Just interpolate: kernel “ridgeless” regression can generalize

Tengyuan Liang and Alexander Rakhlin, AoS 2020

Kunwoong Kim

January 17, 2022

Introduction

- General least-square objective is

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}}^2$$

  where $\mathcal{H}$ is the Hilbert space.
- The regularization parameter λ is a knob for balancing bias and variance.
- However, this paper shows that the test error decreases as λ decreases due to the *implicit regularization* using kernel based regression.
- Implicit regularization occurs by (1) the curvature of the kernel function and (2) data geometry for high-dimensional data.

Notations

Denote the true function as $f_\star(x) = \mathbf{E}(\mathbf{y}|\mathbf{x} = x)$.
The interpolation estimator studied in this paper is defined as

$$\hat{f} = \arg\min_{f \in \mathcal{H}} ||f||_\mathcal{H} \text{ s.t. } f(x_i) = y_i, \forall i.$$

When $K(X, X)$ is full rank, it is equivalent to

$$\hat{f}(x) = K(x, X)K(X, X)^{-1}Y$$

where $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, $K(X, X) = [K(x_i, x_j)]_{i,j} \in \mathbb{R}^{n \times n}$, and, for a new x, we denote $K(x, X) = [K(x, x_1), \ldots, K(x, x_n)] \in \mathbb{R}^{1 \times n}$.

Key quantities

- Denote $\Sigma_d = \mathbf{E}_\mu(x_i x_i^\top)$ the covariance matrix and the operator norm $||\Sigma_d||_{op}$.
- We set the kernel function as

$$K(x, x') = h(\frac{1}{d}\langle x, x' \rangle)$$

for some nonlinear smooth function $h(\cdot) : \mathbb{R} \to \mathbb{R}$ in a neighborhood of 0. Here, define

$$\alpha = h(0) + h''(0)\frac{Tr(\Sigma_d^2)}{d}$$

$$\beta = h'(0)$$

$$\gamma = h(\frac{Tr(\Sigma_d)}{d}) - h(0) - h'(0)\frac{Tr(\Sigma_d)}{d}$$

α, β, and γ are the quantities related to the curvature of $h(\cdot)$.

Main result

Assumptions

1. High dimensionality: $\exists c, C > 0$ such that $c \leq d/n \leq C$. $||\Sigma_d||_{op} \leq 1$
2. (8+m) moments: $|(\Sigma_d)^{-1/2} x_i)_j| \leq C d^{\frac{2}{8+m}}$ for all $1 \leq j \leq d$ and some $m > 0$.
3. Noise condition: $\exists \sigma > 0$ such that $\mathbb{E}((f_\star(\mathbf{x}) - y)^2 | \mathbf{x} = x) \leq \sigma^2$ for all x.
4. Nonlinear kernel: $K(x, x) \leq M$ for any x, where $K(x, x') = h(\frac{1}{d}\langle x, x' \rangle)$.

Main result (Theorem 1)

$$\mathbb{E}_{Y|X}||\hat{f} - f_\star||^2 \leq \phi_{n,d}(X, f_\star) + \epsilon(n, d) \quad (1)$$

with probability at least $1 - 2\delta - d^{-2}$ where

$$\begin{aligned} \phi_{n,d}(X, f_\star) &= \mathbf{V} + \mathbf{B} \\ &= \frac{8\sigma^2||\Sigma_d||_{op}}{d} \sum_j \frac{\lambda_j(\frac{XX^\top}{d} + \frac{\alpha}{\beta}11^\top)^2}{(\frac{\gamma}{\beta} + \lambda_j(\frac{XX^\top}{d} + \frac{\alpha}{\beta}11^\top))^2} \\ &+ ||f_\star||^2_{\mathcal{H}} \inf_{0 \leq k \leq n} \left(\frac{1}{n} \sum_{j>k} \lambda_j(\mathbf{K}_X \mathbf{K}_X^\top) + 2M\sqrt{\frac{k}{n}} \right). \end{aligned} \quad (2)$$

and $\epsilon(n, d) = \mathcal{O}(d^{-\frac{m}{8+m}} \log^{4.1} d) + \mathcal{O}(n^{-\frac{1}{2}} \log^{0.5}(n/\delta))$.

Main result

Assume σ^2 and $||f_\star||_{\mathcal{H}}^2$ are guessed.

Message

- V and B do not depend on λ (only depends on α, β, and γ.).
- V decreases ($\hat{f}$ is generalized) as γ increases and when the data matrix enjoys certain decay of the eigenvalues.
- B decreases as the eigenvalue decay of K is fast.

Example: What if using linear kernel (i.e., $h(a) = a$)?

