RKHS reading group: Notes

Oxford Kernels

July 30, 2015

1 Integral operator and covariance operator

Definition 1 (Integral operator)**.** Let k be a continuous kernel on a compact metric space $\mathcal{X}$, and let ν be a finite Borel measure on $\mathcal{X}$. Let S_k be the "convolution":

$$\begin{aligned} S_k : L_2(\mathcal{X};\nu) &\to \mathcal{H}_k, \\ (S_k f)(x) &= \langle f, I_k k(x,\cdot)\rangle_{L_2(\mathcal{X};\nu)} \\ &= \int k(x,y) f(y) d\nu(y), \quad f \in L_2(\mathcal{X};\nu), \end{aligned}$$

and $T_k = I_k S_k$ its composition with the inclusion $I_k : \mathcal{H}_k \hookrightarrow L_2(\mathcal{X};\nu)$. T_k is said to be the *integral operator* of kernel k.

To see that this definition is well posed, i.e. that $\mathrm{im}(S_k) \subset \mathcal{H}_k$, use the same arguments as in the section on cross-covariance operators below.

Fact 1. *$I_k = S_k^*$. In particular, $T_k = S_k^* S_k$ is self-adjoint.*

Proof. Follows from

$$\begin{aligned} \langle h, S_k f\rangle_{\mathcal{H}_k} &= \left\langle h, \int k(\cdot,y) f(y) d\nu(y)\right\rangle_{\mathcal{H}_k} \\ &= \int \langle h, k(\cdot,y)\rangle_{\mathcal{H}_k} f(y) d\nu(y) \\ &= \int h(y) f(y) d\nu(y) \\ &= \langle I_k h, f\rangle_{L_2(\mathcal{X};\nu)}. \end{aligned}$$

□

Definition 2 (Covariance operator)**.** $C_\nu = S_k S_k^*$ is the (uncentred) covariance operator of ν.

Fact 2. $\langle h, C_\nu g\rangle_{\mathcal{H}_k} = \langle h, S_k S_k^* g\rangle_{\mathcal{H}_k} = \langle I_k h, I_k g\rangle_{L_2(\mathcal{X};\nu)} = \int h(x) g(x) d\nu(x)$.

Even though C_ν and T_k differ only in the order of taking inclusions (which seems irrelevant at first sight), they can be different objects in general!

Example 1. If ν is a probability measure and $\mathcal{X} \subset \mathbb{R}^p$, this is just like a (linear, uncentred) covariance matrix $C_\nu = \mathbb{E}_{X\sim\nu}\left[XX^\top\right]$ of a random vector X. If we consider covariance between scalar projections of X, then

$$\mathbb{E}\left[\left(a^\top X\right)\left(b^\top X\right)\right] \quad = \quad a^\top C_\nu b.$$

Indeed, for a linear kernel, $\mathcal{H}_k$ is the set of linear functionals on $\mathcal{X}$ and can be identified with $\mathbb{R}^p$ and $C_\nu : \mathbb{R}^p \to \mathbb{R}^p$. T_k is a different object though: to every $f \in L_2(\mathcal{X};\nu)$, it associates a linear functional $x \mapsto \left(\int y f(y) d\nu(y)\right)^\top x$. Notice that eigenfunctions of T_k in this case must be linear, so $f(y) = a^\top y$ for some $a \in \mathbb{R}^p$, so eigenvalue equation reads

$$\begin{aligned} \int y a^\top y d\nu(y) &= \lambda a \quad \Leftrightarrow \\ C_\nu a &= \lambda a. \end{aligned}$$

Thus, even though T_k is nominally an operator on L_2 it has at most p non-zero eigenvalues (which are the same as the eigenvalues of C_ν).

