Neural Tangent Kernel

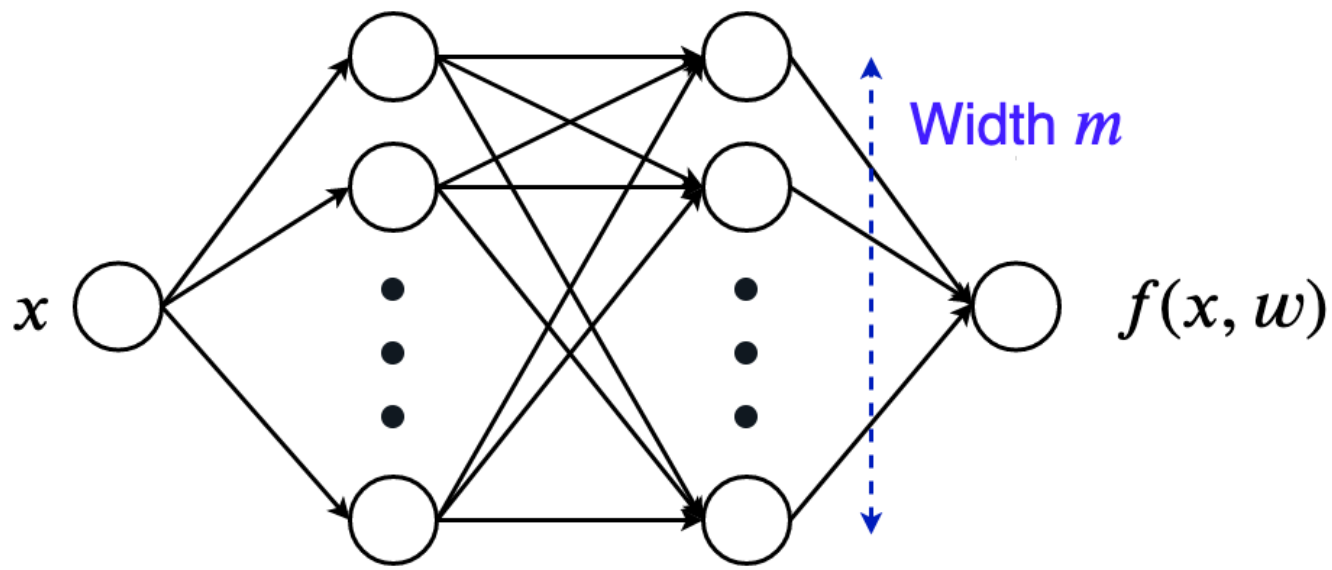
Джаин Никита БПМИ172

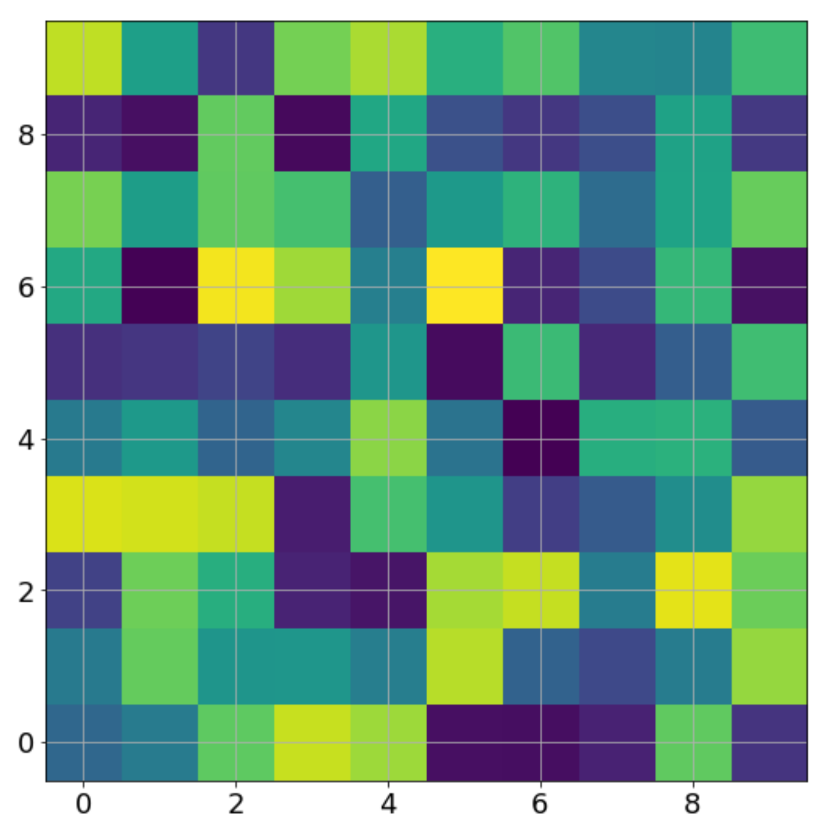
Least squares loss

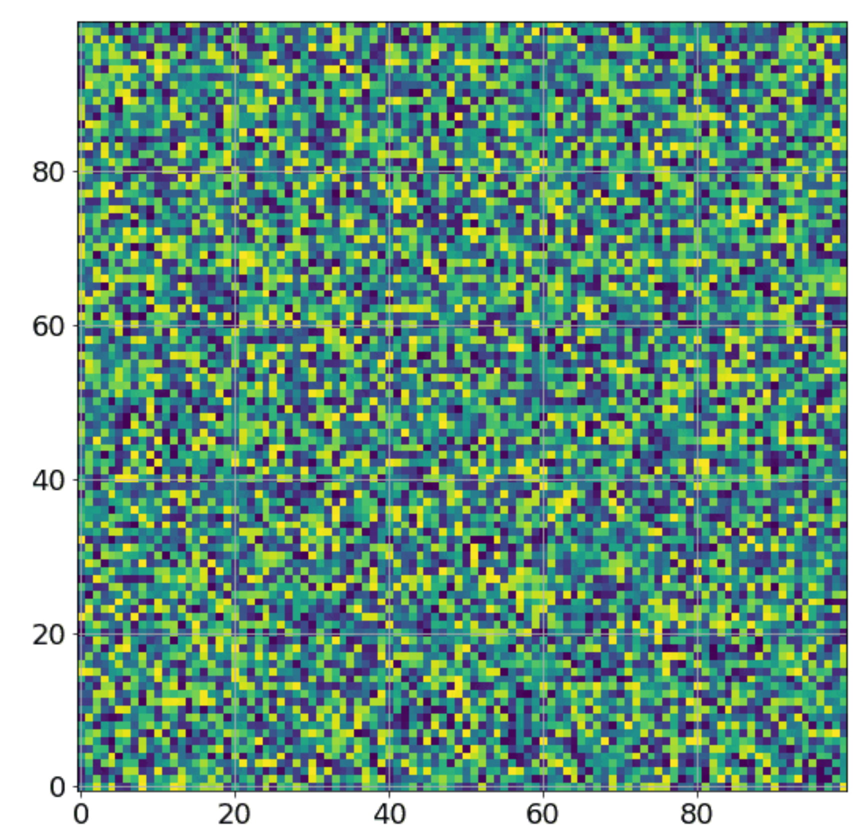
$$L(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(f(\bar{x}_i, \boldsymbol{w}) - \bar{y}_i)^2$$

Пусть $\boldsymbol{y}(\boldsymbol{w})_i = f(\bar{x}_i, \boldsymbol{w})$

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y}(\boldsymbol{w}) - \bar{\boldsymbol{y}}\|_2^2$$

