On the Target-kernel Alignment: a Unified Analysis with Kernel Complexity

Presented by Chao Wang

Joint work with Xin He, Yuwen Wang, and Junhui Wang

November 10, 2024

上海财经大学
Shanghai University of Finance & Economics

Introduction

On what THE LEARNING rate DEPENDS

Goal: This work investigates the impact of alignment between the target function of interest and the model on the performance of the kernel method.

Intuitively, the learning rate of any learning algorithm improves if

- the model complexity of the hypothesis space becomes lower or
- the target-model alignment [1] becomes stronger.

[1] measure of similarity between the hypothesis space and the target function.

Problem setup

- Let $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$ be a collection of covariate-response pairs, where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathscr{X} \subset \mathbb{R}^p$ and $Y_1, \ldots, Y_n \in \mathscr{Y} \subset \mathbb{R}$ are independent with $Y_i \sim \mathbb{P}_{Y|\mathbf{x}_i}$ for each $i \in [n]$. (Fixed design setting is considered.)

- Given a Lipschitz loss function $L(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, the population risk function is defined as

$$\mathscr{E}(f) := \mathbb{E}_{Y^n}\left[\frac{1}{n}\sum_{i=1}^n L\big(Y_i, f(\mathbf{x}_i)\big)\right].$$

- In the field of learning theory, the target function of interest is defined as the minimizer of the population risk

$$f^* := \operatorname{argmin}_f \mathscr{E}(f).$$

Standard kernel Method

Let H_K be the reproducing kernel Hilbert space (RKHS) induced by a positive semi-definite kernel function K and $\|\cdot\|_K$ denote the endowed norm in H_K. We assume $f^* \in H_K$ in this work.

To estimate the underlying target function f^*, we solve the following empirical risk function plus a penalty term that

$$\widehat{f}_\lambda = \operatorname*{argmin}_{f \in H_K} \left\{ \widehat{\mathscr{E}}(f) + \lambda \|f\|_K^2 \right\}.$$

Here, $\widehat{\mathscr{E}}(f)$ is empirical risk function.

Kernel Matrix

- Let $\mathbf{K} = \{\frac{1}{n} K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ be the empirical kernel matrix.

- Eigen-decomposition: $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, where $\mathbf{D} \in \mathbb{R}^{n \times n}$ has diagonal elements $\mu_1, \ldots, \mu_n > 0$ arranging in a descending ordering.

- Polynomial decay case: for $\alpha > 1$, $\mu_j \asymp j^{-\alpha}$. (A decreasing α results in an increasing compacity of the RKHS H_K.)

Target-kernel Alignment

- Let $\xi^* = \mathbf{U}^\top S_{\mathbf{x}}(f^*)$ with $S_{\mathbf{x}}$ denoting the sample operator, defined as

$$S_{\mathbf{x}}(f) := \frac{1}{\sqrt{n}}(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top \quad \text{for } f \in H_K.$$

- There exist some constants $\gamma \geq \frac{1}{2}$ and $u \geq 2$ such that $\sum_{j=1}^{n} {\xi_j^*}^2 {\mu_j}^{-2\gamma} \leq u^2$ for any n. (A greater value of γ implies a stronger target-kernel alignment.)

- Polynomial decay case: for $\alpha > 1$ and $\gamma \geq \frac{1}{2}$, ${\xi_j^*}^2 \asymp j^{-2\gamma\alpha-1}$.

SATURATION EFFECT

Existing result for kernel ridge regression (KRR) [2]:

$$\text{Learning rate } : n^{-\frac{2\eta\alpha}{2\eta\alpha+1}} \quad \text{with} \quad \eta = \min\{\gamma, 1\}.$$

Contradict when γ exceeds 1! This phenomenon is known as the saturation effect[3].

[2] Caponnetto, A., & De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics, 7, 331–368.

[3] Bauer, F., Pereverzev, S., & Rosasco, L. (2007). On regularization algorithms in learning theory. Journal of Complexity, 23, 52–72.

What's new

In this work, we are devoted to

1. offering a comprehensive understanding of the impact of target-kernel alignment on the performance of kernel method from the kernel complexity perspective;
2. providing a theoretically guaranteed solution to eliminate the saturation effect;
3. establishing the minimax lower bound for all $\gamma \geq \frac{1}{2}$.

