High-Dimensional Kernel Methods under Covariate Shift: Data-Dependent Implicit Regularization

Yihang Chen[1] Fanghui Liu[2] Taiji Suzuki[3] Volkan Cevher[1]

[1]EPFL [2]University of Warwick [3]The University of Tokyo

Overview

In this paper, we provide an initial analysis to the following question:

How does IW affect bias-variance trade-off in high-capacity models?

To summarize our contributions:

- We present the asymptotic expansion of high-dimensional kernels $K(\boldsymbol{x},\boldsymbol{x}')$ under covariate shifts, where the nonlinearity in kernels can be eliminated by the kernel function curvature.
- For variance, we demonstrate that the IW strategy can be regarded as an implicit data-dependent regularization on the respective kernel.
- For bias, we demonstrate two cases: 1) near interpolation, and 2) some proper regularization parameter.

Kernel: Asymptotic Expansion

Lemma 1. *Assume the kernel K is the inner-product kernel, $K(\boldsymbol{x},\boldsymbol{x}') := h(\langle \boldsymbol{x},\boldsymbol{x}'\rangle/d)$, or the radial kernel, $K(\boldsymbol{x},\boldsymbol{x}') := h(-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2/d)$, and the training data $\boldsymbol{X}\sim p$.*
(1) Under suitable assumptions, we have $\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})-\boldsymbol{K}^{\text{lin}}(\boldsymbol{X},\boldsymbol{X})\|_2 \to 0$, as $n,d\to\infty, n/d\to\zeta$, where $\boldsymbol{K}^{\text{lin}}(\boldsymbol{X},\boldsymbol{X})$ is defined by $\boldsymbol{K}^{\text{lin}}(\boldsymbol{X},\boldsymbol{X}) := \alpha_p \mathbb{1}\mathbb{1}^\top + \beta_p \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \gamma_p \boldsymbol{I} + \boldsymbol{T}_p$, with non-negative parameters α_p, β_p, γ_p, and the additional matrix $\boldsymbol{T}_p$ given in Table 1.
(2) Under suitable assumptions, with $c_{pq} < 2\theta_q - 1/2$, with the training data $\boldsymbol{X}\sim p$ and a test data $\boldsymbol{x}\sim q$, we have $\mathbb{E}_q\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})-\boldsymbol{K}^{\text{lin}}(\boldsymbol{X},\boldsymbol{x})\|_2\to 0$, as $n,d\to\infty, n/d\to\zeta$, where $\boldsymbol{K}^{\text{lin}}(\boldsymbol{x},\boldsymbol{X})$ is defined by $\boldsymbol{K}^{\text{lin}}(\boldsymbol{X},\boldsymbol{x}) := \beta_{pq}\frac{\boldsymbol{X}\boldsymbol{x}}{d} + \boldsymbol{T}_{pq}(\boldsymbol{X},\boldsymbol{x})$, with non-negative parameters β_{pq}, and the additional vector $\boldsymbol{T}_{pq}$ given in Table 1.

Table 1: Parameters of the linearized kernel $\boldsymbol{K}^{\text{lin}}$ involved with the curvature of h, when $\boldsymbol{X}\sim p$.

Parameters	Inner-Product Kernels	Radial Kernels
α_p	$h(0)+h''(0)\frac{\text{Tr}(\boldsymbol{\Sigma}_p^2)}{2d^2}$	$h(-2\tau_p)+2h''(-2\tau_p)\frac{\text{Tr}(\boldsymbol{\Sigma}_p^2)}{d^2}$
β_p	$h'(0)$	$2h'(-2\tau_p)$
γ_p	$h(\tau_p)-h(0)-\tau_p h'(0)$	$h(0)-2\tau_p h'(-2\tau_p)-h(-2\tau_p)$
$\boldsymbol{T}_p$	$\boldsymbol{0}_{n\times n}$	$-h'(-2\tau_p)\boldsymbol{A}+\frac{1}{2}h''(-2\tau_p)\boldsymbol{A}\odot\boldsymbol{A}$ [1]
β_{pq}	$h'(0)$	$2h'(-(\tau_p+\tau_q))$
$\boldsymbol{T}_{pq}$	$\boldsymbol{0}_{n\times 1}$	$-h(-(\tau_p+\tau_q))\cdot\mathbb{1}-\frac{\beta_{pq}}{2}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})$ [2]

