Applications of Reproducing Kernel Hilbert Spaces to Nonparametric Statistics

Vafa Behnam Roudsari

May 20, 2019

Contents

1 Foundations

1.1 Statistical Motivation

We will concern ourselves with two critical statistical tasks, density estimation and regression. Given a distribution p, the most general setting in statistics and machine learning for density estimation is

$$\min_{\hat{f} \in \mathcal{F}} \mathbb{E}_{x \sim p} \, ||p(x) - \hat{f}(x)||$$

The corresponding setting for regression is, instead of learning a single marginal density, represented by x, we instead try to learn a conditional relationship, i.e. how a variable y varies as x varies.

$$\min_{\hat{f}\in\mathcal{F}} \mathbb{E}_{(x,y)} ||y - \hat{f}(x)||$$

However, in practice, we never have access to the full distribution of x in density estimation or x, y (or else the problem would be trivial). Instead, in statistics, we have access to a finite sample (x_i) or (x_i, y_i) coming from the distributions of interest. So we replace the two equations above with their empirical counterparts.

$$\min_{\hat{f}\in\mathcal{F}} \sum_{i=1}^{N} ||p(x_i) - \hat{f}(x_i)||$$

$$\min_{\hat{f}\in\mathcal{F}} \sum_{i=1}^{N} ||y_i - \hat{f}(x_i)||$$

What objects are under the direct control of the statistician? Basically three things, the norm $||\cdot||$, and the form of $\mathcal{F}$. As we will later see in one example in 3.1, robust kernel density estimation, the choice of norm may be more important than at first sight. The only other thing that appears to under the statistician's control is the function space $\mathcal{F}$.

Of course, one can simply opt for $\mathcal{F} = L^2$ or even the set of all mappings to make sure we are able to learn any function. In this case, our only issue is to yield a tractable algorithm and have representations of functions that can be implemented on a computer with finite precision (or a truncation with acceptable amounts of error). In this scenario, RKHS's have a very important role to play, as infinite-dimensional objects like functions have an exact finite-dimensional representation, greatly facilitating our search for the best function. The core of this is usually what is called the "kernel trick". We will see more of this in the next chapter, as well as in some of the applications.

However, the story does not simply end there. Statistics is not simply applied optimization, since, as stated before, we have access to a finite sample $\{(x_i, y_i)\}_{i=1}^{n}$ as opposed to the full distribution. The intuition here, that will be developed in the future, is that "big" (to be explained exactly what is meant by "big") function sets $\mathcal{F}$ "overfit" to the particular sample, whereas small function sets $\mathcal{F}$ underfit. Why? Big function sets are likelier to contain very complicated functions, so to achieve a good test error they may simply "memorize" our finite sample, achieving excellent test error, but generalizing very poorly to new data.

"Big" can mean cardinality for example, and while it is a half decent heuristic, it turns out the cardinality of the set is not a precise measure of the overfit vs underfit capacity of the set. We will see notios of capacity that try to capture

Here, RKHS's also pose a very interesting option. Regardless of the particular cardinality of the RKHS, they have shown to achieve a decent balance between overfitting and underfitting. We will explore these connections in Section 2, and why regularization techniques, developed to counter overfitting, work well for RKHS's.

This is interesting from a historical standpoint, as kernel methods in machine learning initially began as a way to speed up computation time (or even making an impossible task possible), it now has implications for the underfit vs overfit debate.

1.2 Kernel Methods

In fullest generality, a feature map is simply any map $\phi : \mathcal{H} \to F$ where F is a feature space. Usually F is taken to be a rich higher dimensional space, where complicated manifolds in lower dimensional spaces can be represented linearly in F.

It is clear, if given a kernel k, that many feature maps can be constructed from it. For example, for any point $x \in X$ we can construct a feature map $\phi_x(y) = k_x(y) = k(x, y)$ (this is called the canonical feature map). Some feature maps will be much more useful than others, as we will see.

The next theorem shows the other direction: given a feature map, we can construct a kernel.

Theorem 1. A feature map ϕ defines a positive semi-definite kernel $k(x, y)$ by $k(x, y) = \langle \phi(x), \phi(y) \rangle$

Proof. We prove that k_x is positive semi-definite kernel, and that it is unique. First, we prove k_x is symmetric. Since a kernel is associated to a feature map, and by symmetry of inner products, we have,

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{F}} = \langle \phi(x_2), \phi(x_1) \rangle_{\mathcal{F}} = k(x_2, x_1)$$

Next, we show that this kernel is positive semi-definite.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c_i c_j K(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle = \langle \sum_{i=1}^{n} c_i \phi(x_i), \sum_{j=1}^{m} c_j \phi(x_j) \rangle$$

$$= || \sum_{i=1}^{n} c_i \phi(x_i) ||^2 \geq 0$$

□

The following equality is at the core of the kernel trick. Any algorithm that uses inner products can be altered to inner products in a far more complicated feature space (infinite dimensional in our current setting) induced by ϕ.

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{F}} == \langle k_{x1}, k_{x2} \rangle_{\mathcal{F}}$$

This gives us a sense as to why the kernel trick has been very popular in machine learning. As a quick high-level example, we have efficient algorithms that can find a linear separator for a set of n points, assuming that those points can be linearly separable. If those points were not linearly separable, the algorithm would clearly fail. But there is an immediate way of fixing this without fundamentally changing our linear separator algorithm: if we can find a high-dimensional feature space F such that our n points are linearly separable there, we can simply replace all inner products with the higher dimensional inner product. This higher inner product may be expensive to compute, but we can represent the higher inner product as a kernel in the lower dimensional (and finite) space.

1.3 Constructing a Space of Functions from a Kernel

We begin with the Hilbert Space. A Hilbert space is a complete inner product space that proves to be a very convenient setting for optimization and statistics.

A major reason for this is projections and orthogonality play a crucial role in optimization. These concepts are right at home in an inner product space. The usage of the parallelogram law is most natural in a Hilbert Space.

Beyond this, Hilbert Spaces pose other properties that are extremely useful. For example, the dual of a dual of a Hilbert Space is the original Hilbert Space, and Hilbert Spaces can be decomposed into any closed subspace and its orthogonal complement. These properties will be used in the succedding proofs.

We can now present a definition of the Reproducing Kernel Hilbert Spaces (RKHS):

Definition. A Hilbert Space $\mathcal{H}$ is a RKHS if for any evaluation functional $A_x \in \mathcal{H}^*$ s.t. $A_x(f) = f(x)$, $|A_x(f)| = |f(x)| \leq C||f||_{\mathcal{F}}$, i.e. A_x is bounded.

We have a bit of intuition now about why we care about "reproducibility" in this sense. We see all functions that live in a RKHS have a sort of smoothness property, i.e. that they do not diverge much. This is a key point that will show up over and over again, that making our hypothesis space $\mathcal{F}$ will induce a set of favorable properties. Very precisely, having a RKHS will induce smooth functions, which in turn will exhibit a good balance between overfitting and underfitting. We will go more into depth on this topic in chapter 2.

