Reproducing Kernel Hilbert Space

김지수 (Jisu KIM)

통계적 기계학습(Statistical Machine Learning), 2025 1학기 (spring)

The lecture note is a minor modification of the lecture notes from Prof Larry Wasserman and Ryan Tibshirani's "Statistical Machine Learning", and Arthur Gretton's "Reproducing kernel Hilbert spaces in Machine Learning" (https://www.gatsby.ucl.ac.uk/∼gretton/coursefiles/rkhscourse.html). Also, see Section 5.8 from [1].

1 Review

1.1 Basic Model for Supervised Learning

- Input(입력) / Covariate(설명 변수) : $x \in \mathbb{R}^d$, so $x = (x_1, \ldots, x_d)$.
- Output(출력) / Response(반응 변수): $y \in \mathcal{Y}$. If y is categorical, then supervised learning is "classification", and if y is continuous, then supervised learning is "regression".
- Model(모형) :

  $$y \approx f(x).$$

  If we include the error ϵ to the model, then it can be also written as

  $$y = \phi(f(x), \epsilon).$$

  For many cases, we assume additive noise, so

  $$y = f(x) + \epsilon.$$

- Assumption(가정): f belongs to a family of functions $\mathcal{M}$. This is the assumption of a model: a model can be still used when the corresponding assumption is not satisfied in your data.
- Loss function(손실 함수): $\ell(y, a)$. A loss function measures the difference between estimated and true values for an instance of data.
- Training data(학습 자료): $\mathcal{T} = \{(y_i, x_i), i = 1, \ldots, n\}$, where (y_i, x_i) is a sample from a probability distribution P_i. For many cases we assume i.i.d., or x_i's are fixed and y_i's are i.i.d..
- Goal(목적): we want to find f that minimizes the expected prediction error,

  $$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$

  Here, $\mathcal{F}$ can be different from $\mathcal{M}$; $\mathcal{F}$ can be smaller then $\mathcal{M}$.
- Prediction model(예측 모형): f^0 is unknown, so we estimate f^0 by $\hat{f}$ using data. For many cases we minimizes on the empirical prediction error, that is taking the expectation on the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i)}$.

  $$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_n} \left[\ell(Y, f(X))\right] = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

- Prediction(예측): if $\hat{f}$ is a predicted function, and x is a new input, then we predict unknown y by $\hat{f}(x)$.

1.2 Linear Regression

From the additive noise model

$$y = f(x) + \epsilon,\ f \in \mathcal{M},$$

Linear Regression Model (선형회귀모형) is that

$$\mathcal{M} = \mathcal{F} = \left\{ \beta_0 + \sum_{j=1}^{d} \beta_j x_j : \beta_j \in \mathbb{R} \right\}.$$

For estimating β, we use least squares: suppose the training data is $\{(y_i, x_{ij}) : 1 \le i \le n, 1 \le j \le p\}$. We use square loss

$$\ell(y, a) = (y - a)^2,$$

then the eimpirical loss becomes the residual sum of square (RSS) as

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2. \end{aligned}$$

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d)$ be the nimimizor of RSS, then the predicted function is

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j x_j.$$

1.3 Hölder Spaces and Sobolev Spaces

The class of Lipschitz functions $H(1, L)$ on $T \subset \mathbb{R}$ is the set of functions $g : T \to \mathbb{R}$ such that

$$|g(y) - g(x)| \le L|x - y| \quad \text{for all } x, y \in T.$$

A differentiable function is Lipschitz if and only if it has bounded derivative. Conversely a Lipschitz function is differentiable almost everywhere.

Let $T \subset \mathbb{R}$, let β be a positive integer, and let $L > 0$. The *Hölder class* $H(\beta, L)$ on T is the set of functions $g : T \to \mathbb{R}$ such that g is $\ell = \beta - 1$ times differentiable and satisfies

$$\left| g^{(\ell)}(y) - g^{(\ell)}(x) \right| \le L\,|x - y|\,, \quad \text{for all } x, y \in T.$$

(There is an extension to real valued β but we will not need that.) If $g \in H(\beta, L)$ and $\ell = \beta - 1$, then we can define the Taylor approximation of g at x by

$$\tilde{g}(y) = g(y) + (y - x)g'(x) + \cdots + \frac{(y - x)^\ell}{\ell!} g^{(\ell)}(x)$$

and then

$$|g(y) - \tilde{g}(y)| \le |y - x|^\beta.$$

The definition for higher dimensions is similar. Let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^d$. Let β be a positive integer and $L > 0$. Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

Define the *Hölder class* $H_d(\beta, L)$ on $\mathcal{X}$ as

$$H_d(\beta, L) = \left\{ g :\ |D^s g(x) - D^s g(y)| \le L\,\|x - y\|_2\,, \quad \text{for all } s \text{ such that } |s| = \beta - 1, \text{ and all } x, y \right\}. \tag{1}$$

For example, if $d = 1$ and $\beta = 2$ this means that

$$|g'(x) - g'(y)| \leq L |x - y|, \quad \text{for all } x, y.$$

The most common case is $\beta = 2$; roughly speaking, this means that the functions have bounded second derivatives.

Again, if $g \in H_d(\beta, L)$ then $g(x)$ is close to its Taylor series approximation:

$$|g(u) - g_{x,\beta}(u)| \leq L \|u - x\|_2^\beta, \tag{2}$$

where

$$g_{x,\beta}(u) = \sum_{|s|<\beta} \frac{(u-x)^s}{s!} D^s g(x). \tag{3}$$

In the common case of $\beta = 2$, this means that

$$\left| p(u) - [p(x) + (x-u)^\top \nabla p(x)] \right| \leq L \|x - u\|_2^2.$$

The Sobolev class $S_1(\beta, L)$ on a bounded set $\mathcal{X} \subset \mathbb{R}$ is the set of β times differentiable functions (technically, it only requires weak derivatives) $g : T \to \mathbb{R}$ such that

$$\int_{\mathcal{X}} (g^{(\beta)}(x))^2 dx \leq L^2.$$

Again this extends naturally to $\mathbb{R}^d$. Also, there is an extension to non-integer β.

