Kernel Mean Estimation via Spectral Filtering: Supplementary Material

Krikamol Muandet  
MPI-IS, Tübingen  
krikamol@tue.mpge.de

Bharath Sriperumbudur  
Dep. of Statistics, PSU  
bks18@psu.edu

Bernhard Schölkopf  
MPI-IS, Tübingen  
bs@tue.mpge.de

Abstract

This note contains supplementary materials to Kernel Mean Estimation via Spectral Filtering.

1 Proof of Theorem 1

(i) Since \( \hat{\mu}_\lambda = \frac{\hat{\mu}_p}{\lambda+1} \), we have

\[
\| \hat{\mu}_\lambda - \mu_P \| = \left\| \frac{\hat{\mu}_P}{\lambda+1} - \mu_P \right\| \leq \left\| \frac{\hat{\mu}_P}{\lambda+1} - \frac{\mu_P}{\lambda+1} \right\| + \left\| \frac{\mu_P}{\lambda+1} - \mu_P \right\| \leq \| \hat{\mu}_P - \mu_P \| + \lambda \| \mu_P \|.
\]

From [1], we have that \( \| \hat{\mu}_P - \mu_P \| = O_P(n^{-1/2}) \) and therefore the result follows.

(ii) Define \( \Delta := \mathbb{E}_P \| \hat{\mu}_P - \mu_P \|^2 = \int k(x,x) d\mathbb{P}(x) - \| \mu_P \|^2 \). Consider

\[
\mathbb{E}_P \| \hat{\mu}_\lambda - \mu_P \|^2 - \Delta = \mathbb{E}_P \left[ \left\| \frac{n^\beta}{n^\beta+c} (\hat{\mu}_P - \mu_P) - \mu_P \right\|^2 - \Delta \right] = \left( \frac{n^\beta}{n^\beta+c} \right)^2 \Delta + \frac{c^2}{(n^\beta+c)^2} \| \mu_P \|^2 - \Delta = \frac{c^2 \| \mu_P \|^2 - (c^2 + 2cn^\beta) \Delta}{(n^\beta+c)^2}.
\]

Substituting for \( \Delta \) in the r.h.s. of the above equation, we have

\[
\mathbb{E}_P \| \hat{\mu}_\lambda - \mu_P \|^2 - \Delta = \frac{(nc^2 + c^2 + 2cn^\beta) \| \mu_P \|^2 - (c^2 + 2cn^\beta) \int k(x,x) d\mathbb{P}(x)}{n(n^\beta+c)^2}.
\]

It is easy to verify that \( \mathbb{E}_P \| \hat{\mu}_\lambda - \mu_P \|^2 - \Delta < 0 \) if

\[
\int k(x,x) d\mathbb{P}(x) < \inf \frac{c^2 + 2cn^\beta}{nc^2 + c^2 + 2cn^\beta} = \frac{2^{1/\beta} \beta}{2^{1/\beta} \beta + c^{1/\beta}(\beta - 1)^{(\beta-1)/\beta}}.
\]

Remark. If \( k(x,y) = (x,y) \), then it is easy to check that \( \mathcal{P}_{\epsilon,\beta} = \{ \mathbb{P} \in M_+^1(\mathbb{R}^d) : \frac{\| \theta \|^2}{\text{trace}(\Sigma)} < \frac{\epsilon}{1-\epsilon} \} \) where \( \theta \) and \( \Sigma \) represent the mean vector and covariance matrix. Note that this choice of kernel yields a setting similar to classical James-Stein estimation, wherein for all \( n \) and all \( \mathbb{P} \in \mathcal{P}_{\epsilon,\beta} = \{ \mathbb{P} \in N_{\theta,\sigma} : \| \theta \| < \sigma \sqrt{dA/(1-A)} \} \), \( \hat{\mu}_\lambda \) is admissible for any \( d \), where \( N_{\theta,\sigma} := \{ \mathbb{P} \in M_+^1(\mathbb{R}^d) : d\mathbb{P}(x) = (2\pi \sigma^2)^{-d/2} \frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{2\pi^d} dx, \ \theta \in \mathbb{R}^d, \ \sigma > 0 \} \). On the other hand, the James-Stein estimator is admissible for only \( d \geq 3 \) but for any \( \mathbb{P} \in N_{\theta,\sigma} \).
2 Consequence of Theorem 1 if $k$ is translation invariant

