Minimax estimation in Besov spaces

Yongming Li   Nichakan Loesatapornpipit   Jiyoung Park   Lucas Park

Math 663

Outline

Suppose we are given n noisy samples of a function f:

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

with $t_i = \frac{i}{n} \in [0,1]$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Assuming f belongs to a certain smoothness class $\mathcal{F}$, can we find an estimator $\hat{f}$ depending on data $y_1, \ldots, y_n$ such that $\hat{f}$ minimizes the risk

$$\mathbb{E} \int_0^1 \left(\hat{f}(t) - f(t)\right)^2 \mathrm{d}t. \tag{2}$$

Definition: Minimax risk

$$\mathcal{R}(n, \mathcal{F}) := \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E} \int_0^1 \left(\hat{f}(t) - f(t)\right)^2 \mathrm{d}t. \tag{3}$$

This talk: We consider the smoothness class of Besov spaces $B^s_{p,q}$, and give a quantitative answer to the above problem.

Besov spaces can be defined in various ways.

- Moduli of Smoothness
- Wavelet Coefficients
- Low-Frequency Approximations
- Littlewood-Paley Theory

Let $\Delta_h^r(f)(x)$ be the r difference defined by

$$\Delta_h^{(r)}(f)(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x+kh). \tag{4}$$

Let A be subinterval of $\mathbb{R}$. For $f \in L^p(A), 1 \le p \le \infty$, the rth modulus of smoothness is defined by

$$\omega_r(f,t) \equiv \omega_r(f,t,p) = \sup_{0<h\le t} \left\| \Delta_h^{(r)} f \right\|_p, \quad t > 0. \tag{5}$$

Given $s > 0$ and let $r > s$ be an integer. For $1 \le q \le \infty, 1 \le p \le \infty$, the Besov space is defined by

$$B_{pq}^s \equiv B_{pq}^s(A) = \begin{Bmatrix} f \in L^p(A) : \|f\|_{B_{pq}^s} \equiv \|f\|_p + |f|_{B_{pq}^s < \infty}, & 1 \le p < \infty, \\ f \in C_u(A) : \|f\|_{B_{pq}^s} \equiv \|f\|_\infty + |f|_{B_{pq}^s < \infty}, & p = \infty, \end{Bmatrix} \tag{6}$$

where

$$|f|_{B_{pq}^s} \equiv |f|_{B_{pq}^s(A)} = \begin{cases} \left(\int_0^\infty \left[\frac{\omega_r(f,t)}{t^s} \right]^q \frac{dt}{t} \right)^{\frac{1}{q}}, & 1 \le q < \infty, \\ \sup\limits_{t>0} \frac{\omega_r(f,t)}{t^s}, & q = \infty, \end{cases} \tag{7}$$

is the Besov seminorm.

Definition $\phi \in L^2(\mathbb{R})$ is the scaling function of a multiresolution analysis of $L^2(\mathbb{R})$ if it satisfies the following conditions

- The family $\{\phi(\cdot - k)\}_{k\in\mathbb{Z}}$ is an orthonormal system in $L^2(\mathbb{R})$; that is $<\phi(\cdot - k), \phi(\cdot - l)>= \delta_{k,l}$.
- The linear spaces

$$V_0 = \{f(x) = \sum_{k\in\mathbb{Z}} c_k\phi(x-k), \{c_k\}_{k\in\mathbb{Z}} : \sum_{k\in\mathbb{Z}} c_k^2 < \infty\},$$
$$V_1 = \{f(2x) : f \in V_0\}, \dots,$$
$$V_j = \{f(2^j x) : f \in V_0\}, \dots,$$

  are nested; that is,

$$V_0 \subset V_1 \subset V_2 \subset \dots$$

- $\overline{\cup_{j\geq 0} V_j} = L^2(\mathbb{R})$.

Then,

$$V_j = \text{span}\{\phi_{jk}(x) := 2^{j/2}\phi(2^j x - k)\}_{k=-\infty}^{\infty} \quad (8)$$

Since $V_0 \subset V_1$, the space V_1 can be defined by

$$V_1 = V_0 \oplus W_0 \tag{9}$$

where W_0 is the orthogonal complement of V_0 in V_1.
Since the spaces V_j are nested,

$$V_j = V_0 \oplus \left(\bigoplus_{\ell=0}^{j-1} W_\ell \right), \quad W_\ell := V_{\ell+1} \ominus V_\ell \tag{10}$$

Let $K_j(f)$ be the orthogonal L^2-projections of $f \in L^2$ onto V_j, which is defined by

$$K_j(f) = K_0(f) + \sum_{\ell=0}^{j-1} \text{the projections onto } W_\ell, \tag{11}$$

where

$$K_0(f)(x) = \sum_{k \in \mathbb{Z}} < \phi_k, f > \phi_k(x), \tag{12}$$

and $\phi_k(x) = \phi(x-k)$.

Find basis functions that span the spaces W_ℓ

Assume that there exists a fixed $\psi \in L^2(\mathbb{R})$ such that, for every $\ell \in \mathbb{N} \cup \{0\}$,

$$\{\psi_{\ell,k} := 2^{\ell/2}\psi(2^\ell(\cdot) - k) : k \in \mathbb{Z}\} \tag{13}$$

is an orthonormal set of functions that spans W_ℓ.

- Haar system: if $\phi = \mathbb{1}_{(0,1]}$ the Haar wavelet is $\psi = \mathbb{1}_{[0,\frac{1}{2}]} - \mathbb{1}_{(\frac{1}{2},1]}$.

The projection of f onto W_ℓ is

$$\sum_{k\in\mathbb{Z}} < \psi_{\ell k}, f > \psi_{\ell k} \tag{14}$$

Therefore, the projection $K_j(f)$ of f onto V_j is

$$K_j(f)(x) = \sum_{k\in\mathbb{Z}} < \phi_{jk}, f > \phi_{jk}(x) \tag{15}$$

$$= \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k(x) + \sum_{\ell=0}^{j-1}\sum_{k\in\mathbb{Z}} < \psi_{\ell k}, f > \psi_{\ell k}(x) \tag{16}$$

Since $\cup_{j\geq 0} V_j$ is dense in L^2,

$$L^2 = V_0 \oplus \left(\bigoplus_{\ell=0}^{\infty} W_\ell \right). \tag{17}$$

Hence,

$$\{\phi(x-k), 2^{\ell/2}\psi(2^\ell x - k) : k \in \mathbb{Z}, \ell \in \mathbb{N} \cup \{0\}\} \tag{18}$$

is an orthonormal wavelet basis of the Hilbert space L^2.
Let the scaling function $\phi_k(x) = \phi(x-k)$ be of the first wavelet ψ_{-1k}. This implies that we can abbreviate the wavelet basis as $\{\psi_{\ell k}\}$. Hence, every $f \in L^2$ has the wavelet series expansion

$$\begin{aligned} f &= \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k(x) + \sum_{\ell=0}^{\infty} \sum_{k\in\mathbb{Z}} < \psi_{\ell k}, f > \psi_{\ell k}(x) \\ &= \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k(x) + \sum_{k\in\mathbb{Z}} \sum_{\ell=0}^{\infty} < \psi_{\ell k}, f > \psi_{\ell k}(x) \\ &= \sum_{\ell\geq -1} \sum_{k\in\mathbb{Z}} < \psi_{\ell k}, f > \psi_{\ell k}(x). \end{aligned} \tag{19}$$

