Tutorial:

BO, Double Descent and RKHS ridge regression made easy!

Theodor Misiakiewicz (Stanford)

May 26th, 2022

New Interactions Between Statistics and Optimization workshop, BIRS.

Modern machine learning

- **New regime for statistics:** overparametrized models, no explicit regularization and capacity control, train until (near-)interpolation even with noisy data.

  New phenomenology: benign overfitting, double descent, non-monotonic error curves...

- Phenomena already present in **linear models** [Belkin,Ma,Mandal,'18].

- This talk: **kernel ridge regression** (KRR) in the **high dimension regime**.

Goal of this tutorial is to show:

1. How to derive quickly asymptotics for kernel/random features ridge regression.
2. How the above phenomena have very precise explanations in this regime.

Modern machine learning

- **New regime for statistics:** overparametrized models, no explicit regularization and capacity control, train until (near-)interpolation even with noisy data.

  New phenomenology: benign overfitting, double descent, non-monotonic error curves...

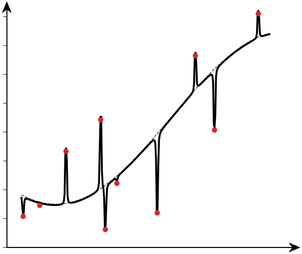
- Phenomena already present in **linear models** [Belkin,Ma,Mandal,'18].

- This talk: **kernel ridge regression** (KRR) in the **high dimension regime**.

Goal of this tutorial is to show:

1. How to derive quickly asymptotics for kernel/random features ridge regression.
2. How the above phenomena have very precise explanations in this regime.

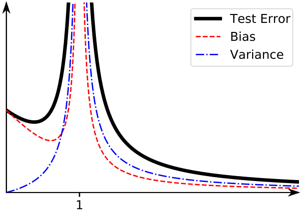
Benign overfitting, self-induced regularization and double descent



- **Benign overfitting:** interpolator generalizes well. Idea: $\hat{f} = f_0 + \Delta$ with f_0 smooth solution + spike part with $\|\Delta\|_{L^2} \ll 1$.

  In linear models: **self-induced regularization**. Non-smooth part of the kernel plays the role of an effective ridge regularization.

  (This is HD phenomena.)



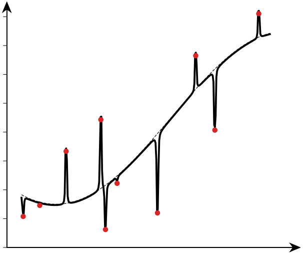
- **Double descent** and non-monotonic curves.

Need exact test error that holds for a given function and is exact up to an additive vanishing constant.

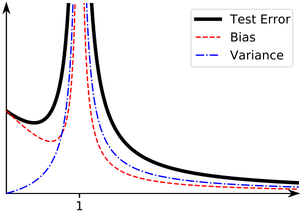
- **Benign overfitting:** interpolator generalizes well. Idea: $\hat{f} = f_0 + \Delta$ with f_0 smooth solution + spike part with $\|\Delta\|_{L^2} \ll 1$.

  In linear models: **self-induced regularization**. Non-smooth part of the kernel plays the role of an effective ridge regularization.

  (This is HD phenomena.)



- **Double descent** and non-monotonic curves.



Need exact test error that holds for a given function and is exact up to an additive vanishing constant.

A subset of references:

- Benign overfitting: [Liang,Rakhlin,'18], [Ghorbani,Mei,M,Montanari,'19], [Bartlett,Long,Lugosi,Tsigler,'20].
- Double descent: [Mei,Montanari,'19], [Hastie,Montanari,Rosset,Tibshirani,'20], [Gerace,Loureiro,Krzakala,Mezard,Zdeborova,'20].
- Linear models: [Tsigler,Bartlett,'20], [Cui,Loureiro,Krzakala,Zdeborova,'21], [Liao,Couillet,Mahoney,'20], [Richards,Mourtada,Rosasco,'21], [Wu,Xu,'20].
- KRR: [Jacot,Simsek,Spadaro,Hongler,Gabriel,'20], [Canatar,Bordelon,Pehlevan,'21], [Mei,M,Montanari,'21], [Bartlett,Montanari,Rakhlin,'21], [Liu,Liao,Suykens,'21], [Liang,Rakhlin,Zhai,'21], [Hu,Lu,'22].

Quick background on KRR (1)

Covariates: $\boldsymbol{x} \in (\mathcal{X}, \nu)$.
Kernel function: $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ PSD, associated kernel operator: $\mathbb{K} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$:

$$\mathbb{K} f(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \nu(\mathrm{d}\boldsymbol{x}') \, .$$

Diagonalization: $\{\phi_j\}_{j \geq 1}$ orthonormal basis of $L^2(\mathcal{X})$ and $\{\lambda_j\}_{j \geq 0}$ nonincreasing ($\lambda_j > 0$)

$$\mathbb{K} = \sum_{j \geq 1} \lambda_j \phi_j \phi_j^* \, , \qquad K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j \geq 1} \lambda_j \phi_j(\boldsymbol{x}_1) \phi_j(\boldsymbol{x}_2) \, .$$

Feature map: $\boldsymbol{x} \mapsto \Phi(\boldsymbol{x}) = (\sqrt{\lambda_j} \phi_j(\boldsymbol{x}))_{j \geq 1}$ so that $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2) \rangle_{\ell_2}$