Claim: Let $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$ where $\psi$ is a bounded continuous positive definite function with $\psi \in L^1(\mathbb{R}^d)$. For $\lambda = cn^{-\beta}$ with $c > 0$ and $\beta > 1$, define

$$\mathcal{P}_{c, \beta, \psi} := \left\{ P \in M_+^1(\mathbb{R}^d) : \| \phi_P \|_{L^2} < \sqrt{\frac{A(2\pi)^d/2}{\| \psi \|_{L^1}}} \right\},$$

where $\phi_P$ is the characteristic function of $P$. Then $\forall n$ and $\forall P \in \mathcal{P}_{c, \beta, \psi}$, we have $E_P[\bar{\mu}_h - \mu_P]^2 < E_P[\bar{\mu}_h - \mu_P]^2$.

Proof. If $k(x, y) = \psi(x - y)$, it is easy to verify that

$$\int \int k(x, y) dP(x) dP(y) = \int |\phi_P(\omega)|^2 \hat{\psi}(\omega) d\omega \leq \sup_{\omega \in \mathbb{R}^d} \hat{\psi}(\omega) \| \phi_P \|^2_{L_2} \leq (2\pi)^{-d/2} \| \psi \|_{L_1} \| \phi_P \|^2_{L_2},$$

where $\hat{\psi}$ is the Fourier transform of $\psi$. On the other hand, since $|\phi_P(\omega)| \leq 1$ for any $\omega \in \mathbb{R}^d$, we have

$$\int \int k(x, y) dP(x) dP(y) = \int |\phi_P(\omega)|^2 \hat{\psi}(\omega) d\omega \leq \int |\phi_P(\omega)| \hat{\psi}(\omega) d\omega \leq \| \phi_P \|_{L^2} \| \hat{\psi} \|_{L^2} \leq \| \phi_P \|_{L^2} \sqrt{\| \hat{\psi} \|_{L^1} \| \hat{\psi} \|_{L^1}},$$

where we used $\psi(0) = |\hat{\psi}|_{L^1}$. As $k(x, x) dP(x) = \psi(0)$, we have that

$$\| \mu_P \|^2 \leq \min \left\{ \frac{\| \phi_P \|^2_{L_2} \| \psi \|_{L_1}}{(2\pi)^{d/2} \psi(0)} : \sqrt{\frac{\| \phi_P \|^2_{L_2} \| \psi \|_{L_1}}{(2\pi)^{d/2} \psi(0)}} \right\}.$$

Since $P \in \mathcal{P}_{c, \beta, \psi}$, we have $P \in \mathcal{P}_{c, \beta}$ and therefore the result follows.

3 Proof of Theorem 2

Since $(e_i)_i$ is an orthonormal basis in $\mathcal{H}$, we have for any $P$ and $f^* \in \mathcal{H}$

$$\mu_P = \sum_{i=1}^{\infty} \mu_i e_i, \quad \bar{\mu}_P = \sum_{i=1}^{\infty} \bar{\mu}_i e_i, \quad \text{and} \quad f^* = \sum_{i=1}^{\infty} f^*_i e_i,$$

where $\mu_i := \langle \mu_P, e_i \rangle$, $\bar{\mu}_i := \langle \bar{\mu}_P, e_i \rangle$, and $f^*_i := \langle f^*, e_i \rangle$. If follows from the Parseval’s identity that

$$\Delta = E_P[|\bar{\mu} - \mu|^2] = E_P \left[ \sum_{i=1}^{\infty} (\bar{\mu}_i - \mu_i)^2 \right] = \sum_{i=1}^{\infty} \Delta_i,$$

$$\Delta_{\alpha} = E_P[|\bar{\mu}_\alpha - \mu|^2] = E_P \left[ \sum_{i=1}^{\infty} (\alpha_i f_i^* + (1 - \alpha_i) \bar{\mu}_i - \mu_i)^2 \right] = \sum_{i=1}^{\Delta_{\alpha}} \Delta_{\alpha, i}.$$

