Nonparametric statistics: Upper bounds, Lower bounds, Adaptation, and Functionals/Semiparametrics

August 30, 2025

1 What is nonparametric statistic all about?

In undergraduate statistics class, you might have learnt some "nonparametric statistical methods", such as permutation test, Wilcoxon rank sum and signed rank tests, and even bootstrapping. These are not the kind of nonparametric statistics we are going to cover in this class. To be pedagogical, they should be called "distribution-free statistical methods".

"Nonparametric statistics" in our class are about testing hypotheses, estimating parameters and making statistical inference on parameters θ when the parameter space Θ is an infinite-dimensional space. This is in contrast to the classical parametric statistics we talked about in the past months.

Another way of thinking about nonparametric statistics is through "optimal rate of convergence". Usually, we call a parameter θ $\underline{\sqrt{n}\text{-estimable}}$ if there exists an estimator $\widehat{\theta}$ such that $\widehat{\theta} = \theta + O_p(n^{-1/2})$ or $\sqrt{n}(\widehat{\theta} - \theta) = O_p(1)$. Note that "$\sqrt{n}$-estimable" does not require the $O_p(1)$ quantity to be centered normal. In this case, we say θ has parametric behavior (parametric rate). Otherwise, we say θ has nonparametric behavior (nonparametric rate). So in this sense, high-dimensional linear regression can be viewed as a nonparametric problem, even though the model itself is parametric linear.

2 Typical models in nonparametric statistics

There are three toy models that people often use as a first step of investigation when they study nonparametric statistics.

1. Density estimation: Given i.i.d. data $X_1, \cdots, X_n$ drawn from some common probability distribution P_f with probability density function f with respect to the Lebesgue measure. The goal is to testing hypotheses of, estimating or making statistical inference on f, where $f \in \mathcal{F}$ and $\mathcal{F}$ is a function space of infinite dimensions.

2. Nonparametric regression: Given i.i.d. data $(X_1, Y_1), \cdots, (X_n, Y_n)$ such that $Y_i = f(X_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ or more generally

$$\mathbb{E}\epsilon_i = 0, \mathbb{E}\epsilon_i^2 < \infty.$$

   The goal is to testing hypotheses of, estimating or making statistical inference on f, where $f \in \mathcal{F}$ and $\mathcal{F}$ is a function space of infinite dimensions.

3. White noise model: Observe one sample path $dY(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t)$, $t \in [0, 1]$ and W is standard Wiener process on $[0, 1]$. Again, the goal is to testing hypotheses of, estimating or making statistical inference on f, where $f \in \mathcal{F}$ and $\mathcal{F}$ is a function space of infinite dimensions.

Remark 1. They are not the only models that people study in research.

In **?**, in Theorem 1.2.1, they proved that when $\mathcal{F} = \mathcal{H}(\alpha; C)$ (Hölder ball), with $\alpha > 1/2$, the Le Cam distances between any two of the above three models are asymptotically zero, that is, they are asymptotically equivalent as statistical experiments. There are many theorems in statistics that have such flavor: e.g. asymptotic equivalence between ergodic diffusions and Gaussian experiments.

3 Common infinite-dimensional statistical models: function spaces and approximation theory

In this class, we focus on infinite-dimensional function spaces as the examples of infinite-dimensional statistical models. Therefore, as a preparatory lecture, we need to cover some basic results in function spaces and the approximation theory in function spaces. In general, functions are quantified via smoothness/sparsity. The more smooth the function f is, the faster the rate of convergence should be. Recently, there is a new attempt of defining function spaces through answering the following question: what functions can be learnt efficiently by neural networks? This philosophy leads to the so-called Barron space and the compositional Barron space [**??**], and we will cover them at the end of this lecture. A large part of this section is based on Chapter 4 of **?**.

First, let's review L_p space equipped with the L_p norm:

$$L_p(\mathbb{X}) := \{f : \mathbb{X} \to \mathbb{R} : \|f\|_p < \infty\}, \|f\|_p := \left\{\int_{\mathbb{X}} |f(x)|^p dx\right\}^{1/p}.$$

A special case: L_2 equipped with the L_2 norm induced by the following inner product:

$$\langle f, g\rangle := \int_{\mathbb{X}} f(x)p(x)dx, \|f\|_2^2 := \int_{\mathbb{X}} f(x)^2 dx.$$

When $\mathbb{X} = [a, b]$, $-\infty < a \leq b < +\infty$, L_2 with the above inner product is a separable complete metric space, which implies that $\forall f \in L_2$, there exists countably many orthonormal basis functions $\{\phi_1, \phi_2, \cdots\}$ such that

$$f = \sum_{j=1}^{\infty} \beta_j \phi_j \equiv \sum_{j=1}^{\infty} \langle f, \phi_j\rangle \phi_j.$$

With the above expansion, we immediately have the famous Parseval's identity in Fourier analysis:

$$\|f\|_2^2 = \int f(x)^2 dx = \sum_{j=1}^{\infty} \beta_j^2 = \|\beta.\|_{\ell_2}^2$$

where $\|\cdot\|_{\ell_2}$ denotes the ℓ_2 norm of a vector.

3.1 Smoothness classes

Traditionally, we define functions by counting the number of derivatives. To this end, we define $\mathcal{C}(\mathbb{X})$ to be the space of continuous functions and $\mathcal{C}_u(\mathbb{X})$ to be the space of uniformly continuous functions. For $m \in \mathbb{N}$, we define

$$\mathcal{C}^m(\mathbb{X}) := \left\{ f \in \mathcal{C}_u(\mathbb{X}) : f^{(j)} \in \mathcal{C}_u(\mathbb{X}) \ \ \forall j = 1, \cdots, m, \|f\|_{\mathcal{C}^m} < \infty \right\}$$

where $\|f\|_{\mathcal{C}^m} := \|f\|_\infty + \|f^{(m)}\|_\infty$. Here we implicitly assume that for $f \in \mathcal{C}^m(\mathbb{X})$, all its derivatives of order m or lower exist. $\mathcal{C}^\infty(\mathbb{X})$ are infinitely differentiable functions. In fields like differential geometry, when people say smooth functions, they usually mean $\mathcal{C}^\infty(\mathbb{X})$.

3.2 Sobolev spaces

Sobolev spaces are extensions of smooth classes by relaxing the assumption of existing (higher-order) derivatives and it is commonly used in PDE. To this end, we will define the concept of "weak differentiable" and weak derivatives. First, locally integrable functions over $\mathbb{R}$ are functions integrable over any Borel measurable sets of $\mathbb{R}$ (formed by unions of open intervals and their complements).

Definition 2. A function f in L_p over $\mathbb{X} \subset \mathbb{R}$ is weakly differentiable if there exists a locally integrable function Df such that

$$\int_{\mathbb{X}} f(u)\phi'(u)du = -\int_{\mathbb{X}} Df(u)\phi(u)du$$

for every $\phi \in \mathcal{C}^\infty(\mathbb{X})$ with compact support in $\mathsf{interior}(A)$. In PDE, ϕ is called "test function".

Then we can define L_p Sobolev space of order $m \in \mathbb{N}$ as

$$\mathcal{W}_p^m(\mathbb{X}) := \left\{ f \in L_p(\mathbb{X}) : D^j f \in L_p(\mathbb{X}) \ \ \forall j = 1, \cdots, m, \|f\|_{\mathcal{W}_p^m} < \infty \right\}$$

where $\|f\|_{\mathcal{W}_p^m} := \|f\|_p + \|D^m f\|_p$.

3.3 Hölder spaces

Another commonly studied space is the Hölder space. It is very similar to Sobolev space but it allows non-integer-valued smoothness index. For $\mathbb{X} \subset \mathbb{R}$, define Hölder space of smoothness s as:

$$\mathcal{H}^s(\mathbb{X}) := \left\{ f \in \mathcal{C}_u(\mathbb{X}) : \|f\|_{\mathcal{H}^s} < \infty \right\}.$$

- When $s \in (0,1)$, $\|f\|_{\mathcal{H}^s} := \|f\|_\infty + \sup_{x,y \in \mathbb{X}, x \neq y} \frac{|f(x) - f(y)|}{|x-y|^s}$
- When $s > 1$ and s is non-integer, $\|f\|_{\mathcal{H}^s} := \|f\|_{\mathcal{C}^{\lfloor s \rfloor}} + \sup_{x,y \in \mathbb{X}, x \neq y} \frac{|f^{(\lfloor s \rfloor)}(x) - f^{(\lfloor s \rfloor)}(y)|}{|x-y|^{s - \lfloor s \rfloor}}$.

Naturally, when s is an integer, $\mathcal{H}^s(\mathbb{X}) = \mathcal{C}^s(\mathbb{X})$.

For both Hölder spaces and Sobolev spaces, we sometimes also restrict to Hölder balls and Sobolev balls by requiring the corresponding norms to be less than some constant C (radius). Then we have

$$\mathcal{W}_p^m(\mathbb{X}; C) \coloneqq \left\{ f \in \mathcal{W}_p^m(\mathbb{X}) : \|f\|_{\mathcal{W}_p^m} \le C \right\},$$
$$\mathcal{H}^s(\mathbb{X}; C) \coloneqq \{ f \in \mathcal{H}^s(\mathbb{X}) : \|f\|_{\mathcal{H}^s} \le C \}.$$

In particular, $\mathcal{W}_2^m(\mathbb{X}; C)$ and $\mathcal{H}^s(\mathbb{X}; C)$, in terms of statistical behavior, are almost the same because:

- In terms of minimax estimation rate of convergence under L_2 loss for $\mathcal{W}_2^m(\mathbb{X}; C)$ ($\mathcal{H}^s(\mathbb{X}; C)$) is $n^{-\frac{m}{1+2m}}$ ($n^{-\frac{s}{1+2s}}$).
- The metric entropies (log of covering number) for $\mathcal{W}_2^m(\mathbb{X}; C)$ ($\mathcal{H}^s(\mathbb{X}; C)$) is $\epsilon^{-1/m}$ ($\epsilon^{-1/s}$).

When $d = 1$, Sobolev spaces with appropriate inner products are reproducing kernel Hilbert spaces (RKHS) because they are at least weakly differentiable (as $m \ge 1$).

Remark 3. Generalizing to $\mathbb{X} \subseteq \mathbb{R}^d$ with d finite (see Appendix B), we have

- In terms of minimax estimation rate of convergence under L_2 loss for $\mathcal{W}_2^m(\mathbb{X}; C)$ ($\mathcal{H}^s(\mathbb{X}; C)$) is $n^{-\frac{m}{d+2m}}$ ($n^{-\frac{s}{d+2s}}$).
- The metric entropies (log of covering number) for $\mathcal{W}_2^m(\mathbb{X}; C)$ ($\mathcal{H}^s(\mathbb{X}; C)$) is $\epsilon^{-d/m}$ ($\epsilon^{-d/s}$).

Therefore both Sobolev spaces and Hölder spaces suffer from the curse of dimensionality (COD). Weinan E (who founded Zhiyuan College at SJTU) wants to ask what kind of function spaces should be well approximated by deep neural networks without suffering from COD so he dislikes Hölder type spaces as models for studying neural networks. See **?**.

4 Approximation of function spaces

For L_2, since it is a separable complete metric space, we can represent any of its members using linear combination of countably many functions. It is relatively easier to think about approximating functions in this space. In practice, we can only compute finitely many functions so we need to truncate the series $f = \sum_{j=1}^{\infty} \beta_j \phi_j$ at some finite dimension k.

One approach is to project f onto the span of the first k basis functions $\bar{\phi}_k \coloneqq \{\phi_1, \phi_2, \cdots, \phi_k\}$ ($\mathsf{span}(\bar{\phi}_k)$). With slight abuse of notation, we write

$$\bar{f}_k \coloneqq \Pi[f|\mathsf{span}(\bar{\phi}_k)] \equiv \Pi[f|\bar{\phi}_k].$$

By orthonormality of the basis functions, we immediately have $\bar{f}_k = \sum_{j=1}^{k} \beta_j \phi_j$. We also call $\bar{f}_k$ as the k-term linear approximation of f because for any two k-term linear approximation $\bar{f}^{(1)}$ and $\bar{f}^{(2)}$, $\bar{f}^{(1)} + \bar{f}^{(2)}$ is also a k-term linear approximation.

Not all function spaces can be well approximated by linear approximation. Thus we sometimes also need nonlinear approximation. To this end, define the following space:

$$\Lambda_{0,k} := \left\{ f \in L_2 : f = \sum_{j=1}^{\infty} \beta_j \phi_j, \|\beta.\|_0 \leq k \right\}.$$

We define the k-term nonlinear approximation of f as

$$\widetilde{f}_k := \sum_{j \in \mathcal{I}_k} \beta_j \phi_j$$

where $\mathcal{I}_k$ is the set of indices corresponding to the largest k $|\beta_j|$'s.

When you assume that the underlying functions are Hölder or Sobolev, then you immediately know which basis functions are the "best" k-term linear approximation basis in those spaces (discussed in the next lecture), i.e. to achieve approximation error $k^{-m/d}$ or $k^{-s/d}$ in L_2 or L_∞ norms. But it is a completely open problem when you do not want to assume $\mathcal{F}$ to be, say, Hölder but you are given a huge dictionary of functions and need to pick which k of them best approximate the underlying function f. In that case, the only known polynomial-time algorithm for finding the best k dictionaries is the greedy forward-selection method and its variants [**??**].

4.1 Common basis functions

There are several orthonormal basis functions (w.r.t. the Lebesgue measure) that we often use.

1. Fourier basis: $\phi_1(x) = 1$, and

$$\phi_{2j}(x) = \frac{1}{\sqrt{2}} \cos(2j\pi x), \phi_{2j+1}(x) = \frac{1}{\sqrt{2}} \sin(2j\pi x), j = 1, 2, \cdots$$

2. Haar basis on $[0, 1]$:

$$\left\{ \phi(x), \psi_{j,\ell}(x), j = 1, 2, \cdots, \ell = 0, 1, 2, \cdots, 2^j - 1 \right\}$$

where $\phi(x) = \mathbb{1}\{x \in [0, 1]\}$ (scaling function/father wavelet) and $\psi_{j,\ell}(x) = 2^{j/2}\psi(2^j x - \ell)$ (dilation and translation), with $\psi(x) = \mathbb{1}\{x \in [0, 1/2]\} - \mathbb{1}\{x \in [1/2, 1]\}$ (wavelet function/mother wavelet). As j gets larger, $\psi_{j,\ell}$ gets more localized. Haar basis is quite special: it is both wavelet basis and spline basis. But the smoothness of Haar basis is low as it consists of step functions only.

Remark 4. Why we only need $\ell = 0, 1, \cdots, 2^j - 1$ but not all natural numbers? Think about the support of ϕ and ψ and compare them to the scaled and shifted $\psi_{j,\ell}$.

3. B-splines, natural splines, cubic splines, Legendre polynomials, Chebyshev polynomials, ... (not covered in this class, but useful to look them up on your own). In particular, splines are very useful when you use generalized additive models (GAMs).

4. Wavelets (will be covered briefly later): before deep neural networks, wavelets are the most power basis function for function approximation and imaging analysis.