[1] $\boldsymbol{A} := \mathbb{1}\boldsymbol{\psi}^\top + \boldsymbol{\psi}\mathbb{1}^\top$, where $\boldsymbol{\psi}\in\mathbb{R}^n$ with $\psi_i := \|\boldsymbol{x}_i\|_2^2/d-\tau_p$.
[2] $\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x}) := \psi_{\boldsymbol{x}}+\boldsymbol{\psi}$, where $\psi_{\boldsymbol{x}} = \|\boldsymbol{x}\|_2^2/d-\tau_q$.

Problem Setting

Notations:

- **Data**.
  - Training distribution: p. Test distribution: q.
  - Re-weighting distribution $\overline{q}$. Re-weighting function $\overline{w}(\boldsymbol{x}) = \mathrm{d}\overline{q}(\boldsymbol{x})/\mathrm{d}p(\boldsymbol{x})$.
  - The label y is generated by f_ρ, $y(\boldsymbol{x}) = f_\rho(\boldsymbol{x})+\varepsilon$, and $\mathbb{E}[\varepsilon]=0$, $\mathbb{V}[\varepsilon]\le\sigma_\varepsilon^2$.
- **Kernel**. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space $\mathcal{H}$ endowed with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ of functions $f:\mathcal{X}\to\mathbb{R}$ with a reproducing kernel $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ where $K(\cdot)\in\mathcal{H}$ and $f(\boldsymbol{x}) = \langle f, K(\boldsymbol{x},\cdot)\rangle_{\mathcal{H}}$. Specifically, we consider the inner-product kernels, $K(\boldsymbol{x},\boldsymbol{x}') := h(\langle\boldsymbol{x},\boldsymbol{x}'\rangle/d)$.
- **Task**: Given n training data $\boldsymbol{Z} = \{(\boldsymbol{x}_i,y_i)\sim p\}_{i=1}^n$, the estimator of KRR in high dimensions under a general IW function $\overline{w}(\boldsymbol{x})$ is given by $\overline{f}_{\lambda,\boldsymbol{Z}} := \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n \overline{w}(\boldsymbol{x}_i)\left(f(\boldsymbol{x}_i)-y_i\right)^2+\lambda\|f\|_{\mathcal{H}}^2\right\}$, where $\lambda>0$ is the regularization parameter.

Assumptions (abbreviated):

Data. Let $\boldsymbol{\Sigma}_p, \Sigma_q$ be the covariance matrix of the distribution p,q.

- (**Sub-Gaussian**) Let $\boldsymbol{x}\sim p$, then $\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x}$ is sub-Gaussian with identity variance and bounded items, and similarly for q.
- (**Covariance**) We assume $\max\{\|\boldsymbol{\Sigma}_p\|,\|\boldsymbol{\Sigma}_q\|\} = O(1)$. Define $\boldsymbol{\Sigma}_{pq} := \boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q$, and $\exists c_{pq}\ge 0$ so that $\text{Tr}(\boldsymbol{\Sigma}_{pq})/d \lesssim d^{c_{pq}}$. To limit the distribution shifts, we additionally assume $c_{pq} < 2\theta_q - \frac{1}{2} = \frac{1}{2}-\frac{4}{8+m_q}$.
- (**Ratio**) The ratio $w := \mathrm{d}q/\mathrm{d}p, \overline{w} := \mathrm{d}\overline{q}/\mathrm{d}p$'s norm in some space is upper bounded by constants dependent on the dimension d.

Model:

- (**Source condition**): We have $f_\rho\in\mathcal{H}$, and there exists $\frac{1}{2}\le\overline{r}<1, \overline{g}_\rho\in\mathcal{L}_q^2$ such that $f_\rho = (L_{\overline{q}})^{\overline{r}}\overline{g}_\rho$. We additionally assume $\max\{\|f_\rho\|_{\mathcal{H}},\|\overline{g}_\rho\|_q,\|f_\rho\|_\infty\}\lesssim d^{c_{\mathcal{H}}}$.
- (**Capacity condition**): For any $\lambda>0$, there exists $E_\mu>0$ and $s_\mu\in[0,1]$ such that for distribution $\mu\in\{q,\overline{q}\}$, $\mathcal{N}_\mu(\lambda) := \text{Tr}((L_\mu+\lambda)^{-1}L_\mu)\le E_\mu^2\lambda^{-s_\mu}, \forall\lambda\in(0,1]$.