Note that the problem has not changed and we are merely looking at it from a different perspective. To estimate $\mu_P$, we may just as well estimate its Fourier coefficient sequence $\mu_i$ with $\bar{\mu}_i$. Based on above decomposition, we may write the risk difference $\Delta_{\alpha} - \Delta$ as $\sum_{i=1}^{\infty} (\Delta_{\alpha, i} - \Delta_i)$. We can thus ask under which conditions on $\alpha = (\alpha_i)$ for which $\Delta_{\alpha, i} - \Delta_i < 0$ uniformly over all $i$.

For each coordinate $i$, we have

$$\Delta_{\alpha, i} - \Delta_i = E_P \left[ [(\alpha_i f_i^* + (1 - \alpha_i) \bar{\mu}_i - \mu_i)^2] - E_P \left[ (\bar{\mu}_i - \mu_i)^2 \right] \right] = E_P[\alpha_i^2 f_i^2 + 2\alpha_i f_i^* (1 - \alpha_i) \bar{\mu}_i + (1 - \alpha_i)^2 \mu_i^2 - 2\alpha_i f_i^* \bar{\mu}_i - 2(1 - \alpha_i) \bar{\mu}_i \mu_i + \mu_i^2] = \alpha_i^2 f_i^2 + 2\alpha_i f_i^* \bar{\mu}_i - 2\alpha_i \bar{\mu}_i \mu_i + (1 - \alpha_i)^2 \mu_i^2.$$
Next, we substitute $E_P[\hat{\mu}^2] = E_P[(\hat{\mu}_i - \mu_i + \mu_i)^2] = \Delta_i + \mu_i^2$ into the last equation to obtain
\[
\Delta_{\alpha,i} - \Delta_i = \alpha_i^2 f_i^2 - 2\alpha_i^2 f_i^* \mu_i + \alpha_i^2 (\Delta_i + \mu_i^2) - 2\alpha_i (\Delta_i + \mu_i^2) + 2\alpha_i \mu_i^2 = \alpha_i^2 (f_i^2 - 2f_i^* \mu_i + \Delta_i + \mu_i^2) - 2\alpha_i \Delta_i = \alpha_i^2 (\Delta_i + (f_i^* - \mu_i)^2) - 2\alpha_i \Delta_i,
\]
which is negative if $\alpha_i$ satisfies
\[
0 < \alpha_i < \frac{2\Delta_i}{\Delta_i + (f_i^* - \mu_i)^2}.
\]
This completes the proof.

