Logistics

Logistics

- Reading Group on **Neural Tangent Kernel**

Logistics

- Reading Group on **Neural Tangent Kernel**
- For me, **generalization** view point but very interested in other insights from optimization etc

Logistics

- Reading Group on **Neural Tangent Kernel**
- For me, **generalization** view point but very interested in other insights from optimization etc
- Weekly meetings every **Wednesday afternoon**, time flexible but "default" at **2-3pm**

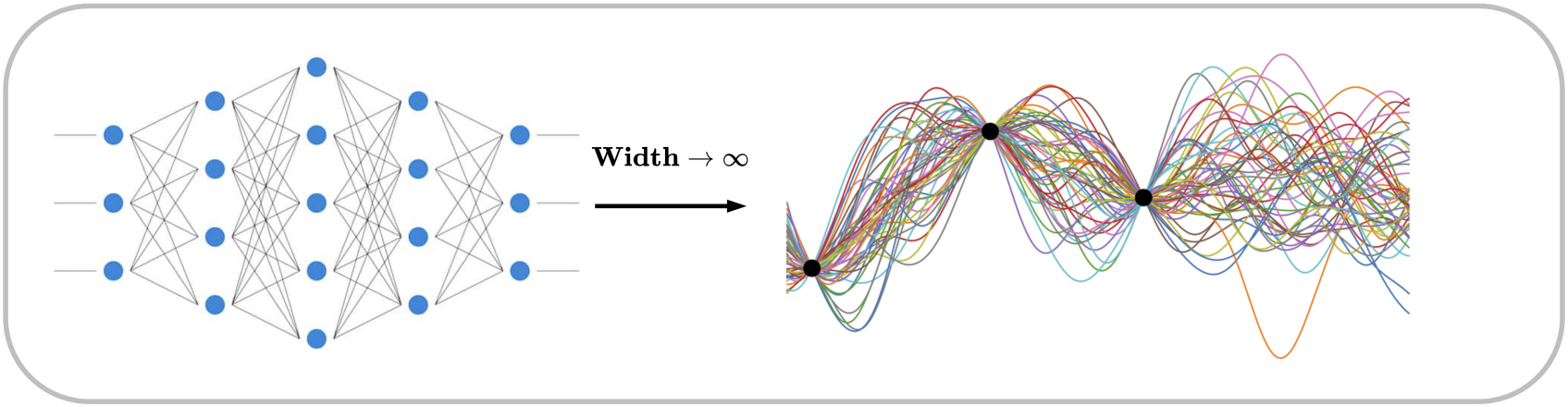
Logistics

- Reading Group on **Neural Tangent Kernel**
- For me, **generalization** view point but very interested in other insights from optimization etc
- Weekly meetings every **Wednesday afternoon**, time flexible but "default" at **2-3pm**
- Sign up for a paper or add your own paper of interest

Logistics

- Reading Group on **Neural Tangent Kernel**
- For me, **generalization** view point but very interested in other insights from optimization etc
- Weekly meetings every **Wednesday afternoon**, time flexible but "default" at **2-3pm**
- Sign up for a paper or add your own paper of interest
- **Virtual** meetings via Zoom (unless of course everyone is back at the lab) and hence slide presentations

Deep Neural Networks as Gaussian Processes



Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, Jascha Sohl-Dickstein

Recap: Gaussian Process

Recap: Gaussian Process

- Mathematical definition of a Gaussian process:

Recap: Gaussian Process

- Mathematical definition of a Gaussian process:

Definition: A **Gaussian process** $\{X_t\}_{t=0}^{\infty} \sim \mathcal{GP}(0, \Sigma)$ is a stochastic process such that for any finite collection $(t_1, \ldots, t_n) \in \mathbb{R}_+^n$, we have $(X_{t_1}, \ldots, X_{t_n}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = \Sigma(X_{t_i}, X_{t_j})$ and $\Sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Recap: Gaussian Process

- Mathematical definition of a Gaussian process:

  Definition: A **Gaussian process** $\{X_t\}_{t=0}^{\infty} \sim \mathcal{GP}(0, \Sigma)$ is a stochastic process such that for any finite collection $(t_1, \ldots, t_n) \in \mathbb{R}_+^n$, we have $(X_{t_1}, \ldots, X_{t_n}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = \Sigma(X_{t_i}, X_{t_j})$ and $\Sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

- Allows to sample random functions where **path continuity** depends solely on Σ (*Kolmogorov's continuity theorem*)

Recap: Gaussian Process

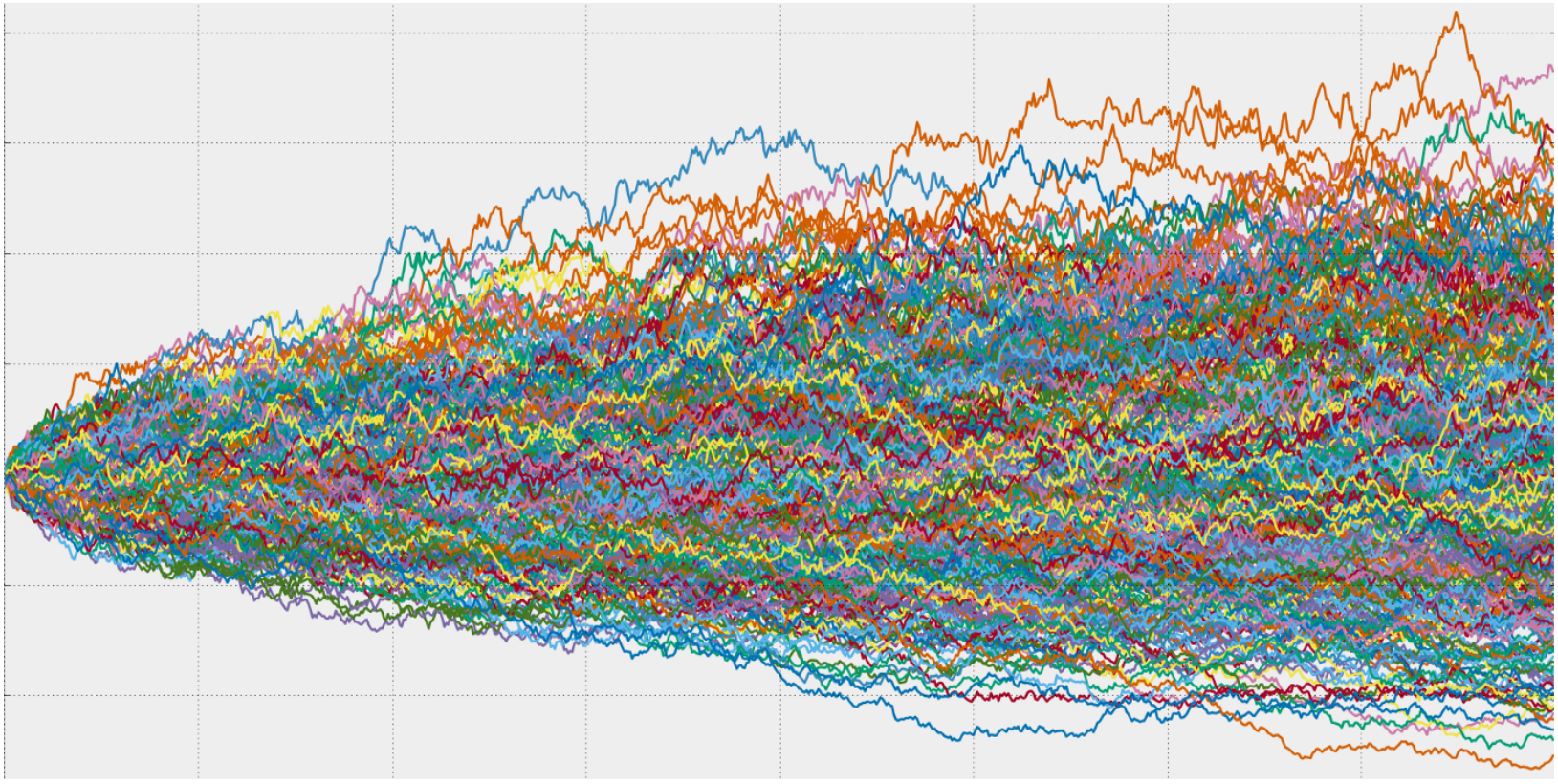
- Mathematical definition of a Gaussian process:

  Definition: A **Gaussian process** $\{X_t\}_{t=0}^{\infty} \sim \mathcal{GP}(0, \Sigma)$ is a stochastic process such that for any finite collection $(t_1, \ldots, t_n) \in \mathbb{R}_+^n$, we have $(X_{t_1}, \ldots, X_{t_n}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = \Sigma(X_{t_i}, X_{t_j})$ and $\Sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

- Allows to sample random functions where **path continuity** depends solely on Σ (*Kolmogorov's continuity theorem*)

- **Example:** Brownian motion with $\Sigma(s, t) = \min(s, t)$

Sample Paths of Brownian Motion



Gaussian Processes in ML (I)

Gaussian Processes in ML (I)

- **Regression Problem:** Observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \subset \mathbb{R}^d$ and a noisy model $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Gaussian Processes in ML (I)

- **Regression Problem:** Observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \subset \mathbb{R}^d$ and a noisy model $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Prior over functions:** $f(\cdot) \sim \mathcal{GP}(0, \Sigma)$

  But how to model Σ?

Gaussian Processes in ML (I)

- **Regression Problem:** Observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \subset \mathbb{R}^d$ and a noisy model $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Prior over functions:** $f(\cdot) \sim \mathcal{GP}(0, \Sigma)$

  But how to model Σ?
- Use information in the features $\boldsymbol{x}_i$ by taking approach

$$\Sigma_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a **symmetric** and **positive semi-definite** kernel

Gaussian Processes in ML (II)

Gaussian Processes in ML (II)

- $[K(\boldsymbol{x}, \boldsymbol{X})]_i = K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[K(\boldsymbol{X}, \boldsymbol{X})]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$

Gaussian Processes in ML (II)

- $[K(\boldsymbol{x}, \boldsymbol{X})]_i = K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[K(\boldsymbol{X}, \boldsymbol{X})]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- **Bayes:** Given new observation $\boldsymbol{x}^* \in \mathbb{R}^d$, how to **update** prior to get $P(y^* | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*)$?

Gaussian Processes in ML (II)

- $[K(\boldsymbol{x}, \boldsymbol{X})]_i = K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[K(\boldsymbol{X}, \boldsymbol{X})]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- **Bayes:** Given new observation $\boldsymbol{x}^* \in \mathbb{R}^d$, how to **update** prior to get $P(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*)$?
- First obtain joint distribution:

Gaussian Processes in ML (II)

- $[K(\boldsymbol{x}, \boldsymbol{X})]_i = K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[K(\boldsymbol{X}, \boldsymbol{X})]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- **Bayes:** Given new observation $\boldsymbol{x}^* \in \mathbb{R}^d$, how to **update** prior to get $P(y^* | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*)$?
- First obtain joint distribution:

$$\begin{pmatrix} \boldsymbol{y} \\ y^* \end{pmatrix} = \underbrace{\begin{pmatrix} h(\boldsymbol{X}) \\ h(\boldsymbol{x}^*) \end{pmatrix}} + \underbrace{\begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon^* \end{pmatrix}}$$

Gaussian Processes in ML (II)

- $[K(\boldsymbol{x}, \boldsymbol{X})]_i = K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[K(\boldsymbol{X}, \boldsymbol{X})]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- **Bayes:** Given new observation $\boldsymbol{x}^* \in \mathbb{R}^d$, how to **update** prior to get $P(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*)$?
- First obtain joint distribution:

$$\begin{pmatrix} \boldsymbol{y} \\ y^* \end{pmatrix} = \underbrace{\begin{pmatrix} h(\boldsymbol{X}) \\ h(\boldsymbol{x}^*) \end{pmatrix}}_{\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{x}, \boldsymbol{X}) \\ K(\boldsymbol{X}, \boldsymbol{x}) & K(\boldsymbol{x}, \boldsymbol{x}) \end{pmatrix}\right)} + \underbrace{\begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon^* \end{pmatrix}}$$

Gaussian Processes in ML (II)

- $[K(\boldsymbol{x}, \boldsymbol{X})]_i = K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[K(\boldsymbol{X}, \boldsymbol{X})]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- **Bayes:** Given new observation $\boldsymbol{x}^* \in \mathbb{R}^d$, how to **update** prior to get $P(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*)$?
- First obtain joint distribution:

$$\begin{pmatrix} \boldsymbol{y} \\ y^* \end{pmatrix} = \underbrace{\begin{pmatrix} h(\boldsymbol{X}) \\ h(\boldsymbol{x}^*) \end{pmatrix}}_{\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{x}, \boldsymbol{X}) \\ K(\boldsymbol{X}, \boldsymbol{x}) & K(\boldsymbol{x}, \boldsymbol{x}) \end{pmatrix}\right)} + \underbrace{\begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon^* \end{pmatrix}}_{\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma^2 \mathbb{1} & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)}$$

