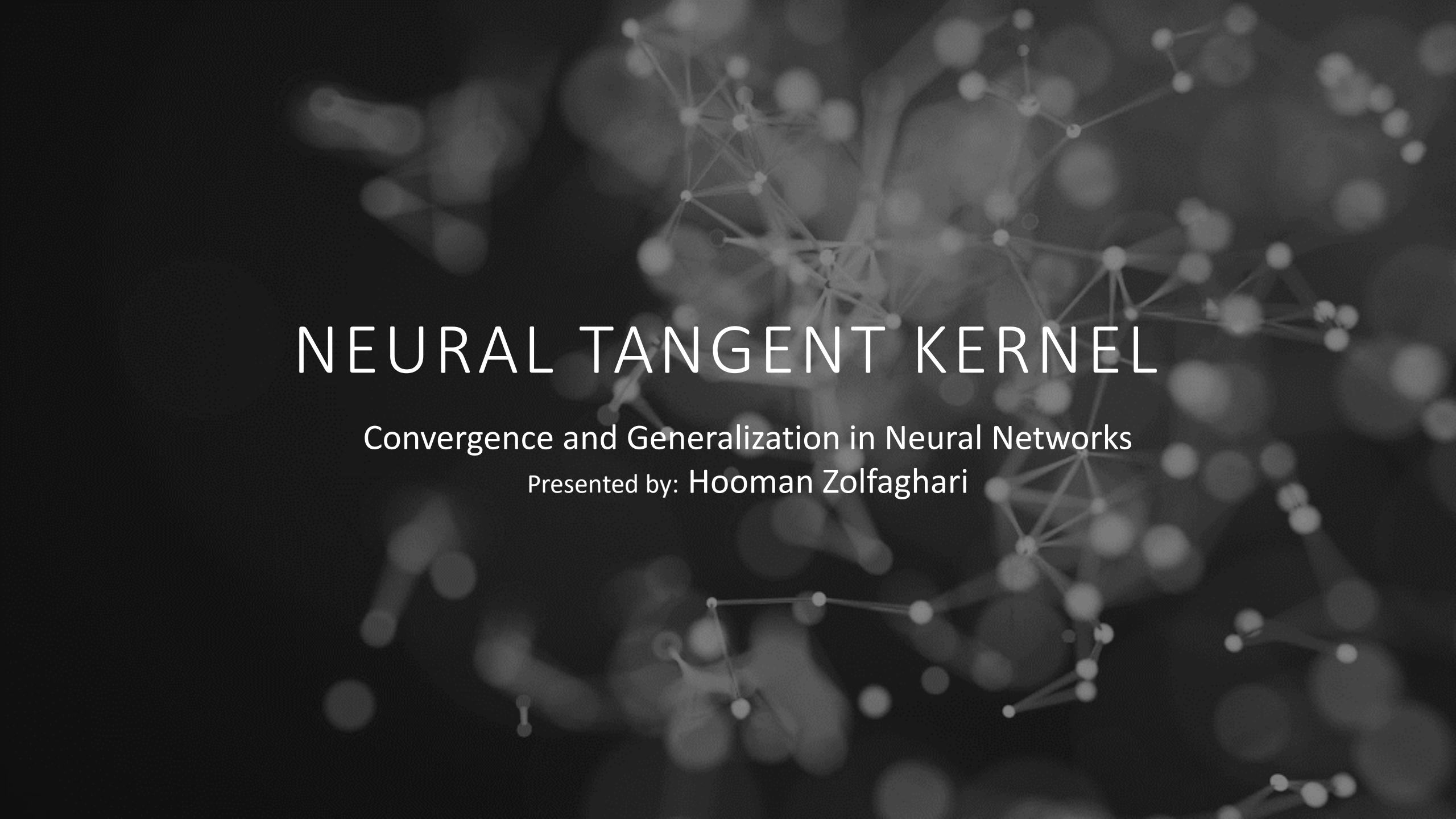


NEURAL TANGENT KERNEL

Convergence and Generalization in Neural Networks

Presented by: Hooman Zolfaghari

AGENDA

INTRODUCTION

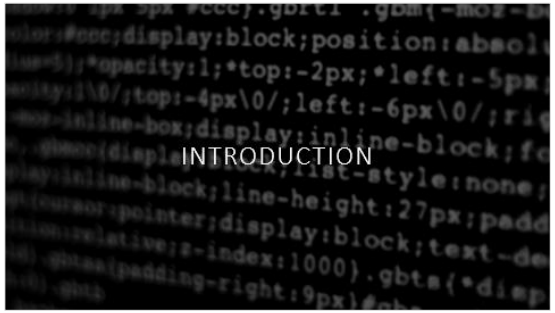
KERNEL GRADIENT

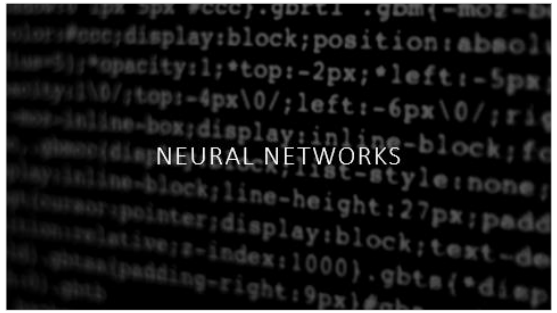
NEURAL TANGENT KERNEL

LEAST SQUARES REGRESSION

NUMERICAL EXPERIMENTS

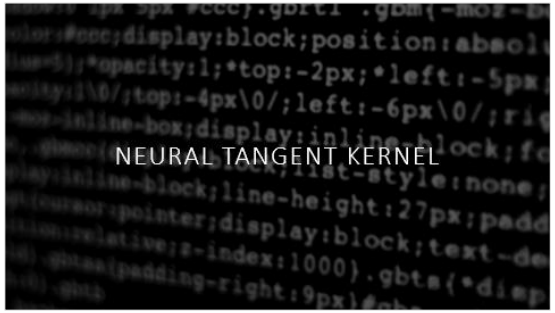
SLIDES ZOOM





KERNEL GRADIENT

OPTIMIZATION OF THE NN



LEAST-SQUARES REGRESSION

WITH NTK

NUMERICAL EXPERIMENTS

LARGE VS INFINITE WIDTH

CONCLUSION

Introducing NTK:

- Describes the local dynamics of ANNs during gradient descent.
- Establishes a connection between ANN training and kernel methods.

Infinite-Width Limit:

- ANNs are represented in function space by the NTK limit, depending on depth, nonlinearity, and parameter initialization variance.
- Gradient descent in ANNs becomes equivalent to kernel gradient descent with respect to the NTK limit.

Significance:

- NTK enables the analysis of generalization properties of ANNs.
- Reveals how depth and nonlinearity influence learning.
- Links convergence of training to the positive-definiteness of the limit NTK.
- Highlights directions favored by early stopping methods.

