# Minimax bounds IV

Kyoung Hee Kim

Sungshin University

1. Rate optimality

2. Examples
   - Gaussian sequence model
   - Nonparametric regression
   - Density estimation
   - Covariance matrix estimation
   - Functional estimation

3. Asymptotic efficiency

4. Asymptotic equivalence

# Minimax optimal rate

- Minimax risk with a *minimax* estimator $\hat{\theta}_{mm}$,

$$R(\Theta) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\tilde{\theta}, \theta) = \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\hat{\theta}_{mm}, \theta).$$

- Minimax optimal rate $\gamma_n$ with a *minimax rate optimal* estimator $\hat{\theta}_{op}$,

$$\text{LB: } R(\Theta) \geq c\gamma_n$$
$$\text{UB: } R(\Theta) \leq \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\hat{\theta}_{op}, \theta) \leq C\gamma_n.$$

where $C/c \leq M$.

# Estimators for a signal in a Gaussian sequence model

- Hard thresholding

$$H_t(x) = x\mathbb{1}\{|x| > t\} = \arg\min_{\mu}(x-\mu)^2 + t^2\mathbb{1}\{\mu \neq 0\}.$$

- Soft thresholding

$$S_t(x) = \mathrm{sign}(x)(|x|-t)\mathbb{1}\{|x| > t\} = \arg\min_{\mu}(x-\mu)^2 + 2t|\mu|.$$

- ...

# Estimators for a density

- Kernel density estimator
- Bayes predictive estimator (theoretical): see p.36 of III.
- Skeleton estimator (theoretical): see p.38 of IV.
- ...

# Estimators for a regression function

- Local polynomial estimator (of order $\ell$)
- Projection estimator
  Idea: Assume $f(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x)$ where $\{\psi_j\}$ is an orthonormal basis of $L_2[0,1]$. Then we approximate $f$ by its projection $\sum_{j=1}^{N} \theta_j \psi_j$ and replace $\theta_j$ by its estimators.
- ...

# Gaussian sequence models

## Example I(3), III(1)

Let $Y_i = \theta_i + \sigma\epsilon_i$ where $\epsilon_i \sim N(0,1)$ and $\sigma \to 0$ with $\theta = (\theta_1, \theta_2, \ldots) \in \Theta$.

1. When $\Theta = \left\{\theta : \sum_i i^{2s}\theta_i^2 \le M\right\}$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 \asymp \sigma^{-4s/(2s+1)}.$$

2. Let $i = 1, \ldots, n$ and $\sigma = 1/\sqrt{n}$. When $\Theta = \{\theta : \|\theta\|_0 \le 1, \|\theta\|_2 \le C\}$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 \asymp \frac{\log n}{n}.$$

# Case 1. $\Theta = \left\{\theta : \sum_i i^{2s}\theta_i^2 \le M\right\}$

- Since $\sum_i i^{2s}\theta_i^2 \le M$, $\theta_i \le Mi^{-s}$, so we estimate $\theta_1, \ldots, \theta_I$ only.
- Let $\hat{\theta}_i = Y_i$ when $i \le I$, and let $\hat{\theta}_i = 0$ when $i > I+1$.
- Then $\mathbb{E}_\theta \sum_{i=1}^{\infty}(\hat{\theta}_i - \theta_i)^2 = \mathbb{E}_\theta \sum_{i=1}^{I}(Y_i - \theta_i)^2 + \sum_{i>I+1}\theta_i^2 \le I\sigma^2 + MI^{-2s}$ since $\sum_{i>I+1}\theta_i^2 = \sum_{i>I+1} i^{-2s}a_i \le I^{-2s}\sum_i a_i \le MI^{-2s}$ where $a_i = i^{2s}\theta_i^2$.
- Take $I \sim \sigma^{-2/(2s+1)}$. Then the supremum risk is bounded above by $\sigma^{-4s/(2s+1)}$.

# Case 2. $\{\theta : \|\theta\|_0 \le 1, \|\theta\|_2 \le C\}$

- Let $\hat{\theta}_i = Y_i \mathbb{1}\{|Y_i| \ge \sqrt{\frac{2\log n}{n}}\}$.
- W.l.o.g, let $\theta_1 = a \neq 0$. That is, $Y_1 \sim N(a, 1/n)$ and $Y_i \sim N(0, 1/n)$ for $i = 2, \ldots, n$.

$$\mathbb{E}_\theta(\hat{\theta}_1 - \theta_1)^2 = \mathbb{E}\left(Y_1 \mathbb{1}\{|Y_1| \ge \sqrt{\frac{2\log n}{n}}\} - a\right)^2$$

$$\mathbb{E}_\theta \sum_{i=2}^{n} (\hat{\theta}_i - \theta_i)^2 = (n-1)\mathbb{E}\left(Y_i^2 \mathbb{1}\left\{|Y_i| \ge \sqrt{\frac{2\log n}{n}}\right\}\right)$$

- Both terms can be bounded by $\frac{\log n}{n}$.
  - With high prob, $Y_1$ is around $a \pm 2/n$ so that the indicator set is satisfied.
  - For the second term, use $\int_b^\infty t^2 \phi(t) \sim b\phi(b)$ when $b \to \infty$ where $\phi(t)$ is a prob density of $N(0,1)$.