Where do kernels come in? It turns out there is a one-to-one correspondence between a kernel and a RKHS. This is very helpful for analysis, as we can learn about a function space by investigating the properties of its corresponding kernel.

We first need to show that all functionals can be expressed as an inner product in terms of the kernel. Before doing so, we need the Riesz Representation Theorem.

Theorem 2. (Riesz Representation Theorem – Luenberger 1997 [1])
Let A be a linear, bounded operator in a Hilbert space $\mathcal{H}$. $\exists g \in \mathcal{H}$ such that $A(f) = \langle f, g \rangle$.

Proof. Let A be a bounded operator acting on $\mathbb{H}$. Let $N = \{f \in \mathcal{H} | Af = 0\}$. N is a subspace, as 1) $\alpha Ax + \beta Ay = 0$ for any $x, y \in N$, and 2) $A0 = 0$ so $0 \in N$.

N is also closed. To prove this, let us take an arbitrary $f \in H$ and a sequence $n_i \to f$ s.t. $n_i \in N$. By continuity of A, $An_i \to Af$, but since $An_i = 0$ as $\forall n_i \in N$, this forces $Af = 0$. If $N = \mathcal{H}$, take $g = 0$ and hence $Af = \langle 0, f \rangle$, proving the theorem.

Otherwise, let us assume $N \neq \mathcal{H}$. Since $\mathcal{H}$ is a Hilbert Space, it can be decomposed into a closed subspace and its orthogonal complement, i.e. $\mathcal{H} = N \bigoplus N^{\perp}$. Furthermore, $\exists h \neq 0 \in N^{\perp}$ such that $Ah \neq 0$, since $N \subsetneq \mathcal{H}$. Without loss of generality, we can take $Ah = 1$.

Let $f \in \mathcal{H}$. $f - Afh \in N$ since $A(f - A(f)h) = Af - AfAh = Af - Af = 0$. Since $h \in N^{\perp}$, $\langle f - Afh, h \rangle = 0$.

$$\langle f, h \rangle - \langle Afh, h \rangle = 0 \implies \langle f, h \rangle = \langle Afh, h \rangle \implies \langle f, h \rangle = Af||h||^2$$

$$\implies \frac{\langle f, h \rangle}{||h||^2} = Af \implies \langle f, \frac{h}{||h||^2} \rangle = Af$$

Therefore, $g = \frac{h}{||h||^2}$ is our target function, proving the theorem.

□

We are now ready to prove a 1-1 correspondence between a positive semi-definite kernel and RKHS. We begin with the RKHS defines a kernel direction first.

Theorem 3. A RKHS $\mathcal{H}$ has a unique reproducing positive semi-definite kernel k.

Proof. By definition of a RKHS, the evaluation operator A_x is bounded, so we can immediately apply the Riesz Representation Theorem to our setting to yield, for $x \in X$,

$$A_x(f) = \langle f, k_x \rangle = f(x)$$

We show that $k_x = k(x, \cdot)$ is the target kernel. Let us define a feature map ϕ by the canonical mapping $x \mapsto k(x, \cdot)$, i.e. $\phi(y) = k_x(y) = k(x, y)$ for $y \in X$. By application of Theorem 1, since a feature mapping defines a positive-semidefinite kernel, $k(x, y)$ must be a positive semi-definite kernel.

All that is left to prove is that k is unique. Suppose there exists two distinct kernels, k and k', such that both are reproducing kernels of $\mathcal{H}$. Let $x, x' \in X$

$$k'(x', x) = \langle k(x, \cdot), k'(x', \cdot) \rangle = k(x, x') = k(x', x)$$

The first and second equalities hold because of the reproducing property applied to k and k' respectively (i.e. in the first equality, we apply the reproducing property to $k(x, \cdot)$ to yield $k'(x', \cdot)(x) = k'(x', x)$).

The third equality holds by symmetry of the kernel. Since x, x' were arbitrary, this results in a contradiction.

□

It is unsurprising that k_x is the reproducing kernel for the point x. This is a crucial equality that will be used again and again in applications of RKHS. It is helpful to think of $k_x = k(x, \cdot)$ as a high dimensional representation of the datapoint x.

We now prove that a positive semidefinite kernel defines a unique RKHS.

Theorem 4. Conversely, given a positive semidefinite kernel k, $\mathcal{H} = \overline{span\{k(x, .)|x \in X\}}$ is a RKHS, with k as its reproducing kernel.

Proof. We begin by proving $\mathcal{H} = \overline{span\{k(x, .)|x \in X\}}$ is a Hilbert space. First we show that the subset $span\{k(x, .)|x \in X\} \subset \mathcal{H}$ is a pre-Hilbert space under the following inner product, then later add in the limits of the Cauchy sequences.

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} c_i b_j K(x_i, y_j)$$

where $f = \sum_{i=1}^{n} c_i K(\cdot, x_i), g = \sum_{j=1}^{m} b_j K(\cdot, y_j)$

Symmetry.

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} c_i b_j K(x_i, y_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} b_j c_i K(x_i, y_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} b_j c_i K(y_j, x_i) = \langle g, f \rangle$$

where third equality holds by symmetry of kernel.

Semidefiniteness. This definition of a positive semi-definite kernel is almost equivalent to the definition of a positive semi-definiteness of the inner product. We just need to check that $\langle f, f\rangle = 0 \implies f = 0$.

$$|f(x)|^2 = \langle k(x,\cdot), f\rangle \underset{\text{Cauchy-Schwarz}}{\leq} \langle k(x,\cdot), k(x,\cdot)\rangle\langle f, f\rangle = 0$$

Linearity. Scalar multiplication. Let $\alpha \in \mathbb{R}$.

$$\alpha\langle f, g\rangle = \alpha \sum_{i=1}^{n}\sum_{j=1}^{m} c_i b_j K(x_i, y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha c_i b_j K(x_i, y_j) = \langle \alpha f, g\rangle$$

Linearity in variables. Let $h \in \mathcal{H}$ and $h = \sum_p d_p k(\cdot, z_p)$. Recall that

$$\langle f, g\rangle = \sum_{j=1}^{m} b_j \sum_{i=1}^{n} c_i K(x_i, y_j) = \sum_{i=1}^{n} b_j f(x_j)$$

Hence,

$$\langle f + h, g\rangle = \sum_{j=1}^{m} b_j (f(x_j) + h(x_j)) = \sum_{j=1}^{m} b_j f(x_j) + \sum_{j=1}^{m} b_j h(x_j) = \langle f, g\rangle + \langle h, g\rangle$$