4 Proof of Proposition 3

Let $K = UDU^T$ be an eigen-decomposition of $K$ where $U = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n]$ consists of orthogonal eigenvectors of $K$ such that $U^T U = I$ and $D = \text{diag}(\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_n)$ consists of corresponding eigenvalues. As a result, the coefficients $\beta(\lambda)$ can be written as
\[
\beta(\lambda) = g_\lambda(K)K1_n = U g_\lambda(D)U^T K1_n = \sum_{i=1}^{n} \hat{u}_i g_\lambda(\hat{\gamma}_i) \hat{u}_i^T K1_n.
\]
Using $K1_n = [(\hat{\mu}, k(x_1, \cdot)), \ldots, (\hat{\mu}, k(x_n, \cdot))]^T$, we can rewrite (1) as
\[
\beta(\lambda) = \sum_{i=1}^{n} \hat{u}_i g_\lambda(\hat{\gamma}_i) \sum_{j=1}^{n} \hat{u}_{ij} (\hat{\mu}, k(x_j, \cdot)) = \sum_{i=1}^{n} \sqrt{\hat{\gamma}_i} \hat{u}_i g_\lambda(\hat{\gamma}_i) \left( \hat{\mu}, \frac{1}{\sqrt{\hat{\gamma}_i}} \sum_{j=1}^{n} \hat{u}_{ij} k(x_j, \cdot) \right),
\]
where $\hat{u}_{ij}$ is the $j$th component of $\hat{u}_i$. Next, we invoke the relation between the eigenvectors of the matrix $K$ and the eigenfunctions of the empirical covariance operator $\hat{C}_K$ in $\mathcal{H}$. That is, it is known that the $i$th eigenfunction of $\hat{C}_K$ can be expressed as $\hat{v}_i = (1/\sqrt{\hat{\gamma}_i}) \sum_{j=1}^{n} \hat{u}_{ij} k(x_j, \cdot)$ [2]. Consequently,
\[
\left( \hat{\mu}, \frac{1}{\sqrt{\hat{\gamma}_i}} \sum_{j=1}^{n} \hat{u}_{ij} k(x_j, \cdot) \right) = \langle \hat{\mu}, \hat{v}_i \rangle
\]
and we can write the Spectral-KMSE as
\[
\hat{\mu}_\lambda = \sum_{j=1}^{n} \left[ \sum_{i=1}^{n} \hat{u}_{ij} \sqrt{\hat{\gamma}_i} g_\lambda(\hat{\gamma}_i) \langle \hat{\mu}, \hat{v}_i \rangle \right] k(x_j, \cdot) = \sum_{i=1}^{n} \sqrt{\hat{\gamma}_i} g_\lambda(\hat{\gamma}_i) \langle \hat{\mu}, \hat{v}_i \rangle \sum_{j=1}^{n} \hat{u}_{ij} k(x_j, \cdot) = \sum_{i=1}^{n} g_\lambda(\hat{\gamma}_i) \hat{\gamma}_i \langle \hat{\mu}, \hat{v}_i \rangle \hat{v}_i.
\]
This completes the proof.

5 Population counterpart of Spectral-KMSE

To obtain the population version of the Spectral-KMSE, we resort to the regression perspective of the kernel mean embedding which has been studied earlier in [3, 4]. The proof techniques used here are similar to those in [3]. Consider
\[
\arg\min_{F \in \mathcal{H} \otimes \mathcal{H}} E_X \left[ \|k(X, \cdot) - Fk(X, \cdot)\|_2^2 \right] + \lambda \|F\|^2_{HS}.
\]
where $F : \mathcal{H} \to \mathcal{H}$ is Hilbert-Schmidt. We can expand the regularized loss (2) as

$$
\mathbb{E}_X \left[ ||k(X, \cdot) - F k(X, \cdot)||_\mathcal{C}^2 + \lambda \|F\|_{HS}^2 \right] = \mathbb{E}_X \langle k(X, \cdot), k(X, \cdot) \rangle_\mathcal{C} - 2 \mathbb{E}_X \langle k(X, \cdot), F k(X, \cdot) \rangle_\mathcal{C} + \mathbb{E}_X \langle F k(X, \cdot), F k(X, \cdot) \rangle_\mathcal{C} + \lambda \langle F, F \rangle_{HS}
$$

$$
= \mathbb{E}_X \langle k(X, \cdot), k(X, \cdot) \rangle_\mathcal{C} - 2 \mathbb{E}_X \langle k(X, \cdot) \otimes k(X, \cdot), F \rangle_{HS} + \mathbb{E}_X \langle k(X, \cdot), F^* F k(X, \cdot) \rangle_\mathcal{C} + \lambda \langle F, F \rangle_{HS}
$$

