A Spectral Filtering Approach for Kernel Mean Estimation

Krikamol Muandet
Empirical Inference Department
Max Planck Institute for Intelligent Systems

April 1, 2014

A Mean Function in RKHS

The kernel mean and its estimator:

$$\begin{aligned} \mu_P &:= \int_{\mathcal{X}} k(x, \cdot) \, \mathrm{d}P(x) \\ \hat{\mu}_P &:= \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot) \, . \end{aligned}$$

The kernel mean has been used in extensively the kernel-based learning algorithms.

$$X \sim \mathcal{N}(\theta, I_d), \quad \theta \in \mathbb{R}^d$$

$$X \sim \mathcal{N}(\theta, I_d), \quad \theta \in \mathbb{R}^d$$

$$X \sim \mathcal{N}(\theta, I_d), \quad \theta \in \mathbb{R}^d$$

$\bullet X$

$\bullet \left(1 - \frac{d-2}{\|X\|^2}\right) X$

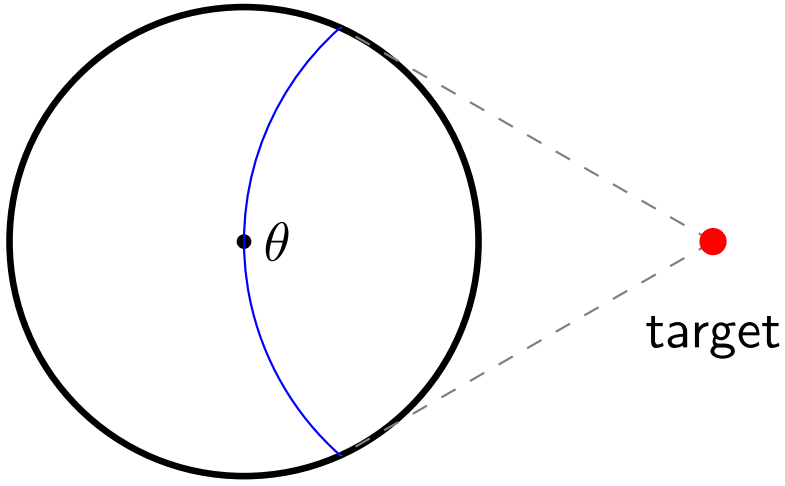
$$X \sim \mathcal{N}(\theta, I_d), \quad \theta \in \mathbb{R}^d$$

$\bullet X$

$\bullet \left(1 - \frac{d-2}{\|X\|^2}\right) X$

$$\mathbb{E}_\theta[\|\hat{\theta}_{JS} - \theta\|^2] \leq \mathbb{E}_\theta[\|\hat{\theta}_{ML} - \theta\|^2]$$

$$X \sim \mathcal{N}(\theta, I)$$



$$\hat{\theta}_{ML} = X, \qquad \hat{\theta}_{JS} = \left(1 - \frac{d-2}{\|X\|^2}\right) X$$

Kernel Mean Shrinkage Estimators

ICML2014, JMLR2014 (In Preparation)

In general, we are interested in the estimators of the form

$$\hat{\mu}_P := \sum_{i=1}^{n} \beta_i k(x_i, \cdot)$$

In the previous work, we proposed two shrinkage estimators:

1. **Simple Shrinkage.**

$$\boldsymbol{\beta} = \left(\frac{1}{1+\lambda}\right)\mathbf{1}_n, \qquad \hat{\mu}_\lambda = (1-\alpha)\hat{\mu}$$