L^2 space: for any $f_* \in L^2(\mathcal{X})$,

$$f_*(\boldsymbol{x}) = \sum_{j \geq 1} \beta_j \phi_j(\boldsymbol{x}) = \langle \boldsymbol{\theta}_*, \Phi(\boldsymbol{x}) \rangle_{\ell_2} \, , \qquad \boldsymbol{\theta}_* = (\theta_j)_{j \geq 1} \, , \qquad \theta_j = \beta_j / \sqrt{\lambda_j} \, .$$

Associated RKHS: $\mathcal{H} = \{ f \in L^2(\mathcal{X}) : \| f \|_{\mathcal{H}} < \infty \}$,

$$\| f \|_{\mathcal{H}}^2 = \| \mathbb{K}^{-1/2} f \|_{L^2}^2 = \sum_{j \geq 1} \frac{\beta_j^2}{\lambda_j} = \| \boldsymbol{\theta}_* \|_{\ell_2}^2 \, .$$

Quick background on KRR (2)

Data: $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$ where $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \nu)$ and $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$, with $f_* \in L^2(\mathcal{X})$ and independent noise ε_i, $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2$.

Kernel ridge regression: fit the data with

$$\hat{f}_\lambda = \arg\min_f \left\{ \sum_{i\in[n]} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

Equivalently: $\hat{f}_\lambda = \langle \hat{\boldsymbol{\theta}}_\lambda, \Phi(\cdot)\rangle_{\ell_2}$ where for $\boldsymbol{\Phi} = [\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_n)]^\mathsf{T} \in \mathbb{R}^{n\times\infty}$,

$$\hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i\in[n]} (y_i - \langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}\rangle)^2 + \lambda \|\boldsymbol{\theta}\|_{\ell_2}^2 \right\} = \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{\Phi}\boldsymbol{\Phi}^\mathsf{T} + \lambda)^{-1}\boldsymbol{y}\,.$$

[**Representer thm**: $\hat{f}_\lambda(\boldsymbol{x}) = \sum_i \hat{a}_i K(\boldsymbol{x}, \boldsymbol{x}_i)$ with $\hat{\boldsymbol{a}} = (\boldsymbol{K} + \lambda)^{-1}\boldsymbol{y}$, $\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij\in[n]}$.]

Goal: compute the test error $R(f_*, \hat{f}_\lambda) = \mathbb{E}_x[(f_*(x) - \hat{f}_\lambda(x))^2]$ in the high dimensional regime $\boldsymbol{x} \in \mathbb{R}^d$ and $\log(n) \asymp \log(d)$.

Quick background on KRR (2)

Data: $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$ where $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \nu)$ and $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$, with $f_* \in L^2(\mathcal{X})$ and independent noise ε_i, $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2$.

Kernel ridge regression: fit the data with

$$\hat{f}_\lambda = \arg\min_f \Big\{ \sum_{i\in[n]} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\} .$$

Equivalently: $\hat{f}_\lambda = \langle \hat{\boldsymbol{\theta}}_\lambda, \Phi(\cdot)\rangle_{\ell_2}$ where for $\boldsymbol{\Phi} = [\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_n)]^{\mathsf{T}} \in \mathbb{R}^{n\times\infty}$,

$$\hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta}} \Big\{ \sum_{i\in[n]} (y_i - \langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}\rangle)^2 + \lambda \|\boldsymbol{\theta}\|_{\ell_2}^2 \Big\} = \boldsymbol{\Phi}^{\mathsf{T}} (\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}} + \lambda)^{-1} \boldsymbol{y} .$$

[**Representer thm**: $\hat{f}_\lambda(\boldsymbol{x}) = \sum_i \hat{a}_i K(\boldsymbol{x}, \boldsymbol{x}_i)$ with $\hat{\boldsymbol{a}} = (\boldsymbol{K} + \lambda)^{-1}\boldsymbol{y}$, $\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij\in[n]}$.]

Goal: compute the test error $R(f_*, \hat{f}_\lambda) = \mathbb{E}_{\boldsymbol{x}}[(f_*(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x}))^2]$ in the high dimensional regime $\boldsymbol{x} \in \mathbb{R}^d$ and $\log(n) \asymp \log(d)$.

Gaussian equivalent model and universality of feature maps (1)

Ridge regression with features $\phi(\boldsymbol{x}_i)$: function of random matrix resolvent.

For some "high dimensional" feature map, expect universality to happen: can replace $\phi(\boldsymbol{x})$ by Gaussian vector $\boldsymbol{z}$ with matching first two moments.

Covariance matrix: $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x}}[\Phi(\boldsymbol{x})\Phi(\boldsymbol{x})^{\mathsf{T}}] = \mathrm{diag}((\lambda_j)_{j\geq 1})$.

