University of Basel

NEURAL INFORMATION PROCESSING SYSTEMS

Learning Curve in Kernel Ridge Regression

Tin Sum Cheng, November 15, 2024

Introduction

A Comprehensive Analysis on the Learning Curve in Kernel Ridge Regression

Tin Sum Cheng
Department of Computer Science
University of Basel
Basel, Switzerland
tinsum.cheng@unibas.ch

Aurelien Lucchi
Department of Computer Science
University of Basel
Basel, Switzerland
aurelien.lucchi@unibas.ch

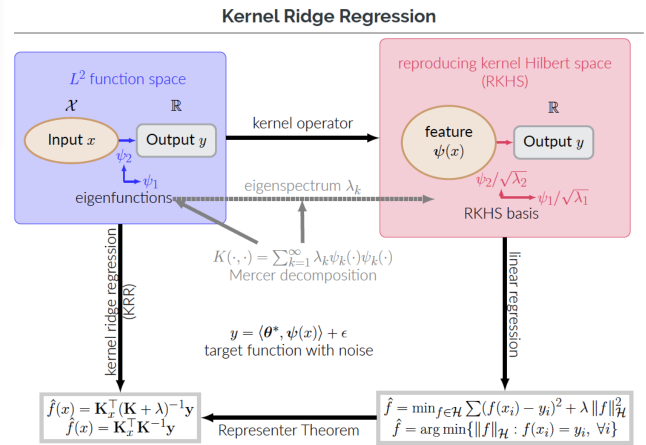
Anastasis Kratsios
Department of Mathematics
McMaster University and The Vector Institute
Ontario, Canada
kratsioa@mcmaster.ca

David Belius
Faculty of Mathematics and Computer Science
UniDistance Suisse
Switzerland david.belius@cantab.ch

Abstract

This paper conducts a comprehensive study of the learning curves of kernel ridge regression (KRR) under minimal assumptions. Our contributions are three-fold: 1) we analyze the role of key properties of the kernel, such as its spectral eigen-decay, the characteristics of the eigenfunctions, and the smoothness of the kernel; 2) we demonstrate the validity of the Gaussian Equivalent Property (GEP), which states that the generalization performance of KRR remains the same when the whitened features are replaced by standard Gaussian vectors, thereby shedding light on the analysis success of previous analyzes under the Gaussian Design Assumption; 3) we derive novel bounds that improve over existing bounds across a broad range of setting such as (in)dependent feature vectors and various combinations of eigen-decay rates in the over/underparameterized regimes.

Kernel Ridge Regression (KRR)



Learning Curve (in number of samples)

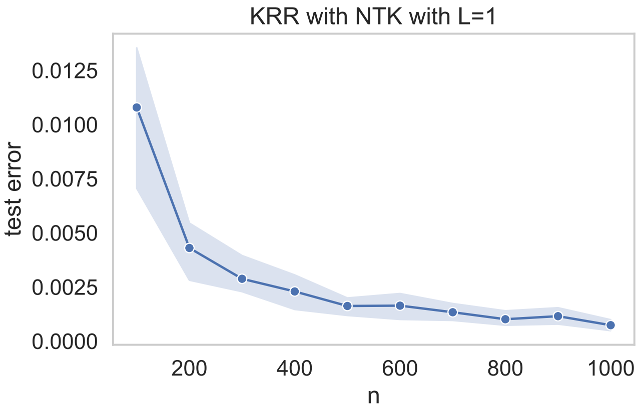


Figure: The test error decreases with sample size n at a certain rate.

Possible Settings

Assumption (IF - independent features) The random feature vector has independent sub-Gaussian entries.
Assumption (GF - generic features) The random feature vector has entries which exhibit some concentration results. Kernels which feature vectors satisfies Assumption **(GF)**:

1. dot-product kernels on hyperspheres;
2. kernels with bounded eigenfunctions;
3. radial base function (RBF) and shift-invariant kernels;
4. kernels on hypercubes.