where $F^*$ denotes the adjoint of $F$ and $C_k = \mathbb{E}_X [k(X, \cdot) \otimes k(X, \cdot)]$. Next, we show that the solution to the above expression is $F := C_k (C_k + \lambda I)^{-1}$. Defining $A := F (C_k + \lambda I)^{1/2}$, the above expression can be rewritten as

$$
\mathbb{E}_X \langle k(X, \cdot), k(X, \cdot) \rangle_\mathcal{C} - 2 \langle C_k, F \rangle_{HS} + \langle C_k, F^* F \rangle_{HS} + \lambda \langle F, F \rangle_{HS}
$$

$$
= \mathbb{E}_X \langle k(X, \cdot), k(X, \cdot) \rangle_\mathcal{C} - 2 \langle C_k, A (C_k + \lambda I)^{-1/2} \rangle_{HS} + \langle A, A \rangle_{HS}
$$

$$
= \mathbb{E}_X \langle k(X, \cdot), k(X, \cdot) \rangle_\mathcal{C} - \|C_k (C_k + \lambda I)^{-1/2} \|_{HS}^2 + \|C_k (C_k + \lambda I)^{-1/2} - A \|_{HS}^2.
$$

As a result, the above expression is minimized when $A = C_k (C_k + \lambda I)^{-1/2}$, implying that $F = C_k (C_k + \lambda I)^{-1}$. As in the sample case, a natural estimate of the Spectral-KMSE is

$$
\mu_\lambda = F \mu_\beta = C_k (C_k + \lambda I)^{-1} \mu_\beta.
$$

\section{Proof of Proposition 4}

The proof employs the relation between the Gram matrix $K$ and the empirical covariance operator $C_k$ shown in Lemma 3. It is known that the operator $C_k$ is of finite rank, self-adjoint, and positive. Moreover, its spectrum has only finitely many nonzero elements [5]. If $\gamma_i$ is a nonzero eigenvalue and $\tilde{v}_i$ is the corresponding eigenfunction of $C_k$, then the following decomposition holds

$$
C_k f = \sum_{i=1}^n \gamma_i \langle f, \tilde{v}_i \rangle \tilde{v}_i, \quad \forall f \in \mathcal{H}.
$$

Note that it may be that $k < n$ where $k$ is the rank of $C_k$. In that case, the above decomposition still holds. Setting $f = \mu$ and applying the definition of the filter function $g_\lambda$ to the operator $C_k$ yield

$$
\mu_\lambda = C_k g_\lambda (C_k) \mu = \sum_{i=1}^n g_\lambda (\gamma_i) \gamma_i \langle \mu, \tilde{v}_i \rangle \tilde{v}_i,
$$

which is exactly the decomposition given in Lemma 3. This completes the proof.

\section{Proof of Theorem 5}

Consider the following decomposition

$$
\mu_\lambda - \mu_\beta = C_k g_\lambda (C_k) \mu_\beta - \mu_\beta = C_k g_\lambda (C_k) (\mu_\beta - \mu_\beta) + C_k g_\lambda (C_k) \mu_\beta - \mu_\beta
$$

$$
= C_k g_\lambda (C_k) (\mu_\beta - \mu_\beta) + (C_k g_\lambda (C_k) - I) \hat{C}_k \beta \hat{h} + (C_k g_\lambda (C_k) - I) (C_k^0 - \hat{C}_k^0) \beta \hat{h}
$$

where we used the fact that there exists $h \in \mathcal{H}$ such that $\mu_\beta = C_k^0 h$ as we assumed that $\mu_\beta \in \mathcal{R}(C_k^0)$ for some $\beta > 0$. Therefore

$$
\|\mu_\lambda - \mu_\beta\| \leq \|C_k g_\lambda (C_k)\|_{op} \|\mu_\beta - \mu_\beta\| + \|C_k g_\lambda (C_k) - I\|_{op} \|\hat{C}_k \beta \hat{h}\| + \|C_k g_\lambda (C_k) - I\|_{op} \|C_k^0 - \hat{C}_k^0\|_{op} \|\hat{h}\|
$$

where we used the fact that $\|A b\| \leq \|A\|_{op} \|b\|$ with $A : \mathcal{H} \to \mathcal{H}$ being a bounded operator, $b \in \mathcal{H}$ and $\|\cdot\|_{op}$ denoting the operator norm defined as $\|A\|_{op} := \sup \{\|A b\| : \|b\| = 1\}$. 