# Nonparametric regression at one point

Hölder class $\Theta^H_{s,L}$ on $[0,1]$: the set of $\ell = \lfloor s \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose derivative $f^{(\ell)}$ satisfies

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \le L|x - x'|^{s-\ell}, \quad \forall x, x' \in [0,1].$$

## Example I(2)

Let $Y_i = \theta(x_i) + \epsilon_i$ for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0,1)$, $x_i = i/n$, $\theta \in \Theta^H_{s,L}$ on $[0,1]$. Then for any $x_0 \in [0,1]$,

$$\inf_{\hat\theta} \sup_{\theta \in \Theta^H_{s,L}} \mathbb{E}_\theta \left(\hat\theta(x_0) - \theta(x_0)\right)^2 \asymp n^{-2s/(1+2s)}.$$

Idea:

- For $\theta \in \Theta^H_{s,L}$ where $s > 1$ and $\ell = \lfloor \beta \rfloor$, if $z$ is sufficiently close to $x$,

$$\begin{aligned}\theta(z) &\approx \theta(x) + h\theta'(x)\frac{(z-x)}{h} + h^2\theta''(x)\frac{(z-x)^2}{2!h^2} + \ldots + h^\ell \theta^{(\ell)}(x)\frac{(z-x)^\ell}{\ell! h^\ell} \\ &= w(x)' U\left(\frac{z-x}{h}\right),\end{aligned}$$

  where

$$\begin{aligned}w(x) &= (\theta(x), h\theta'(x), h^2\theta''(x), \ldots, h^\ell\theta^{(\ell)}(x))^T \\ U(x) &= (1, x, x^2/2!, x^3/3!, \ldots, x^\ell/\ell!)^T.\end{aligned}$$

- Using $Y_i \approx \theta(x_i)$, plug $x_i$ into $z$ above, and estimate $w(x) \in \mathbb{R}^{\ell+1}$ using the least squares idea with an appropriate weight around $x$ via Kernel.
- Then estimate $\hat{\theta}(x) = \hat{w}(x)' U(0)$.

# Local polynomial estimator[1]

- $\hat{\theta}_h(\cdot;\ell)$: *local polynomial* estimator of $\theta$ of degree $\ell$ with kernel $K$ and bandwidth $h$ is constructed at $x$ by fitting a polynomial of degree $\ell$ to the data using weighted least squares.
- $(x_i, Y_i)$ is assigned the weight $K_h(x_i - x)$.
- $\hat{\theta}_h(x;\ell) = \hat{\beta}_0$ where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_\ell)^T$ minimizes

$$\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1(x_i-x)-\beta_2(x_i-x)^2-\ldots-\beta_\ell(x_i-x)^\ell)^2 K_h(x_i-x)$$

over $\beta \in \mathbb{R}^{\ell+1}$.

[1] For equivalent definition, see definition 1.6 of Tsybakov(2009)

- Solve $\hat{\beta} = (X'WX)^{-1}X'WY$ where

$$X_{n\times(\ell+1)} = \begin{bmatrix} 1 & x_1 - x & (x_1 - x)^2 & \dots & (x_1 - x)^\ell \\ 1 & x_2 - x & (x_2 - x)^2 & \dots & (x_2 - x)^\ell \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & (x_n - x)^2 & \dots & (x_n - x)^\ell \end{bmatrix}$$

and

$$W_{n\times n} = \text{diag}(K_h(x_1 - x), K_h(x_2 - x), \dots, K_h(x_n - x))$$

and $Y = (Y_1, \dots, Y_n)^T$.

Explicit formulae exist for $\ell = 0$ and $\ell = 1$ case.

- $\ell = 0$: *local constant*(Nadaraya–Watson) estimator

$$\hat{\theta}_h(x; 0) = \frac{\sum_{i=1}^n K_h(x_i - x) Y_i}{\sum_{i=1}^n K_h(x_i - x)}$$

- $\ell = 1$: *local linear* estimator

$$\hat{\theta}_h(x; 1) = \frac{1}{n} \sum_{i=1}^n \frac{s_{2,h}(x) - s_{1,h}(x)(x_i - x)}{s_{2,h}(x) s_{0,h}(x) - s_{1,h}^2(x)} K_h(x_i - x) Y_i$$

  where $s_{r,h}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - x)^r K_h(x_i - x)$.

- All local polynomial estimators are of the form $\hat{\theta}(x) = \frac{1}{n} \sum_{i=1}^n W(x, x_i) Y_i$, i.e. *linear smoothers*.

# Proof ideas of the upper bound in Example I(2)

Use $\hat{\theta}_h(x;\ell)$ where $h = h_n$ is non-random with $h \to 0$ and $nh \to \infty$ as $n \to \infty$.
Assume $K$ is non-negative, continuous, $K(x) = 0$ for $|x| > 1$, $\int_{-1}^{1} K(x)dx = 1$ and $\int xK(x)dx = 0$.

- $\mathbb{E}_\theta(\hat{\theta}(x) - \theta(x))^2 = \text{Bias}^2 + \text{Var}$.
- $\text{Bias}^2 : (\mathbb{E}_\theta(\hat{\theta}(x)) - \theta(x))^2 \propto h^{2s}$
- $\text{Var} : \mathbb{E}_\theta(\hat{\theta}(x) - \mathbb{E}(\hat{\theta}(x))^2 \propto \frac{1}{nh}$
- Equating these two gives the rate $n^{-2s/(1+2s)}$ by choosing $h \sim n^{-1/(1+2s)}$.

# Nonparametric regression with a global loss

## Example II(2)

Let $Y_i = \theta(X_i) + w_i$ where $X_i \sim Uni[0,1]$, $w_i \sim N(0,1)$ and $X_i \perp w_i$ for $i = 1, \ldots, n$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta^H_{s,L}} \mathbb{E}_\theta \int \left(\hat{\theta}(x) - \theta(x)\right)^2 dx \asymp n^{-\frac{2s}{2s+1}}.$$

Use the local polynomial estimator of degree $\lfloor s \rfloor$ with a bandwidth $h \sim n^{-1/(2s+1)}$ to get the rate $n^{2s/(2s+1)}$.