Assumption (PE - polynomial decay)
$\lambda_k = \Theta_k\left(k^{-(1+a)}\right),\ \theta_k^* = \Theta_k\left(k^{-r}\right)$ for some constants $a, r > 0$. Source coefficient $s = \frac{2r+a}{1+a}$. Ridge $\lambda = \Theta_n\left(n^{-b}\right)$.
Assumption (EE - exponential decay)
$\lambda_k = \Theta_k\left(e^{-ak}\right),\ \theta_k^* = \Theta_k\left(e^{-kr}\right)$ for some constants $a, r > 0$. Source coefficient $s = \frac{2r}{a} + 1$. Ridge $\lambda = \Theta_n\left(e^{-bn}\right)$.

(Partial) Result

$$\text{test error} = \overbrace{\text{bias}}^{\mathcal{B}} + \overbrace{\text{variance}}^{\mathcal{V}}.$$

Ridge		strong		weak	
Feature		**(IF)**	**(GF)**	**(IF)**	**(GF)**
(PE)	$\mathcal{B}$	$\Theta\left(n^{-b\tilde{s}}\right)$	$\mathcal{O}\left(n^{-b\tilde{s}}\right)$	$\Theta\left(n^{-(1+a)\tilde{s}}\right)$	A novel bound
	$\mathcal{V}$	$\Theta\left(\sigma^2 n^{-1+\frac{b}{a+1}}\right)$	$\mathcal{O}\left(\sigma^2 n^{-1+\frac{b}{a+1}}\right)$	$\Theta\left(\sigma^2\right)$	$\tilde{\mathcal{O}}\left(\sigma^2 n^{2a}\right)$
(EE)	$\mathcal{B}$	$\Theta\left(e^{-b\tilde{s}n}\right)$	$\mathcal{O}\left(e^{-b\tilde{s}n}\right)$	$\mathcal{O}\left(e^{-a\tilde{s}n}\right),\ s>1$	$\mathcal{O}\left(e^{-a\tilde{s}n}\right),\ s>1$
	$\mathcal{V}$	$\Theta\left(\sigma^2 n^{-1+\frac{b}{a}}\right)$	$\mathcal{O}\left(\sigma^2 n^{-1+\frac{b}{a}}\right)$	catastrophic overfitting	

Table: *KRR Learning curve: n* is the sample size, $a, r > 0$ define the *eigen-decay rates* of the kernel and target function, $b > 0$ controls the decay rate of the ridge regularization parameter , $\sigma^2 \stackrel{\text{def.}}{=} \mathbb{E}\left[\epsilon^2\right]$ is the *noise level* and source coefficient s defined in Assumptions **(PE)** and **(EE)**. Here $\tilde{s} \stackrel{\text{def.}}{=} \min\{s, 2\}$. Results in blue indicate either previously unstudied regimes or improvements in available rates in a studied regime.

A Novel Bound on the Bias term

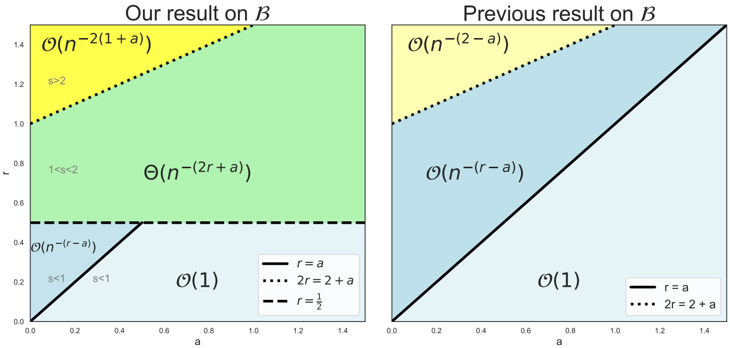


Figure: Phase diagram of the bound of the bias term $\mathcal{B}$ under weak ridge and polynomial eigen-decay. Our result is on the left, which improves over previous result from [1] on the right. On the left plot, the range of the source coefficient $s = \frac{2r+a}{1+a}$ is shown in gray font in each colored region.

