Spectrum-Dependent Learning Curves in Kernel Regression and Wide Neural Networks

Blake Bordelon, Abdulkadir Canatar, Cengiz Pehlevan

Notation

Notation

- We have training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{p} \in \mathcal{X} \times \mathbb{R}$

Notation

- We have training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{p} \in \mathcal{X} \times \mathbb{R}$
- We consider some function space $\mathcal{F}$

Notation

- We have training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{p} \in \mathcal{X} \times \mathbb{R}$
- We consider some function space $\mathcal{F}$
- Assume that the data are labeled by some $f^* \in \mathcal{F}$:

$$y_i = f^*(\boldsymbol{x}_i)$$

Notation

- We have training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{p} \in \mathcal{X} \times \mathbb{R}$
- We consider some function space $\mathcal{F}$
- Assume that the data are labeled by some $f^* \in \mathcal{F}$:

$$y_i = f^*(\boldsymbol{x}_i)$$

- Data follows some distribution $\boldsymbol{x} \sim p$

Reproducing Kernel Hilbert Space

Reproducing Kernel Hilbert Space

- Choose $\mathcal{F}$ to be a Reproducing Kernel Hilbert Space (RKHS), denote $\mathcal{H} = \mathcal{F}$

Reproducing Kernel Hilbert Space

- Choose $\mathcal{F}$ to be a Reproducing Kernel Hilbert Space (RKHS), denote $\mathcal{H} = \mathcal{F}$

Every Cauchy sequence converges

- Hilbert space $\implies$ Banach space (also called complete) and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$

Reproducing Kernel Hilbert Space

- Choose $\mathcal{F}$ to be a Reproducing Kernel Hilbert Space (RKHS), denote $\mathcal{H} = \mathcal{F}$

Every Cauchy sequence converges

- Hilbert space $\implies$ Banach space (also called complete) and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$

- Being RKHS means:
  $\exists!$ $K(\cdot, \cdot)$ symmetric, positive semi-definite fulfilling the reproducing property:

$$\forall f \in \mathcal{F} : f(\boldsymbol{x}) = \langle K(\boldsymbol{x}, \cdot), f \rangle_{\mathcal{H}}$$

Reproducing Kernel Hilbert Space

- Choose $\mathcal{F}$ to be a Reproducing Kernel Hilbert Space (RKHS), denote $\mathcal{H} = \mathcal{F}$

Every Cauchy sequence converges

- Hilbert space $\implies$ Banach space (also called complete) and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
- Being RKHS means:
  $\exists!$ $K(\cdot, \cdot)$ symmetric, positive semi-definite fulfilling the reproducing property:

$$\forall f \in \mathcal{F} : f(\boldsymbol{x}) = \langle K(\boldsymbol{x}, \cdot), f \rangle_{\mathcal{H}}$$

- Equivalent to requiring that the evaluation functional $L_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$ is continuous (which in turn is equivalent to boundedness of the linear functional):

$$\forall \boldsymbol{x} \;\; |L_{\boldsymbol{x}}(f)| \leq M ||f||_{\mathcal{H}}$$

Kernel Ridge(less) Regression

Kernel Ridge(less) Regression

- $\min_{f \in \mathcal{H}} \sum_{i=1}^{p} \left(f(\boldsymbol{x}_i) - y_i\right)^2 + \lambda ||f||_{\mathcal{H}}^2$

Kernel Ridge(less) Regression

- $\min_{f \in \mathcal{H}} \sum_{i=1}^{p} (f(\boldsymbol{x}_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}}^2$
- Define the subspace $\mathbb{L} = \left\{ \sum_{i=1}^{p} \alpha_i K(\boldsymbol{x}_i, \cdot) : \boldsymbol{\alpha} \in \mathbb{R}^p \right\}$ and decompose any $f \in \mathcal{H}$ as $f = f_{\mathbb{L}} + f_{\perp}$ with $f_{\perp}$ orthogonal to $\mathbb{L}$

Kernel Ridge(less) Regression

- $\min_{f \in \mathcal{H}} \sum_{i=1}^{p} (f(\boldsymbol{x}_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}}^2$
- Define the subspace $\mathbb{L} = \left\{ \sum_{i=1}^{p} \alpha_i K(\boldsymbol{x}_i, \cdot) : \boldsymbol{\alpha} \in \mathbb{R}^p \right\}$ and decompose any $f \in \mathcal{H}$ as $f = f_{\mathbb{L}} + f_{\perp}$ with $f_{\perp}$ orthogonal to $\mathbb{L}$
- **RKHS Magic**: $f(\boldsymbol{x}_i) = \langle f_{\mathbb{L}} + f_{\perp}, K(\cdot, \boldsymbol{x}_i) \rangle_{\mathcal{H}} = f_{\mathbb{L}}(\boldsymbol{x}_i)$

  For ridgeless case ($\lambda = 0$), infinite dimensional problem reduced to finding the best p-dimensional parameter $\boldsymbol{\alpha}$

Kernel Ridge(less) Regression

- $\min_{f \in \mathcal{H}} \sum_{i=1}^{p} (f(\boldsymbol{x}_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}}^2$

- Define the subspace $\mathbb{L} = \left\{ \sum_{i=1}^{p} \alpha_i K(\boldsymbol{x}_i, \cdot) : \boldsymbol{\alpha} \in \mathbb{R}^p \right\}$ and decompose any $f \in \mathcal{H}$ as $f = f_{\mathbb{L}} + f_{\perp}$ with $f_{\perp}$ orthogonal to $\mathbb{L}$

- **RKHS Magic**: $f(\boldsymbol{x}_i) = \langle f_{\mathbb{L}} + f_{\perp}, K(\cdot, \boldsymbol{x}_i) \rangle_{\mathcal{H}} = f_{\mathbb{L}}(\boldsymbol{x}_i)$

  For ridgeless case ($\lambda = 0$), infinite dimensional problem reduced to finding the best p-dimensional parameter $\boldsymbol{\alpha}$

- $f(\boldsymbol{x}) = \boldsymbol{y}^T (\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \mathbb{1})^{-1} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})$

Kernel Ridge(less) Regression

- $\min_{f \in \mathcal{H}} \sum_{i=1}^{p} (f(\boldsymbol{x}_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}}^2$
- Define the subspace $\mathbb{L} = \left\{ \sum_{i=1}^{p} \alpha_i K(\boldsymbol{x}_i, \cdot) : \boldsymbol{\alpha} \in \mathbb{R}^p \right\}$ and decompose any $f \in \mathcal{H}$ as $f = f_{\mathbb{L}} + f_{\perp}$ with $f_{\perp}$ orthogonal to $\mathbb{L}$
- **RKHS Magic**: $f(\boldsymbol{x}_i) = \langle f_{\mathbb{L}} + f_{\perp}, K(\cdot, \boldsymbol{x}_i) \rangle_{\mathcal{H}} = f_{\mathbb{L}}(\boldsymbol{x}_i)$

  For ridgeless case ($\lambda = 0$), infinite dimensional problem reduced to finding the best p-dimensional parameter $\boldsymbol{\alpha}$
- $f(\boldsymbol{x}) = \boldsymbol{y}^T (\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \mathbb{1})^{-1} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})$
- **Remember:** $f(\boldsymbol{x}) = \boldsymbol{y}^T \boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{X})$

Predictive function of NTK

Uniform Convergence vs Average Case Analysis

Uniform Convergence vs Average Case Analysis

- Entire statistical learning theory based on a **worst case** analysis:

$$\mathbb{P}_{\boldsymbol{X}\sim p}\left(\sup_{h\in\mathcal{H}}|L_{\text{train}}(h)-L_{\text{gen}}(h)|\right)\geq 1-\delta$$

Uniform Convergence vs Average Case Analysis

- Entire statistical learning theory based on a **worst case** analysis:

$$\mathbb{P}_{\boldsymbol{X}\sim p}\left(\sup_{h\in\mathcal{H}}|L_{\text{train}}(h)-L_{\text{gen}}(h)|\right)\geq 1-\delta$$

- The goal here is to provide an expression for the **average case** analysis of the generalization error. This is very common in statistical mechanics. We will use some of its tools here.

Uniform Convergence vs Average Case Analysis

- Entire statistical learning theory based on a **worst case** analysis:

$$\mathbb{P}_{\boldsymbol{X}\sim p}\left(\sup_{h\in\mathcal{H}}|L_{\text{train}}(h)-L_{\text{gen}}(h)|\right)\geq 1-\delta$$

- The goal here is to provide an expression for the **average case** analysis of the generalization error. This is very common in statistical mechanics. We will use some of its tools here.