Model	KRR:	Gaussian covariates model:
Distribution	$\Phi(\boldsymbol{x}) = (\sqrt{\lambda_j}\phi_j(\boldsymbol{x}))_{j\geq 1}$ with $\boldsymbol{x} \sim \nu$.	$\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{\Sigma})$
Data	$(\Phi(\boldsymbol{x}_i))_{i\in[n]}$ iid, $y_i = \langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}_* \rangle + \varepsilon_i$	$(\boldsymbol{z}_i)_{i\in[n]}$ iid, $y_i = \langle \boldsymbol{z}_i, \boldsymbol{\theta}_* \rangle + \varepsilon_i$
Feature mat.	$\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)]^{\mathsf{T}} \in \mathbb{R}^{n\times\infty}$	$\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^{\mathsf{T}} \in \mathbb{R}^{n\times\infty}$
Kernel fct	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2) \rangle_{\ell_2}$	$K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle_{\ell_2}$
Solution	$\hat{f}_\lambda = \langle \Phi(\boldsymbol{x}), \hat{\boldsymbol{\theta}}_\lambda \rangle_{\ell_2}$ $\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}} + \lambda)^{-1}\boldsymbol{y}$	$\hat{f}_\lambda(\boldsymbol{z}) = \langle \boldsymbol{z}, \hat{\boldsymbol{\theta}}_\lambda^G \rangle_{\ell_2}$ $\hat{\boldsymbol{\theta}}_\lambda^G = \boldsymbol{Z}^{\mathsf{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}} + \lambda)^{-1}\boldsymbol{y}$
Test error	$R(f_*, \hat{f}_\lambda) = \lVert \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_* - \hat{\boldsymbol{\theta}}_\lambda) \rVert_{\ell_2}^2$	$R_G(f_*, \hat{f}_\lambda) = \lVert \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_* - \hat{\boldsymbol{\theta}}_\lambda^G) \rVert_{\ell_2}^2$

Universality: $R(f_*, \hat{f}_\lambda) - R_G(f_*, \hat{f}_\lambda) \xrightarrow{\mathbb{P}} 0$ (already conjectured previously).

Gaussian equivalent model and universality of feature maps (1)

Ridge regression with features $\phi(\boldsymbol{x}_i)$: function of random matrix resolvent.

For some "high dimensional" feature map, expect universality to happen: can replace $\phi(\boldsymbol{x})$ by Gaussian vector $\boldsymbol{z}$ with matching first two moments.

Covariance matrix: $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x}}[\Phi(\boldsymbol{x})\Phi(\boldsymbol{x})^{\mathsf{T}}] = \mathrm{diag}((\lambda_j)_{j\geq 1})$.

Model	**KRR:**	**Gaussian covariates model:**
Distribution	$\Phi(\boldsymbol{x}) = (\sqrt{\lambda_j}\phi_j(\boldsymbol{x}))_{j\geq 1}$ with $\boldsymbol{x} \sim \nu$.	$\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{\Sigma})$
Data	$(\Phi(\boldsymbol{x}_i))_{i\in[n]}$ iid, $y_i = \langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}_* \rangle + \varepsilon_i$	$(\boldsymbol{z}_i)_{i\in[n]}$ iid, $y_i = \langle \boldsymbol{z}_i, \boldsymbol{\theta}_* \rangle + \varepsilon_i$
Feature mat.	$\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \dots, \phi(\boldsymbol{x}_n)]^{\mathsf{T}} \in \mathbb{R}^{n\times\infty}$	$\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_n]^{\mathsf{T}} \in \mathbb{R}^{n\times\infty}$
Kernel fct	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2) \rangle_{\ell_2}$	$K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle_{\ell_2}$
Solution	$\hat{f}_\lambda = \langle \Phi(\boldsymbol{x}), \hat{\boldsymbol{\theta}}_\lambda \rangle_{\ell_2}$ $\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}} + \lambda)^{-1}\boldsymbol{y}$	$\hat{f}_\lambda(\boldsymbol{z}) = \langle \boldsymbol{z}, \hat{\boldsymbol{\theta}}_\lambda^G \rangle_{\ell_2}$ $\hat{\boldsymbol{\theta}}_\lambda^G = \boldsymbol{Z}^{\mathsf{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}} + \lambda)^{-1}\boldsymbol{y}$
Test error	$R(f_*, \hat{f}_\lambda) = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_* - \hat{\boldsymbol{\theta}}_\lambda)\|_{\ell_2}^2$	$R_G(f_*, \hat{f}_\lambda) = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_* - \hat{\boldsymbol{\theta}}_\lambda^G)\|_{\ell_2}^2$

Universality: $R(f_*, \hat{f}_\lambda) - R_G(f_*, \hat{f}_\lambda) \xrightarrow{\mathbb{P}} 0$ (already conjectured previously).

Gaussian equivalent model and universality of feature maps (2)

[Hu, Lu,'20], [Montanari, Saeed,'22],[Gerace,Loureiro,Krzakala,Mezard,Zdeborova,'20] when $n \asymp d$ and or universality of covariates.

Here universality of the entire feature map and polynomial scaling $\log(n) \asymp \log(d)$.

Such an equivalence is not obvious: coordinates of $\Phi(\boldsymbol{x})$ are not subgaussian or weakly dependent. Here will present some cases, where it can be shown rigorously.

Test error: $R_G(f_*, \hat{f}_\lambda) = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_* - \hat{\boldsymbol{\theta}}_\lambda^G)\|_{\ell_2}^2$.

$$\text{Bias} = \|\boldsymbol{\Sigma}^{1/2}\theta_* - \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}^\mathsf{T}(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \lambda)^{-1}\boldsymbol{Z}\theta_*\|_{\ell_2}^2 ,$$
$$\text{Variance} = \sigma_\varepsilon^2 \mathrm{Tr}\big[(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \lambda)^{-2}\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\mathsf{T}\big] .$$

Different than previous Gaussian design work ($n \asymp p$, eigenvalues of same order).