5. Neural networks: not basis functions in the traditional sense, but more likely than not, they are the future.

Example 1. *We use Sobolev ball as an example. We have the following theorem:*

Theorem 5. *Let $\{\phi_j, j = 0, 1, \cdots\}$ be the Fourier basis. Then*

$$W_m(C) \equiv \left\{ f : f = \sum_{j=1}^{\infty} \theta_j \phi_j, \sum_{j=1}^{\infty} \alpha_j^2 \theta_j^2 \leq C^2 \right\}.$$

where $\alpha_j = (\pi j)^m$ for j even and $\alpha_j = (\pi(j-1))^m$ for j odd. The RHS of the above equation is also called a Sobolev ellipsoid with $\alpha_j \sim (\pi j)^{2m}$.

Based on this theorem, it is quite obvious that using the first k Fourier basis to approximate functions in $W_m(C)$, the error rate should be k^{-j}.

5 Besov spaces

Besov spaces are the most general function spaces defined via smoothness and they are widely studied in harmonic analysis. Besov spaces include Sobolev and Hölder spaces as special cases. Another important feature of Besov spaces is that it allows the underlying function space to be non-convex, under which linear estimators are not optimal [**?**] and nonlinear (e.g. shrinkage) estimators must be used. We will offer two equivalent definitions of Besov spaces: (1) through moduli of smoothness just as Sobolev spaces and Hölder spaces; (2) through the decay of wavelet coefficients in the wavelet expansion of f. (Oleg Besov is still alive: Besov's homepage).

5.1 Definition via moduli of smoothness

To define Besov spaces, we need to introduce a new concept: moduli of smoothness. First define the translation operator: $\tau_h(f)(x) = f(x+h)$ and the difference operator $\Delta_h = \tau_h - \mathsf{id}$ so $\Delta_h(f)(x) = f(x+h) - f(x)$. Intuitively, $h^{-1}\Delta_h(f)(x)$ acts like derivative except that we do not send h to zero. Apparently, if Df exists, $\lim_{h\to 0} h^{-1}\Delta_h(f)(x) = Df(x)$.

By induction, we define the r-th order difference operator as $\Delta_h^r = \Delta_h(\Delta_h^{r-1}) = (\tau_h - \mathsf{id})^r$. Then

$$\Delta_h^r(f)(x) = \sum_{k=1}^{r} (-1)^{r-k} \binom{r}{k} f(x+kh).$$

As above, we expect the following to hold:

$$\lim_{h\to 0} h^{-r} \Delta_h^r(f)(x) = D^r f(x).$$

To see this, let's assume Df exists. Then by the fundamental theorem of calculus, we have

$$\begin{aligned}\Delta_h(f)(x) &= \int_x^{x+h} Df(u)du = \int_{\mathbb{R}} Df(u)\mathbb{1}\{x \leq u \leq x+h\}du = h\int_{\mathbb{R}} Df(u)\mathbb{1}\left\{0 \leq \frac{u-x}{h} \leq 1\right\}du \\ &\equiv h\int_{\mathbb{R}} Df(u)N_{1,h}(u-x)du\end{aligned}$$

where $N_1(x) = \mathbb{1}\{x \in [0,1]\}$ (the Haar scaling function) and $N_{1,h}(x) = hN_1(x/h)$.

Let's look at $\Delta_h^2(f)(x)$:

$$\begin{aligned}
\Delta_h^2(f)(x) &= f(x+2h) - 2f(x+h) + f(x) = h\int_{\mathbb{R}} (Df(u+h) - Df(u))N_{1,h}(u-x)du \\
&= h\int_{\mathbb{R}}\int_{\mathbb{R}} D^2f(v)\mathbb{1}\{u \le v \le u+h\}dvN_{1,h}(u-x)du \\
&= h^2\int_{\mathbb{R}}\int_{\mathbb{R}} D^2f(v)\mathbb{1}\left\{0 \le \frac{v-u}{h} \le 1\right\}dvN_{1,h}(u-x)du \\
&= h^2\int_{\mathbb{R}}\int_{\mathbb{R}} D^2f(v)N_{1,h}(v-u)dvN_{1,h}(u-x)du \\
&= h^2\int_{\mathbb{R}} D^2f(v)\int_{\mathbb{R}} N_{1,h}(v-u)N_{1,h}(u-x)dudv \\
&= h^2\int_{\mathbb{R}} D^2f(v)\underbrace{\int_{\mathbb{R}} N_{1,h}((v-x)-(u-x))N_{1,h}(u-x)d(u-x)}_{\equiv N_{2,h}(v-x) := [N_{1,h} * N_{1,h}](v-x)}dv \\
&= h^2\int_{\mathbb{R}} D^2f(v)N_{2,h}(v-x)dv.
\end{aligned}$$

Then by induction

$$\Delta_h^r(f)(x) = h^r\int_{\mathbb{R}} D^rf(u)N_{r,h}(u-x)du$$

where N_r is the $(r-1)$-fold convolution of N_1.

So we are left to show that $\int_{\mathbb{R}} D^rf(u)N_{r,h}(u-x)du \to D^rf(x)$ as $h \to 0$, or equivalently, $N_{r,h}$ asymptotically (in $h \to 0$) becomes the Dirac δ-function (which is not a function, but a generalized function or distribution); for a proof, see Appendix A. Recall that for a δ-function δ_0, we have $\int_{\mathbb{R}} f(u)\delta_0(u-x)du = f(x)$.

With the above analysis, we define the moduli of smoothness

$$\omega_r(f,t,p) := \sup_{0<h\le t} \|\Delta_t^r(f)\|_p$$

and when $p = 2$, we write $\omega_r(f,t) \equiv \omega_r(f,t,2)$.

Finally, we define Besov spaces as follows: given smoothness index s, take any $r > s$ an integer,

$$B_{p,q}^s(\mathbb{X}) := \begin{cases} \left\{f \in L_p(\mathbb{X}) : \|f\|_{B_{p,q}^s} = \|f\|_p + |f|_{B_{p,q}^s} < \infty\right\}, & 1 \le p < \infty, \\ \left\{f \in C_u(\mathbb{X}) : \|f\|_{B_{p,q}^s} = \|f\|_\infty + |f|_{B_{p,q}^s} < \infty\right\}, & p = \infty, \end{cases}$$

where $|f|_{B_{p,q}^s}$ is the so-called Besov semi-norm and is defined as

$$|f|_{B_{p,q}^s} := \begin{cases} \left(\int_0^\infty \left[\frac{\omega_r(f,t,p)}{t^s}\right]^q \frac{dt}{t}\right)^{1/q}, & 1 \le q < \infty, \\ \sup_{t>0} \frac{\omega_r(f,t,p)}{t^s}, & q = \infty. \end{cases}$$

Remark 6.

- The above definition is independent of the choice of r, as long as $r > s$.
- Changing from $[0, \infty]$ to $[0, 1]$ in the definition creates an equivalent norm, so the integration/supremum range does not matter here. Interestingly, for most of the theoretical work in nonparametric statistics, $\mathbb{X}$, the space of the covariates, are assumed to be compactly supported in $\mathbb{R}$ or $\mathbb{R}^d$.
- In the discussion below, we treat the o.n.b. on $\mathbb{R}$ as if we have renormalized and done a bunch of boundary corrections to make them also o.n.b. on $[0, 1]$. We do not go into the nitty gritty about how to transform o.n.b. in one domain into o.n.b. in another domain.

5.2 Definition via decaying rate of wavelet coefficients

Another equivalent definition is through the decay rate of wavelet coefficients of wavelet expansion. This is closely related to the Parseval's identity that we discussed in the beginning. But first, let's briefly introduce wavelet expansion. Here we focus on $\mathbb{X} = [0, 1] \subseteq \mathbb{R}$.

Recall that we said Haar basis is a special wavelet o.n.b. on $[0, 1]$. Haar basis functions are defined as follows:

$$\{\phi, \psi_{jk}, k = 0, 1, \cdots, 2^j - 1, j = 1, 2, \cdots\}$$

where

$$\phi(x) := \mathbb{1}\{0 \leq x \leq 1\}, \psi(x) := \mathbb{1}\{0 \leq x \leq 1/2\} - \mathbb{1}\{1/2 \leq x \leq 1\}$$

and

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

Usually ϕ is called scaling function/father wavelet and ψ is called wavelet function/mother wavelet. It is an exercise to check that Haar wavelets are indeed o.n.b. on $[0, 1]$ with respect to the Lebesgue measure. More importantly, from Haar wavelets, we can easily see an important property of wavelets in general – the multi-resolution analysis (MRA) property. To see this, we must study why we need a function ψ different from ϕ to obtain an o.n.b. Let's first define a subspace V_0 of $L_2([0, 1])$ as follows:

$$V_0 := \mathsf{span}\{\phi\}.$$

Then define V_j iteratively as follows:

$$V_j := \mathsf{span}\{f(2^j \cdot -k), k = 0, 1, \cdots, 2^j - 1 : f \in V_{j-1}\}.$$

Obviously, the subspaces $V_0, V_1, \cdots$ are nested: $V_{j-1} \subseteq V_j$ for every $j = 0, 1, \cdots$ and $V_\infty = L_2([a, b])$. To get o.n.b. out of these nested subspaces of $L_2([0, 1])$, we need to get rid of the redundancy by calculating the difference between V_{j-1} and V_j, denoted as $W_{j-1} = V_j \setminus V_{j-1}$. Again, we start from $W_0 = V_1 \setminus V_0$. In particular, $V_1 = \mathsf{span}\{\phi(2(\cdot)), \phi(2(\cdot) - 1)\} \equiv \mathsf{span}\{\mathbb{1}\{\cdot \in [0, 1/2)\}, \mathbb{1}\{\cdot \in [1/2, 1]\}\}$. Then we can deduce

$$W_0 = V_1 \setminus V_0 = \mathsf{span}\{\mathbb{1}\{\cdot \in [0, 1/2)\} - \mathbb{1}\{\cdot \in [1/2, 1]\}\}.$$

By induction, we can conclude (I omitted many details here and it is not that obvious. But you are recommended to read **?**):

$$W_{j-1} = V_j \setminus V_{j-1} = \mathsf{span}\{2^{j/2}\psi(2^j(\cdot) - k), k = 0, 1, \cdots, 2^j - 1\}.$$

Therefore $V_1 = V_0 \oplus V_1$ and $V_j = V_{j-1} \oplus W_{j-1} = V_0 \oplus \left(\bigoplus_{\ell=0}^{j-1} W_\ell\right) = V_{J_0} \oplus \left(\bigoplus_{\ell=J_0}^{j-1} W_\ell\right)$. This is the so-called MRA property.

Haar is not the only wavelets and it is quite non-smooth (dilation and translation of step functions). We can also define the so-called S-regular Daubechies wavelets by imposing stronger smoothness conditions on the scaling function/father wavelet ϕ and the wavelet function/mother wavelet ψ. S-regular basically means that ϕ and ψ are S-order differentiable (up to some minor technicalities). A good reference on how to construct S-regular wavelet basis functions is **?**. With S-regular scaling and wavelet functions, the corresponding wavelet o.n.b. is

$$\bar{z} := \{\phi_{J_0,\ell}, \ell \in I_{\phi,J_0}, \psi_{j,\ell}, \ell \in I_{\psi,j}, j = J_0, J_0 + 1, \cdots\}. \tag{1}$$

with $2^{J_0} \geq S$ (Now you also see why Haar wavelet o.n.b. use $J_0 = 0$). Here $I_{\phi,j}$ and $I_{\psi,j}$ are the index set of ℓ such that respectively $\phi_{j,\ell}$ and $\psi_{j,\ell}$ might be non-zero. The cardinality of $I_{\phi,j}$ and $I_{\psi,j}$ are at most $O(2^j)$ with the constant depending on ϕ and ψ, respectively. When we truncate the above basis $\bar{z}$ at a certain resolution j, we denote the truncated basis functions as $\bar{z}_j$. In many occasions, denote $k = 2^j$, we also write $\bar{z}_j \equiv \bar{z}_k$. Then the cardinality of $\bar{z}_k$ is $O(k) = O(2^j)$.

All S-regular Daubechies wavelets have the following important structural properties, all derived from the localization property of Daubechies wavelets (similar to Haar):

Lemma 7. *For notational convenience, rewrite $\bar{z}$ in equation (1) as follows*

$$\bar{z} \equiv \{z_1, z_2, \cdots\}$$

where we simply rename every member in equation (1) based on the order of their formal appearance in equation (1). Then the following are quite useful in many theoretical results in nonparametric statistics:

1. *When truncating $\bar{z}$ at resolution j for the scaled and translated wavelet functions/mother wavelets, the number of elements in $\bar{z}_j$ is $O(2^j)$ and from now on, with abuse of notation, we denote $\bar{z}_j$ as $\bar{z}_k$ where $k = 2^j$;*

2. *For every $x \in [0, 1]$, at each resolution j, there is at most $O(1)$ number of functions in $\bar{z}$ that are possibly non-zero;*

3. *$\|\bar{z}_k^\top \bar{z}_k\|_\infty \lesssim k$.*

Proof. Statement 1 is obvious from construction (though we did not cover how to construct Daubechies wavelets in detail, it has very similar localization property to Haar wavelets).

For statement 2, let's examine the part of scaled and translated wavelet functions/mother wavelets. Say at some resolution j and translation ℓ, $\psi_{j\ell}(x) \equiv 2^{j/2}\psi(2^j x - \ell) \neq 0 \Rightarrow \psi(2^j x - \ell) \neq 0$. Then $a \leq 2^j x - \ell \leq b$ for some finite $a < b$ and hence $2^j x - b \leq \ell \leq 2^j x - a$ so $|\ell : 2^j x - b \leq \ell \leq 2^j x - a| = O(1)$.

Statement 3 is a direct corollary of statement 2.

$$\begin{aligned}
\bar{z}_k^\top(x)\bar{z}_k(x) &= \sum_{\ell\in I_{J_0}} 2^{J_0}\phi^2(2^{J_0}x-\ell) + \sum_{m=J_0}^{j}\sum_{\ell\in I_m} 2^m\psi^2(2^m x-\ell) \\
\Rightarrow \|\bar{z}_k^\top\bar{z}_k\|_\infty &\le 2^{J_0}O(1)\sup_t \phi^2(t) + \sum_{m=J_0}^{j} 2^m O(1)\sup_t \psi^2(t) \\
&\lesssim 2^{J_0} + 2^{J_0} + \cdots + 2^j \lesssim 2^{j+1} - 1 \lesssim k
\end{aligned}$$

where in the second line we applied statement 2. □

With these, we now state an equivalent definition of Besov spaces via the decaying property of wavelet coefficients:

Definition 8 (Besov spaces via wavelet coefficients)**.**

$$B^s_{p,q} := \begin{cases} \{f \in L_p([0,1]) : \|f\|_{\ell_{B^s_{p,q}}} < \infty\} & 1 \le p < \infty, \\ \{f \in C_u([0,1]) : \|f\|_{\ell_{B^s_{p,q}}} < \infty\} & p = \infty, \end{cases}$$

where

$$\|f\|_{\ell_{B^s_{p,q}}} := \begin{cases} 2^{J_0(s+\frac{1}{2}-\frac{1}{p})}\|\langle f, \phi_{J_0,\cdot}\rangle\|_p + \left(\sum\limits_{m=J_0}^{\infty} 2^{qm(s+\frac{1}{2}-\frac{1}{p})}\|\langle f, \psi_{m,\cdot}\rangle\|_p^q\right)^{1/q} & 1 \le q < \infty, \\ 2^{J_0(s+\frac{1}{2}-\frac{1}{p})}\|\langle f, \phi_{J_0,\cdot}\rangle\|_p + \sup_{m\ge J_0} 2^{m(s+\frac{1}{2}-\frac{1}{p})}\|\langle f, \psi_{m,\cdot}\rangle\|_p & q = \infty. \end{cases}$$

From the above definition, for $B^s_{2,\infty}$, it is obvious that if we truncate $\bar{z}$ at resolution j, the approximation error is of order 2^{-js}.