1.1 Cross-covariance operator

Let's take $\mathcal{F} = L^2(\mathcal{X}\times\mathcal{Y}, P_{XY})$, for some joint distribution P_{XY} on $\mathcal{X}\times\mathcal{Y}$. Further, assume that $l_y \doteq l(y,\cdot) \in \mathcal{F}$, and $k_x \doteq k(x,\cdot) \in \mathcal{F}$, i.e., that:

$$\begin{aligned} \int l^2(y,\tilde{y}) dP_{XY}(\tilde{x},\tilde{y}) &= \int l^2(y,\tilde{y}) dP_Y(\tilde{y}) \\ &< \infty, \end{aligned}$$

$$\begin{aligned} \int k^2(x,\tilde{x}) dP_{XY}(\tilde{x},\tilde{y}) &= \int k^2(x,\tilde{x}) dP_X(\tilde{x}) \\ &< \infty. \end{aligned}$$

- Denote by$\iota_k : \mathcal{H}_k \to L^2(\mathcal{X}\times\mathcal{Y}, P_{XY})$ and $\iota_l : \mathcal{H}_l \to L^2(\mathcal{X}\times\mathcal{Y}, P_{XY})$ inclusions of the respective RKHSs into $\mathcal{F}$. Consider $R_l : L^2(\mathcal{X}\times\mathcal{Y}, P_{XY}) \to \mathbb{R}^\mathcal{Y}$, given by

  $$\begin{aligned} (R_l f)(y) &= \langle f, \iota_l l_y\rangle_\mathcal{F} \\ &= \int f(\tilde{x},\tilde{y}) l(\tilde{y},y) dP_{XY}(\tilde{x},\tilde{y}), \end{aligned}$$

  and similarly for R_k.

- Let us show that $\mathrm{im}(R_l)$ is inside $\mathcal{H}_l$ (similarly, $\mathrm{im}(R_k)$ is inside $\mathcal{H}_k$). First, function $(x,y) \mapsto f(x,y)l_y$ is Bochner P_{XY}-integrable on $\mathcal{H}_l$, since:

$$\begin{aligned}\int \|f(x,y)l_y\|_{\mathcal{H}_l} dP_{XY}(x,y) &= \int f(x,y)\sqrt{l(y,y)}dP_{XY}(x,y) \\ &\leq \|f\|_{L^2}\left[\int l(y,y)dP_Y(y)\right]^{1/2} \\ &< \infty.\end{aligned}$$

  Second, the evaluation functional $\delta_{y'} : \mathcal{H}_l \to \mathbb{R}$ is bounded, so it commutes with the Bochner integral, i.e.,

$$\begin{aligned}(R_l f)(y') &= \int \delta_{y'}\left[f(x,y)l_y\right] dP_{XY}(x,y) \\ &= \delta_{y'}\left[\int f(y)l_y dP_{XY}(x,y)\right],\end{aligned}$$

  so that $R_l f = \int f(x,y)l_y dP_{XY}(x,y) \in \mathcal{H}_l$.

- Let us show that R_l^* is the inclusion ι_l. For all $h \in \mathcal{H}_l$, $u \mapsto \langle h, u\rangle_{\mathcal{H}_l}$ is a bounded linear operator, and thus:

$$\begin{aligned}\langle h, R_l f\rangle_{\mathcal{H}_l} &= \left\langle h, \int f(x,y)l_y dP_{XY}(x,y)\right\rangle_{\mathcal{H}_l} \\ &= \int f(x,y)\left\langle h, l_y\right\rangle_{\mathcal{H}_k} dP_{XY}(x,y) \\ &= \int f(x,y)h(y)dP(y) \\ &= \langle \iota_l h, f\rangle_{\mathcal{F}}.\end{aligned}$$

- Thereby, we get the cross-covariance operators $C_{YX} = R_l R_k^* : \mathcal{H}_k \to \mathcal{H}_l$, and $C_{XY} = R_k R_l^* : \mathcal{H}_l \to \mathcal{H}_k$. By construction, $C_{YX} = C_{XY}^*$. Now,

$$\begin{aligned}\langle g, C_{YX} f\rangle_{\mathcal{H}_l} &= \langle g, R_l \iota_X f\rangle_{\mathcal{H}_l} \\ &= \langle \iota_Y g, \iota_X f\rangle_{\mathcal{F}} \\ &= \int f(x)g(y)dP_{XY}(x,y) \\ &= \mathbb{E}_{X,Y}\left[f(X)g(Y)\right].\end{aligned}$$

2 Bach's quadrature with importance sampling

This note is based on [1].