It is worth noting that if $\mathcal{X}$ is bounded, then the Sobolev $S_d(\beta, L)$ and Holder $H_d(\beta, L)$ classes are *equivalent* in the following sense: given $S_d(\beta, L)$ for a constant $L > 0$, there are $L_0, L_1 > 0$ such that

$$H_d(\beta, L_0) \subseteq S_d(\beta, L) \subseteq H_d(\beta, L_1).$$

The first containment is easy to show; the second is far more subtle, and is a consequence of the Sobolev embedding theorem.

2 Introduction

A function space is a set of functions $\mathcal{F}$ that has some structure. Often a nonparametric regression function or classifier is chosen to lie in some function space, where the assumed structure is exploited by algorithms and theoretical analysis. Here we review some basic facts about function spaces.

As motivation, consider nonparametric regression. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we want to estimate $f^0(x) = \mathbb{E}(Y|X = x)$. We cannot simply choose f^0 to minimize the training error $\sum_i (Y_i - f^0(X_i))^2$ as this will lead to interpolating the data. One approach is to minimize $\sum_i (Y_i - f^0(X_i))^2$ while restricting f^0 to be in a well behaved function space.

3 Hilbert Spaces

Let V be a vector space over $\mathbb{R}$. A *norm* is a mapping $\|\cdot\| : V \to [0, \infty)$ that satisfies

1. $\|x\| = 0$ if and only if $x = 0$.
2. $\|ax\| = |a| \|x\|$ for all $a \in \mathbb{R}$.
3. $\|x + y\| \leq \|x\| + \|y\|$.

Some examples of normed vector spaces are:

- $(\mathbb{R}, |\cdot|)$.
- $\mathbb{R}^d$: $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p \right)^{1/p}$, $p \geq 1$.

  - $p = 1$: Manhattan
  - $p = 2$: Euclidean
  - $p \to \infty$: maximum norm, $\|x\|_\infty = \max_i |x_i|$
- $C[a,b]$: $\|f\|_p = \left(\int_a^b |f(x)|^p \, dx\right)^{1/p}$, $p \geq 1$.

An *inner product* is a mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies, for all $x, y, z \in V$ and $a, b \in \mathbb{R}$:

1. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$
2. $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$
3. $\langle x, y \rangle = \langle y, x \rangle$

An inner product defines a norm $\|v\| = \sqrt{\langle v, v \rangle}$.

Some examples of inner product spaces are:

- $\mathbb{R}^d$: $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$.
- $C[a,b]$: $\langle f, g \rangle = \int_a^b f(x)g(x)dx$.
- $\mathbb{R}^{d \times d}$: $\langle A, B \rangle = tr\left(AB^\top\right)$.

Two vectors x and y are *orthogonal* if $\langle x, y \rangle = 0$.

Some key relations in inner product space:

- $|\langle x, y \rangle| \leq \|x\| \, \|y\|$ (Cauchy-Schwarz inequality)
- $2\|x\|^2 + 2\|y\|^2 = \|x + y\|^2 + \|x - y\|^2$ (the parallelogram law)
- $4\langle x, y \rangle = \|x + y\|^2 - \|x - y\|^2$ (the polarization identity)
- $x \perp y \implies \|x\|^2 + \|y\|^2 = \|x + y\|^2$ (Pythagorean theorem)

A sequence $\{x_n\}_{n=1}^\infty$ of a normed space is said to *converge* to x if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $\|x_n - x\| < \epsilon$. (which we also say that $\|x_n - x\| \to 0$ as $n \to \infty$) A sequence $\{x_n\}_{n=1}^\infty$ of a normed space is a *Cauchy sequence* if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$, $\|x_n - x_m\| < \epsilon$. (which we also say that $\|x_n - x_m\| \to 0$ as $n, m \to \infty$) Convergent $\Rightarrow$Cauchy, but Cauchy $\not\Rightarrow$Convergent.

A space is *complete* if every Cauchy sequence converges to a limit. A complete, normed space is called a *Banach space*. A *Hilbert space* is a complete, inner product space. Every Hilbert space is a Banach space but the reverse is not true in general. In a Hilbert space, we write $f_n \to f$ to mean that $||f_n - f|| \to 0$ as $n \to \infty$. Note that $||f_n - f|| \to 0$ does NOT imply that $f_n(x) \to f(x)$. For this to be true, we need the space to be a reproducing kernel Hilbert space which we discuss later.

If V is a Hilbert space and L is a closed subspace then for any $v \in V$ there is a unique $y \in L$, called the *projection* of v onto L, which minimizes $\|v - z\|$ over $z \in L$. The set of elements orthogonal to every $z \in L$ is denoted by $L^\perp$. Every $v \in V$ can be written uniquely as $v = w + z$ where z is the projection of v onto L and $w \in L^\perp$. In general, if L and M are subspaces such that every $\ell \in L$ is orthogonal to every $m \in M$ then we define the *orthogonal sum* (or *direct sum*) as

$$L \oplus M = \{\ell + m : \ell \in L, m \in M\}. \tag{4}$$

A set of vectors $\{e_t, t \in T\}$ is *orthonormal* if $\langle e_s, e_t \rangle = 0$ when $s \neq t$ and $\|e_t\| = 1$ for all $t \in T$. If $\{e_t, t \in T\}$ are orthonormal, and the only vector orthogonal to each e_t is the zero vector, then $\{e_t, t \in T\}$ is called an *orthonormal basis*. Every Hilbert space has an orthonormal basis. A Hilbert space is *separable* if there exists a countable orthonormal basis.