- Sometimes, expectations will be denoted in physics-fashion:

$$\langle f(\boldsymbol{X})\rangle_{\boldsymbol{X}}=\mathbb{E}_{\boldsymbol{X}\sim p}[f(\boldsymbol{X})]$$

Generalization Errors

Generalization Errors

- Denote by $f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*)$ the function learnt with kernel regression

Generalization Errors

- Denote by $f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*)$ the function learnt with kernel regression
- $E_g(\boldsymbol{X}, f^*) = \mathbb{E}_{\boldsymbol{x} \sim p}\Big[\left(f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*) - f^*(\boldsymbol{x})\right)^2 \Big]$

Generalization Errors

- Denote by $f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*)$ the function learnt with kernel regression
- $E_g(\boldsymbol{X}, f^*) = \mathbb{E}_{\boldsymbol{x} \sim p}\Big[\left(f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*) - f^*(\boldsymbol{x})\right)^2 \Big]$
- $E_g = \langle E_g(\boldsymbol{X}, f^*) \rangle_{\boldsymbol{X}, f^*}$ (Integrate over data and teacher)

Generalization Errors

- Denote by $f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*)$ the function learnt with kernel regression
- $E_g(\boldsymbol{X}, f^*) = \mathbb{E}_{\boldsymbol{x}\sim p}\Big[\, (f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*) - f^*(\boldsymbol{x}))^2 \,\Big]$
- $E_g = \langle E_g(\boldsymbol{X}, f^*)\rangle_{\boldsymbol{X}, f^*}$ (Integrate over data and teacher)
- **Observe:** Fixed teacher f^* is included in this analysis by setting $p_{f^*} \sim \delta_{f^*}$

Generalization Errors

- Denote by $f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*)$ the function learnt with kernel regression
- $E_g(\boldsymbol{X}, f^*) = \mathbb{E}_{\boldsymbol{x} \sim p}\Big[\left(f_{\boldsymbol{K}}(\boldsymbol{x}; \boldsymbol{X}, f^*) - f^*(\boldsymbol{x}) \right)^2 \Big]$
- $E_g = \langle E_g(\boldsymbol{X}, f^*) \rangle_{\boldsymbol{X}, f^*}$ (Integrate over data and teacher)
- **Observe:** Fixed teacher f^* is included in this analysis by setting $p_{f^*} \sim \delta_{f^*}$
- **Goal:** Calculate E_g for any kernel $\boldsymbol{K}$ and any teacher distribution

Mercer Decomposition

Mercer Decomposition

- Due to Mercer, we can decompose any kernel into a (possibly) infinite eigenbasis:

$\psi_i(x) = \sqrt{\lambda_i}\,\phi_i(x)$

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{M} \lambda_i \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{x}') = \sum_{i=1}^{M} \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}') = \psi(\boldsymbol{x})^T \psi(\boldsymbol{x}')$$

Feature representation of sample x

where ϕ_i fulfills $\int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{y}) \phi_i(\boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} = \lambda_i \phi_i(\boldsymbol{x})$

Mercer Decomposition

- Due to Mercer, we can decompose any kernel into a (possibly) infinite eigenbasis:

$\psi_i(x) = \sqrt{\lambda_i}\,\phi_i(x)$

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{M} \lambda_i \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{x}') = \sum_{i=1}^{M} \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}') = \psi(\boldsymbol{x})^T \psi(\boldsymbol{x}')$$

Feature representation of sample x

where ϕ_i fulfills $\int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{y}) \phi_i(\boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} = \lambda_i \phi_i(\boldsymbol{x})$

- $\{\psi_i(\cdot)\}_{i=1}^{M}$ form a basis of $\mathcal{H}$

Mercer Decomposition

- Due to Mercer, we can decompose any kernel into a (possibly) infinite eigenbasis:

$\psi_i(x) = \sqrt{\lambda_i}\,\phi_i(x)$

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{M} \lambda_i \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{x}') = \sum_{i=1}^{M} \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}') = \psi(\boldsymbol{x})^T \psi(\boldsymbol{x}')$$

Feature representatio of sample x

where ϕ_i fulfills $\int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{y}) \phi_i(\boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} = \lambda_i \phi_i(\boldsymbol{x})$

- $\{\psi_i(\cdot)\}_{i=1}^{M}$ form a basis of $\mathcal{H}$

- Write $f^*(\boldsymbol{x}) = \sum_{i=1}^{M} \bar{w}_i \psi_i(\boldsymbol{x})$ and $f(\boldsymbol{x}) = \sum_{i=1}^{M} w_i \psi_i(\boldsymbol{x})$

Kernel Regression in Feature Space

Kernel Regression in Feature Space

- Define the feature matrix

$$\mathbf{\Psi} = \begin{bmatrix} — & \psi(\boldsymbol{x}_1) & — \\ & \vdots & \\ — & \psi(\boldsymbol{x}_p) & — \end{bmatrix} = \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \dots & \psi_M(\boldsymbol{x}_1) \\ \vdots & & \vdots \\ \psi_1(\boldsymbol{x}_p) & \dots & \psi_M(\boldsymbol{x}_p) \end{bmatrix} \in \mathbb{R}^{M \times p}$$

Kernel Regression in Feature Space

- Define the feature matrix

$$\mathbf{\Psi} = \begin{bmatrix} — & \psi(\boldsymbol{x}_1) & — \\ & \vdots & \\ — & \psi(\boldsymbol{x}_p) & — \end{bmatrix} = \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \dots & \psi_M(\boldsymbol{x}_1) \\ \vdots & & \vdots \\ \psi_1(\boldsymbol{x}_p) & \dots & \psi_M(\boldsymbol{x}_p) \end{bmatrix} \in \mathbb{R}^{M \times p}$$

- Rewrite the kernel as $K(\boldsymbol{X}, \boldsymbol{X}) = \mathbf{\Psi}\mathbf{\Psi}^T$

Kernel Regression in Feature Space

- Define the feature matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} — & \psi(\boldsymbol{x}_1) & — \\ & \vdots & \\ — & \psi(\boldsymbol{x}_p) & — \end{bmatrix} = \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \dots & \psi_M(\boldsymbol{x}_1) \\ \vdots & & \vdots \\ \psi_1(\boldsymbol{x}_p) & \dots & \psi_M(\boldsymbol{x}_p) \end{bmatrix} \in \mathbb{R}^{M \times p}$$

- Rewrite the kernel as $K(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{\Psi}\boldsymbol{\Psi}^T$
- $\min_{\boldsymbol{w} \in \mathbb{R}^M} ||\boldsymbol{\Psi}^T \boldsymbol{w} - \boldsymbol{y}||_2^2 + \lambda ||\boldsymbol{w}||_2^2$

Kernel Regression in Feature Space

- Define the feature matrix

$$\mathbf{\Psi} = \begin{bmatrix} — & \psi(\boldsymbol{x}_1) & — \\ & \vdots & \\ — & \psi(\boldsymbol{x}_p) & — \end{bmatrix} = \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \dots & \psi_M(\boldsymbol{x}_1) \\ \vdots & & \vdots \\ \psi_1(\boldsymbol{x}_p) & \dots & \psi_M(\boldsymbol{x}_p) \end{bmatrix} \in \mathbb{R}^{M \times p}$$

- Rewrite the kernel as $K(\boldsymbol{X}, \boldsymbol{X}) = \mathbf{\Psi}\mathbf{\Psi}^T$
- $\min_{\boldsymbol{w} \in \mathbb{R}^M} ||\mathbf{\Psi}^T \boldsymbol{w} - \boldsymbol{y}||_2^2 + \lambda ||\boldsymbol{w}||_2^2$
- $\boldsymbol{w} = \left(\mathbf{\Psi}\mathbf{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \mathbf{\Psi}\boldsymbol{y} \in \mathbb{R}^M$

Decomposition of Generalization Error

Decomposition of Generalization Error

Lemma 1:
We have the following decompostion of the generalization error:

$$E_g = \sum_{i=1}^{M} E_i$$

where the modes are given by $E_i = \lambda_i \left\langle (w_i - \bar{w}_i)^2 \right\rangle_{\boldsymbol{X},\bar{\boldsymbol{w}}}$

Decomposition of Generalization Error

Lemma 1:
We have the following decompostion of the generalization error:

$$E_g = \sum_{i=1}^{M} E_i$$

where the modes are given by $E_i = \lambda_i \left\langle (w_i - \bar{w}_i)^2 \right\rangle_{\boldsymbol{X},\bar{\boldsymbol{w}}}$

This provides in some sense a **spectral decomposition** of the generalization error as each term is weighted by the corresponding eigenvalue of the kernel.