Gaussian Processes in ML (III)

Gaussian Processes in ML (III)

Adding the two distributions gives

Gaussian Processes in ML (III)

Adding the two distributions gives

$$\begin{pmatrix} \boldsymbol{y} \\ y^* \end{pmatrix} \Bigg| (\boldsymbol{X}, \boldsymbol{x}^*) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \mathbb{1} & K(\boldsymbol{x}, \boldsymbol{X}) \\ K(\boldsymbol{X}, \boldsymbol{x}) & K(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \end{pmatrix}\right)$$

Gaussian Processes in ML (III)

Adding the two distributions gives

$$\begin{pmatrix} \boldsymbol{y} \\ y^* \end{pmatrix} \Bigg| \, (\boldsymbol{X}, \boldsymbol{x}^*) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \mathbb{1} & K(\boldsymbol{x}, \boldsymbol{X}) \\ K(\boldsymbol{X}, \boldsymbol{x}) & K(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \end{pmatrix}\right)$$

- **Fact**: For $\begin{pmatrix} \boldsymbol{z}_A \\ \boldsymbol{z}_B \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right)$ we have

$$\boldsymbol{x}_A | \, \boldsymbol{x}_B \sim \mathcal{N}\left(\Sigma_{AB}\Sigma_{BB}^{-1}\boldsymbol{x}_B, \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$$

Gaussian Processes in ML (III)

Adding the two distributions gives

$$\begin{pmatrix} \boldsymbol{y} \\ y^* \end{pmatrix} \Bigg| (\boldsymbol{X}, \boldsymbol{x}^*) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \mathbb{1} & K(\boldsymbol{x}, \boldsymbol{X}) \\ K(\boldsymbol{X}, \boldsymbol{x}) & K(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \end{pmatrix}\right)$$

- **Fact**: For $\begin{pmatrix} \boldsymbol{z}_A \\ \boldsymbol{z}_B \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right)$ we have

$$\boldsymbol{x}_A | \, \boldsymbol{x}_B \sim \mathcal{N}\left(\Sigma_{AB}\Sigma_{BB}^{-1}\boldsymbol{x}_B, \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$$

- We hence have $y^*|(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*) \sim \mathcal{N}(\mu^*, \sigma^*)$ where

1 $\mu^* = K(\boldsymbol{x}^*, \boldsymbol{X}))^T \left(K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \mathbb{1}\right)^{-1} \boldsymbol{y}$

2 $\sigma^* = K(\boldsymbol{x}^*, \boldsymbol{x}^*) - K(\boldsymbol{x}^*, \boldsymbol{X})^T \left(K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \mathbb{1}\right)^{-1} K(\boldsymbol{x}^*, \boldsymbol{X})$

Gaussian Processes in ML (IV)

Gaussian Processes in ML (IV)

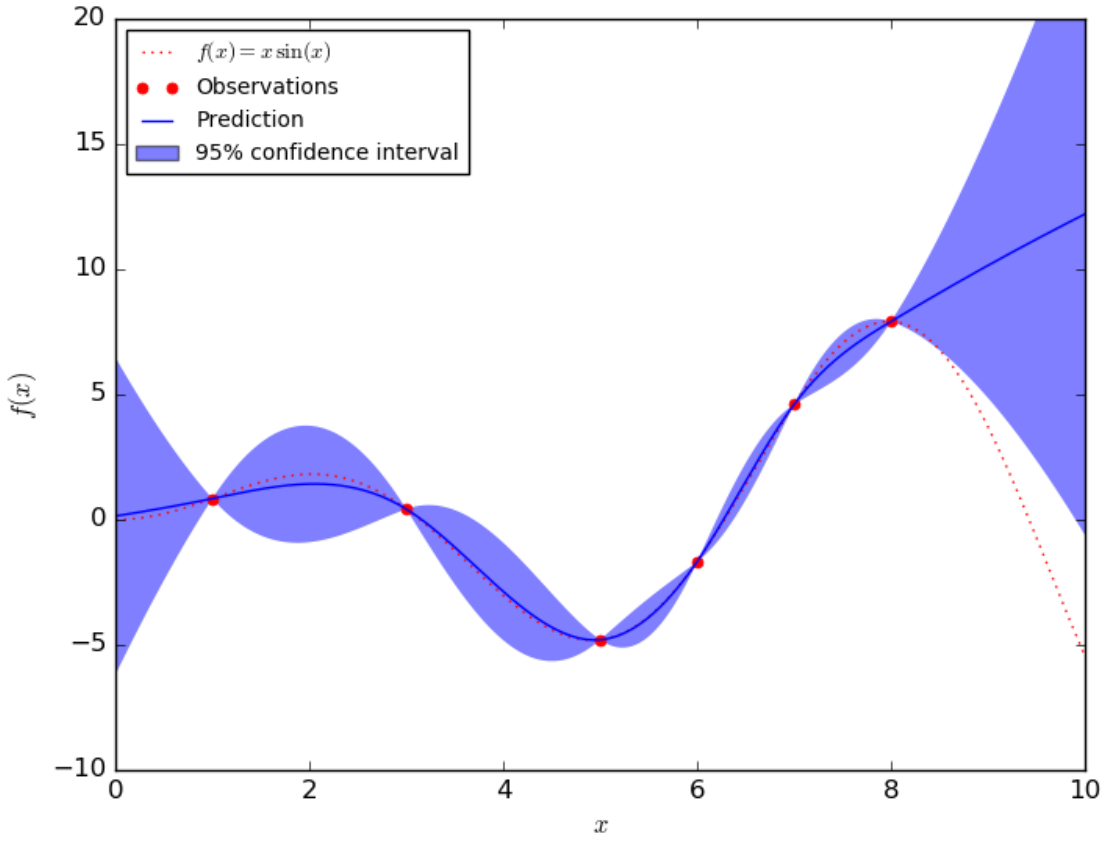
- We have a **probability distribution** as a prediction for y^* and hence a measure for uncertainty

Gaussian Processes in ML (IV)

- We have a **probability distribution** as a prediction for y^* and hence a measure for uncertainty
- Can take the **mean** μ^* or sample from f

Gaussian Processes in ML (IV)