# High-dimensional linear regression

## Example II(1)[2]

Suppose we have $\{(x_i, y_i)\}_{i=1}^n$ from $y_i = x_i^T\theta + w_i$ where $x_i \in \mathbb{R}^p$ and $w_i \sim N(0, \sigma^2)$ is i.i.d., and $\theta \in \mathbb{R}^p$. Assume $p > n$ and let

$$\Theta_s = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s, \|\theta\|_2 \le 1\}.$$

Also we let $\gamma_{2s} = \sup_{\theta \in \{\|\theta\|_0 \le 2s\}} \frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2}$ and $\gamma_0 = \inf_{\{\theta \in \|\theta\|_0 \le 2s\}} \frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2}$. Then

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta_s} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|_2 \asymp \sigma \sqrt{\frac{s}{n} \log\left(\frac{p - s/2}{s}\right)}.$$

[2] Raskutti, Wainwright, and Yu (2011)

# M-estimator

- Theoretical estimator:

$$\hat{\theta} \in \arg\min_{\theta \in \Theta_s} \|Y - X\theta\|_2^2.$$

- Since $\theta^* \in \Theta_s$, $\|Y - X\hat{\theta}\|_2^2 \le \|Y - X\theta^*\|_2^2$. Let $\hat{\Delta} = \hat{\theta} - \theta^*$, then

$$\frac{1}{n}\|X\hat{\Delta}\|_2^2 \le \frac{2|w^T X\hat{\Delta}|}{n}.$$

- Assumption gives $\gamma_0^2\|\hat{\Delta}\|_2^2 \le \frac{1}{n}\|X\hat{\Delta}\|_2^2$.
- Use $|w^T X\hat{\Delta}|/n \le \left\|\frac{w^T X}{n}\right\|_\infty \|\hat{\Delta}\|_1$ where $w^T X/n$ is normal with zero mean and variance $\sim \sigma^2/n$. By union bound, $\left\|\frac{w^T X}{n}\right\|_\infty \lesssim \sqrt{\frac{3\log p}{n}}$ w.h.p.
- Since $\|\hat{\Delta}\|_0 \le 2s$, $\|\hat{\Delta}\|_1 \le \sqrt{2s}\|\hat{\Delta}\|_2$, $\|\hat{\Delta}\|_2^2 \le C\sigma\sqrt{s\log p/n}$.

# Smooth density estimation

## Example I(4)

Let $Y_1, \ldots, Y_n$ be an i.i.d. sample from a density $f \in \Theta^N_{s,B}$ where $\Theta^N_{s,B}$ is a class of $\ell = \lfloor s \rfloor$ times differentiable densities with Hölder type (Nikol'ski class) condition $\int \left(f^{(\ell)}(x+h) - f^{(\ell)}(x)\right)^2 dx \le (B|h|^{s-\ell})^2$. Then

$$\inf_{\hat{f}} \sup_{f \in \Theta^N_{s,B}} \mathbb{E}_f \int (\hat{f}(x) - f(x))^2 dx \asymp n^{-\frac{2s}{1+2s}}.$$

cf. Bayes predictive estimator of YB gives the same rate (see p.34 of III).

# Kernel density estimator

Consider a kernel $K$ of order $\ell = \lfloor s \rfloor$ i.e.
$u \mapsto u^j K(u),\ j = 0, \ldots, \ell$ are integrable and
$\int K(u)du = 1,\ \int u^j K(u)du = 0,\ j = 1, \ldots, \ell$. Also let
$\int |u|^s |K(u)|du < \infty$. Let

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

- Bias$^2 \propto h^{2s}$
- Var $\propto 1/(nh)$
- Equating these two gives the rate $n^{-2s/(1+2s)}$ by choosing $h \sim n^{-1/(1+2s)}$.

# Sparse covariance matrix estimation (Cai & Zhou, 2012)

## Example III(4)

Suppose we have an i.i.d. sample $\mathsf{X}_i$ from $p$ variate Gaussian with a covariance matrix $\Sigma$. Assuming a sparse $\Sigma$ (at most $k(\leq Mn^{1/2}(\log p)^{-3/2}) + 1$ nonzero elements on each row and column),

$$\sup_{\Sigma \in \mathcal{G}_0(k)} \mathbb{E}_\Sigma \left\| \hat{\Sigma} - \Sigma \right\|^2 \asymp \left( k^2 \frac{\log p}{n} \right).$$

# Thresholding estimator

- Let $S = (s_{ij})_{1 \le i,j \le p} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T$ be close to a sample covariance matrix.
- Let $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p \times p}$ where

$$\hat{\sigma}_{ij} = s_{ij} \mathbb{1} \left\{ |s_{ij}| \ge \gamma \sqrt{\frac{\log p}{n}} \right\}.$$

- Define the event $E_{ij}$,

$$E_{ij} := \left\{ |\hat{\sigma}_{ij} - \sigma_{ij}| \le 4 \min \left( |\sigma_{ij}|, \gamma \sqrt{\frac{\log p}{n}} \right) \right\}$$

s.t. $\mathbb{P}(E_{ij}) \ge 1 - cp^{-9/2}$.

- Let $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_\infty = \max_i \sum_j |a_{ij}|$, then

$$\|\|A\|\|^2 \le \|A\|_1 \|A\|_\infty.$$

  When $A$ is symmetric, $\|A\|_1 = \|A\|_\infty$, so $\|\|A\|\|^2 \le \|A\|_1^2$

- Then

$$\mathbb{E}\left\|\left\|\hat{\Sigma} - \Sigma\right\|\right\|^2 = \mathbb{E}\left(\max_j \sum_i |\hat{\sigma}_{ij} - \sigma_{ij}| E_{ij}\right)^2 + \text{neg.}$$

  where the first term can be bounded by

$$\left(\max_j \sum_{i \ne j} \min\left\{|\sigma_{ij}|, \gamma\sqrt{\frac{\log p}{n}}\right\}\right)^2 + C\frac{\log p}{n}.$$

- Since there exists at most $\sim k$ nonzero elements in each column, the first term above is bounded by $k^2 \frac{\log p}{n}$.

