# Kernel regression in high dimensions: Refined analysis beyond double descent

**Fanghui Liu (KU Leuven), Zhenyu Liao (UC Berkeley), Johan A.K. Suykens (KU Leuven)**

ESAT-STADIUS, KU Leuven

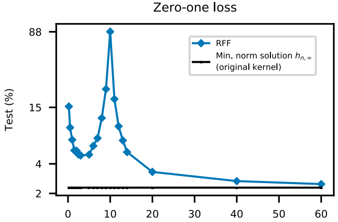
March 14, 2021

# Outline

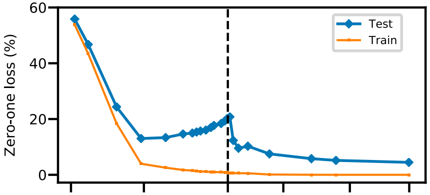
# Research Overview

## Understanding large dimensional machine learning

- high dimensions: large $n$ and $d$
- abnormal phenomena: training error can be zero but still generalize well




(a) Random features




(b) A fully connected neural network

Figure: Experiments on MNIST from [Belkin et al. PNAS2019.]

# Research Overview

## Understanding large dimensional machine learning

- double descent
- exist in over-parameterized models, e.g., neural networks, random features

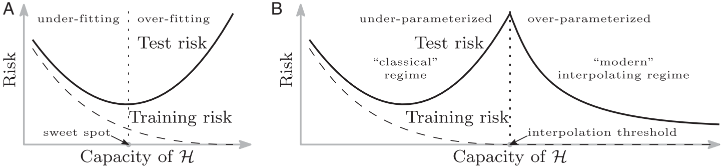



Figure: A cartoon by [Belkin et al. PNAS2019.]

# Research Overview

## Understanding large dimensional machine learning

- Kernel methods? different from random features:

**formulation: *primal* vs. *dual***

$$\text{RFF:} \quad k(\boldsymbol{x}, \boldsymbol{x}') \approx \varphi^\top(\boldsymbol{x})\varphi(\boldsymbol{x}')\,,$$

where $\varphi(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^s$ is an **explicit** feature mapping in $\mathbb{R}^s$ space.

**eigenvalue gap**

$\boldsymbol{Z} = \varphi(\boldsymbol{X}) \in \mathbb{R}^{n \times s}$, for large $d$ and take $s \to \infty$

$$\|\boldsymbol{K} - \boldsymbol{Z}^\top \boldsymbol{Z}\|_{\mathrm{F}} \to 0$$

$$\|\boldsymbol{K} - \boldsymbol{Z}^\top \boldsymbol{Z}\|_2 \nrightarrow 0$$

# Research Overview

## Interpolation learning generalizes well[1]

**Kernel "ridegeless" regression**

$$f_{\boldsymbol{z}} := \operatorname*{argmin}_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}, \quad \text{s.t.} \quad \underbrace{f(\boldsymbol{x}_i) = y_i}_{\mathcal{E}_{\boldsymbol{z}}(f)=0} .$$

**(Informal) Definition of Implicit regularization**

The property that an algorithm (solving the un-regularized problem) always pick up solutions with small excess risk.

Implicit regularization

- optimization: SGD, early stopping
- intrinsic structure: the curvature of kernel functions

[1] Liang and Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. Annals of Statistics, 2020.

# Research Overview

## Explicit regularization *vs.* Implicit regularization

### Kernel ridge regression (KRR)

Given a training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ and a kernel function $k$ in RKHS $\mathcal{H}$, KRR aims to solve the following empirical risk minimization (ERM)

$$f_{\boldsymbol{z},\lambda} := \operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^n \big(f(\boldsymbol{x}_i)-y_i\big)^2 + \lambda\langle f, f\rangle_{\mathcal{H}} \right\}. \tag{1}$$

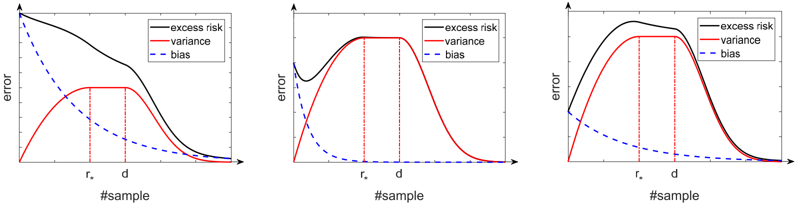
- closed-from solution: $f_{\boldsymbol{z},\lambda}(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{X})^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1}\boldsymbol{y}$.
- explicit regularization: $\lambda := \bar{c}n^{-\vartheta}$ with some $\vartheta \geq 0$ and $0 \leq \bar{c} \leq 1$.
- In KRR, the expected excess risk

$$\mathbb{E}_{y|\boldsymbol{x}}[\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho)] = \mathbb{E}_{y|\boldsymbol{x}}\|f_{\boldsymbol{z},\lambda} - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} := \text{Bias} + \text{Variance}$$