Definition A multiresolution wavelet basis

$$\{\phi_k = \phi(x-k), \psi_{\ell k} = 2^{\ell/2}\psi(2^\ell(x) - k) : k \in \mathbb{Z}, \ell \in \mathbb{N} \cup \{0\}\}$$

of $L^2(\mathbb{R})$ with the projection kernel $K(x,y) = \sum_{k\in\mathbb{Z}} \phi(x-k)\phi(y-k)$ is said to be S-regular for some $S \in \mathbb{N}$ if the following conditions are satisfied:

- $\int_{\mathbb{R}} \psi(u)u^\ell du = 0 \ \forall \ell = 0, 1, \ldots, S-1, \quad \int_{\mathbb{R}} \phi(x)dx = 1$, and for all $v \in \mathbb{R}$,

$$\int_{\mathbb{R}} K(v, v+u)du = 1, \int_{\mathbb{R}} K(v, v+u)u^\ell du = 0 \ \forall \ell = 1, \ldots, S-1,$$

- $\sum_{k\in\mathbb{Z}} |\phi(x-k)| \in L^\infty(\mathbb{R}), \ \sum_{k\in\mathbb{Z}} |\psi(x-k)| \in L^\infty(\mathbb{R})$, and
- For $\kappa(x,y)$ equal to either $K(x,y)$ or $\sum \psi(x-k)\psi(y-k)$,

$$\sup_{v\in\mathbb{R}} |\kappa(v, v-u)| \le c\Phi(c_2|u|), \text{ for some } 0 < c_1, c_2 < \infty \text{ and every } u \in \mathbb{R},$$

  for some bounded integrable $\Phi : [0,\infty) \to \mathbb{R}$ such that $\int_{\mathbb{R}} |u|^S \Phi(|u|)du < \infty$.

Definition by Wavelet Coefficients

We will use a wavelet basis of regularity $S > s, S \in \mathbb{N}$ satisfying $\phi, \psi \in C^S(\mathbb{R})$ with $D^S\phi, D^S\psi$ dominated by some integrable function. Starting from the wavelet series

$$f = \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k + \sum_{\ell=0}^{\infty} \sum_{k\in\mathbb{Z}} < \psi_{\ell k}, f > \psi_{\ell k}, \text{ in } L^p, 1 \le p \le \infty, \tag{20}$$

of $f \in L^p(\mathbb{R})(p < \infty)$ and of $f \in C_u(\mathbb{R})(p = \infty)$. The idea is to use the decay, as $\ell \to \infty$, of the L^p norms

$$\left\| \sum_k < f, \psi_{\ell k} > \psi_{\ell k} \right\|_{L^p} \simeq 2^{\ell(\frac{1}{2} - \frac{1}{p})} \|\{< f, \psi_{\ell k} >\}_k\|_{\ell^p}$$

to describe the regularity of a function f.
Fot $1 \le p \le \infty, 1 \le q \le \infty, 0 < s < S$, we set

$$B^{s,W}_{pq} \equiv \begin{cases} f \in L^p(\mathbb{R}) : \|f\|_{B^{s,W}_{pq}} < \infty, & 1 \le p < \infty, \\ f \in C_u(\mathbb{R}) : \|f\|_{B^{s,W}_{pq}} < \infty, & p = \infty, \end{cases} \tag{21}$$

with wavelet-sequence norm, given, for $s \in \mathbb{R}$, by

$$\|f\|_{B^{s,W}_{pq}} \equiv \begin{cases} \|\{< f, \phi_k >\}_k\|_p + \left(\sum_{\ell=0}^{\infty} 2^{q\ell(s + \frac{1}{2} - \frac{1}{p})} \|\{< f, \psi_{\ell k} >\}_k\|_p^q \right)^{1/q}, & 1 \le q < \infty \\ \|\{< f, \phi_k >\}_k\|_p + \sup_{\ell > 0} 2^{\ell(s + \frac{1}{2} - \frac{1}{p})} \|\{< f, \psi_{\ell k} >\}_k\|_p, & q = \infty. \end{cases} \tag{22}$$

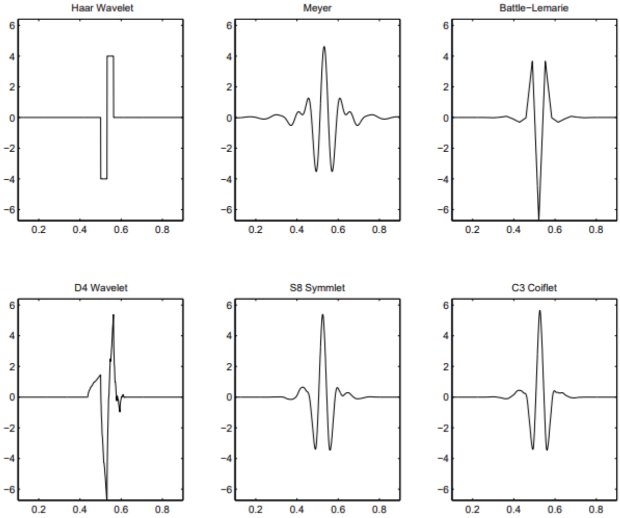


Figure 1: Wavelet basis. Source: Iain M. Johnstone's textbook pg. 192

Theorem (Donoho-Johnstone '98)

Let $\mathcal{F} = B_{B^s_{p,q}([0,1])}(0,1)$ *be a unit ball in the Besov space* $B^s_{p,q}$*, where*

$$s > \frac{1}{p}, \quad 1 \le p, q \le \infty \quad \text{or} \quad s = p = q = 1. \tag{23}$$

Let $\mathcal{R}(n,\mathcal{F}) = \inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}_{D_n} \left\| \widehat{f} - f \right\|^2_{L^2(P_x)}$ *denote the minimax risk from observations and let* $\mathcal{R}_L(n,\mathcal{F})$ *denote the minimax risks when the infimum is restricted to be linear in the data* (y_i)*. Here,* $D_n = (x_i, f(x_i) + \eta_i)_{i=1,\dots,n}$ *with* $x_i \overset{i.i.d}{\sim} P_x$ *an uniform measure over* $[0,1]$ *and* $\eta_i \overset{i.i.d}{\sim} N(0,\sigma^2)$*. Then,*

1. *Minimax rate:*

$$\mathcal{R}(n,\mathcal{F}) \asymp n^{-\frac{2s}{2s+1}} \qquad \mathcal{R}_L(n,\mathcal{F}) \asymp n^{-\frac{2s-2\gamma}{2s+1-2\gamma}} \quad \gamma := \frac{1}{p} - \frac{1}{p \vee 2}$$

2. *Optimality of the wavelet shrinkage estimator: If* $p \le q$*, there exists a wavelet estimator with a proper thresholding* $\widehat{f}$ *such that*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{D_n} \left\| \widehat{f} - f \right\|^2_{L^2(P_x)} \lesssim \mathcal{R}(n,\mathcal{F})(1 + o(1)).$$

Preliminaries: Minimax (lower)-bound techniques

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $\boxed{r_n \asymp n^{-\frac{s}{2s+1}}}$

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\Longrightarrow \exists f_1, \ldots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \ldots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.

TEXAS A&M UNIVERSITY

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\Longrightarrow \exists f_1, \dots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \dots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.
2. Step 1 and Chebyshev's inequality imply:

$$\inf_{\widehat{f}} \max_{j=1,\dots,N} \mathbb{P}_{f_j}\left[\left\|\widehat{f} - f_j\right\| \geq r_n\right] \leq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{P}_f\left[\left\|\widehat{f} - f\right\| \geq r_n\right] \leq \frac{1}{r_n^2} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left\|\widehat{f} - f\right\|^2$$

Notation: probability measure $\mathbb{P}_g$ is relative to the Gaussian distribution $\{g(t_i) + \epsilon_i : i = 1, \dots, n\}$.