Now suppose that $span\{k(x,.)|x \in X\} \neq \mathcal{H}$ (If they are equal we have already proved $\mathcal{H}$ is a pre-Hilbert space). Let $f \in \mathcal{H} \setminus span\{k(x,.)|x \in X\}$ (so f is the limit of a Cauchy sequence of the span of the kernel). Let $g \in \mathcal{H}$. We "generalize" our inner product in this case to be,

$$\langle f, g\rangle = \lim_{l\to\infty}\langle f_l, g\rangle = \lim_{l\to\infty} \sum_{i=1}^{n_l}\sum_{j=1}^{m} (c_i)_l b_j K((x_i)_l, y_j)$$

where $f = \lim_{l\to\infty} f_l(x) = \lim_{l\to\infty}(\sum_{i=1}^{n} c_i K(\cdot, x_i))_l = \sum_{i=1}^{n_l}(c_i)_l b_j K(\cdot, (x_i)_l)$,
and $g = \sum_{j=1}^{m} b_j K(\cdot, y_j)$

Since limits can be safely interchanged with finite sums, and since limits weakly preserve inequalities, all of the preceding proofs can be easily amended to include the inequality case. Therefore, $\mathcal{H}$ is a pre-Hilbert space.

If we take $\mathcal{H}$ to be a subset of L^p for $1 \leq p \leq \infty$, (which we will do almost always) then, as the L^p spaces are all complete metric spaces, closedness and completeness of subsets are equivalent, hence proving $\mathcal{H}$ is a Hilbert Space.

We have proven that $\mathcal{H}$ is a Hilbert space. To make this into a RKHS, we need to check that the evaluation functionals on this space are bounded. Let A_x be an evaluation functional.

$$||A_x(f)|| \underset{\text{rep. prop}}{=} \langle f, k(\cdot, x)\rangle \underset{\text{CS ineq.}}{\leq} \langle f, f\rangle^{\frac{1}{2}}\langle k(\cdot,x), k(\cdot,x)\rangle^{\frac{1}{2}} \underset{\text{kernel trick eq. 2}}{=} ||f||k(x,x)^{\frac{1}{2}} < \infty$$

□

1.4 Constructing a Space of Functions from a Measure

In the last chapter, we took a point $x \in \mathbb{R}^d$ and mapped it into a RKHS using $k(x, \cdot)$, and this kernel generated the entire space RKHS.

What if, in a similar vein, we take a probability measure, $\mathbb{P}$, and mapped it into $\int k(x, \cdot)d\mathbb{P}(x)$? Can we, like the previous case, generate a RKHS, and does $x \mapsto \int k(x, \cdot)d\mathbb{P}$ itself belong to it?

Before proceeding, why would we want to do this? The answer is just for similar reasons reasons as before, we can represent complicated and very useful objects much more simply in terms of inner products or kernels (such as moments in statistics)

The target of this chapter is to prove that the range of $\mathbb{P} \mapsto \int k(x, \cdot)d\mathbb{P}(x)$ is – in **some** cases – a RKHS, and in fact the RKHS that corresponds to the reproducing kernel k as we discussed in the previous section.

One of the major issues is that we need to make sure the integral $\int k(x, \cdot)d\mathbb{P}(x)$ is a valid integral. Lebesgue integral theory is insufficient because it only applies to real-valued functions. We now have Hilbert space-valued functions (notice the free variable in the integral means the integral spits out another function, therefore for each infinitesmal dx we have an entire function).

We use the Bochner integral, and to avoid excessive measure theory we will suppose our functions are Bochner integrable from now on.

Definition. A measurable function $k(x, \cdot) : X \to \mathcal{H}$ is Bochner integrable if there exists a sequence of simple functions s_n such that $\lim_{n\to\infty} \int_X ||k(x, \cdot) - s_n||_{\mathcal{H}} d\mathbb{P}(x) = 0$. If this holds, we define the Bochner integral by $\int_X k(x, \cdot)d\mathbb{P}(x) = \lim_{n\to\infty} \int_X s_n d\mathbb{P}(x)$.

We first require injectivity. This might be less restrictive than at first seems, as a number characteristic kernels such as the Dirac, Gaussian and Laplacian kernels, are injective.

Theorem 5. (Muandet et al. 2017 Section 3.3.1 [3])
For characteristic kernels, the map $\mu_{\mathbb{P}}$ is injective.

With this, all we need is now the surjectivity of the map $\mu_{\mathbb{P}}$. This will then allow us to move seamlessly between Hilbert space functions and measures. We need one more definition before we can proceed with the final proof.

Definition. (Berlinet & Thomas-Agnan 2004 [2])
Let X be a random variable in $\mathcal{H}$, and A a subset of the associated sigma-algebra of the space. X is said to be weakly integrable with respect to A if 1) for all $f \in \mathcal{H}$, $\int_A \langle X, f\rangle d\mathbb{P} = \langle \int_A X d\mathbb{P}, f\rangle < \infty$, and 2) if there exists $z \in \mathcal{H}$ such that for all $f \in \mathcal{H}$, $\mathbb{E}_A(\langle X, f\rangle) = \langle z, f\rangle$.
If this holds, we call z the weak integral of X on A and $z = \int_A X d\mathbb{P} = \mathbb{E}_A(X)$

(A side comment: it turns out even weak integrability is not strictly necessary, but then we would have to modify $\mu_{\mathbb{P}}$ which looks very nice in its "canonical" form. If this condition is not satisfied, we would need to compose this canonical map with another function as our mapping from the measures to the Hilbert space, which can get complicated and loses some

of the simplicity (like representation of the inner products). More details in Berlinet and Thomas-Agnan Theorem 99.)

With that comment out of the way, we now proceed with the main proof, which is basically that the map μ defines a RKHS.

Theorem 6. (Berlinet & Thomas-Agnan 2004 [2])

Suppose $\delta : X \to \mathcal{M}$ by $x \mapsto \delta_x$, the Dirac delta, is weakly integrable for all $\mathbb{P} \in \mathcal{M}$. Further suppose that $\mu : \mathbb{P} \mapsto \int k(x, \cdot) d\mathbb{P}(x)$ is injective (which holds for universal kernels, see Theorem 5). Finally, suppose $k(x, y) = \langle \delta_x, \delta_y \rangle$ is a measurable and bounded function.

Then $Im(\mu_{\mathbb{P}}(\mathcal{M})) = \mathcal{H}$, where $\mathcal{H}$ is the RKHS induced by the kernel k.

Proof. Let $\mathcal{M}_0$ be a the set of measures in $\mathcal{M}$ with finite support, and define an inner product $\langle \mathbb{P}, v \rangle$ such that $\mathcal{M}_0$ is dense in $\mathcal{M}$. For elements $\mathbb{P}, v \in \mathcal{M}_0$,

$$\langle \mathbb{P}, v \rangle_{\mathcal{M}} = \int k(x, y) d(\mathbb{P} \otimes v)(x, y) = \langle \mu(\mathbb{P}), \mu(v) \rangle_{\mathcal{H}}$$

Since $\langle \mathbb{P}, v \rangle_{\mathcal{M}} = \langle \mu(\mathbb{P}), v \rangle_{\mathcal{H}}$ this implies μ is an isometry.

