Conditional KRR: Injecting Unpenalized Features into Kernel Methods

5-minute overview

Rustem Takhanov Zhenisbek Assylbekov

Nazarbayev University

Purdue University Fort Wayne

Classical KRR

Given data $(x_i, y_i)_{i=1}^N$ and a positive definite kernel K, standard KRR solves

$$\min_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_K}^2.$$

All components of f are regularized.

But sometimes we already know important features:

$$f_1, \ldots, f_k.$$

Conditionally positive definite kernels

A kernel K is conditionally positive definite w.r.t. $\mathcal{F}$ if

$$\sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \geq 0$$

whenever

$$\sum_{i=1}^{N} \alpha_i f(x_i) = 0 \qquad \text{for all } f \in \mathcal{F}.$$

Interpretation
The directions in $\mathcal{F}$ form a null space: they are not penalized.

Residual kernel

Let Π_P be the $L_2(\mathcal{X}, P)$-projection onto $\mathcal{F}$.

$$K_P = \big((I - \Pi_P) \otimes (I - \Pi_P)\big)[K].$$

Equivalently, K_P removes from K all components lying in $\mathcal{F}$.

Theorem
If K is conditionally positive definite w.r.t. $\mathcal{F}$, then the residual kernel

$$\begin{aligned} K_P(x, y) = K(x, y) - \Pi_P[K(x, \cdot)](x, y) - \\ - \Pi_P[K(\cdot, y)](x, y) + \Pi_P[\Pi_P[K(x, \cdot)](\cdot, y)](x, y). \end{aligned}$$

is positive definite.

Native space viewpoint

Conditional KRR works in the semi-Hilbert space where every function decomposes as

$$f = f_{\|} + f_{\perp}, \qquad f_{\|} \in \mathcal{F}, \quad f_{\perp} \in \mathcal{H}_{K_P}.$$

The norm only sees the residual part:

$$\|f\|^2_{\mathcal{H}^{\mathcal{F}}_K} = \|f_{\perp}\|^2_{\mathcal{H}_{K_P}}.$$

Actually, this is not a canonical definition, but we prove that they are equivalent.
Therefore, the definition of the native space does not depend on P.

$\mathcal{F}$ is free; only $\mathcal{F}^{\perp}$ is regularized.

Conditional KRR

The estimator is defined by

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K^{\mathcal{F}}} \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K^{\mathcal{F}}}.$$

This looks like KRR, but with one essential difference:

features in $\mathcal{F}$ are unpenalized.

Equivalent two-stage procedure

Let

$$F = [f_a(x_i)]_{a=1,\dots,k}^{i=1,\dots,N}.$$

Project the labels onto the orthogonal complement of the row space of F:

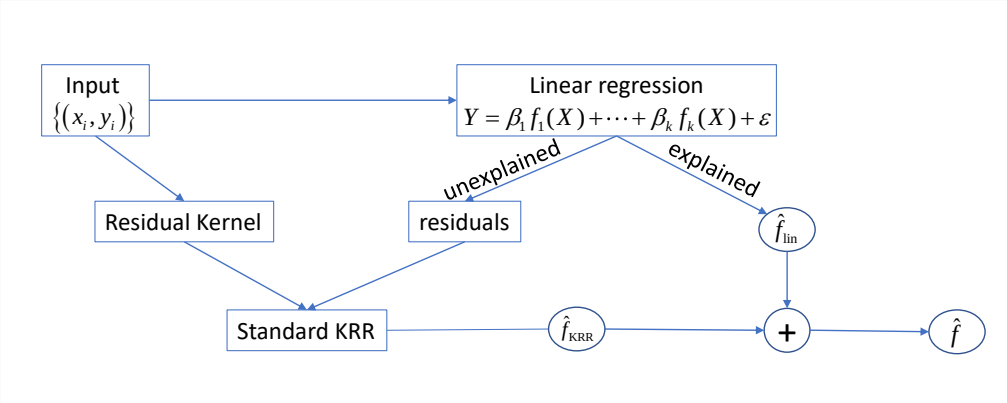
$$r = y - \mathrm{Proj}_{\mathrm{row}(F)}\, y.$$

Then conditional KRR is equivalent to

$$\min_{g \in \mathcal{H}_{K_P}} \frac{1}{N} \sum_{i=1}^{N} (g(x_i) - r_i)^2 + \lambda \|g\|^2_{\mathcal{H}_{K_P}}.$$

Linear regression first, KRR on residuals second.