6 Barron space and compositional space: A step towards deep learning?

As mentioned, **?** offer a new perspective of defining function spaces that are arguably better suited to explain the success of deep learning than the traditional function spaces (culminated at Besov spaces) because they suffer from the curse of dimensionality. In particular, they define the so-called Barron space to be the space of functions that they conjecture to be learnable without suffering from curse of dimensionality using two-layer neural networks.

Definition 9 (Barron space)**.** Barron space is defined as a continuous version of a two-layer neural network $f_m(x) = \frac{1}{m}\sum_{j=1}^m a_j\sigma(\omega_j^\top x + b_j)$ (i.e. a neural network with one hidden layer) by taking the width of the hidden layer to ∞:

$$\mathcal{F}_{B_p} := \left\{ \begin{array}{c} f : f(x) = \int a\sigma(\omega^\top x + b)\rho(da, d\omega, db), \|f\|_{B_p} \le \infty \\ \rho \text{ is a probability measure of the neural network parameters } a,\ \omega, \text{ and } b \end{array} \right\}$$

where

$$\|f\|_{B_p} := \inf_\rho \{\mathbb{E}_\rho[|a|^p(\|\omega\|_1 + |b|)^p]\}^{1/p}$$

with $1 \le p \le \infty$.

Then **?** showed the following theorem:

Theorem 10. *$\forall f \in \mathcal{F}_{B_p}$, there exists a two-layer neural network $f_m(\cdot;\theta)$ such that*

$$\|f(\cdot) - f_m(\cdot;\theta)\|_{B_p}^2 \lesssim \frac{\|f\|_{B_p}^2}{m}$$

with $\|\theta\|_p \coloneqq \frac{1}{m}\sum_{j=1}^m |a_j|\left(\|\omega_j\|_1 + |b_j|\right) \leq 2\|f\|_{B_p}$.

Define $\mathcal{F}_{B_p}(C) \coloneqq \left\{f \in \mathcal{F}_{B_p} : \|f\|_{B_p} \leq C\right\}$ to be the Barron ball. Then

$$\mathsf{Rad}_n(\mathcal{F}_{B_p}(C)) \lesssim \left(\frac{\log(d)}{n}\right)^{1/2}$$

where Rad_n denotes the Rademacher complexity and d is the dimension of X.

Next **?** generalize the Barron space to space of functions that are conjectured to be learnable by deep neural networks. We only mention the formalization for deep residual networks. **?** define the space that is learnable by deep residual networks by taking the number of layers L to ∞ and view the parameter propagation through the residual networks as a dynamical system. In particular deep residual network is constructed as follows:

$$\begin{aligned}
&Z_{0,L}(x) = Vx, \text{ the input layer}\\
&Z_{\ell+1,L}(x) = Z_{\ell,L}(x) + \frac{1}{L}U_\ell\sigma(W_\ell Z_{\ell,L}(x))\\
&f_L(x;\theta) = \alpha^\top Z_{L,L}(x).
\end{aligned}$$

Then $V \in \mathbb{R}^{D\times d}$, $W_\ell \in \mathbb{R}^{m\times D}$, $U_\ell \in \mathbb{R}^{D\times m}$ and α are the parameters in the deep residual network. Then let $L \to \infty$, we have the following corresponding dynamical system:

$$\begin{aligned}
&Z(x,0) = Vx\\
&\dot{Z}(x,t) = \mathbb{E}_{(U,W)\sim\rho_t}U\cdot\sigma(WZ(x,t))\\
&f_{\alpha,\{\rho_t\}_{t\in[0,1]}}(x) = \alpha^\top Z(x,1).
\end{aligned}$$

Remark 11. Final remark on these “deep learning” spaces. The idea is quite novel even in the applied mathematics community. But somehow I feel like it makes things too easy; it will be interesting to see (1) if deep learning can adaptively learn the “simpler” Barron spaces when they are embedded in a “more difficult” space like Besov and (2) if the true data generating law is not the simpler Barron spaces what deep learning would learn. Very few papers consider model misspecification in deep learning because (i) it is often overparameterized and (ii) there are many universal approximation theorems for neural networks. But if deep learning is really leveraging the dynamical system evolution governed by stochastic gradient descent, there will be a bias towards certain structure.

7 Optimal rate of estimation: Upper bound

Let’s consider the following observation scheme:

$$(X_i, Y_i)_{i=1}^n \overset{i.i.d.}{\sim} P_f, Y = f(X) + \epsilon, \epsilon \sim N(0,\sigma^2), X \sim \mathsf{Unif}([0,1])$$

where $f \in \text{Hölder}(\alpha; C)$. We want to understand the rate of convergence of an estimator $\widehat{f}$ to the truth f in the above setup in squared error loss:

$$\mathbb{E}\|\widehat{f} - f\|_2^2 = \mathbb{E}\int_0^1 (\widehat{f}(x) - f(x))^2 dx.$$

Here f is an infinite dimensional object but we can only compute finite-dimensional quantity. Therefore we need to first approximate f by truncating its linear wavelet expansion at certain resolution $j \equiv j(n)$ with $j \to \infty$ as $n \to \infty$. Denote $k(n) = 2^{j(n)}$. Here we need the resolution to depend on the sample size n, which might be not very natural if it is the first time that you see this. Such "changing with n" estimation procedure is called "sieve methods" in statistical literature. By the previous discussion, we know that the best linear approximation of f is

$$\bar{f}_{k(n)}(x) = \Pi[f|\bar{z}_{k(n)}](x) = \beta_{k_n}^\top \bar{z}_{k(n)}(x) \equiv \langle f, \bar{z}_{k(n)}\rangle^\top \bar{z}_{k(n)}(x)$$

and we immediately know that $\|f - \bar{f}_{k(n)}\|_\infty^2 \asymp k(n)^{-2\alpha}$, which also gives $\|f - \bar{f}_{k(n)}\|_2^2 \lesssim k(n)^{-2\alpha}$ where $\|f - \bar{f}_{k(n)}\|_2^2 \coloneqq \int_0^1 (f(x) - \bar{f}_{k(n)}(x))^2 dx$. This gives us the bias of estimating f by $\bar{f}_{k(n)}(x)$.

In class, we pretend not knowing the marginal distribution of X and construct the following unbiased estimator of $\bar{f}_{k_n}(x)$:

$$\widehat{f}_{k(n)}(x) = \left\{\frac{1}{n}\sum_{i=1}^n Y_i \bar{z}_{k(n)}(X_i)\right\}^\top \bar{z}_{k(n)}(x).$$

Let's now compute the variance of $\widehat{f}_{k(n)}(x)$:

$$\begin{aligned}
&\mathbb{E}\int_0^1 \left(\widehat{f}_{k(n)}(x) - \bar{f}_{k(n)}(x)\right)^2 dx = \mathbb{E}\int_0^1 \left\{\left[\frac{1}{n}\sum_{i=1}^n Y_i \bar{z}_{k(n)}(X_i) - \langle f, \bar{z}_{k(n)}\rangle\right]^\top \bar{z}_{k(n)}(x)\right\}^2 dx \\
=&\mathbb{E}\int_0^1 \frac{1}{n^2}\sum_{i=1}^n \left(Y_i \bar{z}_{k(n)}(X_i) - \langle f, \bar{z}_{k(n)}\rangle\right)^\top \bar{z}_{k(n)}(x)\bar{z}_{k(n)}(x)^\top \left(Y_i \bar{z}_{k(n)}(X_i) - \langle f, \bar{z}_{k(n)}\rangle\right) dx \\
=&\frac{1}{n}\mathbb{E}\left(Y_i \bar{z}_{k(n)}(X_i) - \langle f, \bar{z}_{k(n)}\rangle\right)^\top \left(Y_i \bar{z}_{k(n)}(X_i) - \langle f, \bar{z}_{k(n)}\rangle\right) \\
\leq&\frac{1}{n}\mathbb{E}Y_i^2 \bar{z}_{k(n)}(X_i)^\top \bar{z}_{k(n)}(X_i) \\
\leq&\frac{1}{n}\|\bar{z}_{k(n)}^\top \bar{z}_{k(n)}\|_\infty \left(\mathbb{E}\mathbb{E}[Y^2|X]\right) \\
\lesssim&\frac{k(n)}{n}
\end{aligned}$$

where in the last line we use Lemma 7.3. Therefore we have squared bias $k(n)^{-2s}$ and variance $k(n)/n$. To obtain optimal rates under squared error loss, we need to equate $k(n)^{-2s} = k(n)/n \Rightarrow k(n) = n^{\frac{1}{1+2s}}$. This gives us the optimal rate of convergence $n^{-\frac{2s}{1+2s}}$ in squared error loss or $n^{-\frac{s}{1+2s}}$ in L_2 norm.

Remark 12. When X is unknown, we need to perform linear regression between Y and $\bar{z}_{k(n)}(X)$ and have an extra term $\left\{\frac{1}{n}\sum\limits_{i=1}^n \bar{z}_{k(n)}(X_i)\bar{z}_{k(n)}(X_i)^\top\right\}^{-1}$ and we need to use some very basic random matrix theory to prove variance bound.

There are several questions left unanswered: (1) is the upper bound tight? (2) can we adapt without knowing the smoothness?

8 Minimax lower bound

There are several good references for techniques of proving minimax lower bound: books like **??**, notes including Siva Balakrishnan's notes and Larry Wasserman's notes, a survey paper by **?** and an application in privacy data analysis. In particular, Siva's notes have many examples of lower bound calculations outside the smoothness classes that we will focus on in this note.

8.1 Reduction scheme

Lower bounds are generally difficult to derive because to show something to be impossible is hard. Fortunately, statisticians have come up with a set of tools that have been proven to be very useful in problem solving. The main idea is the following reduction scheme: from estimation to hypothesis testing. In particular, as summarized in **?**, the reduction scheme is comprised of the following four steps:

1. Markov inequality turns risk (expectation) into tail probability. Recall that the minimax risk is defined as

$$R_n^* \coloneqq \inf_{\widehat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}[d(\widehat{\theta}_n, \theta)].$$

   A straightforward application of Markov inequality tells us:

$$\begin{aligned}\inf_{\widehat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \geq \frac{s}{2}\right) &\leq \inf_{\widehat{\theta}_n} \sup_{\theta\in\Theta} \frac{2}{s}\mathbb{E}[d(\widehat{\theta}_n, \theta)] \\ \Rightarrow R_n^* \geq \frac{s}{2} \inf_{\widehat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \geq \frac{s}{2}\right).\end{aligned}$$

2. "Supremum" reduced to "maximum". A second reduction is quite natural: $\sup_{\theta\in\Theta}$ is quite difficult to handle when Θ is an infinite set so we lower bound the worst case over all Θ by the worst case over finitely many elements in Θ:

$$\inf_{\widehat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \geq \frac{s}{2}\right) \geq \inf_{\widehat{\theta}_n} \max_{\theta\in\{\theta_1,\cdots,\theta_M\}} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \geq \frac{s}{2}\right)$$

   where $\theta_j \in \Theta$ for all $j = 1, \cdots, M$.

3. Reducing "tail probability of the loss" to "error probability of the best test statistic". Here we need to put constraints on the finite subset $\{\theta_1, \cdots, \theta_M\}$ of Θ so that they are not too close: $d(\theta_j, \theta_k) \geq s$ for all $1 \leq j \neq k \leq M$. Here s will be the minimax rate we are looking for but we leave it as it is for now and will determine its value through information theoretical calculations. The distance lower bound between every pair θ_j, θ_k implies:

$$s \leq d(\theta_j, \theta_k) \leq d(\widehat{\theta}_n, \theta_j) + d(\widehat{\theta}_n, \theta_k).$$

Furthermore define a test statistic $\Psi_n : \mathbb{X}^n \mapsto \{1, 2, \cdots, M\}$ is a mapping from the data to one of the classes from 1 to M and $\Psi^* := \arg \min_{1 \le \ell \le M} d(\widehat{\theta}_n, \theta_\ell)$ is a test based on the estimator $\widehat{\theta}_n$. For every j, when the true data generating distribution is $\mathbb{P}_j$ corresponding to θ_j

$$\mathbb{P}_j(\Psi^* \neq j) \equiv \mathbb{P}_j\left(\exists\ k \neq j \text{ s.t. } d(\widehat{\theta}_n, \theta_k) \le d(\widehat{\theta}_n, \theta_j)\right) \le \mathbb{P}_j\left(d(\widehat{\theta}_n, \theta_j) \ge \frac{s}{2}\right).$$

Hence we have

$$\inf_{\widehat{\theta}_n} \max_{\theta \in \{\theta_1, \cdots, \theta_M\}} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \ge \frac{s}{2}\right) \ge \inf_{\widehat{\theta}_n} \max_{j \in \{1, \cdots, M\}} \mathbb{P}_j(\Psi^* \neq j) \ge \inf_{\Psi_n} \max_{j \in \{1, \cdots, M\}} \mathbb{P}_j(\Psi_n \neq j).$$

In summary, we have the following strings of inequalities: suppose that $\{\theta_1, \cdots, \theta_M\}$ satisfies for every $j \neq k$, $d(\theta_j, \theta_k) \ge s$, then

$$\begin{aligned} R_n^* &\ge \frac{s}{2} \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \ge \frac{s}{2}\right) \\ &\ge \frac{s}{2} \inf_{\widehat{\theta}_n} \max_{\theta \in \{\theta_1, \cdots, \theta_M\}} \mathbb{P}\left(d(\widehat{\theta}_n, \theta) \ge \frac{s}{2}\right) \\ &\ge \frac{s}{2} \inf_{\Psi_n} \max_{j \in \{1, \cdots, M\}} \mathbb{P}_j(\Psi_n \neq j). \end{aligned} \tag{2}$$

Thus if s is the minimax rate we are going after, we only need $\inf_{\Psi_n} \max_{j \in \{1, \cdots, M\}} \mathbb{P}_j(\Psi_n \neq j) \ge c'$ for some $c' \in (0, 1)$.