- We have a **probability distribution** as a prediction for y^* and hence a measure for uncertainty
- Can take the **mean** μ^* or sample from f



Neural Networks: Notation

Neural Networks: Notation

- Only consider **fully-connected** networks of depth L:

Neural Networks: Notation

- Only consider **fully-connected** networks of depth L:
  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

Neural Networks: Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, σ is a coordinate-wise activation function and $\beta > 0$

Neural Networks: Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, σ is a coordinate-wise activation function and $\beta > 0$

- $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(L)}(\boldsymbol{x}) \in \mathbb{R}^{d_L}$ is the **output** of the network

Random Initialization

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- Made **width dependence** of random initialization explicit in the architecture

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- Made **width dependence** of random initialization explicit in the architecture
- **Effectively:** $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{d_l})$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \beta^2)$

  $\implies$ Usual Lecun initialization

Multidimensional Central Limit Theorem

Multidimensional Central Limit Theorem

- CLT will be major **ingredient** for the proofs

Multidimensional Central Limit Theorem

- CLT will be major **ingredient** for the proofs
- Let us restate it for completeness:

Multidimensional Central Limit Theorem

- CLT will be major **ingredient** for the proofs
- Let us restate it for completeness:

Theorem: *For* $\boldsymbol{X}_1 = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1d} \end{pmatrix}, \ldots, \boldsymbol{X}_n = \begin{pmatrix} X_{n1} \\ \vdots \\ X_{nd} \end{pmatrix}$ *i.i.d. multi-dim. random variables with* $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ *and* $cov(\boldsymbol{X}_i) = \boldsymbol{\Sigma}$*, it holds that in the limit*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{X}_i \xrightarrow{(d)} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Main Result

Main Result

Let's state the **main result** of this paper:

Main Result

Let's state the **main result** of this paper:

Theorem: *Given a NN $f_{\boldsymbol{\theta}}$, it holds that for $d_1, \ldots, d_{L-1} \to \infty$:*

$$f^{\infty}_{\boldsymbol{\theta},i}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(L)})$$

Main Result

Let's state the **main result** of this paper:

Theorem: *Given a NN $f_{\boldsymbol{\theta}}$, it holds that for $d_1, \ldots, d_{L-1} \to \infty$:*

$$f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(L)})$$

i.i.d. in i, where $\Sigma^{(L)}$ has the recursive structure

- $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \frac{1}{\sqrt{d_0}} \boldsymbol{x}^T \boldsymbol{x'} + \beta^2$
- $\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}^{(l-1)})} \left[\sigma(z_1) \sigma(z_2) \right] + \beta^2$

Main Result

Let's state the **main result** of this paper:

Theorem: *Given a NN $f_{\boldsymbol{\theta}}$, it holds that for $d_1, \ldots, d_{L-1} \to \infty$:*

$$f^{\infty}_{\boldsymbol{\theta},i}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(L)})$$

i.i.d. in i, where $\Sigma^{(L)}$ has the recursive structure

- $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \frac{1}{\sqrt{d_0}} \boldsymbol{x}^T \boldsymbol{x'} + \beta^2$
- $\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}^{(l-1)})} \left[\sigma(z_1) \sigma(z_2) \right] + \beta^2$

for $\tilde{\boldsymbol{\Sigma}}^{(l-1)} = \begin{pmatrix} \Sigma^{(l-1)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(l-1)}(\boldsymbol{x}, \boldsymbol{x'}) \\ \Sigma^{(l-1)}(\boldsymbol{x'}, \boldsymbol{x}) & \Sigma^{(l-1)}(\boldsymbol{x'}, \boldsymbol{x'}) \end{pmatrix}$

Some Comments

Some Comments

- Consider the **infinite-width** limit $d_1, \ldots, d_{L-1} \to \infty$ (all limits here denote convergence in **law**)

Some Comments

- Consider the **infinite-width** limit $d_1, \ldots, d_{L-1} \to \infty$ (all limits here denote convergence in **law**)

- **In words**: Each component of the output converges to a Gaussian process with a known covariance function modulo an integral over σ (known for $ReLU$)

Some Comments

- Consider the **infinite-width** limit $d_1, \ldots, d_{L-1} \to \infty$ (all limits here denote convergence in **law**)
- **In words**: Each component of the output converges to a Gaussian process with a known covariance function modulo an integral over σ (known for $ReLU$)
- Allows to perform full **Bayesian-inference** with an infinitely wide neural network (no Monte Carlo methods)

Some Comments

- Consider the **infinite-width** limit $d_1, \ldots, d_{L-1} \to \infty$ (all limits here denote convergence in **law**)
- **In words**: Each component of the output converges to a Gaussian process with a known covariance function modulo an integral over σ (known for $ReLU$)
- Allows to perform full **Bayesian-inference** with an infinitely wide neural network (no Monte Carlo methods)
- Only study network at **initialization**, no training at all

Some Comments

- Consider the **infinite-width** limit $d_1, \ldots, d_{L-1} \to \infty$ (all limits here denote convergence in **law**)
- **In words**: Each component of the output converges to a Gaussian process with a known covariance function modulo an integral over σ (known for $ReLU$)
- Allows to perform full **Bayesian-inference** with an infinitely wide neural network (no Monte Carlo methods)
- Only study network at **initialization**, no training at all
- Theory would actually break down since weights become **correlated** after just one step of SGD

Proof of Theorem

Proof of Theorem

- Proof uses **induction** on the depth L

Proof of Theorem

- Proof uses **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ are taken **sequentially**

Proof of Theorem

- Proof uses **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ are taken **sequentially**
- Central Limit Theorem performs all the heavy-lifting

Proof of Theorem

- Proof uses **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ are taken **sequentially**
- Central Limit Theorem performs all the heavy-lifting
- Use $\Sigma^{(l)}$ for covariance **function** and $\boldsymbol{\Sigma}^{(l)}$ for the resulting covariance **matrix** for data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$

Proof of Theorem

- Proof uses **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ are taken **sequentially**
- Central Limit Theorem performs all the heavy-lifting
- Use $\Sigma^{(l)}$ for covariance **function** and $\boldsymbol{\Sigma}^{(l)}$ for the resulting covariance **matrix** for data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$
- Write limits as $f_{\boldsymbol{\theta}}^{\infty}$ and $\tilde{\alpha}_{i,\infty}^{(l)}$