Theorem. *Let V be a separable Hilbert space with countable orthonormal basis $\{e_1, e_2, \ldots\}$. Then, for any $x \in V$, we have $x = \sum_{j=1}^\infty \theta_j e_j$ where $\theta_j = \langle x, e_j \rangle$. Furthermore, $\|x\|^2 = \sum_{j=1}^\infty \theta_j^2$, which is known as* Parseval's identity.

The coefficients $\theta_j = \langle x, e_j \rangle$ are called *Fourier coefficients*.

Some examples of inner product spaces are:

- $\mathbb{R}^d$: $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$.

- $L_2(\mathcal{X})$: If ν is a measure on $\mathcal{X} \subset \mathbb{R}^d$, then the space

$$L_2(\mathcal{X};\nu) = \left\{ f : \mathcal{X} \to \mathbb{R} : \|f\|_2 = \left(\int_{\mathcal{X}} |f(x)|^2 \, d\nu(x) \right)^{1/2} < \infty \right\}$$

  is a Hilbert space with inner product

$$\langle f, g \rangle = \int_{\mathcal{X}} f(x) g(x) d\nu(x).$$

.

Theorem (Riesz representation theorem)**.** *In a Hilbert space $\mathcal{X}$, for every continuous linear functional $L : \mathcal{X} \to \mathbb{R}$, there exists a unique $y \in \mathcal{X}$ such that*

$$Lx = \langle x, y \rangle .$$

4 L_p Spaces

Let $\mathcal{F}$ be a collection of functions taking $[a, b]$ into $\mathbb{R}$. The L_p norm on $\mathcal{F}$ is defined by

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{1/p} \tag{5}$$

where $0 < p < \infty$. For $p = \infty$ we define

$$\|f\|_\infty = \sup_x |f(x)|. \tag{6}$$

Sometimes we write $\|f\|_2$ simply as $\|f\|$. The space $L_p(a, b)$ is defined as follows:

$$L_p(a, b) = \left\{ f : [a, b] \to \mathbb{R} : \ \|f\|_p < \infty \right\}. \tag{7}$$

Every L_p is a Banach space. Some useful inequalities are:

Cauchy-Schwartz	$\left(\int f(x)g(x)dx\right)^2 \leq \int f^2(x)dx \int g^2(x)dx$
Minkowski	$\|f + g\|_p \leq \|f\|_p + \|g\|_p$ where $p > 1$
Hölder	$\|fg\|_1 \leq \|f\|_p \|g\|_q$ where $(1/p) + (1/q) = 1$.

Special Properties of L_2. As we mentioned earlier, the space $L_2(a, b)$ is a Hilbert space. The *inner product* between two functions f and g in $L_2(a, b)$ is $\int_a^b f(x)g(x)dx$ and the *norm* of f is $\|f\|^2 = \int_a^b f^2(x)\, dx$. With this inner product, $L_2(a, b)$ is a separable Hilbert space. Thus we can find a countable orthonormal basis $\phi_1, \phi_2, \ldots$; that is, $\|\phi_j\| = 1$ for all j, $\int_a^b \phi_i(x)\, \phi_j(x)dx = 0$ for $i \neq j$ and the only function that is orthogonal to each ϕ_j is the zero function. (In fact, there are many such bases.) It follows that if $f \in L_2(a, b)$ then

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x) \tag{8}$$

where

$$\theta_j = \int_a^b f(x)\, \phi_j(x)\, dx \tag{9}$$

are the coefficients. Also, recall Parseval's identity

$$\int_a^b f^2(x)dx = \sum_{j=1}^{\infty} \theta_j^2. \tag{10}$$

The set of functions

$$\left\{ \sum_{j=1}^{n} a_j \phi_j(x) : \ a_1, \ldots, a_n \in \mathbb{R} \right\} \tag{11}$$

is the called the *span* of $\{\phi_1, \ldots, \phi_n\}$. The projection of $f = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$ onto the span of $\{\phi_1, \ldots, \phi_n\}$ is $f_n = \sum_{j=1}^{n} \theta_j \phi_j(x)$. We call f_n the n-term *linear approximation* of f. Let Λ_n denote all functions of the form $g = \sum_{j=1}^{\infty} a_j \phi_j(x)$ such that at most n of the a_j's are non-zero. Note that Λ_n is not a linear space, since if $g_1, g_2 \in \Lambda_n$ it does not follow that $g_1 + g_2$ is in Λ_n. The best approximation to f in Λ_n is $f_n = \sum_{j \in A_n} \theta_j \, \phi_j(x)$ where A_n are the n indices corresponding to the n largest $|\theta_j|$'s. We call f_n the n-term *nonlinear approximation* of f.

The *Fourier basis* on $[0, 1]$ is defined by setting $\phi_1(x) = 1$ and

$$\phi_{2j}(x) = \frac{1}{\sqrt{2}} \cos(2j\pi x), \quad \phi_{2j+1}(x) = \frac{1}{\sqrt{2}} \sin(2j\pi x), \quad j = 1, 2, \ldots \tag{12}$$

The *cosine basis* on $[0, 1]$ is defined by

$$\phi_0(x) = 1, \qquad \phi_j(x) = \sqrt{2} \cos(2\pi j x), \qquad j = 1, 2, \ldots. \tag{13}$$

The *Legendre basis* on $(-1, 1)$ is defined by

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x), \ \ldots \tag{14}$$

These polynomials are defined by the relation

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n. \tag{15}$$

The Legendre polynomials are orthogonal but not orthonormal, since

$$\int_{-1}^{1} P_n^2(x) dx = \frac{2}{2n + 1}. \tag{16}$$

However, we can define modified Legendre polynomials $Q_n(x) = \sqrt{(2n + 1)/2} \, P_n(x)$ which then form an orthonormal basis for $L_2(-1, 1)$.