2. **Flexible Shrinkage.**

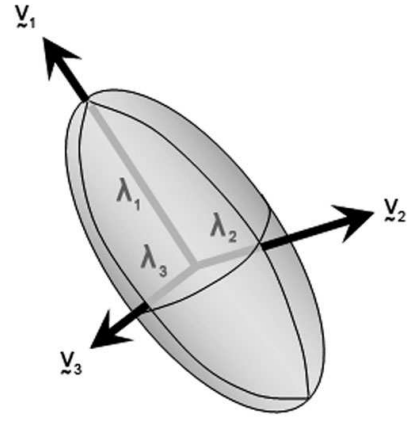
$$\boldsymbol{\beta} = (K+\lambda I)^{-1} K \mathbf{1}_n, \qquad \hat{\mu}_\lambda = \sum_{i=1}^{n} \frac{d_i}{d_i+\lambda} \langle \hat{\mu}, \mathbf{v}_i \rangle \mathbf{v}_i$$

Kernel Mean Shrinkage Estimators

ICML2014, JMLR2014 (In Preparation)

In general, we are interested in the estimators of the form

$$\hat{\mu}_P := \sum_{i=1}^{n} \beta_i k(x_i, \cdot)$$



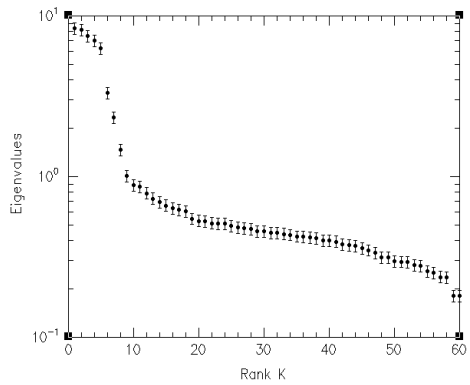
In the previous work, we proposed two shrinkage estimators:

1. **Simple Shrinkage.**

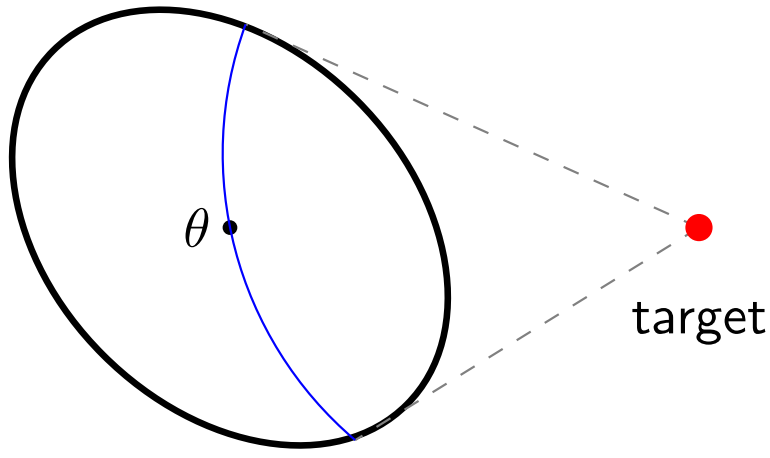
$$\boldsymbol{\beta} = \left(\frac{1}{1+\lambda}\right)\mathbf{1}_n, \qquad \hat{\mu}_\lambda = (1-\alpha)\hat{\mu}$$

2. **Flexible Shrinkage.**

$$\boldsymbol{\beta} = (K+\lambda I)^{-1} K \mathbf{1}_n, \qquad \hat{\mu}_\lambda = \sum_{i=1}^{n} \frac{d_i}{d_i+\lambda} \langle \hat{\mu}, \mathbf{v}_i \rangle \mathbf{v}_i$$