2.1 Setup

We would like to compute:

$$\rho_g[h] = \int h(x)g(x)d\rho(x) = \langle h, \mu_{k,g}\rangle_{\mathcal{H}_k}.$$

Here, $d\rho$ is a Borel probability measure, $h \in \mathcal{H}_k$, $g \in L_2(d\rho)$. Moreover, $\mu_{k,g} \in \mathcal{H}_k$ is the convolution with k, given by

$$\mu_{k,g} = \int k(\cdot, x)g(x)d\rho(x) = S_k g.$$

We consider estimators of form $\tilde{\rho}_g[h] = \sum_{i=1}^n \alpha_i h(x_i)$ or equivalently the estimators of μ_g of form $\tilde{\mu}_{k,g} = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$. From Cauchy-Schwarz

$$\sup_{\|h\|_{\mathcal{H}_k}\leq 1} |\tilde{\rho}_g[h] - \rho_g[h]| \quad = \quad \|\tilde{\mu}_{k,g} - \mu_{k,g}\|_{\mathcal{H}_k}.$$

Let us fix the choice $\{x_i\}$. We can find α by solving

$$\begin{aligned}\arg\min_\alpha \left\|\sum_{i=1}^n \alpha_i k(\cdot, x_i) - \mu_{k,g}\right\|_{\mathcal{H}_k}^2 + n\lambda\|\alpha\|_2^2 = \\ \arg\min_\alpha \alpha^\top (K + n\lambda I)\alpha - 2\alpha^\top \mu_{k,g}(\mathbf{x}) \qquad (1)\end{aligned}$$

solution of which is $\alpha = (K + n\lambda I)^{-1}\mu_{k,g}(\mathbf{x})$. These are in some sense "best weights" for fixed locations and also can be derived from Bayesian quadrature interpretation. Like this: treat h as a random function with a GP prior and compute its posterior given observations $h(\mathbf{x}) = (h(x_i))_{i=1}^n$: $\rho_g[h]$ is then a random variable induced by h. Posterior mean of $\rho_g[h]|h(\mathbf{x})$ is then:

$$\begin{aligned}\mathbb{E}[\rho_g[h]|h(\mathbf{x})] &= \int \left[\int h(x)g(x)d\rho(x)\right] p(h|h(\mathbf{x}))dh \\ &= \int \left[\int h(x)p(h|h(\mathbf{x}))dh\right] g(x)d\rho(x) \\ &= \rho_g[\mathbb{E}[h|h(\mathbf{x})]] \\ &= \underbrace{\mu_{k,g}(\mathbf{x})^\top (K + n\lambda I)^{-1}}_{\alpha^\top} h(\mathbf{x}).\end{aligned}$$

Denote by $\mathcal{T}_k^{1/2}$ the isometric isomorphism between $L_2(d\rho)$ and $\mathcal{H}_k$. Deliberately making a different notation since $T_k^{1/2}$ will denote the operator on $L_2(d\rho)$. Note that $\mathcal{T}_k^{-1/2}$ is well defined and $\mathcal{T}_k^{-1/2} = T_k^{-1/2} I_k$ (with some abuse of notation since $T_k^{-1/2}$ is not defined everywhere on $L_2(d\rho)$. Denote $\psi(\cdot, x) = \mathcal{T}_k^{-1/2} k(\cdot, x)$. Then

$$\langle \psi(\cdot, x), \psi(\cdot, y)\rangle_{L_2(d\rho)} = k(x, y). \qquad (2)$$

Note that

$$\begin{aligned}\mathcal{T}_k^{-1/2}\mu_{k,g} &= T_k^{-1/2} I_k S_k g \\ &= T_k^{1/2} g \\ &= \int \psi(\cdot,x) g(x) d\rho(x)\end{aligned}$$

$\psi(x,y)$ is a "rougher" positive definite kernel whose integral operator is precisely $T_\psi = T_k^{1/2}$. If $k(x,y) = \sum \eta_j e_j(x) e_j(y)$, then $\psi(x,y) = \sum \sqrt{\eta_j} e_j(x) e_j(y)$. Here, $\{\eta_j, e_j\}$ are eigenvalue-eigenfunction pairs of T_k, i.e., $T_k e_j = \eta_j e_j$.