Let $\mathbb{P}, v \in \mathcal{M}$. Then, by density of $\mathcal{M}_0$, $\exists$ a sequence of $\mathbb{P}_i, v_i \in \mathcal{M}_0$ that converge to $\mathbb{P}, v$ respectively. Since μ is an isometry, it preserves Cauchy sequences and hence by completeness of $\mathcal{H}$ there exists functions $f, g \in \mathcal{H}$ such that $\mu(\mathbb{P}_i) \to f, \mu(v) \to g$.

Define a map $h : \mu(\mathcal{M}) \to \mathcal{H}$ by $h(\mu(\mathbb{P})) = f$, i.e. associate each member with the image of the convergent sequence in $\mathcal{M}_0$.

$$\langle h(\mu(\mathbb{P})), h(\mu(v)) \rangle_{\mathcal{H}} = \lim_{n \to \infty} \langle \mu(\mathbb{P}_n), \mu(v_n) \rangle_{\mathcal{H}} = \lim_{n \to \infty} \langle \mathbb{P}_n, v_n \rangle_{\mathcal{M}} = \langle \mathbb{P}, v \rangle_{\mathcal{M}}$$

Now we turn to proving h is an identity map.

$$\langle \mathbb{P}, v \rangle_{\mathcal{M}} = \langle \int \delta_x d\mathbb{P}(x), \delta_y dv(y) \rangle_{\mathcal{M}} = \int \langle \delta_x, \int \delta_y dv(y) \rangle_{\mathcal{M}} d\mathbb{P}(x)$$

$$= \int \Big(\int \langle \delta_x, \delta_y \rangle_{\mathcal{M}} dv(y) \Big) d\mathbb{P}(x) = \int \Big(\int k(x, y) dv(y) \Big) d\mathbb{P}(x))$$

$$= \int \big\langle k(x, \cdot) d\mathbb{P}(x), k(y, \cdot) dv(y) \big\rangle_{\mathcal{H}}$$

Since this implies $\langle h(\mu) =$ for arbitrary μ, h is the identity map, hence $\mu = h \circ \mu$ and therefore μ is surjective on $\mathcal{H}$

□

Unlike the mapping $x \mapsto k(x, \cdot)$ which trivially belongs to $\mathcal{H} = \overline{span\{k(x, .) | x \in X\}}$, it is not guaranteed that $x \mapsto \int k(x, \cdot) d\mathbb{P}(x) \in \mathcal{H}$. This following statement gives a sufficient condition for when this is the case, and in addition let's us express expectations in a very simple inner product (a reproducing property for kernel mean embeddings).

Theorem 7. ((Muandet et al. 2017 Section 3.1 [3]))

If $\mathbb{E}_{X \sim P} \sqrt{k(X, X)} < \infty, \mu_{\mathbb{P}} \in \mathcal{H}$ and $\mathbb{E}_{X \sim P}(f(X)) = \langle f, \mu_{\mathbb{P}} \rangle$

Proof. For any $f \in \mathcal{H}$,

$$|\mathbb{E}_{X\sim P} f(X)| \underset{\text{Jensen's ineq.}}{\leq} \mathbb{E}_{X\sim P}|(f(X))| = \mathbb{E}_{X\sim P}|\langle f, k(X,\cdot)\rangle|$$

$$\underset{\text{C-S ineq.}}{\leq} \mathbb{E}_{X\sim P}||f||\sqrt{k(X,X)}$$

Where the last line holds by Cauchy Schwarz and the same trick we used in the proof of the semidefiniteness of the RKHS.

Since this is bounded, we can apply Riesz representation theorem (Thm 2) once again to represent the expectation as an inner product.

$$\exists h \in \mathcal{H} \text{ s.t. } |\mathbb{E}_{X\sim P} f(X)| = \langle f, h\rangle$$

Let $f = k(x,\cdot)$ for some $x \in X$. $h(x) = |\mathbb{E}_{X\sim P} k(x,\cdot)| = \int k(x,x')d\mathbb{P}(x')$ which implies $h = \int k(\cdot,x')d\mathbb{P}(x') = \mu_{\mathbb{P}}$.

□

We now have all the tools we need to begin looking into applications. Before we turn away, we want to demonstrate what kind of information the kernel mean embedding keeps. Borrowing from Muandet et al 2017 Section 3.1, if we set the kernel to be the linear kernel, i.e. $k(x,x') = \langle x, x'\rangle$, $\mu_{\mathbb{P}} = E_{x\sim\mathbb{P}}(x)$, the mean embedding is the expected value of the distribution $\mathbb{P}$.

A slightly more complicated kernel is the exponential kernel, $k(x,x') = \exp(\langle x, x'\rangle)$. In this case, $\mu_{\mathbb{P}} = E_{x\sim\mathbb{P}}(e^{\langle\cdot,\cdot\rangle})$ which is the moment generating function (if it exists for this distribution), which uniquely identifies distributions. So we see critical quantities are represented by the kernel mean embedding.

2 Regularization and Semiparametrics

2.1 Generalizing the James-Stein Shrinkage Effect to Arbitrary Distributions

Suppose we are trying to estimate the mean of an arbitrary Gaussian $\mathcal{N}(\mu, \sigma^2 I_d)$ for which we have a number of samples x_i. The obvious estimator, $\hat{\mu} = \frac{1}{N}\sum_i x_i$, the empirical mean, has favorable statistical properties, like being consistent, unbiased, asmyptotically normal with appropriate scaling and being the maximum likelihood estimate.

But in a stunning result, James and Stein (1961) introduce a biased estimator...

$$\hat{\mu}_{JS} = (1 - \frac{(m-2)\sigma^2/2}{||\hat{\mu}||^2})\hat{\mu}$$

...which, over any value of $\mu \in \mathbb{R}$, dominates the standard estimator under the mean squared loss function. i.e.

$$\mathbb{E}(||\hat{\mu}_{JS} - \mu||^2) \leq \mathbb{E}(||\hat{\mu} - \mu||^2)$$

It is important to note that μ_{JS} itself is dominated by a whole host of estimators. But what is important is that "shrinking" our estimate towards zero has favorable statistical properties, which we will investigate further in the regularization section.

Does this result hold for parameters other than the mean, and for arbitrary distributions beyond Gaussians? It depends on some conditions. But these conditions can be formulated with the formalism we have developed.

In Muandet et al. 2014 [4], two conditions are needed, one on the kernel, and one of the probability measure $\mathbb{P}$.

Condition (i): $k(x_i, x_j) = \phi(x_i - x_j)$ for some ϕ positive definite with $\phi(0) > 0$ Condition (ii): Boundedness of characteristic function.