INTRODUCTION

CONVERGENCE OF NEURAL NETWORKS

- Loss surface is highly non-convex.
- high number of saddle points may slow down the convergence
- Recent studies on loss landscape geometry at initialization and dynamics of training large-width limit for shallow networks
- The dynamics of **deep** networks has however remained an open problem.
- ANNs have good generalization properties despite their usual over-parametrization
- Note that Kernel methods have the same properties

- In the infinite-width limit, ANNs have a Gaussian distribution described by a kernel.
- In the same limit, the behavior of ANNs during training is described by a related kernel:

NEURAL TANGENT KERNEL (NTK)

NEURAL NETWORKS

NEURAL NETWORKS

We consider fully-connected ANNs

- layers numbered from 0 (input) to L (output), each containing $n_0, \dots, n_L$ neurons.
- Lipschitz, twice-differentiable activation $\sigma: \mathbb{R} \to \mathbb{R}$, with bounded second derivative.
- ANN realization function $F^{(L)}: \mathbb{R}^P \to \mathcal{F}$
- Connection matrices $W^{(l)} \in \mathbb{R}^{n_l \times n_{l+1}}$ and bias vectors $b^{(l)} \in \mathbb{R}^{n_{l+1}}$
- In our setup, the parameters are initialized as i.i.d Gaussians $\mathcal{N}(01)$