Base Case: $L = 1$

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Remains fixed

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Remains fixed

Considering just one component leads to

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Remains fixed

Considering just one component leads to

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \overbrace{\sum_{j=1}^{d_0} W_{ij}^{(0)} x_j} + \beta b_i^{(0)}$$

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Remains fixed

Considering just one component leads to

Lin.comb.of i.i.d. $\mathcal{N}$

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \overbrace{\sum_{j=1}^{d_0} W_{ij}^{(0)} x_j}^{} + \beta b_i^{(0)}$$

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Remains fixed

Considering just one component leads to

Lin.comb.of i.i.d. $\mathcal{N}$

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \overbrace{\sum_{j=1}^{d_0} W_{ij}^{(0)} x_j} + \beta b_i^{(0)}$$

$$\sim \mathcal{N}\left(0, \frac{1}{d_0} \sum_{j=1}^{d_0} \mathrm{Var}\left(W_{ij}^{(0)}\right) x_j^2 + \beta^2 \,\mathrm{Var}(b_i^{(0)})\right)$$

Base Case: $L = 1$

Network takes the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \boldsymbol{W}^{(0)} \boldsymbol{x} + \beta \boldsymbol{b}^{(0)}$$

Remains fixed

Considering just one component leads to

Lin.comb.of i.i.d. $\mathcal{N}$

$$\begin{aligned} f_{\boldsymbol{\theta},i}(\boldsymbol{x}) &= \frac{1}{\sqrt{d_0}} \overbrace{\sum_{j=1}^{d_0} W_{ij}^{(0)} x_j} + \beta b_i^{(0)} \\ &\sim \mathcal{N}\left(0, \frac{1}{d_0} \sum_{j=1}^{d_0} \mathrm{Var}\left(W_{ij}^{(0)}\right) x_j^2 + \beta^2 \,\mathrm{Var}(b_i^{(0)})\right) \\ &\overset{(d)}{=} \mathcal{N}\left(0, \frac{1}{d_0} ||\boldsymbol{x}||_2^2 + \beta^2\right) \end{aligned}$$

Covariance Structure

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^{d_0}$ we have:

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^{d_0}$ we have:

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x'})]$$

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^{d_0}$ we have:

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x'})]$$

$$= \mathbb{E}\Big[\left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j + \beta b_i^{(0)} \right) \left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x'_j + \beta b_i^{(0)} \right)$$

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^{d_0}$ we have:

$\mathbb{E}[W_{ij}^{(0)} b_i^{(0)}] = 0$

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x'})]$$

$$= \mathbb{E}\Big[\left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j + \beta b_i^{(0)} \right) \left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j' + \beta b_i^{(0)} \right)$$

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^{d_0}$ we have:

$\mathbb{E}[W_{ij}^{(0)} b_i^{(0)}] = 0$

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x'})]$$

$$= \mathbb{E}\Big[\left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j + \beta b_i^{(0)} \right) \left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j' + \beta b_i^{(0)} \right)$$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{E}\Big[W_{ij}^{(0)} W_{il}^{(0)} \Big] x_j x_l' + \beta^2$$

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^{d_0}$ we have:

$\mathbb{E}[W_{ij}^{(0)} b_i^{(0)}] = 0$

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x'})]$$

$$= \mathbb{E}\Big[\left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j + \beta b_i^{(0)} \right) \left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j' + \beta b_i^{(0)} \right)$$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{E}\Big[W_{ij}^{(0)} W_{il}^{(0)} \Big] x_j x_l' + \beta^2$$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{1}_{\{j=l\}} x_j x_l' + \beta^2$$

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$ we have:

$\mathbb{E}[W_{ij}^{(0)} b_i^{(0)}] = 0$

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x}')]$$

$$= \mathbb{E}\Big[\left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j + \beta b_i^{(0)} \right) \left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j' + \beta b_i^{(0)} \right)$$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{E}\Big[W_{ij}^{(0)} W_{il}^{(0)} \Big] x_j x_l' + \beta^2$$

$\mathbb{E}[W_{ij}^{(0)} W_{il}^{(0)}] = \mathbb{1}_{\{j=l\}}$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{1}_{\{j=l\}} x_j x_l' + \beta^2$$

Covariance Structure

For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$ we have:

$\mathbb{E}[W_{ij}^{(0)} b_i^{(0)}] = 0$

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},i}(\boldsymbol{x}')]$$

$$= \mathbb{E}\Big[\left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j + \beta b_i^{(0)} \right) \left(\frac{1}{\sqrt{d_0}} \sum_{j=1}^{d_0} W_{ij}^{(0)} x_j' + \beta b_i^{(0)} \right)$$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{E}\Big[W_{ij}^{(0)} W_{il}^{(0)} \Big] x_j x_l' + \beta^2$$

$\mathbb{E}[W_{ij}^{(0)} W_{il}^{(0)}] = \mathbb{1}_{\{j=l\}}$

$$= \frac{1}{d_0} \sum_{j=1}^{d_0} \sum_{l=1}^{d_0} \mathbb{1}_{\{j=l\}} x_j x_l' + \beta^2$$

$$= \frac{1}{d_0} \boldsymbol{x}^T \boldsymbol{x}' + \beta^2$$

Gaussian Process Property

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have:

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have:

$$\mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x'})] = 0$$

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have:

$$\mathbb{E}[W_{ik}^{(0)} W_{jk}^{(0)}] = 0$$

$$\mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x'})] = 0$$

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have:

$$\mathbb{E}[W_{ik}^{(0)} W_{jk}^{(0)}] = 0$$

$$\mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x'})] = 0$$

which implies **independence** due to Gaussianity.

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have:

$$\mathbb{E}[W_{ik}^{(0)} W_{jk}^{(0)}] = 0$$

$$\mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x'})] = 0$$

  which implies **independence** due to Gaussianity.

- Finally, we have for any $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k \in \mathbb{R}^{d_0}$:

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have:

$$\mathbb{E}[W_{ik}^{(0)} W_{jk}^{(0)}] = 0$$

$$\mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x'})] = 0$$

  which implies **independence** due to Gaussianity.

- Finally, we have for any $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k \in \mathbb{R}^{d_0}$:

$$\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(1)}\right)$$

Gaussian Process Property

- Moreover, for $i \neq j$ and all $\boldsymbol{x}, \boldsymbol{x'}$ we have: $\mathbb{E}[W_{ik}^{(0)} W_{jk}^{(0)}] = 0$

$$\mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x'})] = 0$$

  which implies **independence** due to Gaussianity.