Derivation

Derivation

$$E_g = \langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \rangle_{\boldsymbol{x}, \boldsymbol{X}, f^*}$$

Derivation

$$E_g = \langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \rangle_{\boldsymbol{x},\boldsymbol{X},f^*} = \left\langle \left(\sum_{i=1}^{M} (w_i - \bar{w}_i)\psi_i(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*}$$

Derivation

$$E_g = \langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \rangle_{\boldsymbol{x},\boldsymbol{X},f^*} = \left\langle \left(\sum_{i=1}^{M} (w_i - \bar{w}_i)\psi_i(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*}$$

$$= \left\langle \left(\sum_{i,j=1}^{M} (w_i - \bar{w}_i)(w_j - \bar{w}_j)\psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*}$$

Derivation

$$
\begin{aligned}
E_g = \langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \rangle_{\boldsymbol{x},\boldsymbol{X},f^*} &= \left\langle \left(\sum_{i=1}^{M} (w_i - \bar{w}_i)\psi_i(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*} \\
&= \left\langle \left(\sum_{i,j=1}^{M} (w_i - \bar{w}_i)(w_j - \bar{w}_j)\psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*} \\
&= \sum_{i,j=1}^{M} \Big\langle (w_i - \bar{w}_i)(w_j - \bar{w}_j) \big\langle \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \big\rangle_{\boldsymbol{x}} \Big\rangle_{\boldsymbol{X},f^*}
\end{aligned}
$$

Derivation

$$
\begin{aligned}
E_g = \langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \rangle_{\boldsymbol{x},\boldsymbol{X},f^*} &= \left\langle \left(\sum_{i=1}^{M} (w_i - \bar{w}_i)\psi_i(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*} \\
&= \left\langle \left(\sum_{i,j=1}^{M} (w_i - \bar{w}_i)(w_j - \bar{w}_j)\psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*} \\
&= \sum_{i,j=1}^{M} \left\langle (w_i - \bar{w}_i)(w_j - \bar{w}_j) \big\langle \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \big\rangle_{\boldsymbol{x}} \right\rangle_{\boldsymbol{X},f^*}
\end{aligned}
$$

Now $\big\langle \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \big\rangle_{\boldsymbol{x}} = \int_{\mathcal{X}} \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{1}_{\{i=j\}} \lambda_i$

Derivation

$$E_g = \langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \rangle_{\boldsymbol{x},\boldsymbol{X},f^*} = \left\langle \left(\sum_{i=1}^{M} (w_i - \bar{w}_i)\psi_i(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*}$$

$$= \left\langle \left(\sum_{i,j=1}^{M} (w_i - \bar{w}_i)(w_j - \bar{w}_j)\psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \right)^2 \right\rangle_{\boldsymbol{x},\boldsymbol{X},f^*}$$

$$= \sum_{i,j=1}^{M} \left\langle (w_i - \bar{w}_i)(w_j - \bar{w}_j) \left\langle \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} \right\rangle_{\boldsymbol{X},f^*}$$

Now $\left\langle \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} = \int_{\mathcal{X}} \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \mathbb{1}_{\{i=j\}}\lambda_i$

$$E_g = \sum_{i=1}^{M} \lambda_i \left\langle (w_i - \bar{w}_i)^2 \right\rangle_{\boldsymbol{X},f^*} = \sum_{i=1}^{M} E_i$$

Theorem 2:

Theorem 2:

We have the following expression for the generalization error under the minimizer $\boldsymbol{w}$:

$$E_g = \operatorname{tr}\left(\boldsymbol{D}\left\langle \boldsymbol{G}^2 \right\rangle_{\boldsymbol{X}}\right)$$

Theorem 2:

We have the following expression for the generalization error under the minimizer $\boldsymbol{w}$*:*

$$E_g = \mathrm{tr}\left(\boldsymbol{D}\left\langle \boldsymbol{G}^2 \right\rangle_{\boldsymbol{X}}\right)$$

where we define the matrices

- $\boldsymbol{G} = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1}\right)^{-1}$
- $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi}$
- $\boldsymbol{D} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \right\rangle_{\bar{\boldsymbol{w}}} \boldsymbol{\Lambda}^{-\frac{1}{2}}$

$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_M)$

Theorem 2:

We have the following expression for the generalization error under the minimizer $\boldsymbol{w}$:

$$E_g = \operatorname{tr}\left(\boldsymbol{D}\left\langle \boldsymbol{G}^2\right\rangle_{\boldsymbol{X}}\right)$$

where we define the matrices

- $\boldsymbol{G} = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1}\right)^{-1}$
- $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi}$ $\quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_M)$
- $\boldsymbol{D} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\right\rangle_{\bar{\boldsymbol{w}}}\boldsymbol{\Lambda}^{-\frac{1}{2}}$

Observe that:

Theorem 2:
We have the following expression for the generalization error under the minimizer $\boldsymbol{w}$*:*

$$E_g = \mathrm{tr}\left(\boldsymbol{D}\left\langle \boldsymbol{G}^2\right\rangle_{\boldsymbol{X}}\right)$$

where we define the matrices

- $\boldsymbol{G} = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1}\right)^{-1}$
- $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi}$ $\quad\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_M)$
- $\boldsymbol{D} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\right\rangle_{\bar{\boldsymbol{w}}}\boldsymbol{\Lambda}^{-\frac{1}{2}}$

Observe that:

- Nice separation of teacher-dependence ($\boldsymbol{D}$) and data-dependence ($\boldsymbol{G}$).

Theorem 2:

We have the following expression for the generalization error under the minimizer $\boldsymbol{w}$*:*

$$E_g = \operatorname{tr}\left(\boldsymbol{D}\left\langle \boldsymbol{G}^2 \right\rangle_{\boldsymbol{X}}\right)$$

where we define the matrices

- $\boldsymbol{G} = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1}\right)^{-1}$
- $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi}$
- $\boldsymbol{D} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \right\rangle_{\bar{\boldsymbol{w}}}\boldsymbol{\Lambda}^{-\frac{1}{2}}$

$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_M)$

Observe that:

- Nice separation of teacher-dependence ($\boldsymbol{D}$) and data-dependence ($\boldsymbol{G}$).
- Calculating $\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \right\rangle_{\bar{\boldsymbol{w}}}$ is relatively easy (if the distribution is centered, this is just the covariance matrix)

Theorem 2:

We have the following expression for the generalization error under the minimizer $\boldsymbol{w}$*:*

$$E_g = \operatorname{tr}\left(\boldsymbol{D}\left\langle \boldsymbol{G}^2\right\rangle_{\boldsymbol{X}}\right)$$

where we define the matrices

- $\boldsymbol{G} = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1}\right)^{-1}$
- $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi}$
- $\boldsymbol{D} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\right\rangle_{\bar{\boldsymbol{w}}}\boldsymbol{\Lambda}^{-\frac{1}{2}}$

$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)$

Observe that:

- Nice separation of teacher-dependence ($\boldsymbol{D}$) and data-dependence ($\boldsymbol{G}$).
- Calculating $\left\langle \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\right\rangle_{\bar{\boldsymbol{w}}}$ is relatively easy (if the distribution is centered, this is just the covariance matrix)
- Calculating $\left\langle \boldsymbol{G}^2\right\rangle_{\boldsymbol{X}}$ is difficult

Derivation

Derivation

same steps as in Lemma

$$E_g(\boldsymbol{X}, f^*) = \left\langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Derivation

same steps as in Lemma

$$E_g\left(\boldsymbol{X}, f^*\right) = \left\langle\, (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \,\right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

Derivation

same steps as in Lemma

$$E_g\left(\boldsymbol{X}, f^*\right) = \left\langle \left(f(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2 \right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

$$\boldsymbol{w} = \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^T \bar{\boldsymbol{w}} = \bar{\boldsymbol{w}} - \lambda \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Derivation

same steps as in Lemma

$$E_g(\boldsymbol{X}, f^*) = \left\langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

$$\boldsymbol{w} = \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^T \bar{\boldsymbol{w}} = \bar{\boldsymbol{w}} - \lambda \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Then we get that

Derivation

same steps as in Lemma

$$E_g(\boldsymbol{X}, f^*) = \left\langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

$$\boldsymbol{w} = \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^T \bar{\boldsymbol{w}} = \bar{\boldsymbol{w}} - \lambda \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Then we get that

$$E_g(\boldsymbol{X}, f^*) = \lambda^2 \bar{\boldsymbol{w}}^T \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Derivation

same steps as in Lemma

$$E_g\left(\boldsymbol{X}, f^*\right) = \left\langle\, (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \,\right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

$$\boldsymbol{w} = \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^T \bar{\boldsymbol{w}} = \bar{\boldsymbol{w}} - \lambda \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Then we get that

$$\begin{aligned} E_g\left(\boldsymbol{X}, f^*\right) &= \lambda^2 \bar{\boldsymbol{w}}^T \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}} \\ &= \lambda^2 \bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \bar{\boldsymbol{w}} \end{aligned}$$

Derivation

same steps as in Lemma

$$E_g\left(\boldsymbol{X}, f^*\right) = \left\langle \left(f(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2 \right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

$$\boldsymbol{w} = \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^T \bar{\boldsymbol{w}} = \bar{\boldsymbol{w}} - \lambda \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Then we get that

$$\begin{aligned}
E_g\left(\boldsymbol{X}, f^*\right) &= \lambda^2 \bar{\boldsymbol{w}}^T \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}} \\
&= \lambda^2 \bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \bar{\boldsymbol{w}} \\
&= \bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{G}\boldsymbol{G} \boldsymbol{\Lambda}^{-\frac{1}{2}} \bar{\boldsymbol{w}} = \operatorname{tr}\left(\bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{G}\boldsymbol{G} \boldsymbol{\Lambda}^{-\frac{1}{2}} \bar{\boldsymbol{w}}\right)
\end{aligned}$$

Derivation

same steps as in Lemma

$$E_g(\boldsymbol{X}, f^*) = \left\langle (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \right\rangle_{\boldsymbol{x}} = (\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Lambda} (\boldsymbol{w} - \bar{\boldsymbol{w}})$$

Let's recall the minimal solution (using $\boldsymbol{y} = f^*(\boldsymbol{x}) = \boldsymbol{\Psi}^T \bar{\boldsymbol{w}}$)

$$\boldsymbol{w} = \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^T \bar{\boldsymbol{w}} = \bar{\boldsymbol{w}} - \lambda \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}}$$

Then we get that

$$\begin{aligned}
E_g(\boldsymbol{X}, f^*) &= \lambda^2 \bar{\boldsymbol{w}}^T \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \bar{\boldsymbol{w}} \\
&= \lambda^2 \bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T + \lambda \mathbb{1}\right)^{-1} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{\frac{1}{2}} \bar{\boldsymbol{w}} \\
&= \bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{G}\boldsymbol{G} \boldsymbol{\Lambda}^{-\frac{1}{2}} \bar{\boldsymbol{w}} = \operatorname{tr}\left(\bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{G}\boldsymbol{G} \boldsymbol{\Lambda}^{-\frac{1}{2}} \bar{\boldsymbol{w}}\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}} \bar{\boldsymbol{w}} \bar{\boldsymbol{w}}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{G}\boldsymbol{G}\right) = \operatorname{tr}\left(\boldsymbol{D}\boldsymbol{G}^2\right)
\end{aligned}$$

How to calculate $\langle G^2 \rangle_X$?