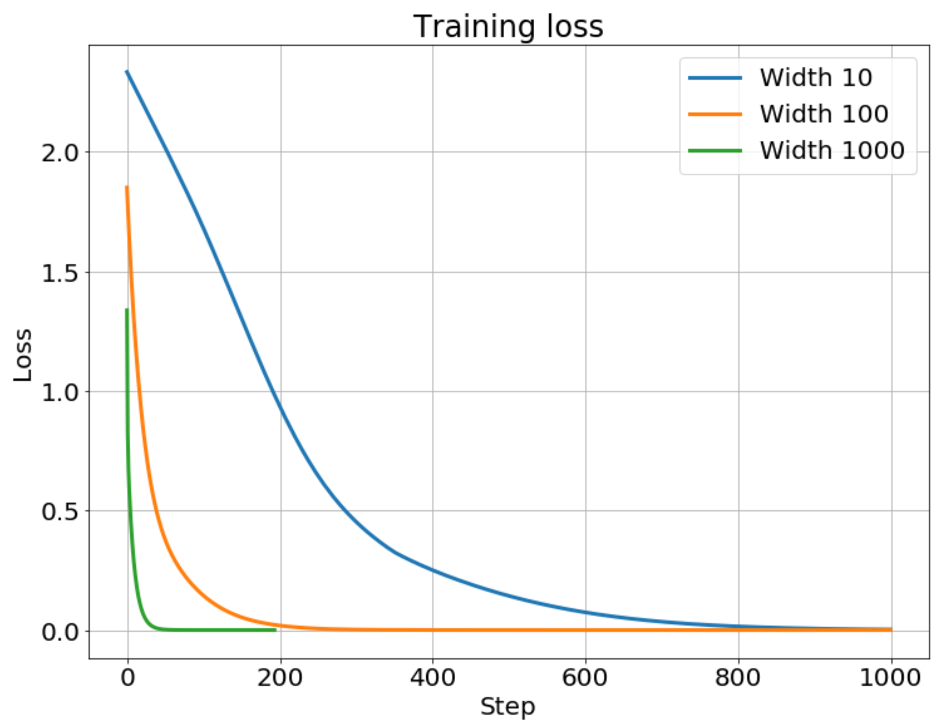


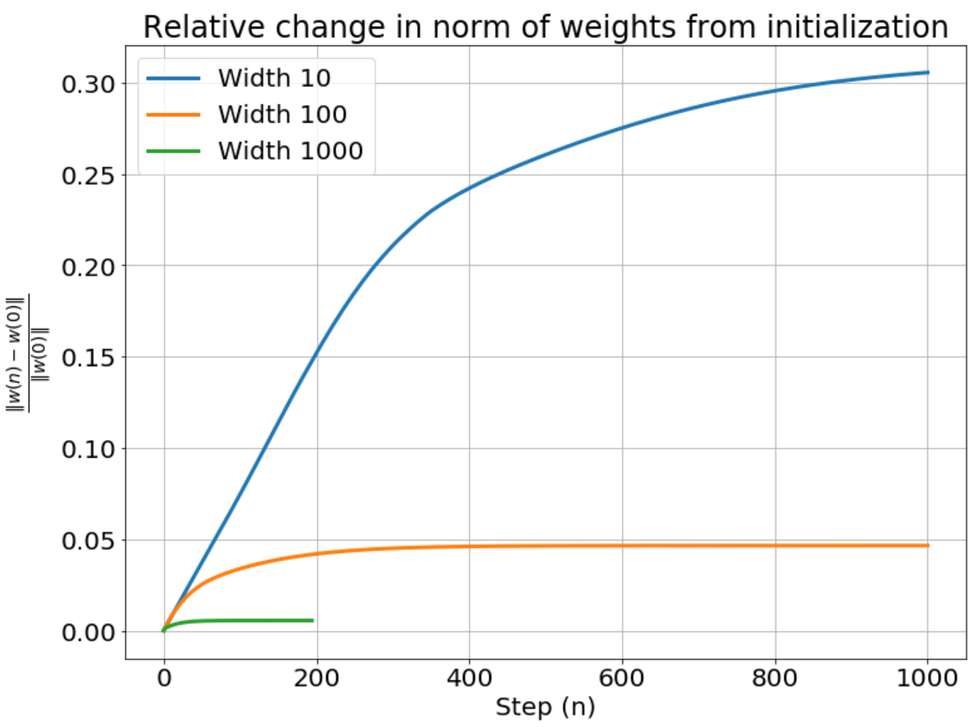




Least squares loss

Рассмотрим простую модель с 1-D входом и выходом, с 2 полносвязными слоями:





Можно заметить, что с ростом ширины слоев относительное изменение нормы весов почти не меняется

$$\frac{\|\boldsymbol{w}(n) - \boldsymbol{w}_0\|_2}{\|\boldsymbol{w}_0\|_2}$$

Linear approximation

Мы можем линейно приблизить нашу модель разложив ее по Тейлору:

$$f(x, \boldsymbol{w}) \approx f(x, \boldsymbol{w_0}) + \nabla_{\boldsymbol{w}} f(x, \boldsymbol{w_0})^T (\boldsymbol{w} - \boldsymbol{w_0})$$

Мы линейно аппроксимировали нашу модель относительно весов.
Данное приближение можно представить, как обычную линейную модель с отображением $\phi(x)$

$$\boldsymbol{\phi}(x) = \nabla_{\boldsymbol{w}} f(x, \boldsymbol{w_0})$$

Насколько справедливо использовать такое приближение?

