1 Support Vector Machines

Support vector machine is a kernelized optimal margin linear classifier (original paper [1] and a nice summary [2]). It distinguishes itself from classifiers minimizing empirical risk as it favors classifier which makes confident predictions. For binary classification problem $\mathcal{Y} = \{-1, +1\}$, we are interested in finding a linear decision bondary, parameterized by $w \in \mathbb{R}^d, b \in \mathbb{R}$, that separates the training data points by maximizing the worst case distance (margin) of each data point to the decision boundary. We first assume that training set can be linearly separated. Given dataset $\{(x_i, y_i)\}_{i=1}^n$, we are interested in solving the following quadratic programming problem,

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \|w\|^2 \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \cdots, n \end{aligned}$$

To derive the dual problem, we write the Lagrangian,

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \alpha_i \left[1 - y_i(w^T x_i + b)\right] \tag{1}$$

where $\alpha = \{\alpha_i\}_{i=1}^n$ are the dual variables. Solve for $\inf_{w,b} \mathcal{L}(w, b, \alpha)$ to arrive at the dual objective. In particular, first order optimality condition gives $w = \sum_{i=1}^n \alpha_i y_i x_i$ and it must be that $0 = \sum_{i=1}^n \alpha_i y_i$. Therefore,

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t. } \alpha_i \geq 0 \quad i = 1, 2, \cdots, n && \text{(dual feasibility)} \\ &\sum_{i=1}^n \alpha_i y_i = 0 && \text{(from } \nabla_b \mathcal{L} = 0\text{)} \end{aligned}$$

The dual can be solved more efficiently than the primal problem using coordinate descent. The decision rule is linear w.r.t support vectors (those x_i right on margin with $\alpha_i > 0$)

$$\hat{y}(x) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i x_i^T x + b\right) \quad \text{for} \quad b = y_i - \sum_{j=1}^n \alpha_j y_j x_j^T x_i \tag{2}$$

for any support vector x_i. We observe that optimization as well as prediction uses input vectors via dot products only. We are motivated to use feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ to map input vectors to a higher dimensional possibly infinite feature space in hope that the lifted space is linearly separable. The kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ allows us to compute dot products $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ efficiently and represents a notion of similarity over two instance of aribtrary objects, e.g. vectors in $\mathbb{R}^n$, graphs, texts. We can substitute k whenever inner product is used and arrive at a optimal margin classifier over implicitly defined nonlinear feature mapping ϕ. In case when training dataset is not linearly separable, we can introduce slack variable $\{\xi_i\}_{i=1}^n$ where $x_i \geq 0$ to relax the inequality constraints and penalize misclassified or within margin points with $C \sum_{i=1}^n \xi_i$ for some $C \in \mathbb{R}$. In this case, we have the following Lagrangian,

$$\mathcal{L}(w, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha \left(1 - y_i(\langle w, \phi(x_i) \rangle_{\mathcal{H}} + b) - \xi_i\right) + \sum_{i=1}^n \beta_i \xi_i \tag{3}$$

First order condition $0 = \frac{\partial}{\partial \xi_i}\mathcal{L} = C - \alpha_i + \beta_i$ together with dual feasibility $\beta_i \geq 0$ yield $\alpha_i \leq C$ for all $i = 1, 2, \cdots, n$. Therefore, we optimize for the following dual problem,

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2}\alpha^T Q \alpha - 1^T \alpha \\ \text{s.t. } & y^T \alpha = 0 \\ & 0 \leq \alpha_i \leq C \quad i = 1, 2, \cdots, n \end{aligned}$$

where $Q \in \mathbb{R}^{n \times n}$ with $Q_{ij} = \alpha_i \alpha_j k(x_i, x_j)$, with optimal decision rule as

$$\hat{y}(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b\right) \quad \text{for} \quad b = y_i - \sum_{j=1}^{n} \alpha_j y_j k(x_j, x_i) \tag{4}$$

for any support vector i, i.e. $0 < \alpha_i < C$.

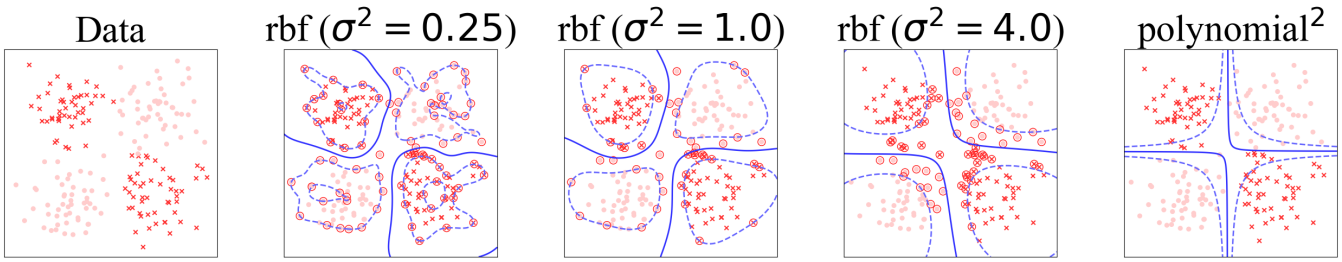

Figure 1: SVM on simulated 2D Gaussian with degree two polynomial kernel and radial basis function kernel with varying bandwidth. Larger bandwidth corresponds to smoother decision boundary and in the limit approaches the decision boundary of a linear kernel

2 Reproducing Kernel Hilbert Space

[3] provides a rigorous introduction to RKHS while [4] provides pretty good intuitions. Intuitively, RKHS can be considered as (1) a space of well-behaved functions whose smoothness is determined by its kernel. This view is useful when trying to think about regularization in terms of $\|f\|_{\mathcal{H}}^2$ (2) an inner product space for features. This view is useful when trying to apply kernel trick. (3) a space of functions spanned by representors $\{k(x, \cdot)\}_{x \in \mathcal{X}}$. We think of functions in RKHS as a simple function class, i.e. linear with respect to $\phi(x)$, and also flexible due to various choice of kernels. (4) a space of functions spanned by countably many eigenfunctions of the integral operator of kernel.