# Estimation of a functional (Bickel & Ritov, 1988)

## Example III(2)

Suppose we have i.i.d. sample $X_1, \ldots, X_n$ from a density $\theta \in \Theta$ where

$$\Theta := \{\theta \text{ on } [0,1],\ 0 < c_0 \leq \theta(x) \leq c_1 < \infty, |\theta^{(2)}(x)| \leq c_2 < \infty\}.$$

and let $T(\theta) = \int_0^1 (\theta'(x))^2 dx$ for $\theta \in \Theta$. Then

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta |\hat{T} - T(\theta)| \geq c n^{-4/9}.$$

# Ideas obtaining an estimator

Distributed Taylor expansion:
$\psi(\mathbb{P}) = \psi(\tilde{\mathbb{P}}) - \int \phi(z, \tilde{\mathbb{P}}) d(\tilde{\mathbb{P}}(z) - \mathbb{P}(z)) + R(\tilde{\mathbb{P}}, \mathbb{P})$ where
$\int \phi(z, \mathbb{P}) d\mathbb{P}(z) = 0$ where $\phi(z, \mathbb{P})$ is influence function, (similar to) gradient $\frac{\partial \psi(\mathbb{P})}{\partial p(z)}$ where $p$ is the density of $\mathbb{P}$.

- If $\psi(\mathbb{P}) = \int \theta(x)^2 dx$, then $\phi(z, \tilde{\mathbb{P}}) = 2\hat{\theta}(z) - 2\int \tilde{\theta}(x)^2 dx$ and use the empirical distribution in place of $\mathbb{P}$.
- If $\psi(\mathbb{P}) = \int \left(\theta^{(1)}(x)\right)^2 dx = -\int \theta^{(2)}\theta(x) dx$, then $\phi(z, \tilde{\mathbb{P}}) = -2\hat{\theta}^{(2)}(z) + 2\int \tilde{\theta}^{(2)}\theta(x) dx$.

# Estimator for $\int \theta^2(x)dx$

- Divide the sample into two $X_1, \ldots, X_{n_1}$ and $X_{n_1+1}, \ldots, X_n$.
- Take a suitable kernel $K_h(x) = h^{-1}K(x/h)$.
- Let $\hat{T}_0 = \frac{n_1}{n}\hat{T}_{01} + \frac{n_2}{n}\hat{T}_{02}$ where

$$\hat{T}_{01} = \int \hat{f}_2^2 dx + \frac{2}{n_1}\sum_{i=1}^{n_1}\left(\hat{f}_2(X_i) - \int \hat{f}_2^2(x)dx\right) + \frac{1}{n_2}\int K_h^2(x)dx$$

$$= 2\int K_h(x-t)d\hat{F}_1(t)d\hat{F}_2(x) - \frac{1}{n_2^2}\sum_{n_1+1\le i\ne j\le n}\int K_h(x-X_i)K_h(x-X_j)dx$$

where $\hat{f}_i(x) = \int K_h(x-y)d\hat{F}_i(y) = \frac{1}{n_1}\sum_{i=1}^{n_!} K_h(x-X_i)$ with $\hat{F}_1$ and $\hat{F}_2$ are empirical distribution fuctions of each sample.

- $\hat{T}_{02}$ is obtained by interchanging the roles of the two subsamples.

# Estimator for $\int(\theta^{(1)}(x))^2dx$

- Use similar ideas as before, then

$$\hat{T}_1 = -2\int K_h^{(2)}(x-t)d\hat{F}_1(t)d\hat{F}_2(x)$$
$$-\frac{n_2}{n(n_1(n_1-1))}\sum_{1\le i<j\le n_1}\int K_h^{(1)}(x-X_i)K_h^{(1)}(x-X_j)dx$$
$$-\frac{n_1}{n(n_2(n_2-1))}\sum_{n_1+1\le i<j\le n}\int K_h^{(1)}(x-X_i)K_h^{(1)}(x-X_j)dx$$

# Non-smooth functional (Cai & Low, 2011)

## Example III(6)

Let $Y_i \sim N(\theta_i, 1)$ and our interest is to estimate $T(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\theta_i|$ assuming $|\theta_i| \leq 1$. Let $\Theta = \{\theta = (\theta_1, \ldots \theta_n), |\theta_i| \leq 1\}$. Then

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{E}(\hat{T} - T(\theta))^2 \asymp \beta_*^2 \left( \frac{\log \log n}{\log n} \right)^2 .$$

# Remarks

Model: $Y_i \sim N(\theta_i, 1)$ for $i = 1, \ldots, n$ with $|\theta_i| \le 1$.

- Different from estimating $\theta \in \mathbb{R}^n$ (assuming various conditions including tail decay, sparsity, etc.)
- If we estimate $\bar{\theta} := \frac{1}{n}\sum_{i=1}^n \theta_i$, then we just use $\bar{Y}$. Then

$$\mathbb{E}_\theta(\bar{Y} - \bar{\theta})^2 = \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^n (Y_i - \theta_i)\right)^2 = \frac{1}{n}$$

- If we estimate $\bar{\theta^2} = \frac{1}{n}\sum_{i=1}^n \theta_i^2$, we can use UE $\tilde{\theta^2} := \frac{1}{n}\sum_{i=1}^n (Y_i^2 - 1)$. Then

$$\mathbb{E}_\theta(\tilde{\theta^2} - \bar{\theta^2})^2 = \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^n (Y_i^2 - 1 - \theta_i^2)\right)^2 \sim \frac{1}{n}.$$

- These two don't need any assumption on $\theta$.