- Finally, we have for any $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k \in \mathbb{R}^{d_0}$:

$$\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(1)}\right)$$

  where again $\boldsymbol{\Sigma}_{ij}^{(1)} = \Sigma^{(1)}(\boldsymbol{x}_i, \boldsymbol{x}_j)$

- **All in all:** $f_{\boldsymbol{\theta},i} \sim \mathcal{GP}(0, \Sigma^{(1)})$

Induction Step

Induction Step

- **Assume:** $\tilde{\alpha}_{i,\infty}^{(l)}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(l)})$ for $d_1, \ldots, d_{l-1} \to \infty$

Induction Step

- **Assume:** $\tilde{\alpha}_{i,\infty}^{(l)}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(l)})$ for $d_1, \ldots, d_{l-1} \to \infty$
- Consider network of depth $l + 1$:

Induction Step

- **Assume:** $\tilde{\alpha}_{i,\infty}^{(l)}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(l)})$ for $d_1, \ldots, d_{l-1} \to \infty$
- Consider network of depth $l + 1$:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)}$$

Induction Step

- **Assume:** $\tilde{\alpha}_{i,\infty}^{(l)}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(l)})$ for $d_1, \ldots, d_{l-1} \to \infty$
- Consider network of depth $l+1$:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)}$$

- Component-wise:

Induction Step

- **Assume:** $\tilde{\alpha}_{i,\infty}^{(l)}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(l)})$ for $d_1, \ldots, d_{l-1} \to \infty$
- Consider network of depth $l+1$:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)}$$

- Component-wise:

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} W_{ik}^{(l)} \alpha_k^{(l)}(\boldsymbol{x}) + \beta b_i^{(l)}$$

Induction Step

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$
- We know that for $d_1, \ldots, d_{l-1} \to \infty$, $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ is a sum of i.i.d. random variables:

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$
- We know that for $d_1, \ldots, d_{l-1} \to \infty$, $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ is a sum of i.i.d. random variables:

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}) + \beta b_i^{(l)}$$

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$
- We know that for $d_1, \ldots, d_{l-1} \to \infty$, $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ is a sum of i.i.d. random variables:

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}) + \beta b_i^{(l)}$$

- The covariance structure needed for CLT can be written as

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$
- We know that for $d_1, \ldots, d_{l-1} \to \infty$, $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ is a sum of i.i.d. random variables:

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}) + \beta b_i^{(l)}$$

- The covariance structure needed for CLT can be written as

$$\hat{\boldsymbol{\Sigma}}_{st} = \operatorname{cov}\left(W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^s), W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^t)\right)$$

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$
- We know that for $d_1, \ldots, d_{l-1} \to \infty$, $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ is a sum of i.i.d. random variables:

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}) + \beta b_i^{(l)}$$

- The covariance structure needed for CLT can be written as

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{st} &= \operatorname{cov}\left(W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^s), W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^t)\right) \\ &= \mathbb{E}[W_{ik}^{(L)} W_{ik}^{(l)}] \mathbb{E}[\alpha_{k,\infty}^{(l)}(\boldsymbol{x}^s) \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^t)]\end{aligned}$$

Induction Step

- Take $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k) \subset \mathbb{R}^{d_0}$ and look at $\left(f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1), \ldots, f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k)\right)$
- We know that for $d_1, \ldots, d_{l-1} \to \infty$, $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ is a sum of i.i.d. random variables:

$$f_{\boldsymbol{\theta},i}(\boldsymbol{x}) \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}) + \beta b_i^{(l)}$$

- The covariance structure needed for CLT can be written as

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{st} &= \operatorname{cov}\left(W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^s), W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^t)\right) \\
&= \mathbb{E}[W_{ik}^{(L)} W_{ik}^{(l)}] \mathbb{E}[\alpha_{k,\infty}^{(l)}(\boldsymbol{x}^s) \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^t)] \\
&= \mathbb{E}[\alpha_{k,\infty}^{(l)}(\boldsymbol{x}^s) \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^t)]
\end{aligned}$$

Covariance Structure

Covariance Structure

- By the multidimensional CLT we now obtain

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix}$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix}$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

- Moreover, we can write out the covariance function as

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

- Moreover, we can write out the covariance function as

$$\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x'}) = \hat{\Sigma}(\boldsymbol{x}, \boldsymbol{x'}) + \beta^2$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

- Moreover, we can write out the covariance function as

$$\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x'}) = \hat{\Sigma}(\boldsymbol{x}, \boldsymbol{x'}) + \beta^2 = \mathbb{E}[\alpha_{k,\infty}^{(l+1)}(\boldsymbol{x}) \alpha_{k,\infty}^{(l+1)}(\boldsymbol{x'})] + \beta^2$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

- Moreover, we can write out the covariance function as

$$\begin{aligned} \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x'}) = \hat{\Sigma}(\boldsymbol{x}, \boldsymbol{x'}) + \beta^2 &= \mathbb{E}[\alpha_{k,\infty}^{(l+1)}(\boldsymbol{x}) \alpha_{k,\infty}^{(l+1)}(\boldsymbol{x'})] + \beta^2 \\ &= \mathbb{E}\left[\sigma\left(\tilde{\alpha}_{k,\infty}^{(l+1)}(\boldsymbol{x})\right) \sigma\left(\tilde{\alpha}_{k,\infty}^{(l+1)}(\boldsymbol{x'})\right)\right] + \beta^2 \end{aligned}$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

- Moreover, we can write out the covariance function as

$$\begin{aligned} \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x'}) = \hat{\Sigma}(\boldsymbol{x}, \boldsymbol{x'}) + \beta^2 &= \mathbb{E}[\alpha_{k,\infty}^{(l+1)}(\boldsymbol{x}) \alpha_{k,\infty}^{(l+1)}(\boldsymbol{x'})] + \beta^2 \\ &= \mathbb{E}\left[\sigma\left(\tilde{\alpha}_{k,\infty}^{(l+1)}(\boldsymbol{x})\right) \sigma\left(\tilde{\alpha}_{k,\infty}^{(l+1)}(\boldsymbol{x'})\right)\right] + \beta^2 \\ &= \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\sigma\left(z_1\right) \sigma\left(z_2\right)\right] + \beta^2 \end{aligned}$$