Definition 1. *(Hilbert space) A Hilbert space is a complete inner product space, i.e.* $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$

Definition 2. *(Separable Hilbert space) A Hilbert space is separable if it has a countable basis*

Definition 3. *(Reproducing kernel Hilbert space) A Hilbert space* $\mathcal{H}$ *of functions* $f : \mathcal{X} \to \mathbb{R}$ *is a RKHS if its evaluation functional,* $\delta_x : \mathcal{H} \to \mathbb{R}$ *where* $\delta_x(f) = f(x)$ *is continuous* $\forall\, x \in \mathcal{X}$.

Intuitively, RKHS is a space of well-behaved functions. In particular, norm convergence in $\mathcal{H}$ yield pointwise convergence, i.e. $\lim_{n \to \infty} \|f_n - f\| = 0$ implies $f_n \to f$. Note evaluation functional is linear. The condition that δ_x is continuous is equivalent to δ_x be bounded [3]. Regularization of the form $\|f\|_{\mathcal{H}}$ leads to regularization on function values.

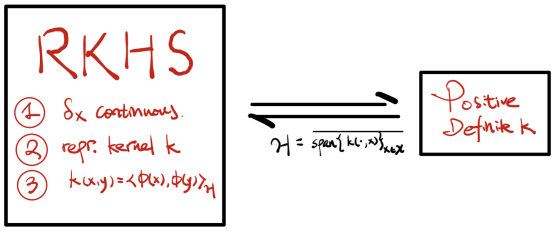

Figure 2: Equivalent views of RKHS

Definition 4. *(Reproducing kernel) Let $\mathcal{H}$ be Hilbert space defined on non-empty $\mathcal{X}$, then a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if*

1. $\forall\, x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$
2. $\forall\, x \in \mathcal{X}$, $\forall\, f \in \mathcal{H}$, $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$ *(reproducing property)*

In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k(y, \cdot), k(y, \cdot)\rangle_{\mathcal{H}}$.

Intuitively, the definition of reproducing kernel implies that $k(x, \cdot)$ is (1) a high dimensional representer of x, and (2) as a representer of evaluation for any function in $\mathcal{H}$ on data point x. We also note that we can always find a feature map associated with a reproducing kernel, namely the canonical feature map $\phi : x \to k(x, \cdot)$, and represent k as an inner product in feature space $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$. It turns out that RKHS has $\mathcal{H}$ is a RKHS if and only if $\mathcal{H}$ has a reproducing kernel. Definition of RKHS based on continuous evaluation functional is equivalent to existence of a (unique) reproducing kernel.

Definition 5. *(Positive definite functions) A symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if for all $n \geq 1$ and for all $(a_1, \cdots, a_n) \in \mathbb{R}^n$ for all $(x_1, \cdots, x_n) \in \mathcal{X}^n$,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0 \tag{5}$$

If you consider a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, k is positive definite if any kernel matrix $K \in \mathbb{R}^{n\times n}$ where $K_{ij} = k(x_i, x_j)$ is positive definite, i.e. $K \succeq 0$.

Definition 6. *(Kernel defined via feature map) Let $\mathcal{X}$ be non-empty set. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a real Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall\, x, y \in \mathcal{X}$, $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$. ϕ is said to be the feature map and $\mathcal{H}$ be the feature space.*

Note there maybe more than one feature map yield for any one kernel. Intuitively, a kernel is a function that can be represented as inner product. It turns out that RKHS with reproducing kernel k is positive definite,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left\langle k(x_i, \cdot), k(x_j, \cdot)\right\rangle = \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{i=1}^{n} a_i \phi(x_i)\right\rangle_{\mathcal{H}} \geq 0 \tag{6}$$

and, conversely, we can show that for every positive definite function k there is an unqiue RKHS whose reproducing kernel is k (Moore-Aronsajn). Intuitively, we construct a RKHS $\mathcal{H} = \overline{\mathcal{H}_0}$, the completion of a pre-RKHS space $\mathcal{H}_0 = \text{span}(\{k(x, \cdot)\}_{x\in\mathcal{X}})$. In particular, the choice of inner product

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) \tag{7}$$

for $f = \sum_i \alpha_i k(x_i, \cdot), g = \sum_j \beta_j k(y_j, \cdot)$ makes $\mathcal{H}_0$ a valid pre-RKHS. Intuitively, this construction implies that RKHS is a space spanned by representers $\{k(x, \cdot)\}_{x \in \mathcal{X}}$.

Definition 7. *(Integral Operator [3, 5]) Let X be compact Hausdorff space and μ be finite Borel measure over $\mathcal{X}$. Let k be a continuous kernel. The integral operator of kernel k is $T_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ defined*

$$T_k f = \int_{\mathcal{X}} k(x, \cdot) f(x) d\mu(x) \tag{8}$$

If k is a positive definite function, then T_k is self-adjoint ($\langle f, T_k g \rangle = \langle T_k f, g \rangle \ \forall \ f, g \in L_2(\mathcal{X}, \mu)$) positive ($\langle f, T_k f \rangle \geq 0 \ \forall \ f \in L_2(\mathcal{X}, \mu)$) and compact.

Theorem 1. *(Mercer's theorem) Assume $\mathcal{X}, k, T_k$ defined as previous. Suppose for all $f \in L_2(\mathcal{X}, \mu)$, we have $\int_{\mathcal{X}} k(x, y) f(x) f(y) dx dy \geq 0$. Then there exists a countable orthonormal basis $\{e_i\}$ of $L_2(\mathcal{X}, \mu)$ consisting of eigenfunctions of T_k whose corresponding eigenvalues $\{\lambda_i\}$ are non-negative. Then,*

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_j(y) \tag{9}$$

for all $x, y \in \mathcal{X}$. The convergence is absolute and uniform.