Note that now μ refers to the kernel mean embedding, as opposed to a mean of a multivariate distribution. The "naive" estimator in this setting is now $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot)$

Theorem 8. (Muandet et al. 2014 [4])
For the following shrinkage estimator,

$$\hat{\mu}_{\text{JS}} = \hat{\mu}\big(1 - \frac{2\mathbb{E}||\hat{\mu} - \mu||^2||}{\mathbb{E}(||\hat{\mu} - \mu||^2) + ||\mu||^2}\big)$$

and for all kernels k, and probability measures $\mathbb{P}$ satisfying conditions (i) + (ii), $\mathbb{E}||\hat{\mu}_{\mathbb{P}}^{JS} - \mu_{\mathbb{P}}||_{\mathcal{H}}^2 < \mathbb{E}||\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}||_{\mathcal{H}}^2$ where $\mathcal{H}$ is the RKHS induced by kernel k.

Proof. Let $\alpha = \frac{2\mathbb{E}||\hat{\mu}-\mu||^2||}{\mathbb{E}(||\hat{\mu}-\mu||^2)+||\mu||^2}$

$$\mathbb{E}||\hat{\mu}_{\text{JS}} - \mu||_{\mathcal{F}}^2 = ||\mathbb{E}[\hat{\mu}_{\text{JS}}] - \mu||_{\mathcal{F}}^2 + \mathbb{E}||\hat{\mu}_{\text{JS}} - \mathbb{E}\hat{\mu}_{\text{JS}}||_{\mathcal{F}}^2$$

$$= ||\text{Bias}(\hat{\mu}_{\text{JS}})||_{\mathcal{F}}^2 + ||\text{Var}(\hat{\mu}_{\text{JS}})||$$

where

$$\text{Bias}(\hat{\mu}_{\text{JS}}) = \mathbb{E}[\hat{\mu}_{\text{JS}}] - \mu = \mathbb{E}[(1-\alpha)\hat{\mu}] - \mu = \mathbb{E}[\hat{\mu}] - \alpha\mathbb{E}[\hat{\mu}] - \mu = -\alpha\mu$$

and

$$\text{Var}(\hat{\mu}_{\text{JS}}) = (1-\alpha)^2\mathbb{E}||\hat{\mu} - \mu||_{\mathcal{F}}^2$$

Hence,

$$\mathbb{E}||\hat{\mu}_{JS} - \mu||_{\mathcal{F}}^2 = \alpha^2||\mu||_{\mathcal{F}}^2 + (1-\alpha^2)(\mathbb{E}(||\hat{\mu} - \mu||^2)$$

Therefore,

$$\mathbb{E}||\hat{\mu}_{JS} - \mu||_{\mathcal{F}}^2 - \mathbb{E}||\hat{\mu} - \mu||_{\mathcal{F}}^2 = \alpha^2[\mathbb{E}||\hat{\mu} - \mu||_{\mathcal{F}}^2 + ||\mu||_{\mathcal{F}}^2] - 2\alpha(\mathbb{E}||\hat{\mu} - \mu||_{\mathcal{F}}^2) < 0$$

when $\alpha = \frac{2\mathbb{E}||\hat{\mu}-\mu||^2||}{\mathbb{E}(||\hat{\mu}-\mu||^2)+||\mu||^2}$

□

2.2 Tikhonov Regularization

Why does shrinkage work? The answer to this question may be provided by Tikhonov regularization, a sort of a "generalization" of shrinkage.

$$\mathcal{R}(f) + \lambda\Omega(f)$$

where Ω is any monotonic function of f, and $\lambda \in \mathbb{R}$ is a regularization parameter, controlling how much we should bias our solution to be smooth. We will see ways of selecting a regularization parameter.

Notice that by the reproducing property, $f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$. Hence, setting $\Omega(f(\alpha)) = ||\alpha||$ is still a strictly monotonic function of f, and for λ, we recover the James-Stein shrinkage estimator, where the coefficient values are incentivized to go to 0

Let us now use $\Omega(f) = ||f||^2$. This is easy to interpret, simple and yields the benefits of regularization. It also has a good interpretation in terms of the kernel, as the norm of a function is nothing more than the inner product in high-dimensional space of that function with itself, which in turn is simply a kernel. Therefore, the choice of kernel has a critical impact on regularization, and hence on the function outputted.

We can be a bit more clearer and precise if we investigate the Fourier transform.

The Fourier transform for a function f and $\omega \in \mathbb{R}^d$,

$$T[f](\omega) = \frac{1}{2\pi^{d/2}} \int f(x) e^{-\sqrt{-1}x^T\omega} dx$$

Suppose our kernel $k(x, y)$ is shift invariant, i.e. there is some continuous ψ such that $k(x, y) = \psi(x - y)$. By Kanagawa et al. 2018 [5] the RKHS induced by such shift invariant kernels are,

$$\mathcal{F} = \{f \in L^2 | ||f||_{\mathcal{F}}^2 = \frac{1}{(2\pi^{d/2}} \int \frac{|T[f](\omega)|^2}{T[\psi](\omega)} < \infty\}$$

If ψ is smooth, then its the tail of its Fourier transform ought to decay quickly as $|\omega| \to \infty$, implying that f must be smooth in order for the integral to not diverge. So we see that only a certain set of smooth functions are admitted on the space $\mathcal{F}$ in the first place.

Now, going back to our discussion on Tikhonov Regularization, as we are now penalizing high values of the norm, this means, for a fixed ψ, we punish f that do not decay as quickly as ψ, and hence we bias our selection towards "smoother" functions.

2.3 Connection to Inverse Problems and Well-Posedness

Tikhonov regularization originated in the inverse problem literature, which we will explain in a bit. Oftentimes in the learning or statistics literature, there is a lot of imprecisely defined terms with regards to the inverse problem literature. (for example, using Hadamard well-posedness when they actually mean Tikhonov well-posedness). In the next section we will discuss inverse problems, why Tikhonov regularization was able to fix the inverse problems, and finally, the connections between the inverse problems literature and statistics/learning. On the latter point, RKHS have appeared to make the connections explicit.

In the inverse problems literature, it is desired that problems follow the following three "well-posedness" properties of Hadamard:

1) Solution exists
2) Solution is unique
3) Continuous in the data

And indeed, Tikhonov regularization is a way to make these three conditions hold, as we will see.

The non-regularized risk functional is currently:

$$\min_{\hat{f}\in\mathcal{F}} \sum_{i=1}^{N} ||y_i - \hat{f}(x_i)|| = \min_{\hat{f}\in\mathcal{F}} \mathcal{R}(\hat{f})$$

In this current formulation, there is no guarantee that any of the three conditions of Hadamard hold. However, suppose for a moment that that $\mathcal{F}$ is a compact set. Suddenly, all three properties of Hadamard appear. The continuity on the data follows from the continuity of norms, whereas the existence of a minima is guaranteed by the fact that a continuous function on a compact set attains its minimum. The third, uniqueness of solution, will be proven below.