- Since $\gamma = 0$, $\mathbf{V}$ becomes very large if $\lambda_j(\frac{XX^\top}{d} + \frac{\alpha}{\beta}11^\top)$ are small. In contrast, curvature of h introduces *implicit regularization* through a nonzero γ.

Behavior of the data-dependent bound

Let $K(x, x') = \exp(\frac{2||x-x'||}{d})$ with $\mathbf{r} = \gamma/\beta \asymp (\frac{Tr(\Sigma_d)}{d})^2$.

- Case $n > d$. The bounds are

$$\mathbf{V} \precsim \frac{1}{n} \sum_{j=1}^{d} \frac{\lambda_j(XX^\top/n)}{(\frac{d}{n}\mathbf{r} + \lambda_j(XX^\top/n))^2}$$

  and

$$\mathbf{B} \precsim \mathbf{r} + \frac{1}{d} \sum_{j=1}^{d} \lambda_j(XX^\top/n).$$

  Here, $\mathbf{r}$ controls the trade-off between $\mathbf{V}$ and $\mathbf{B}$.

- Case $d > n$. The bounds are

$$\mathbf{V} \precsim \frac{1}{d} \sum_{j=1}^{n} \frac{\lambda_j(XX^\top/d)}{(\mathbf{r} + \lambda_j(XX^\top/d))^2}$$

  and

$$\mathbf{B} \precsim \mathbf{r} + \frac{1}{n} \sum_{j=1}^{n} \lambda_j(XX^\top/d).$$

Confirmation of trade-off using synthetic data

- Parametrize the eigenvalues of covariance as $\lambda_j(\Sigma_d) = (1 - ((j-1)/d)^{\kappa})^{1/\kappa}$ where κ controls approximate "low-rankness" of the data: the closer κ is to 0, the faster does the spectrum of the data decay.
- Use the RBF kernel $k(x, x') = \exp(-||x - x'||^2/d)$.
- Target nonlinear function $f_\star(x) = \sum_{l=1}^{100} K(x, \theta_l)$ where $\theta_l \sim \mathcal{N}(0, I_d)$.
- Then, generate as $x_i \sim \mathcal{N}(0, \Sigma_{d,k}), y_i = f_\star(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some σ^2.

Small κ (fast spectral decay) $\rightarrow$ Large $\mathbf{V}$.
Big κ (slow spectral decay) $\rightarrow$ Large $\mathbf{B}$.

Examples

- $n > d$
  (Low rank) $\Sigma_d = diag(1, \ldots, 1, 0, \ldots, 0)$ with ϵd ones. Then, $\mathbf{r} = \epsilon^2$ and $\lambda_j(XX^\top/n) \geq (1 - \sqrt{\epsilon d/n})^2$ with high prob. Then,
  $$\mathbf{V} \precsim \frac{d}{n}\epsilon \text{ and } \mathbf{B} \precsim \epsilon^2 + \epsilon.$$
  Thus $\mathbf{V}, \mathbf{B} \to 0$ as $\epsilon \to 0$ for $n > d$.
- $d > n$
  (Favorable spectral decay) If $\mathbf{r}^{1/2} = Tr(\Sigma_d)/d = \mathcal{O}((n/d)^{1/3})$,
  $$\mathbf{V} \precsim \frac{n}{d}\frac{1}{4\mathbf{r}} \text{ and } \mathbf{B} \precsim \mathbf{r}^{1/2},$$
  thus $\mathbf{V}, \mathbf{B} \to 0$ for $d >> n$.

Confirmation of trade-off using synthetic data $(n > d)$

TABLE 1
Case $n > d$: variance bound **V** (4.1), *bias bound* **B** (4.2)

Spectral Decay	Implicit Reg	$n/d = 5$		$n/d = 20$	
		V	**B**	**V**	**B**
$\kappa = e^{-1}$	0.005418	14.2864	0.07898	9.4980	0.07891
$\kappa = e^{0}$	0.2525	0.4496	0.7535	0.1748	0.7538
$\kappa = e^{1}$	0.7501	0.1868	1.6167	0.05835	1.6165

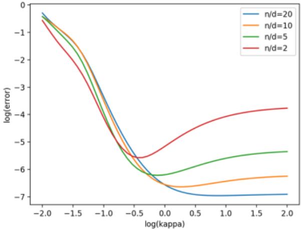

FIG. 2. *Generalization error as a function of varying spectral decay. Here, $d = 200$, $n = 400, 1000, 2000, 4000$.*

Confirmation of trade-off using synthetic data ($d > n$)

TABLE 2
Case $d > n$: variance bound **V** (4.3), *bias bound* **B** (4.4)

Spectral Decay	Implicit Reg	$d/n = 5$		$d/n = 20$	
		V	**B**	**V**	**B**
$\kappa = e^{-1}$	0.005028	3.9801	0.07603	0.7073	0.07591
$\kappa = e^{0}$	0.2503	0.1746	0.7513	0.04438	0.7502
$\kappa = e^{1}$	0.7466	0.06329	1.6106	0.01646	1.6102