Mercer's theorem provides a construction of rkhs using eigenfunctions of integral operator. In particular $\mathcal{H} = \{\sum_i a_i e_i\}$ satisfying some integrability constraints with $\left\langle \sum_i a_i e_i, \sum_j b_j e_j \right\rangle_{\mathcal{H}} = \sum_i \frac{a_i b_i}{\lambda_i}$ is the RKHS with k as reproducing kernel.

Mercer's representation gives another feature map (not canonical) for kernel k,

$$k(x, y) = \left\langle \sqrt{\lambda_i} e_i(x), \sqrt{\lambda_i} e_i(y) \right\rangle_{\ell^2} \tag{10}$$

where $\varphi(x) = \left[\sqrt{\lambda_1} e_1(x), \sqrt{\lambda_2} e_2(x), \cdots\right]$.

3 Kernel Approximation

Theorem 2. *(Bochner's theorem) A bounded continous shift-invariant kernel $k(x, x') = k(x - x') = k(\delta)$ on $\mathbb{R}^d$ is positive definite if and only if there exists a finite non-negative Borel measure Λ on $\mathbb{R}^d$ such that*

$$k(\delta) = \int_{\mathbb{R}^d} e^{i\langle \omega, \delta \rangle} d\Lambda(\omega) \tag{11}$$

We can think of Λ as the fourier transform of $k(\delta)$, i.e. $k = \mathcal{F}[\Lambda]$. If kernel is normalized such that $k(0) = 1$, then Λ will be a probability measure and k corresponds to its characteristic function $k = \mathbb{E}\left[e^{i\langle \omega, \cdot \rangle}\right]$. If k is Gaussian rbf kernel $k(x, x') = \exp(-\frac{1}{2\sigma^2} \|x - x'\|^2)$, then $\Lambda \sim \mathcal{N}(0, \frac{1}{\sigma^2} I)$.

Computing the kernel matrix is $\mathcal{O}(n^2)$ while inversion requires $\mathcal{O}(n^3)$, infeasible for large n. [6] proposes to map data into a relatively low-dimensional randomized feature space via $z : \mathbb{R}^d \to \mathbb{R}^L$ such that inner products between transformed pair of points approximates their kernel evaluation, i.e. $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \approx z(x)^T z(y)$. Let $p(\omega)$ be Fourier transform of a continuous shift-invariant kernel $k(\delta)$, whose existence is guaranteed by Bochner's theorem. Let $z_\omega : \mathbb{R}^d \to \mathbb{R}$ $z_\omega(x) = \sqrt{2} \cos(\omega^T x + b)$ be the randomized feature map where $\omega \sim p(\omega)$, $b \sim \text{Unif}[0, 2\pi]$. Then for any $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}\left[z_\omega(x) z_\omega(y)\right] &= \mathbb{E}\left[\sqrt{2} \cos(\omega^T x + b) \sqrt{2} \cos(\omega^T y + b)\right] \\
&= \mathbb{E}\left[\cos(\omega^T (x + y) + 2b)\right] + \mathbb{E}\left[\cos(\omega^T (x - y))\right] \\
&= \mathbb{E}\left[\cos(\omega^T (x - y))\right] \\
&= \mathbb{E}\left[e^{i\omega^T (x - y)}\right] = k(x - y)
\end{aligned}$$

where we have used fact $2\cos(a)\cos(b) = \cos(a+b)+\cos(a-b)$ and that $\int e^{i\omega^T x}p(\omega)d\omega = \int \cos(\omega^T x)p(\omega)d\omega$ when $p(\omega)$ is real-valued. We can estimate $k(\delta)$ via Monte Carlo estimates. Now define random Fourier feature map $z : \mathbb{R}^d \to \mathbb{R}^L$ where L is the number of particles,

$$z(x) = \frac{1}{\sqrt{L}}(z_{\omega_1}(x), \cdots, z_{\omega_L}(x)) = \sqrt{\frac{2}{L}}(\cos(\omega_1^T x + b_1), \cdots, \cos(\omega_L^T x + b_L)) \tag{12}$$

where $(\omega_1, \cdots, \omega_L) \overset{iid}{\sim} p(\omega)$ and $(b_1, \cdots, b_L) \overset{iid}{\sim} \text{Unif}[0, 2\pi]$. Then,

$$z(x)^T z(y) = \frac{1}{L}\sum_{l=1}^{L} z_{\omega_l}(x)z_{\omega_l}(y) \quad \xrightarrow{p} \quad \mathbb{E}\left[z_\omega(x)z_\omega(y)\right] = k(x-y) \tag{13}$$

where convergence is superlinear via Hoeffding's Inequality using the fact that $z_\omega(x)z_\omega(y) \in [-2, 2]$, i.e. $\mathbb{P}\left[z(x)^T z(y) - k(x-y) > \epsilon\right] \leq \exp(-\frac{L\epsilon^2}{8})$. Given data $\{x_i\}_{i=1}^n$, and let $Z \in \mathbb{R}^{n\times L}$ where i-th row is $z(x_i)$. We can then approximate the kernel matrix $K \in \mathbb{R}^{n\times n}$ via low-rank decomposition ZZ^T since $(ZZ^T)_{ij} = z(x_i)^T z(x_j)$. Intuitively, random Fourier feature map projects a pair of data points x, y to a set of random lines in $\mathbb{R}^d$ parameterized by $\{\omega_l, b_l\}_{l=1}^L$, and then pass the resulting projection element-wise through a sinusoidal function. If ω_l are drawn from a particular probability distribution, then the inner products of resulting vectors in $\mathbb{R}^d$ approximates $k(x, y)$ well. Figure (3) illustrates the effect of varying number of particles L when trying to approximate the kernel matrix and its effect on classification.