4
By (C1), (C2) and (C3), we have \( \| \hat{C}_k g_\lambda(\hat{C}_k) \|_{op} \leq B, \| \hat{C}_k g_\lambda(\hat{C}_k) - I \|_{op} \leq C \) and \( \| (\hat{C}_k g_\lambda(\hat{C}_k) - I)^{\beta \gamma} \|_{op} \leq D^{\lambda \mu} \) respectively. Denoting \( \| h \| = \| C_k^{-\beta} \mu_p \| \), we therefore have

\[
\| \hat{\mu}_\lambda - \mu_p \| \leq B \| \hat{\mu}_p - \mu_p \| + D^{\lambda \mu} \| C_k^{-\beta} \mu_p \| + C \| C_k^{-\beta} \|_{op} \| C_k^{-\beta} \mu_p \|.
\]

(3)

For \( 0 \leq \beta \leq 1 \), it follows from Theorem 1 in [6] that there exists a constant \( \tau_1 \) such that

\[
\| C_k^\beta - \hat{C}_k^\beta \|_{op} \leq \tau_1 \| C_k - \hat{C}_k \|_{op}^{\beta} \leq \tau_1 \| C_k - \hat{C}_k \|_{HS}^{\beta}.
\]

On the other hand, since \( \alpha \to \alpha^\beta \) is Lipschitz on \([0, \kappa^2]\) for \( \beta \geq 1 \), the following lemma yields that

\[
\| C_k^\beta - \hat{C}_k^\beta \|_{op} \leq \| C_k^\beta - \hat{C}_k^\beta \|_{HS} \leq \tau_2 \| C_k - \hat{C}_k \|_{HS}
\]

where \( \tau_2 \) is the Lipschitz constant of \( \alpha \to \alpha^\beta \) on \([0, \kappa^2]\). In other words,

\[
\| C_k^\beta - \hat{C}_k^\beta \|_{op} \leq \max \{ \tau_1, \tau_2 \} \| C_k - \hat{C}_k \|_{HS}^{\min \{1, \beta\}}.
\]

(4)

**Lemma 1** (Contributed by Andreas Maurer, see Lemma 5 in [7]). Suppose \( A \) and \( B \) are self-adjoint Hilbert-Schmidt operators on a separable Hilbert space \( H \) with spectrum contained in the interval \([a, b]\), and let \( (\sigma_i)_{i \in I} \) and \( (\tau_j)_{j \in J} \) be the eigenvalues of \( A \) and \( B \), respectively. Given a function \( r : [a, b] \to \mathbb{R} \), if there exists a finite constant \( L \) such that

\[
| r(\sigma_i) - r(\tau_j) | \leq L | \sigma_i - \tau_j |, \quad \forall i \in I, j \in J,
\]

then

\[
\| r(A) - r(B) \|_{HS} \leq L \| A - B \|_{HS}.
\]

Using (4) in (3), we have

\[
\| \hat{\mu}_\lambda - \mu_p \| \leq B \| \hat{\mu}_p - \mu_p \| + D^{\lambda \mu} \| C_k^{-\beta} \mu_p \| + C \| C_k^{-\beta} \|_{op} \| C_k^{-\beta} \mu_p \|,
\]

(5)

where \( \tau := \max \{ \tau_1, \tau_2 \} \). We now obtain bounds on \( \| \hat{\mu}_p - \mu_p \| \) and \( \| C_k - \hat{C}_k \|_{HS} \) using the following results.