The *Haar basis* on [0,1] consists of functions

$$\left\{ \phi(x), \psi_{jk}(x) : \ j = 0, 1, \ldots, \ k = 0, 1, \ldots, 2^j - 1 \right\} \tag{17}$$

where

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \le x < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

$\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$ and

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \le x \le \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < x \le 1. \end{cases} \tag{19}$$

This is a doubly indexed set of functions so when f is expanded in this basis we write

$$f(x) = \alpha \phi(x) + \sum_{j=1}^{\infty} \sum_{k=1}^{2^j - 1} \beta_{jk} \psi_{jk}(x) \tag{20}$$

where $\alpha = \int_0^1 f(x) \, \phi(x) \, dx$ and $\beta_{jk} = \int_0^1 f(x) \, \psi_{jk}(x) \, dx$. The Haar basis is an example of a *wavelet basis*.

Let $[a, b]^d = [a, b] \times \cdots \times [a, b]$ be the d-dimensional cube and define

$$L_2\left([a, b]^d\right) = \left\{ f : [a, b]^d \to \mathbb{R} : \ \int_{[a,b]^d} f^2(x_1, \ldots, x_d) \, dx_1 \ldots dx_d < \infty \right\}. \tag{21}$$

Suppose that $\mathcal{B} = \{\phi_1, \phi_2, \ldots\}$ is an orthonormal basis for $L_2([a, b])$. Then the set of functions

$$\mathcal{B}^d = \mathcal{B} \otimes \cdots \otimes \mathcal{B} = \left\{ \phi_{i_1}(x_1) \, \phi_{i_2}(x_2) \cdots \phi_{i_d}(x_d) : \ i_1, i_2, \ldots, i_d \in \{1, 2, \ldots, \} \right\}, \tag{22}$$

is called the *tensor product* of $\mathcal{B}$, and forms an orthonormal basis for $L_2([a, b]^d)$.

5 Hölder Spaces

Let β be a positive integer.[1] Let $T \subset \mathbb{R}$. The Holder space $H(\beta, L)$ is the set of functions $g : T \to \mathbb{R}$ such that

$$|g^{(\beta-1)}(y) - g^{(\beta-1)}(x)| \leq L|x-y|, \quad \text{for all } x, y \in T. \tag{23}$$

The special case $\beta = 1$ is sometimes called the Lipschitz space. If $\beta = 2$ then we have

$$|g'(x) - g'(y)| \leq L\,|x-y|, \quad \text{for all } x, y.$$

Roughly speaking, this means that the functions have bounded second derivatives.

There is also a multivariate version of Holder spaces. Let $T \subset \mathbb{R}^d$. Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and

$$D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

The Hölder class $H(\beta, L)$ is the set of functions $g : T \to \mathbb{R}$ such that

$$|D^s g(x) - D^s g(y)| \leq L\|x-y\|^{\beta-|s|} \tag{24}$$

for all x, y and all s such that $|s| = \beta - 1$.

If $g \in H(\beta, L)$ then $g(x)$ is close to its Taylor series approximation:

$$|g(u) - g_{x,\beta}(u)| \leq L\|u-x\|^{\beta} \tag{25}$$

where

$$g_{x,\beta}(u) = \sum_{|s|\leq\lfloor\beta\rfloor} \frac{(u-x)^s}{s!} D^s g(x). \tag{26}$$

In the case of $\beta = 2$, this means that

$$|g(u) - [g(x) + (x-u)^T \nabla g(x)]| \leq L||x-u||^2.$$

We will see that in function estimation, the optimal rate of convergence over $H(\beta, L)$ under L_2 loss is $O(n^{-2\beta/(2\beta+d)})$.

6 Sobolev Spaces

Let f be integrable on every bounded interval. Then f is *weakly differentiable* if there exists a function f' that is integrable on every bounded interval, such that $\int_x^y f'(s)ds = f(y) - f(x)$ whenever $x \leq y$. We call f' the *weak derivative* of f. Let $D^j f$ denote the j^{th} weak derivative of f.

The *Sobolev space of order m* is defined by

$$S_{m,p} = \Big\{ f \in L_p(0,1) : \ \|D^m f\| \in L_p(0,1) \Big\}. \tag{27}$$

The *Sobolev ball of order m and radius c* is defined by

$$S_{m,p}(c) = \Big\{ f : \ f \in S_{m,p}, \ \|D^m f\|_p \leq c \Big\}. \tag{28}$$

For the rest of this section we take $p = 2$ and write S_m instead of $S_{m,2}$

Theorem. *The Sobolev space S_m is a Hilbert space under the inner product*

$$\langle f, g \rangle = \sum_{k=0}^{m-1} f^{(k)}(0) g^{(k)}(0) + \int_0^1 f^{(k)}(x) g^{(k)}(x)\, dx. \tag{29}$$

[1] It is possible to define Holder spaces for non-integers but we will not need this generalization.

Define

$$K(x,y) = \sum_{k=1}^{m-1} \frac{1}{k!} x^k y^k + \int_0^{x \wedge y} \frac{(x-u)^{m-1}(y-u)^{m-1}}{(m-1)!^2} du. \tag{30}$$

Then, for each $f \in S_m$ we have

$$f(y) = \langle f, K(\cdot, y) \rangle \tag{31}$$

and

$$K(x,y) = \langle K(\cdot, x), K(\cdot, y) \rangle. \tag{32}$$

We say that K is a kernel for the space and that S_m is a *reproducing kernel Hilbert space* or *RKHS*.