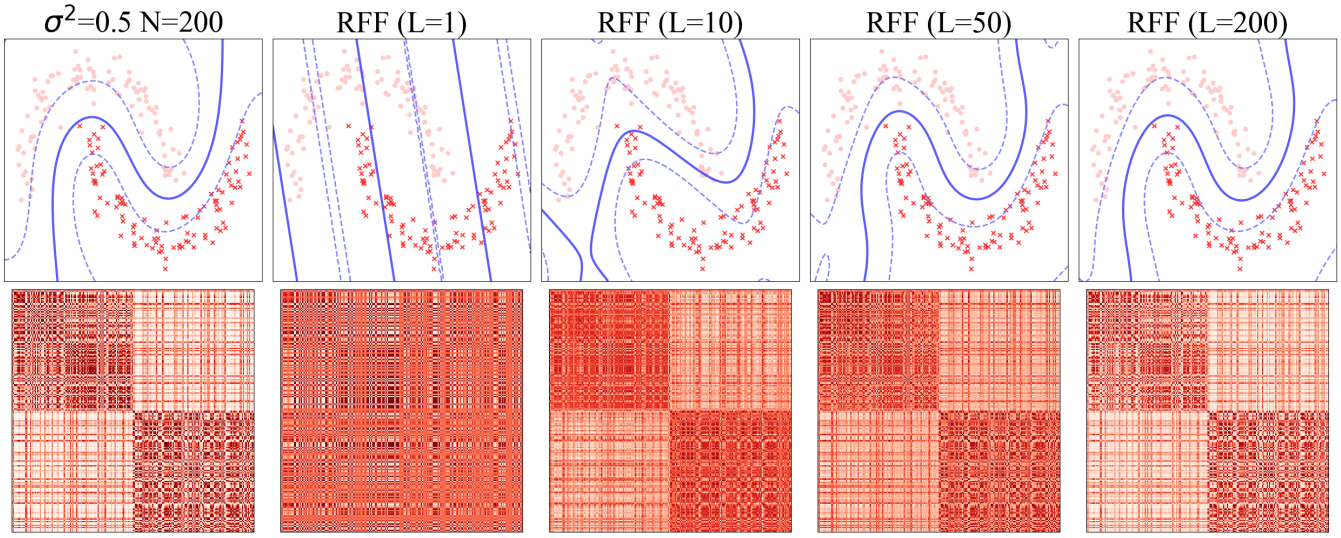

Figure 3: Radial basis kernel matrix ($\sigma^2 = 0.5$) and its random Fourier approximation (varying L) over moon dataset. Higher number of particles L makes kernel matrix approximation better.

4 Kernel Mean Embedding of Distributions

[5] chapter 3 provides a really clear generalization from feature map over points $x \in \mathcal{X}$ to measures over the measurable space $(\mathcal{X}, \mathcal{F})$ where $\mathcal{F}$ is sigma algebra of $\mathcal{X}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel k. Consider the mean map $\mu : \mathcal{P} \to \mathcal{H}$

$$\mu(\mathbb{P}) = \int k(x, \cdot)d\mathbb{P}(x) \tag{14}$$

and write $\mu_{\mathbb{P}} := \mu(\mathbb{P})$. Intuitively, the mean map $\mu_{\mathbb{P}}$ is a representation of probability measure $\mathbb{P}$ in $\mathcal{H}$. Not rigorously, we noticed that we can compute $\mathbb{E}_{\mathbb{P}}[f(X)]$ as inner products in $\mathcal{H}$,

$$\int f d\mathbb{P} = \int \langle f, k(x,\cdot)\rangle_{\mathcal{H}} \, d\mathbb{P}(x) = \left\langle f, \int k(x,\cdot) d\mathbb{P}(x) \right\rangle_{\mathcal{H}} = \langle f, \mu_{\mathbb{P}} \rangle$$

In effect, we are interchanging expectation and inner products, $\mathbb{E}[\langle f, k(X,\cdot)\rangle_{\mathcal{H}}] = \langle f, \mathbb{E}[k(X,\cdot)]\rangle_{\mathcal{H}}$.

Lemma 1. *If* $\mathbb{E}\left[\sqrt{k(X,X)}\right] < \infty$*, then* $\mu_{\mathbb{P}} \in \mathcal{H}$ *and* $\mathbb{E}_{\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ *for all* $f \in \mathcal{H}$*.*

Proof. For any $\mathbb{P} \in \mathcal{P}$, define a linear functional $T_{\mathbb{P}} : \mathcal{H} \to \mathbb{R}$ where $T_{\mathbb{P}}(f) = \mathbb{E}_{\mathbb{P}}[f(X)]$ as the operation of taking expectation with respect to function $f \in \mathcal{H}$ over $\mathbb{P}$. We show $T_{\mathbb{P}}$ is bounded (to use Rietz)

$$|T_{\mathbb{P}} f| = |\mathbb{E}[f(X)]| \leq \mathbb{E}[|f(X)|] = \mathbb{E}[|\langle f, k(X,\cdot)\rangle|] \leq \mathbb{E}\left[\sqrt{k(X,X)}\, \|f\|_{\mathcal{H}}\right] < \infty$$

By Rietz representation theorem, exists $g \in \mathcal{H}$ such that $T_{\mathbb{P}} f = \langle f, g\rangle_{\mathcal{H}}$. The choice of $f = k(x,\cdot)$ implies.

$$g(x) = \langle k(x,\cdot), g\rangle_{\mathcal{H}} = T_{\mathbb{P}}[k(x,\cdot)] = \int k(x,x') d\mathbb{P}(x') = \mu_{\mathbb{P}}(x)$$

Therefore, $\mu_{\mathbb{P}} \in \mathcal{H}$ is guaranteed by Rietz and we can re-write the Rietz's result and considered it as a reproducing property for $T_{\mathbb{P}}$, i.e. $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$. □

In effect, what this lemma states is under some mild assumptions, the mean map $\mu_{\mathbb{P}}$ ends up in RKHS $\mathcal{H}$ and that we can compute the expectation of any function in RKHS with respect to distribution $\mathbb{P}$ by computing inner product between the function f and the mean map $\mu_{\mathbb{P}}$ in $\mathcal{H}$. This is analogous to previous definition of a reproducing kernel, that the cannical feature map ends up in RKHS $k(x,\cdot) \in \mathcal{H}$ and we can evaluate the evaluation functional as an inner product in RKHS, i.e. $\delta_x(f) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$.