**Lemma 2** ([8]). Suppose that \( \kappa = \sup_{x \in X} \sqrt{k(x, x)} \). For any \( \delta > 0 \), the following inequality holds with probability at least \( 1 - e^{-\delta} \)

\[
\| \hat{\mu}_p - \mu_p \| \leq \frac{2 \kappa + \kappa \sqrt{2\delta}}{\sqrt{n}}.
\]

**Lemma 3** (e.g., see Theorem 7 in [5]). Let \( \kappa := \sup_{x \in X} \sqrt{k(x, x)} \). For \( n \in \mathbb{N} \) and any \( \delta > 0 \), the following inequality holds with probability at least \( 1 - 2e^{-\delta} \):

\[
\| \hat{C}_k - C_k \|_{HS} \leq \frac{2 \sqrt{2} \kappa \sqrt{\delta}}{\sqrt{n}}.
\]

Using Lemmas 2 and 3 in (5), for any \( \delta > 0 \), with probability \( 1 - 3e^{-\delta} \), we obtain

\[
\| \hat{\mu}_\lambda - \mu_p \| \leq \frac{2 \kappa B + \kappa B \sqrt{2\delta}}{\sqrt{n}} + D^{\lambda \mu} \| C_k^{-\beta} \mu_p \| + C \| C_k^{-\beta} \|_{op} \| C_k^{-\beta} \mu_p \|,
\]

(5)

**8 Shrinkage parameter** \( \lambda = cn^{-\beta} \)

In this section, we provide supplementary results that demonstrate the effect of the shrinkage parameter \( \lambda \) presented in Theorem 1. That is, if we choose \( \lambda = cn^{-\beta} \) for some \( c > 0 \) and \( \beta > 1 \), the estimator \( \hat{\mu}_\lambda \) is a proper estimator of \( \mu \). Unfortunately, the true value of \( \beta \), which characterizes the smoothness of the true kernel mean \( \mu_p \), is not known in practice. Nevertheless, we provide simulated experiments that illustrate the convergence of the estimator \( \hat{\mu}_\lambda \) for different values of \( c \) and \( \beta \).

The data-generating distribution used in this experiment is identical to the one we consider in our previous experiments on synthetic data. That is, the data are generated as follows: \( x \sim \)
Figure 1: The risk of shrinkage estimator $\hat{\mu}_\lambda$ when $\lambda = cn^{-\beta}$. The left figure shows the risk of the shrinkage estimator as sample size increases while fixing the value of $\beta$, whereas the right figure shows the same plots while fixing the value of $c$. See text for more explanation.

$$\sum_{i=1}^{4} \pi_i N(\theta_i, \Sigma_i) + \varepsilon, \theta_{ij} \sim U(-10, 10), \Sigma_i \sim W(3 \times I_d, 7), \varepsilon \sim N(0, 0.2 \times I_d)$$

where $U(a, b)$ and $W(\Sigma, df)$ are the uniform distribution and Wishart distribution, respectively. We set $\pi = [0.05, 0.3, 0.4, 0.25]$. We use the Gaussian RBF kernel $k(x, x') = \exp(-\|x-x'\|^2/2\sigma^2)$ whose bandwidth parameter is calculated using the median heuristic, i.e., $\sigma^2 = \text{median}\{\|x_i - x_j\|^2\}$.

Figure 1 depicts the comparisons between the standard kernel mean estimator and the shrinkage estimators with varying values of $c$ and $\beta$.

As we can see in Figure 1, if $c$ is very small or $\beta$ is very large, the shrinkage estimator $\hat{\mu}_\lambda$ behaves like the empirical estimator $\hat{\mu}_P$. This coincides with the intuition given in Theorem 1. Note that the value of $\beta$ specifies the smoothness of the true kernel mean $\mu$ and is unknown in practice. Thus, one of the interesting future directions is to develop procedure that can adapt to this unknown parameter automatically.

References