Here for simplicity, we assume: $\exists \delta > 0$ and a sequence $m(n)$ such that $m \leq n^{1-\delta}$ and

$$\lambda_{m+1} \cdot n^{1+\delta} \leq \sum_{j=m+1}^{\infty} \lambda_j .$$

(a **'spectral gap'**, which will happen for models with a lot of symmetries.)

Random kernel matrix: $\boldsymbol{K} = (\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle)_{ij\in[n]} \in \mathbb{R}^{n\times n}$.

$$\boldsymbol{K} = \sum_{j=1}^{\infty} \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^{\mathsf{T}} = \boldsymbol{K}_{\le m} + \boldsymbol{K}_{>m}\,, \qquad \boldsymbol{u}_j = (z_{ij})_{i\in[n]}\,.$$

Main intuition:

- **High-frequency part:** $\boldsymbol{K}_{>m} = \boldsymbol{Z}_{>m}\boldsymbol{Z}_{>m}^{\mathsf{T}}$ with $\boldsymbol{Z}_{>m}$ iid Gaussian rows. Denote $\lambda_{>m} = \sum_{j>m} \lambda_j$. Then w.h.p.,

$$\|\boldsymbol{Z}_{>m}\boldsymbol{Z}_{>m}^{\mathsf{T}} - \lambda_{>m}\boldsymbol{I}\|_{\mathrm{op}} \lesssim \lambda_{m+1} + \lambda_{>m}/n \lesssim n^{-1} \cdot \lambda_{>m}\,.$$

- **Low-frequency part:** $\boldsymbol{K}_{\le m} = \boldsymbol{Z}_{\le m}\boldsymbol{Z}_{\le m}^{\mathsf{T}} = \boldsymbol{G}_m \boldsymbol{\Sigma}_m \boldsymbol{G}_m^{\mathsf{T}}$ where $\boldsymbol{G}_m = \boldsymbol{Z}_{\le m}\boldsymbol{\Sigma}_m^{-1/2}$, $\boldsymbol{G}_m \in \mathbb{R}^{n\times m}$ iid $\mathrm{N}(0,1)$ with $m \ll n$. Then $\boldsymbol{G}_m$ almost orthogonal:

$$\|\boldsymbol{G}_m^{\mathsf{T}}\boldsymbol{G}_m/n - \boldsymbol{I}_m\|_{\mathrm{op}} \lesssim \frac{m}{n} = n^{-\delta}\,.$$

The kernel matrix: $\lambda_{\mathrm{eff}} = \lambda_{>m} + \lambda$,

$$\boldsymbol{K} + \lambda \boldsymbol{I} = \boldsymbol{G}_m \boldsymbol{\Sigma}_m \boldsymbol{G}_m^{\mathsf{T}} + \lambda_{\mathrm{eff}} \boldsymbol{I}_n\,.$$

Asymptotics:

$$\begin{aligned}\text{Bias} &= \|\boldsymbol{\beta}_m - (\boldsymbol{\Sigma}_m + (\lambda_{\text{eff}}/n)\boldsymbol{I}_m)^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\beta}_m\|_2^2 + \|\boldsymbol{\beta}_{>m}\|_{\ell_2}^2 + o_{d,\mathbb{P}}(1)\,,\\ \text{Variance} &= o_{d,\mathbb{P}}(1)\,.\end{aligned}$$

Test error:

$$R_G(f_*, \hat{f}_\lambda) = \|\boldsymbol{\beta} - (\boldsymbol{\Sigma} + (\lambda_{\text{eff}}/n)\boldsymbol{I})^{-1}\boldsymbol{\Sigma}\boldsymbol{\beta}\|_{\ell_2}^2 + o_{d,\mathbb{P}}(1)\,.$$

For KRR:

$$R(f_*, \hat{f}_\lambda) = \|f_* - \mathbb{S}_\lambda f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\,,$$

with shrinkage operator:

$$\hat{f}_\lambda(\boldsymbol{x}) \approx \mathbb{S}_\lambda f_*(\boldsymbol{x}) = \sum_{j\geq 1} \frac{\lambda_j}{\lambda_j + \lambda_{\text{eff}}/n} \langle f_*, \phi_j\rangle_{L^2}\phi_j(\boldsymbol{x})\,.$$

KRR acts as a shrinkage operator

With spectral gap, KRR with finite data :

$$\hat{f}_\lambda = \arg\min_f \left\{ \frac{1}{n} \sum_{i \in [n]} (y_i - f(\boldsymbol{x}_i))^2 + \frac{\lambda}{n} \|f\|_{\mathcal{H}}^2 \right\},$$

replaced by effective problem with $n = \infty$ and $\lambda_{\text{eff}} = \lambda_{>m} + \lambda$:

$$\hat{f}_\lambda^{\text{eff}} = \arg\min_f \left\{ \mathbb{E}[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2] + \frac{\lambda_{\text{eff}}}{n} \|f\|_{\mathcal{H}}^2 \right\}.$$

Components: $\lambda_j \gg \lambda_{\text{eff}}/n$ perfectly fitted, $\lambda_j \ll \lambda_{\text{eff}}/n$ not fitted at all.