# Research Overview

## Our findings

- in high-dimensions, eigenvalue decay equivalence: $\boldsymbol{K}$ and $\boldsymbol{X}\boldsymbol{X}^\top/d$
- bias: independent of $d$, converges at a $\mathcal{O}(\lambda)$ rate
- variance: depends on $n, d$, can be unimodal or monotonic decreasing
- regularization: affects the position and value of the peak point




(a) decreasing (b) double descent (c) bell-shaped

# Outline

# (Basic) Assumptions

- **existence of $f_\rho$**: $f_\rho \in \mathcal{H}$
- **noise condition**: $\exists \sigma$ such that $\mathbb{E}[(f_\rho(\boldsymbol{x}) - y)^2 \mid \boldsymbol{x}] \leq \sigma^2$.
  uniformly bounded noise, sub-Gaussian noise
- **kernel functions**:
  1) inner-product kernels: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = h\left(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle / d\right)$
  2) *radial* kernels: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = h\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 / d\right)$
  Here $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is a nonlinear function that is assumed to be (locally) smooth.
- (8+$m$)-moments in high-dimensional statistics:
  Let $\boldsymbol{x}_i = \boldsymbol{\Sigma}_d^{1/2} \boldsymbol{t}_i$, satisfying i.i.d entries with $\mathbb{E}[\boldsymbol{t}_i(j)] = 0$, $\mathbb{V}[\boldsymbol{t}_i(j)] = 1$, and $\mathbb{E}(|\boldsymbol{t}_i(j)|) \leq C d^{\frac{2}{8+m}}$ such that $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] = \boldsymbol{\Sigma}_d$ with $\|\boldsymbol{\Sigma}_d\|_2 < \infty$

# Linearization of $\boldsymbol{K}$ in high dimensions

In high dimensions[2], $\|\boldsymbol{K} - \widetilde{\boldsymbol{K}}^{\mathrm{lin}}\|_2 \to 0$ as $n, d \to \infty$

$$\widetilde{\boldsymbol{K}}^{\mathrm{lin}} := \underbrace{\alpha \mathbf{1}\mathbf{1}^\top + \beta \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}}_{\triangleq \widetilde{\boldsymbol{X}}} + \underbrace{\gamma \boldsymbol{I}}_{\text{implicit regularization}} + \boldsymbol{T}\,, \tag{2}$$

| parameters | inner-product kernels | radial kernels |
|---|---|---|
| $\alpha$ | $h(0) + h''(0)\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{2d^2}$ | $h(2\tau) + 2h''(2\tau)\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{d^2}$ |
| $\beta$ | $h'(0)$ | $-2h'(2\tau)$ |
| $\gamma$ | $h(\tau) - h(0) - \tau h'(0)$ | $h(0) + 2\tau h'(2\tau) - h(2\tau)$ |
| $\boldsymbol{T}$ | $\mathbf{0}_{n\times n}$ | $h'(2\tau)\boldsymbol{A} + \frac{1}{2}h''(2\tau)\boldsymbol{A} \odot \boldsymbol{A}$ [1] |

[1] $\boldsymbol{A} := \mathbf{1}\boldsymbol{\psi}^\top + \boldsymbol{\psi}\mathbf{1}^\top$, where $\boldsymbol{\psi} \in \mathbb{R}^n$ with $\psi_i := \|\boldsymbol{x}_i\|_2^2/d - \tau$ and $\tau := \mathrm{tr}(\boldsymbol{\Sigma}_d)/d$.

[2] Karoui. The spectrum of kernel random matrices. Annals of Statistics, 2010.