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\Longrightarrow \exists f_1, \ldots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \ldots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.
2. Step 1 and Chebyshev's inequality imply:
$$\inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}\left[\left\|\widehat{f} - f_j\right\| \geq r_n\right] \leq \inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{P}_f\left[\left\|\widehat{f} - f\right\| \geq r_n\right] \leq \frac{1}{r_n^2} \inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}_f \left\|\widehat{f} - f\right\|^2$$
Notation: probability measure $\mathbb{P}_g$ is relative to the Gaussian distribution $\{g(t_i) + \epsilon_i : i = 1, \ldots, n\}$.
3. Set $j_* = \operatorname{argmin}_j \|f_j - \hat{f}\|$. Then
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\left\|\widehat{f} - f_{j_*}\right\| \leq \left\|\widehat{f} - f_j\right\|\right)$$

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\Longrightarrow \exists f_1, \dots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \dots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.
2. Step 1 and Chebyshev's inequality imply:
$$\inf_{\widehat{f}} \max_{j=1,\dots,N} \mathbb{P}_{f_j}\left[\left\|\widehat{f} - f_j\right\| \geq r_n\right] \leq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{P}_f\left[\left\|\widehat{f} - f\right\| \geq r_n\right] \leq \frac{1}{r_n^2} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left\|\widehat{f} - f\right\|^2$$
Notation: probability measure $\mathbb{P}_g$ is relative to the Gaussian distribution $\{g(t_i) + \epsilon_i : i = 1, \dots, n\}$.
3. Set $j_* = \text{argmin}_j \|f_j - \hat{f}\|$. Then
$$\forall j = 1, \dots, N: \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\left\|\widehat{f} - f_{j_*}\right\| \leq \left\|\widehat{f} - f_j\right\|\right)$$
4. By separation hypothesis and triangle inequality on the preceeding event,
$$\forall j \neq j_*: \quad r_n \leq \|\hat{f} - f_j\|.$$
Conclude that
$$\forall j = 1, \dots, N: \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\|\hat{f} - f_j\| \geq r_n\right)$$

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\Longrightarrow \exists f_1, \ldots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \ldots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.
2. Step 1 and Chebyshev's inequality imply:
$$\inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}\left[\left\|\widehat{f} - f_j\right\| \geq r_n\right] \leq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{P}_f\left[\left\|\widehat{f} - f\right\| \geq r_n\right] \leq \frac{1}{r_n^2} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left\|\widehat{f} - f\right\|^2$$
Notation: probability measure $\mathbb{P}_g$ is relative to the Gaussian distribution $\{g(t_i) + \epsilon_i : i = 1, \ldots, n\}$.
3. Set $j_* = \operatorname{argmin}_j \|f_j - \hat{f}\|$. Then
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\left\|\widehat{f} - f_{j_*}\right\| \leq \left\|\widehat{f} - f_j\right\|\right)$$
4. By separation hypothesis and triangle inequality on the preceeding event,
$$\forall j \neq j_* : \quad r_n \leq \|\hat{f} - f_j\|.$$
Conclude that
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\|\hat{f} - f_j\| \geq r_n\right)$$
5. Step 2 $\Longrightarrow \inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}(j_* \neq j) \leq \frac{1}{r_n^2}\mathcal{R}(n, \mathcal{F})$

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\implies \exists f_1, \ldots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \ldots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.
2. Step 1 and Chebyshev's inequality imply:
$$\inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}\left[\left\|\widehat{f} - f_j\right\| \geq r_n\right] \leq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{P}_f\left[\left\|\widehat{f} - f\right\| \geq r_n\right] \leq \frac{1}{r_n^2} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f\left\|\widehat{f} - f\right\|^2$$
Notation: probability measure $\mathbb{P}_g$ is relative to the Gaussian distribution $\{g(t_i) + \epsilon_i : i = 1, \ldots, n\}$.
3. Set $j_* = \operatorname{argmin}_j \|f_j - \hat{f}\|$. Then
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\left\|\widehat{f} - f_{j_*}\right\| \leq \left\|\widehat{f} - f_j\right\|\right)$$
4. By separation hypothesis and triangle inequality on the preceeding event,
$$\forall j \neq j_* : \quad r_n \leq \|\hat{f} - f_j\|.$$
Conclude that
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\|\hat{f} - f_j\| \geq r_n\right)$$
5. Step 2 $\implies \inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}(j_* \neq j) \leq \frac{1}{r_n^2}\mathcal{R}(n, \mathcal{F})$
6. **Information-Theoretic Lower Bound via Kullback-Leibler distance** [Theorem 6.3.2, GN]: if there exists $C > 0$ such that
$$\sum_{j=1}^{N} \mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{f_j} \| \mathbb{P}_{f_1}\right) \leq CN \log N$$
then $\inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}(\widehat{j} \neq j)$ can be lower-bounded by C.

General reduction principle via multiple hypothesis testing [Giné-Nickl] Set $r_n \asymp n^{-\frac{s}{2s+1}}$

1. $\mathcal{F}$ compact $\Longrightarrow \exists f_1, \ldots, f_N \in \mathcal{F}$ such that $\{B(f_j, r_n) : j = 1, \ldots, N\}$ covers $\mathcal{F}$ and **separation hypothesis** holds: $\|f_j - f_{j'}\| \geq 2r_n$ for $\forall j \neq j'$.
2. Step 1 and Chebyshev's inequality imply:
$$\inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}\left[\left\|\widehat{f} - f_j\right\| \geq r_n\right] \leq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{P}_f\left[\left\|\widehat{f} - f\right\| \geq r_n\right] \leq \frac{1}{r_n^2} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left\|\widehat{f} - f\right\|^2$$
Notation: probability measure $\mathbb{P}_g$ is relative to the Gaussian distribution $\{g(t_i) + \epsilon_i : i = 1, \ldots, n\}$.
3. Set $j_* = \mathrm{argmin}_j \|f_j - \hat{f}\|$. Then
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\left\|\widehat{f} - f_{j_*}\right\| \leq \left\|\widehat{f} - f_j\right\|\right)$$
4. By separation hypothesis and triangle inequality on the preceeding event,
$$\forall j \neq j_* : \quad r_n \leq \|\hat{f} - f_j\|.$$
Conclude that
$$\forall j = 1, \ldots, N : \quad \mathbb{P}_{f_j}(j_* \neq j) \leq \mathbb{P}_{f_j}\left(\|\hat{f} - f_j\| \geq r_n\right)$$
5. Step 2 $\Longrightarrow \inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}(j_* \neq j) \leq \frac{1}{r_n^2}\mathcal{R}(n, \mathcal{F})$
6. **Information-Theoretic Lower Bound via Kullback-Leibler distance** [Theorem 6.3.2, GN]: if there exists $C > 0$ such that
$$\sum_{j=1}^{N} \mathrm{D_{KL}}\left(\mathbb{P}_{f_j} \| \mathbb{P}_{f_1}\right) \leq CN \log N$$
then $\inf_{\widehat{f}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}(\widehat{j} \neq j)$ can be lower-bounded by C.
7. Conclusion: if separation hypothesis and KL bound holds, we can conclude that $\mathcal{R}(n, \mathcal{F}) \gtrsim r_n^2$

The proof now reduces to verifying the **separation hypothesis** and the **KL bound**. Fix $N \in \mathbb{N}$ for the number of r_n-balls to cover $\mathcal{F}$.

The proof now reduces to verifying the **separation hypothesis** and the **KL bound**. Fix $N \in \mathbb{N}$ for the number of r_n-balls to cover $\mathcal{F}$.