Linear approximation

Насколько справедливо использовать такое приближение?

$$w_1 = w_0 - \eta \nabla_w L(w_0)$$

При достаточно маленьком значении η, ошибка всегда будет уменьшаться. Тогда:

net change in $y(w) \lesssim \|y(w_0) - \bar{y}\|$

$$\textbf{distance d moved in w space} \approx \frac{\textbf{net change in } y(w)}{\textbf{rate of change of } y \textbf{ w.r.t } w} = \frac{\|y(w_0) - \bar{y}\|}{\|\nabla_w y(w_0)\|}$$

$$\textbf{relative change in model Jacobian} \approx \frac{d \cdot \textbf{rate of change of Jacobian}}{\textbf{norm of Jacobian}} = \frac{d \cdot \|\nabla_w^2 y(w_0)\|}{\|\nabla_w y(w_0)\|} = \|(y(w_0) - \bar{y})\| \frac{\|\nabla_w^2 y(w_0)\|}{\|\nabla_w y(w_0)\|^2} = \kappa(w_0) \ll 1$$

Если выполняется последнее условие, то мы можем гарантировать, что наше линейное приближение близко к самой модели.

Gradient flow

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta \nabla_{\boldsymbol{w}} L(\boldsymbol{w}_k)$$

Перепишем как:

$$\frac{\boldsymbol{w}_{k+1} - \boldsymbol{w}_k}{\eta} = - \nabla_{\boldsymbol{w}} L(\boldsymbol{w}_k)$$

Разницу весов можно проинтерпретировать, как изменение весов в момент времени:

$$\frac{d\boldsymbol{w}(t)}{dt} = - \nabla_{\boldsymbol{w}} L(\boldsymbol{w}(t))$$

Подставляя ls loss, получим:

$$\dot{\boldsymbol{w}} = - \nabla \boldsymbol{y}(\boldsymbol{w})(\boldsymbol{y}(\boldsymbol{w}) - \bar{\boldsymbol{y}})$$

Теперь мы можем вывести изменение модели за итерацию:

$$\dot{\boldsymbol{y}}(\boldsymbol{w}) = \nabla \boldsymbol{y}(\boldsymbol{w})^T \dot{\boldsymbol{w}} = - {\color{red}\nabla \boldsymbol{y}(\boldsymbol{w})^T \nabla \boldsymbol{y}(\boldsymbol{w})}(\boldsymbol{y}(\boldsymbol{w}) - \bar{\boldsymbol{y}})$$

Часть выражения, выделенная красным - NTK

Gradient flow

$$H(w) = \nabla \boldsymbol{y}(\boldsymbol{w})^T \nabla \boldsymbol{y}(\boldsymbol{w})$$

Если наша модель близка к линейной аппроксимации ($\kappa(w_0) \ll 1$), то Якобиан не сильно меняется:

$$\nabla \boldsymbol{y}(\boldsymbol{w}(t)) \approx \nabla \boldsymbol{y}(\boldsymbol{w}_0) \implies \boldsymbol{H}(\boldsymbol{w}(t)) \approx \boldsymbol{H}(\boldsymbol{w}_0)$$