# Main results

## Basic Results

### Theorem

*Under the above assumptions, for $d$ large enough,* $\lambda := \bar{c} n^{-\vartheta}$ *with* $0 \le \vartheta \le 1/2$, *for any given* $\varepsilon > 0$, *it holds with probability at least* $1 - 2\delta - d^{-2}$ *with respect to the draw of* $\boldsymbol{X}$ *that*

$$\mathbb{E}_{y|\boldsymbol{x}} \left\| f_{\boldsymbol{z},\lambda} - f_\rho \right\|^2_{\mathcal{L}^2_{\rho_X}} \lesssim \underbrace{\lambda \log^4\left(\frac{2}{\delta}\right)}_{\text{bounds for bias}} + \underbrace{\mathtt{V}_1 + \text{residual term}}_{\text{bounds for variance}}, \tag{3}$$

*where* $\mathtt{V}_1 := \frac{\sigma^2 \beta}{d} \mathcal{N}^{n\lambda + \gamma}_{\widetilde{\boldsymbol{X}}}$ *with*

$$\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} := \operatorname{tr}\left[ (\widetilde{\boldsymbol{X}} + b\boldsymbol{I}_n)^{-2} \widetilde{\boldsymbol{X}} \right] = \sum_{i=1}^n \frac{\lambda_i(\widetilde{\boldsymbol{X}})}{\left[ b + \lambda_i(\widetilde{\boldsymbol{X}}) \right]^2}.$$

# Main results

## Refined results

Refined results with two additional assumptions in approximation theory

- source condition: $f_\rho = L_K^r g_\rho, \quad$ with some $0 < r \leq 1$ and $g_\rho \in \mathcal{L}^2_{\rho_X}$
- capacity condition[3]: $\mathcal{N}(\lambda) := \operatorname{tr}\left((L_K + \lambda I)^{-1} L_K\right) \leq Q^2 \lambda^{-\eta}$ with $\eta \in [0, 1]$. (**corresponds to RKHS and eigenvalue decay**)

The bias $\mathtt{B}$ can be improved as ($r = 1/2$ and $\eta = 1$)

$$\mathtt{B} \lesssim \mathcal{O}(\lambda) \quad \rightarrow \quad \mathtt{B} \lesssim \mathcal{O}\left(\lambda n^{-2r}\right).$$

Note that $\eta$ is nearly independent of the learning rates.

[3]Strictly speaking, this would depend on $d$.

# Discussion on error bounds

## Eigenvalue decay of $\boldsymbol{K}$ or $\boldsymbol{X}\boldsymbol{X}^\top/d$ or $\widetilde{\boldsymbol{X}}$

$\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} = \sum_{i=1}^n \frac{\lambda_i(\widetilde{\boldsymbol{X}})}{\left[b+\lambda_i(\widetilde{\boldsymbol{X}})\right]^2}$ with $b := n\lambda + \gamma$, and $r_* := \operatorname{rank}(\widetilde{\boldsymbol{X}})$

| | $\lambda_i(\widetilde{\boldsymbol{X}})$ | | $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}}$ | |
|---|---|---|---|---|
| | $i \leq r_*$ | $i > r_*$ | $n < d$ | $n > d$ |
| *harmonic decay* | $n/i$ | 0 | $\mathcal{O}(\frac{n}{b^2})$ | 0 [1] |
| *polynomial decay* | $ni^{-2a}$ with $a > 1/2$ | 0 | $\mathcal{O}(\frac{1}{b}\left(\frac{n}{b}\right)^{\frac{1}{2a}})$ | 0 [1] |
| *exponential decay* | $ne^{-ai}$ with $a > 0$ | 0 | Bound[2] | 0 [1] |

[1] $\lim_{n\to\infty} \mathcal{N}^b_{\widetilde{\boldsymbol{X}}} = 0$

[2] $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} \leq \mathcal{O}\left(\frac{1}{b+ne^{-a(r_*+1)}} - \frac{1}{b+ne^{-a}}\right)$

# Discussion on error bounds

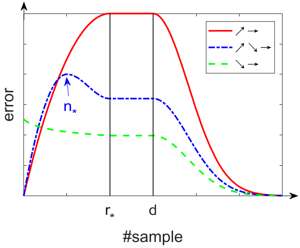
## harmonic decay

**Harmonic decay**: $\mathtt{V}_1 \leqslant \mathcal{O}(\frac{n}{b^2 d})$ $b := n\lambda + \gamma$ and $r_* = \mathrm{rank}(\boldsymbol{X}\boldsymbol{X}^\top)$