Furthermore, we note that inner products of mean maps is expectation of kernel, i.e. for $X \sim \mathbb{P}, Y \sim \mathbb{Q}$,

$$\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} = \mathbb{E}[\mu_{\mathbb{Q}}(X)] = \mathbb{E}[k(X,Y)] \tag{15}$$

where we have used fact that evaluating the mean map at $x \in \mathcal{X}$ computes the expectation of a kernel,

$$\mu_{\mathbb{Q}}(x) = \langle \mu_{\mathbb{Q}}, k(x,\cdot)\rangle_{\mathcal{H}} = \mathbb{E}[k(x,Y)] \tag{16}$$

For appropriately chosen kernel, specifically a characteristic kernel, the mean map $\mu_{\mathbb{P}}$ completely characterizes a distribution. For example, the choice of $k(x,x') = e^{\langle x,x'\rangle}$ implies that the mean embedding is the moment generating function $\mu_{\mathbb{P}} = \mathbb{E}\left[e^{\langle X,\cdot\rangle}\right]$.

Definition 8. *(Characteristic kernel) A kernel is characteristic if the map* $\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}$ *is injective*

A characteristic kernel ensures that the induced RKHS is rich enough to represent higher order momemts of $\mathbb{P}$. In particular, it ensures that $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. As an example, Gaussian and Laplacian kernels are characteristic. Characteristic kernels are important when trying to distinguish distributions, might be less useful when trying to do predictive tasks with distributional data.

Definition 9. *(Universal kernel) A continuous positive definite kernel on compact metric space* $\mathcal{X}$ *is universal if the corresponding RKHS* $\mathcal{H}$ *is dense in* $C(\mathcal{X})$*, space of bounded continous functions over* $\mathcal{X}$*.*

Universal kernel are characteristic kernels. See [5] for a classification of kernels.

We can derive empirical estimate of the mean map $\hat{\mu}_{\mathbb{P}} = \frac{1}{m}\sum_{i=1}^{m} k(\cdot, x_i)$ for $x_i \overset{iid}{\sim} \mathbb{P}$.

5 Maximum Mean Discrepancy

The kernel mean embedding defines a natural metric for probability distributions, so called maximum mean discrepancy, as $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$. In particular, evaluation can be deligated to kernels,

$$\begin{aligned}\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} &(17)\\ &= \mathbb{E}\left[k(X,X')\right] + \mathbb{E}\left[k(Y,Y')\right] - 2\mathbb{E}\left[k(X,Y)\right] &(18)\end{aligned}$$

where $X, X' \overset{iid}{\sim} \mathbb{P}$ and $Y, Y' \overset{iid}{\sim} \mathbb{Q}$. The unbiased quadratic $\mathcal{O}(mn)$ estimate is given by

$$\widehat{\mathrm{MMD}}_u^2(X,Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{mn} \sum_{i,j} k(x_i, y_j) \tag{19}$$

where $\mathbf{X} = (X_1, \cdots, X_m) \overset{iid}{\sim} \mathbb{P}$ and $\mathbf{Y} = (Y_1, \cdots, Y_n) \overset{iid}{\sim} \mathbb{Q}$. Alternatively, we can consider maximum mean discrepancy as a class of integral probability metric over functions in the unit ball in RKHS $\mathcal{H}$,

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f \in \mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \left\{ \int f d\mathbb{P} - \int f d\mathbb{Q} \right\} \tag{20}$$

which is equivalent to the view of MMD as distance between kernel mean embeddings

$$\sup_{f \in \mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \left\{ \int f d\mathbb{P} - \int f d\mathbb{Q} \right\} = \sup_{f \in \mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \tag{21}$$

where optimal witness function is normalized difference of feature means $f^* = \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} / \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$. Figure (4) illustrates an empirical estimate of witness function, using kernels.

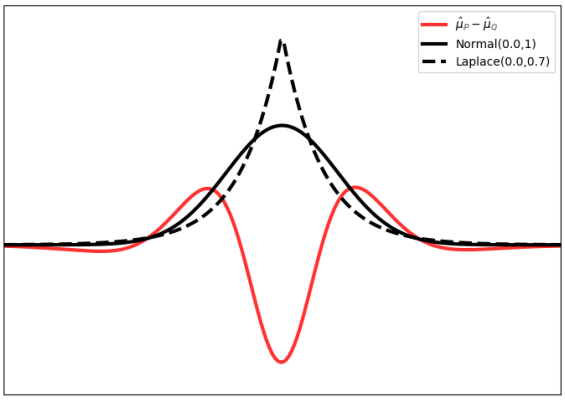

Figure 4: The unnormalized witness function $\hat{f}^* = \widehat{\mu_{\mathbb{P}}} - \widehat{\mu_{\mathbb{Q}}}$ where $\mathbb{P} \equiv \mathrm{N}(0,1)$ is compared with $\mathbb{Q} \equiv \mathrm{Lap}(0, \sqrt{.5})$. Both distribution has same mean and variance. The witness is computed from 4×10^4 samples, using a Gaussian kernel with $\sigma^2 = 0.5$. The witness function is large whenever $\mathbb{P} - \mathbb{Q}$ is positive and small when $\mathbb{P} - \mathbb{Q}$ is large. This plot reproduces one figure in [7].

Example 1. *Consider* $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1), X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$. *Let* $\mathcal{H}$ *be RKHS with* $k(x,y) = x^T y$ *as its reproducing kernel. We note that MMD*^{2} *with linear witness is equivalent to difference in means,*

$$MMD^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2}) = \left\| \mu_{\mathbb{P}_{X_1}} - \mu_{\mathbb{P}_{X_2}} \right\|_{\mathcal{H}}^2 = \|k(\mu_1, \cdot) - k(\mu_2, \cdot)\|_{\mathcal{H}}^2 = \|k(\mu_1 - \mu_2, \cdot)\|_{\mathcal{H}}^2 = \|\mu_1 - \mu_2\|_2^2 \tag{22}$$

where $\mu_{\mathbb{P}_{X_1}}(x) = \mathbb{E}\left[k(x, X_1)\right] = x^T \mathbb{E}\left[X_1\right]$ *implies* $\mu_{\mathbb{P}_{X_1}} = k(\mu_1, \cdot)$.