То есть, мы можем свести динамику тренировки к решению простого ДУ

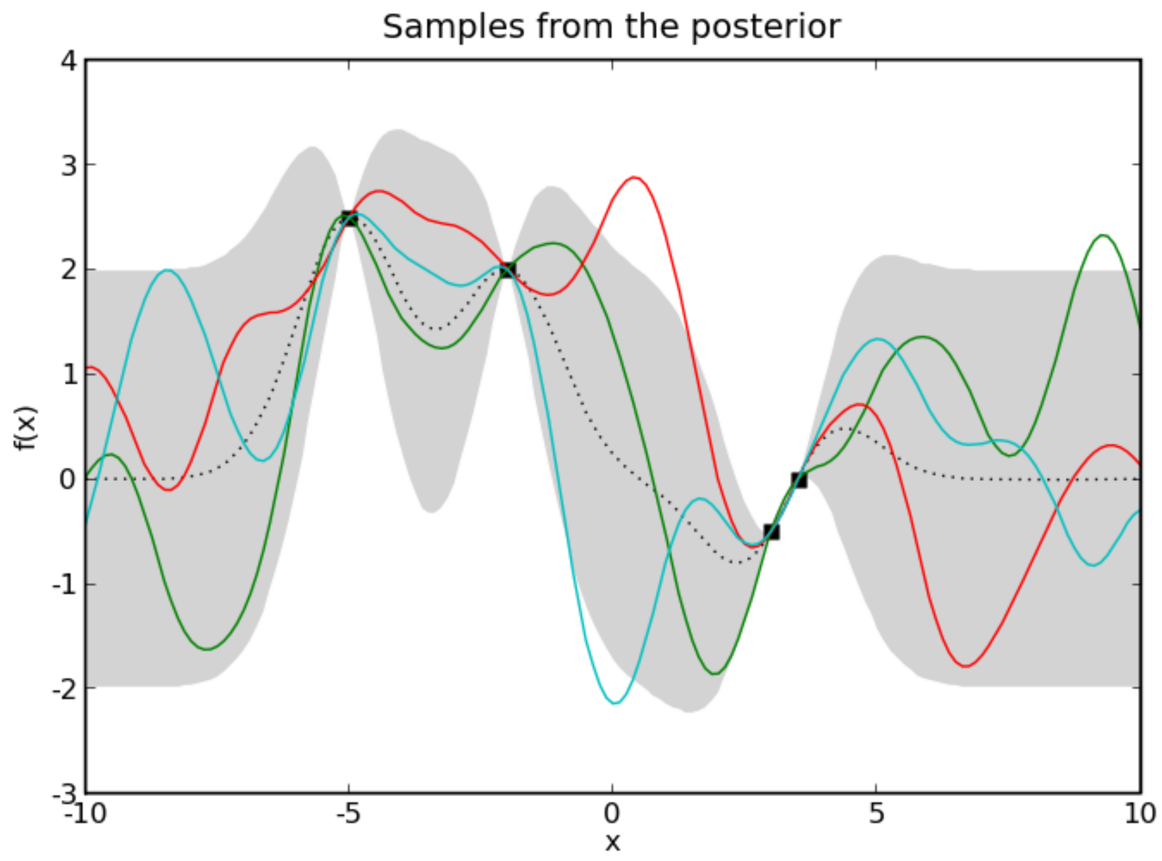
$$\dot{\boldsymbol{y}}(\boldsymbol{w}) = -\boldsymbol{H}(\boldsymbol{w}_0)(\boldsymbol{y}(\boldsymbol{w}) - \bar{\boldsymbol{y}})$$

$$\boldsymbol{u}(t) = \boldsymbol{u}(0)e^{-\boldsymbol{H}(\boldsymbol{w}_0)t}$$ **, где** $\boldsymbol{u} = \boldsymbol{y}(\boldsymbol{w}) - \bar{\boldsymbol{y}}$

Gaussian process

$$f(\mathbf{x}) \sim \mathcal{GP}(\cdot|\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$$