2.2 Φ-operator

We will draw samples $\{x_i\}_{i=1}^n \sim q d\rho$ (importance distribution with density q wrt $d\rho$). Also, we will aim for an estimator of the form $\tilde{\rho}_g[h] = \sum_{i=1}^n \frac{\beta_i}{\sqrt{q(x_i)}} h(x_i)$. Now, consider operator $\Phi : \mathbb{R}^n \to L_2(d\rho)$ defined as

$$\Phi\beta = \sum_{i=1}^n \beta_i q(x_i)^{-1/2} \psi(\cdot, x_i).$$

It adjoint is $\Phi^* : L_2(d\rho) \to \mathbb{R}^n$, $\Phi^* f = \left(q(x_i)^{-1/2} \langle \psi(\cdot, x_i), f \rangle_{L_2(d\rho)} \right)_{i=1}^n$. We claim that $\Phi^*\Phi$ is related to the kernel matrix. Indeed,

$$\begin{aligned}(\Phi^*\Phi\beta)_i &= q(x_i)^{-1/2} \left\langle \psi(\cdot, x_i), \sum_{j=1}^n \beta_j q(x_j)^{-1/2} \psi(\cdot, x_j) \right\rangle_{L_2(d\rho)} \\ &= \sum_{j=1}^n \beta_j q(x_j)^{-1/2} q(x_i)^{-1/2} k(x_i, x_j),\end{aligned}$$

so $[\Phi^*\Phi]_{ij} = \frac{k(x_i,x_j)}{\sqrt{q(x_i)q(x_j)}}$. On the other hand

$$\Phi\Phi^* f = \sum_{i=1}^n q(x_i)^{-1} \langle \psi(\cdot, x_i), f \rangle_{L_2(d\rho)} \psi(\cdot, x_i),$$

means that we can write

$$\Phi\Phi^* = \sum_{i=1}^n q(x_i)^{-1} \psi(\cdot, x_i) \otimes_{L_2(d\rho)} \psi(\cdot, x_i).$$

Another way to think about T_k is as $\int \left[\psi(\cdot, x) \otimes_{L_2(d\rho)} \psi(\cdot, x) \right] d\rho(x)$. Why is this? Because for every $f \in L_2(d\rho)$

$$\begin{aligned}\left[\int \left[\psi(\cdot, x) \otimes_{L_2(d\rho)} \psi(\cdot, x) \right] d\rho(x) \right] f &= \int \langle \psi(\cdot, x), f \rangle_{L_2(d\rho)} \psi(\cdot, x) \, d\rho(x) \\ &= \int \left[T_k^{1/2} f \right](x) \psi(\cdot, x) \, d\rho(x) \\ &= T_k^{1/2} T_k^{1/2} f = T_k f.\end{aligned}$$

Thus $\hat{T}_k = \frac{1}{n}\Phi\Phi^*$is an importance sampling estimator of T_k and is unbiased. Now, action of Φ^* on $f_{k,g} = \int \psi(\cdot,x)g(x)d\rho(x) = \mathcal{T}_k^{-1/2}\mu_{k,g}$ is

$$\begin{aligned}
\Phi^* f_{k,g} &= \left(q\left(x_i\right)^{-1/2}\left\langle \psi\left(\cdot,x_i\right), f_{k,g}\right\rangle_{L_2(d\rho)}\right)_{i=1}^n \\
&= \left(q\left(x_i\right)^{-1/2}\left\langle \mathcal{T}_k^{-1/2}k\left(\cdot,x_i\right), \mathcal{T}_k^{-1/2}\mu_{k,g}\right\rangle_{L_2(d\rho)}\right)_{i=1}^n \\
&= \left(q\left(x_i\right)^{-1/2}\left\langle k\left(\cdot,x_i\right), \mu_{k,g}\right\rangle_{\mathcal{H}_k}\right)_{i=1}^n \\
&= \left(q\left(x_i\right)^{-1/2}\mu_{k,g}\left(x_i\right)\right)_{i=1}^n.
\end{aligned}$$

Since [1] considers estimators of the form $\tilde{\rho}_g[h] = \sum_{i=1}^n \frac{\beta_i}{\sqrt{q(x_i)}}h(x_i)$ this results in an optimization problem slightly different from (1):

$$\begin{aligned}
\arg\min_\beta \left\|\sum_{i=1}^n \frac{\beta_i}{\sqrt{q(x_i)}}k(\cdot,x_i) - \mu_{k,g}\right\|_{\mathcal{H}_k}^2 + n\lambda\left\|\beta\right\|_2^2 = \\
\arg\min_\beta \beta^\top\left(\tilde{K}_q + n\lambda I\right)\beta - 2\beta^\top\left(q\left(x_i\right)^{-1/2}\mu_{k,g}\left(x_i\right)\right)_{i=1}^n,
\end{aligned}$$