How to calculate $\langle G^2 \rangle_X$?

Two approaches outlined in the paper:

How to calculate $\langle G^2 \rangle_X$?

Two approaches outlined in the paper:

- Turn this expectation into a **PDE** by deriving a recursion in the number of samples. Viewing the sample size p as a **continuous parameter** will lead to the PDE formulation

How to calculate $\langle G^2 \rangle_{\boldsymbol{X}}$?

Two approaches outlined in the paper:

- Turn this expectation into a **PDE** by deriving a recursion in the number of samples. Viewing the sample size p as a **continuous parameter** will lead to the PDE formulation
- Use the (infamous) **Replica trick** to brute-force calculate/approximate the expectation

How to calculate $\langle G^2 \rangle_X$?

Two approaches outlined in the paper:

- Turn this expectation into a **PDE** by deriving a recursion in the number of samples. Viewing the sample size p as a **continuous parameter** will lead to the PDE formulation
- Use the (infamous) **Replica trick** to brute-force calculate/approximate the expectation
- Both approximations **agree**!

PDE Formulation (I)

PDE Formulation (I)

- **Trick:** Introduce $\tilde{\boldsymbol{G}}(p, v) = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$, then

PDE Formulation (I)

- **Trick:** Introduce $\tilde{\boldsymbol{G}}(p, v) = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$, then

$$\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}} = -\frac{\partial}{\partial v} \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{X}} \Big|_{v=0}$$

PDE Formulation (I)

- **Trick:** Introduce $\tilde{\boldsymbol{G}}(p, v) = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$, then

$$\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}} = -\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{X}}\Big|_{v=0}$$

Just the basis matrix from Mercer

- **Recall**: $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi} \in \mathbb{R}^{M\times p}$. Add a sample: $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{M\times(p+1)}$

PDE Formulation (I)

- **Trick:** Introduce $\tilde{\boldsymbol{G}}(p, v) = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$, then

$$\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}} = -\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{X}}\Big|_{v=0}$$

Just the basis matrix from Mercer

- **Recall**: $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi} \in \mathbb{R}^{M \times p}$. Add a sample: $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{M \times (p+1)}$

$$\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T = \begin{bmatrix} | & & | & | \\ \phi_1 & \dots & \phi_p & \phi_{p+1} \\ | & & | & | \end{bmatrix} \begin{bmatrix} — & \phi_1^T & — \\ & \vdots & \\ — & \phi_p^T & — \\ — & \phi_{p+1}^T & — \end{bmatrix}$$

PDE Formulation (I)

- **Trick:** Introduce $\tilde{\boldsymbol{G}}(p,v) = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$, then

$$\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}} = -\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v)\right\rangle_{\boldsymbol{X}}\Big|_{v=0}$$

Just the basis matrix from Mercer

- **Recall**: $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi} \in \mathbb{R}^{M\times p}$. Add a sample: $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{M\times(p+1)}$

$$\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T = \begin{bmatrix} | & & | & | \\ \phi_1 & \dots & \phi_p & \phi_{p+1} \\ | & & | & | \end{bmatrix} \begin{bmatrix} — & \phi_1^T & — \\ & \vdots & \\ — & \phi_p^T & — \\ — & \phi_{p+1}^T & — \end{bmatrix}$$

$$= \begin{bmatrix} & | \\ \boldsymbol{\Phi} & \phi_{p+1} \\ & | \end{bmatrix} \begin{bmatrix} & \boldsymbol{\Phi}^T & \\ — & \phi_{p+1}^T & — \end{bmatrix}$$

PDE Formulation (I)

- **Trick:** Introduce $\tilde{\boldsymbol{G}}(p,v) = \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$, then

$$\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}} = -\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v) \right\rangle_{\boldsymbol{X}}\Big|_{v=0}$$

Just the basis matrix from Mercer

- **Recall**: $\boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi} \in \mathbb{R}^{M\times p}$. Add a sample: $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{M\times(p+1)}$

$$\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T = \begin{bmatrix} | & & | & | \\ \phi_1 & \dots & \phi_p & \phi_{p+1} \\ | & & | & | \end{bmatrix} \begin{bmatrix} — & \phi_1^T & — \\ & \vdots & \\ — & \phi_p^T & — \\ — & \phi_{p+1}^T & — \end{bmatrix}$$

$$= \begin{bmatrix} & | \\ \boldsymbol{\Phi} & \phi_{p+1} \\ & | \end{bmatrix} \begin{bmatrix} & \boldsymbol{\Phi}^T & \\ — & \phi_{p+1}^T & — \end{bmatrix} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \phi_{p+1}\phi_{p+1}^T$$

PDE Formulation (II)

PDE Formulation (II)

- We can use the Sherman-Woodbury formula:

PDE Formulation (II)

- We can use the Sherman-Woodbury formula:

$$\tilde{\boldsymbol{G}}(p+1, v) = \left(\frac{1}{\lambda} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Phi}}^{T} + \boldsymbol{\Lambda}^{-1} + v\mathbb{1} \right)^{-1}$$

PDE Formulation (II)

$(A+\underline{u}\underline{v}^T)^{-1} = A^{-1} + \frac{A^{-1}\underline{u}\underline{v}^T A^{-1}}{1+\underline{v}^T A^{-1}\underline{u}}$

- We can use the Sherman-Woodbury formula:

$$\tilde{\boldsymbol{G}}(p+1, v) = \left(\frac{1}{\lambda}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$$

$$= \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$$

PDE Formulation (II)

- We can use the Sherman-Woodbury formula:

$$\tilde{\boldsymbol{G}}(p+1, v) = \left(\frac{1}{\lambda}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$$

$$= \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1}$$

$$= \left(\tilde{\boldsymbol{G}}(p, v)^{-1} + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T\right)^{-1}$$

PDE Formulation (II)

- We can use the Sherman-Woodbury formula:

$$
\begin{aligned}
\tilde{\boldsymbol{G}}(p+1, v) &= \left(\frac{1}{\lambda}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} \\
&= \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} \\
&= \left(\tilde{\boldsymbol{G}}(p, v)^{-1} + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T\right)^{-1} \\
&= \tilde{\boldsymbol{G}}(p, v) - \frac{\tilde{\boldsymbol{G}}(p, v)\frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T\tilde{\boldsymbol{G}}(p, v)}{1 + \phi_{p+1}^T\tilde{\boldsymbol{G}}(p, v)\phi_{p+1}}
\end{aligned}
$$

PDE Formulation (II)

- We can use the Sherman-Woodbury formula:

$$\begin{aligned}
\tilde{\boldsymbol{G}}(p+1, v) &= \left(\frac{1}{\lambda}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} \\
&= \left(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T + \boldsymbol{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} \\
&= \left(\tilde{\boldsymbol{G}}(p, v)^{-1} + \frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T\right)^{-1} \\
&= \tilde{\boldsymbol{G}}(p, v) - \frac{\tilde{\boldsymbol{G}}(p, v)\frac{1}{\lambda}\phi_{p+1}\phi_{p+1}^T\tilde{\boldsymbol{G}}(p, v)}{1 + \phi_{p+1}^T\tilde{\boldsymbol{G}}(p, v)\phi_{p+1}}
\end{aligned}$$

- Taking the expectation over the data $\boldsymbol{X}$:

PDE Formulation (II)

- We can use the Sherman-Woodbury formula:

$$
\begin{aligned}
\tilde{\boldsymbol{G}}(p+1, v) &= \left(\frac{1}{\lambda} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Phi}}^T + \boldsymbol{\Lambda}^{-1} + v \mathbb{1} \right)^{-1} \\
&= \left(\frac{1}{\lambda} \boldsymbol{\Phi} \boldsymbol{\Phi}^T + \frac{1}{\lambda} \phi_{p+1} \phi_{p+1}^T + \boldsymbol{\Lambda}^{-1} + v \mathbb{1} \right)^{-1} \\
&= \left(\tilde{\boldsymbol{G}}(p, v)^{-1} + \frac{1}{\lambda} \phi_{p+1} \phi_{p+1}^T \right)^{-1} \\
&= \tilde{\boldsymbol{G}}(p, v) - \frac{\tilde{\boldsymbol{G}}(p, v) \frac{1}{\lambda} \phi_{p+1} \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v)}{1 + \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v) \phi_{p+1}}
\end{aligned}
$$