# Constructing estimator

Idea:

- Nonexistence of an UE of $|\theta_i|$. Consider the best polynomial approximation to the absolute value function
- Smooth at 0 by a polynomial approximation and construct an UE for each term in the expansion using the Hermite polynomials.
- Recall: $H_k$ be the Hermite polynomial defined by

$$\frac{d^k}{dy^k}\phi(y) = (-1)^k H_k(y)\phi(y).$$

Then $\int H_k^2(y)\phi(y)dy = k!$ and $\int H_k(y)H_j(y)\phi(y)dy = 0$ if $k \neq j$.

# Constructing estimator

- $G_K^*(x) = \sum_{k=0}^K g_{2k}^* x^{2k}$: best polynomial approximation of degree $2K$ to $|x|$, then

$$\max_{x\in[-1,1]} \left|G_K^*(x) - |x|\right| \le \frac{\beta_*}{2K}(1+o(1)).$$

- Approximate $\frac{1}{n}\sum_{i=1}^n |\theta_i|$ by a smooth functional $\tilde{T}(\theta) = \frac{1}{n}\sum_{i=1}^n G_K^*(\theta_i) = \sum_{k=0}^K g_{2k}^* b_{2k}(\theta)$, where $b_{2k}(\theta) := \frac{1}{n}\sum_{i=1}^n \theta_i^{2k}$.
- Estimate $b_{2k}(\theta)$ by its UE: $\frac{1}{n}\sum_{i=1}^n H_{2k}(y_i)$.
- Choose $K \sim \frac{\log n}{2\log\log n}$, then the bias is bounded by $\sim \frac{\beta_*}{2K}$ and the variance is bounded by $K^{2K}/n$.

# Estimating smooth functional

## Cai & Low (2005)

Let $Y_i = \theta_i + n^{-1/2}\epsilon_i$ where $\epsilon_i \sim N(0,1)$. Let $\Theta = \{\theta : \sum_i i^{2s}\theta_i^2 \leq M\}$. We estimate $Q(\theta) := \sum_{i=1} \theta_i^2$.

$$\inf_{\hat{Q}} \sup_{\theta \in \Theta} \mathbb{E}(\hat{Q} - Q(\theta))^2 \asymp n^{\gamma}$$

where $\gamma = -1$ when $s \geq 1/4$ and $\gamma = -8s/(1+4s)$ when $s < 1/4$.

UB: estimate $\theta_i$ by its UE until $i \leq m$ (where $m$ is chosen later), i.e. define

$$\hat{Q}_m := \sum_{i=1}^{m} (Y_i^2 - 1/n).$$

$Y_i = \theta_i + n^{-1/2}\epsilon_i$ and we estimate $Q(\theta) := \sum_i \theta_i^2$ where $\sum_i i^{2s}\theta_i^2 \le M$. Let $\hat{Q}_m := \sum_{i=1}^m (Y_i^2 - 1/n)$.

- $\mathbb{E}\hat{Q}_m = \sum_{i=1}^m \theta_i^2$.
- $\mathrm{Var}(Q_m) = \sum_{i=1}^m \mathrm{Var}(Y_i^2 - 1/n) = \sum_{i=1}^m \mathrm{Var}(\frac{\epsilon_i^2 - 1}{n} - \frac{2}{\sqrt{n}}\epsilon_i\theta_i) = \frac{2m}{n^2} + \frac{4}{n}\sum_{i=1}^m \theta_i^2 \le \frac{2m}{n^2} + \frac{4M}{n}$
- $\mathrm{Bias}^2 = \left(\sum_{i>m} \theta_i^2\right)^2 \le \frac{M^2}{m^{4s}}$.
- Equate $m/n^2 \sim m^{4s}$, then $m \sim n^{2/(1+4s)}$, which gives the upper bound

$$\mathbb{E}_\theta(\hat{Q}_m - Q(\theta))^2 \lesssim n^{-8s/(1+4s)} + n^{-1}.$$

LB: use Le Cam II. Let

$$\theta_0 = 0, \quad \Theta_1 = \{\theta : |\theta_i| = a, 1 \le i \le m, \ \theta_i = 0, i \ge m+1\}$$

Let $\mathbb{P}_0 = \otimes_{i=1}^m N(0, 1/n)$ and $\mathbb{P}_1 = \frac{1}{2^m} \sum_{\theta \in \Theta_1} \mathbb{P}_\theta$.

- (loss) For $\theta_1 \in \Theta_1$, $(Q(\theta_1) - Q(\theta_0))^2 = (\sum_{i=1}^m a^2)^2 = m^2 a^4$.
- (test) The uniform mixture over $2^m$ measures is also a product measure $\otimes_{i \le m} \left(\frac{1}{2} N(a, 1/n) + \frac{1}{2} N(-a, 1/n)\right)$,

$$\chi^2(\mathbb{P}_0, \mathbb{P}_1) = \int \frac{(d\mathbb{P}_1)^2}{d\mathbb{P}_0} - 1 \le \exp\left(\frac{1}{2} m n^2 a^4\right) - 1$$

- Take $ma^4 \sim n^{-2}$ to obtain the lower bound $mn^{-2}$.

- Check for $\theta \in \Theta_1$, $\sum_i i^{2s}\theta_i^2 = \sum_{i \le m} i^{2s} a^2 \sim m^{2s+1}a^2 \le M$ by choosing $m \sim a^{-2/(2s+1)}$. That is, $ma^4 \sim a^{(8s+2)/(2s+1)} \sim n^{-2}$ which yields $a \sim n^{-(2s+1)/(4s+1)}$ and $m \sim n^{2/(4s+1)}$.
- The first lower bound is $n^{-8s/(4s+1)}$.
- The other lower bound is via the information: $I(Q(\theta)) = \frac{n}{4\sum_i \theta_i^2} \le \frac{n}{4M}$ to obtain $1/n$ rate.