Diagram


Input
$\{(x_i, y_i)\}$
Linear regression
$Y = \beta_1 f_1(X) + \cdots + \beta_k f_k(X) + \varepsilon$
unexplained
explained
Residual Kernel
residuals
$\hat{f}_{\text{lin}}$
Standard KRR
$\hat{f}_{\text{KRR}}$
+
$\hat{f}$

Cost of conditioning

Assume the true regression function decomposes as

$$f = f_{\parallel} + f_{\perp}, \qquad f_{\parallel} \in \mathcal{F}, \quad f_{\perp} \in \mathcal{H}_{K_P}.$$

Define the ideal $\mathcal{F}$-conditional learner:

$$h = \text{KRR with } K_P \text{ applied to } Y_i - f_{\parallel}(X_i).$$

The cost of conditioning is

$$c_{\text{con}} = \mathbb{E}\big[(\hat{f}(X) - f_{\parallel}(X) - h(X))^2\big].$$

It measures how close conditional KRR is to the ideal two-stage learner.

Main statistical theorem

Under orthogonality and boundedness assumptions on $f_1, \ldots, f_k$, with probability at least $1 - \delta$ over randomness in $X_1, \cdots, X_N$, we have

$$\mathbb{E}_\varepsilon[c_{\mathrm{con}}] \leq C_1 \|f\|^2_{\mathcal{H}_K^{\mathcal{F}}} \frac{k\sqrt{\log(2k/\delta)}}{\sqrt{N}} + C_2 \frac{\sigma^2}{N}.$$

Meaning
Conditional KRR behaves like residual KRR, up to a controlled statistical price.

Important consequence

If the signal is mostly inside $\mathcal{F}$, then

$$f_{\perp} \approx 0.$$

Then the main term disappears or becomes small:

$$\mathbb{E}_{\varepsilon}[c_{\mathrm{con}}] \approx O\left(\frac{\sigma^2 k}{N}\right).$$

Unpenalized features help when they capture a strong signal component.

Application I: hard thresholding

Suppose K is positive definite with Mercer expansion

$$K(x, y) = \sum_{i \geq 1} \lambda_i \phi_i(x) \phi_i(y).$$

Choose

$$\mathcal{F} = \text{span}\{\phi_1, \ldots, \phi_k\}.$$

Then conditional KRR unpenalizes the leading eigenfunctions.

This is kernel thresholding: remove penalty from the top eigenspace.

Application II: soft thresholding

Let us assume that K is a Mercer kernel that is given through the random features mapping $f : \Omega \times \mathcal{X} \to \mathbb{R}$ and $(\Omega, \Sigma, \mathcal{P})$ is a probabilistic space, that is $K(x, y) = \mathbb{E}_{\omega \sim \mathcal{P}}[f(\omega, x) f(\omega, y)]$. Instead of taking the first Mercer eigenfunctions, choose random features

$$f_1, \ldots, f_k$$

sampled from a random feature representation of K. I.e. $f_i(x) = f(\omega_i, x)$ for i.i.d. samples $\omega_1, \ldots, \omega_k \sim \mathcal{P}$.

This produces a soft version of thresholding.

Prediction
The test risk can have a U-shaped dependence on k: too few unpenalized features underfit, too many may overfit.

Experiments with synthetic kernels on $[0, 2\pi]$

$$K(x, y) = 1 + \sum_{i=1}^{\infty} i^{-2s} \big(\cos(ix)\cos(iy) + \sin(ix)\sin(iy) \big)$$

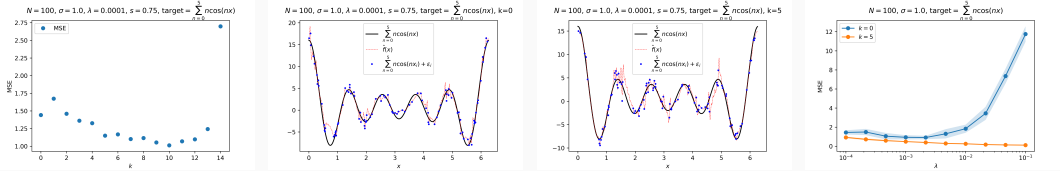


Figure 1: Comparison of test MSE for Conditional KRR with $k = 0$ (standard KRR) and $k = 5$. The last plot compares test MSEs across a range of regularization parameters λ.

Observed behavior

- conditional KRR often improves test MSE;
- risk as a function of k is frequently U-shaped;
- empirically, c_{con} typically decays like $1/N$.