$$Cov(f(x), f(x')) = K(x, x')$$



Neural tangent kernel

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

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

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

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

Neural tangent kernel

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

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

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

Neural tangent kernel

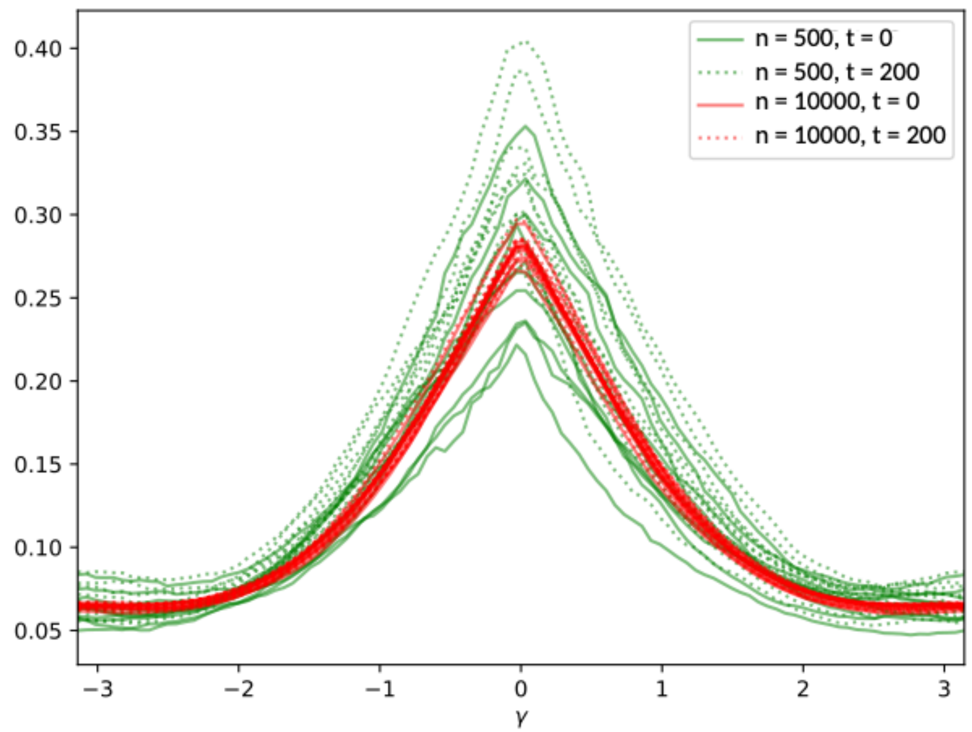


Figure 1: Convergence of the NTK to a fixed limit for two widths n and two times t.

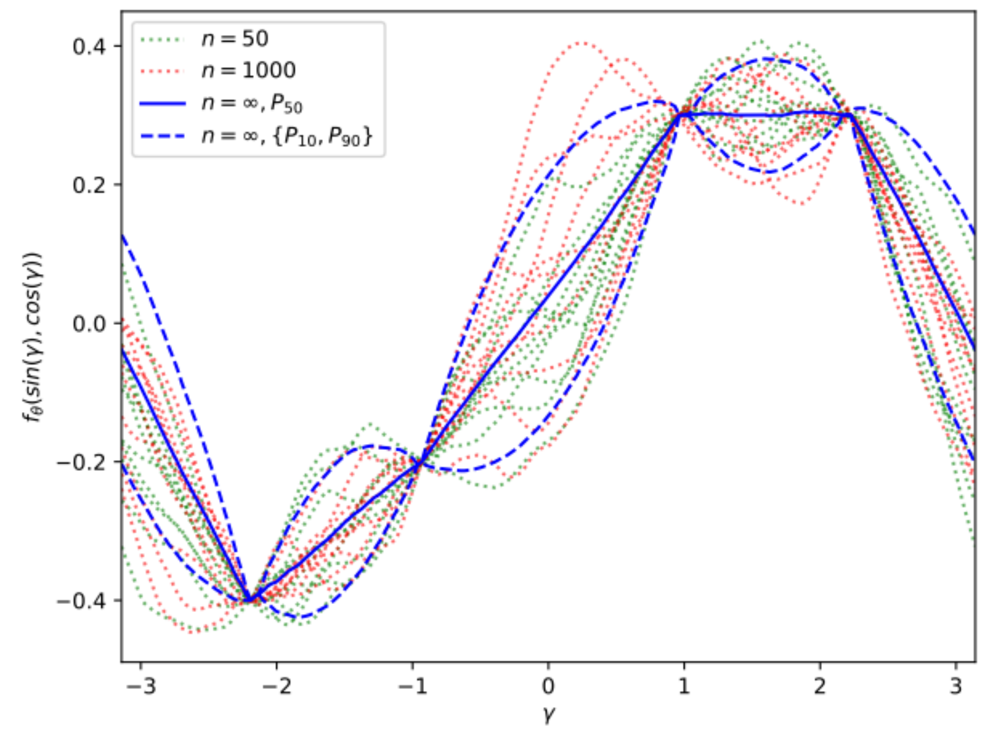


Figure 2: Networks function f_θ near convergence for two widths n and 10th, 50th and 90th percentiles of the asymptotic Gaussian distribution.

Источники

- https://arxiv.org/pdf/1806.07572.pdf
- https://rajatvd.github.io/NTK/