Reduced RKHS

From Amini et al. (2022), [4] let $\{\psi_k\}_{k\in[n]} \subset H_K$ be defined such that

$$\psi_k := \operatorname{argmin}\{\|\psi\|_K : \psi \in H_K, S_{\mathbf{x}}(\psi) = \mathbf{u}_k\},$$

where $\mathbf{u}_k$ is the k-column of $\mathbf{U}$. For a given r, define the reduced function space

$$H_{K_r} := \left\{ \sum_{k=1}^{r} \alpha_k \psi_k : \alpha = (\alpha_1, \ldots, \alpha_r)^\top \in \mathbb{R}^r \right\},$$

which is an r-dimensional reduced RKHS associated with kernel

$$K_r(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{r} \mu_k \psi_k(\mathbf{x}) \psi_k(\mathbf{x}').$$

[4] Amini, A., Baumgartner, R., & Feng, D. (2022). Target alignment in truncated kernel ridge regression. NeurIPS, 2024.

Truncated kernel method

The reduced RKHS H_{K_r} can be treated as a smaller approximation of the full RKHS H_K. Based on H_{K_r}, a truncated estimator can be obtained by solving

$$\widehat{f}_{\lambda,r} = \operatorname*{argmin}_{f \in H_{K_r}} \left\{ \widehat{\mathscr{E}}(f) + \lambda \|f\|_{K_r}^2 \right\},$$

where $\|\cdot\|_{K_r}$ denotes the endowed norm in H_{K_r}.

- Let $\mathbf{K}_r = \{\frac{1}{n} K_r(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ be the empirical kernel matrix w.r.t. K_r.
  - $\mathbf{K}_r = \mathbf{U} \mathbf{D}_r \mathbf{U}^\top$, where $\mathbf{D}_r$ is diagonal matrix with elements $\mu_1, ..., \mu_r, 0, ..., 0$.

Theoretical Results

Kernel complexity

A measure of model complexity for RKHS: kernel complexity function, defined as

$$R(\delta) := \Big(\frac{1}{n}\sum_{j=1}^{n} \min\{\delta^2, \mu_j\}\Big)^{1/2}.$$

Fix any $\iota \in (0,1)$. A crucial quantity appearing in the error bounds is the critical radius δ_n, defined as the smallest positive value δ satisfying

$$C \log \iota^{-1} R(\delta) \le \delta^{2\eta+1} \quad \text{with} \quad \eta = \min\{\gamma, 1\}.$$

Here and throughout, c, C are some universal constants with varying values line by line.

Remark

The existence and uniqueness of δ_n can be verified.

Result for standard kernel method

Theorem

Fix any $\iota \in (0,1)$. Let $\eta = \min\{\gamma, 1\}$. Then under certain conditions (specified in our paper), with probability at least $1-\iota$, one has

$$\max\left\{ \|\widehat{f}_\lambda - f^*\|_n^2,\ \mathscr{E}(\widehat{f}_\lambda) - \mathscr{E}(f^*) \right\} \leq C\left(\delta_n^{4\eta} + \lambda^{2\eta}\right).$$

For the polynomial decay case, with λ properly chosen, one has a simpler bound that

$$\mathscr{E}(\widehat{f}_\lambda) - \mathscr{E}(f^*) \asymp \|\widehat{f}_\lambda - f^*\|_n^2 \leq C\left(\frac{(\log \iota^{-1})^2}{n}\right)^{\frac{2\eta\alpha}{2\eta\alpha+1}}.$$

Kernel Complexity of the reduced RKHS

- Kernel complexity of the reduced RKHS:

$$R_r(\delta) := \left(\frac{1}{n} \sum_{j=1}^{r} \min\{\delta^2, \mu_j\} \right)^{1/2}.$$

- Fix any $\iota \in (0,1)$. The critical radius $\delta_{n,r}$ is defined as the smallest positive value δ satisfying

$$C \log \iota^{-1} R_r(\delta) \leq \delta^{2\eta+1}.$$

- $R_r(\delta) \leq R(\delta)$ implies $\delta_{n,r} \leq \delta_n$.
- Learning rate $\asymp$ Lower complexity $+$ approximation error .

Result for truncated standard kernel method

Theorem

Fix any $\iota \in (0,1)$. Let $\eta = \min\{\gamma, 1\}$. Then under certain conditions (specified in our paper), with probability at least $1-\iota$, one has

$$\max\left\{\left\|\widehat{f}_{\lambda,r} - f^*\right\|_n^2,\ \mathscr{E}(\widehat{f}_{\lambda,r}) - \mathscr{E}(f^*)\right\} \le C\Bigg(\underbrace{\delta_{n,r}^{4\eta} + \lambda^{2\eta}}_{\textit{Estimation error}} + \underbrace{\sum_{j=r+1}^{n} \xi_j^{*2}}_{\textit{Approximation bias}} \Bigg).$$