It follows from Mercer's theorem (Theorem 7.5) that there is an orthonormal basis $\{e_1, e_2, \ldots, \}$ for $L_2(a,b)$ and real numbers $\lambda_1, \lambda_2, \ldots$ such that

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \, e_j(x) \, e_j(y). \tag{33}$$

The functions e_j are eigenfunctions of K and the λ_j's are the corresponding eigenvalues,

$$\int K(x,y) \, e_j(y) \, dy = \lambda_j e_j(x). \tag{34}$$

Hence, the inner product defined in (29) can be written as

$$\langle f, g \rangle = \sum_{j=0}^{\infty} \frac{\theta_j \, \beta_j}{\lambda_j} \tag{35}$$

where $f(x) = \sum_{j=0}^{\infty} \theta_j e_j(x)$ and $g(x) = \sum_{j=0}^{\infty} \beta_j e_j(x)$.

Next we discuss how the functions in a Sobolev space can be parameterized by using another convenient basis. An *ellipsoid* is a set of the form

$$\Theta = \left\{ \theta : \ \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{36}$$

where a_j is a sequence of numbers such that $a_j \to \infty$ as $j \to \infty$. If Θ is an ellipsoid and if $a_j^2 \sim (\pi j)^{2m}$ as $j \to \infty$, we call Θ a *Sobolev ellipsoid* and we denote it by $\Theta_m(c)$.

Theorem. *Let* $\{\phi_j, j = 0, 1, \ldots\}$ *be the* Fourier basis*:*

$$\phi_1(x) = 1, \quad \phi_{2j}(x) = \frac{1}{\sqrt{2}} \cos(2j\pi x), \quad \phi_{2j+1}(x) = \frac{1}{\sqrt{2}} \sin(2j\pi x), \quad j = 1, 2, \ldots \tag{37}$$

Then,

$$S_m(c) = \left\{ f : \ f = \sum_{j=1}^{\infty} \theta_j \phi_j, \ \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{38}$$

where $a_j = (\pi j)^m$ *for* j *even and* $a_j = (\pi(j-1))^m$ *for* j *odd. Thus, a Sobolev space corresponds to a Sobolev ellipsoid with* $a_j \sim (\pi j)^{2m}$.

Note that (38) allows us to define the Sobolev space S_m for fractional values of m as well as integer values. A multivariate version of Sobolev spaces can be defined as follows. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be non-negative integers and define $|\alpha| = \alpha_1 + \cdots + \alpha_d$. Given $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ write $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}. \tag{39}$$

Then the Sobolev space is defined by

$$S_{m,p} = \left\{ f \in L_p\left([a,b]^d\right) : \ D^\alpha f \in L_p([a,b]^d) \text{ for all } |\alpha| \leq m \right\}. \tag{40}$$

We will see that in function estimation, the optimal rate of convergence over $S_{\beta,2}$ under L_2 loss is $O(n^{-2\beta/(2\beta+d)})$.

7 Mercer Kernels and Reproducing Kernel Hilbert Spaces

Intuitively, a reproducing kernel Hilbert space (RKHS) is a class of smooth functions defined by an object called a Mercer kernel. Here are the details.

7.1 Motivating Example: Nonparametric Regression

We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we want to estimate $f^0(x) = \mathbb{E}(Y|X = x)$. The approach we used earlier was based on **smoothing kernels**:

$$\hat{f}(x) = \frac{\sum_i Y_i \; K\left(\frac{||x-X_i||}{h}\right)}{\sum_i K\left(\frac{||x-X_i||}{h}\right)}.$$

Another approach is regularization: choose f to minimize

$$\sum_i (Y_i - f(X_i))^2 + \lambda J(f)$$

for some penalty J. This is equivalent to: choose $f \in \mathcal{M}$ to minimize $\sum_i (Y_i - f(X_i))^2$ where $\mathcal{M} = \{f : \; J(f) \leq L\}$ for some $L > 0$.

We would like to construct $\mathcal{M}$ so that it contains smooth functions. We shall see that a good choice is to use a RKHS.

7.2 Evaluation Functional

Definition 1. Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \neq \emptyset$. For each $x \in \mathcal{X}$, the (Dirac) evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ is defined as $\delta_x f = f(x)$.

Evaluation functional is always linear: for $f, g \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$, $\delta_x(\alpha f + \beta g) = (\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x) = \alpha \delta_x(f) + \beta \delta_y(f)$.

However in general, the evaluation functional is not continuous. This means we can have $f_n \to f$ but $\delta_x f_n$ does not converge to $\delta_x f$. For example, let $f(x) = 0$ and $f_n(x) = \sqrt{n} I(x < 1/n^2)$. Then $||f_n - f|| = 1/\sqrt{n} \to 0$. But $\delta_0 f_n = \sqrt{n}$ which does not converge to $\delta_0 f = 0$. Intuitively, this is because Hilbert spaces can contain very unsmooth functions.

We define RKHS be the Hilbert spaces where the evaluation functional is continous. Intuitively, this means that the functions in the space are well-behaved.

Definition 2. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \neq \emptyset$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if δ_x is continuous for any $x \in \mathcal{X}$.

If two function f, g are close in the norm, then $f(x)$ and $g(x)$ are also close.

Theorem. *If* $\lim_{n\to\infty} \|f_n - f\| = 0$, *then* $\lim_{n\to\infty} f_n(x) = f(x)$, *for all* $x \in \mathcal{X}$.

7.3 Reproducing Kernel

Definition 3. Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \neq \emptyset$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies

1. For all $x \in \mathcal{X}$, $K_x = K(\cdot, x) \in \mathcal{H}$.
2. For all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $\langle f, K_x \rangle = f(x)$ (reproducing property).

This implies that K_x is the **representer** of the evaluation functional: think of Riesz representation theorem. In particular, for any $x, y \in \mathcal{X}$,

$$K(x, y) = \langle K_y, K_x \rangle = \langle K_x, K_y \rangle = K(y, x).$$

Theorem. *If it exists, reproducing kernel is unique.*

Definition 4. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \neq \emptyset$ is RKHS if it has a reproducing kernel.