Phenomenology:

1) $\lambda_{>m}$ self-induced regularization from high-degree part of kernel,

2) Interpolator $\lambda = 0$ are optimal,

3) KRR learns $\mathrm{P}_{\leq m} f_*$ (smooth part) and doesn't learn at all $\mathrm{P}_{>m} f_*$ (and $\mathrm{P}_{>m} \hat{f}_\lambda$ spiky part for interpolation).

KRR acts as a shrinkage operator

With spectral gap, KRR with finite data :

$$\hat{f}_\lambda = \arg\min_f \left\{ \frac{1}{n} \sum_{i\in[n]} (y_i - f(\boldsymbol{x}_i))^2 + \frac{\lambda}{n} \|f\|_{\mathcal{H}}^2 \right\},$$

replaced by effective problem with $n = \infty$ and $\lambda_{\text{eff}} = \lambda_{>m} + \lambda$:

$$\hat{f}_\lambda^{\text{eff}} = \arg\min_f \left\{ \mathbb{E}[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2] + \frac{\lambda_{\text{eff}}}{n} \|f\|_{\mathcal{H}}^2 \right\}.$$

Components: $\lambda_j \gg \lambda_{\text{eff}}/n$ perfectly fitted, $\lambda_j \ll \lambda_{\text{eff}}/n$ not fitted at all.

Phenomenology:
1) $\lambda_{>m}$ self-induced regularization from high-degree part of kernel,
2) Interpolator $\lambda = 0$ are optimal,
3) KRR learns $\mathsf{P}_{\leq m} f_*$ (smooth part) and doesn't learn at all $\mathsf{P}_{>m} f_*$ (and $\mathsf{P}_{>m} \hat{f}_\lambda$ spiky part for interpolation).

Proving universality

What properties on $\phi(\boldsymbol{x})$ allow you to show universality with Gaussian model?

[Mei,**M.**,Montanari,'21]: hypercontractivity of the top eigenfunctions.

Assumptions on $(\phi^{(n)})_{n\geq 1}$:

- **Spectral gap:** exists $m(n)_{n\geq 1}$ and $\delta > 0$ such that $m \leq n^{1-\delta}$ and
$$\lambda_{m+1} \cdot n^{1+\delta} \leq \sum_{j>m} \lambda_j \, .$$
- **Hypercontractivity:** for any $q \geq 1$, there exists C_q such that
$$\|h\|_{L^{2q}} \leq C_q \|h\|_{L^2} \, , \qquad \forall h \in \text{span}(\psi_s^{(n)} : 1 \leq s \leq m\} \, .$$

E.g., low-degree polynomials for $\boldsymbol{x}$ Gaussian vector/uniform on hypercube/uniform on hypersphere.

Other abstract assumptions that show universality. However, difficult to check these assumptions in practice.

One example: inner-product kernel on the sphere

- $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $h : [-1, +1] \to \mathbb{R}$ PD, non-polynomial. Eigendecomposition of the kernel:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d) = \sum_{k=0}^{\infty} \xi_k \sum_{s \in [B(d,k)]} Y_{ks}(\boldsymbol{x}_1) Y_{ks}(\boldsymbol{x}_2),$$

where Y_{ks} degree-k spherical harmonics and $\xi_k = \Theta_d(B(d,k)^{-1}) = \Theta_d(d^{-k})$.

- For $d^{\ell+\delta} \leq n \leq d^{\ell+1-\delta}$, can take $m = \sum_{k \leq \ell} B(d,k) = \Theta(d^{\ell})$:

$$n \cdot \lambda_{m+1} = n \cdot \xi_{\ell+1} \geq d^{\delta}, \qquad \sum_{j>m} \lambda_j = \sum_{k>\ell} \xi_k B(d,k) = \Theta_d(1).$$

- For $j \leq m$, $\lambda_j \gg \lambda^{\text{eff}}/n$ are perfectly learned, $\lambda_j \ll \lambda^{\text{eff}}/n$ are not learned at all, i.e.,

$$\hat{f}_\lambda(\boldsymbol{x}) = \mathrm{P}_{\leq \ell} f_*(\boldsymbol{x}) + o_{d,\mathbb{P}}(1).$$

What about $n \asymp d^\ell$?

No spectral gap when $n \asymp d^\ell$:

$$\boldsymbol{K} \approx \boldsymbol{K}_{\leq \ell - 1} + \mu_\ell \frac{\boldsymbol{Y}_\ell \boldsymbol{Y}_\ell^\mathsf{T}}{B(d, \ell)} + mu_{>\ell} \boldsymbol{I}_n \, .$$

- $\boldsymbol{K}_{\leq \ell - 1}$: low-rank spike matrix.
- $\boldsymbol{K}_{>\ell} \approx \mu_{>\ell} \boldsymbol{I}_n$: self-induced reg from high-degree part.
- $\boldsymbol{Y}_\ell = (Y_{\ell s}(\boldsymbol{x}_i))_{i \in [n], s \in [B(d,\ell)]} \in \mathbb{R}^{n \times B(d,\ell)}$, iid rows. Covariance matrix. Spectrum converges to a Marchenko-Pastur law.