1. Consider the S-regular Daubechies wavelets. At the j-th level, there are $c_0 2^j$ wavelet functions $\{\psi_{jk} : k = 1, \ldots, c_0 2^j\}$.

The proof now reduces to verifying the **separation hypothesis** and the **KL bound**. Fix $N \in \mathbb{N}$ for the number of r_n-balls to cover $\mathcal{F}$.

1. Consider the S-regular Daubechies wavelets. At the j-th level, there are $c_0 2^j$ wavelet functions $\{\psi_{jk} : k = 1, \ldots, c_0 2^j\}$.
2. Let $\beta_{\mathbf{m}} = (\beta_{mk}) \in \{-1, 1\}^{c_0 2^j}$ (Hamming cube), and set

$$f_0 = 0, \quad f_{\mathbf{m}}(x) = 2^{-j(s+\frac{1}{2})} \sum_{k=1}^{c_0 2^j} \beta_{mk} \psi_{jk}(x).$$

Note that $\|f_{\mathbf{m}}\|_{B^s_{p,q}} \leq 1$ for all $\mathbf{m} = 1, \ldots, c_0 2^j$.

The proof now reduces to verifying the **separation hypothesis** and the **KL bound**. Fix $N \in \mathbb{N}$ for the number of r_n-balls to cover $\mathcal{F}$.

1. Consider the S-regular Daubechies wavelets. At the j-th level, there are $c_0 2^j$ wavelet functions $\{\psi_{jk} : k = 1, \dots, c_0 2^j\}$.
2. Let $\beta_{\mathbf{m}} = (\beta_{mk}) \in \{-1, 1\}^{c_0 2^j}$ (Hamming cube), and set
$$f_0 = 0, \quad f_{\mathbf{m}}(x) = 2^{-j(s+\frac{1}{2})} \sum_{k=1}^{c_0 2^j} \beta_{mk} \psi_{jk}(x).$$
Note that $\|f_{\mathbf{m}}\|_{B^s_{p,q}} \leq 1$ for all $\mathbf{m} = 1, \dots, c_0 2^j$.
3. Parseval's identity:
$$\forall \beta_{\mathbf{m}} \neq \beta_{\mathbf{m}'} \in \{-1, 1\}^{2^j} : \quad \|f_{\mathbf{m}} - f_{\mathbf{m}'}\|^2_{L^2[0,1]} = 2^{-j(2s+1)} \sum_{k=1}^{2^j} (\beta_{\mathbf{m}k} - \beta_{\mathbf{m}'k})^2$$

The proof now reduces to verifying the **separation hypothesis** and the **KL bound**. Fix $N \in \mathbb{N}$ for the number of r_n-balls to cover $\mathcal{F}$.

1. Consider the S-regular Daubechies wavelets. At the j-th level, there are $c_0 2^j$ wavelet functions $\{\psi_{jk} : k = 1, \dots, c_0 2^j\}$.
2. Let $\beta_{\mathbf{m}} = (\beta_{mk}) \in \{-1, 1\}^{c_0 2^j}$ (Hamming cube), and set
$$f_0 = 0, \quad f_{\mathbf{m}}(x) = 2^{-j(s+\frac{1}{2})} \sum_{k=1}^{c_0 2^j} \beta_{mk} \psi_{jk}(x).$$
Note that $\|f_{\mathbf{m}}\|_{B^s_{p,q}} \leq 1$ for all $\mathbf{m} = 1, \dots, c_0 2^j$.
3. Parseval's identity:
$$\forall \beta_{\mathbf{m}} \neq \beta_{\mathbf{m}'} \in \{-1, 1\}^{2^j} : \quad \|f_{\mathbf{m}} - f_{\mathbf{m}'}\|^2_{L^2[0,1]} = 2^{-j(2s+1)} \sum_{k=1}^{2^j} (\beta_{\mathbf{m}k} - \beta_{\mathbf{m}'k})^2$$
4. Using coding theory (**Gilbert-Shannon-Varshamov bound**), $\exists c_1, c_2 > 0$ and subset $\mathcal{M} \subset \{-1, 1\}^{c_0 2^j}$ such that $\#\mathcal{M} = 2^{c_1 2^j}$ and
$$\forall \mathbf{m} \neq \mathbf{m}' \in \mathcal{M} : \quad \sum_{\mathbf{m}} |\beta_{\mathbf{m}} - \beta_{\mathbf{m}'}|^2 \geq c_2 2^{j+2}.$$

TEXAS A&M UNIVERSITY

The proof now reduces to verifying the **separation hypothesis** and the **KL bound**. Fix $N \in \mathbb{N}$ for the number of r_n-balls to cover $\mathcal{F}$.

1. Consider the S-regular Daubechies wavelets. At the j-th level, there are $c_0 2^j$ wavelet functions $\{\psi_{jk} : k = 1, \dots, c_0 2^j\}$.
2. Let $\beta_{\mathbf{m}} = (\beta_{mk}) \in \{-1, 1\}^{c_0 2^j}$ (Hamming cube), and set
$$f_0 = 0, \quad f_{\mathbf{m}}(x) = 2^{-j(s+\frac{1}{2})} \sum_{k=1}^{c_0 2^j} \beta_{mk} \psi_{jk}(x).$$
Note that $\|f_{\mathbf{m}}\|_{B^s_{p,q}} \leq 1$ for all $\mathbf{m} = 1, \dots, c_0 2^j$.
3. Parseval's identity:
$$\forall \beta_{\mathbf{m}} \neq \beta_{\mathbf{m}'} \in \{-1, 1\}^{2^j} : \quad \|f_{\mathbf{m}} - f_{\mathbf{m}'}\|^2_{L^2[0,1]} = 2^{-j(2s+1)} \sum_{k=1}^{2^j} (\beta_{\mathbf{m}k} - \beta_{\mathbf{m}'k})^2$$
4. Using coding theory (**Gilbert-Shannon-Varshamov bound**), $\exists c_1, c_2 > 0$ and subset $\mathcal{M} \subset \{-1, 1\}^{c_0 2^j}$ such that $\#\mathcal{M} = 2^{c_1 2^j}$ and
$$\forall \mathbf{m} \neq \mathbf{m}' \in \mathcal{M} : \quad \sum_{\mathbf{m}} |\beta_{\mathbf{m}} - \beta_{\mathbf{m}'}|^2 \geq c_2 2^{j+2}.$$
5. For large enough $n \in \mathbb{N}$, choose $\boxed{j = \frac{\log(n)}{2s+1}}$ and $N \leq \#\mathcal{M} \leq N^N$. Hence separation hypothesis holds:
$\|f_{\mathbf{m}} - f_{\mathbf{m}'}\|_{L^2} \gtrsim 2r_n$

KL bound: Since $\mathbb{P}_f$ is drawn from i.i.d. Gaussian samples $\{(x_i, f(x_i)+\eta_i) : i = 1,\dots,n\}$, it tensorizes and gives (via Radon-Nikodym)

$$\begin{aligned} \mathrm{D_{KL}}\left(\mathbb{P}_{f_{\mathbf{m}}} \| \mathbb{P}_{f_0}\right) &= n\left(\underbrace{\mathrm{D_{KL}}\left(\mathbb{P}_x \| \mathbb{P}_x\right)}_{=0} + \mathbb{E}_{x\sim\mathbb{P}_x}\mathrm{D_{KL}}\left(f_{\mathbf{m}}(x)+\eta \,\|\, f_0(x)+\eta\right)\right) \\ &= \frac{n}{2\sigma^2}\mathbb{E}_{x\sim\mathbb{P}_x}\|f_{\mathbf{m}}(x)-f_0(x)\|_2^2 = \frac{n}{2\sigma^2}\|f_{\mathbf{m}}-f_0\|^2_{L^2(\mathbb{P}_x)} = \frac{n}{2\sigma^2}\|f_{\mathbf{m}}\|^2_{L^2(\mathbb{P}_x)}. \end{aligned}$$

where $\mathbb{P}_x$ is Lebesgue (uniform) measure on $[0,1]$.