In fact it is not difficult to see that two definitions of RKHS are equivalent: If the evaluational functionals are continuous, then for each $x \in \mathcal{X}$ we can find $K_x \in \mathcal{H}$ with $\delta_x(f) = f(x) = \langle f, K_x \rangle$ for all $f \in \mathcal{H}$. Conversely, suppose that $f_n \to f$. Then

$$\delta_x f_n = \langle f_n, K_x \rangle \to \langle f, K_x \rangle = f(x) = \delta_x f$$

so the evaluation functional is continuous.

7.4 Positive Semidefinite function

Definition 5. A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive (semi)definite if for all $n \geq 1$, $(a_1, \ldots, a_n) \in \mathbb{R}^n$, $(x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) = a^\top \mathbb{K} a \geq 0.$$

The function K is strictly positive definite if for mutually distinct x_i, the equality holds only when $a_1 = \cdots = a_n = 0$.

Then we can see that every inner product is positive semidefinite, and every reproducing kernel is positive semidefinite.

Given a positive semidefinite function K, let $K_x(\cdot)$ be the function obtained by fixing the second coordinate. That is, $K_x(y) = K(y, x)$. For the Gaussian kernel, K_x is a Normal, centered at x. We can create functions by taking liner combinations of the kernel:

$$f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x).$$

Let $\mathcal{H}_0$ denote all such functions:

$$\mathcal{H}_0 = \left\{ f : \sum_{j=1}^{k} \alpha_j K_{x_j}(x) \right\}.$$

Given two such functions $f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x)$ and $g(x) = \sum_{j=1}^{m} \beta_j K_{y_j}(x)$ we define an inner product

$$\langle f, g \rangle = \langle f, g \rangle_K = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j).$$

In general, f (and g) might be representable in more than one way. You can check that $\langle f, g \rangle_K$ is independent of how f (or g) is represented. The inner product defines a norm:

$$||f||_K = \sqrt{\langle f, f \rangle} = \sqrt{\sum_j \sum_k \alpha_j \alpha_k K(x_j, x_k)} = \sqrt{\alpha^T \mathbb{K} \alpha}$$

Definition 6. The completion of $\mathcal{H}_0$ with respect to $|| \cdot ||_K$ is denoted by $\mathcal{H}_K$ and is called the RKHS generated by K.

To verify that this is a well-defined Hilbert space, you should check that the following properties hold:

$$\begin{aligned} \langle f, g \rangle &= \langle g, f \rangle \\ \langle cf + dg, h \rangle &= c\langle f, h \rangle + d\langle g, h \rangle \\ \langle f, f \rangle = 0 \quad &\text{iff} \quad f = 0. \end{aligned}$$

The last one is not obvious so let us verify it here. It is easy to see that $f = 0$ impies that $\langle f, f \rangle = 0$. Now we must show that $\langle f, f \rangle = 0$ implies that $f(x) = 0$. So suppose that $\langle f, f \rangle = 0$. Pick any x. Then

$$\begin{aligned} 0 &\leq f^2(x) = \langle f, K_x \rangle^2 = \langle f, K_x \rangle \, \langle f, K_x \rangle \\ &\leq ||f||^2 \, ||K_x||^2 = \langle f, f \rangle^2 \, ||K_x||^2 = 0 \end{aligned}$$

where we used Cauchy-Schwartz. So $0 \leq f^2(x) \leq 0$ which means that $f(x) = 0$.

7.5 Mercer Kernels

A RKHS can be also defined by a **Mercer kernel**. A Mercer kernel $K(x, y)$ is a continuous function of two variables that is symmetric and positive semidefinite. This means that, $K(x, y) = K(y, x)$, and for any function f,

$$\int\int K(x, y)f(x)f(y)dx\, dy \geq 0.$$

(This is like the definition of a positive semidefinite matrix: $x^T Ax \geq 0$ for each x.)

The function

$$K(x, y) = \sum_{k=1}^{m-1} \frac{1}{k!} x^k y^k + \int_0^{x \wedge y} \frac{(x-u)^{m-1}(y-u)^{m-1}}{(m-1)!^2} du \tag{41}$$

introduced in the Section 6 on Sobolev spaces is an example of a Mercer kernel. The most commonly used kernel is the Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

Theorem (Mercer's theorem)**.** *Let $\mathcal{X}$ be compact, and μ be a measure on $\mathcal{X}$ with $\mu(\mathcal{X}) < \infty$ and $\text{supp}(\mu) = \mathcal{X}$. Suppose that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous, symmetric, and satisfies $\sup_{x,y} K(x, y) < \infty$. Define*

$$T_K f(x) = \int_{\mathcal{X}} K(x, y)\, f(y)\, d\mu(y) \tag{42}$$

suppose that $T_k : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ is positive semidefinite; thus,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} K(x, y)\, f(x)\, f(y)\, d\mu(x)\, d\mu(y) \geq 0 \tag{43}$$

for any $f \in L^2(\mathcal{X})$. Then there exists a countable eigenvalues and eigenfunctions λ_i, Ψ_i, i.e.,

$$\int_{\mathcal{X}} K(x, y)\Psi_i(y)\, d\mu(y) = \lambda_i \Psi_i(x). \tag{44}$$

where $\{\Psi_i\}$ is an orthonormal basis, and if $\lambda_i > 0$ then Ψ_i is continuous. Further, $\sum_i \lambda_i < \infty$, $\sup_x \Psi_i(x) < \infty$, and

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \Psi_i(x)\, \Psi_j(y), \tag{45}$$

where the convergence is uniform in x, y.