6 Kernel Dependence Measures

Intuitively, we can compare dependence between two random variables by computing the distance in a RKHS. [8] first proposes HSIC as Hilbert-Schmidt norm of covariance operator to measure dependence between two probability measures. [5] provides a good summary on the topic. [9] provides good summary of several related kernel dependency measures like COCO and KMI.

6.1 Covariance Operators

[10] gives a pretty good introduction on covariance operators.

Definition 10. *(Hilbert-Schmidt operator) Let $\mathcal{G}, \mathcal{F}$ be separable Hilbert space and $\{f_j\}_{j\in J}$, $\{e_i\}_{i\in I}$ be orthonormal basis for $\mathcal{G}$ and $\mathcal{F}$, respectively. A Hilbert-Schmidt operator is a bounded operator $A : \mathcal{G} \to \mathcal{F}$ whose Hilbert-Schmidt norm $\|A\|_{HS}$ is finite,*

$$\|A\|_{HS}^2 = \sum_{j\in J} \|Af_j\|_{\mathcal{F}}^2 = \sum_{i\in I}\sum_{j\in J} |\langle Af_j, e_i\rangle_{\mathcal{F}}|^2 \tag{23}$$

Hilbert-Schmidt norm reduces to Frobenius norm when $\mathcal{G}, \mathcal{F}$ is $\mathbb{R}^n$ since $\|A\|_{\text{HS}}^2 = \sum_{j=1}^n \|a_i\|_2^2 = \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 = \|A\|_F^2$.

Definition 11. *(Space of Hilbert-Schmidt operators) The Hilbert-Schmidt operators from $\mathcal{G}$ to $\mathcal{F}$ with inner product defined below is a Hilbert space denoted as $HS(\mathcal{G}, \mathcal{F})$.*

$$\langle L, M\rangle_{HS} = \sum_{j\in J} \langle Lf_j, Mf_j\rangle_{\mathcal{F}} = \sum_{i\in I}\sum_{j\in J} \langle Lf_j, e_i\rangle_{\mathcal{F}} \langle Mf_j, e_i\rangle_{\mathcal{F}} \tag{24}$$

Although the inner product is defined based on $(f_j)_{j\in J}$, it is in fact independent of the orthonormal basis chosen, as shown in [10]. Additionally, we see that this inner product recovers the Hilbert-Schmidt norm, $\langle L, L\rangle_{\text{HS}} = \sum_{j\in J} \langle Lf_j, Lf_j\rangle_{\mathcal{F}} = \sum_{j\in J} \|Lf_j\|_{\mathcal{F}}^2 = \|L\|_{\text{HS}}^2$. Another thing to note is that $\text{HS}(\mathcal{G}, \mathcal{F}) \simeq \mathcal{G} \otimes \mathcal{F}$.

Definition 12. *(Tensor product on Hilbert space) For $b \in \mathcal{G}, a \in \mathcal{F}$, the tensor product $a \otimes b$ is a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$ defined as $(a \otimes b)(g) = \langle g, b\rangle_{\mathcal{G}} a$ for $g \in \mathcal{G}$.*

Tensor product of spaces can be defined via a bilinear map fo the form $\mathcal{F} \times \mathcal{G} \to \mathbb{R}$ that satisfies the property of being linear in each input coordinate. We can define $(a \otimes b) : \mathcal{G} \times \mathcal{F} \to \mathbb{R}$ with $(a \otimes b)(g, f) = \langle g, b\rangle_{\mathcal{G}} \langle f, a\rangle_{\mathcal{F}}$ and define the resulting tensor product space as $\mathcal{G} \otimes \mathcal{F} = \overline{\{a \otimes b\}}_{a\in\mathcal{F}, b\in\mathcal{G}}$. Alternatively, we can define a map from $\mathcal{G}$ to $\mathcal{F}$ as previously and arrive at the same result. Intuitively, tensor product on function space generalizes outer product in Euclidean vector spaces. For example for $a, b \in \mathbb{R}^n$, we have $(ab^T)g = a(b^T g)$, similar to definition (12)

Proposition 1. *(Properties of tensor product [10])*

1. $\|a \otimes b\|_{HS}^2 = \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2$ *(and therefore $a \otimes b \in HS(\mathcal{G}, \mathcal{F})$)*
2. $\langle L, a \otimes b\rangle_{HS} = \langle a, Lb\rangle_{\mathcal{F}}$ *for For $L \in HS(\mathcal{G}, \mathcal{F})$ ($L = u \otimes v$ implies $\langle u \otimes v, a \otimes b\rangle_{HS} = \langle u, a\rangle_{\mathcal{F}} \langle v, b\rangle_{\mathcal{G}}$)*

The take-away is that tensor products are Hilbert-Schmidt operators and we can evaluate HS inner product of two tensor products as multiple of two inner products in $\mathcal{G}, \mathcal{F}$.

Definition 13. *(Cross-covariance operator) Let (X, Y) be random variables over $\mathcal{X} \times \mathcal{Y}$ and $(\mathcal{F}, k, \phi)$ and $(\mathcal{G}, l, \psi)$ be RKHS, its reproducing kernel and feature map, respectively. $\widetilde{C}_{XY} : \mathcal{G} \to \mathcal{F}$ is called the (uncentered) cross-covariance operator*

$$\widetilde{C}_{XY} = \mathbb{E}\left[\phi(X) \otimes \psi(Y)\right] \tag{25}$$

The corresponding centered covariance operator is

$$C_{XY} = \mathbb{E}\left[(\phi(X) - \mu_X) \otimes (\psi(Y) - \mu_Y)\right] = \widetilde{C}_{XY} - \mu_X \otimes \mu_Y \tag{26}$$

where $\mu_X = \mathbb{E}\left[\phi(X)\right], \mu_Y = \mathbb{E}\left[\psi(Y)\right]$ are mean embeddings of $\mathbb{P}_X, \mathbb{P}_Y$.