By our wavelet construction,

$$\frac{n}{2\sigma^2}\|f_{\mathbf{m}}\|^2_{L^2} \le \frac{n}{2\sigma^2}\cdot 2^{-j(2s+1)}\|\beta_{\mathbf{m}}\|^2 \le \log\#\mathcal{M} \le N\log N$$

Thus,

$$\mathrm{D_{KL}}\left(\mathbb{P}_{f_{\mathbf{m}}} \| \mathbb{P}_{f_0}\right) \le N\log N$$

Conclusion:

$$\mathcal{R}(n,\mathcal{F}) \gtrsim n^{-\frac{2s}{2s+1}} \tag{24}$$

Assuming $\mathcal{R}(n,\mathcal{F}) \asymp n^{-\frac{2s}{2s+1}}$, we prove that $\mathcal{R}_L(n,\mathcal{F}) \asymp n^{-\frac{2s-2\gamma}{2s+1-2\gamma}}$ where $\gamma = \frac{1}{p} - \frac{1}{p\vee 2}$.
By wavelet theory, there is a correspondence between Besov functions and wavelet coefficients:

$$\mathcal{R}_*(n,F) \simeq \tilde{\mathcal{R}}_*(\tfrac{\sigma}{\sqrt{n}}, \Theta^s_{p,q}) := \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}} \mathbb{E}\|\hat{\theta}-\theta\|_2^2$$

Assuming $\mathcal{R}(n,\mathcal{F}) \asymp n^{-\frac{2s}{2s+1}}$, we prove that $\mathcal{R}_L(n,\mathcal{F}) \asymp n^{-\frac{2s-2\gamma}{2s+1-2\gamma}}$ where $\gamma = \frac{1}{p} - \frac{1}{p\vee 2}$.
By wavelet theory, there is a correspondence between Besov functions and wavelet coefficients:

$$\mathcal{R}_*(n,F) \simeq \tilde{\mathcal{R}}_*\left(\frac{\sigma}{\sqrt{n}}, \Theta^s_{p,q}\right) := \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}} \mathbb{E}\|\hat{\theta}-\theta\|_2^2$$

1. Quadratic hull:

$$\mathcal{Q}\mathrm{Hull}(\Theta) = \{\theta : \theta^2 \in \mathrm{Hull}(\Theta^2_+)\}$$

Assuming $\mathcal{R}(n,\mathcal{F}) \asymp n^{-\frac{2s}{2s+1}}$, we prove that $\mathcal{R}_L(n,\mathcal{F}) \asymp n^{-\frac{2s-2\gamma}{2s+1-2\gamma}}$ where $\gamma = \frac{1}{p} - \frac{1}{p\vee 2}$.
By wavelet theory, there is a correspondence between Besov functions and wavelet coefficients:

$$\mathcal{R}_*(n,F) \simeq \tilde{\mathcal{R}}_*(\tfrac{\sigma}{\sqrt{n}}, \Theta^s_{p,q}) := \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}} \mathbb{E}\|\hat{\theta}-\theta\|_2^2$$

1. Quadratic hull:

$$\mathcal{Q}\mathrm{Hull}(\Theta) = \{\theta : \theta^2 \in \mathrm{Hull}(\Theta^2_+)\}$$

2. [Donoho-Liu-MacGibbon, Annals-Stat '90] showed that

$$\mathcal{Q}\mathrm{Hull}(\Theta^s_{p,q}) = \Theta^{s-\gamma}_{p\vee 2, q\vee 2}$$

and

$$\tilde{\mathcal{R}}_L(\varepsilon,\Theta) = \tilde{\mathcal{R}}_L(\varepsilon,\mathcal{Q}\mathrm{Hull}(\Theta)), \quad \text{and} \quad \tilde{\mathcal{R}}_L(\varepsilon,\mathcal{Q}\mathrm{Hull}(\Theta)) \simeq \tilde{\mathcal{R}}(\varepsilon,\mathcal{Q}\mathrm{Hull}(\Theta))$$

Proof of linear minimax rate

Assuming $\mathcal{R}(n,\mathcal{F}) \asymp n^{-\frac{2s}{2s+1}}$, we prove that $\mathcal{R}_L(n,\mathcal{F}) \asymp n^{-\frac{2s-2\gamma}{2s+1-2\gamma}}$ where $\gamma = \frac{1}{p} - \frac{1}{p\vee 2}$.
By wavelet theory, there is a correspondence between Besov functions and wavelet coefficients:

$$\mathcal{R}_*(n,F) \simeq \tilde{\mathcal{R}}_*(\tfrac{\sigma}{\sqrt{n}}, \Theta^s_{p,q}) := \inf_{\hat\theta} \sup_{\Theta^s_{p,q}} \mathbb{E}\|\hat\theta - \theta\|_2^2$$

1. Quadratic hull:

$$\mathcal{Q}\mathrm{Hull}(\Theta) = \{\theta : \theta^2 \in \mathrm{Hull}(\Theta^2_+)\}$$

2. [Donoho-Liu-MacGibbon, Annals-Stat '90] showed that

$$\mathcal{Q}\mathrm{Hull}(\Theta^s_{p,q}) = \Theta^{s-\gamma}_{p\vee 2, q\vee 2}$$

and

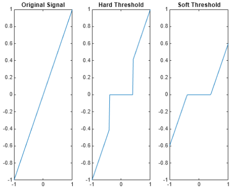
$$\tilde{\mathcal{R}}_L(\varepsilon,\Theta) = \tilde{\mathcal{R}}_L(\varepsilon,\mathcal{Q}\mathrm{Hull}(\Theta)), \quad \text{and} \quad \tilde{\mathcal{R}}_L(\varepsilon,\mathcal{Q}\mathrm{Hull}(\Theta)) \simeq \tilde{\mathcal{R}}(\varepsilon,\mathcal{Q}\mathrm{Hull}(\Theta))$$

3. Thus, for $p \le q < 2$, we have the (suboptimal) linear rate

$$\mathcal{R}_L(n,\mathcal{F}) \simeq \tilde{\mathcal{R}}_L(\varepsilon_n, \Theta^s_{p,q}) = \tilde{\mathcal{R}}_L(\varepsilon_n, \mathcal{Q}\mathrm{Hull}(\Theta^s_{p,q})) = \tilde{\mathcal{R}}_L(\varepsilon_n, \Theta^{s-\gamma}_{2,2}) \asymp \tilde{\mathcal{R}}(\varepsilon_n, \Theta^{s-\gamma}_{2,2}) \asymp n^{-\frac{2s-2\gamma}{2s+1-2\gamma}}$$

along the sequence $\varepsilon_n = \frac{\sigma}{\sqrt{n}}$

- Construction of the estimator:
  1. Apply discrete wavelet transform on y_i with suitable wavelet (e.g., Daubechies, Meyer), *i.e.*, $\theta_j = W_j Y$; W_j: an orthogonal matrix corresponding to the discrete wavelet transform operator at jth level. Obtain θ_j up to $j = -1, \ldots, (\log n - 1)$th level.
  2. Apply thresholding to θ_{jk} with the certain threshold λ; denote $\delta_\lambda(\theta_{jk})$. See Figure 2.
     - Hard thresholding:
     $$\delta_\lambda(z) = z\mathbb{1}_{\{|z| \geq \lambda\}}.$$
     - Soft thresholding:
     $$\delta_\lambda(z) = \mathrm{sgn}(z)\sigma(|z| - \lambda).$$
  3. Apply inverse wavelet transform using $\delta_\lambda(\theta_{jk})$, *i.e.*,
  $$\widehat{f}^\lambda(x) = \sum_{j=-1}^{\log n - 1} \sum_{k=1}^{2^j} \delta_\lambda(\theta_{jk})\psi_{jk}(x).$$
  for $\lambda = (\lambda_{jk})_{j,k}$.
- Thresholding works like 'denoising'.