Catastrophic Overfitting with (EE)

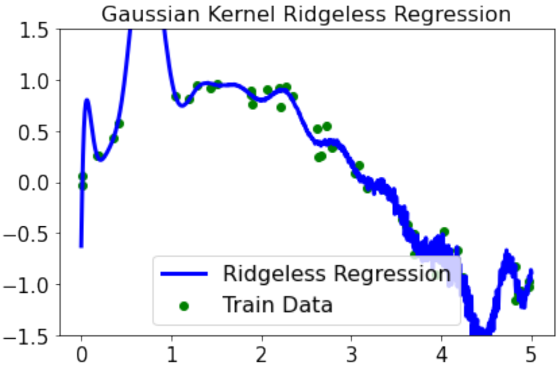


Figure: It is well known that kernels with exponential eigen-decay suffers from catastrophic overfitting.

Catastrophic/tempered Overfitting with (PE)

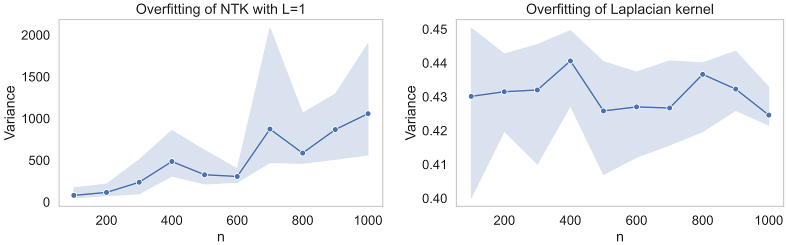


Figure: **Kernels with polynomial eigen-decay fitting pure noise on unit 2-disk.** (left): Neural tangent kernel (with 1 hidden layer) exhibits catastrophic overfitting. (right): Laplacian exhibits tempered overfitting.

Gaussian Equivalence Property (GEP)

Previous literature [2]–[4] replace feature vectors by Gaussian random vectors to obtain KRR learning curve, which agree with the empirical results. This phenomenon is called GEP.

When and why does the Gaussian Equivalence Property (GEP) exist?
we provide the same non-asymptotic bounds for both cases under a strong ridge. However, GEP does not hold under weak ridge!

Matching Lower Bound

Ridge		strong		weak	
Feature		**(IF)**	**(GF)**	**(IF)**	**(GF)**
(PE) or **(EE)**	$\mathcal{B}$	✓	✓	✓	✓ (when $1 \leq s \leq 2$)
	$\mathcal{V}$	✓	unknown	✓	× see Figure 4

Table: The table shows whether the lower bound is matching the upper bound deduced in this paper.

Master Inequalities

Using results from [1], [5]:

$$\mathcal{B} \leq \left(\frac{1+\rho^2\zeta^2\xi^{-1}+\rho}{\delta}\right)\|\boldsymbol{\theta}^*_{>k}\|^2_{\boldsymbol{\Sigma}_{>k}} + (\zeta^2\xi^{-2}+\rho\zeta^2\xi^{-1})\frac{s_1(\mathbf{A}_k)^2}{n^2}\left\|\boldsymbol{\theta}^*_{\leq k}\right\|^2_{\boldsymbol{\Sigma}^{-1}_{\leq k}}$$

$$\mathcal{V}/\sigma^2 \leq \rho^2\left(\zeta^2\xi^{-1}\frac{k}{n} + \frac{\mathrm{Tr}[\mathbf{Z}_{>k}\boldsymbol{\Sigma}^2_{>k}\mathbf{Z}^\top_{>k}]}{n\,\mathrm{Tr}[\boldsymbol{\Sigma}^2_{>k}]}\frac{r_k(\boldsymbol{\Sigma})^2}{nR_k(\boldsymbol{\Sigma})}\right)$$

- the "probably constant" part: random matrix theory
- the "decay" part: simple calculus