8.2 Le Cam's two-points method

Le Cam's method discretizes Θ by a set of two parameters $\{\theta_0, \theta_1\}$ so $M = 2$. Then equation (2) reduces to

$$R_n^* \ge \frac{s}{2} \inf_{\Psi_n} \max\left\{\mathbb{P}_0^{\otimes n}(\Psi_n = 1), \mathbb{P}_1^{\otimes n}(\Psi_n = 0)\right\} \ge \frac{s}{4} \inf_{\Psi_n}\left\{\mathbb{P}_0^{\otimes n}(\Psi_n = 1) + \mathbb{P}_1^{\otimes n}(\Psi_n = 0)\right\}.$$

Then recall the following equivalent characterizations of total variation distance $\mathsf{TV}(\mathbb{P}_0^{\otimes n}, \mathbb{P}_1^{\otimes n})$:

$$1 - \mathsf{TV}(\mathbb{P}_0^{\otimes n}, \mathbb{P}_1^{\otimes n}) = \inf_{\Psi_n}\left\{\mathbb{P}_0^{\otimes n}(\Psi_n = 1) + \mathbb{P}_1^{\otimes n}(\Psi_n = 0)\right\} = \int f_0^n(x) \wedge f_1^n(x) dx$$

where f_0^n and f_1^n are the p.d.f. of the product measures $\mathbb{P}_0^{\otimes n}$ and $\mathbb{P}_1^{\otimes n}$. Here $f_0 = \frac{d\mathbb{P}_0}{d\mu}$ and $f_1 = \frac{d\mathbb{P}_1}{d\mu}$ with μ a dominating measure of $\mathbb{P}_0$ and $\mathbb{P}_1$.

Thus we have the following Le Cam's lemma:

Lemma 13.

$$\begin{aligned} R_n^* &\ge \frac{s}{4} \int f_0^n(x) \wedge f_1^n(x) dx = \frac{s}{4}\{1 - \mathsf{TV}(\mathbb{P}_0^{\otimes n}, \mathbb{P}_1^{\otimes n})\} \\ &\ge \frac{s}{8} e^{-n\mathsf{KL}(\mathbb{P}_0 \| \mathbb{P}_1)} \ge \frac{s}{8} e^{-n\chi^2(\mathbb{P}_0 \| \mathbb{P}_1)}. \end{aligned} \tag{3}$$

Remark 14. We further derive lower bounds based on KL- and χ^2-divergences because they tensorize for product measures.

Proof. The first line of inequality (3) is a consequence of equivalent characterization of TV distance. For the first inequality in the second line, we need some new results. First let's consider the famous Pinsker inequality.

Lemma 15. $\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) \geq 2\mathsf{TV}^2(\mathbb{P}_0,\mathbb{P}_1) \equiv 2\left(\frac{1}{2}\int|f_0-f_1|\right)^2 = \frac{1}{2}\|\mathbb{P}_0-\mathbb{P}_1\|_1^2$.

Proof. This proof is due to David Pollard. Consider the elementary inequality for the function that we have used in proving Bernstein's inequality:

$$\phi(t) = (1+t)\log(1+t) - t \geq \frac{1}{2}\frac{t^2}{1+t/3}, \text{ for } t \geq -1.$$

Take $\frac{f_0(x)}{f_1(x)} = 1+r(x)$, obviously $r(x) \geq -1$. Obviously, we have the following identities

$$\int r(x)f_1(x)dx = \int\left(\frac{f_0}{f_1}(x)-1\right)f_1(x)dx = 0, \int|r(x)|f_1(x)dx = \int\left|\frac{f_0}{f_1}(x)-1\right|f_1(x)dx = \|\mathbb{P}_0-\mathbb{P}_1\|_1.$$

Next

$$\begin{aligned}
\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) &= \int f_0(x)\log\left(\frac{f_0}{f_1}(x)\right)dx = \int(1+r(x))\log(1+r(x))f_1(x)dx \\
&= \int\{(1+r(x))\log(1+r(x)) - r(x)\}f_1(x)dx \\
&\geq \int\frac{1}{2}\frac{r(x)^2}{1+r(x)/3}f_1(x)dx \\
&= \frac{1}{2}\int\frac{r(x)^2}{1+r(x)/3}f_1(x)dx\cdot\int(1+r(x)/3)f_1(x)dx \\
&\geq \frac{1}{2}\left[\int\frac{|r(x)|}{\sqrt{1+r(x)/3}}\sqrt{1+r(x)/3}f_1(x)dx\right]^2 \\
&= \frac{1}{2}\left[\int|r(x)|f_1(x)dx\right]^2 = \frac{1}{2}\|\mathbb{P}_0-\mathbb{P}_1\|_1^2 \\
&\equiv \frac{1}{2}(2\mathsf{TV}(\mathbb{P}_0,\mathbb{P}_1))^2 = 2\mathsf{TV}(\mathbb{P}_0,\mathbb{P}_1)^2.
\end{aligned}$$

□

But Pinsker inequality is not good enough for us to obtain the exponential bound on KL-divergence. To get such bound, we actually need the following lemma:

Lemma 16. $\int f_0\wedge f_1 \geq \frac{1}{2}\exp(-\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1))$.

Proof. Recall the following fact from our previous lecture $\int f_0\vee f_1 + \int f_0\wedge f_1 = 2$. Using this fact, we have

$$\begin{aligned}\int f_0 \wedge f_1 &\geq \frac{1}{2}\left(\int \sqrt{f_0 f_1}\right)^2 = \frac{1}{2}\exp\left(2\log\int\sqrt{f_0 f_1}\right)\\ &= \frac{1}{2}\exp\left(2\log\int\sqrt{\frac{f_1}{f_0}}f_0\right) \geq \frac{1}{2}\exp\left(2\int\log\sqrt{\frac{f_1}{f_0}}f_0\right)\\ &= \frac{1}{2}\exp\left(\int\log\frac{f_1}{f_0}f_0\right) = \frac{1}{2}\exp\left(-\int\log\frac{f_0}{f_1}f_0\right)\\ &\equiv \frac{1}{2}\exp\left(-\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1)\right).\end{aligned}$$

Applying Lemma 16 with f_0, f_1 replaced by f_0^n, f_1^n, and using the following fact:

$$\mathsf{KL}(\mathbb{P}_0^{\otimes n}\|\mathbb{P}_1^{\otimes n}) = n\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1),$$

we immediately have $R_n^* \geq \frac{s}{8}e^{-n\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1)}$. □

To get the last lower bound with χ^2-divergence, we recall the following result:

Lemma 17. $\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) \leq \chi^2(\mathbb{P}_0\|\mathbb{P}_1)$.

Proof. $\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) = \int f_0 \log\dfrac{f_0}{f_1} \leq \log\int\dfrac{f_0^2}{f_1} = \log\left\{\chi^2(\mathbb{P}_0\|\mathbb{P}_1)+1\right\} \leq \chi^2(\mathbb{P}_0\|\mathbb{P}_1)$. □

□

Corollary 1. *If* $\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) \leq \dfrac{\log(2)}{n}$, *then* $R_n^* \geq \dfrac{s}{16}$.

Proof.

$$R_n^* \geq \frac{s}{8}\exp\left\{-n\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1)\right\} \geq \frac{s}{8}e^{-\log(2)} = \frac{s}{16}.$$

□

Now we use the above results to show that for estimating a point on the function $f(x)$, with f Lipschitz continuous (Hölder with smoothness $s = 1$), the minimax lower bound under L_2 loss is $n^{-1/3}$, which is simply $n^{-\frac{s}{1+2s}}$ with $s = 1$.

Example 2. *Observation:* $(X_i, Y_i)_{i=1}^n \overset{i.i.d.}{\sim} P$ *where* $Y_i = f(X_i) + \varepsilon_i$, $\varepsilon_i \sim N(0,1)$, $X \sim \mathsf{Unif}([0,1])$, *and* $f \in \mathsf{Lipschitz}(L)$. *We are interested in estimating* $\theta = f(0)$. *The loss function is then simply* $d(\theta_0, \theta_1) = |\theta_0 - \theta_1|$.

The key is to construct the two points θ_0, θ_1. *Since we are interested in deriving lower bound, we only need to exhibit a counterexample for which the rate of convergence must be slower than* s *and therefore we should try to find the simplest possible counterexample so the analysis is easy. Another constraint is that* θ_0 *and* θ_1 *cannot be too close – they need to be* s *apart or* $n^{-1/3}$ *apart!*

Based on the above reasoning, we consider the following two point: $f_0(x) \equiv 0$ *and*

$$f_1(x) = \begin{cases} L(\delta - x) & 0 \leq x \leq \delta \\ 0 & x \geq \delta \end{cases}$$

Then $|\theta_0 - \theta_1| = |f_0(0) - f_1(0)| = L\delta \geq s$. *Similarly, by the property of KL-divergence between two normals*

$$\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) = \frac{1}{2}\int_0^1 (f_1(x) - f_0(x))^2 dx = \frac{1}{2}\int_0^\delta L^2(\delta - x)^2 dx = \frac{L^2}{6}\delta^3 \leq \frac{\log(2)}{n}$$

which gives us $\delta \lesssim n^{-1/3}$ *and thus* $s \asymp n^{-1/3}$.

8.3 Fano's "multiple-points" method

It is not hard to see that Le Cam's two-point method seems to simplify the problem a bit too much. For instance, in Example 2, if we change our parameter of interest to the entire function $\theta = f(\cdot)$ and the loss function as the L_2 norm $d(\theta_0, \theta_1) = \left\{\int_0^1 (f_1(x) - f_0(x))^2 dx\right\}^{1/2}$. Then based on Le Cam's two point, we have

$$d(\theta_0, \theta_1) = \left\{L^2 \int_0^\delta (\delta - x)^2 dx\right\}^{1/2} \asymp \delta^{3/2} \asymp n^{-1/2}.$$

But we know that this bound is not tight (should be $n^{-1/3}$ instead). Le Cam's two point method for such global estimation problem is deficient and we need some modification to get a tighter lower bound.

The modification is simple: instead of using two points, using M points and this is the so-called Fano's method. First, we need the following Fano's lemma as our theoretical foundation, which is proved by techniques with very heavy information theoretical flavor.

Lemma 18 (Fano's lemma)**.** $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$ *with* $\mathbb{P} \in \{\mathbb{P}_1, \cdots, \mathbb{P}_M\}$. *Define* $\beta \coloneqq \max_{1\leq k\neq j\leq M} \mathsf{KL}(\mathbb{P}_j\|\mathbb{P}_k)$. *Then*

$$\frac{1}{M}\sum_{j=1}^M \mathbb{P}_j(\Psi_n \neq j) \geq 1 - \frac{n\beta + \log(2)}{\log(M)}.$$

Proof. We assume $n = 1$ without loss of generality. Define $U \sim \mathsf{Unif}\{1, \cdots, M\}$ and $X|U = j \sim \mathbb{P}_j$. Then for any measurable event E

$$\mathbb{P}(X \in E, U = j) = \mathbb{P}(U = j)\,\mathbb{P}(X \in E|U = j) = \frac{1}{M}\mathbb{P}_j(E).$$

Then for the test statistic $\Psi \equiv \Psi(X)$,

$$\mathbb{P}(\Psi \neq U) = \frac{1}{M}\sum_{j=1}^M \mathbb{P}_j(\Psi \neq j).$$

Further define $Z \coloneqq \mathbb{1}\{\Psi \neq U\}$ is a function of (X, U). Then

$$\begin{aligned}
H(Z, U|X) &= H(U|X) + \underbrace{H(Z|U, X)}_{\equiv 0} = H(U|X),\\
\Rightarrow H(U|X) = H(Z, U|X) &= H(Z|X) + H(U|Z, X)\\
&\leq H(Z) + H(U|Z, X)\\
&= -\,\mathbb{P}(\Psi \neq U)\log\mathbb{P}(\Psi \neq U) - \mathbb{P}(\Psi = U)\log\mathbb{P}(\Psi = U) + H(U|Z, X)\\
&\leq \log(2) + H(U|Z, X).
\end{aligned}$$

By definition

$$H(U|Z,X) = \mathbb{P}(Z=0)H(U|Z=0,X) + \mathbb{P}(Z=1)H(U|Z=1,X).$$

But when $Z=0$, $\Psi(X)=U$ so given X we have all the information on U thus $H(U|Z=0,X)=0$. When $Z=1$, $\Psi(X)\neq U$ so the possible number of values that Ψ can take is $M-1$ (we do not have to use this though). Then

$$H(U|Z,X) = \mathbb{P}(Z=1)\log(M-1) = \mathbb{P}(\Psi\neq U)\log(M-1),$$

hence $H(U|X) \leq \log(2) + \mathbb{P}(\Psi\neq U)\log(M-1)$. But $H(U|X) = H(U) - I(U;X)$ where $I(U;X)$ is the mutual information between U and X, where

$$\begin{aligned} I(U;X) = \mathsf{KL}(\mathbb{P}_{U,X}\|\mathbb{P}_U\otimes\mathbb{P}_X) &= \frac{1}{M}\sum_{j=1}^{M}\mathsf{KL}\left(\mathbb{P}_j\|\frac{1}{M}\sum_{k=1}^{M}\mathbb{P}_k\right) \\ &\leq \frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\mathsf{KL}(\mathbb{P}_j\|\mathbb{P}_k) \leq \beta. \end{aligned}$$

Hence

$$\frac{1}{M}\sum_{j=1}^{M}\mathbb{P}(\Psi\neq j) = \mathbb{P}\left(\Psi\neq U\right) \geq \frac{H(U)-I(U;X)-\log(2)}{\log(M-1)} \geq \frac{\log(M)-\beta-\log(2)}{\log(M-1)} \geq 1-\frac{\beta+\log(2)}{\log(M)}.$$

□

Lemma 18 has the following immediate corollary:

Corollary 2 (Fano's minimax bound)**.** *If* $\beta \leq \dfrac{\log(M)}{4n}$*, then* $R_n^* \geq \dfrac{s}{4}$ *as long as* $M \geq 16$.

Finally, we revisit Example 2 using Fano's minimax bound. A common strategy is to parameterize $\{1,\cdots,M\}$ through Boolean hypercubes $\Omega = \{\omega : \omega\in\{\pm1\}^N\}$ with $2^N = M$. For any $\omega,\nu\in\Omega$, define their Hamming distance as $\mathsf{Ham}(\omega,\nu) := \sum_{i=1}^{N}\mathbb{1}\{\omega_i\neq\nu_i\}$. We consider the following function space within the Lipschitz class:

$$\mathcal{F} := \{f_\omega = \sum_{j=1}^{N}\omega_j B_j, \omega\in\Omega\}$$

where $N=\frac{1}{h}$ and we define

$$B(x) := \begin{cases} x & 0\leq x\leq\frac{1}{2} \\ 1-x & \frac{1}{2}\leq x\leq 1 \end{cases}, B(\frac{x}{h}) := \begin{cases} \frac{x}{h} & 0\leq x\leq\frac{h}{2} \\ \frac{1-x}{h} & \frac{h}{2}\leq x\leq h \end{cases} \text{ and } B_j(x) := LhB(\frac{x-j}{h}).$$

Then we need to upper bound the KL-divergence and lower bound the distance between any member of $\mathcal{F}$. But apparently, Ω (and hence $\mathcal{F}$) is too large for the distance lower bound to be any useful. Therefore we consider $\mathcal{F}'$ to be the pruned space

$$\mathcal{F}' := \{f_\omega = \sum_{j=1}^{N}\omega_j B_j, \omega\in\Omega'\}.$$

By the following Varshamov-Gilbert bound (Lemma 19), we know that we can always have Ω' such that $\Omega' = \{\omega^{(1)},\cdots,\omega^{(M)}\}$

1. $M \geq 2^{N/8}$;

2. $\mathsf{Ham}(\omega^{(j)}, \omega^{(k)}) \geq \dfrac{N}{8}\ \forall\ 0 \leq j \neq k \leq M$.