# Revisit Gaussian sequence models

## Example I(3)

Let $Y_i = \theta_i + \sigma\epsilon_i$ where $\epsilon_i \sim N(0,1)$ and $\sigma \to 0$ with $\theta = (\theta_1, \theta_2, \ldots) \in \Theta$, where $\Theta = \left\{\theta : \sum_i i^{2s}\theta_i^2 \le M\right\}$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 \asymp M^{1/(2s+1)}\sigma^{-4s/(2s+1)}.$$

Q. Can we obtain *sharp* minimax optimal estimator? That is, can we prove

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 = C^*(s,M)\sigma^{-4s/(2s+1)}(1+o(1))?$$

# Pinsker Theorem

## Informal version

Let $Y_i = \theta_i + \sigma\epsilon_i$ where $\epsilon_i$ are i.i.d. $N(0,1)$ where $\theta \in \Theta = \{\theta : \sum_i b_i^2\theta_i^2 \leq M\}$ and $b_i \to \infty$. Denote a linear estimator by $\hat{\theta}_i^L = w_i Y_i$ to estimate $\theta_i$ . Then

$$\inf_{\hat{\theta}^L} \sup_{\theta\in\Theta} \mathbb{E}\|\theta - \hat{\theta}^L\|_2^2 \sim \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}\|\theta - \hat{\theta}\|_2^2$$

as $\epsilon \to 0$.

First we find the best linear estimator and obtain the supremum risk.
Then we get the lower bound using Bayes idea.

# Upper bound

- Consider a linear estimator $\hat{\theta}_i = w_i Y_i$ for $i = 1, \ldots,$ where $w_i$ is chosen s.t. the supremum risk over $\Theta$ is minimized later.
- Bias-variance decomposition gives

$$\mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 = \mathbb{E}_\theta \sum_i (w_i Y_i - \theta_i)^2 = \sum_{i=1}^{\infty} \sigma^2 w_i^2 + (1 - w_i)^2 \theta_i^2.$$

- We consider the supremum risk using $w_i Y_i$:

$$\inf_{\{w_i\}} \sup_{\{\theta : \sum_i b_i^2 \theta_i^2 \le M\}} \sum_{i=1}^{\infty} \left(\sigma^2 w_i^2 + (1 - w_i)^2 \theta_i^2\right)$$

$$= \inf_{\{w_i\}} \left( \sum_{i=1}^{\infty} \sigma^2 w_i^2 + \sup_{\sum_i a_i \le M} \sum_{i=1}^{\infty} (1 - w_i)^2 b_i^{-2} a_i \right)$$

- Second term is maximized by putting $a_j = M$ where $(1-w_j)^2 b_i^{-2}$ is the largest. Thus

$$\inf_{\{w_i\}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \sum_{i=1}^{\infty} (w_i\theta_i - \theta)^2 = \inf_{\{w_i\}} \sum_{i=1}^{\infty} \left( \sigma^2 w_i^2 + \sup_j (1-w_j)^2 b_j^{-2} M \right)$$

$$= \inf_{\lambda} \inf_{\{\sup_j (1-w_j)^2 b_j^{-2} = \lambda^{-2}\}} \left( \sum_{i=1}^{\infty} \sigma^2 w_i^2 + \lambda^{-2} M \right)$$

- When $\sup_j (1-w_j)^2 b_j^{-2} = \lambda^{-2}$, then $w_i \geq (1 - \frac{b_i}{\lambda})_+$ for all $i$ and equality holds for at least one $i$, which implies

$$\inf_{\{\sup_i (1-w_i)^2 b_i^{-2} = \lambda^{-2}\}} \left( \sum_{i=1}^{\infty} \sigma^2 w_i^2 + \lambda^{-2} M \right) = \lambda^{-2} M + \sum_i \sigma^2 (1 - \frac{b_i}{\lambda})_+^2.$$

- RHS in the above is minimized when $\lambda$ is satisfied

$$\lambda = \frac{M + \sigma^2 \sum_{i \leq \lambda^{1/s}} b_i^2}{\sigma^2 \sum_{i \leq \lambda^{1/s}} b_i}.$$

- Note that when $b_i \sim i^s$,
  - $\sum_{i \le I} i^{2s} \sim I^{2s+1}/(2s+1)$
  - $\sum_{i \le I} i^{s} \sim I^{s+1}/(s+1)$
- Need to solve (where $I = \lambda^{1/2}$)

$$\lambda \sim \frac{M + \sigma^2 \lambda^{(2s+1)/s}/(2s+1)}{\sigma^2 \lambda^{(s+1)/s}/(s+1)} \sim \frac{M}{s+1}\sigma^{-2}\lambda^{-(s+1)/s} + \frac{s+1}{2s+1}\lambda.$$

That is,

$$\lambda \sim (M\sigma^{-2})^{\frac{s}{2s+1}}.$$

- By careful calculation, using $\hat{\theta}_i = w_i \theta_i$, by choosing $w_i = (1 - b_i/\lambda)_+$, we get

$$\inf_{\{w_i\}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 = \sigma^2 (I - \frac{\sum_i i^s}{\lambda}) \sim P_\gamma M^{(1-2\gamma)} \pi^{-2\gamma} \sigma^{4\gamma},$$

where $\gamma = s/(2s+1)$ and $P_\gamma = \gamma(1-\gamma)^{\gamma-1}(2-\gamma)^{-\gamma}$ is Pinsker constant.