- The proof consists of the following steps:
  1. Instead of the original problem *i.e.*, estimating f^* given D_n, consider a 'Gaussian white noise model' in a sequence space; $\widehat{f}^\lambda$ has a counterpart $\widehat{\theta}^\lambda$ in the Gaussian white noise model.
  2. Show $\exists\lambda^*$ s.t. $\widehat{\theta}^{\lambda^*}$: minimax optimal for the Gaussian white noise model.
     - Here, use equivalence between minimax risk and minimax Bayes risk:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta^s_{p,q}} \left\| \widehat{\theta} - \theta \right\|^2 = \mathcal{R}(\epsilon, \Theta^s_{p,q}) \asymp \mathcal{B}(\epsilon, \Theta^s_{p,q}) := \inf_{\widehat{\theta}} \sup_{\mu \in \mathcal{P}(\Theta^s_{p,q})} \mathbb{E}_{\theta\sim\mu} \left\| \widehat{\theta} - \theta \right\|^2,$$

$$\mathcal{R}_T(\epsilon, \Theta^s_{p,q}) \asymp \mathcal{B}_T(\epsilon, \Theta^s_{p,q}) := \inf_{\lambda} \sup_{\mu \in \mathcal{P}(\Theta^s_{p,q})} \mathbb{E}_{\theta\sim\mu} \left\| \widehat{\theta}^\lambda - \theta \right\|^2.$$

     - In fact, Bayes risk $\leq$ minimax risk always holds ($\because$ mean $\leq$ max).
     - $\asymp$ holds for this problem $\because$ $\Theta^s_{p,q}$ is a convex set w.r.t. measures and ℓ^2 loss is lower semi-continuous.
  3. Show Gaussian white noise model $\approx$ the original problem.
- $\Rightarrow \mathcal{R}_T(n, \mathcal{F}) \lesssim \mathcal{R}_T(\epsilon, \Theta^s_{p,q}) \lesssim \mathcal{R}(\epsilon, \Theta^s_{p,q}) \asymp \mathcal{R}(n, \mathcal{F})$ when $\epsilon = \sigma/\sqrt{n}$.

- Gaussian white noise model in sequence space:

$$y_I = \theta_I + z_I$$
$$I \in \mathcal{I} = \cup_{j \geq -1} \mathcal{I}_j, \qquad \mathcal{I}_j = \left\{ I_{jk} = \left[\frac{k-1}{2^j}, \frac{k}{2^j} \right] \right\}_{k=1,\dots,2^j}$$
$$z_I \overset{i.i.d}{\sim} N(0, \epsilon^2).$$

- $\widehat{\theta}_I^{\lambda} = \delta_{\lambda_I}(y_I)$ in the sequence model corresponds to $\widehat{f}^{\lambda}$ in the previous slides.
- We first show

$$\mathcal{R}_T(\epsilon, \Theta^s_{p,q}) := \min_{\lambda} \sup_{\widehat{\theta}^{\lambda} \in \Theta^s_{p,q}} \mathbb{E} \left\| \widehat{\theta}^{\lambda} - \theta \right\|_2^2 \leq C_p \mathcal{R}(\epsilon, \Theta^s_{p,q})$$

as $\epsilon \to 0$. Here, Θ is some ball of wavelet sequence w.r.t. wavelet sequence Besov norm; see (DJ98)[Equation (6)].

- Proof for the sequence model: $\mathcal{R}_T(\epsilon, \Theta^s_{p,q}) \lesssim \mathcal{R}(\epsilon, \Theta^s_{p,q})$
  1. There exists a prior distribution $\mu^* \in \mathcal{P}(\Theta^s_{p,q})$ such that $\mathcal{R}_T(\epsilon, \Theta)$ is equal to a Bayes risk w.r.t. $\mu^* = (\mu^*_I)_{I \in \mathcal{I}}$, *i.e.*,

$$\mathcal{R}_T(\epsilon, \Theta^s_{p,q}) \asymp \mathcal{B}_T(\epsilon, \Theta^s_{p,q}) = \sum_I \max_{\mu_I \in \mathcal{P}(\Theta)} \underbrace{\inf_{\lambda_I} \mathbb{E}_{\theta_I \sim \mu_I} \left|\delta_{\lambda_I}(y_I) - \theta_I\right|^2}_{:=\rho(\mu_I)} = \sum_I \rho(\mu^*_I).$$

     This is proven by showing the target functional has a saddle point (λ^*_I, μ^*_I).
  2. $\Theta^s_{p,q}$ being some ball implies $\mu \in \mathcal{P}(\Theta^s_{p,q})$ has a finite pth moment, *i.e.*, $\tau = (\tau_I)_{I \in \mathcal{I}}$ has a finite Besov norm, where $\tau_I = \mathbb{E}\left|\theta_I\right|_p$. This implies $\rho(\mu^*_I) \le \inf_{\lambda_I} \sup_{\mathbb{E}_{\mu_I}\left|\theta_I\right|_p \le \tau_I} \mathbb{E}_{\theta_I} \left|\delta_{\lambda_I}(y) - \theta_I\right|^2$.
  3. (DJ94) showed for some $C_p > 0$, the minima λ^*_I satisfies

$$\sup_{\mathbb{E}_{\mu_I}\left|\theta_I\right|_p \le \tau_I} \mathbb{E}_{\theta_I} \left|\delta_{\lambda^*_I}(y_I) - \theta_I\right|^2 \le C_p \sup_{\mathbb{E}_{\mu_I}\left|\theta_I\right|_p \le \tau_I} \mathbb{E}_{\theta_I \sim \mu_I} \left|y_I - \theta_I\right|^2.$$

  4. Lastly, minimax risk upper bounds Bayes risk, *i.e.*, the $\sum_I$ RHS $\le C_p \mathcal{R}(\epsilon, \Theta^s_{p,q})$.
- $\therefore \mathcal{R}_T(\epsilon, \Theta^s_{p,q}) \le C_p \mathcal{R}(\epsilon, \Theta^s_{p,q})$.