Covariance Structure

- By the multidimensional CLT we now obtain

$$\begin{pmatrix} f_{\boldsymbol{\theta},i}(\boldsymbol{x}^1) \\ \vdots \\ f_{\boldsymbol{\theta},i}(\boldsymbol{x}^k) \end{pmatrix} = \frac{1}{\sqrt{d_l}} \sum_{k=1}^{d_l} \begin{pmatrix} W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^1) \\ \vdots \\ W_{ik}^{(l)} \alpha_{k,\infty}^{(l)}(\boldsymbol{x}^k) \end{pmatrix} \xrightarrow{d_l \to \infty} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} + \beta^2\right)$$

- **Hence:** $f_{\boldsymbol{\theta},i}^{\infty}(\cdot) \sim \mathcal{GP}\left(0, \Sigma^{(l+1)}\right)$

- Moreover, we can write out the covariance function as

$$\begin{aligned} \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') &= \hat{\Sigma}(\boldsymbol{x}, \boldsymbol{x}') + \beta^2 = \mathbb{E}[\alpha_{k,\infty}^{(l+1)}(\boldsymbol{x}) \alpha_{k,\infty}^{(l+1)}(\boldsymbol{x}')] + \beta^2 \\ &= \mathbb{E}\left[\sigma\left(\tilde{\alpha}_{k,\infty}^{(l+1)}(\boldsymbol{x})\right) \sigma\left(\tilde{\alpha}_{k,\infty}^{(l+1)}(\boldsymbol{x}')\right)\right] + \beta^2 \\ &= \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\sigma\left(z_1\right) \sigma\left(z_2\right)\right] + \beta^2 \end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}^{(l)} = \begin{pmatrix} \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \\ \Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}) & \Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') \end{pmatrix}$

Independence

Independence

- Finally we have to check **independence**:

Independence

- Finally we have to check **independence**:

$$Q_{ij} = \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x}))$$

Independence

- Finally we have to check **independence**:

$$Q_{ij} = \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x})) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}) f_{\boldsymbol{\theta},j}(\boldsymbol{x})]$$

Independence

- Finally we have to check **independence**:

$$
\begin{aligned}
Q_{ij} &= \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x})) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}) f_{\boldsymbol{\theta},j}(\boldsymbol{x})] \\
&= \frac{1}{d_l} \sum_{k=1}^{d_l} \sum_{s=1}^{d_l} \mathbb{E}[W_{ik}^{(l)} W_{js}^{(l)}] \mathbb{E}[\alpha_k^{(l)}(\boldsymbol{x}) \alpha_s^{(l)}(\boldsymbol{x})] + 0
\end{aligned}
$$

Independence

- Finally we have to check **independence**:

$$
\begin{aligned}
Q_{ij} &= \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x})) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}) f_{\boldsymbol{\theta},j}(\boldsymbol{x})] \\
&= \frac{1}{d_l} \sum_{k=1}^{d_l} \sum_{s=1}^{d_l} \mathbb{E}[W_{ik}^{(l)} W_{js}^{(l)}] \mathbb{E}[\alpha_k^{(l)}(\boldsymbol{x}) \alpha_s^{(l)}(\boldsymbol{x})] + 0 \\
&= 0
\end{aligned}
$$

Independence

- Finally we have to check **independence**:

$$\begin{aligned} Q_{ij} &= \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x})) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}) f_{\boldsymbol{\theta},j}(\boldsymbol{x})] \\ &= \frac{1}{d_l} \sum_{k=1}^{d_l} \sum_{s=1}^{d_l} \mathbb{E}[W_{ik}^{(l)} W_{js}^{(l)}] \mathbb{E}[\alpha_k^{(l)}(\boldsymbol{x}) \alpha_s^{(l)}(\boldsymbol{x})] + 0 \\ &= 0 \end{aligned}$$

- Again by CLT, the two variables $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ and $f_{\boldsymbol{\theta},j}(\boldsymbol{x})$ converge **jointly** to the limit $\mathcal{N}(\mathbf{0}, \boldsymbol{Q})$.

Independence

- Finally we have to check **independence**:

$$\begin{aligned} Q_{ij} &= \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x})) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}) f_{\boldsymbol{\theta},j}(\boldsymbol{x})] \\ &= \frac{1}{d_l} \sum_{k=1}^{d_l} \sum_{s=1}^{d_l} \mathbb{E}[W_{ik}^{(l)} W_{js}^{(l)}] \mathbb{E}[\alpha_k^{(l)}(\boldsymbol{x}) \alpha_s^{(l)}(\boldsymbol{x})] + 0 \\ &= 0 \end{aligned}$$

- Again by CLT, the two variables $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ and $f_{\boldsymbol{\theta},j}(\boldsymbol{x})$ converge **jointly** to the limit $\mathcal{N}(\mathbf{0}, \boldsymbol{Q})$.

- Thus due to Gaussianity, $Q_{ij} = 0$ also implies independence in the limit.

Independence

- Finally we have to check **independence**:

$$\begin{aligned} Q_{ij} &= \mathrm{cov}(f_{\boldsymbol{\theta},i}(\boldsymbol{x}), f_{\boldsymbol{\theta},j}(\boldsymbol{x})) = \mathbb{E}[f_{\boldsymbol{\theta},i}(\boldsymbol{x}) f_{\boldsymbol{\theta},j}(\boldsymbol{x})] \\ &= \frac{1}{d_l} \sum_{k=1}^{d_l} \sum_{s=1}^{d_l} \mathbb{E}[W_{ik}^{(l)} W_{js}^{(l)}] \mathbb{E}[\alpha_k^{(l)}(\boldsymbol{x}) \alpha_s^{(l)}(\boldsymbol{x})] + 0 \\ &= 0 \end{aligned}$$

- Again by CLT, the two variables $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$ and $f_{\boldsymbol{\theta},j}(\boldsymbol{x})$ converge **jointly** to the limit $\mathcal{N}(\mathbf{0}, \boldsymbol{Q})$.
- Thus due to Gaussianity, $Q_{ij} = 0$ also implies independence in the limit.
- This **concludes** the proof.