Generic Feature

Let $\mathbf{x} \in \mathbb{R}^p$ be the random feature vector with covariance $\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$. Let $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ be the whitened feature. Assumption (GF): for all $k \in \mathbb{N}$, assume that

$$\alpha_k \stackrel{\text{def.}}{=} \operatorname*{ess\,inf}_{\mathbf{z}} \frac{\|\mathbf{z}_{>k}\|^2_{\boldsymbol{\Sigma}_{>k}}}{\mathrm{Tr}[\boldsymbol{\Sigma}_{>k}]} = \Theta_k(1),$$

$$\beta_k \stackrel{\text{def.}}{=} \operatorname*{ess\,sup}_{\mathbf{z}} \max\left\{ \frac{\|\mathbf{z}_{\le k}\|_2^2}{k}, \frac{\|\mathbf{z}_{>k}\|^2_{\boldsymbol{\Sigma}_{>k}}}{\mathrm{Tr}[\boldsymbol{\Sigma}_{>k}]}, \frac{\|\mathbf{z}_{>k}\|^2_{\boldsymbol{\Sigma}^2_{>k}}}{\mathrm{Tr}[\boldsymbol{\Sigma}^2_{>k}]} \right\} = \Theta_k(1).$$

Reason: $\mathbb{E}_{\mathbf{z}}\left[\frac{\|\mathbf{z}_{\le k}\|_2^2}{k}\right] = \mathbb{E}_{\mathbf{z}}\left[\frac{\|\mathbf{z}_{>k}\|^2_{\boldsymbol{\Sigma}_{>k}}}{\mathrm{Tr}[\boldsymbol{\Sigma}_{>k}]}\right] = \mathbb{E}_{\mathbf{z}}\left[\frac{\|\mathbf{z}_{>k}\|^2_{\boldsymbol{\Sigma}^2_{>k}}}{\mathrm{Tr}[\boldsymbol{\Sigma}^2_{>k}]}\right] = 1.$

Implicit Regularization

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the input block. Recall the ridge regressor:

$$\hat{\theta} = \mathbf{X}^\top (\underbrace{\mathbf{X}\mathbf{X}^\top + n\lambda \mathbf{I}_n}_{\mathbf{A}})^{-1} \mathbf{y} \in \mathbb{R}^p$$

Write $\mathbf{X} = (\mathbf{X}_{\leq k} | \mathbf{X}_{>k})$ and

$$\mathbf{A} = \underbrace{\mathbf{X}_{\leq k}\mathbf{X}_{\leq k}^\top}_{\text{fit target}} + \overbrace{\underbrace{\mathbf{X}_{>k}\mathbf{X}_{>k}^\top}_{\text{implicit reg.}} + \underbrace{n\lambda \mathbf{I}_n}_{\text{explicit reg.}}}^{\mathbf{A}_k}$$

Concentration Coefficients

Master inequalities:

$$\mathcal{B} \leq \left(\frac{1+\rho^2\zeta^2\xi^{-1}+\rho}{\delta}\right)\|\theta^*_{>k}\|^2_{\mathbf{\Sigma}_{>k}} + (\zeta^2\xi^{-2}+\rho\zeta^2\xi^{-1})\frac{s_1(\mathbf{A}_k)^2}{n^2}\left\|\theta^*_{\leq k}\right\|^2_{\mathbf{\Sigma}^{-1}_{\leq k}}$$

$$\mathcal{V}/\sigma^2 \leq \rho^2\left(\zeta^2\xi^{-1}\frac{k}{n} + \frac{\mathrm{Tr}[\mathbf{Z}_{>k}\mathbf{\Sigma}^2_{>k}\mathbf{Z}^\top_{>k}]}{n\,\mathrm{Tr}[\mathbf{\Sigma}^2_{>k}]}\frac{r_k(\Sigma)^2}{nR_k(\Sigma)}\right)$$

Concentration Coefficients:

$$\xi_{n,k} \overset{\text{def.}}{=} \frac{s_1(\mathbf{Z}^\top_{\leq k}\mathbf{Z}_{\leq k})}{n}; \quad \zeta_{n,k} \overset{\text{def.}}{=} \frac{s_1(\mathbf{Z}^\top_{\leq k}\mathbf{Z}_{\leq k})}{s_k(\mathbf{Z}^\top_{\leq k}\mathbf{Z}_{\leq k})}; \quad \rho_{n,k} \overset{\text{def.}}{=} \frac{n\left\|\mathbf{\Sigma}_{>k}\right\|_{op}+s_1(\mathbf{A}_k)}{s_n(\mathbf{A}_k)}$$

where $\mathbf{Z}_{\leq k} \overset{\text{def.}}{=} \mathbf{X}_{\leq k}\mathbf{\Sigma}^{-1/2}_{\leq k} \in \mathbb{R}^{n\times k}$.