For a fixed distribution p^{in}, on $\mathbb{R}^{n_0}$, denote function space $\mathcal{F} = \{f: \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}\}$. On this space:

- The semi-norm $|| \, . \, ||_{p^{in}}$
- From bilinear form:

$$< f, g >_{p^{in}} = \mathbb{E}_{p^{in}}[f(x)^T g(x)]$$

- We assume p^{in} is the Empirical Distribution on finite dataset $\{x_1, \dots, x_N\}$

NEURAL NETWORKS

Define $f_\theta(x) \coloneqq \tilde{\alpha}(x;\theta)$ as:

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

The factors $\frac{1}{\sqrt{n_l}}$ are key to obtaining a consistent asymptotic behavior of neural networks as the widths of the hidden layers grow to infinity.

KERNEL GRADIENT

OPTIMIZATION OF THE NN

KERNEL GRADIENT

- Functional Cost $C: \mathcal{F} \to \mathbb{R}$
- Multi-dimensional kernel $K: \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L \times n_L}$
- K is Symmetric
- Defines a bilinear map:

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

- Positive Definite with respect to $|| \, . \, ||_{p^{in}}$:

$$|| \, f \, ||_{p^{in}} > 0 \implies || \, f ||_K > 0$$

- Denote $\mathcal{F}^*$ as dual of $\mathcal{F}$ wth respect to $< . , . >_{p^{in}}$
- It's the set of linear forms $\mu: \mathcal{F} \to \mathbb{R}$
- Where $\mu = < d, . >_{p^{in}}$ for some $d \in \mathcal{F}$.
- Two elements of $\mathcal{F}$ define the same linear form if and only if they are equal on the data.
- Define map $\Phi_K: \mathcal{F}^* \to \mathcal{F}$, as $f_\mu = \Phi_K(\mu)$

KERNEL GRADIENT

- The cost $C: \mathcal{F} \to \mathbb{R}$ only depends on values of $f \in \mathcal{F}$ at the data points.
- The functional derivative $\partial_f^{in} C|_{f_0} \in \mathcal{F}^*$
- $\partial_f^{in} C|_{f_0} := < d|_{f_0}, . >_{p^{in}}$

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

- The *Kernel Gradient*:

$$\nabla_K C \Big|_{f_0} := \Phi_K(\partial_f^{in} C \Big|_{f_0})$$

- In contrast to $\partial_f^{in} C|_{f_0}$ which is only defined on the dataset,
- the kernel gradient generalizes to values x outside the dataset thanks to the kernel K:

$$\nabla_K C|_{f_0}(x) = \frac{1}{N} \sum_{j=1}^{N} K(x, x_j) d|_{f_0}(x_j)$$

KERNEL GRADIENT

A time-dependent function $f(t)$ follows the *kernel gradient descent* with respect to K if it satisfies the differential equation:

$$\partial_t f(t) = -\nabla_K C \Big|_{f(t)}$$

- During kernel gradient descent, the cost $C(f(t))$ evolves as:

$$\partial_t C|_{f(t)} = -\left\langle d|_{f(t)}, \nabla_K C|_{f(t)} \right\rangle_{p^{in}} = -\left\| d|_{f(t)} \right\|_K^2$$

- Convergence to a critical point of C is hence guaranteed if the kernel K is positive definite with respect to $||\,.\,||_{p^{in}}$

- The cost is then strictly decreasing except at points such that $|| \, d|_{f(t)}||_{p^{in}} = 0$. If the cost is convex and bounded from below, the function $f(t)$, therefore converges to a global minimum as $t \to \infty$

