Precise Learning Curves and Higher-Order Scaling Limits for Dot Product Kernel Regression

Lechao Xiao[1], Hong Hu[2], Theodor Misiakiewicz[3], Yue M. Lu [4] and Jeffrey Pennington[1]
xlc@google.com, huhong@wharton.upenn.edu, misiakie@stanford.edu, yuelu@seas.harvard.edu, jpennin@google.com
1. Google Research, Brain Team, 2. University of Pennsylvania, 3. Stanford University, 4. Harvard University

Abstract

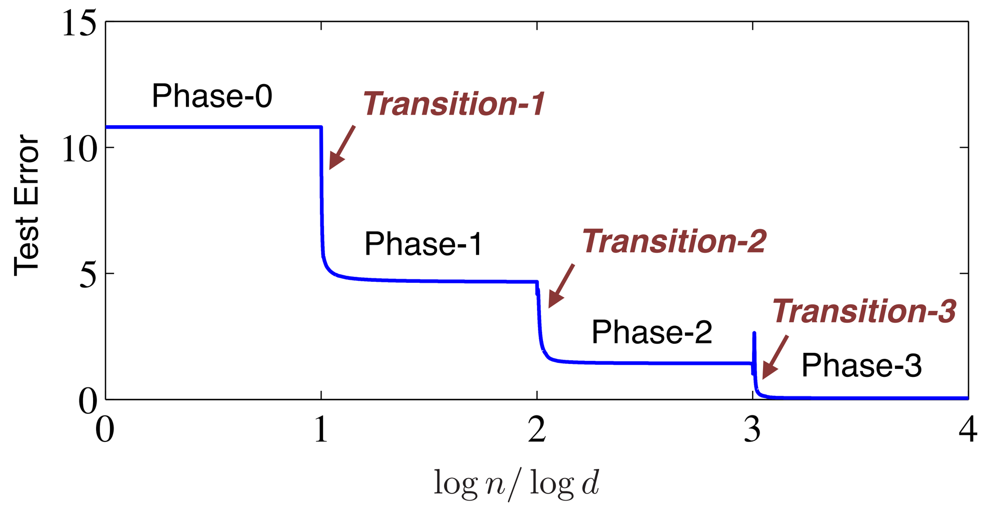
We present an *exact* analysis of learning performance of kernel ridge regression (KRR) in the *polynomial scaling regime*.

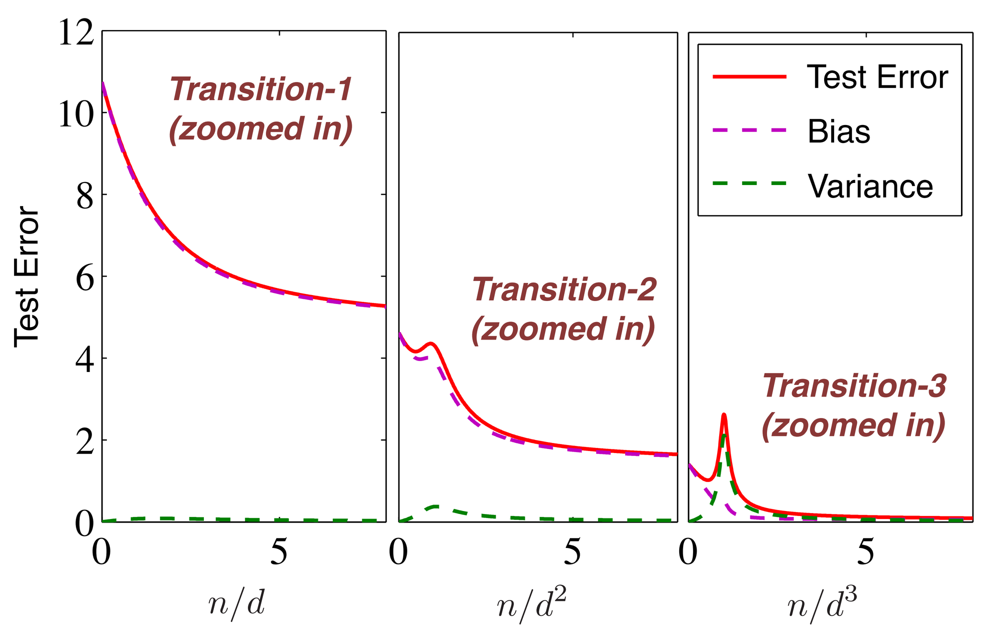
Our main contributions:

- Precise formula for the sample-wise KRR learning curves for dot-product kernel
- Characterizations of limiting empirical spectrum of the dot-product matrices
- An extension of the above results to the convolutional kernel

Multi-phase Learning Curve

Hierarchical Learning Process of KRR:





Top figure: Test error as a function of $\log n/\log d$, where n is the sample size and d is the input dimension.
The test error appears to remain unchanged when $d^{k-1} \ll n \ll d^k$, for $k \in \mathbb{Z}^+$, while transitions occur at $n \asymp d^k$.

Bottom figure: Zoomed-in views of the test error within each transition region, corresponding to $n \asymp d^k$ for $k = 1, 2, 3$.
The learning curves exhibit delicate non-monotonic behavior at different polynomial scaling regimes.

Challenge: Previous works analyzed test error when $d^{k-1} \ll n \ll d^k$, but the precise characterization of transitions among different phases is left unaddressed. The main challenge lies in the *non-linearity* of kernel function.

Kernel Ridge Regression (KRR)

Input: A collection of i.i.d. training samples:

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{P}.$$

Method: Learn a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ in a reproducing kernel Hilbert space (RKHS) by solving:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^n [y_i - f(\boldsymbol{x}_i)]^2 + \lambda \|f\|_K^2. \tag{1}$$

where $\|\cdot\|_K$ is the RKHS norm associated with kernel function $K(\cdot,\cdot)$.

Test Error: The performance of KRR can be captured by the test error:

$$\mathrm{Err} = \mathbb{E}_{\text{new}}[\mathbb{E}(y_{\text{new}} \mid \boldsymbol{x}_{\text{new}}) - \hat{f}(\boldsymbol{x}_{\text{new}})]^2,$$

where $(\boldsymbol{x}_{\text{new}}, y_{\text{new}}) \sim \mathcal{P}$ is an independent test sample.

A High-dimensional Model

Polynomial Scaling Regime: We consider the setting when $d \to \infty$, while for some $r \in \mathbb{Z}^+$,

$$\frac{N(d,r)}{n} \to \alpha_r \in (0, \infty).$$

Data: We consider

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} \tau_{d-1} \quad \text{and} \quad y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

where τ_{d-1} is the uniform distribution over d-dimensional unit sphere $\mathcal{S}^{d-1}$ and $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$.

Label function: The label function f has the following spectral decomposition:

$$f(\boldsymbol{x}) = \sum_{k \geq 0} \sum_{j=1}^{N(d,k)} \mu_{kj} Y_{kj}(\boldsymbol{x}) \tag{2}$$

where the eigenfunction $Y_{kj}(\boldsymbol{x})$ is the jth order-k spherical harmonics, μ_{kj} are the eigenvalues and $N(d,k)$ is the total number of order-k spherical harmonics.
For $k < r$, $\{\mu_{kj}\}$ are fixed and for $k \geq r$, $\{\mu_{kj}\}$ are random satisfying

$$\mathbb{E}(\mu_{kj}\mu_{k'j'}) = \frac{\hat{f}_k^2 \mathbb{1}_{kk'} \mathbb{1}_{jj'}}{N(d,k)}.$$

Kernel: We consider dot-product kernel on $\mathcal{S}^{d-1}$:

$$K(\boldsymbol{x}, \boldsymbol{x}') = h(\boldsymbol{x}^\top \boldsymbol{x}') \tag{3}$$

and $h(t)$ has the following spectral decomposition:

$$h(t) = \sum_{k=0}^{\infty} \hat{h}_k P_k(t)$$

where $P_k(t)$ is the order-k Legendre polynomials.