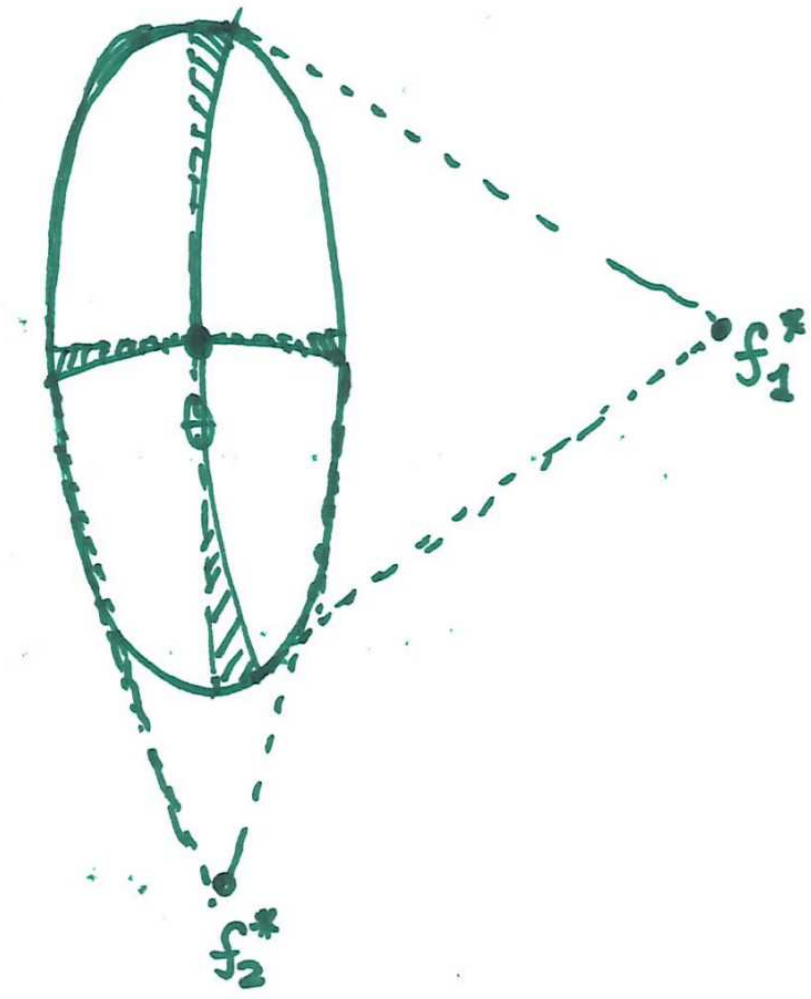
$$X \sim \mathcal{N}(\theta, \Sigma)$$



$$\hat{\theta}_{ML} = X, \qquad \hat{\theta}_{JS} = \left(1 - \frac{d-2}{\|X\|_{\Sigma}^2}\right) X$$

Where to Shrink?

$$X \sim \mathcal{N}(\theta, \Sigma)$$



Kernel Mean Estimation vs. Least Square Regression

Least Square Regression

$$\begin{aligned} \hat{f}_P &= \sum_{i=1}^{n} \beta_i k(x_i, \cdot) \\ K\beta_{LS} &= \mathbf{y} \\ \beta_{LS} &= K^{-1}\mathbf{y} \end{aligned}$$

Spectral Filtering Method

$$\tilde{\beta}_{LS} = g_\lambda(K)\mathbf{y}$$

$g_\lambda(\cdot)$ is a filter function. For Tikhonov regularization, we have

$$g_\lambda(K) = (K + \lambda I)^{-1}.$$

numerical stability $\rightarrow$ generalization

Kernel Mean Estimation

$$\begin{aligned} \hat{\mu}_P &= \sum_{i=1}^{n} \beta_i k(x_i, \cdot) \\ K\beta &= K\mathbf{1}_n \\ \beta &= \mathbf{1}_n \end{aligned}$$

Spectral Filtering Method

$$\tilde{\beta} = g_\lambda(K)K\mathbf{1}_n$$

$g_\lambda(\cdot)$ is a filter function. For flexible shrinkage, we have

$$g_\lambda(K) = (K + \lambda I)^{-1}.$$

shrinkage

Both correspond to interpolation between RKHS and L2, but perhaps from opposite direction (ack. Ingo Steinwart).