- Taking the expectation over the data $\boldsymbol{X}$:

$$
\langle \tilde{\boldsymbol{G}}(p+1, v) \rangle_{\tilde{\boldsymbol{\Phi}}} = \langle \tilde{\boldsymbol{G}}(p, v) \rangle_{\boldsymbol{\Phi}} - \left\langle \frac{\tilde{\boldsymbol{G}}(p, v) \phi_{p+1} \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v)}{\lambda + \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v) \phi_{p+1}} \right\rangle_{\tilde{\boldsymbol{\Phi}}}
$$

PDE Formulation (III)

PDE Formulation (III)

- First **approximation**, take expectation of quotient separately:

PDE Formulation (III)

- First **approximation**, take expectation of quotient separately:

$$\langle \tilde{\boldsymbol{G}}(p+1, v) \rangle_{\tilde{\boldsymbol{\Phi}}} \approx \langle \tilde{\boldsymbol{G}}(p, v) \rangle_{\boldsymbol{\Phi}} - \frac{\left\langle \tilde{\boldsymbol{G}}(p, v) \phi_{p+1} \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\tilde{\boldsymbol{\Phi}}}}{\left\langle \lambda + \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v) \phi_{p+1} \right\rangle_{\tilde{\boldsymbol{\Phi}}}}$$

PDE Formulation (III)

- First **approximation**, take expectation of quotient separately:

$$
\begin{aligned}
\left\langle \tilde{\boldsymbol{G}}(p+1, v) \right\rangle_{\tilde{\boldsymbol{\Phi}}} &\approx \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}} - \frac{\left\langle \tilde{\boldsymbol{G}}(p, v) \phi_{p+1} \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\tilde{\boldsymbol{\Phi}}}}{\left\langle \lambda + \phi_{p+1}^T \tilde{\boldsymbol{G}}(p, v) \phi_{p+1} \right\rangle_{\tilde{\boldsymbol{\Phi}}}} \\
&= \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}} - \frac{\left\langle \tilde{\boldsymbol{G}}(p, v) \left\langle \phi_{p+1} \phi_{p+1}^T \right\rangle_{\phi_{p+1}} \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\left\langle \phi_{p+1} \phi_{p+1}^T \right\rangle_{\phi_{p+1}} \tilde{\boldsymbol{G}}(p, v) \right) \right\rangle_{\boldsymbol{\Phi}}}
\end{aligned}
$$

PDE Formulation (III)

- First **approximation**, take expectation of quotient separately:

$$
\begin{aligned}
\langle \tilde{\boldsymbol{G}}(p+1,v)\rangle_{\tilde{\boldsymbol{\Phi}}} &\approx \langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} - \frac{\left\langle \tilde{\boldsymbol{G}}(p,v)\phi_{p+1}\phi_{p+1}^T\tilde{\boldsymbol{G}}(p,v)\right\rangle_{\tilde{\boldsymbol{\Phi}}}}{\left\langle \lambda + \phi_{p+1}^T\tilde{\boldsymbol{G}}(p,v)\phi_{p+1}\right\rangle_{\tilde{\boldsymbol{\Phi}}}} \\
&= \langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} - \frac{\left\langle \tilde{\boldsymbol{G}}(p,v)\left\langle \phi_{p+1}\phi_{p+1}^T\right\rangle_{\phi_{p+1}}\tilde{\boldsymbol{G}}(p,v)\right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\left\langle \phi_{p+1}\phi_{p+1}^T\right\rangle_{\phi_{p+1}}\tilde{\boldsymbol{G}}(p,v)\right)\right\rangle_{\boldsymbol{\Phi}}} \\
&= \langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} - \frac{\left\langle \tilde{\boldsymbol{G}}(p,v)^2\right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right)\right\rangle_{\boldsymbol{\Phi}}}
\end{aligned}
$$

$\left\langle \phi_{p+1}\phi_{p+1}^T\right\rangle_{\phi_{p+1}} = \mathbb{1}$

PDE Formulation (IV)

PDE Formulation (IV)

Equivalently:

PDE Formulation (IV)

Equivalently:

$$\langle \tilde{\boldsymbol{G}}(p+1, v) \rangle_{\tilde{\boldsymbol{\Phi}}} - \langle \tilde{\boldsymbol{G}}(p, v) \rangle_{\boldsymbol{\Phi}} \approx - \frac{\left\langle \tilde{\boldsymbol{G}}(p, v)^2 \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p, v) \right) \right\rangle_{\boldsymbol{\Phi}}}$$

PDE Formulation (IV)

Equivalently:

$$\langle \tilde{\boldsymbol{G}}(p+1,v)\rangle_{\tilde{\boldsymbol{\Phi}}} - \langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\left\langle \tilde{\boldsymbol{G}}(p,v)^2 \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

$$\overset{"1\rightarrow 0"}{\Longleftrightarrow} \quad \frac{\partial}{\partial p}\langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v)\right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

PDE Formulation (IV)

Equivalently:

$$\langle \tilde{\boldsymbol{G}}(p+1, v)\rangle_{\tilde{\boldsymbol{\Phi}}} - \langle \tilde{\boldsymbol{G}}(p, v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\left\langle \tilde{\boldsymbol{G}}(p, v)^2 \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p, v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

$$\overset{\text{"}1\longrightarrow 0\text{"}}{\Longleftrightarrow} \frac{\partial}{\partial p} \langle \tilde{\boldsymbol{G}}(p, v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p, v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

where we used that $\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}} = \left\langle \tilde{\boldsymbol{G}}(p, v)^2 \right\rangle_{\boldsymbol{\Phi}}$

PDE Formulation (IV)

Equivalently:

$$\langle \tilde{\boldsymbol{G}}(p+1,v)\rangle_{\tilde{\boldsymbol{\Phi}}} - \langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\left\langle \tilde{\boldsymbol{G}}(p,v)^2 \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

$$\overset{"1\to 0"}{\Longleftrightarrow} \; \frac{\partial}{\partial p}\langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v) \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

where we used that $\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v) \right\rangle_{\boldsymbol{\Phi}} = \left\langle \tilde{\boldsymbol{G}}(p,v)^2 \right\rangle_{\boldsymbol{\Phi}}$

We hence reduced the problem to the PDE

PDE Formulation (IV)

Equivalently:

$$\langle \tilde{\boldsymbol{G}}(p+1,v)\rangle_{\tilde{\boldsymbol{\Phi}}} - \langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\left\langle \tilde{\boldsymbol{G}}(p,v)^2 \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

$$\overset{"1\longrightarrow 0"}{\Longleftrightarrow} \frac{\partial}{\partial p}\langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} \approx -\frac{\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v) \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

where we used that $\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v) \right\rangle_{\boldsymbol{\Phi}} = \left\langle \tilde{\boldsymbol{G}}(p,v)^2 \right\rangle_{\boldsymbol{\Phi}}$

We hence reduced the problem to the PDE

$$\frac{\partial}{\partial p}\langle \tilde{\boldsymbol{G}}(p,v)\rangle_{\boldsymbol{\Phi}} = -\frac{\frac{\partial}{\partial v}\left\langle \tilde{\boldsymbol{G}}(p,v) \right\rangle_{\boldsymbol{\Phi}}}{\lambda + \left\langle \operatorname{tr}\left(\tilde{\boldsymbol{G}}(p,v)\right) \right\rangle_{\boldsymbol{\Phi}}}$$

PDE Formulation (V)

PDE Formulation (V)

- **Observe:** This is a matrix PDE

PDE Formulation (V)

- **Observe:** This is a matrix PDE

no samples $\Rightarrow \Phi\Phi^T = 0$

- We have the initial condition $\tilde{\boldsymbol{G}}(0, v) = (\boldsymbol{\Lambda}^{-1} + v\mathbb{1})^{-1}$

PDE Formulation (V)

- **Observe:** This is a matrix PDE
- We have the initial condition $\tilde{\boldsymbol{G}}(0, v) = (\boldsymbol{\Lambda}^{-1} + v\mathbb{1})^{-1}$
- $\tilde{\boldsymbol{G}}(0, v)$ is a diagonal matrix for any v $\implies \tilde{\boldsymbol{G}}(0, v)_{ij} = 0$

PDE Formulation (V)

- **Observe:** This is a matrix PDE
- We have the initial condition $\tilde{\boldsymbol{G}}(0, v) = (\boldsymbol{\Lambda}^{-1} + v\mathbb{1})^{-1}$
- $\tilde{\boldsymbol{G}}(0, v)$ is a diagonal matrix for any v $\implies \tilde{\boldsymbol{G}}(0, v)_{ij} = 0$

  $\implies \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}}$ remains diagonal

PDE Formulation (V)