- Asymptotic equivalence between function estimation problem and sequence model: Our final step is to show
$$\mathcal{R}_T(n,\mathcal{F}) \lesssim \mathcal{R}_T(\sigma/\sqrt{n},\Theta^s_{p,q}).$$
In fact, $\asymp$ holds, but we omit $\gtrsim$ part.
  1. Given $x_i = i/n$, there exists a smooth interpolation of $f(x_i)$ called Deslauriers-Dubuc interpolant $\widetilde{f} : [0,1] \to \mathbb{R}$. Such interpolant satisfies
$$\sup_{f^*\in\mathcal{F}} \mathbb{E}_{D_n} \left\|\widehat{f} - f^*\right\|^2_{L^2([0,1])} \approx \sup_{f\in\mathcal{F}} \mathbb{E} \left\|\widehat{f} - \widetilde{f}\right\|^2_{L^2([0,1])}$$
as $n\to\infty$ (P_x being uniform is used here).
  2. Isometry property of the wavelet gives $\left\|\widehat{f} - \widetilde{f}\right\|^2_{L^2} = \left\|\widehat{\theta}^\lambda - \widetilde{\theta}\right\|^2_2$.
  3. $\widetilde{\theta}$ is the Gaussian white noise model of
$$\widetilde{y}_I = \widetilde{\theta}_I + \epsilon_n \widetilde{z}_I, \quad I \in \cup_{j=-1}^{\log n-1}\mathcal{I}_j.$$
The specific choice of the interpolation and the optimal $\lambda=\lambda^*$ induces $\epsilon_n = C\sigma/\sqrt{n}$ and $\widetilde{\theta}_I \in C\Theta^s_{p,q}$ for some (possibly different) $C>0$.
  4. $\therefore \sup_{\theta\in\Theta^s_{p,q}} \mathbb{E}\left\|\widehat{\theta}^{\lambda^*} - \widetilde{\widetilde{\theta}}\right\|^2_2 = \mathcal{R}_T(\epsilon_n, C\Theta^s_{p,q}) \asymp \mathcal{R}_T(\sigma/\sqrt{n},\Theta^s_{p,q}) \lesssim \mathcal{R}(\sigma/\sqrt{n},\Theta^s_{p,q}) \asymp \mathcal{R}(n,\mathcal{F})$.
Here, the optimality of λ^* and the scaling property of $\mathcal{R}$ was used.
- In conclusion, there exists λ^* s.t. $\sup_{f^*}\mathbb{E}\left\|\widehat{f}^{\lambda^*} - f^*\right\|^2 \lesssim \mathcal{R}(n,\mathcal{F})$.

TEXAS A&M UNIVERSITY

- To finalize the result, it is sufficient to obtain $\mathcal{R}(\epsilon, \Theta^s_{p,q})$'s upper bound.
  1. Again, recall $\mathcal{R}(\epsilon, \Theta^s_{p,q}) \asymp \mathcal{B}(\epsilon, \Theta^s_{p,q}) = \inf_{\widehat{\theta}} \sup_{\mu \in \mathcal{P}(\Theta^s_{p,q})} \mathbb{E}_{\theta \sim \mu} \left\| \widehat{\theta} - \theta \right\|_2^2$.
  2. Observe this structure can be decomposed, *i.e.*,

$$\inf_{\widehat{\theta}} \sup_{\mu \in \mathcal{P}(\Theta^s_{p,q})} \mathbb{E}_{\theta \sim \mu} \mathbb{E}_{y_I} \left\| \widehat{\theta} - \theta \right\|_2^2 \leq \sum_I \inf_{\widehat{\theta}} \sup_{\mu_I} \mathbb{E}_{\theta_I \sim \mu_I} \left| \widehat{\theta}_I - \theta_I \right|^2$$

$$\leq \sum_I \underbrace{\inf_{\widehat{\theta}} \sup_{\mu_I \text{ s.t. } p\text{th moment of } \mu_I \text{ is } \tau_I \in \Theta^s_{p,q}} \mathbb{E}_{\theta_I \sim \mu_I} \mathbb{E}_{y_I} \left| \widehat{\theta}_I - \theta_I \right|^2}_{:= \rho(\tau_I, \epsilon)}.$$

  3. The above problem becomes solving the constraint optimization min $\rho(t, \epsilon)$ given $|t|_{b^s_{p,q}} \leq C$. One can solve such optimziation using the calculation rule for $\rho(t, \epsilon)$ for this specific Gaussian white noise model (DJ94).
  4. Solving the optimization problem induces $\mathcal{B}(\epsilon, \Theta^s_{p,q}) \lesssim \epsilon^{\frac{4s}{2s+1}}$.
- $\therefore n^{-\frac{2s}{2s+1}} \lesssim \mathcal{R}(n, \mathcal{F}) \lesssim \mathcal{R}(\sigma/\sqrt{n}, \Theta^s_{p,q}) \lesssim n^{-\frac{2s}{2s+1}}$.
- Remark: This is 'not' a typical strategy in Statistics. Instead, one directly calculates the upper bound of the certain estimator and show it matches to the lower bound; for the wavelet threshold estimator, e.g., (GN15)[Proposition 5.1.7].

- $n^{-\frac{2s}{2s+1}} \overset{(1)}{\lesssim} \mathcal{R}(n,\mathcal{F}) \overset{(2)}{\lesssim} \mathcal{R}_T(n,\mathcal{F}) \overset{(3)}{\lesssim} \mathcal{R}_T(\sigma/\sqrt{n},\Theta^s_{p,q}) \overset{(4)}{\lesssim} \mathcal{R}(\sigma/\sqrt{n},\Theta^s_{p,q}) \overset{(5)}{\lesssim} n^{-\frac{2s}{2s+1}}$.
- (1): the lower bound proof.
- (2): the definition of the minimax rate.
- (3): the equivalence between estimation and Gaussian white noise model.
- (4): optimality of the wavelet estimator.
- (5): upper bound analysis.
- $\therefore \mathcal{R}(n,\mathcal{F}) \asymp \mathcal{R}_T(n,\mathcal{F}) \asymp n^{-\frac{2s}{2s+1}}$

Theorem

If we consider $\Omega = [0,1]^d$ instead of $[0,1]$, the minimax rate for $B^s_{p,q}(\Omega)$ with $s > d/p$ is

$$\mathcal{R}(n, \mathcal{F}) \gtrsim n^{-\frac{2s}{2s+d}}.$$

- The proof goes the same as earlier minimax proof; consider $f_m(x) = \epsilon 2^{-j(s+1/2)} \sum_k \beta_{mk} \psi_{jk}(x)$, where ψ_{jm} forming a wavelet basis of $L^2(\Omega)$. In this case, at jth resolution level there are now $C2^{jd}$ wavelet coefficients. The rest goes the same.
- The minimax optimality of the wavelet threshold estimator can be analyzed in $d > 1$ as well, but practically such estimator is not desirable as the number of wavelet coefficients grows exponentially w.r.t d.
- $\therefore$ we want to consider the alternative estimator.

Estimation error with DNN estimators.

- Let $\Phi(L, W, S, B)$ be a L-layer W-width ReLU Deep neural network with the following structure:

$$\Phi(L, W, S, B)(x) = \left[\left(W^{(L)}(\cdot) + b^{(L)}\right) \circ \sigma \cdots \circ \sigma \circ \left(W^{(1)}(\cdot) + b^{(1)}\right)\right](x). \tag{25}$$

  - L: Neural network depth.
  - W: Neural network width, *i.e.*, $W^{(l)} \in \mathbb{R}^{W \times W}$, $b^{(l)} \in \mathbb{R}^{W}$ for all $l = 1, \ldots, L$.
  - S: Sparsity parameter, *i.e.*, $\sum_{l=1}^{L} \left[\left\|W^{(l)}\right\|_0 + \left\|b^{(l)}\right\|_0\right] \leq S$.
  - B: Norm constraint, *i.e.*, $\max_{l=1,\ldots,L} \left[\left\|W^{(l)}\right\|_\infty, \left\|b^{(l)}\right\|_\infty\right] \leq B$.
  - σ: ReLU activation.

- Consider the problem of estimating $f^* \in B_{B^s_{p,q}(\Omega)}(0,1) \cap B_{L^\infty(\Omega)}(0,F)$ for some $F > 0$, with the data $y_i = f^*(x_i) + \eta_i$ with $\eta_i \overset{i.i.d}{\sim} N(0,\sigma^2)$ and $x_i \overset{i.i.d}{\sim} P_x$ where $supp(P_x) \subseteq \Omega = [0,1]^d$.