- $\lambda = 0$, $\mathtt{V}_1 \leq \mathcal{O}(\frac{n}{d})$
- $\lambda \neq 0$, $\mathtt{V}_1 \leqslant \mathcal{O}(\frac{n}{d(\bar{c}n^{1-\vartheta}+\gamma)^2})$, define $n_* = \operatorname{argmin}_n \frac{n}{d(\bar{c}n^{1-\vartheta}+\gamma)^2}$

  1. $\vartheta \geq \frac{1}{2(2-\bar{c})}$: $\nearrow \rightarrow$
  2. $\vartheta \leq \frac{1}{2(2-\bar{c})}$

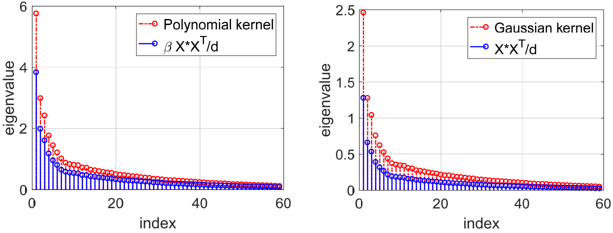
  1) $d < n_*$ 2) $r_* < n_* < d$ 3) $n_* < r* < d$ 4) $n_*$ is small enough



# Outline

# Numerical results

## Eigenvalue decay equivalence



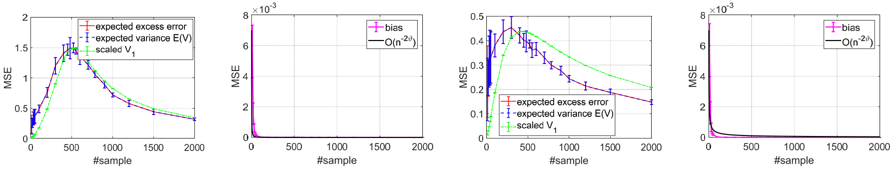

(d) *poly kernel*  (e) *Gaussian kernel*

Figure: Top 60 eigenvalues on the subset of the *YearPredictionMSD* dataset.

# Numerical results

Risk curve on synthetic dataset with $d = 500$ and set $\gamma = 0$ (implicit regularization)

we assume $y_i = f_\rho(\boldsymbol{x}_i) + \varepsilon$ with $f_\rho(\boldsymbol{x}) = \sin(\|\boldsymbol{x}\|_2^2)$ and Gaussian noise $\varepsilon \sim \mathcal{N}(0,1)$. The samples are generated from $\boldsymbol{x}_i = \boldsymbol{\Sigma}_d^{1/2} \boldsymbol{t}_i$ by
(i) take $\boldsymbol{\Sigma}_d$ as a diagonal matrix: $(\boldsymbol{\Sigma}_d)_{ii} \propto n/i$ in *harmonic decay*
(ii) take $\boldsymbol{T}$ as a random orthogonal matrix such that $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{T}^\top \boldsymbol{\Sigma}_d \boldsymbol{T}$ also has a harmonic eigendecay with $\boldsymbol{T}$ having almost i.i.d entries.



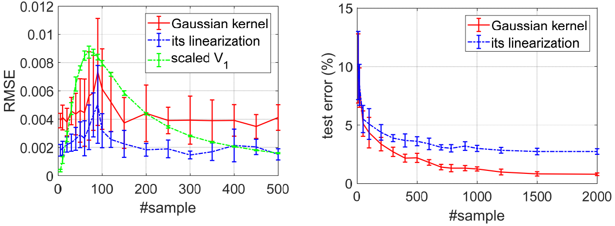

(a) $\vartheta = 2/3$ (b) $\vartheta = 2/3$ (c) $\vartheta = 1/3$ (d) $\vartheta = 1/3$

Figure: MSE of variance and bias $\mathcal{O}(n^{-2\vartheta r})$ with $r = 1$.

# Numerical results

## Risk curve on real-world datasets

We take $\lambda = 0$ and study implicit regularization $\gamma$




(a) *YearPredictionMSD* (b) *MNIST* (digits 3 vs. 7)

Figure: The test performance of the kernel interpolation estimator and its linearization one.

# Outline

Conclusion

- the **eigenvalue decay equivalence** between the kernel matrix and the data matrix in high-dimensions
- the **monotonic bias** and **unimodal variance**
- explicit and implicit regularization of kernel regression in high-dimensions

Future work

- extend $(8+m)$-moment assumption to distribution-free analysis
- the scale width, affect eigenvalue, $\mathcal{N}(\lambda)$

# Thanks for your attention!

# Q & A