So for any $\omega^{(j)}, \omega^{(k)} \in \Omega'$,

$$\begin{aligned} d(f_j, f_k)^2 &= \int_0^1 (f_j(x) - f_k(x))^2 dx = \int_0^1 \left(\sum_{i=1}^N (\omega_i^{(j)} - \omega_i^{(k)}) B_i(x) \right)^2 dx \\ &= \sum_{i=1}^N (\omega_i^{(j)} - \omega_i^{(k)})^2 \int_0^1 L^2 h^2 B^2 \left(\frac{x-j}{h} \right) dx \\ &\asymp h^3 \sum_{i=1}^N (\omega_i^{(j)} - \omega_i^{(k)})^2 = h^3 \mathsf{Ham}(\omega^{(j)}, \omega^{(k)}) \geq h^3 \frac{N}{8} = \frac{h^2}{8}, \end{aligned}$$

and following Corollary 2

$$\mathsf{KL}(\mathbb{P}_j \| \mathbb{P}_k) \lesssim h^3 \mathsf{Ham}(\omega^{(j)}, \omega^{(k)}) \leq h^3 N = h^2 \leq \frac{\log(2^{N/8})}{4n} \lesssim N/n = \frac{1}{nh} \Rightarrow h \lesssim n^{-1/3}.$$

Hence $s^2 \asymp n^{-2/3}$.

Finally, we will use this following important result in combinatorics:

Lemma 19 (Varshamov-Gilbert bound). *Let Ω be a Boolean hypercube with dimension N with $N \geq 8$. Then there exists a "pruned" hypercube $\Omega' = \{\omega^{(1)}, \cdots, \omega^{(M)}\}$ s.t.*

1. $M \geq 2^{N/8}$;

2. $\mathsf{Ham}(\omega^{(j)}, \omega^{(k)}) \geq \dfrac{N}{8}\ \forall\ 0 \leq j \neq k \leq M$.

Proof. Define $n \coloneqq N/8$, say choose N such that n is an integer. We do the following steps to get the pruned hypercube Ω': First, define $\omega^{(0)} = (1, 1, \cdots, 1, 1)$.

- Define $\Omega_0 = \Omega$ and $\Omega_1 = \{\omega \in \Omega : \mathsf{Ham}(\omega, \omega^{(0)}) > n\}$ and take any element from Ω_1 to be $\omega^{(1)}$.
- Repeat the above recursively and define $\Omega_j = \{\omega \in \Omega_{j-1} : \mathsf{Ham}(\omega, \omega^{(j-1)}) > n\}$ until we cannot find any such ω any more, say $j = 1, \cdots, M$. At each step, the pruned out space is

$$\Gamma_j = \{\omega \in \Omega_j : \mathsf{Ham}(\omega, \omega^{(j)}) \leq n\}.$$

Define $n_j \coloneqq |\Gamma_j|$. Apparently $\Gamma_0, \cdots, \Gamma_M$ form a partition of Ω. Thus

$$2^N = n_0 + \cdots + n_M.$$

Also we observe that $n_j \leq \sum\limits_{i=1}^n \binom{N}{i}$ because fixing $\omega^{(j)}$, the ω's s.t. $\mathsf{Ham}(\omega, \omega^{(j)}) \leq n$ are the ω's that differ from $\omega^{(j)}$ in at most n dimensions out of the total N dimensions. Therefore

$$\begin{aligned} 2^N = n_0 + \cdots + n_M &\leq (M+1) \sum_{i=1}^n \binom{N}{i} \\ \Rightarrow M + 1 &\geq \frac{1}{\sum\limits_{i=1}^n \binom{N}{i} 2^{-N}} = \frac{1}{\sum\limits_{i=1}^n \binom{N}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{N-i}}. \end{aligned}$$

Thus the denominator of the RHS of the above display is the c.d.f. at n of a Binomial random variable $Z \in \text{Binom}(N, 1/2)$. Then by a simple Hoeffding's inequality, we have $\mathbb{P}(Z \leq n) = \mathbb{P}(Z - N/2 \leq n - N/2) \leq \exp\{-9N/32\}$. Thus

$$M + 1 \geq \exp\{9N/32\} \Rightarrow M \geq 2^{N/8}.$$

□

9 Adaptive estimation in L_2 risk

Recall that for $f \in \mathcal{H}(\alpha; C)$, in nonparametric regression with Gaussian error and uniform on $[0, 1]$ distribution for covariates X, the minimax squared error risk is $n^{-\frac{2\alpha}{1+2\alpha}}$, achieved by the wavelet projection estimator

$$\widehat{f}_{j(n)}(x) = \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i \bar{z}_{j(n)}(X_i) \right\}^{\top} \bar{z}_{j(n)}(x)$$

but $2^{j(n)} = k(n) = n^{\frac{1}{1+2\alpha}}$ explicitly depends on the possibly unknown smoothness index α. However, in general we do not want our statistical procedure to depend on knowledge of α. Adaptive methods are designed to achieve such goal by building data-adaptive choice of $\widehat{j}$.

In this note, we introduce one general adaptation scheme called Lepski's method [**?**]. To convey the main idea, we approximate f using dilated and shifted Haar father wavelets/scaling functions as the basis. In particular, denote

$$\bar{z}_j(\cdot) = \left\{ z_{j,\ell}(\cdot) \equiv 2^{j/2} \phi(2^j \cdot - \ell), \ell = 1, 2, \cdots, 2^j - 1 \right\}.$$

A key feature that we will use is no two different functions in the above set have overlap in their support: if $z_{j,\ell}(x) \neq 0$ then $z_{j,m} = 0 \; \forall m \neq \ell, m = 1, 2, \cdots, 2^j - 1$. For a fixed j, we know that $\widehat{f}_j$ has variance of order $2^j/n$ and squared bias of order $(2^j)^{-2s}$.

Lepski's method:

> "chooses j as the smallest resolution with $\|\widehat{f}_j - \widehat{f}_\ell\|_2^2 \lesssim \frac{2^\ell}{n} \; \forall \; \ell > j$."

The algorithm can be defined as, for some $\tau > 0$,

$$\widehat{j} \coloneqq \mathsf{min} \left\{ j \in \mathcal{J} : \|\widehat{f}_j - \widehat{f}_\ell\|_2^2 \leq \tau \frac{2^\ell}{n} \; \forall \; \ell > j, \ell \in \mathcal{J} \right\}$$

where $\mathcal{J} = \{1, \cdots, j_{\max}\}$ with $2^{j_{\max}} \leq n$. We will prove the following theorem in this note, which says adaptation does not have a penalty on the rate of convergence when estimating functions from Hölder balls in L_2 norm (also true for L_∞ norm; but this is generally not true for functional estimation, which we will consider later).

Theorem 20.

$$\sup_{f \in \textit{Hölder}(s;C)} \mathbb{E}\|\widehat{f}_{\widehat{j}} - f\|_2 \lesssim n^{-\frac{s}{1+2s}}.$$

Proof. To facilitate the proof, we need to define the oracle resolution

$$j^* \coloneqq \left\{ j \in \mathcal{J} : (2^{-j})^{2s} \lesssim \frac{2^j}{n} \right\}.$$

Thus for any $j \geq j^*$, we have $(2^{-j})^{2s} \lesssim \frac{2^j}{n}$ and for any $j < j^*$, we have $(2^{-j})^{2s} \gg \frac{2^j}{n}$ and therefore $2^{j^*} \asymp n^{\frac{1}{1+2s}}$. We will compare $\widehat{f}_{\widehat{j}}$ with $\widehat{f}_{j^*}$ in our proof. The common strategy in Lepski's proof is the following decomposition:

$$\mathbb{E}\|\widehat{f}_{\widehat{j}} - f\|_2 = \underbrace{\mathbb{E}\|\widehat{f}_{\widehat{j}} - f\|_2 \mathbb{1}\{\widehat{j} \leq j^*\}}_{I} + \underbrace{\mathbb{E}\|\widehat{f}_{\widehat{j}} - f\|_2 \mathbb{1}\{\widehat{j} > j^*\}}_{II}.$$

We now upper bound I and II respectively. I is easy:

$$\begin{aligned} I &\lesssim \mathbb{E}\left(\|\widehat{f}_{\widehat{j}} - \widehat{f}_{j^*}\|_2 + \|\widehat{f}_{j^*} - f\|_2\right) \mathbb{1}\{\widehat{j} \leq j^*\} \\ &\lesssim \sqrt{\frac{2^{j^*}}{n}} + \mathbb{E}\|\widehat{f}_{j^*} - \bar{f}_{j^*}\|_2 + \mathbb{E}\|\bar{f}_{j^*} - f\|_2 \\ &\lesssim n^{-\frac{s}{1+2s}}. \end{aligned}$$

Term II is more complicated. But Lepski's method has the following important property:

Lemma 21. *$\forall j > j^*$, $\mathbb{P}\left(\widehat{j} = j\right) \lesssim e^{-C2^j}$ and $\mathbb{P}\left(\widehat{j} > j^*\right) \lesssim e^{-C'2^{j^*}}$.*

Proof. Apparently, the first statement implies the second because

$$\mathbb{P}\left(\widehat{j} > j^*\right) = \sum_{j > j^*} \mathbb{P}\left(\widehat{j} = j\right) \lesssim \sum_{j > j^*} e^{-C2^j} \lesssim e^{-C'2^{j^*}}.$$

For some $j > j^*$, define $j^- = j - 1$. Then $\widehat{j} = j$ implies $\widehat{j} > j^-$, i.e. the Lepski's criterion must fail for j^- i.e.

$$\begin{aligned} \mathbb{P}\left(\widehat{j} = j\right) &= \mathbb{P}\left(\bigcup_{\ell \geq j} \|\widehat{f}_{j^-} - \widehat{f}_\ell\|_2^2 > \tau \frac{2^\ell}{n}\right) \\ &\leq \sum_{\ell \geq j} \mathbb{P}\left(\|\widehat{f}_{j^-} - \widehat{f}_\ell\|_2^2 > \tau \frac{2^\ell}{n}\right). \end{aligned}$$

We decompose $\|\widehat{f}_{j^-} - \widehat{f}_\ell\|_2^2$ as follows:

$$\begin{aligned}
&\|\widehat{f}_{j^-} - \widehat{f}_\ell\|_2^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n Y_i \left[\bar{z}_{j^-}(X_i)^\top \bar{z}_{j^-} - \bar{z}_\ell(X_i)^\top \bar{z}_\ell \right] \right\|_2^2 \\
&= \int_0^1 \left\{ \frac{1}{n} \sum_{i=1}^n Y_i \left[\bar{z}_{j^-}(X_i)^\top \bar{z}_{j^-}(x) - \bar{z}_\ell(X_i)^\top \bar{z}_\ell(x) \right] \right\}^2 dx \\
&= \int_0^1 \left\{ \begin{array}{c} \sum_{m=1}^{2^{j^-}} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{j^-,m}(X_i) - \beta_{j^-,m} \right) z_{j^-,m}(x) - \sum_{m=1}^{2^\ell} \frac{1}{n} \left(\sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m} \right) z_{\ell,m}(x) \\ + \bar{f}_{j^-}(x) - \bar{f}_\ell(x) \end{array} \right\}^2 dx \\
&\lesssim \underbrace{\|\bar{f}_{j^-} - \bar{f}_\ell\|_2^2}_{\lesssim \frac{2^\ell}{n} \text{ because } \ell > j^{-1} \geq j^*} \\
&+ \int_0^1 \left\{ \sum_{m=1}^{2^{j^-}} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{j^-,m}(X_i) - \beta_{j^-,m} \right) z_{j^-,m}(x) - \sum_{m=1}^{2^\ell} \frac{1}{n} \left(\sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m} \right) z_{\ell,m}(x) \right\}^2 dx.
\end{aligned}$$

The second term can be further upper bounded as follows:

$$\begin{aligned}
&\int_0^1 \left\{ \sum_{m=1}^{2^{j^-}} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{j^-,m}(X_i) - \beta_{j^-,m} \right) z_{j^-,m}(x) - \sum_{m=1}^{2^\ell} \frac{1}{n} \left(\sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m} \right) z_{\ell,m}(x) \right\}^2 dx \\
\lesssim\ &\int_0^1 \left\{ \sum_{m=1}^{2^{j^-}} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{j^-,m}(X_i) - \beta_{j^-,m} \right) z_{j^-,m}(x) \right\}^2 dx + \int_0^1 \left\{ \sum_{m=1}^{2^\ell} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m} \right) z_{\ell,m}(x) \right\}^2 dx.
\end{aligned}$$

We only need to deal with the second term: by father wavelets (no overlap in support and orthonormality)

$$\begin{aligned}
&\int_0^1 \left\{ \sum_{m=1}^{2^\ell} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m} \right) z_{\ell,m}(x) \right\}^2 dx \\
= &\sum_{m=1}^{2^\ell} \left(\frac{1}{n} \sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m} \right)^2.
\end{aligned}$$

Let $\widehat{\beta}_{\ell,m} \coloneqq \frac{1}{n} \sum_{i=1}^n Y_i z_{\ell,m}(X_i)$. From the assumptions and the particular choice of the basis functions, one can show (left as an exercise) that $\widehat{\beta}_{\ell,m} - \beta_{\ell,m} = \frac{1}{n} \sum_{i=1}^n Y_i z_{\ell,m}(X_i) - \beta_{\ell,m}$ is mean-zero sub-Gaussian with sub-Gaussian proxy $1/\sqrt{n}$. As a result, $(\widehat{\beta}_{\ell,m} - \beta_{\ell,m})^2 - \mathbb{E}(\widehat{\beta}_{\ell,m} - \beta_{\ell,m})^2$ is mean-zero sub-exponential with sub-exponential proxy $1/n$. It is also easy to see $\mathbb{E}(\widehat{\beta}_{\ell,m} - \beta_{\ell,m})^2 \lesssim 1/n$.

Then to get the desired tail probability, we need to compute

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{m=1}^{2^\ell}(\widehat{\beta}_{\ell,m}-\beta_{\ell,m})^2 \geq \tau'\frac{2^\ell}{n}\right)\\
=\ &\mathbb{P}\left(\sum_{m=1}^{2^\ell}(\widehat{\beta}_{\ell,m}-\beta_{\ell,m})^2 - \mathbb{E}(\widehat{\beta}_{\ell,m}-\beta_{\ell,m})^2 \geq \tau''2^\ell\right)\\
\leq\ &\exp\left\{-\frac{C_1 2^{2\ell}}{C_2 2^\ell}\right\} = \exp\left\{-C'2^\ell\right\}
\end{aligned}
$$

which Bernstein's inequality for sum of sub-exponential random variables.