- **Observe:** This is a matrix PDE
- We have the initial condition $\tilde{\boldsymbol{G}}(0, v) = (\boldsymbol{\Lambda}^{-1} + v\mathbb{1})^{-1}$
- $\tilde{\boldsymbol{G}}(0, v)$ is a diagonal matrix for any v $\implies \tilde{\boldsymbol{G}}(0, v)_{ij} = 0$

  $\implies \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}}$ remains diagonal
- Hence introduce $g_i(p, v) = \left\langle \tilde{\boldsymbol{G}}(p, v_{ii}) \right\rangle_{\boldsymbol{\Phi}}$ and $t(p, v) = \sum_{i=1}^{M} g_i(t, v)$

PDE Formulation (V)

- **Observe:** This is a matrix PDE
- We have the initial condition $\tilde{\boldsymbol{G}}(0, v) = (\boldsymbol{\Lambda}^{-1} + v\mathbb{1})^{-1}$
- $\tilde{\boldsymbol{G}}(0, v)$ is a diagonal matrix for any v $\implies \tilde{\boldsymbol{G}}(0, v)_{ij} = 0$

  $\implies \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}}$ remains diagonal
- Hence introduce $g_i(p, v) = \left\langle \tilde{\boldsymbol{G}}(p, v_{ii}) \right\rangle_{\boldsymbol{\Phi}}$ and $t(p, v) = \sum_{i=1}^{M} g_i(t, v)$

By summing up all the individual PDEs we get the PDE

PDE Formulation (V)

- **Observe:** This is a matrix PDE
- We have the initial condition $\tilde{\boldsymbol{G}}(0, v) = (\boldsymbol{\Lambda}^{-1} + v\mathbb{1})^{-1}$
- $\tilde{\boldsymbol{G}}(0, v)$ is a diagonal matrix for any v $\implies \tilde{\boldsymbol{G}}(0, v)_{ij} = 0$

  $\implies \left\langle \tilde{\boldsymbol{G}}(p, v) \right\rangle_{\boldsymbol{\Phi}}$ remains diagonal
- Hence introduce $g_i(p, v) = \left\langle \tilde{\boldsymbol{G}}(p, v_{ii}) \right\rangle_{\boldsymbol{\Phi}}$ and $t(p, v) = \sum_{i=1}^{M} g_i(t, v)$

By summing up all the individual PDEs we get the PDE

$$\frac{\partial t(p, v)}{\partial p} = \frac{1}{\lambda + t(p, v)} \frac{\partial t(p, v)}{\partial v}$$

Method of Characteristics (I)

Method of Characteristics (I)

- **Assume** we have the general PDE

$$a(p,v)\frac{\partial t}{\partial p} + b(p,v)\frac{\partial t}{\partial v} = c(p,v)$$

Method of Characteristics (I)

- **Assume** we have the general PDE

$$a(p,v)\frac{\partial t}{\partial p} + b(p,v)\frac{\partial t}{\partial v} = c(p,v)$$

- The solution forms a **surface**

$$S = \{(t(p,v),p,v) : (p,v) \in \mathbb{R}^2\}$$

Method of Characteristics (I)

- **Assume** we have the general PDE

$$a(p,v)\frac{\partial t}{\partial p} + b(p,v)\frac{\partial t}{\partial v} = c(p,v)$$

- The solution forms a **surface**

$$S = \{(t(p,v),p,v) : (p,v) \in \mathbb{R}^2\}$$

- We have the **tangent** vectors
  1) $\frac{\partial}{\partial p}(t,p,v) = (\frac{\partial t}{\partial p},1,0)$
  2) $\frac{\partial}{\partial v}(t,p,v) = (\frac{\partial t}{\partial v},0,1)$

Method of Characteristics (I)

- **Assume** we have the general PDE

$$a(p,v)\frac{\partial t}{\partial p} + b(p,v)\frac{\partial t}{\partial v} = c(p,v)$$

- The solution forms a **surface**

$$S = \{(t(p,v), p, v) : (p,v) \in \mathbb{R}^2\}$$

- We have the **tangent** vectors
  1) $\frac{\partial}{\partial p}(t,p,v) = (\frac{\partial t}{\partial p}, 1, 0)$
  2) $\frac{\partial}{\partial v}(t,p,v) = (\frac{\partial t}{\partial v}, 0, 1)$

- We can form the **normal vector** $\boldsymbol{n}(p,v) = \left(-1, \frac{\partial t}{\partial p}, \frac{\partial t}{\partial v}\right)$

Method of Characteristics (II)

Method of Characteristics (II)

- **Observe:** $(c(p,v), a(p,v), b(p,v)) \bullet \boldsymbol{n}(p,v) = 0$

  $\Longrightarrow$ $(c(p,v), a(p,v), b(p,v))$ is in the tangent plane

Method of Characteristics (II)

- **Observe:** $(c(p,v), a(p,v), b(p,v)) \bullet \boldsymbol{n}(p,v) = 0$

  $\implies (c(p,v), a(p,v), b(p,v))$ is in the tangent plane

- **Idea:** Construct S such that $\forall (z,p,v) \in S : (c(p,v), a(p,v), b(p,v))$ is in the tangent plane to S

Method of Characteristics (II)

- **Observe:** $(c(p,v), a(p,v), b(p,v)) \bullet \boldsymbol{n}(p,v) = 0$ $\quad \forall (p,v)$

  $\implies (c(p,v), a(p,v), b(p,v))$ is in the tangent plane

- **Idea:** Construct S such that $\forall (z,p,v) \in S : (c(p,v), a(p,v), b(p,v))$ is in the tangent plane to S

- **Assume:** We have initial data $t(p, v_0) = f(p, v_0)$

Method of Characteristics (III)

Method of Characteristics (III)

- If we want to construct a curve $\Gamma(s) = (z(s), p(s), v(s))$ in S, we can start from the initial data and make sure that its tangent vector agrees with $(c(p,v), a(p,v), b(p,v))$:

  1) $(z(0), p(0), v(0)) = (f(p_0, v_0), p_0, v_0)$

  2) $\left(\frac{\partial z}{\partial s}, \frac{\partial p}{\partial s}, \frac{\partial z}{\partial v}\right) = (c(p,v), a(p,v), b(p,v))$

Method of Characteristics (III)

- If we want to construct a curve $\Gamma(s) = (z(s), p(s), v(s))$ in S, we can start from the initial data and make sure that its tangent vector agrees with $(c(p,v), a(p,v), b(p,v))$:

  1) $(z(0), p(0), v(0)) = (f(p_0, v_0), p_0, v_0)$

  2) $\left(\frac{\partial z}{\partial s}, \frac{\partial p}{\partial s}, \frac{\partial z}{\partial v}\right) = (c(p,v), a(p,v), b(p,v))$

- This is an ODE, so called **characteristic ODE**

Method of Characteristics (III)

- If we want to construct a curve $\Gamma(s) = (z(s), p(s), v(s))$ in S, we can start from the initial data and make sure that its tangent vector agrees with $(c(p, v), a(p, v), b(p, v))$:

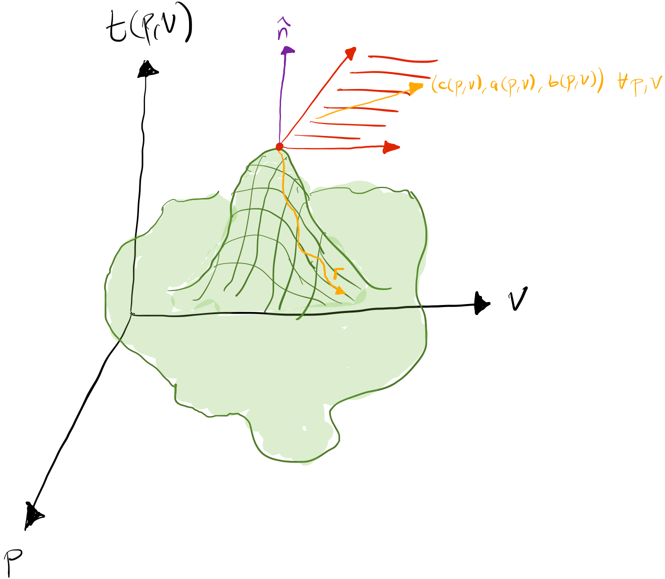
  1) $(z(0), p(0), v(0)) = (f(p_0, v_0), p_0, v_0)$

  2) $\left(\frac{\partial z}{\partial s}, \frac{\partial p}{\partial s}, \frac{\partial z}{\partial v}\right) = (c(p, v), a(p, v), b(p, v))$

- This is an ODE, so called **characteristic ODE**

- Its solution is called **characteristic curve**, taking the unions of all such curves constructs entire S

Visualization



Method of Characteristics: Applied Here (I)

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda + t} \frac{\partial t}{\partial v}$ and $t(0, v) = \mathrm{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda + t} \frac{\partial t}{\partial v}$ and $t(0, v) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$
- The parameter vector is given by $\left(0, 1, -\frac{1}{\lambda + t}\right)$

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda+t}\frac{\partial t}{\partial v}$ and $t(0,v) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$
- The parameter vector is given by $\left(0, 1, -\frac{1}{\lambda+t}\right)$
- **Initial data** $(f(v), 0, v)$