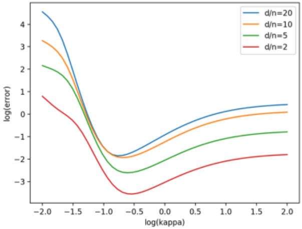

FIG. 3. *Generalization error as a function of varying spectral decay. Here, $n = 200$, $d = 400, 1000, 2000, 4000$.*

Experiments

MNIST

- Use the RBF kernel $k(x, x') = \exp(-||x - x'||^2/d)$ where $d = 784$.
- Binary classification: $10C_2$ experiments, with many λs.
- Measure: out-of-sample test error $\frac{\sum_i (\hat{f}(x_i) - y_i)^2}{\sum_i (\bar{y} - y_i)}$

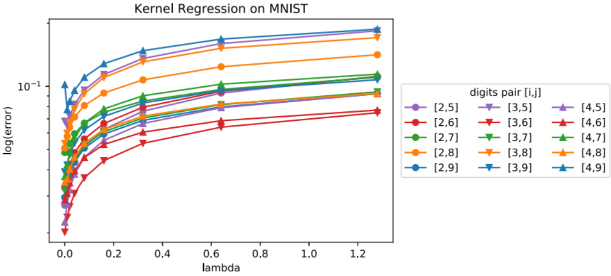

FIG. 4. *Test error, normalized as in (6.1). The y-axis is on the log scale.*

Experiments

Synthetic dataset

- Use the RBF kernel $k(x, x') = \exp(-||x - x'||^2/d)$.
- Target nonlinear function $f_\star(x) = \sum_{l=1}^{100} K(x, \theta_l)$ where $\theta_l \sim \mathcal{N}(0, I_d)$.
- Generating data: $x_i \sim \mathcal{N}(0, \Sigma_{d,k}), y_i = f_\star(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1, 0.5$.
- Measure: out-of-sample test error $\frac{\sum_i (\hat{f}(x_i) - y_i)^2}{\sum_i (\bar{y} - y_i)}$

Experiments

Synthetic dataset

- For a general pair of high dimensionality ratio n/d, there is a sweet spot of κ (favorable geometric structure) such that the trade-off is optimized.

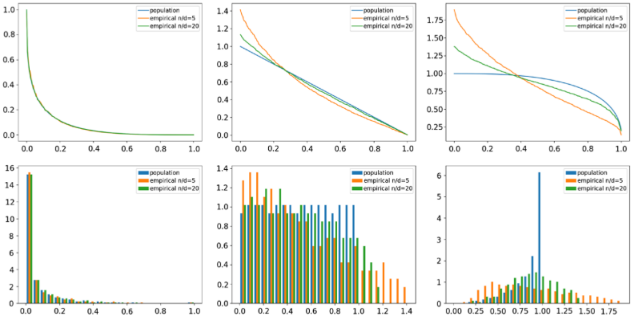

FIG. 6. *Varying spectral decay: case $n > d$. Columns from left to right: $\kappa = e^{-1}, e^{0}, e^{1}$. Rows from top to bottom: ordered eigenvalues, and the histogram of eigenvalues. Here, we plot the population eigenvalues for Σ_d, and the empirical eigenvalues for X^*X/n. In this simulation, $d = 100$, $n = 500, 2000$.*

Experiments

Synthetic dataset

- For a general pair of high dimensionality ratio d/n, there is a sweet spot of κ (favorable geometric structure) such that the trade-off is optimized.

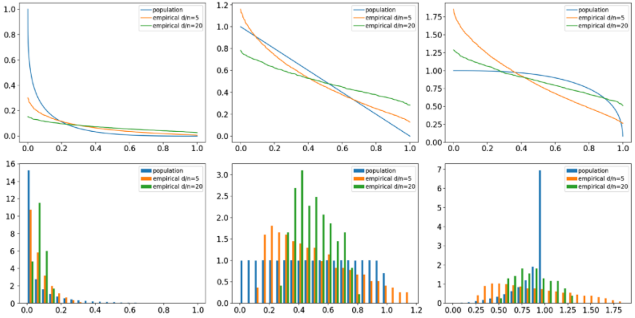

FIG. 7. *Varying spectral decay: case $d > n$. Columns from left to right: $\kappa = e^{-1}, e^0, e^1$. Rows from top to bottom: ordered eigenvalues, and the histogram of eigenvalues. Here, we plot the population eigenvalues for Σ_d, and the empirical eigenvalues for XX^*/d. In this simulation, $d = 2000$, $n = 400, 100$.*