The positive semidefinite requirement for Mercer kernels is generally difficult to verify. But the following basic results show how one can build up kernels in pieces.

If $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are Mercer kernels then so are the following:

$$\begin{aligned}
K(x, y) &= K_1(x, y) + K_2(x, y) && (46)\\
K(x, y) &= c\, K_1(x, y) + K_2(x, y) \quad \text{for } c \in \mathbb{R}_+ && (47)\\
K(x, y) &= K_1(x, y) + c \quad \text{for } c \in \mathbb{R}_+ && (48)\\
K(x, y) &= K_1(x, y)\, K_2(x, y) && (49)\\
K(x, y) &= f(x)\, f(y) \quad \text{for } f : \mathcal{X} \longrightarrow \mathbb{R} && (50)\\
K(x, y) &= (K_1(x, y) + c)^d \quad \text{for } \theta_1 \in \mathbb{R}_+ \text{ and } d \in \mathbb{N} && (51)\\
K(x, y) &= \exp\left(K_1(x, y)/\sigma^2\right) \quad \text{for } \sigma \in \mathbb{R} && (52)\\
K(x, y) &= \exp\left(-(K_1(x, x) - 2K_1(x, y) + K_1(y, y))/2\sigma^2\right) && (53)\\
K(x, y) &= K_1(x, y)/\sqrt{K_1(x, x)\, K_1(y, y)} && (54)
\end{aligned}$$

7.6 Spectral Representation

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel with the conditions from Mercer's theorem satisfied. Note that

$$\begin{aligned} K(x,y) &= \sum_{i=1}^{\infty} \lambda_i \Psi_i(x) \Psi_j(y) \\ &= \left\langle \sum_{i=1}^{\infty} \sqrt{\lambda_i} \Psi_i(x), \sum_{i=1}^{\infty} \sqrt{\lambda_i} \Psi_i(y) \right\rangle_{\ell^2(\mathbb{N})}. \end{aligned}$$

Hence we can define the **feature map** Φ by

$$\Phi(x) = (\sqrt{\lambda_1}\Psi_1(x), \sqrt{\lambda_2}\Psi_2(x), \ldots).$$

We can expand f either in terms of K or in terms of the basis $\Psi_1, \Psi_2, \ldots$:

$$f(x) = \sum_i \alpha_i K(x_i, x) = \sum_{j=1}^{\infty} \beta_j \Psi_j(x).$$

Furthermore, if $f(x) = \sum_j a_j \Psi_j(x)$ and $g(x) = \sum_j b_j \Psi_j(x)$, then

$$\langle f, g \rangle_K = \sum_{j=1}^{\infty} \frac{a_j b_j}{\lambda_j}.$$

Roughly speaking, when $||f||_K$ is small, then f is smooth.
This is since

$$K_x(\cdot) = K(\cdot, x) = \sum_{i=1}^{\infty} \lambda_i \Psi_i(x) \Psi_i(\cdot),$$

and hence from $\langle f, K_x \rangle = f(x)$, and if we additionally impose that $\langle \Psi_i, \Psi_j \rangle_K = 0$, then

$$\Psi_j(x) = \langle \Psi_j, K_x \rangle_K = \left\langle \Psi_j, \sum_{i=1}^{\infty} \lambda_i \Psi_i(x) \Psi_i(\cdot) \right\rangle_{\ell^2(\mathbb{N})} = \lambda_j \Psi_j(x) \langle \Psi_j, \Psi_j \rangle,$$

and hence $\langle \Psi_j, \Psi_j \rangle_K$ should be $1/\lambda_j$.

7.7 Examples

Example 7. Let $\mathcal{H}$ be all functions f on $\mathbb{R}$ such that the support of the Fourier transform of f is contained in $[-a, a]$. Then

$$K(x,y) = \frac{\sin(a(y-x))}{a(y-x)}$$

and

$$\langle f, g \rangle = \int fg.$$

Example 8. Let $\mathcal{H}$ be all functions f on $(0,1)$ such that

$$\int_0^1 (f^2(x) + (f'(x))^2) x^2 dx < \infty.$$

Then

$$K(x,y) = (xy)^{-1} \left(e^{-x} \sinh(y) I(0 < x \leq y) + e^{-y} \sinh(x) I(0 < y \leq x) \right)$$

and

$$||f||^2 = \int_0^1 (f^2(x) + (f'(x))^2) x^2 dx.$$

Example 9. The Sobolev space of order m is (roughly speaking) the set of functions f such that $\int (f^{(m)})^2 < \infty$. For $m = 1$ and $\mathcal{X} = [0, 1]$ the kernel is

$$K(x, y) = \begin{cases} 1 + xy + \frac{xy^2}{2} - \frac{y^3}{6} & 0 \leq y \leq x \leq 1 \\ 1 + xy + \frac{yx^2}{2} - \frac{x^3}{6} & 0 \leq x \leq y \leq 1 \end{cases}$$

and

$$||f||_K^2 = f^2(0) + f'(0)^2 + \int_0^1 (f''(x))^2 dx.$$

7.8 Representer Theorem

Let ℓ be a loss function depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$ and on $f(X_1), \ldots, f(X_n)$. Let $\hat{f}$ minimize

$$\ell + g(||f||_K^2)$$

where g is any monotone increasing function. Then $\hat{f}$ has the form

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

for some $\alpha_1, \ldots, \alpha_n$.

7.9 RKHS Regression

Define $\hat{f}$ to minimize

$$R = \sum_i (Y_i - f(X_i))^2 + \lambda ||f||_K^2.$$

By the representer theorem, $\hat{f}(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$. Plug this into R and we get

$$R = ||Y - \mathbb{K}\alpha||^2 + \lambda \alpha^T \mathbb{K} \alpha$$

where $\mathbb{K}_{jk} = K(X_j, X_k)$ is the Gram matrix. The minimizer over α is

$$\hat{\alpha} = (\mathbb{K} + \lambda I)^{-1} Y$$

and $\hat{m}(x) = \sum_j \hat{\alpha}_j K(X_i, x)$. The fitted values are

$$\hat{Y} = \mathbb{K}\hat{\alpha} = \mathbb{K}(\mathbb{K} + \lambda I)^{-1} Y = LY.$$

So this is a linear smoother.