Main Results

Bias-variance decomposition: We have the following bias-variance decomposition:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}}-f_\rho\|_q^2 = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}}-\overline{f}_{\lambda,\boldsymbol{X}}\|_q^2 + \|\overline{f}_{\lambda,\boldsymbol{X}}-f_\rho\|_q^2 := \mathsf{V}+\mathsf{B}^2\,.$$

where $\overline{f}_{\lambda,\boldsymbol{X}} := \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n\overline{w}(\boldsymbol{x}_i)\left(f(\boldsymbol{x}_i)-f_\rho(\boldsymbol{x}_i)\right)^2+\lambda\|f\|_{\mathcal{H}}^2\right\}$.

Variance:

Theorem 2 (Data-dependent regularization). *Let $\delta\in(0,1)$, then for large d, with probability at least $1-\delta-2d^{-2}$ with respect to a draw of $\boldsymbol{X}\sim p$ and $\epsilon>0$, the variance can be estimated by*

$$\mathsf{V}\le\frac{8\sigma_\varepsilon^2\|\boldsymbol{\Sigma}_q\|}{d}\underbrace{\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}+\frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1};\frac{\gamma_p}{\beta_p}\right)}_{\textit{dominated term } \mathsf{V}_{\boldsymbol{x}}}+\frac{8\sigma_\varepsilon^2}{\gamma_p^2}d^{-(4\theta_q-1-2c_{pq})}\log^{4(1+\epsilon)}d\,.$$

The dominated term in Theorem 2 can be represented as

$$\mathsf{V}_{\boldsymbol{x}}\asymp\frac{1}{d}\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}+\frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1};\frac{\gamma_p}{\beta_p}\right)\,,$$

which implies that the variance is well controlled by the capacity of $\boldsymbol{K}^{\text{lin}}+\lambda n\overline{\boldsymbol{W}}^{-1}$.

Bias:

Theorem 3 (Bias under arbitrary λ). *Let $\delta\in(0,1)$, we have the bias B is upper bounded as $\mathsf{B}\le\mathsf{B}_{\text{in}}+\mathsf{B}_{\text{iw}}$, where B_{in} is the* intrinsic bias *that only depends on the problem of covariate shift from p to q via the ratio $w(\boldsymbol{x})$, and $\mathsf{B}_{\text{in}} := \text{Tr}\left(\boldsymbol{K}^{\text{lin}}\boldsymbol{W}\right)/n$. The second term is the re-weighting bias B_{iw} that depends on the choice of $\overline{w}(\boldsymbol{x})$, $w(\boldsymbol{x})$, and λ. When $w=\overline{w}$, we have $\mathsf{B}_{\text{iw}} := \lambda^2 n\mathcal{N}\left(\boldsymbol{K}^{\text{lin}}\boldsymbol{W}, n\lambda\right)+o(1)$, with probability at least $1-4\delta$ for sufficiently large d.*

Theorem 4 (Bias under some λ). *Under some assumptions on the data and model, with proper selection of c_λ and C_λ, when choosing $\lambda := \cdot C_\lambda n^{-c_\lambda}$, then with probability at least $1-\delta$, for sufficiently large d, when $c_{\mathcal{H}}<\overline{r}c_\lambda$, it holds that*

$$\mathsf{B}\lesssim n^{-\overline{r}c_\lambda+c_{\mathcal{H}}}\|L_q(L_{\overline{q}}+\lambda)^{-1}\|^{1/2}\,.$$

For general λ, we have, with $\lesssim$ here hiding the dependence on n,

$$\mathsf{B}\lesssim(\lambda^{\overline{r}}+\lambda^{-\frac{1}{2}})\|L_q(L_{\overline{q}}+\lambda)^{-1}\|^{\frac{1}{2}}\,.$$

Related works:

1. Liang, T., Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize.
2. Liu, F., Liao, Z., Suykens, J. (2021, March). Kernel regression in high dimensions: Refined analysis beyond double descent. In International Conference on Artificial Intelligence and Statistics (pp. 649-657). PMLR.