# Ideas of lower bound

- Least favorable prior $\theta_i \sim N(0, \tau_i^2)$ with $\tau_i^2 = \sigma^2(\lambda/b_i - 1)_+$ since $\tau_i^2/(\tau_i^2 + \sigma^2) \times y_i = (1 - b_i/\lambda)_+ y_i$ is the optimal linear estimator.
- This prior gives the following Bayes risk $\sum_i \tau_i^2\sigma^2/(\tau_i^2 + \sigma^2) = \sigma^2 \sum_i (1 - b_i/\lambda)_+$ is equal to the maximum risk with the optimal linear estimator.
- Consider the prior $G_\sigma = \prod_{i=1} N(0, \sigma^2(\lambda/b_i - 1)_+)$, but then

$$\mathbb{E} \sum_i b_i^2\theta_i^2 = \sum \sigma^2 b_i^2 (\lambda/b_i - 1)_+ = \sigma^2 \sum_i b_i(\lambda - b_i)_+ = M.$$

Thus, this prior is not supported in $\Theta$ that we consider.

- Strategy: find a sequence of priors $Q_\sigma$ supported in $\Theta$ s.t. the Bayes risk is $(1+o(1))\times$ best linear minimax risk.
- Natural prior $G_\sigma$ is not supported in $\Theta$
- What about $\kappa_\sigma G_\sigma$ with $\kappa_\sigma \to 1$? That is,

$$G_\sigma^* = \prod_i N(0, \kappa_\sigma^2 \sigma^2 (\lambda/b_i - 1)_+)$$

  where $\kappa_\sigma^2 < 1$ and increase to 1 with a certain rate.
- Hope $G_\sigma^*$ is concentrated on $\Theta$ by choosing $\kappa_\sigma$ appropriately. Then define $Q_\sigma = G_\sigma^*(\cdot|\Theta)$.
- If we can show $G_\sigma^*$ is concentrated on $\Theta$, then $G_\sigma^*$ must be very close to $Q_\sigma$, so these two Bayes risk is very close. $Q_\sigma$ Bayes risk must be the lower bound of the general minimax risk.
- Then $\inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\theta - \hat\theta\|_2^2 \geq$
  $\kappa_\sigma \inf_{\{w_i\}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\theta - \hat\theta\|_2^2 (1+o(1))$

# Conventional gaussian models

- (WN) White Noise model: observe $Y_n(t)$ from

$$dY_n(t) = f(t)dt + n^{-1/2}dW(t), \quad 0 \le t \le 1 \qquad (1)$$

where $f : [0,1] \to \mathbb{R}$, $W(t)$ is a standard Brownian motion.

- (GS) Gaussian Sequence model: observe $y_1, \ldots, y_n$ from

$$y_i = \theta_i + n^{-1/2}\epsilon_i, \quad \epsilon_i \sim N(0,1), \quad i \in I$$

- (NR) Nonparametric Regression model: observe $Y_1, \ldots, Y_n$ from

$$Y_i = f(i/n) + \eta_i, \quad \eta_i \sim N(0,1), \quad i = 1, \ldots, n. \qquad (2)$$

# A) Connection between WN & GS

- Suppose $\{\psi_i(t) : i \in I\}$ is an orthonormal basis of $\mathcal{L}^2[0,1]$.
- Equation (1) gives

$$\int_0^1 \psi_i(t) dY_n(t) = \int_0^1 \psi_i(t) f(t) dt + n^{-1/2} \int_0^1 \psi_i(t) dW(t),$$

  that is, $y_i = \theta_i + n^{-1/2}\epsilon_i$ where $y_i := \int \psi_i(t) dY_n(t)$, $\theta_i := \int f(t)\psi_i(t)dt$, $\epsilon_i := \int \psi_i(t) dW(t) \sim N(0,1)$ since $\psi_i(t)$ is orthonormal.
- That is, observing WN gives information from $y_i = \theta_i + n^{-1/2}\epsilon_i$ where $\epsilon_i \sim N(0,1)$, i.e. GS model.
- Estimator $\hat{\theta}_i$ of $\theta_i$ gives $\hat{f}(t) = \sum_{i=1}^{I} \hat{\theta}_i \psi_i(t)$ so that $\int \hat{f}(t)\psi_i(t)dt = \hat{\theta}_i$. Thus $\mathbb{E}_f \|\hat{f}(t) - f(t)\|^2 dt = \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2$

# B) Connection between WN & NR

- $dY_n(t) = f(t)dt + n^{-1/2}dW(t)$ can be discretized: integrating over $[t, t+\delta]$,

$$\frac{Y_n(t+\delta) - Y_n(t)}{\delta} = \frac{1}{\delta}\int_t^{t+\delta} f(s)ds + \frac{n^{-1/2}}{\delta}(W(t+\delta) - W(t)).$$

- Let $y(t) := \frac{Y_n(t+\delta)-Y_n(t)}{\delta}$ and $\eta(t) := \frac{n^{-1/2}}{\delta}(W(t+\delta) - W(t))$, then $\eta(t) \sim N(0, \frac{1}{n\delta})$.
- Take $\delta = 1/n$, then $y(t) \approx f(t) + \eta(t)$ where $\eta(t) \sim N(0,1)$ and the approximation error is $\frac{1}{\delta}\int_t^{t+\delta} f(s)ds - f(t)$.
- Take $t = i/n$, then the above is

$$Y_i \approx f(i/n) + \eta_i, \quad \eta_i \sim N(0,1).$$

# C) Connection between NR & GS

- Suppose NR model (2) consider a basis $\{\psi_i : i \in I\}$ satisfying $\frac{1}{n}\sum_{i=1}^{n} \psi_j(i/n)\psi_k(i/n) = \mathbb{1}\{j = k\}$ for $1 \le k, j \le n-1$.
- Let $y_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_j(i/n)$, $f_j = \frac{1}{n}\sum_{i=1}^{n} f(i/n)\psi_j(i/n)$, and $\epsilon_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_i\psi_j(i/n)$. Then $\epsilon_j \sim N(0,1)$.
- Then $y_j = f_j + n^{-1/2}\epsilon_j$, for $j = 1, \ldots, n$. Since $\theta_j = \int f(t)\psi_j(t)dt \approx \frac{1}{n}\sum_{i=1}^{n} f(i/n)\psi_j(i/n)$, we have $\theta_j \approx f_j$. Hence, $y_j \approx \theta_j + n^{-1/2}\epsilon_j$.