Thus

$$
\mathbb{P}\left(\widehat{j}=j\right) \lesssim \sum_{\ell\geq j}\exp\left\{-C'2^\ell\right\} \lesssim \exp\left\{-C'2^j\right\}.
$$

□

Then applying Lemma 21, we have

$$
\begin{aligned}
II =\ & \sum_{j\in\mathcal{J}:j>j^*}\mathbb{E}\|\widehat{f}_{\widehat{j}}-f\|_2\mathbb{1}\{\widehat{j}=j\}\\
\text{Cauchy Schwarz} \leq\ & \sum_{j\in\mathcal{J}:j>j^*}\left(\mathbb{E}\|\widehat{f}_{\widehat{j}}-f\|_2^2\right)^{1/2}\mathbb{P}\left(\widehat{j}=j\right)^{1/2}\\
\lesssim\ & \sum_{j\in\mathcal{J}:j>j^*}\left(\mathbb{E}\|\widehat{f}_{j_{\max}}-f\|_2^2\right)^{1/2}\mathbb{P}\left(\widehat{j}=j\right)^{1/2}\\
\lesssim\ & \sum_{j\in\mathcal{J}:j>j^*}\mathbb{P}\left(\widehat{j}=j\right)^{1/2}\\
=\ & |\{j\in\mathcal{J}:j>j^*\}|\frac{1}{|\{j\in\mathcal{J}:j>j^*\}|}\sum_{j\in\mathcal{J}:j>j^*}\mathbb{P}\left(\widehat{j}=j\right)^{1/2}\\
\text{Jensen} \leq\ & |\{j\in\mathcal{J}:j>j^*\}|\left\{\frac{1}{|\{j\in\mathcal{J}:j>j^*\}|}\sum_{j\in\mathcal{J}:j>j^*}\mathbb{P}\left(\widehat{j}=j\right)\right\}^{1/2}\\
=\ & \left\{|\{j\in\mathcal{J}:j>j^*\}|\mathbb{P}\left(\widehat{j}>j^*\right)\right\}^{1/2}\\
\lesssim\ & j_{\max}^{1/2}\exp\{-C2^{j^*}\} = o\left(n^{-\frac{s}{1+2s}}\right).
\end{aligned}
$$

□

Other adaptive estimation strategies include but are not limited to: (1) wavelet thresholding [**????**], (2) model selection via penalized empirical risk minimization (see the monumental works by **?** and **?** and the review [**?**]), (3) aggregation [**??**] or in some sense (Bayesian) model averaging, and (4) Bayesian nonparametrics [**?**]. In terms of the robustness of these proof techniques, Bayesian nonparametrics $\succeq$ Lepskii's methods $\succeq$ Model selection $\succeq$ Aggregation $\succeq$ Wavelet Thresholding. Bayesian nonparametric proofs tend to be quite technical (need to check many many regularity conditions).

10 Minimax optimal confidence sets and functional estimation

In this section, we consider a more difficult problem than estimation: How to build honest confidence set for $f \in \mathcal{F}$, the size of which shrinks to zero at an optimal rate in the minimax sense? In this section, we again assume that the Hölder smoothness s is known. We only consider confidence sets in L_2-norm.

Consider the following oracle setting. If we have a minimax optimal nonparametric estimator $\widehat{f}_n$ of $f \in \mathcal{H}(s; B)$, an asymptotically honest nominal $(1 - \alpha)$ confidence set centered around $\widehat{f}_n$ is a set $\widehat{C}_\alpha(\widehat{f}_n)$ that satisfies:

$$\liminf_{n} \inf_{f \in \mathcal{H}(s;B)} \Pr_f \left(f \in \widehat{C}_\alpha(\widehat{f}_n) \right) \geq 1 - \alpha. \tag{4}$$

To tackle this problem, we can consider the following "oracle" confidence set[1]:

$$\widetilde{C}_\alpha(\widehat{f}_n) = \left\{ f \in \mathcal{H}(s; B) : |f - \widehat{f}_n|^2 \leq c_\alpha \mathbb{E}_f[(\widehat{f}_n - f)^2] \right\} \tag{5}$$

where c_α is some appropriately chosen constant that depends on α and some other universal constants such as the Hölder ball radius B and the wavelets basis used in the procedure.

Then it is obvious

$$\begin{aligned}
& \Pr_f \left(f \in \widetilde{C}_\alpha(\widehat{f}_n) \right) \\
= {} & 1 - \Pr_f \left((\widehat{f}_n(X) - f(X))^2 \geq c_\alpha \mathbb{E}_f[(\widehat{f}_n - f)^2] \right) \\
\geq {} & 1 - \frac{\mathbb{E}_f[(\widehat{f}_n - f)^2]}{c_\alpha \mathbb{E}_f[(\widehat{f}_n - f)^2]} = 1 - c_\alpha^{-1}.
\end{aligned}$$

If we can estimate $\mathbb{E}_f[(\widehat{f}_n - f)^2]$ with an error order $o(n^{-\frac{2\alpha}{1+2\alpha}})$[2], then we can build an asymptotically honest and minimax optimal confidence interval for f.

Can you think of why it is impossible for the size of the interval to be shorter than $\{\mathbb{E}_f[(\widehat{f}_n - f)^2]\}^{1/2}$?

At this point, it is not difficult to see in order to estimate $\mathbb{E}_f[(\widehat{f}_n - f)^2]$, we need to estimate the unknown $\mathbb{R}$-valued parameter $\int_0^1 f(x)^2 \mathrm{d}x$. This parameter can be viewed as a functional (function of functions):

$$\psi(f) = \int f(x)^2 \mathrm{d}x : \mathcal{H}(\alpha; B) \to \mathbb{R}.$$

Therefore, we will spend some time to discuss how to estimate functionals.

[1] An oracle procedure is a procedure that might depend on the unknown parameter but very helpful to build the actual procedure. For example, by assuming the smoothness s to be known or the density of X to be known is an implicitly oracle procedure.

[2] Why this is true? You can try to think what the order of $\mathbb{E}_f[(\widehat{f}_n - f)^2]$, the target of interest.

10.1 Estimation of statistical functionals

$\psi(f)$ is called a quadratic functional. In high-dimensional linear regression, when people posit the model $Y = X^\top \beta + \text{noise}$, people are often interested in estimating β or β_j or $\sum_{j=1}^d \beta_j$, which are in fact linear functionals. If you have heard of debiased lasso, you can essentially re-develop the entire theory of debiased lasso (independently developed by Andrea Montanari and Cun-Hui Zhang) based on what we will cover in this section. For linear regression problems, a quadratic functional of interest is $\beta^\top \Sigma \beta$, where $\Sigma = \mathbb{E} X X^\top$ is the population Gram matrix.

In this section, to make things more interesting, I will assume the marginal distribution of X to have an unknown density g on $[0, 1]$, instead of uniform distribution. Therefore we consider the following quadratic functional:

$$\psi(\theta) = \mathbb{E} f(X)^2 = \int_0^1 f(x)^2 p(x) \mathrm{d}x. \tag{6}$$

If we assume $\theta = (f, p)$ to lie in some infinite-dimensional function spaces $\Theta = \mathcal{F} \times \mathcal{P}$, then $\psi(\theta)$ is a low-dimensional parameter of infinite-dimensional statistical models. Low-dimensional parameters of infinite-dimensional statistical models can also be viewed as a semiparametric statistical problem. One very unique philosophy that is temporarily only advocated in statistics is that *in data analysis (more importantly in the design stage), we should try hard to think about the low-dimensional parameters that we want to infer from some complex probabilistic models and only consider optimally learn those parameters instead of the entire model.* This philosophy is very much against our instinct in particular for physicists or biologists because in those fields it seems that developing a model that can explain the phenomenon is much more important. But in social science, economics, and political sciences, this philosophy is an easy sell.

First let's assume p to be known. How to estimate a statistical functional like $\psi(\theta)$? We can consider the following naïve approach first: the so-called plug-in estimator.

$$\psi(\widehat{\theta}_n) = \psi(\widehat{f}_n) = \int \widehat{f}_n(x)^2 p(x) \mathrm{d}x = \widehat{\beta}_k^\top \int z_k(x) z_k(x)^\top p(x) \mathrm{d}x \widehat{\beta}_k = \widehat{\beta}_k^\top \Sigma_k \widehat{\beta}_k$$

where we estimate f by the usual wavelet projection $\widehat{f}_n$ with $k = n^{\frac{1}{1+2\alpha}}$. If p is unknown, we could "estimate" the unknown p in the plug-in estimator $\psi(\widehat{f}_n)$ by the empirical measure of X and construct the following estimator

$$\psi(\widehat{\theta}_n) = \psi(\widehat{f}_n, \widehat{p}_{n,emp}) = \frac{1}{n} \sum_{i=1}^n \widehat{f}_n(X_i)^2$$

where the empirical measure estimator $\widehat{p}_{n,emp}$ of p is computed using a sample of size n independent from the n samples used to compute $\widehat{f}_n$. This is the so-called "sample splitting" strategy, which is very useful to simplify the mathematical analysis. Many people nowadays also advocate the use of sample splitting in practice as a default statistical principle[3].

[3] But unfortunately this strategy seems to be frowned upon by people outside statistics, with the "argument" that this strategy is not fully using the data...

We simply look at the bias of $\psi(\widehat{\theta}_n)$:

$$\begin{aligned}
\mathbb{E}\left[\psi(\widehat{\theta}_n) - \psi(\theta)\right] &= \mathbb{E}\left[\int_0^1 \widehat{f}_n(x)^2 - f(x)^2 \mathrm{d}x\right] \\
&= \mathbb{E}\left[\int_0^1 (\widehat{f}_n(x) - f(x))(\widehat{f}_n(x) + f(x))\mathrm{d}x\right] \\
&\leq \mathbb{E}\left[\|\widehat{f}_n - f\|_2 \|\widehat{f}_n + f\|_2\right] \\
&\leq \left\{\mathbb{E}\left[\|\widehat{f}_n - f\|_2^2\right]\right\}^{1/2} \left\{\mathbb{E}\left[\|\widehat{f}_n + f\|_2^2\right]\right\}^{1/2} \\
&\lesssim \left\{\mathbb{E}\left[\|\widehat{f}_n - f\|_2^2\right]\right\}^{1/2} \asymp n^{-\frac{\alpha}{1+2\alpha}}.
\end{aligned}$$

Now the question is if we can construct an estimator with an improved bias.

10.2 Bias correction by using first-order influence functions/functional gradients

I will describe an estimator first, followed by showing its underlying mathematical meaning.

Here for simplicity, we again consider to have $2n$ samples. The first half of the sample is used to compute $\widehat{f}_n$ and the second half of the sample is used to compute the estimator of $\psi(\theta)$. Consider the following (first-order) estimator of $\psi(\theta)$.

$$\begin{aligned}
\widehat{\psi}_1(\widehat{f}_n) &= \frac{2}{n}\sum_{i=1}^n Y_i \widehat{f}_n(X_i) - \int_0^1 \widehat{f}_n(x)^2 p(x)\mathrm{d}x \\
\text{or } \widehat{\psi}_1(\widehat{\theta}_n) &= \frac{2}{n}\sum_{i=1}^n Y_i \widehat{f}_n(X_i) - \frac{1}{n}\sum_{i=1}^n \widehat{f}_n(X_i)^2.
\end{aligned} \tag{7}$$

We first analyze its bias, conditioning on the data used to compute $\widehat{f}_n$, which we denote as O_{nuis}[4]:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\psi}_1(\widehat{f}_n) - \psi(\theta)|O_{nuis}\right] &= 2\mathbb{E}\left[Y\widehat{f}_n(X)\right] - \int_0^1 \widehat{f}_n(x)^2 p(x)\mathrm{d}x - \int_0^1 f(x)^2 p(x)\mathrm{d}x \\
&= \int \{2f(x)\widehat{f}_n(x) - \widehat{f}_n(x)^2 - f(x)^2\} p(x)\mathrm{d}x \\
&= -\int \{f(x) - \widehat{f}_n(x)\}^2 p(x)\mathrm{d}x = -\|f - \widehat{f}_n\|_2^2.
\end{aligned}$$

Then obviously, the marginal bias is

$$\mathbb{E}\left[\widehat{\psi}_1(\widehat{f}_n) - \psi(\theta)\right] = -\mathbb{E}\left[\|f - \widehat{f}_n\|_2^2\right] \asymp n^{-\frac{2s}{1+2s}}$$

which is $o(n^{-\frac{s}{1+2s}})$, a much improved rate (still worse than the ideal case $o(n^{-\frac{2s}{1+2s}})$). You may calculate the variance on your own. Now let us dissect what happens here and what is the underlying mathematical structure.

[4] This is because in semiparametric statistics, $\psi(\theta)$ is the parameter of interest and f is a nuisance parameter.

We first look at

$$\mathbb{E}\left[\widehat{\psi}_1(\theta) - \psi(\theta)\right] = \mathbb{E}\left[2Yf(X) - f(X)^2 - \int_0^1 f(x)^2 p(x)\mathrm{d}x\right].$$

It turns out that the random variable inside the expectation, $2Yf(X) - f(X)^2 - \psi(\theta)$, is the influence function/influence curve/canonical gradient $\dot{\psi}_\theta(O)$ of $\psi(\theta)$, where $O = (X, Y)$, evaluated at the nuisance parameter θ. And the estimator $\widehat{\psi}_1(\widehat{f}_n)$ is nothing but correcting the bias of the plug-in estimator $\psi(\widehat{f}_n)$ by the influence function $\dot{\psi}_{\widehat{f}_n}$ with the f component of the nuisance parameter θ evaluated at $\widehat{f}_n$:

$$\psi(\widehat{f}_n) + \frac{1}{n}\sum_{i=1}^n \dot{\psi}_{\widehat{f}_n}(O_i) = \int \widehat{f}_n(x)^2 p(x)\mathrm{d}x + \frac{2}{n}\sum_{i=1}^n Y_i \widehat{f}_n(X_i) - \int_0^1 \widehat{f}_n(x)^2 p(x)\mathrm{d}x - \int_0^1 \widehat{f}_n(x)^2 p(x)\mathrm{d}x.$$

Similarly,

$$\psi(\widehat{\theta}_n) + \frac{1}{n}\sum_{i=1}^n \dot{\psi}_{\widehat{\theta}_n}(O_i) = \frac{1}{n}\sum_{i=1}^n \widehat{f}_n(X_i)^2 + \frac{2}{n}\sum_{i=1}^n Y_i \widehat{f}_n(X_i) - \frac{1}{n}\sum_{i=1}^n \widehat{f}_n(X_i)^2 - \frac{1}{n}\sum_{i=1}^n \widehat{f}_n(X_i)^2.$$

The influence function $\dot{\psi}_\theta(O)$ has the following important property: $\mathbb{E}_\theta \dot{\psi}_\theta(O) \equiv 0$ for any θ, as long as θ in the expectation and θ in the influence function coincide.

To explain what is an influence function, we use the following functional Taylor expansion when assuming the underlying functional is sufficiently smooth (related to but not the smoothness nuisance parameter): Plug-in estimator can be viewed as a 0-th order Taylor expansion of the functional $\psi(\theta)$

$$\psi(\theta) = \psi(\widehat{\theta}_n) + \mathcal{O}(\|\theta - \widehat{\theta}_n\|)$$

whereas $\widehat{\psi}_1(\widehat{\theta}_n)$, by the above reasoning, can be viewed as a 1-st order Taylor expansion:

$$\psi(\theta) = \psi(\widehat{\theta}_n) + \psi'_{\widehat{\theta}_n}(\theta - \widehat{\theta}_n) + \mathcal{O}(\|\theta - \widehat{\theta}_n\|^2). \tag{8}$$

But for a functional $\psi(\theta)$, what is its first-order functional derivative $\psi'_{\widehat{\theta}_n}(\theta - \widehat{\theta}_n)$ whose nuisance parameter is taken a value $\widehat{\theta}_n$, mapping $\theta - \widehat{\theta}_n$ to $\mathbb{R}$? In fact, it is

$$\int \dot{\psi}_{\widehat{\theta}_n}(o)\left(\mathrm{d}P(o;\theta) - \mathrm{d}P(o;\widehat{\theta}_n)\right)$$

for some (mean-zero under law $P_{\widehat{\theta}_n}$) variable $\dot{\psi}_{\widehat{\theta}_n}(o)$. You can learn the above results in any books related to Calculus of Variation, e.g. **?**.