with solution

$$\begin{aligned}
\beta &= \left(\tilde{K}_q + n\lambda I\right)^{-1}\left(q\left(x_i\right)^{-1/2}\mu_{k,g}\left(x_i\right)\right)_{i=1}^n \\
&= \left(\Phi^*\Phi + n\lambda I\right)^{-1}\Phi^* f_{k,g} \\
&= \frac{1}{n}\Phi^*\left(\hat{T}_k + \lambda I\right)^{-1} f_{k,g}.
\end{aligned} \tag{3}$$

where we set $f_{k,g} = \mathcal{T}_k^{-1/2}\mu_{k,g}$ and denoted $\tilde{K}_q = \Phi^*\Phi$, i.e., $\left[\tilde{K}_q\right]_{ij} = \frac{k(x_i,x_j)}{\sqrt{q(x_i)q(x_j)}}$

At optimality,

$$\begin{aligned}
\left\|\sum_{i=1}^n \frac{\beta_i}{\sqrt{q(x_i)}}k(\cdot,x_i) - \mu_{k,g}\right\|_{\mathcal{H}_k}^2 &= \left\|\sum_{i=1}^n \frac{\beta_i}{\sqrt{q(x_i)}}\psi(\cdot,x_i) - \mathcal{T}_k^{-1/2}\mu_{k,g}\right\|_{L_2(d\rho)}^2 \\
&= \left\|\Phi\beta - f_{k,g}\right\|_{L_2(d\rho)}^2 \\
&= \left\|\left(\hat{T}_k\left(\hat{T}_k + \lambda I\right)^{-1} - I\right)f_{k,g}\right\|_{L_2(d\rho)}^2 \\
&= \lambda^2\left\|\left(\hat{T}_k + \lambda I\right)^{-1} f_{k,g}\right\|_{L_2(d\rho)}^2. \\
&= \lambda\left\langle f_{k,g}, \lambda\left(\hat{T}_k + \lambda I\right)^{-2} f_{k,g}\right\rangle_{L_2(d\rho)}.
\end{aligned} \tag{4}$$

Moreover:

$$\begin{aligned} n\lambda \|\beta\|_2^2 &= n\lambda \left\langle \frac{1}{n}\Phi^* \left(\hat{T}_k + \lambda I\right)^{-1} f_{k,g}, \frac{1}{n}\Phi^* \left(\hat{T}_k + \lambda I\right)^{-1} f_{k,g} \right\rangle_{\mathbb{R}^n} \\ &= \lambda \left\langle f_{k,g}, \left(\hat{T}_k + \lambda I\right)^{-1} \hat{T}_k \left(\hat{T}_k + \lambda I\right)^{-1} f_{k,g} \right\rangle_{L_2(d\rho)}. \end{aligned} \tag{5}$$

Now, operators inside both inner products (4) and (5) are $\preceq \left(\hat{T}_k + \lambda I\right)^{-1}$.To see this, consider spectra: if we denote eigenvalues of $\hat{T}_k$ by $\{\hat{\eta}_j\}$, then

$$\begin{aligned} \frac{\lambda}{(\eta_j+\lambda)^2} \leq \frac{1}{\eta_j+\lambda},\ \forall j \quad &\Rightarrow \quad \lambda \left(\hat{T}_k + \lambda I\right)^{-2} \preceq \left(\hat{T}_k + \lambda I\right)^{-1} \\ \frac{\eta_j}{(\eta_j+\lambda)^2} \leq \frac{1}{\eta_j+\lambda},\ \forall j \quad &\Rightarrow \quad \left(\hat{T}_k + \lambda I\right)^{-1} \hat{T}_k \left(\hat{T}_k + \lambda I\right)^{-1} \preceq \left(\hat{T}_k + \lambda I\right)^{-1}. \end{aligned}$$

Thus, to bound $\left\| \sum_{i=1}^n \frac{\beta_i}{\sqrt{q(x_i)}} k(\cdot, x_i) - \mu_{k,g} \right\|_{\mathcal{H}_k}^2 + n\lambda \|\beta\|_2^2$ it suffices to bound $\lambda \left\langle f_{k,g}, \left(\hat{T}_k + \lambda I\right)^{-1} f_{k,g} \right\rangle_{L_2(d\rho)} = \lambda \left\langle g, T_k^{1/2} \left(\hat{T}_k + \lambda I\right)^{-1} T_k^{1/2} g \right\rangle_{L_2(d\rho)}$.