RANDOM FUNCTIONS APPROXIMATION

- A kernel K can be approximated by a choice of P random functions $f^{(p)}$ sampled independently from any distribution on $\mathcal{F}$ whose (non-centered) covariance is given by the kernel K:

$$\mathbb{E}\left[f_k^{(p)}(x) f_{k'}^{p}(x')\right] = K_{kk'}(x, x')$$

- These functions define $F^{lin}: \mathbb{R}^P \to \mathcal{F}$ as:

$$\theta \to f_\theta^{lin} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \theta_p f^{(p)}$$

$$\partial_{\theta_p} F^{lin}(\theta) = \frac{1}{\sqrt{P}} f^{(p)}$$

RANDOM FUNCTIONS APPROXIMATION

- Optimizing the cost $C \circ F^{lin}$ through gradient descent, the parameters follow the ODE:

$$\partial_t \theta_p(t) = -\partial_{\theta_p}(C \circ F^{lin})(\theta(t)) = -\frac{1}{\sqrt{P}} \partial_f^{in} C|_{f^{lin}_{\theta(t)}} \, f^{(p)} = -\frac{1}{\sqrt{P}} \left\langle d|_{f^{lin}_{\theta(t)}}, f^{(p)} \right\rangle_{p^{in}}.$$

- As a result the function $f^{lin}_{\theta(t)}$ evolves according to

$$\partial_t f^{lin}_{\theta(t)} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \partial_t \theta_p(t) f^{(p)} = -\frac{1}{P} \sum_{p=1}^{P} \left\langle d|_{f^{lin}_{\theta(t)}}, f^{(p)} \right\rangle_{p^{in}} f^{(p)},$$

- where the right-hand side is equal to the kernel gradient $-\nabla_{\tilde{K}} C$ with respect to the *tangent kernel*

$$\tilde{K} = \sum_{p=1}^{P} \partial_{\theta_p} F^{lin}(\theta) \otimes \partial_{\theta_p} F^{lin}(\theta) = \frac{1}{P} \sum_{p=1}^{P} f^{(p)} \otimes f^{(p)}.$$

- This is a random n_L-dimensional kernel with values $\tilde{K}_{ii'}(x, x') = \frac{1}{P} \sum_{p=1}^{P} f_i^{(p)}(x) f_{i'}^{(p)}(x')$.

RANDOM FUNCTIONS APPROXIMATION

- Performing gradient descent on the cost $C \circ F^{lin}$ is equivalent to kernel gradient descent with the tangent kernel $\widetilde{K}$ in the function space.
- As $P \to \infty$, by the law of large numbers, The (random) tangent kernel $\widetilde{K}$ tends to the fixed kernel K
- Makes this method an approximation of kernel gradient descent with respect to the limiting kernel K

NEURAL TANGENT KERNEL

NEURAL TANGENT KERNEL

For ANNs trained with gradient descent on $C \circ F^{(L)}$, the behavior is similar.

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

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

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

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

NEURAL TANGENT KERNEL

However:

- Unlike F^{lin}, for ANNs $F^{(L)}$ is not linear.
- So, Derivatives $\partial_{\theta p} F^{(L)}(\theta)$ and NTK depend on the parameters θ.
- Therefore, NTK is random at initialization and changes during training,
- complicating f_θ's convergence analysis.

They show that, in the **infinite-width limit**:

- NTK becomes **deterministic** at initialization and **remains constant** during training.
- f_θ is Gaussian at initialization
- So, In the limit, asymptotic behavior of f_θ during training can be explicitly analyzed in the function space $\mathcal{F}$.

INITIALIZATION

- It is known that the output functions $f_{\theta,i}$ for $i = 1, \ldots, n_L$ tend to i.i.d Gaussian processes in the infinite-width limit

Proposition 1. *For a network of depth L at initialization, with a Lipschitz nonlinearity σ, and in the limit as $n_1, \ldots, n_{L-1} \to \infty$, the output functions $f_{\theta,k}$, for $k = 1, \ldots, n_L$, tend (in law) to iid centered Gaussian processes of covariance $\Sigma^{(L)}$, where $\Sigma^{(L)}$ is defined recursively by:*

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

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

taking the expectation with respect to a centered Gaussian process f of covariance $\Sigma^{(L)}$.