(Generalization of [El Karoui,'10] to the polynomial scaling.)

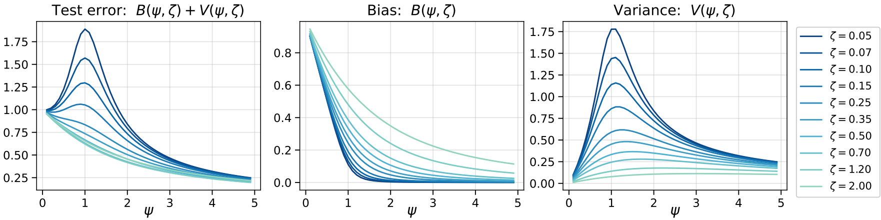
Fits completely degree-$(\ell - 1)$ polynomial approximation, none of the degree $> \ell$ components, and partially degree-ℓ components.

Test error $= \|P_{>\ell} f_*\|_{L^2}^2$ + the test error of ridge regression model with $\boldsymbol{x}_i \sim N(0, \boldsymbol{I}_B)$ and $y_i = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}\|_2 = \|P_\ell f_*\|_{L^2}$ and noise $\mathbb{E}[\varepsilon_i^2] = \|P_{>\ell} f_*\|_{L^2}^2 + \sigma_\varepsilon^2$ and regularization $\xi_\ell = (\mu_{>\ell} + \lambda)/\mu_\ell$:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \xi_\ell \|\boldsymbol{\beta}\|_2^2 \right\}.$$

As $n/B(d,\ell) \to \psi$:

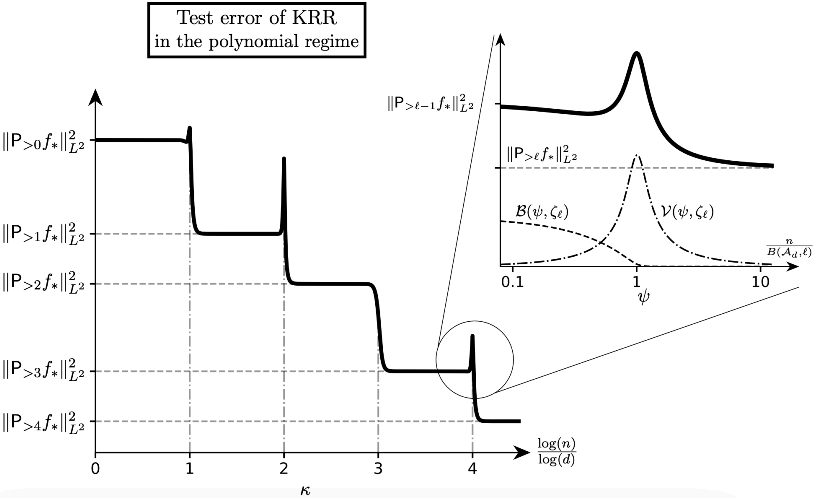
$$R(f_*; \hat{f}_\lambda) = \|P_\ell f_*\|_{L^2}^2 \cdot \mathcal{B}(\psi, \zeta_\ell) + (\|P_{>\ell} f_*\|_{L^2}^2 + \sigma_\varepsilon^2) \cdot \mathcal{V}(\psi, \zeta_\ell) + \|P_{>\ell} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1).$$



Asymptotics of KRR on the sphere in polynomial scaling

$\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $K(\boldsymbol{x};\boldsymbol{z}) = h(\langle \boldsymbol{x}, \boldsymbol{z}\rangle / d)$.

Asymptotics in polynomial scaling $n/d^\kappa \to \psi$ for any $\kappa, \psi > 0$:



Random features ridge regression

$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{w}\sim\nu}[\sigma(\boldsymbol{x}_1; \boldsymbol{w})\sigma(\boldsymbol{x}_2; \boldsymbol{w})]$ with $\sigma \in L^2(\mathcal{X} \times \mathcal{V})$.

Random feature approx: sample $(\boldsymbol{w}_s) \sim_{iid} \nu$ and replace K by

$$K_N(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{N} \sum_{s\in[N]} \sigma(\boldsymbol{x}_1; \boldsymbol{w}_s)\sigma(\boldsymbol{x}_2; \boldsymbol{w}_s) \, .$$

Random Features Ridge Regression (RFRR): fit a model $\hat{f}_{N,\lambda}(\boldsymbol{x}) = \frac{1}{N} \sum_{s\in[N]} \hat{a}_s \sigma(\boldsymbol{x}; \boldsymbol{w}_s)$ with

$$\hat{\boldsymbol{a}} = \arg\min_{a\in\mathbb{R}^N} \left\{ \sum_{i\in[n]} (y_i - f_N(\boldsymbol{x}_i; \boldsymbol{a}))^2 + \frac{\lambda}{N} \|\boldsymbol{a}\|_2^2 \right\} .$$

When $N \to \infty$, $\hat{f}_{N,\lambda} \to \hat{f}_\lambda$.