Intuitively, we can think of $\phi(X) \otimes \psi(Y)$ as a random variable in space of operators $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ or in the tensor product feature space $\mathcal{G} \otimes \mathcal{F}$. The definition generalizes finite dimensional (uncentered) covariance matrix $\mathbb{E}\left[XY^T\right]$ to be the corresponding covariance matrix in the tensor product feature space. Equivalently, we can consider cross-covariance operator as mean embedding of $\mathbb{P}_{XY}$ in $\mathcal{G} \otimes \mathcal{F}$ induced by product kernel $v((x, y), (x', y') = k(x, x')l(y, y')$, i.e. $\widetilde{C}_{XY} = \mu_{\mathbb{P}_{XY}}$.

Lemma 2. *If $\mathbb{E}\left[\|\phi(X) \otimes \psi(Y)\|_{HS}\right] < \infty$, then $\widetilde{C}_{XY} \in HS(\mathcal{G}, \mathcal{F})$ and $\widetilde{C}_{XY}$ uniquely satisfies*

$$\left\langle \widetilde{C}_{XY}, A \right\rangle_{HS} = \mathbb{E}\left\langle \phi(X) \otimes \psi(Y), A \right\rangle_{HS} \tag{27}$$

Proof. Detail in [10]. Define linear operator $T_{XY} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) \to \mathbb{R}$ as $T_{XY}(A) \mapsto \mathbb{E}\left\langle \phi(X) \otimes \psi(Y), A \right\rangle_{\mathrm{HS}}$ and show it is bounded. Use Riesz representation theorem to guarantee existence of $\widetilde{C}_{XY}$ satisfying the equation and show that it has the form defined previously. □

Again, we can consider (27) as exchanging expectation and inner products over $\mathrm{HS}(\mathcal{G}, \mathcal{F})$. In particular, if we let $A = f \otimes g$, then we can use properties of tensor product and the previous lemma and get

$$\left\langle f, \widetilde{C}_{XY} g \right\rangle_{\mathcal{F}} = \left\langle \widetilde{C}_{XY}, f \otimes g \right\rangle_{\mathrm{HS}} = \mathbb{E}\left\langle \phi(X) \otimes \psi(Y), f \otimes g \right\rangle_{\mathrm{HS}} = \mathbb{E}\left[\left\langle f, \phi(X) \right\rangle_{\mathcal{F}} \left\langle g, \psi(Y) \right\rangle_{\mathcal{G}}\right] \tag{28}$$

$$= \mathbb{E}\left[f(X)g(Y)\right] \tag{29}$$

This is used to show HSIC as IPM with two witness functions $f \in \mathcal{F}, g \in \mathcal{G}$.

6.2 Hilbert-Schmidt Independence Criterion

[8] uses Hilbert-Schmidt norm of (centered) covariance operator to measure dependence,

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{XY}) = \left\| \widetilde{C}_{XY} \right\|_{\mathrm{HS}} = \left\| C_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathrm{HS}} \tag{30}$$

Equivalently, we can consider HSIC as MMD between kernel embedding of joint $\mathbb{P}_{XY}$ with that of product of marginals $\mathbb{P}_X \mathbb{P}_Y$ over the tensor product feature space $\mathcal{G} \otimes \mathcal{F}$,

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{XY}) = \left\| \mu_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathcal{G} \otimes \mathcal{F}} \tag{31}$$

Proposition 2. *(HSIC and Independence) $HSIC(\mathcal{F}, \mathcal{G}, \mathbb{P}_{XY}) = 0$ if and only if $X \perp\!\!\!\perp Y$.*

Computation of HSIC can be delegated to kernels again (see [10] for detailed derivation), implying that we can compute a measure of dependence between X and Y via expectation over kernels without needing to perform density estimation.

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{XY}) = \mathbb{E}_{X,Y}\mathbb{E}_{X',Y'} k(X, X') l(Y, Y') + \mathbb{E}\left[k(X, X')\right] \mathbb{E}\left[k(Y, Y')\right] \tag{32}$$

$$- 2\mathbb{E}_{X,Y}\left[\mathbb{E}\left[k(X, X')\right] \mathbb{E}\left[l(Y, Y')\right]\right] \tag{33}$$

Given samples $Z = \{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \mathbb{P}_{XY}$, we have a simple biased estimate,

$$\widehat{\text{HSIC}}_b^2(\mathcal{F}, \mathcal{G}, Z) = \frac{1}{n^2}\operatorname{tr}(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}) = \frac{1}{n^2}\operatorname{tr}(\mathbf{KHLH}) \tag{34}$$

where $\widetilde{\mathbf{K}}, \widetilde{\mathbf{L}}$ are centered kernel matrices computed from kernels k, l and $\mathbf{H}$ is the centering matrix (Refer to example (2) for centering in feature space). The last equality given by fact that $\mathbf{H}$ is indempotent and cyclic property of trace. [11] derived an unbiased $\mathcal{O}(n^2)$ estimator,

$$\text{HSIC}^2(\mathcal{F}, \mathcal{G}, Z) = \frac{1}{n(n-3)}\left[\operatorname{tr}(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}) + \frac{\mathbf{1}^T\widetilde{\mathbf{K}}\mathbf{1}\mathbf{1}^T\widetilde{\mathbf{L}}\mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2}\mathbf{1}^T\widetilde{K}\widetilde{L}\mathbf{1}\right] \tag{35}$$

