Target alignment in truncated kernel ridge regression

Arash Amini,
Richard Baumgartner
Dai Feng

UCLA
Merck & Co., Inc.
AbbVie Inc.

Kernel ridge regression (KRR)

- Kernel ridge regression (KRR) has recently attracted a lot interest.
- Connections to neural networks via the neural tangent kernel (NTK).
- Potential for explaining transient effects, double descent, etc.
- Connections to minimum-norm interpolating solutions.

- This paper: Target alignment and spectral truncation in KRR.
- High level messages:

1. More alignment $\implies$ lower the generalization error (if taken advantage of).
2. Truncated KRR better takes advantage of the alignment compared to KRR.
3. Multiple descent phenomena can happen in multi-band "alignment spectra".
4. There is an Over-aligned regime that TKRR beats usual KRR.

Setup

- Consider the usual setup of nonparametric regression:

$$y_i = f^*(x_i) + w_i,\ i = 1, \ldots, n \tag{1}$$

- A natural estimator is the kernel ridge regression (KRR):

$$\widehat{f}_{n,\lambda} := \underset{f \in \mathbb{H}}{\operatorname{argmin}}\ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathbb{H}}^2, \tag{2}$$

- By the representer theorem (Kimeldorf and Wahba 1971), the problem reduces to

$$\widehat{\omega} = \underset{\omega \in \mathbb{R}^n}{\operatorname{argmin}}\ \frac{1}{n} \|y - \sqrt{n} K \omega\|^2 + \lambda \omega^T K \omega, \tag{3}$$

involving the kernel matrix:

$$K = \frac{1}{n} \big(\mathbb{K}(x_i, x_j) \big) \in \mathbb{R}^{n \times n}$$

Truncated KRR (TKRR)

- The kernel matrix K is a dense $n \times n$ matrix.
- Often K is approximated by Nyström, sketching, etc.
- The simplest approximation is spectral (or rank) truncation:

$$K = \sum_{k=1}^{n} \mu_k u_k u_k^T \implies \widetilde{K} = \sum_{k=1}^{r} \mu_k u_k u_k^T$$

- First result, TKRR is an exact KRR in a smaller $\widetilde{\mathbb{H}} \subset \mathbb{H}$.

- The target alignment (TA) spectrum of f^*:

$$\xi_k^* = \frac{1}{\sqrt{n}} u_k^T \begin{pmatrix} f^*(x_1) \\ f^*(x_2) \\ \cdots \\ f^*(x_n) \end{pmatrix}, \quad k = 1, \ldots, n$$

Theorem 1 (Exact MSE)

For any TKRR solution $\widetilde{f}_{r,\lambda}$, we have

$$\mathbb{E}\|\widetilde{f}_{r,\lambda} - f^*\|_n^2 = \sum_{i=1}^{r} \frac{\lambda^2}{(\mu_i+\lambda)^2}(\xi_i^*)^2 + \sum_{i=r+1}^{n}(\xi_i^*)^2 + \frac{\sigma^2}{n}\sum_{i=1}^{r}\frac{\mu_i^2}{(\mu_i+\lambda)^2} \tag{4}$$

$$= \|f^*\|_n^2 + \sum_{i=1}^{r}\frac{1}{(\mu_i+\lambda)^2}\left[-a_i(\lambda)(\xi_i^*)^2 + \frac{\sigma^2}{n}\mu_i^2\right] \tag{5}$$

where $a_i(\lambda) = (\mu_i+\lambda)^2 - \lambda^2$ and the expectation is w.r.t. the randomness in the noise vector w.

Proposition 1 (Bandlimited model, informal statement)

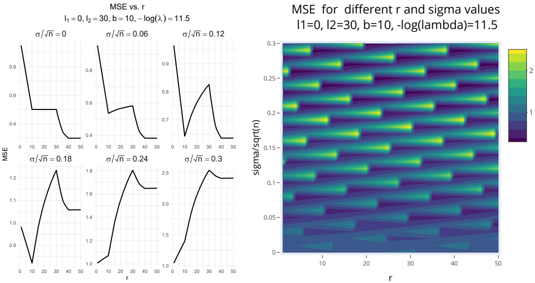
For a single-band alignment spectra supported on $[\ell, \ell + b]$:

(a) *There is j^* such that MSE as a function of r*

1. *increases in $[1, j^*)$,*
2. *decreases in $[j^*, \ell + b)$,*
3. *increases in $[\ell + b, n]$.*

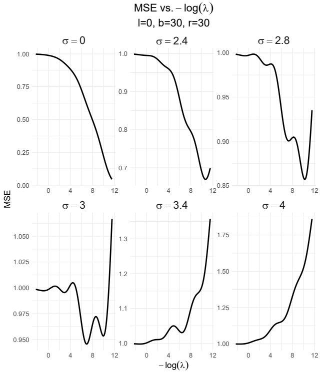
(b) *Alignment spectra that are concentrated near lower indices are better.*

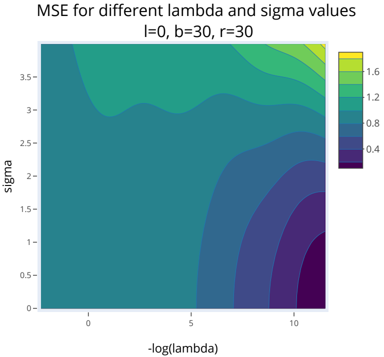
(c) *Concentrated alignment spectra are better than diffuse ones.*



Two non-overlapping bands of length b, starting at indices $\ell_1 + 1$ and $\ell_2 + 1$.

Simulations: MSE versus λ





Polynomial alignment

- The case of polynomially decaying kernel eigenvalues and TA scores:

$$\mu_i \asymp i^{-\alpha}, \quad (\xi_i^*)^2 \asymp i^{-2\gamma\alpha-1} \tag{6}$$

Theorem 1

Let $\eta = \min(r, \lambda^{-1/\alpha})$. *Under the polynomial decay assumption* (6),

$$\mathbb{E}\|\widetilde{f}_{r,\lambda} - f^*\|_n^2 \asymp \lambda^2 \max(1, \eta^{-2(\gamma-1)\alpha}) + r^{-2\gamma\alpha}1\{r < n\} + \frac{\sigma^2}{n}\eta. \tag{7}$$

(a) *Taking* $\lambda \asymp (\sigma^2/n)^{\gamma\alpha/(2\gamma\alpha+1)}$ *and* $r \asymp (n/\sigma^2)^{1/(2\gamma\alpha+1)}$, *TKRR achieves the following rate*

$$\mathbb{E}\|\widetilde{f}_{r,\lambda} - f^*\|_n^2 \asymp \left(\frac{\sigma^2}{n}\right)^{2\gamma\alpha/(2\gamma\alpha+1)} \quad \text{for } \gamma > 1. \tag{8}$$

(b) *Assume* $n^{-2\alpha} \lesssim \sigma^2 \lesssim n$, *and let* $\delta := \min(1, \gamma)$. *Then, the best rate achievable by the full KRR is obtained for regularization choice* $\lambda \asymp (\sigma^2/n)^{\alpha/(2\delta\alpha+1)}$ *and is*

$$\mathbb{E}\|\widetilde{f}_{r,\lambda} - f^*\|_n^2 \asymp \left(\frac{\sigma^2}{n}\right)^{2\delta\alpha/(2\delta\alpha+1)} \quad \text{for } \gamma > 0. \tag{9}$$

Summary of the theorem

- To summarize, let us define the rate exponent function,

$$s(\gamma) := 2\gamma\alpha/(2\gamma\alpha + 1). \tag{10}$$

- There are four regimes of target alignment, implied by Theorem 1:

(i) Under-aligned regime, $\gamma \in (0, \frac{1}{2})$: The target is not in the RKHS ...

(ii) Just-aligned regime, $\gamma = \frac{1}{2}$: Target in the RKHS, no extra alignment ...

(iii) Weakly-aligned regime, $\gamma \in (\frac{1}{2}, 1]$: ...

(iv) Over-aligned regime, $\gamma > 1$: Target in RKHS and strongly aligned with the kernel.
  - The best achievable rate is $(\sigma^2/n)^{s(\gamma)}$ which is achieved by TKRR:
  - The full KRR can only achieve the rate $(\sigma^2/n)^{s(1)}$, which is the best achievable in the weakly-aligned regime.

Rate exponent function $s(\gamma)$