Therefore, compactness of the set we optimize over achieves Hadamard's well-posedness criteria. It turns out, that adding the Tikhonov regularization penalty is as if we are optimizing a compact set, even if our original set $\mathcal{F}$ is not compact! This is formalized in the following theorem.

Theorem 9. Suppose that a solution f^* exists for both sides of the below equation. Then, given a not necessarily compact RKHS $\mathcal{F}$ and a regularization parameter $\lambda \in \mathbb{R}$, there exists a compact ball $K \subset \mathcal{F}$ with radius R such that the equation below holds.

Conversely, for any compact ball K with radius R, there exists a regularization parameter λ and a RKHS $\mathcal{F}$ such that the equation below holds.

$$\operatorname*{argmin}_{\hat{f}\in K} \mathcal{R}(\hat{f}) = \operatorname*{argmin}_{\hat{f}\in\mathcal{F}} \mathcal{R}(\hat{f}) + \lambda||\hat{f}||^2_{\mathcal{H}}$$

Proof. For direction. Suppose, by way of contradiction, that $f^* = \operatorname*{argmin}_{\hat{f}\in\mathcal{F}} \mathcal{R}(\hat{f}) + \lambda||\hat{f}||^2_{\mathcal{H}}$, yet there is some f' in a compact ball K with radius $||f^*||^2$ such that $\mathcal{R}(\hat{f^*}) > \mathcal{R}(\hat{f'})$. Since $f' \in K$, $||f'||^2 \leq ||f*||^2$ as well, so $\mathcal{R}(\hat{f}*) + \lambda||\hat{f}*||^2_{\mathcal{H}} > \mathcal{R}(\hat{f'}) + \lambda||\hat{f'}||^2_{\mathcal{H}}$, leading to a contradiction.

Converse. Let K be any compact ball with radius R and let f^* be its optimal solution. We formulate the Lagrangian dual of the constrained optimization problem.

$$\begin{aligned} \operatorname*{argmin}_{\hat{f}\in\mathcal{F}} \max_{\lambda\geq 0} \quad & \mathcal{R}(\hat{f}) + \lambda(||\hat{f}||^2_{\mathcal{F}} - R^2) \\ \text{subject to} \quad & \lambda(||f||^2 - R^2) \leq 0 \end{aligned}$$

In order for this dual to be equivalent to the original primal problem, we will prove strong Lagrangian duality. Since, $\mathcal{R}$ is a convex (indeed strongly convex) functional on a

convex domain K of a vector space $\mathcal{F}$, and the constraint is convex when viewed from the perspective of $||f||^2$, strong Lagrangian duality holds.

Now, suppose we have found a λ^* that optimizes the inner maximum. Now, plugging this in into the dual yields the following optimization problem.

$$\underset{\hat{f}\in\mathcal{F}}{\operatorname{argmin}}\, \mathcal{R}(\hat{f}) + \lambda(||\hat{f}||_{\mathcal{H}}^2 - R^2)$$

Going back to the Tihkonov formulation, if we select as our regularization parameter λ^*,

$$\underset{\hat{f}\in\mathcal{F}}{\operatorname{argmin}}\, \mathcal{R}(\hat{f}) + \lambda^*||\hat{f}||_{\mathcal{H}}^2$$

we can simply subtract R^2 as it is a constant and does not affect the arg minimization.

$$\underset{\hat{f}\in\mathcal{F}}{\operatorname{argmin}}\, \mathcal{R}(\hat{f}) + \lambda^*(||\hat{f}||_{\mathcal{H}}^2 - R^2)$$

Now the two optimization problems are identical.

□

This theorem gives us a hint as to why Tikhonov Regularization has been effective in practice. Tikhonov Regularization is as if we are optimizing over a compact space, and compactness has many excellent topological properties.

However, while intuitive to an extent, this does not paint a full picture of why topological properties like compactness ought to be very useful in statistics. It is a bit of a mystery why solutions, such as well-posedness, that originated in the inverse problems literature have applications to statistics. In the next section, we will rigorously show that for certain nice spaces, of which RKHS are one, that well-posedness of the optimization problem, and good learning properties such as consistency and generalizability are equivalent. This a deep result, philosophically but also in application, as it suggests the full force of the inverse problems literature can be used profitably in statistics and learning theory.

2.4 Why Well-Posed Optimization Problems Have Good Statistical Properties

Regularization originated in the inverse problems literature, so it will be useful to briefly review this literature.

We begin with a canonical inverse problem, (called of the first kind). Given some g in some Hilbert space $\mathcal{H}$, and A an operator mapping from $\mathcal{H}$ into some other Hilbert space, we seek a u that will satisfy the following equation.

$$g = Au$$

There is no guarantee in general that any of the three of the well-posed properties of Hadamard hold. However, we replace the exact equation above with an "approximation".

$$\min_u ||g - Au||^2$$

If Pg is within the range of A, then existence and uniqueness are restored by the generalized Moore-Penrose solution.

Solution to the above is

$$Au = Pg$$

However, the issue of continuity on the data g is not guaranteed. For this we turn to Tihkonov Regularization.

$$\min_u ||g - Au||^2 + \lambda ||u||^2$$

This makes the above continuous for similar reasons as we saw above (and that we will see again below).

Now, let us merge this inverse problem notation with what we have previously. In order to do so, we introduce an operator from (Smale and Zhou 2004 [6]), the "sampling" operator which is a function of sampled data as opposed to the function. Let $x \in \mathbb{R}^n$. For any $y \in \mathbb{R}$ Let $A_x : \mathcal{H} \to \mathbb{R}^n$ by $A_x(f) = (f(x_i))_{i=1}^n$, basically the evaluation operator that we previously saw in Section 1.2 but this time x is a vector as opposed to a single point.

$$y = A_x f$$

The inverse of A_x has an interesting interpretation. Since A_x is continuous (by boundedness of evaluation operators in an RKHS), by operator inversion lemma (), A_x^{-1} exists and is also continuous. For $y \in \mathbb{R}$ it is equal to $A_x^{-1}(y) = \min_{f \in \mathcal{F}} \mathcal{R}(f; x, y) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$

Now our learning problem is in this notation

In this form, there is no general guarantee that any of the three criteria for well-posedness holds in this scenario. However, if we instead switch this to a minimal norm problem and use Tikhonov Regularization, which was precisely used to solve inverse problems of this form. As we showed above, Tikhonov Regularization is tantamount to maximizing over a compact set $K \subset \mathcal{F}$, and we use this form here.

$$\min_{f \in K} ||A_x^{-1} y - f||^2$$

Now all properties of Hadamard well-posedness appear. The continuity of A_x on the data follows from the continuity of norms, whereas the existence of a minima is guaranteed by the fact that a continuous function on a compact set attains its minimum. Finally, since in the original Tikhonov form the norm was strongly convex, the sum of a convex and strongly convex function is strongly convex, meaning there exists a unique minimum (Shalev-Shwartz & Ben-David 2014 [7])

Now, finally, we connect the stability property of the Tikhonov solution to statistical property, that of consistency.