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda+t}\frac{\partial t}{\partial v}$ and $t(0,v) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$
- The parameter vector is given by $\left(0, 1, -\frac{1}{\lambda+t}\right)$
- **Initial data** $(f(v), 0, v)$
- Let's take a characteristic curve $(z(s), p(s), v(s))$:

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda+t}\frac{\partial t}{\partial v}$ and $t(0,v) = \mathrm{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$
- The parameter vector is given by $\left(0, 1, -\frac{1}{\lambda+t}\right)$
- **Initial data** $(f(v), 0, v)$
- Let's take a characteristic curve $(z(s), p(s), v(s))$:

  1) $(z(0), p(0), v(0)) = (f(v_0), 0, v_0)$

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda + t}\frac{\partial t}{\partial v}$ and $t(0, v) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$

- The parameter vector is given by $\left(0, 1, -\frac{1}{\lambda+t}\right)$

- **Initial data** $(f(v), 0, v)$

- Let's take a characteristic curve $(z(s), p(s), v(s))$:

  1) $(z(0), p(0), v(0)) = (f(v_0), 0, v_0)$
  2) $\left(\frac{\partial z}{\partial s}, \frac{\partial p}{\partial s}, \frac{\partial z}{\partial v}\right) = \left(0, 1, -\frac{1}{\lambda+t}\right)$

Method of Characteristics: Applied Here (I)

- $\frac{\partial t}{\partial p} = \frac{1}{\lambda+t}\frac{\partial t}{\partial v}$ and $t(0,v) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} + v\mathbb{1}\right)^{-1} = f(v)$
- The parameter vector is given by $\left(0, 1, -\frac{1}{\lambda+t}\right)$
- **Initial data** $(f(v), 0, v)$
- Let's take a characteristic curve $(z(s), p(s), v(s))$:

  1) $(z(0), p(0), v(0)) = (f(v_0), 0, v_0)$
  2) $\left(\frac{\partial z}{\partial s}, \frac{\partial p}{\partial s}, \frac{\partial z}{\partial v}\right) = \left(0, 1, -\frac{1}{\lambda+t}\right)$

- $$\begin{aligned}(z(s), p(s), v(s)) &= \left(c_1(v_0), s + c_2(v_0), -\frac{s}{\lambda+t} + c_3(v_0)\right)\\ &= (\mathrm{f}(\mathrm{v}_0), s, -\frac{s}{\lambda+t} + v_0)\end{aligned}$$

Method of Characteristics: Applied Here (II)

Method of Characteristics: Applied Here (II)

- $(z(s), p(s), v(s)) = (f(v_0), s, -\frac{s}{\lambda+t} + v_0)$

Method of Characteristics: Applied Here (II)

- $(z(s), p(s), v(s)) = (f(v_0), s, -\frac{s}{\lambda+t} + v_0)$
- Inverting the equations gives:
  1) $v_0 = v + \frac{s}{\lambda+t}$
  2) $s = p$
  3) $z = f(v + \frac{p}{\lambda+t})$

Method of Characteristics: Applied Here (II)

- $(z(s), p(s), v(s)) = (f(v_0), s, -\frac{s}{\lambda+t} + v_0)$
- Inverting the equations gives:
  1) $v_0 = v + \frac{s}{\lambda+t}$
  2) $s = p$
  3) $z = f(v + \frac{p}{\lambda+t})$
- We hence get the solution finally:

Method of Characteristics: Applied Here (II)

- $(z(s), p(s), v(s)) = (f(v_0), s, -\frac{s}{\lambda+t} + v_0)$
- Inverting the equations gives:
  1) $v_0 = v + \frac{s}{\lambda+t}$
  2) $s = p$
  3) $z = f(v + \frac{p}{\lambda+t})$
- We hence get the solution finally:

$$t(p,v) = f\left(v + \frac{p}{\lambda + t}\right) = \mathrm{tr}\left(\left(\mathbf{\Lambda}^{-1} + (v + \frac{p}{\lambda + t(p,v)})\mathbb{1}\right)^{-1}\right)$$

$\implies$ Still an implicit equation

Generalization Error Formula

Generalization Error Formula

Plugging in the PDE approximation to obtain $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$, after some algebraic manipulations, leads to:

Generalization Error Formula

Plugging in the PDE approximation to obtain $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$, after some algebraic manipulations, leads to:

Theorem 3:
The average generalization error can be approximated like

$$E_i(p) = \frac{\langle \bar{w}_i^2 \rangle}{\lambda_i} \left(\frac{1}{\lambda_i} + \frac{p}{\lambda + t(p)} \right)^{-2} \left(1 - \frac{p\gamma(p)}{(\lambda + t(p))^2} \right)^{-1}$$

where $\gamma(p) = \sum_{i=1}^{M} \left(\frac{1}{\lambda_i} + \frac{p}{\lambda + t(p)} \right)^{-2}$

Algorithm

Algorithm

The implicit equation needs to be solved numerically. If one is able to do that, we have the following algorithm:

Algorithm

The implicit equation needs to be solved numerically. If one is able to do that, we have the following algorithm:

Algorithm 1 Computing Theoretical Learning Curves

Input: RKHS spectrum $\{\lambda_\rho\}$, target function weights $\{\overline{w}_\rho\}$, regularizer λ, sample sizes $\{p_i\}$, $i = 1, ..., m$;

for $i = 1$ **to** m **do**

Solve numerically $t_i = \sum_\rho \left(\frac{1}{\lambda_\rho} + \frac{p_i}{\lambda + t_i}\right)^{-1}$

Compute $\gamma_i = \sum_\rho \left(\frac{1}{\lambda_\rho} + \frac{p_i}{\lambda + t_i}\right)^{-2}$

$E_{\rho,i} = \frac{\langle \overline{w}_\rho^2 \rangle}{\lambda_\rho} \left(\frac{1}{\lambda_\rho} + \frac{p_i}{\lambda + t_i}\right)^{-2} \left(1 - \frac{p_i \gamma_i}{(\lambda + t_i)^2}\right)^{-1}$

end for

Algorithm

The implicit equation needs to be solved numerically. If one is able to do that, we have the following algorithm:

Algorithm 1 Computing Theoretical Learning Curves

Input: RKHS spectrum $\{\lambda_\rho\}$, target function weights $\{\overline{w}_\rho\}$, regularizer λ, sample sizes $\{p_i\}$, $i = 1, ..., m$;

for $i = 1$ **to** m **do**

Solve numerically $t_i = \sum_\rho \left(\frac{1}{\lambda_\rho} + \frac{p_i}{\lambda + t_i}\right)^{-1}$

Compute $\gamma_i = \sum_\rho \left(\frac{1}{\lambda_\rho} + \frac{p_i}{\lambda + t_i}\right)^{-2}$

$E_{\rho,i} = \frac{\langle \overline{w}_\rho^2 \rangle}{\lambda_\rho} \left(\frac{1}{\lambda_\rho} + \frac{p_i}{\lambda + t_i}\right)^{-2} \left(1 - \frac{p_i \gamma_i}{(\lambda + t_i)^2}\right)^{-1}$

end for

This allows one to plot learning curves with the sample size as the varying parameter

Statistical Mechanics

Statistical Mechanics

- Use variation of **Replica trick** to estimate $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$

Statistical Mechanics

- Use variation of **Replica trick** to estimate $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$
- Idea is to write the inverse of the partition function as

$$Z^{-1} = \lim_{n \longrightarrow 0} Z^{n-1}$$

Statistical Mechanics

- Use variation of **Replica trick** to estimate $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$
- Idea is to write the inverse of the partition function as

$$Z^{-1} = \lim_{n \longrightarrow 0} Z^{n-1}$$

- Then calculate the right-hand side for **integer** n and then analytically continue to found formula to take the limit

Statistical Mechanics

- Use variation of **Replica trick** to estimate $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$
- Idea is to write the inverse of the partition function as

$$Z^{-1} = \lim_{n \longrightarrow 0} Z^{n-1}$$

- Then calculate the right-hand side for **integer** n and then analytically continue to found formula to take the limit
- As usual the calculations are **very bizarre**, order parameter definitions are enforced via Dirac deltas, those get replaced by their Fourier integral representation and then the whole thing simplifies somehow...