Some Empirical Illustrations

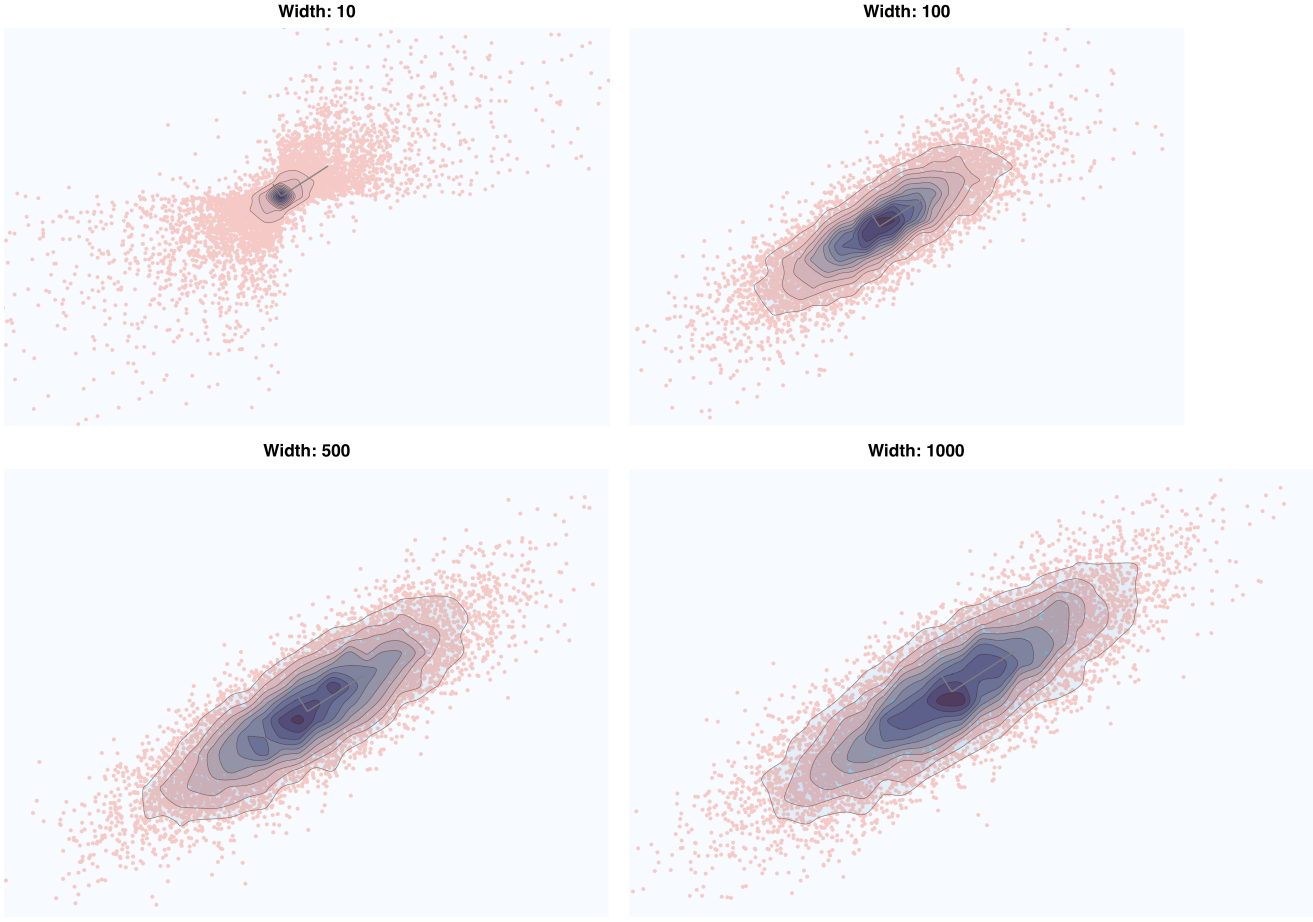
- 8-layer neural network with fixed width d on two inputs $\boldsymbol{x}, \boldsymbol{x'}$ from *MNIST*

Some Empirical Illustrations

- 8-layer neural network with fixed width d on two inputs $\boldsymbol{x}, \boldsymbol{x'}$ from *MNIST*
- Propagated the inputs through 10000 random networks and plotted $f_{\boldsymbol{\theta},1}(\boldsymbol{x})$ vs. $f_{\boldsymbol{\theta},1}(\boldsymbol{x'})$ for $d \in \{10, 100, 500, 1000\}$

Some Empirical Illustrations

- 8-layer neural network with fixed width d on two inputs $\boldsymbol{x}, \boldsymbol{x'}$ from *MNIST*
- Propagated the inputs through 10000 random networks and plotted $f_{\boldsymbol{\theta},1}(\boldsymbol{x})$ vs. $f_{\boldsymbol{\theta},1}(\boldsymbol{x'})$ for $d \in \{10, 100, 500, 1000\}$



Most Important Takeaway (Regarding NTK)

- We can fully characterize the distribution of $f_{\boldsymbol{\theta}}^{\infty}(\cdot)$ **at initialization**

Most Important Takeaway (Regarding NTK)

- We can fully characterize the distribution of $f_{\boldsymbol{\theta}}^{\infty}(\cdot)$ **at initialization**
- We needed a scaling of $\frac{1}{\sqrt{d_l}}$ at each layer for the **asymptotics** to work

Most Important Takeaway (Regarding NTK)

- We can fully characterize the distribution of $f_{\boldsymbol{\theta}}^{\infty}(\cdot)$ **at initialization**
- We needed a scaling of $\frac{1}{\sqrt{d_l}}$ at each layer for the **asymptotics** to work
- We have a **recursive** covariance structure which will reappear for NTK:

$$\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(0, \tilde{\Sigma}^{(l-1)})}\left[\sigma(z_1)\sigma(z_2)\right] + \beta^2$$

Most Important Takeaway (Regarding NTK)

- We can fully characterize the distribution of $f_{\boldsymbol{\theta}}^{\infty}(\cdot)$ **at initialization**
- We needed a scaling of $\frac{1}{\sqrt{d_l}}$ at each layer for the **asymptotics** to work
- We have a **recursive** covariance structure which will reappear for NTK:

$$\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(0, \tilde{\Sigma}^{(l-1)})}\left[\sigma(z_1)\sigma(z_2)\right] + \beta^2$$

- Can we describe $f_{\boldsymbol{\theta}}^{\infty}(\cdot)$ during **training**?

Discussion

Discussion

- All results only hold at **initialization**

Discussion

- All results only hold at **initialization**
- Seems hard to assess how the correlation structure between weights changes over course of training

Discussion

- All results only hold at **initialization**
- Seems hard to assess how the correlation structure between weights changes over course of training
- Even if we know how correlation evolves, CLT with correlated variables are hard to apply

Discussion

- All results only hold at **initialization**
- Seems hard to assess how the correlation structure between weights changes over course of training
- Even if we know how correlation evolves, CLT with correlated variables are hard to apply
- Infinite-width limit interesting because generalization gets better with **increasing width** (empirically)