The adjoint of our sampling operator A_x is $A_x^*(c) = \sum_{i=1}^n c_i k(x_i, \cdot)$ for $c \in \mathbb{R}^n$.

Theorem 10. (De Vito et al. 2005)
Let f^* be the Tikhonov solution, i.e. $f^* = (A_x^* A_x + \lambda I)^{-1} A_x^* y$ (De Vito et al. 2005)

$$\lim_{n \to \infty} \mathbb{P}[(\mathcal{R}(f^*) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)) \geq \epsilon] = 0$$

The proof is very involved, so we refer the reader to De Vito et al. 2005.

2.5 Representing Tikhonov Solutions In Terms of a Kernel

Now we go to one of the most fundamental and important results in RKHS (very possible the most fundamental result), the so-called Representer Theorem (Wahba Kimeldorf), where we can represent the Tikhonov solution of a nonparametric set in terms of a sum of the kernel.

Theorem 11. Let$f^* = \min_{\hat{f} \in \mathcal{F}} \mathcal{R}(\hat{f}) + ||\hat{f}||_{\mathcal{F}}^2$.
Then, $\forall x \in X, \;\; f^*(x) = \sum_{i=1}^N \gamma_i k(x, x_i)$, where k is the reproducing kernel of $\mathcal{F}$.

We delay the proof of this theorem just for a moment, as we prove a stronger and more "general" form, that of the semiparameteric Representer Theorem (which consists of both a nonparametric and a parametric portion).

We could have parametric assumptions for two different reasons. The first is a reflection of prior knowledge, which can be intuition, theory or prior studies on a topic. The second is that of model interpretability: the deficiencies of ordinary least squares are known, but the reason it permeates is that it gives a very simple and intuitive explanation. We can treat the infinite dimensional and ultra-flexible part as a "nuisance parameter", while focusing on the easy to interpret OLS coefficient. This is important for decision makers.

Let $\{\psi\}_i$ be a set of parametric functions and $h \in span\{\psi_p(x_i)\}$. This is the parametric part.

Hence, we need to adapt our risk function to now take two arguments. Let $\mathcal{R} : \mathcal{F} \times \mathbb{R}^l \to \mathbb{R}$.

$$\mathcal{R}((\hat{f}, \beta)) = \sum_{i=1}^N ||y_i - (\hat{f}(x_i) + \beta_i \psi(z_i))||^2$$

Let us use one function $\tilde{f} = (\hat{f} + h)$ so we can optimize over one variable as opposed to two.

This is Theorem 11, but now in semiparametric form.

Theorem 12. (Scholkopf and Smola 2001 [8])
Let$\tilde{f} = \min_{\hat{f}, \beta \in \mathcal{F}, \mathbb{R}^l} (\mathcal{R}(\hat{f}, \beta) + \lambda ||\hat{f}||_{\mathcal{F}}^2)$.
Then, $\forall x \in X, \;\; \tilde{f}(x) = \sum_{i=1}^m \gamma_i k(x, x_i) + \sum_{j=1}^l \beta_j z_j$, where k is the reproducing kernel of $\mathcal{F}$.

Proof. Our strategy is to show, separately for both the vanilla loss function and for the regularized term, the term that cannot be described by the span of the kernels must vanish. In other words, the orthogonal complement must vanish to achieve a minimzer.

Since the $span\{k(x_i, \cdot)\} = \mathcal{H}$ is a closed subspace , H, a Hilbert space, can be decomposed into the direct sum $\mathcal{H}_\perp \bigoplus \mathrm{H}$

Hence, the nonparametric portion of $\tilde{f}$, $\hat{f} \in \mathcal{H}$, can be uniquely represented by $\hat{f}_\perp + \hat{f}_\parallel$.

Let x_i be an arbitrary datapoint in our finite sample.

$$\tilde{f}(x_i) \underset{\text{rep prop. eq. 2}}{=} \langle \tilde{f}, k(x_j, \cdot)\rangle = \langle \hat{f}_\perp + \hat{f}_\parallel(x_i) + h(x_i), k(x_j, \cdot)\rangle$$

$$= \langle \hat{f}_\perp, k(x_j, \cdot)\rangle + \langle \hat{f}_\parallel, k(x_j, \cdot)\rangle + \langle h, k(x_j, \cdot)\rangle$$

Since $\hat{f}_\perp \in H_\perp$, the middle inner product is 0. Furthermore, by reproducing property,

Hence, as x_i was arbitrary, the vanilla loss function can be rewritten as. (the upshot is $\hat{f}_\perp$ does not contribute anything to the vanilla loss function)

$$\mathcal{R}((\hat{f}, \beta)) = \sum_{i=1}^{N} ||y_i - (\hat{f}_\parallel(x_i) + h(x_i)||^2$$

Now,

$$||\hat{f}||^2 = ||\hat{f}_\perp + \hat{f}_\parallel||^2 \underset{\text{parall. law}}{=} ||\hat{f}_\perp||^2 + ||\hat{f}_\parallel||^2 \geq ||\hat{f}_\parallel||^2$$

Since $H_\perp$ is a closed subspace, $0 \in H_\perp$ and therefore $f_\perp = 0$ is the minimizer of the regularizer.

We have proved for both the vanilla loss function and the regularizer that $_\perp$ vanishes, the optimum $\tilde{f}$ can be represented as, after applying the reproducing property and the defintion of h,

$$\tilde{f}(x) = \sum_j \alpha_j k(x, x_j) + \sum_p \beta_p \psi_p(x)$$

□

This is very useful result, as the optimal function in an infinite dimensional space can be represented by a finite sum.

A particularly useful and ubiquitous semiparametric model is when the parametric part consists of only one linear function $h(x) = \phi_1(x) = x$, i.e. a partially linear model.

Corollary 1.

$$\tilde{f}(x) = \sum_{i=1}^{m} \gamma_i k(x, x_i) + \beta x$$

3 Frequentist Nonparametrics

3.1 Robust Kernel Density Estimation

The workhorse of frequentist nonparametrics is the kernel density estimator. Most applications of kernel density estimators appear in the form.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x_i, x)$$

where $(x_i)_{i=1}^n$ is a set of sampled data. This is a specific form of a more general weighting scheme, with $w_i = \frac{1}{n}$ uniformly.

$$\hat{f}(x) = \sum_{i=1}^{n} w_i f(x_i, x) \text{ where } \sum_{i=1}^{n} w_i = 1$$

You might guess that, after the section on shrinkage, that perhaps the uniform weights $\frac{1}{n}$ are not the best set of weights. Indeed, in a certain sense, the uniform $\frac{1}{n}$ weights are much more sensitive to outliers (not robust).