INITIALIZATION

INITIALIZATION

- The first key result of the paper is: In the same limit, the Neural Tangent Kernel (NTK) converges in probability to an explicit deterministic limit, only depending on σ, depth, and variance of initialization.

Theorem 1. *For a network of depth L at initialization, with a Lipschitz nonlinearity σ, and in the limit as the layers width $n_1, ..., n_{L-1} \to \infty$, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel:*

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

The scalar kernel $\Theta_\infty^{(L)} : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$ is defined recursively by

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

where

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

taking the expectation with respect to a centered Gaussian process f of covariance $\Sigma^{(L)}$, and where $\dot{\sigma}$ denotes the derivative of σ.

TRAINING

- The second key result is that the NTK stays asymptotically constant during training
- The parameters are updated according to a training direction $d_t \in \mathcal{F}$: $\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}}$
- In the case of gradient descent, $d_t = -d|_{f_{\theta(t)}}$

Theorem 2. *Assume that σ is a Lipschitz, twice differentiable nonlinearity function, with bounded second derivative. For any T such that the integral $\int_0^T \|d_t\|_{p^{in}} dt$ stays stochastically bounded, as $n_1, \ldots, n_{L-1} \to \infty$, we have, uniformly for $t \in [0, T]$,*

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

As a consequence, in this limit, the dynamics of f_θ is described by the differential equation

$$\partial_t f_{\theta(t)} = \Phi_{\Theta_\infty^{(L)} \otimes Id_{n_L}} \left(\langle d_t, \cdot \rangle_{p^{in}} \right).$$

LEAST-SQUARES REGRESSION

WITH NTK

LEAST-SQUARES REGRESSION

- Given a goal function f^* and input distribution p^{in}, the least-squares regression cost is

$$C(f) = \frac{1}{2}||f - f^*||^2_{p^{in}} = \frac{1}{2}\mathbb{E}_{x\sim p^{in}}\left[\|f(x) - f^*(x)\|^2\right].$$

- The behavior of a function f_t during kernel gradient descent with NTK limit kernel:

$$\partial_t f_t = \Phi_K\left(\langle f^* - f, \cdot\rangle_{p^{in}}\right)$$

- The solution of this differential equation:

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

LEAST-SQUARES REGRESSION

- For a finite dataset $x_1, ..., x_N$ of size N, the map Π takes the form

$$\Pi(f)_k(x) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k'=1}^{n_L} f_{k'}(x_i) K_{kk'}(x_i, x)$$

- The map has at most Nn_L positive eigenfunctions: The kernel K's principal components $f^{(1)}, \ldots, f^{(Nn_L)}$, of the data.
- Decomposing the difference $(f^* - f_0) = \Delta_f^0 + \Delta_f^1 + ... + \Delta_f^{Nn_L}$ along the eigenspaces of Π, the trajectory of the function f_t reads

$$f_t = f^* + \Delta_f^0 + \sum_{i=1}^{Nn_L} e^{-t\lambda_i} \Delta_f^i,$$

where Δ_f^0 is in the kernel (null-space) of Π and $\Delta_f^i \propto f^{(i)}$.

LEAST-SQUARES REGRESSION

- Decomposing the difference $(f^* - f_0) = \Delta_f^0 + \Delta_f^1 + ... + \Delta_f^{Nn_L}$ along the eigenspaces of Π, the trajectory of the function f_t reads

$$f_t = f^* + \Delta_f^0 + \sum_{i=1}^{Nn_L} e^{-t\lambda_i} \Delta_f^i,$$

where Δ_f^0 is in the kernel (null-space) of Π and $\Delta_f^i \propto f^{(i)}$.

- This motivates the use of early stopping: The convergence is indeed faster along the eigenspaces corresponding to larger eigenvalues λ_i.
- Early stopping hence focuses the convergence on the most relevant kernel principal components, while avoiding to fit the ones in eigenspaces with lower eigenvalues.