Concentration Coefficients

Let $k \in \mathbb{N}$ be an integer. Recall that $\xi_{n,k} \overset{\text{def.}}{=} \frac{s_1(\mathbf{Z}_{\leq k}^\top \mathbf{Z}_{\leq k})}{n}$. If Assumption (GF) (or resp. (IF)) holds, then with probability at least $1 - 2\exp(-\frac{1}{2\beta_k^2} n)$ (or resp. $1 - 2\exp(-c_1 kn)$), it holds that

$$\xi_{n,k} \geq \frac{1}{2}.$$

Proof:
Since the largest singular value is larger than the average of the singular values,

$$\xi_{n,k} \overset{\text{def.}}{=} \frac{s_1(\mathbf{Z}_{\leq k}^\top \mathbf{Z}_{\leq k})}{n} \geq \frac{\frac{1}{k}\operatorname{Tr}[\mathbf{Z}_{\leq k}^\top \mathbf{Z}_{\leq k}]}{n} = \frac{\operatorname{Tr}[\mathbf{Z}_{\leq k}^\top \mathbf{Z}_{\leq k}]}{kn}.$$

Concentration Coefficients

If Assumption (GF) holds, then

$$\mathrm{Tr}[\mathbf{Z}_{\le k}^\top \mathbf{Z}_{\le k}] = \mathrm{Tr}[\mathbf{Z}_{\le k}\mathbf{Z}_{\le k}^\top] = \sum_{i=1}^{n} \|(\mathbf{z}_i)_{\le k}\|_2^2 \le \beta_k kn.$$

Set $M = \beta_k k$ and by Hoeffding's inequality, the above trace concentrates:

$$\mathbb{P}\left(\left|\mathrm{Tr}[\mathbf{Z}_{\le k}\mathbf{Z}_{\le k}^\top] - kn\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{nM^2}\right)$$

Set $t = nk/2$ to conclude the statement.
Analogously, if Assumption (IF) holds, for $i = 1, ..., n$ and $l = 1, ..., k$, $(z_i^{(l)})^2 - 1$ is centered sub-exponential variable with sub-exponential norm $\left\|(z_i^{(l)})^2 - 1\right\|_{\psi_1} \lesssim G^2$. With probability at least $1 - 2\exp(-c_1 kn)$,

$$\left|\mathrm{Tr}[\mathbf{Z}_{\le k}^\top \mathbf{Z}_{\le k}] - kn\right| = \left|\sum_{i=1}^{n}\sum_{l=1}^{k}(z_i^{(l)})^2 - kn\right| \le \frac{1}{2}kn.$$

Our Team

Tin Sum Cheng
University of Basel

Aurelien Lucchi
University of Basel

Anastasis Kratsios
McMaster University

David Belius
UniDistance Suisse

University of Basel

Thank you
for your attention.

References I

[1] D. Barzilai and O. Shamir, *Generalization in kernel regression under realistic assumptions*, 2024. arXiv: 2312.15995 [cs.LG].

[2] N. Mallinar, J. B. Simon, A. Abedsoltan, P. Pandit, M. Belkin, and P. Nakkiran, „Benign, tempered, or catastrophic: A taxonomy of overfitting“, *Annual Conference on Neural Information Processing Systems*, 2022.

[3] H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborová, „Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime“, in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 10 131–10 143.

References II

[4] J. B. Simon, M. Dickens, D. Karkada, and M. R. DeWeese, *The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks*, 2021. arXiv: 2110.03922 [cs.LG].

[5] A. Tsigler and P. L. Bartlett, „Benign overfitting in ridge regression.“, *J. Mach. Learn. Res.*, vol. 24, pp. 123–1, 2023.