Precise Formula

Define

$$\chi_B(\alpha) = \int (1 + \xi t)^{-2} \mu_\alpha(t) dt$$

and

$$\chi_V(\alpha) = \alpha \xi^2 \int t(1 + \xi t)^{-2} \mu_\alpha(t) dt$$

where μ_α is the PDF of Marchenko–Pastur distribution with ratio α and ξ is defined as:

$$\xi := \frac{\hat{h}_r^2}{\alpha(\lambda + \hat{h}_{>r}^2)}$$

Asymptotic bias and variance: Define the asymptotic bias and variance associated with the rth order component as:

$$B_r(\alpha_r) = \chi_B(\alpha_r) \hat{f}_r^2 + \hat{f}_{>r}^2$$

and

$$V_r(\alpha_r) = \chi_V(\alpha_r)(\hat{f}_{>r}^2 + \sigma_\epsilon^2)$$

Theorem 1. *Under the main assumptions of the paper, the test error satisfies:*

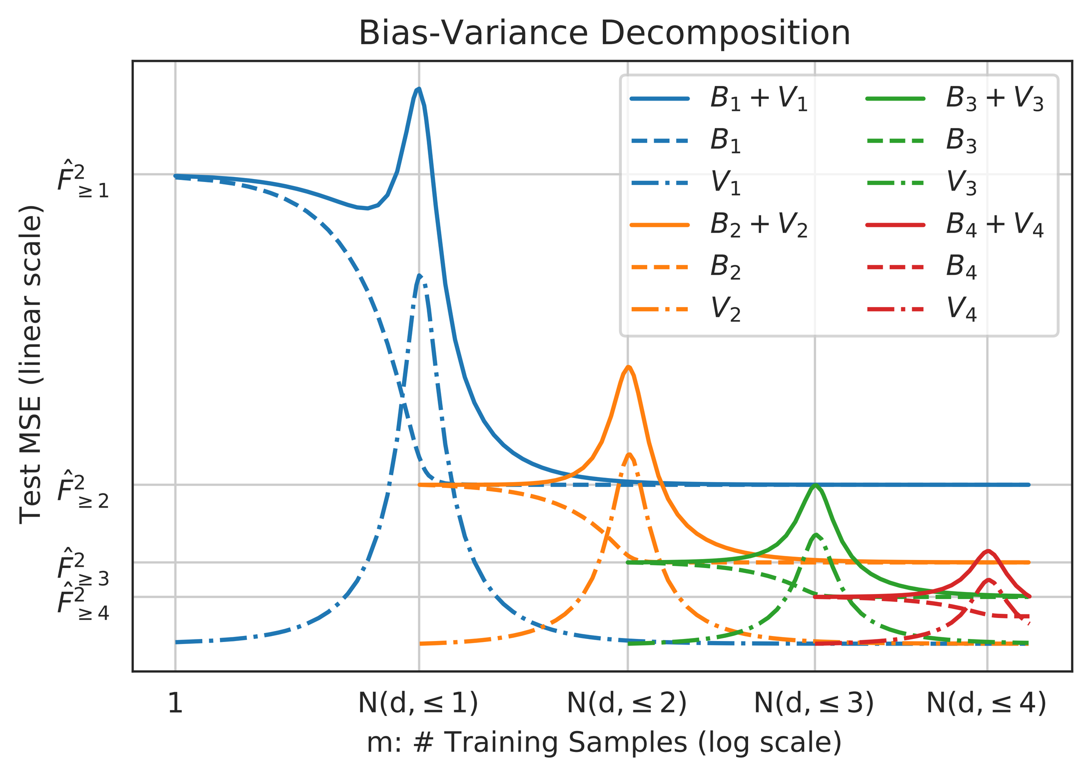
$$Err \xrightarrow{\mathbb{P}} B_r(\alpha_r) + V_r(\alpha_r). \tag{4}$$

Bias-variance Trade-off

Bias: KRR *perfectly* learns all low-frequency ($k < r$) modes; *partially* learns the critical frequency ($k = r$) modes; while *does not* learn any high-frequency mode ($k > r$).

Variance: All the high-frequency modes play the roles as the *additive noise*, while all the low-frequency modes do not contribute to the variance.

Non-monotonicity of learning curves: The bias is monotonically *decreasing*, while the variance exhibits *multiple descents*.