Again, σ can be seen as a compact operator and is diagonalizable:

$$\sigma(\boldsymbol{x};\boldsymbol{w}) = \sum_{j\geq 1} \sqrt{\lambda_j}\phi_j(\boldsymbol{x})\psi_j(\boldsymbol{x}) = \langle \Phi(\boldsymbol{x}), \Psi(\boldsymbol{w})\rangle_{\ell_2} ,$$

$$\Phi(\boldsymbol{x}) = (\lambda_j^{1/4}\phi_j(\boldsymbol{x}))_{j\geq 1} , \qquad \Psi(\boldsymbol{x}) = (\lambda_j^{1/4}\psi_j(\boldsymbol{x}))_{j\geq 1} ,$$

with $\{\phi_j\}$ orthonormal basis of $L^2(\mathcal{X})$ and $\{\psi_j\}$ orthonormal basis of $L^2(\mathcal{V})$.

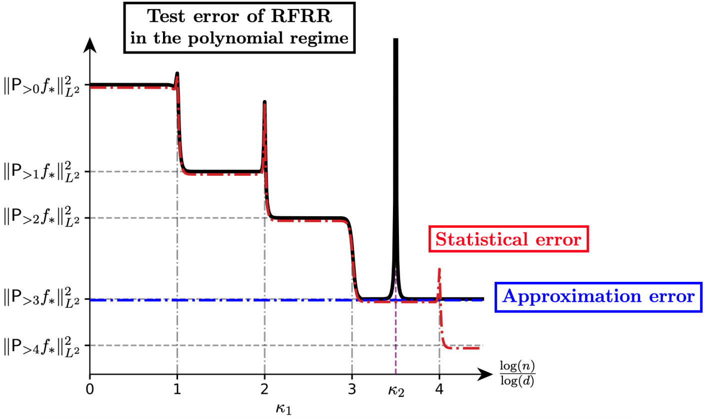
Gaussian equivalent model: $\Phi(\boldsymbol{x}_i) \leftrightarrow \boldsymbol{z}_i$, $\Psi(\boldsymbol{w}_j) \leftrightarrow \boldsymbol{g}_j$ and $\sigma(\boldsymbol{x}_i;\boldsymbol{w}_j) \leftrightarrow \langle \boldsymbol{z}_i, \boldsymbol{g}_j\rangle_{\ell_2}$.

$\boldsymbol{F} = (\sigma(\boldsymbol{x}_i;\boldsymbol{w}_j))_{i\in[n],j\in[N]}$, $\boldsymbol{Z} = [\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n]^{\mathsf{T}} \in \mathbb{R}^{n\times\infty}$, $\boldsymbol{G} = [\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n]^{\mathsf{T}} \in \mathbb{R}^{N\times\infty}$.

$$\hat{f}^G_{N,\lambda}(\boldsymbol{z}) = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{G}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z}^{\mathsf{T}}(\boldsymbol{Z}\boldsymbol{G}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z} + \lambda\boldsymbol{I}/N)^{-1}\boldsymbol{y}/N .$$

$\boldsymbol{x} \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\hat{f}_{RF}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i \in [N]} a_i \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle)$.

Asymptotics in polynomial scaling $n/d^{\kappa_1} \to \psi_1$, $N/d^{\kappa_2} \to \psi_2$ for any $\kappa_1, \kappa_2, \psi_1, \psi_2 > 0$:



Test error $\approx$ max(approzimation error, statistical error).

Application I: learning with group-invariance (1)

- Data invariant by group action $\mathcal{G}_d$ (subgroup of $\mathcal{O}(d)$):
  i.e., $f_*(g \cdot \boldsymbol{x}) = f_*(\boldsymbol{x})$ for all $g \in \mathcal{G}_d, \boldsymbol{x} \in \mathbb{S}^{d-1}(\sqrt{d})$.

- Comparison between learning with:
  - standard kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d)$;
  - $\mathcal{G}_d$-invariant kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathcal{G}_d} h(\langle \boldsymbol{x}_1, g \cdot \boldsymbol{x}_2 \rangle / d) \pi_d(\mathrm{d}g)$.

- **Group $\mathcal{G}_d$ of degeneracy α:** if for any $k \geq \alpha$,

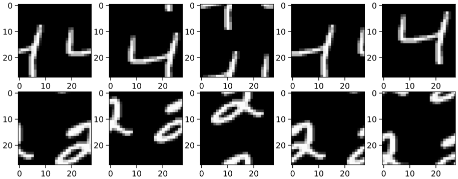
$$\frac{\dim(V_{d,k})}{\dim(V_{d,k}(\mathcal{G}_d))} \asymp d^{\alpha} \, ,$$

  $V_{d,k}$: space of degree-k polynomials; $V_{d,k}(\mathcal{G}_d)$: degree-k $\mathcal{G}_d$-invariant polynomials.
  Cyclic group: $\alpha = 1$ ($g_r \cdot \boldsymbol{x} = (x_{1+r}, x_{2+r}, \ldots, x_d, x_1, \ldots, x_r)$).

- To learn a degree-ℓ polynomial approximation needs $d^{\ell - \alpha}$ samples.
  (Gain of a factor d^{α} in sample size and number of random features.)

Application I: learning with group-invariance (2)

Cyclic invariant MNIST:



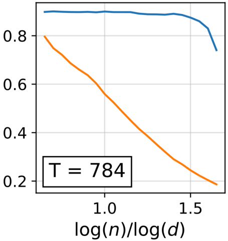


Figure: Test accuracy against number of samples (orange: cyclic kernel, blue: standard kernel).