We can use cross-validation to choose λ. **Compare this with smoothing kernel regression.**

7.10 Logistic Regression

Let

$$f^0(x) = \mathbb{P}(Y = 1 | X = x) = \frac{e^{f(x)}}{1 + e^{f(x)}}.$$

We can estimate f^0 by minimizing

$$-\text{loglikelihood} + \lambda ||f||_K^2.$$

Then $\hat{f} = \sum_j K(x_j, x)$ and α may be found by numerical optimization. In this case, smoothing kernels are much easier.

7.11 Support Vector Machines

Suppose $Y_i \in \{-1, +1\}$. The linear SVM minimizes the penalized hinge loss:

$$J = \sum_i [1 - Y_i(\beta_0 + \beta^T X_i)]_+ + \frac{\lambda}{2}||\beta||_2^2.$$

The dual is to maximize

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle$$

subject to $0 \leq \alpha_i \leq C$.

The RKHS version is to minimize

$$J = \sum_i [1 - Y_i f(X_i)]_+ + \frac{\lambda}{2}||f||_K^2.$$

The dual is the same except that $\langle X_i, X_j \rangle$ is replaced with $K(X_i, X_j)$. This is called the kernel trick.

7.12 The Kernel Trick

This is a fairly general trick. In many algorithms you can replace $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$ and get a nonlinear version of the algorithm. This is equivalent to replacing x with $\Phi(x)$ and replacing $\langle x_i, x_j \rangle$ with $\langle \Phi(x_i), \Phi(x_j) \rangle$. However, $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ and $K(x_i, x_j)$ is much easier to compute.

In summary, by replacing $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$ we turn a linear procedure into a nonlinear procedure without adding much computation.

7.13 Hidden Tuning Parameters

There are hidden tuning parameters in the RKHS. Consider the Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

For nonparametric regression we minimize $\sum_i (Y_i - m(X_i))^2$ subject to $||m||_K \leq L$. We control the bias variance tradeoff by doing cross-validation over L. But what about σ?

This parameter seems to get mostly ignored. Suppose we have a uniform distribution on a circle. The eigenfunctions of $K(x, y)$ are the sines and cosines. The eigenvalues λ_k die off like $(1/\sigma)^{2k}$. So σ affects the bias-variance tradeoff since it weights things towards lower order Fourier functions. In principle we can compensate for this by varying L. But clearly there is some intercation between L and σ. The practical effect is not well understood.

Now consider the polynomial kernel $K(x, y) = (1 + \langle x, y \rangle)^d$. This kernel has the same eigenfunctions but the eignvalues decay at a polynomial rate depending on d. So there is an interaction between L, d and, the choice of kernel itself.

7.14 Example: Two Sample Test

Gretton, Borgwardt, Rasch, Scholkopf and Smola (GBRSS 2008) show how to use kernels for two sample testing. Suppose that

$$X_1, \ldots, X_m \sim P \qquad Y_1, \ldots, Y_n \sim Q.$$

We want to test the null hypothesis $H_0 : P = Q$.

Let $\mathcal{F} = \{f : \ ||f||_K \leq 1\}$. Define

$$M = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] \right|.$$

Under weak regulaarty conditions on K, it can be shown that $M = 0$ if and only if $P = Q$. Thus we can test H_0 by estimating M.

Define

$$\hat{M} = \sup_{f \in \mathcal{F}} \left| \frac{1}{m}\sum_{i=1}^m f(X_i) - \frac{1}{n}\sum_{i=1}^m f(Y_i) \right|.$$

Some calcculations show that

$$\hat{M}^2 = \frac{1}{m^2}\sum_{j,k} K(X_j, X_k) - \frac{2}{mn}\sum_{j,k} K(X_j, Y_k) + \frac{1}{n^2}\sum_{j,k} K(Y_j, Y_k).$$

We reject H_0 if $\hat{M} > t$. **We can determine t exactly using a permutation test.**

Using McDiarmmid's inequality and a Rademacher bound, GBRSS shows that

$$\mathbb{P}\left(|\hat{M} - M| > 2\left(\sqrt{\frac{C}{m}} + \sqrt{\frac{C}{n}}\right) + \epsilon\right) \leq \exp\left(-\frac{\epsilon^2 mn}{C(m+n)}\right).$$

There is a connection with smoothing kernels. Let

$$\hat{f}_X(u) = \frac{1}{m}\sum_{i=1}^{n} \kappa(X_i - u)$$

and similarly for $\hat{f}_Y$. Then

$$\int |\hat{f}_X(u) - \hat{f}_Y(u)|^2 du = \hat{M}^2$$

where $\hat{M}$ is based on the kernel $K(x,y) = \int \kappa(x-z)\kappa(y-z)dz$. So they are really the same!

In practice, one would use the Gaussian kernel $K_\sigma(x,y) = e^{-\frac{||x-y||^2}{\sigma^2}}$. Call the resuling statistic $\hat{M}_\sigma$. For hypothesis testing, there is no need to choose a bandwidth σ. Just define

$$\hat{M} = \sup_\sigma \hat{M}_\sigma.$$

Again, the critical value can be obtained using permutation methods. This is needed since the distribution of $\hat{M}$ under H_0 is very complex and involved unknown quantities. (See Rosenbaum (2005, *Biometrika*) for a cool, two-sample test with an exact, known, distribution free null distribution.)

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.