Example 2. *(Centering Matrix [12] A.1) Given φ and a set of observations $x_1, \cdots, x_n$, we might want to centered features $\widetilde{\varphi}(x_i) = \varphi(x_i) - \overline{\varphi}$ where $\overline{\varphi} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)$ is the centroid. We want to find centered kernel matrix $\widetilde{\mathbf{K}}$ such that $\widetilde{\mathbf{K}}_{ij} = \langle\widetilde{\varphi}(x_i), \widetilde{\varphi}(x_j)\rangle_{\mathcal{H}}$. Note,*

$$\begin{aligned}
\widetilde{\mathbf{K}}_{ij} &= \langle\widetilde{\varphi}(x_i), \widetilde{\varphi}(x_j)\rangle_{\mathcal{H}} && (36)\\
&= \left\langle \varphi(x_i) - \frac{1}{n}\sum_{s=1}^n \varphi(x_s), \varphi(x_j) - \frac{1}{n}\sum_{t=1}^n \varphi(x_t)\right\rangle_{\mathcal{H}} && (37)\\
&= k(x_i, x_j) - \frac{1}{n}\sum_{s=1}^n k(x_s, x_j) - \frac{1}{n}\sum_{t=1}^n k(x_t, x_i) + \frac{1}{n^2}\sum_{s=1}^n\sum_{t=1}^n k(x_s, x_t) && (38)\\
&= \mathbf{K}_{ij} - (\mathbf{1}_n\mathbf{K})_{ij} - (\mathbf{K}\mathbf{1}_n)_{ij} + (\mathbf{1}_n\mathbf{K}\mathbf{1}_n)_{ij} && (39)\\
&= ((\mathbf{I} - \mathbf{1}_n)\mathbf{K}(\mathbf{I} - \mathbf{1}_n))_{ij} && (40)\\
&= (\mathbf{HKH})_{ij} && (41)
\end{aligned}$$

where $\mathbf{1}_n \in \mathbb{R}^{n\times n}$ defined as $\mathbf{1}_n = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Therefore, $\widetilde{\mathbf{K}} = \mathbf{HKH}$ where $\mathbf{H} = \mathbf{I} - \mathbf{1}_n$.

Example 3. *(2D Gaussian) Let $X \sim \mathcal{N}(\mu, \Sigma)$ over $\mathbb{R}^2$ where $\mu = \mathbf{1}$, $\Sigma = \mathbf{I}_2$. Consider $A \in \mathbb{R}^{2\times 2}$ where a_1, a_2 are rows of A. Let $(y_1, y_2) := Y = AX$ where $Y \sim \mathcal{N}(A\mu, A\Sigma A^T)$. We want to find coefficients of linear map that makes the transformed variables independent subject to constraint that the $\sum_j a_i j = 1$ for $i = 1, 2$ to avoid degenerate zero mapping solution,*

$$\min_{a_1, a_2} \mathit{HSIC}_k(a_1 x, a_2 x) - \lambda\left\|A\mathbf{1} - \mathbf{1}\right\|^2 \tag{42}$$

Note $y_1 \perp\!\!\!\perp y_2$ if and only if $A\Sigma A^T = I$, which is what we observe in simulation. See Figure (5)

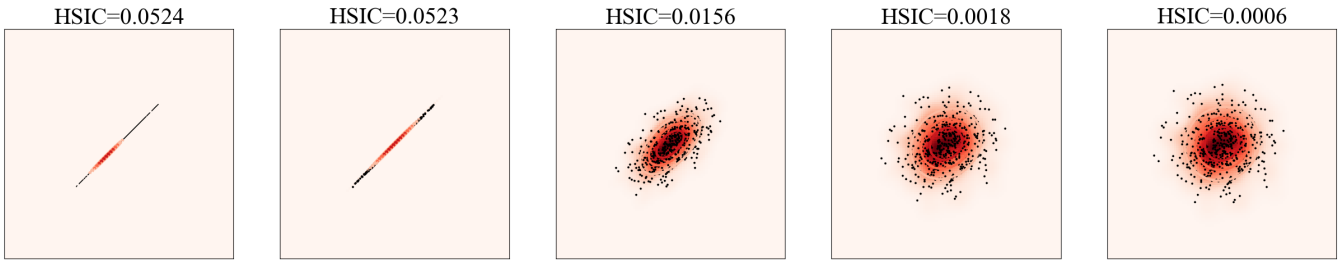

Figure 5: Use stochastic gradient descent to optimize for (42) where we have used RBF kernel with bandwidth chosen using median heuristic. We plotted transformed point by the linear mappings and its estimated density using Gaussian KDE. We also plotted contours of density of Y. Note as optimization progresses, the transformed points look as if they are sampled from an isotropic Gaussian.

References

[1] Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: (1992).

[2] Jean-Philippe Vert, Koji Tsuda, and Bernhard Scholkopf. "A primer on kernel methods". In: (2004), p. 42.

[3] Arthur Gretton. "What is an RKHS?" In: (2012). URL: http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_2014.pdf.

[4] Arthur Gretton. "Introduction to RKHS, and some simple kernel algorithms". In: (2019), p. 33. URL: http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf.

[5] Krikamol Muandet et al. "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends® in Machine Learning* 10.1 (2017), pp. 1–141. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000060. arXiv: 1605.09522. URL: http://arxiv.org/abs/1605.09522 (visited on 12/25/2020).

[6] Ali Rahimi and Ben Recht. "Random Features for Large-Scale Kernel Machines". In: (2007), p. 8.

[7] Arthur Gretton et al. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773. URL: http://jmlr.org/papers/v13/gretton12a.html (visited on 06/21/2020).

[8] Arthur Gretton et al. "Measuring statistical dependence with hilbert-schmidt norms". In: *Proceedings of the 16th international conference on Algorithmic Learning Theory*. ALT'05. Berlin, Heidelberg: Springer-Verlag, Oct. 8, 2005, pp. 63–77. ISBN: 978-3-540-29242-5. DOI: 10.1007/11564089_7. URL: https://doi.org/10.1007/11564089_7 (visited on 01/01/2021).

[9] Arthur Gretton et al. "Kernel Methods for Measuring Independence". In: (2005), p. 55.

[10] Arthur Gretton. "Notes on mean embeddings and covariance operators". In: (2019), p. 15. URL: http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture5_covarianceOperator.pdf.

[11] Le Song. "Feature Selection via Dependence Maximization". In: (2012), p. 42.

[12] B. Schölkopf, A. Smola, and K. Müller. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10.5 (July 1998). Conference Name: Neural Computation, pp. 1299–1319. ISSN: 0899-7667. DOI: 10.1162/089976698300017467.