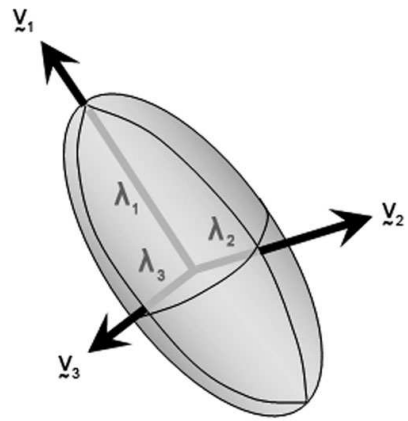
Spectral KMSE

Lemma

The spectral-KMSE satisfies

$$\hat{\mu}_\lambda = \sum_{i=1}^{n} g_\lambda(d_i) d_i \langle \hat{\mu}, \mathbf{v}_i \rangle \mathbf{v}_i,$$

where $d_i, \mathbf{v}_i$ are eigenvalue and eigenvector pairs of the empirical covariance operator $\hat{C}_{xx}$ in $\mathcal{H}$.



Remark

- The function $g_\lambda(d_i)$ should approximate $\frac{1}{d_i}$ as λ goes to zero.
- For flexible KMSE, we have $g_\lambda(d_i)d_i = \frac{d_i}{d_i+\lambda}$.

Spectral Filtering Methods

- **L2 Boosting.** This is also known as gradient descent or Landweber iteration:

$$\beta^t \leftarrow \beta^{t-1} + \eta(K\mathbf{1}_n - K\beta^{t-1})$$

  The filter function is $g_\lambda(d) = \eta \sum_{i=1}^{t-1}(I - \eta d)^i$.

- **Accelerated L2 Boosting.** This is also known as ν-method:

$$\beta^t \leftarrow \beta^{t-1} + \omega_t(\beta^{t-1} - \beta^{t-2}) + \frac{\kappa_t}{n}(K\mathbf{1}_n - K\beta^{t-1})$$

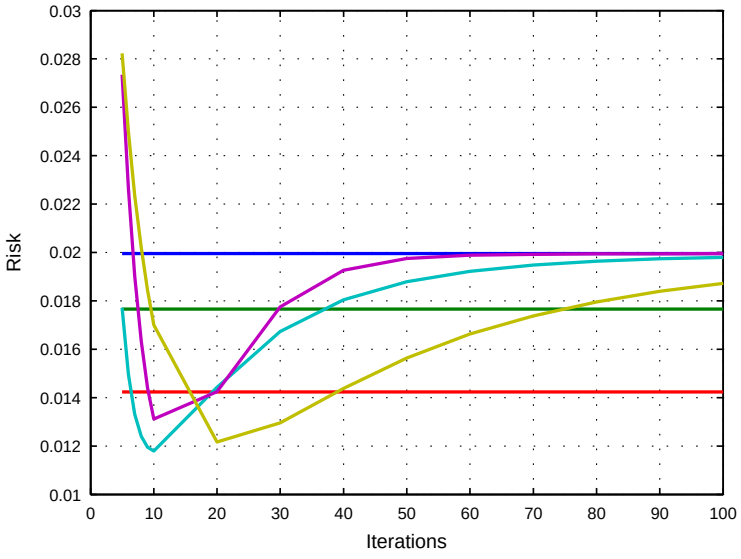
  The filter function is $g_t(d) = p_t(d)$ where p_t is a polynomial of degree $t - 1$.