2.3 Choice of q

Consider

$$\begin{aligned} d(\lambda) &= \operatorname{Tr}\left[T_k \left(T_k + \lambda I\right)^{-1}\right] \\ &= \sum_j \frac{\eta_j}{\eta_j + \lambda} \\ &= \sum_j \left\langle \left(T_k + \lambda I\right)^{-1} e_j, T_k e_j \right\rangle_{L_2(d\rho)} \\ &= \sum_j \left\langle \left(T_k + \lambda I\right)^{-1} e_j, \int \langle \psi(\cdot, x), e_j \rangle_{L_2(d\rho)} \psi(\cdot, x)\, d\rho(x) \right\rangle_{L_2(d\rho)} \\ &= \int \sum_j \left\langle \left(T_k + \lambda I\right)^{-1} e_j, \langle \psi(\cdot, x), e_j \rangle_{L_2(d\rho)} \psi(\cdot, x) \right\rangle_{L_2(d\rho)} d\rho(x) \\ &= \int \left\langle \left(T_k + \lambda I\right)^{-1} \psi(\cdot, x), \psi(\cdot, x) \right\rangle_{L_2(d\rho)} d\rho(x) \\ &\leq d_{\max}(q, \lambda) := \sup_x \frac{\left\langle \left(T_k + \lambda I\right)^{-1} \psi(\cdot, x), \psi(\cdot, x) \right\rangle_{L_2(d\rho)}}{q(x)}. \end{aligned}$$

To simplify, we have:

$$\left\langle \psi(\cdot,x), (T_k+\lambda I)^{-1}\psi(\cdot,x)\right\rangle_{L_2(d\rho)} =$$
$$\left\langle I_k k(\cdot,x), T_k^{-1/2}(T_k+\lambda I)^{-1}T_k^{-1/2} I_k k(\cdot,x)\right\rangle_{L_2(d\rho)} =$$
$$\left\langle \sum_j \eta_j e_j(x)e_j, \sum_i \eta_i e_i(x) T_k^{-1/2}(T_k+\lambda I)^{-1}T_k^{-1/2} e_i\right\rangle_{L_2(d\rho)} =$$
$$\left\langle \sum \eta_j e_j(x)e_j, \sum \frac{1}{\eta_i+\lambda} e_i(x)e_i\right\rangle_{L_2(d\rho)} =$$
$$\sum_j \frac{\eta_j}{\eta_j+\lambda} e_j^2(x) =$$
$$\check{k}_\lambda(x,x),$$

where $\check{k}_\lambda$ is the kernel of operator $T_k(T_k+\lambda I)^{-1}$ (marginal kernel from DPP). The bound on error in Proposition 1 of [1] requires to have $n = \mathcal{O}(d_{\max}\log(d_{\max}))$. Therefore, we seek to minimize $d_{\max}(q,\lambda) = \sup_x \frac{\check{k}_\lambda(x,x)}{q(x)}$ so we set:

$$q(x) \propto \check{k}_\lambda(x,x),$$

for which $d_{\max}(q,\lambda) = \int \check{k}_\lambda(x,x)d\rho(x) = d(\lambda)$.

3 Notes following meeting on July 28, 2015.

Some notation. We want to compute $\int h(x)d\rho(x)$, where $h \in \mathcal{H}_k$ for some kernel k and some probability measure ρ on domain $\mathcal{X}$. We need the following:

1. A way to compute/approximate the optimal importance density (wrt ρ) $q(x) \propto \check{k}_\lambda(x,x)$ where $\check{k}_\lambda(x,x) = \sum_m \frac{\eta_m}{\eta_m+\lambda} e_m^2(x)$, provided that we have Mercer expansion $k(x,y) = \sum_m \eta_m e_m(x)e_m(y)$ of k *with respect to* ρ, i.e., $\{e_m\}$ are ONS in $L_2(\rho)$.
2. A way to compute/approximate mean embedding $\mu_\rho = \int k(\cdot,x)d\rho(x)$ (needed for computing weights).

Methods to compare.