Now the question becomes what is this mysterious $\dot{\psi}_{\widetilde{\theta}}(o)$ for any $\widetilde{\theta}$: Consider a perturbation $\widetilde{\theta}_t$ such that $\widetilde{\theta}_{t=0} = \widetilde{\theta}$. We consider the following functional differentiation operation, which applies to all examples we consider in this note and almost all examples people are dealing with for smooth functionals in practice,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\psi(\widetilde{\theta}_t)\bigg|_{t=0} &= \mathbb{E}_{\widetilde{\theta}}\left[\dot{\psi}_{\widetilde{\theta}}(O)\cdot S_{\widetilde{\theta}}(O)\right] = \int \dot{\psi}_{\widetilde{\theta}}(o)\,\frac{\partial \log p(o;\widetilde{\theta}_t)}{\partial t}\bigg|_{t=0} p(o;\widetilde{\theta})\mathrm{d}o = \int \dot{\psi}_{\widetilde{\theta}}(o)\,\frac{\partial p(o;\widetilde{\theta}_t)}{\partial t}\bigg|_{t=0}\mathrm{d}o \\
&= \int \dot{\psi}_{\widetilde{\theta}}(o)\lim_{t\to 0}\frac{p(o;\widetilde{\theta}_t)-p(o;\widetilde{\theta})}{t}\mathrm{d}o = \lim_{t\to 0}\int \dot{\psi}_{\widetilde{\theta}}(o)\frac{p(o;\widetilde{\theta}_t)-p(o;\widetilde{\theta})}{t}\mathrm{d}o \\
&= \lim_{t\to 0}\frac{1}{t}\left\{\int \dot{\psi}_{\widetilde{\theta}}(o)\mathrm{d}P(o;\widetilde{\theta}_t) - \int \dot{\psi}_{\widetilde{\theta}}(o)\mathrm{d}P(o;\widetilde{\theta})\right\} \\
&= \lim_{t\to 0}\frac{1}{t}\left\{\mathbb{E}_{\widetilde{\theta}_t}\dot{\psi}_{\widetilde{\theta}}(O) - \underbrace{\mathbb{E}_{\widetilde{\theta}}\dot{\psi}_{\widetilde{\theta}}(O)}_{\equiv 0}\right\}
\end{aligned} \tag{9}$$

where $S_{\widetilde{\theta}}(O)$ is the score of the model and $\dot{\psi}_{\widetilde{\theta}}(O)$ is the first-order influence function of the functional. The above influence function representation is in fact a manifestation of the celebrated Riesz representation theorem in functional analysis.

Below we will use the quadratic functional as an example to show you how to derive the influence function $\dot{\psi}_{\widetilde{\theta}}(O)$ using the above Riesz representation:

1. Write down $\psi(\theta_t)$ explicitly:

$$\psi(\theta_t) = \int f_t(x)^2 p_t(x)\mathrm{d}x = \int_x \left\{\int_y y p_t(y|x)\mathrm{d}y\right\}^2 p_t(x)\mathrm{d}x.$$

2. Take derivative over t and set $t = 0$; "create" scores for the joint law $p_t(x,y) = p_t(y|x)p_t(x)$ (thus $\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\log p_t(x,y) = \frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\log p_t(y|x) + \frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\log p_t(x)$ in turn gives $S_\theta(X,Y) = S_\theta(Y|X) + S_\theta(X)$):

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\psi(\theta_t) \\
&= \int_x f(x)^2\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}p_t(x)\mathrm{d}x + \int_x 2f(x)p(x)\left\{\int_y y\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}p_t(y|x)\mathrm{d}y\right\}\mathrm{d}x \\
&= \int_x f(x)^2\left(\frac{1}{p(x)}\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}p_t(x)\right)p(x)\mathrm{d}x + \int_x 2f(x)p(x)\left\{\int_y y\left(\frac{1}{p(y|x)}\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}p_t(y|x)\right)p(y|x)\mathrm{d}y\right\}\mathrm{d}x \\
&= \int_x f(x)^2\left(\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\log p_t(x)\right)p(x)\mathrm{d}x + \int_x 2f(x)p(x)\left\{\int_y y\left(\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\log p_t(y|x)\right)p(y|x)\mathrm{d}y\right\}\mathrm{d}x \\
&= \int_x f(x)^2 S_\theta(x)p(x)\mathrm{d}x + \int_x 2f(x)p(x)\left\{\int_y yS_\theta(y|x)p(y|x)\mathrm{d}y\right\}\mathrm{d}x \\
&= \int_x f(x)^2 S_\theta(x)p(x)\mathrm{d}x + \int_x\int_y 2f(x)yS_\theta(y|x)p(y|x)p(x)\mathrm{d}y\mathrm{d}x \\
&= \mathbb{E}_\theta\left[f(X)^2 S_\theta(X)\right] + \mathbb{E}_\theta\left[2Yf(X)S_\theta(Y|X)\right].
\end{aligned}$$

This seems to be quite close to what we eventually want ($\mathbb{E}[\dot{\psi}_\theta(X,Y)S_\theta(X,Y)]$).

3. At this step, recall the following facts about the marginal and conditional scores:

$$\mathbb{E}_\theta[S_\theta(X)] \equiv 0, \ \mathbb{E}_\theta\left[S_\theta(Y|X)|X\right] \equiv 0, \ \text{and} \ S_\theta(X,Y) = S_\theta(X) + S_\theta(Y|X)$$

   Now for the first term, we can directly add $S_\theta(Y|X)$ because

$$\mathbb{E}_\theta[f(X)^2 S_\theta(Y|X)|X] = f(X)^2 \mathbb{E}_\theta[S_\theta(Y|X)|X] = 0.$$

   Thus $\mathbb{E}_\theta\left[f(X)^2 S_\theta(X)\right] = \mathbb{E}_\theta\left[f(X)^2(S_\theta(X) + S_\theta(Y|X))\right] = \mathbb{E}_\theta\left[f(X)^2 S_\theta(X,Y)\right]$. For the second term, we can further rewrite it as follows

$$\mathbb{E}_\theta\left[2Yf(X)S_\theta(Y|X)\right] = \mathbb{E}_\theta\left[(2Yf(X) - 2f(X)^2)S_\theta(Y|X)\right].$$

   Though we can subtract any function of X from $2Yf(X)$, we choose this function to be $2f(X)^2$ because $f(X)^2 = f(X)\mathbb{E}_\theta[Y|X]$, which means $\mathbb{E}_\theta\left[2Yf(X) - 2f(X)^2|X\right] \equiv 0$. Thus we can now add any function of X to $S_\theta(Y|X)$ and we choose this function to be $S_\theta(X)$:

$$\begin{aligned}\mathbb{E}_\theta\left[2Yf(X)S_\theta(Y|X)\right] &= \mathbb{E}_\theta\left[(2Yf(X) - 2f(X)^2)S_\theta(Y|X)\right] \\ &= \mathbb{E}_\theta\left[(2Yf(X) - 2f(X)^2)(S_\theta(X) + S_\theta(Y|X))\right] \\ &= \mathbb{E}_\theta\left[(2Yf(X) - 2f(X)^2)S_\theta(X,Y)\right].\end{aligned}$$

   Combining the above two terms we have

$$\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\psi(\theta_t) = \mathbb{E}_\theta\left[(2Yf(X) - 2f(X)^2 + f(X)^2)S_\theta(X,Y)\right] = \mathbb{E}_\theta\left[(2Yf(X) - f(X)^2)S_\theta(X,Y)\right].$$

4. We are one step away. Since we also need $\mathbb{E}_\theta\dot{\psi}_\theta(X,Y) = 0$, we can simply recenter $2Yf(X) - f(X)^2$ by a constant (its expectation $\int f(x)^2 p(x)\mathrm{d}x = \psi(\theta)$). Thus eventually, we have

$$\dot{\psi}_\theta(X,Y) = 2Yf(X) - f(X)^2 - \psi(\theta).$$

For a more comprehensive introduction to semiparametric theory, see **??????** and the notes written by Yen-Chi Chen.

Remark 22. In the beginning of this section, I mentioned that the approach in this section is essentially what is going on for debiased lasso. Can you see why after reading the following papers **??**?

10.3 Higher-order influence functions – A unified framework for smooth functional estimation

In this section, we investigate if the above first-order estimator can still be improved. Amazingly, I will show an estimator, under the assumption that g is known, to achieve the following rate of convergence that is conjectured to be minimax optimal for $\psi(\theta) = \mathbb{E}[f(X)^2]$:

$$\inf_{\widehat{\psi}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[(\widehat{\psi}_n - \psi(\theta))^2\right] \lesssim \begin{cases} n^{-1} & s > \frac{1}{4}, \\ n^{-\frac{8s}{1+4s}} & s \leq \frac{1}{4}. \end{cases} \tag{10}$$

So the optimal rate exhibits a phase transition phenomenon and very interestingly, it is possible to obtain the parametric n^{-1} rate for functionals even under an infinite-dimensional statistical model. Thus when $s > \frac{1}{4}$, this nonparametric problem has parametric behavior.

Consider the following estimator: choose $k' = \max\{n, n^{\frac{2}{1+4s}}\}$,

$$\widehat{\psi}_{2,k'}(\widehat{\theta}_n) = \widehat{\psi}_1(\widehat{\theta}_n) + \frac{1}{n(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} (Y_{i_1} - \widehat{f}_n(X_{i_1}))\underline{z}_{k'}(X_{i_1})^\top \Sigma_{k'}^{-1} \underline{z}_{k'}(X_{i_2})(Y_{i_2} - \widehat{f}_n(X_{i_2}))$$

where $\Sigma_{k'} = \mathbb{E}[\underline{z}_{k'}(X)\underline{z}_{k'}(X)^\top]$. The added term is actually the second-order influence function of $\psi(\theta)$ evaluated at $\widehat{\theta}_n$[5]. To see how $\widehat{\psi}_{2,k'}(\widehat{\theta}_n)$ improves upon $\widehat{\psi}_1(\widehat{\theta}_n)$, we first analyze its bias:

$$\begin{aligned}
&\mathbb{E}_\theta\left[\widehat{\psi}_{2,k'}(\widehat{\theta}_n) - \psi(\theta)|O_{nuis}\right] \\
= {}& \mathbb{E}_\theta\left[\widehat{\psi}_1(\widehat{\theta}_n) - \psi(\theta)|O_{nuis}\right] + \mathbb{E}_\theta\left[(f(X) - \widehat{f}_n(X))\underline{z}_{k'}(X)^\top|O_{nuis}\right]\Sigma_{k'}^{-1}\mathbb{E}_\theta\left[\underline{z}_{k'}(X)(f(X) - \widehat{f}_n(X))|O_{nuis}\right] \\
= {}& \left\{\mathbb{E}_\theta\left[(f(X) - \widehat{f}_n(X))\underline{z}_{k'}(X)^\top|O_{nuis}\right]\Sigma_{k'}^{-1}\right\}\Sigma_{k'}\left\{\Sigma_{k'}^{-1}\mathbb{E}_\theta\left[\underline{z}_{k'}(X)(f(X) - \widehat{f}_n(X))|O_{nuis}\right]\right\} \\
&- \mathbb{E}_\theta\left[(f(X) - \widehat{f}_n(X))^2\right].
\end{aligned}$$

Speculating the red term carefully, it is not difficult to see it is actually the inner product of the following quantity:

$$\left\{\mathbb{E}_\theta\left[(f(X) - \widehat{f}_n(X))\underline{z}_{k'}(X)^\top|O_{nuis}\right]\Sigma_{k'}^{-1}\right\}\underline{z}_{k'}(x) \equiv \beta_{k'}^\top \underline{z}_{k'}(x)$$

but $\beta_{k'}$ is now the regression coefficient of the regression (or L_2 linear projection) between $f(X) - \widehat{f}_n(X)$ (or equivalently $Y - \widehat{f}_n(X)$) and the k'-dimensional basis $\underline{z}_{k'}$. Since it is not difficult to see $f - \widehat{f}_n$ belongs to $\mathcal{H}(s; B)$, we have $\|f - \widehat{f}_n - \beta_{k'}^\top \underline{z}_{k'}\|_\infty \asymp k'^{-s}$ and in fact, finishing the above calculations, we have

$$\begin{aligned}
&\mathbb{E}_\theta\left[\widehat{\psi}_{2,k'}(\widehat{\theta}_n) - \psi(\theta)|O_{nuis}\right] \\
= {}& \mathbb{E}_\theta\left[\beta_{k'}^\top \underline{z}_{k'}(X)\underline{z}_{k'}(X)^\top \beta_{k'} - (f(X) - \widehat{f}_n(X))^2\right] \\
= {}& \mathbb{E}_\theta\left[\left\{f(X) - \widehat{f}_n(X) - \beta_{k'}^\top \underline{z}_{k'}(X)\right\}^2\right] \lesssim k'^{-2s}.
\end{aligned}$$

We are left to show the variance of $\widehat{\psi}_{2,k'}(\widehat{\theta}_n)$, which is a sum between a sum of i.i.d. (of order $1/n$) and a second-order U-statistic (of order $(1/n) \vee (k'/n^2)$). The variance of U-statistic in general can be computed via Hoeffding's decomposition. This part will be left as an exercise. By balancing the bias-variance trade-off, we have $k'^{-4s} \asymp \frac{1}{n} \vee \frac{k'}{n^2}$. Thus $k' \asymp n^{\frac{2}{1+4s}}$, which in turn gives us

$$n^{-\frac{8s}{1+4s}} \vee n^{-1}.$$

When g is known, it is easy to show the above rate to be tight; see **?**.

When g is unknown, looking at $\widehat{\psi}_{2,k'}(\widehat{\theta}_n)$, the only unknown term is $\Sigma_{k'}^{-1} = \{\mathbb{E}_\theta[\underline{z}_{k'}(X)\underline{z}_{k'}(X)^\top]\}^{-1} = \left\{\int \underline{z}_{k'}(x)\underline{z}_{k'}(x)^\top p(x)\mathrm{d}x\right\}^{-1}$. How to solve this problem? If we assume $g \in \mathcal{H}(s_g; B)$, then we

[5] Here I am not being very rigorous. If you are interested in this theory, you should read **??**. But be cautious with typos.

can estimate g by $\widehat{g}$ again using the sub-data O_{nuis} and optimal wavelet bases and estimate $\int \underline{z}_{k'}(x)\underline{z}_{k'}(x)^\top p(x)\mathrm{d}x$ by $\widehat{\Sigma}_{k'} \equiv \int \underline{z}_{k'}(x)\underline{z}_{k'}(x)^\top \widehat{g}(x)\mathrm{d}x$. But now one needs to be careful about the error introduced by $\widehat{g} - g$. Eventually, one needs to correct such bias by adding more bias correction terms, by

$$\begin{aligned}&\widehat{\psi}_{m,k'}(\widehat{\theta}_n)\\&= \widehat{\psi}_{2,k'}(\widehat{\theta}_n) + \sum_{j=3}^{m}(-1)^j \frac{(n-j)!}{n!} \sum_{1 \leq i_1 \neq \cdots \neq i_j \leq n} (Y_{i_1} - \widehat{f}(X_{i_1}))\underline{z}_{k'}(X_{i_1})^\top \left\{ \prod_{\ell=3}^{j} \widehat{\Sigma}_{k'}^{-1} \left(\underline{z}_{k'}(X_\ell)\underline{z}_{k'}(X_\ell)^\top - \widehat{\Sigma}_{k'} \right) \right\} \widehat{\Sigma}_{k'}^{-1} \underline{z}_{k'}(X_{i_2})(Y_{i_2} - \widehat{f}(X_{i_2})).\end{aligned}$$

Then we choose m appropriately. But if we do not want to assume any smoothness on g, we eventually need $m \to \infty$.