Statistical Mechanics

- Use variation of **Replica trick** to estimate $\langle \boldsymbol{G}^2 \rangle_{\boldsymbol{X}}$
- Idea is to write the inverse of the partition function as

$$Z^{-1} = \lim_{n \longrightarrow 0} Z^{n-1}$$

- Then calculate the right-hand side for **integer** n and then analytically continue to found formula to take the limit
- As usual the calculations are **very bizarre**, order parameter definitions are enforced via Dirac deltas, those get replaced by their Fourier integral representation and then the whole thing simplifies somehow...
- Using a saddle point approximation they finally find the same solution as with the PDE approach

Experiments (I)

Experiments (I)

- Experiment mainly with NTK

Experiments (I)

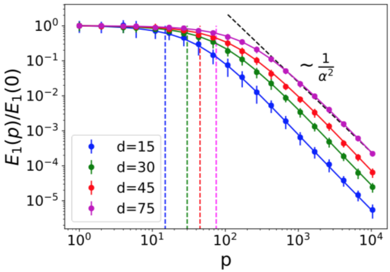
- Experiment mainly with NTK
- Assume target is given by $f^*(\boldsymbol{x}) = \sum_{i=1}^{p'} \bar{\alpha}_i \boldsymbol{\Theta}(\boldsymbol{x}, \bar{\boldsymbol{x}}_i)$, $\bar{\alpha}_i \sim$ Bernoulli($\frac{1}{2}$) and $\boldsymbol{x} \sim \mathbb{S}^{d-1}$

Experiments (I)

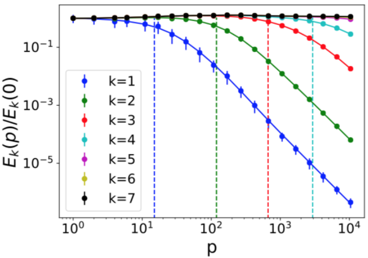
- Experiment mainly with NTK
- Assume target is given by $f^*(\boldsymbol{x}) = \sum_{i=1}^{p'} \bar{\alpha}_i \boldsymbol{\Theta}(\boldsymbol{x}, \bar{\boldsymbol{x}}_i)$, $\bar{\alpha}_i \sim \text{Bernoulli}(\frac{1}{2})$ and $\boldsymbol{x} \sim \mathbb{S}^{d-1}$
- This form allows for perfect calculation of E_i leveraging the spherical harmonics $Y_{k,m}$

Experiments (I)

- Experiment mainly with NTK
- Assume target is given by $f^*(\boldsymbol{x}) = \sum_{i=1}^{p'} \bar{\alpha}_i \boldsymbol{\Theta}(\boldsymbol{x}, \bar{\boldsymbol{x}}_i)$, $\bar{\alpha}_i \sim \text{Bernoulli}(\frac{1}{2})$ and $\boldsymbol{x} \sim \mathbb{S}^{d-1}$
- This form allows for perfect calculation of E_i leveraging the spherical harmonics $Y_{k,m}$



(b) 3-layer NTK $k = 1$ $\lambda = 1$



(a) 3-layer NTK $d = 15$ $\lambda = 0$

Experiments (II)

Experiments (II)

- Compare NTK learning curves with finite-width NNs

Experiments (II)

- Compare NTK learning curves with finite-width NNs
- $f^*(\boldsymbol{x}) = \sum_{i=1}^{p'} \bar{\alpha}_i Q_k(\boldsymbol{x}^T \bar{\boldsymbol{x}}_i)$ where

$$Q_k(\boldsymbol{x}^T \boldsymbol{x}') = \frac{1}{N(d,k)} \sum_{m=1}^{N(d,k)} Y_{km}(\boldsymbol{x}) Y_{km}(\boldsymbol{x}')$$

Experiments (II)

- Compare NTK learning curves with finite-width NNs
- $f^*(\boldsymbol{x}) = \sum_{i=1}^{p'} \bar{\alpha}_i Q_k(\boldsymbol{x}^T \bar{\boldsymbol{x}}_i)$ where

$$Q_k(\boldsymbol{x}^T \boldsymbol{x}') = \frac{1}{N(d,k)} \sum_{m=1}^{N(d,k)} Y_{km}(\boldsymbol{x}) Y_{km}(\boldsymbol{x}')$$

- Recall that the Mercer decomposition for NTK for $\boldsymbol{x}_i \sim \mathcal{U}(\mathbb{S}^{d-1})$ is

$$\boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=0}^{\infty} \lambda_k \sum_{m=1}^{N(d,k)} Y_{km}(\boldsymbol{x}) Y_{km}(\boldsymbol{x}')$$

Experiments (II)

- Compare NTK learning curves with finite-width NNs
- $f^*(\boldsymbol{x}) = \sum_{i=1}^{p'} \bar{\alpha}_i Q_k(\boldsymbol{x}^T \bar{\boldsymbol{x}}_i)$ where

$$Q_k(\boldsymbol{x}^T \boldsymbol{x}') = \frac{1}{N(d,k)} \sum_{m=1}^{N(d,k)} Y_{km}(\boldsymbol{x}) Y_{km}(\boldsymbol{x}')$$

- Recall that the Mercer decomposition for NTK for $\boldsymbol{x}_i \sim \mathcal{U}(\mathbb{S}^{d-1})$ is

$$\boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=0}^{\infty} \lambda_k \sum_{m=1}^{N(d,k)} Y_{km}(\boldsymbol{x}) Y_{km}(\boldsymbol{x}')$$

- f^* is hence composed of harmonics belonging to the same eigenvalue

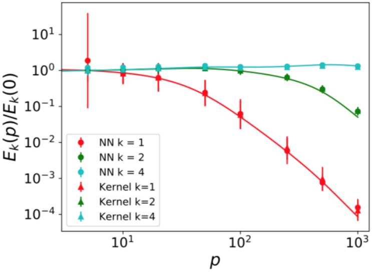
Experiments (III)

Experiments (III)

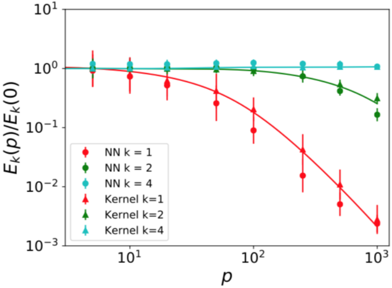
- Train 2 and 4-layer NNs with widths 500, $d = 30$.

Experiments (III)

- Train 2 and 4-layer NNs with widths 500, $d = 30$.



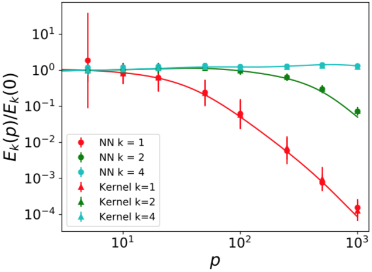
(a) 2-layer NN $N = 10000$



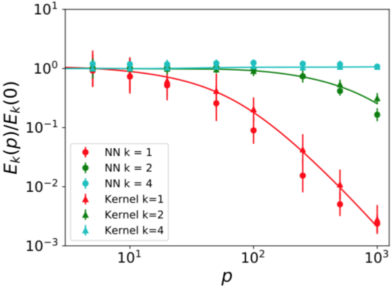
(b) 4-layer NN $N = 500$

Experiments (III)

- Train 2 and 4-layer NNs with widths 500, $d = 30$.



(a) 2-layer NN $N = 10000$



(b) 4-layer NN $N = 500$

- Good agreement for two layer network, but seems to get worse with depth. No deeper depth experiments included.

Experiments (IV)

Experiments (IV)

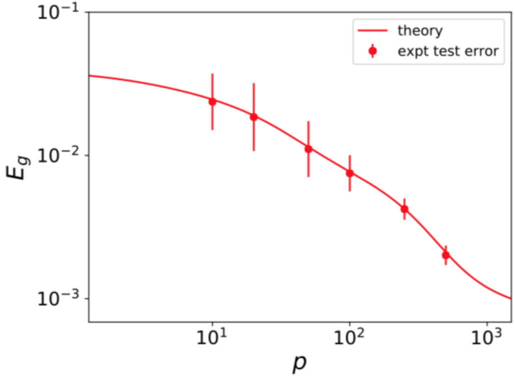
- Consider more complicated teacher functions:

$$f^*(\boldsymbol{x}) = \bar{\boldsymbol{r}}^T \sigma(\bar{\boldsymbol{W}} \boldsymbol{x}) \text{ and } f(\boldsymbol{x}) = \boldsymbol{r}^T \sigma(\boldsymbol{W} \boldsymbol{x})$$

Experiments (IV)

- Consider more complicated teacher functions:

$$f^*(\boldsymbol{x}) = \bar{\boldsymbol{r}}^T \sigma(\bar{\boldsymbol{W}}\boldsymbol{x}) \text{ and } f(\boldsymbol{x}) = \boldsymbol{r}^T \sigma(\boldsymbol{W}\boldsymbol{x})$$



(c) 2-Layer NN Student-Teacher; $N = 8000$

Discussion

Discussion

- I really enjoyed the **mathematics/tools** in this paper

Discussion

- I really enjoyed the **mathematics/tools** in this paper
- Learning curves very general as they hold for **any kernel**

Discussion

- I really enjoyed the **mathematics/tools** in this paper
- Learning curves very general as they hold for **any kernel**
- Also a bit disappointing, would have been nicer to really **specialize** the theory to NTK

Discussion

- I really enjoyed the **mathematics/tools** in this paper
- Learning curves very general as they hold for **any kernel**
- Also a bit disappointing, would have been nicer to really **specialize** the theory to NTK
- Would also have been nicer to have an **analytical** expression (instead of these implicit equations) but probably hard to do

Discussion

- I really enjoyed the **mathematics/tools** in this paper
- Learning curves very general as they hold for **any kernel**
- Also a bit disappointing, would have been nicer to really **specialize** the theory to NTK
- Would also have been nicer to have an **analytical** expression (instead of these implicit equations) but probably hard to do
- Nice that the same solution pops out under two different approximations. But maybe they are **more similar** than one can see at first glance. In both cases, two integer parameters are **made continuous** and limits are taken