1. Sampling from $d\rho(x)$ and weighting appropriately
2. Sampling from $\frac{k(x,x)}{\int k(x,x)d\rho(x)}d\rho(x)$ and weighting appropriately
3. Sampling from $q(x)d\rho(x)$ and weighting appropriately
4. Sampling from DPP and weighting appropriately

Questions to answer.

1. Is improvement from 3 to 4 larger than from 1-2 to 3?
2. How do these improvements change with increasing dimensionality

3.1 Details of examples

Example 1: Sobolev spaces and Beta distributions We consider domain $\mathcal{X} = [0, 1]$. Kernel is

$$\begin{aligned} k(x,y) &= \sum_{m=0}^{\infty} m^{-2s} \cos\left(2\pi m\left(x-y\right)\right) \\ &= \sum_{m=0}^{\infty} m^{-2s} \left[\cos\left(2\pi mx\right)\cos\left(2\pi my\right) + \sin\left(2\pi mx\right)\sin\left(2\pi my\right)\right]. \end{aligned}$$

Consider the simple case of ρ being Uniform on $[0, 1]$: $d\rho(x) = dx$. Then

1. Because

$$\begin{aligned} \int_0^1 \cos\left(2\pi mx\right)\sin\left(2\pi m'x\right) dx &= 0 \\ \int_0^1 \cos\left(2\pi mx\right)\cos\left(2\pi m'x\right) dx &= \frac{1}{2}\delta(m-m') \\ \int_0^1 \sin\left(2\pi mx\right)\sin\left(2\pi m'x\right) dx &= \frac{1}{2}\delta(m-m'), \end{aligned}$$

   we have the desired Mercer expansion with eigenfunctions $\left\{\sqrt{2}\cos\left(2\pi mx\right), \sqrt{2}\sin\left(2\pi mx\right)\right\}$ and corresponding eigenvalues $\eta_m = \frac{1}{2}m^{-2s}$ (each with multiplicity 2), so can compute q:

$$\check{k}_\lambda(x,x) = \sum_m \frac{\eta_m}{\eta_m + \lambda} e_m^2(x) = \sum_m \frac{m^{-2s}/2}{m^{-2s}/2 + \lambda} \propto 1, \tag{6}$$

   which is just uniform ($qd\rho = d\rho$)

2. Because $\forall y$

$$\begin{aligned} \mu_p(y) = \int k(y,x) d\rho(x) &= 1 + \int_0^1 \left(\sum_{m=1}^{\infty} m^{-2s}\cos\left(2\pi m\left(x-y\right)\right)\right) dx \\ &= 1 + \sum_{m=0}^{\infty} m^{-2s} \underbrace{\int_0^1 \cos\left(2\pi m\left(x-y\right)\right) dx}_{=0} \\ &= 1, \end{aligned}$$

   mean embedding is also trivial to compute.

Remark 1. Is it possible to consider DPP with "true marginal kernel" $\check{k}_\lambda$ which we have analytically in this case (supported on all of $[0,1]$)? We have eigenvalues analytically so can sample Bernoullis and can explicitly evaluate eigenfunctions.

If now $d\rho$ is no longer uniform, neither 1 nor 2 are trivial anymore. Say $d\rho(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}dx$, i.e., Beta$(a,b)$ distribution.

1. Finding Mercer's expansion now involves finding eigenfunctions of

$$f \quad \mapsto \quad \frac{1}{B(a,b)} \sum_{m=0}^{\infty} m^{-2s} \int_0^1 f(x) \cos\left(2\pi m \left(x - \cdot\right)\right) x^{a-1}(1-x)^{b-1} dx,$$

   which does not look obvious. Alternatively, we could use a generic procedure to approximate marginal kernel $\check{k}_\lambda$ by using the fact that it is the kernel of the operator $T_k \left(T_k + \lambda I\right)^{-1}$. First we would sample $\{z_i\}_{i=1}^N \overset{iid}{\sim} d\rho$, and form a kernel matrix K on these points. $\hat{K}_\lambda = K\left(K + \lambda I\right)^{-1}$ would then be the approximate kernel matrix of the marginal kernel and so we can subsample from $\{z_i\}$ with probabilities proportional to $\left[\hat{K}_\lambda\right]_{ii}$ (which results approximately in a sample from $qd\rho$). This is exactly sampling marginally from a DPP supported on $\{z_i\}$, so there is effectively no additional cost incurred by sampling from the full DPP here. Computing weights using Bach's approach in 3 also involves computing $q(x_i)$ for the subsampled $x_i = z_\ell$ for some $\ell \in \{1, \ldots, N\}$. Here, we can also substitute $\hat{q}(x_i) = \left[\hat{K}_\lambda\right]_{\ell\ell} / \mathrm{Tr}\left[\hat{K}_\lambda\right]$.