NUMERICAL EXPERIMENTS

LARGE VS INFINITE WIDTH

CONVERGENCE OF THE NTK

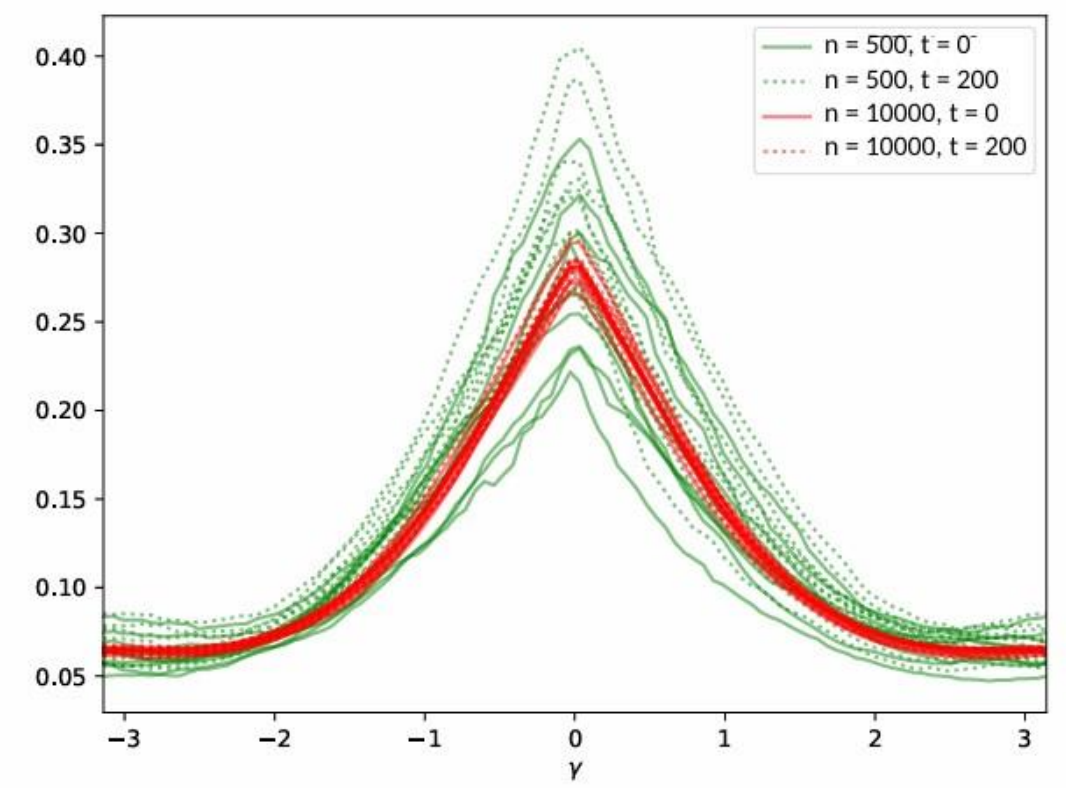


Figure 1: Convergence of the NTK to a fixed limit for two widths n and two times t.