Application II: learning with convolutional kernels

Covariates $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, patches of size q: $\boldsymbol{x}_{(k)} = (x_k, \ldots, x_{k+q-1})$.

NTK of 1-layer convolutional kernel followed by local average pooling:

$$H_{q,\omega}(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{d\omega} \sum_{k \in [d]} \sum_{s,s' \in [\omega]} h(\langle \boldsymbol{x}_{k+s}, \boldsymbol{z}_{k+s'} \rangle / q) \, .$$

For $\boldsymbol{x} \sim \text{Unif}(\{+1, -1\}^d)$, $H_{q,\omega}$ can be diagonalized and we can compute sharp asymptotics for the test error.

E.g., target function: $f_*(\boldsymbol{x}) = \frac{1}{d} \sum_{k \in [d]} P_\ell(\boldsymbol{x}_{(k)})$.

To fit f_*	H^{FC}	H^{FC}_{GP}	H^{CK}	H^{CK}_{ω}	H^{CK}_{GP}
Sample complexity	d^ℓ	$d^{\ell-1}$	$dq^{\ell-1}$	$dq^{\ell-1}/\omega$	$q^{\ell-1}$

H^{FC}: $q = d, \omega = 1$; H^{FC}_{GP}: $q = d, \omega = d$;
H^{CK}: $q, \omega = 1$; H^{CK}_{ω}: q, ω; H^{CK}_{GP}: $q, \omega = d$.

Application III: learning with anisotropic data (1)

Spiked covariates model: orthogonal matrix $[\boldsymbol{U}, \boldsymbol{U}^\perp]$

$$\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z}_1 + \boldsymbol{U}^\perp \boldsymbol{z}_2, \qquad \boldsymbol{z}_1 \in \mathbb{R}^{d_s}, \boldsymbol{z}_2 \in \mathbb{R}^{d-d_s}.$$

Signal part: $\boldsymbol{z}_1 \sim \mathrm{Unif}\left(\mathbb{S}^{d_s-1}(\sqrt{\mathrm{snr}_c \cdot d_s})\right)$.

Noise part: $\boldsymbol{z}_2 \sim \mathrm{Unif}\left(\mathbb{S}^{d-d_s-1}(\sqrt{d-d_s})\right)$

d_s = signal dimension.

snr_c = covariate SNR.

Target function: $f_\star(\boldsymbol{x}) = \varphi(\boldsymbol{z}_1)$.

Define effective dimension: $d_{\mathrm{eff}} = d_s \vee (d/\mathrm{snr}_c)$

$$\text{for } {d_{\mathrm{eff}}}^{\ell+\delta} \leq n \leq {d_{\mathrm{eff}}}^{\ell+1-\delta}, \qquad R(f_*, \hat{f}) = \|\mathrm{P}_{>\ell} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\,.$$

- Approx. isotropic data: $\mathrm{snr}_c \approx 1$, $d_{\mathrm{eff}} \approx d$.
- Very anisotropic data: $\mathrm{snr}_c \gg 1$, $d_{\mathrm{eff}} \approx d_s \ll d$. KRR much more efficient.

Application III: learning with anisotropic data (2)

For images: (1) Spectrum concentrates on low-frequencies;
(2) Labels depend predominantly on low-frequencies.

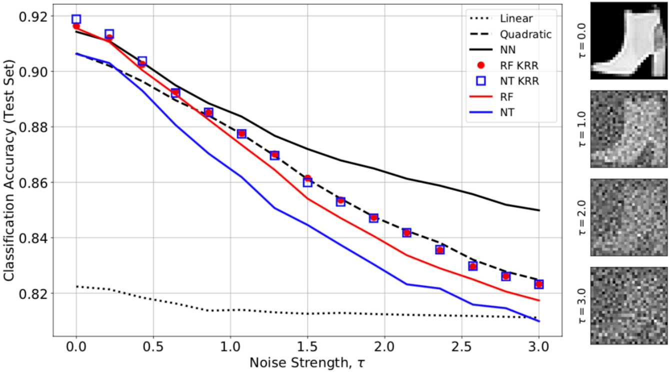


Figure: Test accuracy on FMNIST: adding noise to the high frequency components (decreasing snr_c, increasing $d_{\rm eff} = d/snr_c$).

Summary

- Gaussian equivalent model for "high-dimensional enough models".
- Can give very precise results which give clear conceptual picture.
- **More general type of universality:** entire feature maps + polynomial scaling $\log(n) \asymp \log(d)$.
- **Limitations:** hard to apply it to specific setting (most of the time, no explicit diagonalization).
- More general directions (ERM universality).

Thank you!

1. *Linearized two-layers neural networks in high dimension*. Ghorbani, Mei, **M.**, Montanari (2019).

2. *When do neural networks outperform kernel methods?* Ghorbani, Mei, **M.**, Montanari (2020).

3. *Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration*. Mei, **M.**, Montanari (2021).

4. *Learning with invariances in random features and kernel models*. Mei, **M.**, Montanari (2021).

5. *Learning with convolution and pooling operations in kernel methods*. **M.**, Mei (2021).

6. *Spectrum of inner-product kernel matrices in the polynomial scaling and multiple descent phenomenon in kernel ridge regression*. **M.** (2022).

+ some ongoing work.