2. Finding mean embedding is also not obvious. Again, we can simply work with empirical mean embedding on a sample $\{z_i\} \overset{iid}{\sim} \rho$, $\hat{\mu}_\rho = \frac{1}{N}\sum_{i=1}^N k(\cdot, z_i)$.

Example 2: Gaussian RBF kernel and Gaussian ρ Gaussian RBF kernel on $\mathbb{R}^p$: $k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$. For the case where ρ is Gaussian, we can find Mercer's expansion analytically. If we consider univariate case and ρ is $\mathcal{N}(0, \tau^2)$ then we obtain expansion [3, Section 4.3.1] with

$$\eta_m = \sqrt{\frac{2a}{A}} B^m, \; e_m(x) = \exp\left(-\left(c-a\right)x^2\right) H_m\left(\sqrt{2c}x\right), \quad m = 0, 1, \ldots \tag{7}$$

where $a = 1/4\tau^2$, $b = 1/2\sigma^2$, $c = \sqrt{a^2 + 2ab}$, $A = a + b + c$, $B = b/A$ and $H_m = (-1)^m \exp\left(x^2\right) \frac{d^m}{dx^m} \exp\left(-x^2\right)$ is the m-th order Hermite polynomial. If working with isotropic kernels and distributions, readily generalizes to the multivariate case, since one would simply work with the products of eigenvalues and eigenfunctions across dimensions. Due to the geometric decrease of eigenvalues, truncating sum $\check{k}_\lambda(x,x) = \sum_{m=0}^\infty \frac{\eta_m}{\eta_m + \lambda} e_m^2(x)$ should give a good estimate of the optimal importance density $q(x)$.

Remark 2. What does importance distribution $qd\rho$ look like in this case? Is it very far from the original Gaussian? If it looks very different, and there

are substantial improvements for using q rather than ρ, an unbiased estimator of $\check{k}_\lambda(x,x)$ can be constructed for each x, so one could sample from q with pseudo-marginal MCMC.

Mean embedding amounts to solving a Gaussian integral so can also be computed analytically. For simplicity, if ρ is $\mathcal{N}(0,\tau^2 I)$, then

$$\mu_\rho(x) = \left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)^{p/2} \exp\left(-\frac{\|x\|_2^2}{2\left(\sigma^2+\tau^2\right)}\right).$$

Example 3: Brownian covariance kernel (and Gaussian or uniform ρ?) Here $k(x,y) = \frac{1}{2}\left(\|x\|_2^{2H} + \|y\|_2^{2H} - \|x-y\|_2^{2H}\right)$, for $H \in (0,1]$ (recovers linear kernel for $H=1$ and covariance of standard Brownian motion for $H=1/2$). Not translation invariant. See [2, p.319] for its RKHS. Also, not a bounded kernel, so mean embeddings are only defined for measures with finite fractional moments $\int \|x\|_2^H \, d\rho(x)$. Can consider a compact domain to get around this, e.g., $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\| \leq 1\}$. Mercer's expansions should be known in some cases, e.g., $H=1/2$, $p=1$, $\mathcal{X}=[0,1]$ and ρ uniform (note $k(x,y)=\min(x,y)$ in this case) should have

$$\eta_m = \frac{1}{\left(m+\frac{1}{2}\right)^2 \pi^2}, \; e_m(x) = \sqrt{2}\sin\left(\left(m+\frac{1}{2}\right)\pi x\right), \quad m = 0,1,\ldots \tag{8}$$

Example 4: Vovk's infinite polynomial kernel (and Gaussian ρ?) Here $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\| \leq c < 1\}$ and $k(x,y) = \left(1 - x^\top y\right)^{-\alpha}$, so also not translation invariant. Due to flat spectrum, expected to have poor generalization properties[4, p.113], so not often used in practice. However, it is known to be universal [5].

References

[1] F. Bach. On the equivalence between quadrature rules and random features. Technical Report HAL-01118276, 2015.

[2] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[3] C.E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[4] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[5] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2001.