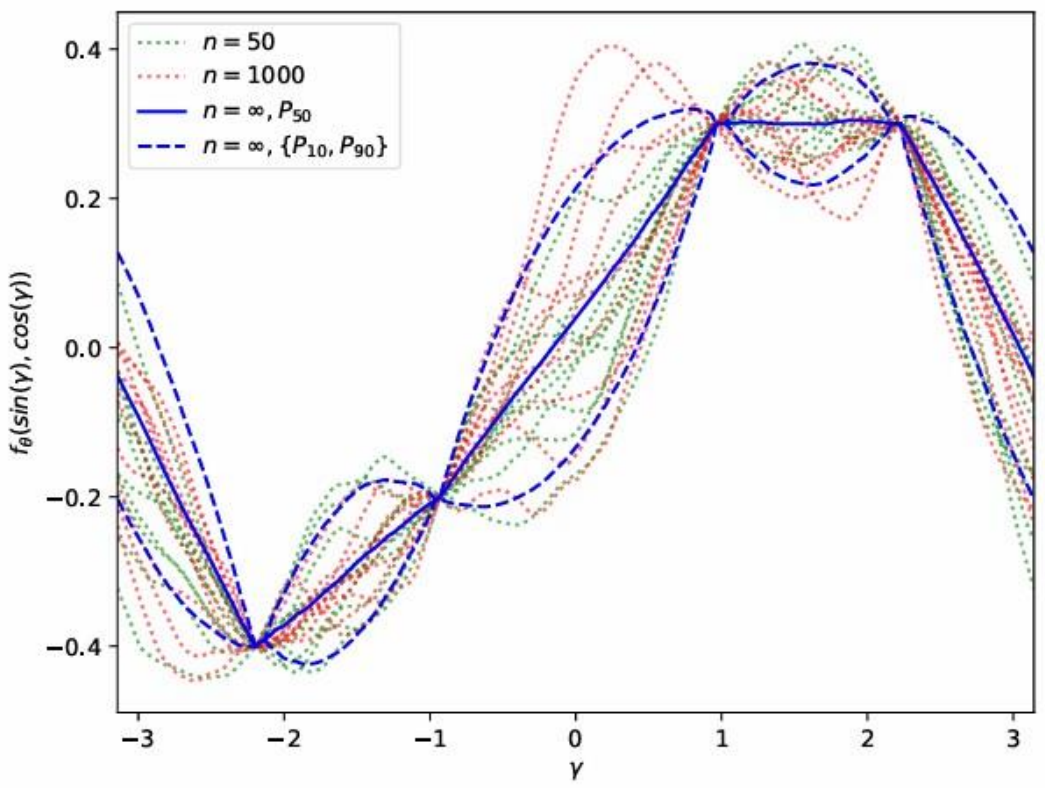
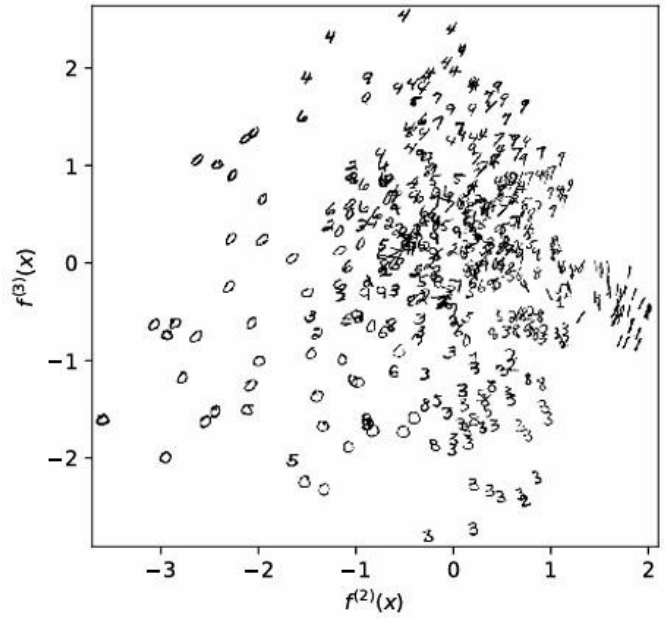
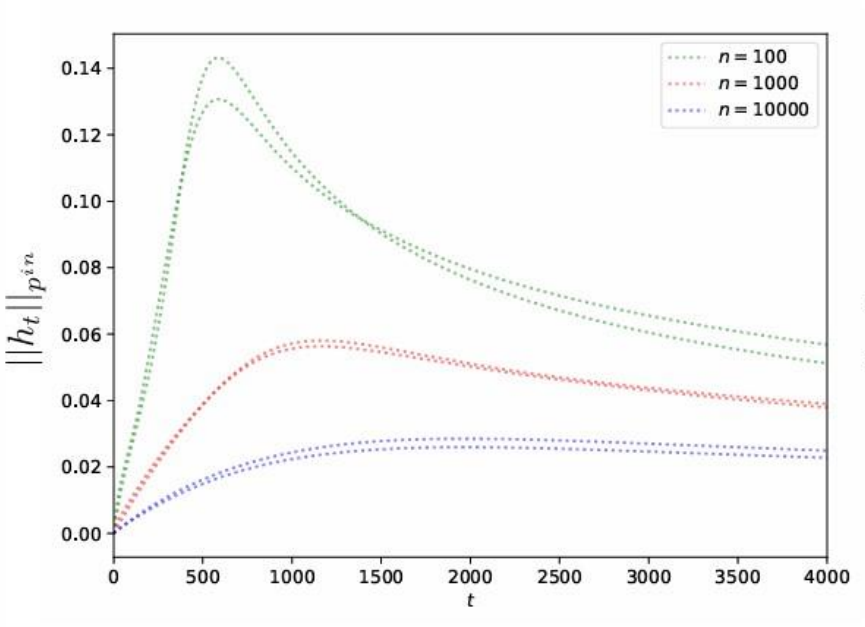