Kim and Scott 2011 [10] proposed to embed the kernel into the RKHS, and consider. This is an excellent paper that demonstrates the algorithmic benefits of using kernel methods, and also the benefits to theoretical analysis.

Borrowing ideas from M-estimation in finite dimensions, they use a loss-function that is known to be more robust.

Letting ϕ as always be the feature map associated with k,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x, x_i) = \frac{1}{n}\sum_{i=1}^{n} \langle \phi(x), \phi(x_i) \rangle = \langle \phi(x), \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \rangle$$

So we see that the kernel density estimator is basically the inner product between the new point in high dimensional feature space with the centroid of the already existing points.

$$\hat{f}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$$

This would then imply that $\hat{f}(\cdot)$ is a minimizer (not necessarily the only one) of the following loss function,

$$\underset{g \in \mathcal{F}}{\operatorname{argmin}} \; ||\phi(x_i) - g||^2$$

Borrowing ideas from M-estimation theory, which was developed in the finite and parametric case to guard against outliers, the authors propose a modification of the above, but with a robust loss function instead.

$$\hat{f}(\cdot) = \underset{g \in \mathcal{F}}{\operatorname{argmin}} \; \rho(||\phi(x_i) - g||) \quad (1)$$

One example of such a loss function is Huber's loss function.

$$\rho(x) = \begin{cases} x^2/2 & \text{if } 0 \leq x \leq a \\ ax - a^2/2 & \text{if } a < x \end{cases}$$

The critical thing to notice is that if x exceeds a constant a then its influence decreases on our loss function and hence any optimizing algorithm will weigh it less.

Now, this equation is only useful if it can be implemented in an algorithm tractably. The first step is to see whether there is a convenient expression for robust KDE, a sort of a representer theorem (like the ones we saw for Tikhonov Regularization).

Theorem 13. Let $\hat{f}(\cdot)$ be the minimizer of (1). Then $\hat{f}(x)$ admits a representation of the form $\sum_{i=1}^{n} w_i k(x, x_i)$

Proof. By Lemma 1 of Scott and Kim 2011 [10] the following holds,

$$\frac{1}{n}\sum_{i=1}^{n} \rho'(||\phi(x_i) - \hat{f})(\phi(x_i) - \hat{f}) = 0$$

Solving for $\hat{f}$, we have $\hat{f} = \sum_{i=1}^{n} w_i \phi(x_i)$ where $w_i = (\sum_{j=1}^{n} \rho'(||\phi(x_j) - \hat{f}||_{\mathcal{F}}))^{-1} \rho'(||\phi(x_i) - \hat{f}||_{\mathcal{F}})$

By the fact that is non-decreasing, $w_i \geq 0$ and $\sum_i w_i = 1$ since we are each w_i is being normalized.

□

The authors then demonstrate the robustness of their scheme both with empirical experiments and also with theoretical analysis of robustness. The theoretical analysis of robustness is beyond the scope of this paper.

3.2 Hypothesis Testing for Distributions (Without the Curse of Dimensionality)

Do we have distributional equivalents to parametric statistical hypothesis testing?

We know, for example, that if the Central Limit Theorem holds, then we can test if a sample mean comes has a hypothesized value.

Why would we want to test two distributions with one another? Because we might be interested in comparing the full shapes of two distributions as opposed to a single or two parameters like the mean and standard deviation.

While indeed normal distributions are fully characterized by these two elements, it would be insufficient if we need to estimate other types of distributions.

Gretton et al. 2008 [11] develop a test statistic for comparing two distributions using the kernel mean embedding, called the maximum mean discrepancy.

The major innovation of this paper over previous versions of hypothesis testing (Anderson et al. 1994 [12]) is that two expensive computations of a kernel density estimate (which suffers from the curse of dimensionality, i.e. the computational costs grow exponential in the dimension of the data) are not required by this test, by usage of the kernel trick.

We begin with a general setting for estimating how much two distributions diverge. (This is called the integral probability metric by Gretton et al. 2008 [11])

$$\sup_{||f||_{\mathcal{F}} \leq 1} \Big[(\mathbb{E}_x(f(x) - \mathbb{E}_y(f(y)) \Big]^2$$

Now, let us set $\mathcal{F}$ to a RKHS $\mathcal{H}$ and see what happens. We can write the above as, using the Representer Theorem 11,

$$= \sup_{||f||_{\mathcal{H}} \leq 1} \Big[\langle \mu_p - \mu_q, f \rangle_{\mathcal{H}}\Big]^2 = ||\mu_p - \mu_q||^2_{\mathcal{H}} = ||\mu_p||^2_{\mathcal{H}} - 2\langle \mu_p \mu_q \rangle_{\mathcal{H}} + ||\mu_q||^2_{\mathcal{H}}$$

$$\underset{\text{kernel trick}}{=} \mathbb{E}_{x,x'} \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} - 2\mathbb{E}_{x,y} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} + \mathbb{E}_{y,y'} \langle \phi(y), \phi(y') \rangle_{\mathcal{H}} = MMD^2$$

Since we do not have access to the true expecations, we can replace each of the elements above with their empirical counterparts.

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

Our first theorem in this section proves that this empirical counterpart converges to the true MMD.

(Thm 10)

Theorem 14. Suppose $0 \leq k(x_i, x_j) \leq K$.

$$P(|\widehat{MMD}^2 - MMD^2| \geq \epsilon) \to 0$$

Proof. This is a simple application of Hoeffding's inequality, which bounds how much sample means deviate from the true mean. Since our random variables are bounded by assumption, we applying Hoeffding's to our current context:

$$P(\widehat{MMD}^2 - MMD^2) \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{K^2}}$$

As $n \to \infty$, the RHS goes to 0.

□

Hence, for a given significance α, and a bound on the kernel k, if we select the threshold $4K/\sqrt{n}\sqrt{\log \alpha^{-1}}$, then the hypothesis test is a consistent one (i.e. if the two distributions p, q are actually equal, then we never reject the null for any significance level α.)

References

[1] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. USA: John Wiley Sons, Inc., 1997.

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

[3] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, p. 1–141, 2017. [Online]. Available: http://dx.doi.org/10.1561/2200000060

[4] K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf, "Kernel mean shrinkage estimators," 2014.

[5] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, "Gaussian processes and kernel methods: A review on connections and equivalences," 2018.

[6] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, pp. 153–172, 2007.

[7] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.

[8] E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone, "Learning from examples as an inverse problem," *J. Mach. Learn. Res.*, vol. 6, pp. 883–904, 2005.

[9] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[10] J. Kim and C. D. Scott, "Robust kernel density estimation," 2011.

[11] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola, "A kernel method for the two-sample problem," 2008.

[12] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *Journal of Multivariate Analysis*, vol. 50, no. 1, pp. 41–54, 1994. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:jmvana:v:50:y:1994:i:1:p:41-54