For the polynomial decay case, with λ and r properly chosen, one has a simpler bound that

$$\mathscr{E}(\widehat{f}_{\lambda,r}) - \mathscr{E}(f^*) \asymp \left\|\widehat{f}_{\lambda,r} - f^*\right\|_n^2 \le C\left(\frac{(\log \iota^{-1})^2}{n}\right)^{\frac{2\gamma\alpha}{2\gamma\alpha+1}}.$$

Algorithm-free lower bounds

Theorem

Let f^* defined with squared loss specified, satisfying $\sum_{j=1}^{n} {\xi_j^*}^2 {\mu_j}^{-2\gamma} \leq u^2$. Suppose that the RKHS is induced by the regular kernel, and $\widetilde{f}$ is any estimator based on the data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. For $\frac{1}{2} \leq \gamma \leq 1$, one has

$$\inf_{\widetilde{f}} \sup_{f^* \in H_K} \mathbb{P}\left(\|\widetilde{f} - f^*\|_n^2 \geq c\delta_n^{4\gamma} \right) \geq \frac{1}{2}.$$

For $\gamma > 1$, with r properly chosen, one has

$$\inf_{\widetilde{f}} \sup_{f^* \in H_K} \mathbb{P}\left(\|\widetilde{f} - f^*\|_n^2 \geq c\delta_{n,r}^{4} \right) \geq \frac{1}{2}.$$

Illustration of the established results

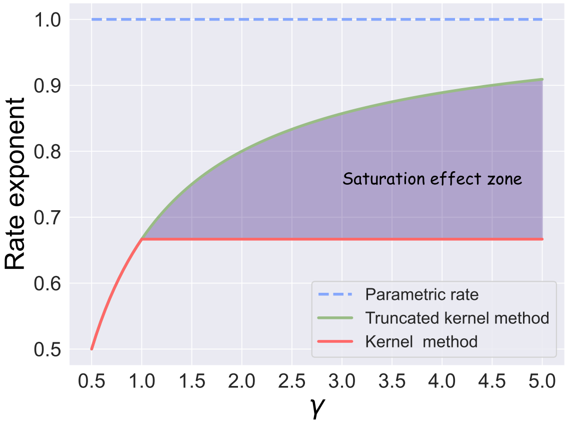

Figure: The exponent rate θ of the learning rate $n^{-\theta}$ versus the alignment level γ for different methods.

Theoretical suggestions

1. The learning rate for the truncated method can be consistently improved as γ increases, eliminating the phenomena of saturation effect;

2. An optimal trade-off between the model complexity and approximation bias can be attained by the truncated kernel method with r properly chosen;

3. The truncated kernel method has a stronger ability to capture the alignment so that a faster rate compared to the standard kernel method is achieved; (price to pay: an additional truncated parameter r to tune.)

4. The truncated kernel method can be treated as optimal tackling whenever the alignment level is.

Numerical Results

Numerical studies

Sub-goals in this part:

- verify the improvement of the truncated kernel method over the standard kernel method.
- verify a conjecture: lower complexity of the RKHS may result in a potential mismatch between the model space and the target, consequently weakening the target-kernel alignment which undermines the learning efficiency.

Numerical studies

Recall:

$$\text{Learning rate of the truncated estimator} \asymp n^{-\frac{2\gamma\alpha}{2\gamma\alpha+1}}.$$

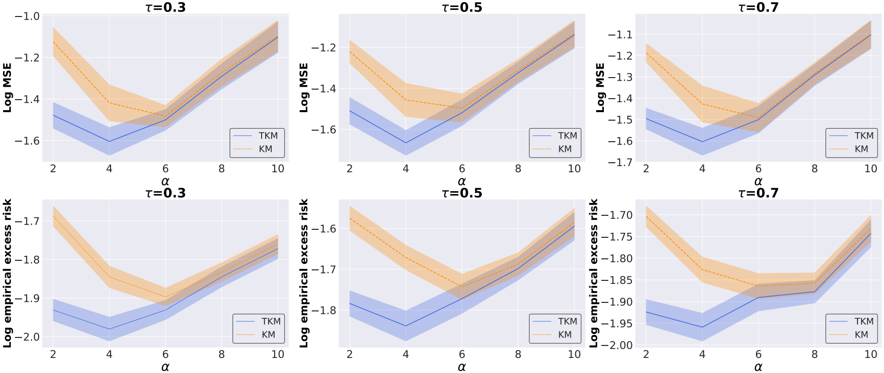

Figure: Quantile regression; averaged log MSE and log empirical excess risk for (kernel method) KM and (Truncated kernel method) TKM versus α for different quantile level τ.

Future directions

There are several directions for future research, two of which are mentioned here to conclude.

1. general spectral kernel method;
2. from fixed design setting to random design setting. (non-trivial!)

The End

Questions? Comments?