Figure 2: Networks function f_θ near convergence for two widths n and 10th, 50th and 90th percentiles of the asymptotic Gaussian distribution.

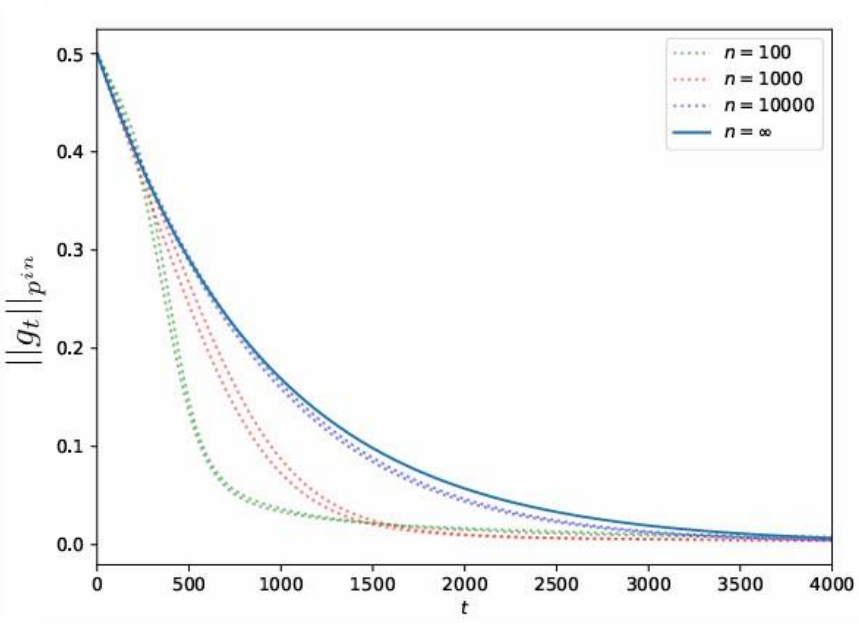
CONVERGENCE ALONG A PRINCIPAL COMPONENT



(a) The 2nd and 3rd principal components of MNIST.



(b) Deviation of the network function f_θ from the straight line.



(c) Convergence of f_θ along the 2nd principal component.

Figure 3

CONCLUSION

Introducing NTK:

- Describes the local dynamics of ANNs during gradient descent.
- Establishes a connection between ANN training and kernel methods.

Infinite-Width Limit:

- ANNs are represented in function space by the NTK limit, depending on depth, nonlinearity, and parameter initialization variance.
- Gradient descent in ANNs becomes equivalent to kernel gradient descent with respect to the NTK limit.

Significance:

- NTK enables the analysis of generalization properties of ANNs.
- Reveals how depth and nonlinearity influence learning.
- Links convergence of training to the positive-definiteness of the limit NTK.
- Highlights directions favored by early stopping methods.

THANK YOU

Hooman Zolfaghari

Email: hoomanzolfaghari84@gmail.com

Website: hoomanzolfaghari84.github.io/