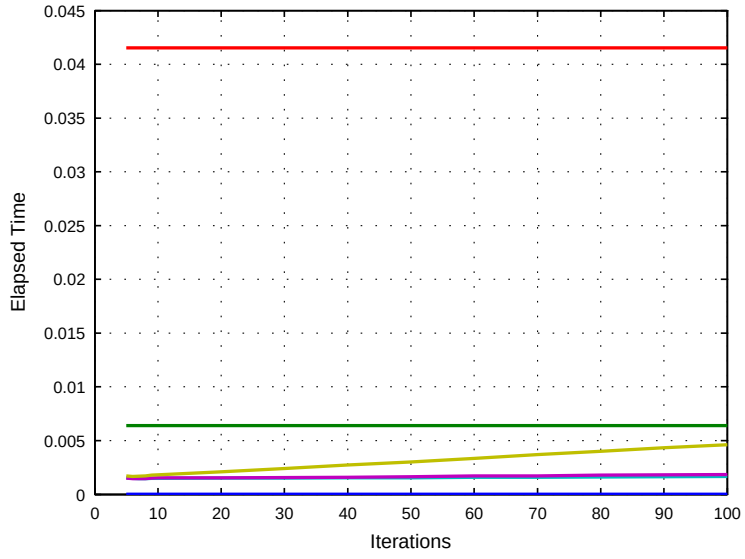
- **Iterated Tikhonov.** This method can be viewed as a combination of Tikhonov regularization and gradient descent. The filter function is given by

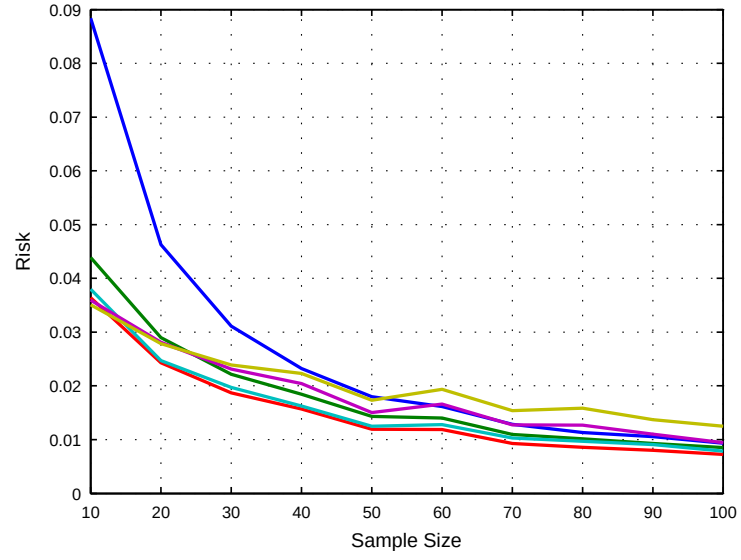
$$g_{\lambda,t}(d) = \frac{(d+\lambda)^t - d^t}{\lambda(d+\lambda)^t}$$

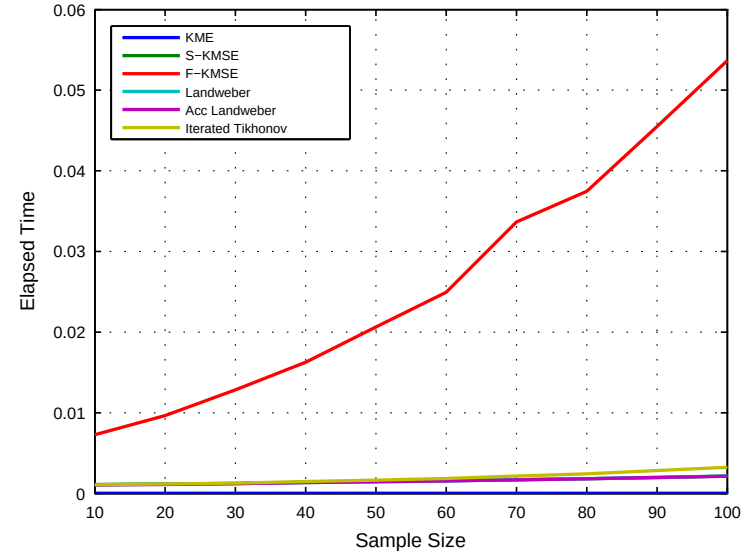
- **Truncated SVD.** The filter function is defined as $g_\lambda(d) = 1/d$ if $d \geq \lambda$ and 0 otherwise.

Preliminary Results









Question?