One final caveat: when letting m be large, the variance of $\widehat{\psi}_{m,k'}(\widehat{\theta}_n)$ will increase when $k' > n$ (i.e. when $s < 1/4$). Then one also needs to cut out certain terms in the basis function $\underline{z}_{k'}$ through a very delicate hyperbolic cut scheme detailed in Section 4 of **?**.

Remark 23. The concept of higher-order influence functions or higher-order functional gradients first appeared in a series of works by Johan Pfanzagl [**??**], who generalized the information bound calculations for parametric-nonparametric (nowadays called semiparametric) models initiated (again!) by Charles Stein [**?**]; also see **?** and **?**. In 2004, when studying the problem of optimal sequential decision making (nowadays called reinforcement learning), **?** first reported partial results on higher-order influence functions, a then on-going collaborating project with Aad van der Vaart. These results eventually culminated in Lingling Li and Eric Tchetgen Tchetgen's joint PhD Thesis [**??**]. **?** developed an interesting potentially alternative strategy by diverging numbers of bootstrapping. An earlier version of this bootstrap idea could be traced back to Guang Cheng's PhD Thesis [**?**]; also see the Bayesian version by **?**. Cun-Hui Zhang and Pierre Bellec recently developed a second-order theory for high-dimensional sparse linear regression problems, coined as "Second-Order Stein" [**??**]. Higher order influence functions, and all these above works, are about higher-order accuracy of statistical procedures. Yet most contemporary statistical methods and theory are about first-order accuracy.

In a nutshell, the higher-order influence function approach simply keeps Taylor-expanding the underlying functional $\psi(\theta)$ up to a certain order, and hope that

$$\psi(\theta) = \psi(\widehat{\theta}_n) + \sum_{j=1}^{m} \psi_{\widehat{\theta}_n}^{(j)}(\theta - \widehat{\theta}_n)^{\otimes j} + \mathcal{O}(\|\theta - \widehat{\theta}_n\|^{m+1}). \tag{11}$$

An m-th order influence function is the first-order influence function of an $(m-1)$-th order influence function, and can be calculated via the same calculation as that of the first-order influence function.

Higher-order influence functions suggest the following interesting phenomenon: to obliviate any assumptions on the marginal density of X, the estimator is in the computational complexity class Exp.

Remark 24. The following open problem, though extremely simple to describe, has puzzled Robins for almost 20 years. Many people have tried but failed. Now no one is working on this problem because the risk is too high and there are tons of new problems due to deep learning. People who have tried this problem believe some significant intellectual leap is required to solve this problem. I simply paraphrase the problem statement from **?**.

$X \in [0,1]^d$ is random (random design) and the effective smoothness $s/d < 1/4$ but $s > 1$. Does there exist an estimator of $\mathbb{E}[f(X)^2] = \int_{[0,1]^d} \{\mathbb{E}[Y|X = x]\}^2 p(x) \mathrm{d}x$ that converges at the conjectured optimal rate $n^{-\frac{4s}{d+4s}}$ when $s/d < 1/4$ without any condition on g except that g exists and is bounded above and below? Since one does not want to put any structural assumptions on g, it might be reasonable to conjecture that the behavior should be similar to the case when g has zero smoothness, i.e. X is fixed. When X is fixed, interestingly **?** have shown that the rate of convergence should be at a much slower order $n^{-2s/d}$ instead.

However, when $s/d < 1/4$ and $s < 1$, **?** construct a clever but non-generalizable estimator that converges at rate $n^{-\frac{4s}{d+4s}}$, without requiring any smoothness assumption on the density of X (except that X is random instead of fixed). So it seems that even just a little bit randomness should help a lot!

Robins has been trying to find an answer to this question without success for a number of years; one of Robins' students Lingling Li conjectured that when $s > 1$ the rate $n^{-\frac{4s}{d+4s}}$ is not achievable and should depend on the smoothness of g, unlike the case $s \leq 1$ but no one has any idea how to establish a matching lower bound for the rate conjectured by Lingling. He suggested that it is now time for some crowd-sourcing. At a first sight, a recent paper by **?** comes very close, but it turns out they are equally far from the final results as most people are.

There are several questions left unanswered (actually open till today):

1. How to adapt over unknown smoothness s without knowledge on g? **?** partially answered this question by assuming g to be sufficiently smooth. But the result below this smoothness condition is very difficult to obtain (this is because exponential tail inequalities for higher-order U-statistics are too loose to use).

2. Another open problem is if it is possible to construct optimal adaptive confidence intervals for low-dimensional functionals.

11 Adaptive minimax optimal inference

As I mentioned in class, it is very difficult for Lespki's method to pinpoint the "correct" resolution j^* or equivalently estimate the "correct" smoothness. If on the contrary it is possible to do so, we would be able to construct the so-called honest and minimax adaptive confidence sets for f in Hölder balls, centered at an adaptive minimax rate-optimal estimator $\widehat{f}_{\widehat{j}}$. But it has been proven to be impossible to construct such adaptive confidence sets for the entire Hölder ball; see Chapter 8.3 of **?**. The state-of-the-art is that we have to remove a lot of functions from Hölder ball and focus only on the so-called "self-similar" classes. Due to time constraint, for now I am planning to leave this part for self-study.

12 Final remarks

There are other problems with a similar flavor to the problems considered in this short note: e.g. estimating the sparsity of a high-dimensional sparse linear model [**?**]. Also, one can actually embed high-dimensional linear models into the nonparametric statistics framework by considering the so-called re-arranged Besov-type spaces. Unlike classical Besov-type spaces, for which the Fourier coefficients decay geometrically with the energy/frequency of the basis functions, re-arranged Besov-type spaces do not have an *a-priori* order on the basis functions and thus selection is necessary. The key distinction between high-dimensional statistics and classical non-parametric statistics can be understood via this mere fact.

In this chapter, we have explored several techniques of proving (adaptive) minimax upper and lower bounds through the lens of the theory of function spaces. All these results now have analogues in Bayesian settings and deep learning settings. Most of the literature focused on estimation of a regression function or a conditional expectation. A more challenging problem is what if the function is a solution to a complicated stochastic PDE. This direction will become more important as nowadays large-scale/high-dimensional PDEs start to be routinely solved by deep learning and stochastic gradient descent. But the data that are used to parameterize the PDEs are often noisy physical measurements and how to quantify the uncertainty in a statistically rigorous way will become an important topic in statistics, e.g. **??**.

A Proof of $N_{r,h}$ approximating δ-function

We need to apply the following result:

Proposition 1. *$f : \mathbb{R} \to \mathbb{R}$ a measurable function and $K \in L_1$ a kernel function s.t. $\int_{\mathbb{R}} K(x)dx = 1$. Then, denoting $K_h(\cdot) = h^{-1}K(\cdot/h)$*

1. *f bounded on $\mathbb{R}$ and continuous at $x \in \mathbb{R} \Rightarrow K_h * f(x) \to f(x)$ as $h \to 0$ pointwise.*
2. *f bounded and uniformly continuous on $\mathbb{R} \Rightarrow \|K_h * f - f\|_\infty \to 0$ as $h \to 0$.*
3. *$f \in L_p$ for some $1 \leq p < \infty \Rightarrow \|K_h * f - f\|_p \to 0$ as $h \to 0$.*

Now by the definition of $N_{r,h}$, it is easy to see that $\|N_{r,h}\| = 1$ so $N_{r,h}$ is a kernel like K in Proposition 1. Then applying Proposition 1 completes the proof. We are left to show Proposition 1.

Proof of Proposition 1.

$$\begin{aligned} K_h * f(x) - f(x) &= \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\ &= \int_{\mathbb{R}} K(u)(f(x-hu) - f(x))du. \end{aligned}$$

Then parts 1 and 2 are obvious from the above result (consult the proof of Proposition 4.1.1 of **?** if you are not convinced yet). For part 3, we need to use a famous result from functional analysis – the Minkowski integral inequality that helps you exchange integrals; see Lemma 25. Then we have

$$\left\| \int_{\mathbb{R}} K(u)(f(\cdot - hu) - f(\cdot))du \right\|_p \leq \int_{\mathbb{R}} |K(u)| \|f(\cdot - hu) - f(\cdot)\|_p du.$$

Since $\|f(\cdot - hu) - f(\cdot)\|_p \to 0$ as $h \to 0$ pointwise for u and $\|f(\cdot - hu) - f(\cdot)\|_p$ is uniformly bounded by $2\|f\|_p$, then by DCT, the above display will converge to 0. □

Lemma 25 (Minkowski's integral inequality)**.** *For a bivariate function $f(x,y)$, the following holds:*

$$\left\{\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(x,y)dx\right|^{p} dy\right\}^{1/p} \leq \int_{\mathbb{X}}\left\{\int_{\mathbb{Y}} |f(x,y)|^{p} dy\right\}^{1/p} dx.$$

Proof. When $p=1$, it follows from Fubini's theorem. When $p>1$, then

$$\begin{aligned}
\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(x,y)dx\right|^{p} dy &= \int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(x,y)dx\right|^{p-1}\int_{\mathbb{X}} f(x,y)dxdy \\
&\leq \int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(t,y)dt\right|^{p-1}\int_{\mathbb{X}} |f(x,y)|dxdy \\
&= \int_{\mathbb{Y}}\int_{\mathbb{X}}\left|\int_{\mathbb{X}} f(t,y)dt\right|^{p-1} |f(x,y)|dxdy \\
&= \int_{\mathbb{X}}\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(t,y)dt\right|^{p-1} |f(x,y)|dydx \\
&\overset{\star}{\leq} \int_{\mathbb{X}}\left\{\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(t,y)dt\right|^{q(p-1)} dy\right\}^{1/q}\left\{\int_{\mathbb{Y}} |f(x,y)^{p}|dy\right\}^{1/p} dx \\
&= \left\{\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(t,y)dt\right|^{p} dy\right\}^{1/q}\int_{\mathbb{X}}\left\{\int_{\mathbb{Y}} |f(x,y)^{p}|dy\right\}^{1/p} dx
\end{aligned}$$

where in step $\star$ we use Hölder's inequality with $q = p/(p-1)$. Finally dividing both sides by $\left\{\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(t,y)dt\right|^{p} dy\right\}^{1/q}$, we have, by $1-1/q = 1/p$,

$$\begin{aligned}
\left\{\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(x,y)dx\right|^{p} dy\right\}^{1/p} &= \left\{\int_{\mathbb{Y}}\left|\int_{\mathbb{X}} f(x,y)dx\right|^{p} dy\right\}^{1-1/q} \\
&\leq \int_{\mathbb{X}}\left\{\int_{\mathbb{Y}} |f(x,y)^{p}|dy\right\}^{1/p} dx.
\end{aligned}$$

□

B Multidimensional extension

B.1 Sobolev spaces

Let $m = (m_1, \cdots, m_d)$ be non-negative integers and define $|m| = m_1 + \cdots + m_d$. Given $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, write $x^m = x^{m_1} \cdots x^{m_d}$ and

$$D^m = \frac{\partial^{|m|}}{\partial x_1^{m_1} \cdots \partial x_d^{m_d}}.$$

Then Sobolev space on $([a,b])^d$ is

$$W_{m,p}^{d} \coloneqq \left\{f \in L_p([a,b]^d) : D^{\alpha} f \in L_p([a,b]^d) \;\; \forall \; |\alpha| \leq m\right\}.$$

B.2 Besov spaces

Two good references for multidimensional extension of Besov spaces are **?** and **?**.

C Reproducing Kernel Hilbert Space (RKHS)

RKHS has been a quite important topic in recent history of statistical learning theory [**??**], in which people try to look for spaces that are learnable in high dimensional or nonlinear settings. RKHS is an extremely small space compared to what we have covered in this note (Hölder, Besov, and Sobolev spaces with $d > 1$). But recent progress in deep learning theory (such as Neural Tangent Kernel [**???**] and Weinan E's works) definitely suggests that RKHS should still be quite useful in the future.

RKHS is a class of very smooth functions defined via the so-called Mercer kernel.

Definition 26. A Mercer kernel is a continuous function $K : [a, b] \times [a, b] \to \mathbb{R}$ s.t. $K(x, y) = K(y, x)$ (symmetric) and K is p.s.d. in the following sense:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j) c_i c_j \geq 0$$

for all finite sets of points $x_1, \cdots, x_n \in [a, b]$ and all reals $c_1, \cdots, c_n$.

We then have the famous Mercer's theorem:

Theorem 27. *Suppose (1) $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is symmetric and $\sup_{x,y} K(x, y) < \infty$, and define the operator*

$$T_K f(x) = \int_{\mathbb{X}} K(x, y) f(y) dy,$$

and (2) $T_K : L_2(\mathbb{X}) \to L_2(\mathbb{X})$ is p.s.d.: $\int_{\mathbb{X}} \int_{\mathbb{X}} K(x, y) f(x) f(y) dx dy \geq 0 \ \forall \ f \in L_2(\mathbb{X})$. Let λ_i, Ψ_i be the eigenvalues and eigenfunctions of T_K, i.e.

$$\int_K K(x, y) \Psi_i(y) dy = \lambda_i \Psi_i(x).$$

Then $\sum_i \lambda_i < \infty$, $\sup_x \Psi_i(x) < \infty$ and

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

So with a Mercer kernel, we can always use eigen-decomposition of the Mercer kernel to obtain a set of basis functions.

Definition 28 (RKHS)**.** Given a kernel K, denote $K_x(y) = K(x, y)$. Let

$$\mathcal{H}_{0,K} = \mathsf{span} \left\{ f : f(x) = \sum_{j=1}^{k} \beta_j^f K_{x_j}(x) \right\},$$

with the following inner product: Given any two f and g from $\mathcal{H}_0$, define

$$\langle f, g \rangle_K = \sum_i \sum_j \beta_i^f \beta_j^g K(x_i, x_j).$$

Then $\mathcal{H}_{0,K}$ is an RKHS generated by K.

Key property of RKHS – reproducing! $\langle f, K_x \rangle = \sum_i \beta_i^f K(x_i, x) = f(x)$ and $\langle K_x, K_x \rangle = K(x, x)$. Here $T_K[\cdot](x) = \int K(x, y) \cdot (y) dy$ is called an "evaluation functional": input is a function and output is a point on that function. RKHS has continuous evaluational functional but general Hilbert spaces do not (because the evaluation functional has the δ function as the kernel).

D Other references on function spaces

Once you have a good understanding of function spaces, you can try to read this post from Terence Tao and a paper by Richard Nickl [**?**], together with its reference, in particular the papers/books by Triebel and colleagues.