Theorem (DNN estimator of Besov function)

Let $\widehat{f} := \operatorname{argmin}_{h \in \Phi(L,W,S,B)} \sum_{i=1}^n |y_i - h(x_i)|^2$, *called Empirical Risk Minimizer (ERM) with* L, W, S, B *that depends on* n, s, d, p. *For all* $f^* \in B_{B^s_{p,q}(\Omega)}(0,1) \cap B_{L^\infty(\Omega)}(0,F)$ *with some* $F > 0$,

$$\mathbb{E}_{D_n} \left\| f^* - \widehat{f} \right\|^2_{L^2(P_x)} \lesssim n^{-\frac{2s}{2s+d}} (\log n)^3.$$

- The proof consists of two ingredients:
  - Approximation of a Besov function f^* by some DNN $\widetilde{f}$. $\widetilde{f}$ may depends on f^*
  - Statistical learning theory to control the error between $\widehat{f}$ and the best approximator $\widetilde{f}$ for any choice of $\widetilde{f}$.
  - Total error $\left\| f^* - \widehat{f} \right\|$ is bounded by the above two errors: $\left\| f^* - \widetilde{f} \right\|$, $\left\| \widetilde{f} - \widehat{f} \right\|$.

- Optimal approximation error: For sufficiently large $N \in \mathbb{N}$, there exists L, W, S, B that depends on N, d, s, p s.t.

$$\sup_{f^* \in B_{B^s_{p,q}(\Omega)}(0,1)} \inf_{\widetilde{f} \in \Phi(L,W,S,B)} \left\| \widetilde{f} - f^* \right\| \lesssim N^{-\frac{s}{d}}.$$

- Basic strategy: two-stage approximation: $B^s_{p,q}(\Omega) \approx$ B-spline functions $\approx \Phi(L, W, S, B)$.
  - B-spline functions:
    - Fix m and consider

$$N_m(x_i) := \left(\underbrace{\mathbb{1}_{[0,1]} * \mathbb{1}_{[0,1]} * \cdots * \mathbb{1}_{[0,1]}}_{(m+1) \text{ times}} \right)(x_i).$$

    - $N_m(x)$ is a piecewise polynomial of the order m.
    - The following basis is called B-spline.

$$M^{m,d}_{k,j}(x) := \prod_{i=1}^{d} N_m(2^{k_i} x_i - j_i).$$

One can think of j as a location parameter and k as spatial resolution (just like a wavelet).

- $B^s_{p,q}(\Omega) \approx$ B-spline is well established in (DP88).
- B-Spline $\approx \Phi(L, W, S, B)$ is from the following observations:
  - For some $M > 0$, write $\phi_{(0,M)}(x) := \sigma(x) - \sigma(x - M) = M \wedge \sigma(x)$.
  - Observe $N_m(x)$ has the form

$$N_m(x) = \frac{1}{m!} \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} (m+1)^m \left(\phi_{(0,1-\frac{j}{m+1})} \left(\frac{x-j}{m+1} \right) \right)^m .$$

  We focus on approximating $\left(\phi_{(0,1-\frac{j}{m+1})} \left(\frac{x-j}{m+1} \right) \right)^m$. Once this is possible, approximating the linear combination is doable.
  - (Yar17) showed for some $D \in \mathbb{N}$ there exists $\psi : \mathbb{R}^D \to \mathbb{R} \in \Phi(L_1, W_1, S_1, B_1)$ for some L_1, W_1, S_1, B_1 that depends on m and ϵ such that

$$\sup_{x\in[0,M]} \left| \psi \underbrace{\left(\phi_{(0,M)} \left(\frac{x}{M} \right), \dots, \phi_{(0,M)} \left(\frac{x}{M} \right) \right)}_{m \text{ times. Write this function as } \psi \circ \phi_{(0,M)}(x/M).} - \left(\phi_{(0,M)} \left(\frac{x}{M} \right) \right)^m \right| \le \epsilon$$

  - Therefore, the reasonable construction of the approximator of $N_m(x)$ will be

$$f(x) = \frac{1}{m!} \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} (m+1)^m \left(\psi \circ \phi_{(0,1-\frac{j}{m+1})} \left(\frac{x-j}{m+1} \right) \right) .$$

  - Then, one can appropriately manipulate ψ and f to make $M^{m,d}_{0,0}(x)$.

- For any $F > 0$ and any function space $\mathcal{F} \subseteq B_{L^\infty(\Omega)}(0, F)$, there exists the following generalization gap type bound:

$$\mathbb{E}_{D_n} \left\| f^* - \widehat{f} \right\|_{L^2(P)}^2 \leq C \left(\underbrace{\inf_{f \in \mathcal{F}} \|f^* - f\|_{L^2(P)}^2}_{\approx \|f^* - \widetilde{f}\|} + \underbrace{(F^2 + \sigma^2) \frac{\log N(\mathcal{F}, \delta, \|\cdot\|_\infty)}{n} + \delta(F + \sigma)}_{\approx \|\widehat{f} - \widetilde{f}\|} \right).$$

  Proof strategy:

  1. Substitute $\widehat{f}$ to the closest δ-minimal covering of $\mathcal{F}$ and use the fact $\mathcal{F} \subseteq B_{L^\infty(\Omega)}(0, F)$ to bound the population risk by the empirical risk (Hardest part).
  2. Bound the empirical risk in terms of the optimal recovery error: By using the fact that $\widehat{f}$ is ERM.

- Set $\mathcal{F} = \Phi(L, W, S, B) \cap B_{L^\infty(\Omega)}(0, F)$, and then the covering number analysis will give the following:

$$\log N \left(\Phi(L, W, S, B), \delta, \|\cdot\|_\infty \right) \leq 2SL \log\left((B \vee 1)(W + 1)\right) + S \log\left(\frac{L}{\delta}\right).$$

  - This result is from using Lipschitz continuity of ReLU repeatedly for each layer.

- Set $\delta = 1/n$ in Step 1's RHS;

- Apply (1) the approximation result to get $\inf_{f \in \mathcal{F}} \|f^* - f\|_{L^2(P)}^2 \lesssim N^{-\frac{s}{d}}$, and (2) the covering number bound obtained in Step 2 with specific L, W, S, B in approximation result.
- Then, optimizing the RHS w.r.t. N will induce the claimed bound with $N \asymp n^{\frac{d}{2s+d}}$.

- How to actually train such neural network?
  - Solving constraint optimization.
  - Can we solve this is plain unconstraint optimization, possibly with a regularizer?
  - Constraints are used in two parts:
    1. Approximation: This is to avoid the overfitting to the noise.
    2. Learning: To bound the covering number, which controls the generalization bound.
  - It is not immediate how to avoid such constraints in approximation stage. On the other hand, there are alternative approaches to obtain a generalization bound (e.g., Rademacher complexity, VC dimensions) to avoid the constraint. Can we utilize those?
- Adaptivity
  - Constructing $\Phi(L, W, S, B)$ requires the prior knowledge on the regularity of the P^*; e.g., choices of L, W, S, B require s, p. This makes the estimation non-adaptive.

Thank You For Your Attention!

[DJ94] David L. Donoho and Iain M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994), 425–455.

[DJ98] ______, *Minimax estimation via wavelet shrinkage*, The Annals of Statistics **26** (1998), no. 3, 879 – 921.

[DP88] R. A. Devore and V. Popov, *Interpolation of besov spaces*, American Mathematical Society **305** (1988), 397 – 414.

[GN15] Evarist Giné and Richard Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2015.

[Yar17] Dmitry Yarotsky, *Error bounds for approximations with deep relu networks*, 2017.