Experiments on the Hard thresholding (real world data case)

- MNIST 7-vs-9 classification with Gaussian, Laplace, and NNGP kernels;

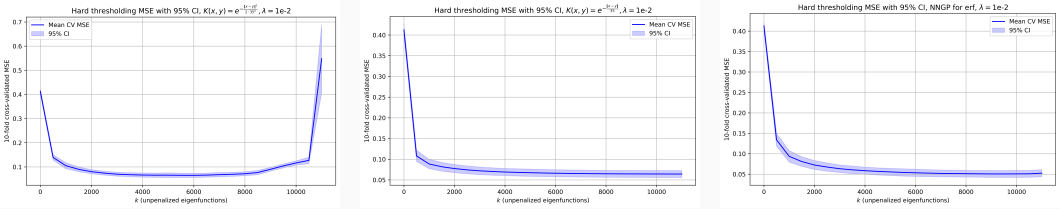


Figure 2: U-shaped Test MSE in the hard thresholding setup for the 7-vs-9 MNIST dataset (with standardization).

Experiments with random feature models

A random field on $\mathcal{X} = \mathbb{S}^{d-1}$ with covariance K was defined as follows:

(a) $f(x, [\omega, b]) = \cos(\omega^\top x + b)$ with $\omega \sim \mathcal{N}(\mathbf{0}, I_d)$ and $b \sim U([0, 2\pi])$;
(b) $f(x, [\omega, b]) = \mathrm{ReLU}(\omega^\top x + b)$ with $\omega \sim \mathcal{N}(\mathbf{0}, I_d)$ and $b \sim U([-1, 1])$;
(c) $f(x, [\omega, b]) = \tanh(\omega^\top x + b)$ with $\omega \sim \mathcal{N}(\mathbf{0}, I_d)$ and $b \sim U([-1, 1])$.

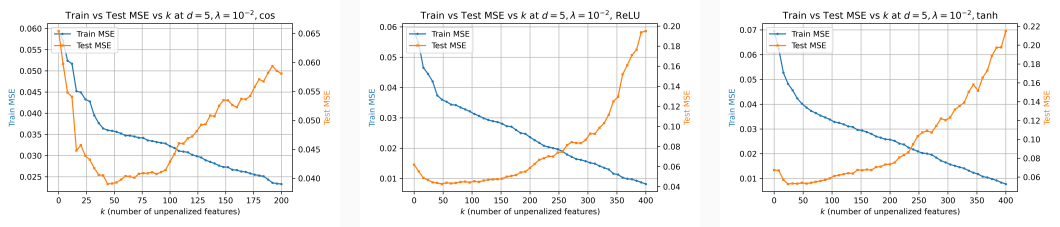


Figure 3: The effect of the soft thresholding for the cosine, ReLU and tanh activation functions and the regression function $f(x_1, \cdots, x_d) = \sin(x_1) + \frac{1}{2}\cos(x_2)$.

Experiments with random feature models (real world data case)

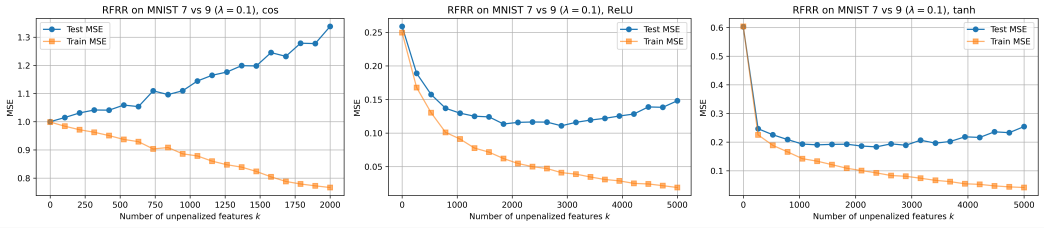


Figure 4: The effect of the soft thresholding for the cosine, ReLU and tanh activation functions on the 7-vs-9 MNIST dataset.

Takeaway

Conditional KRR gives a principled way to combine

$$\text{explicit features} \quad + \quad \text{kernel learning.}$$

It is equivalent to residual KRR, but the equivalence has a statistical cost.

If the chosen features explain an important part of the target,
conditional KRR can outperform standard KRR.

Open direction

The theory assumes

$$\lambda > 0.$$

But perfect memorization corresponds to the interpolation regime

$$\lambda = 0.$$

Open problem
Extend the statistical theory of conditional KRR to the interpolation limit and understand its relation to benign overfitting.