Spectrum of Kernel Matrix

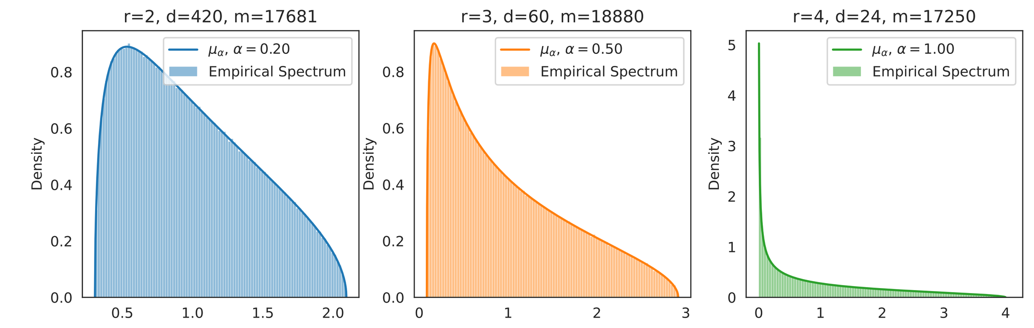
The key of proof of Theorem 1 is computing the limiting Stieltjes transform of empirical spectral distribution of the rth component of the kernel matrix:

$$R(\gamma) = \frac{1}{n} \mathrm{Tr}(\gamma \boldsymbol{I} + \boldsymbol{K}_r)^{-1}$$

where $\gamma > 0$ and $\boldsymbol{K}_r = \frac{1}{N(d,r)} \boldsymbol{Y}_r(\boldsymbol{X}) \boldsymbol{Y}_r(\boldsymbol{X}^\top)$.

Theorem 2. *Under the main assumption, the empirical spectrum of $\boldsymbol{K}_r$ converges in distribution to the Marchenko-Pastur distribution with ratio $\alpha = \alpha_r$.*

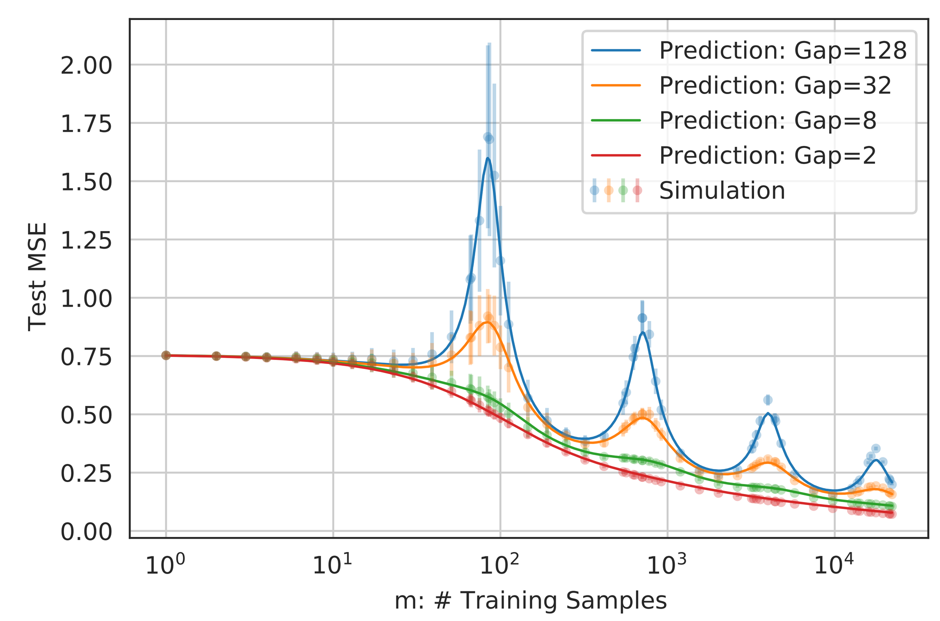
Gaussian Equivalence: The limiting spectrum of $\boldsymbol{K}_r$ remains unchanged, if $\boldsymbol{Y}_r(\boldsymbol{X})$ is replaced with an i.i.d Gaussian matrix $\boldsymbol{G}$ of same size.



Convolutional Kernel

- Similar results can be obtained for convolutional kernel.
- The eigenstructure of convolutional kernel matrix not only depends on the order of eigenfunctions but also on the topologies of neural network.

One-layer CNN Kernel:



T.M. was supported by NSF through award DMS-2031883 and the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning. T.M. also acknowledge the NSF grant CCF-2006489 and the ONR grant N00014-18-1-2729. The work of Yue M. Lu is supported by a Harvard FAS Dean's competitive fund award for promising scholarship, and by the US National Science Foundation under grant CCF-1910410.