# More about B) NR and WN

(1) WN: $\{Y_n(t) = \int_0^t f(x)dx + n^{-1/2}W(t), \quad 0 \le t \le 1\}$
(2) NR: $\{Y_i = f(i/n) + \eta_i, \quad i = 1, \ldots, n\}$

Brown and Low(1996, AS): "To any NR problem, $\exists$ corresponding WN problem which is asymptotically *equivalent*."

- Two statistical problems $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ with sample spaces $\mathcal{X}^{(i)}$, $i = 1, 2$ with the same parameter space $\Theta$ .
- Family of distributions $\{P_\theta^{(i)} : \theta \in \Theta\}$.
- $\mathcal{A}$: action space, $L : \Theta \times \mathcal{A} \to [0, \infty)$: loss function, $\delta^{(i)}$ : decision procedure
- $R_\theta^{(i)}(\delta^{(i)}, L) = \int L(\theta, \delta^{(i)})dP_\theta^{(i)}$: risk
- $\|L\| = \sup\{L(\theta, a) : \theta \in \Theta, a \in \mathcal{A}\}$

# Le Cam's metric

- Risk difference
  $D := D(\delta^{(1)}, \delta^{(2)}, \theta, L) := \left| R_\theta^{(1)}(\delta^{(1)}, L) - R_\theta^{(2)}(\delta^{(2)}, L) \right|$
- Le Cam's metric between two experiments

  $$\Delta := \Delta(\mathcal{P}^{(1)}, \mathcal{P}^{(2)}) = \max \left[ \inf_{\delta^{(1)}} \sup_{\delta^{(2)}} \sup_{\theta} \sup_{L:\|L\|=1} D, \inf_{\delta^{(2)}} \sup_{\delta^{(1)}} \sup_{\theta} \sup_{L:\|L\|=1} D \right].$$

- If $\Delta < \epsilon$, for every procedure $\delta^{(i)}$ in problem $i$, $\exists$ a procedure $\delta^{(j)}$ in problem $j$ $(j \neq i)$, with risk $\leq \epsilon$, uniformly over all $L$ s.t. $\|L\| = 1$ and $\theta \in \Theta$.
- $\{\mathcal{P}_n^{(1)} : n = 1, \ldots, \}$ and $\{\mathcal{P}_n^{(2)} : n = 1, \ldots, \}$ are asymptotically equivalent if for any sequence $\delta_n^{(1)}$ in $\mathcal{P}_n^{(1)}$, $n = 1, \ldots$, $\exists$ a sequence $\delta_n^{(2)}$ in $\mathcal{P}_n^{(2)}$ for which

  $$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left| R_\theta^{(1)}(\delta_n^{(1)}, L) - R_\theta^{(2)}(\delta_n^{(2)}, L) \right| = 0.$$

# Key steps

Construct

1. randomized versions $\tilde{\mathcal{P}}_n^{(i)}$ of $\mathcal{P}_n^{(i)}$, characterized by families of prob. measures $\{\tilde{P}_\theta^{(i,n)}, \theta \in \Theta_n\}$ in certain spaces $\tilde{\mathcal{X}}_{i,n}$
2. equivalence mappings $T_{1,n} : \tilde{\mathcal{X}}_{1,n} \to \mathcal{X}_{2,n}$ and $T_{2,n} : \tilde{\mathcal{X}}_{2,n} \to \mathcal{X}_{1,n}$, independent of $\theta$, s.t.

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} h^2(P_\theta^{1,n}, \tilde{P}_\theta^{(2,n)} \circ T_{2,n}^{-1}) = 0$$

and

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} h^2(P_\theta^{2,n}, \tilde{P}_\theta^{(1,n)} \circ T_{1,n}^{-1}) = 0.$$

# Asymptotic equivalence

By considering a (randomized) synthetic procedure $\tilde{\delta}^{(2,n)} = \delta^{(1,n)} \circ T_{2,n}$,

$$\sup_{\theta\in\Theta}\left|\int L(\theta,\delta^{(1,n)})dP_\theta^{(1,n)} - \int L(\theta,\tilde{\delta}^{(2,n)})dP_\theta^{(2,n)}\right| =$$

$$\sup_{\theta\in\Theta}\left|\int_{\mathcal{X}_{1,n}} L(\theta,\delta^{(1,n)})\left\{dP_\theta^{(j,n)} - d\tilde{P}_\theta^{(2,n)}\circ T_{2,n}^{-1}\right\}\right|$$

$$\leq \|L\|_\infty \sup_{\theta\in\Theta}\int_{\mathcal{X}_{1,n}}\left|dP_\theta^{(j,n)} - d\tilde{P}_\theta^{(2,n)}\circ T_{2,n}^{-1}\right|$$

$$\leq 2\|L\|_\infty h(P_\theta^{(1,n)}, \tilde{P}_\theta^{(2,n)}\circ T_{2,n}^{-1}),$$

since $L_1(P,Q) = 2TV(P,Q) \leq 2h(P,Q)$. Do similarly for $\tilde{\delta}^{(1,n)}$.

# Equivalence of B) WN & NR

(1) WN: $\{dZ_n(t) = f(t)dt + n^{-1/2}dW(t), \quad 0 \le t \le 1\}$
(2) NR: $\{Y_i = f(i/n) + \eta_i, \quad i = 1, \ldots, n\}$ with $\eta_i \sim N(0,1)$.

## Theorem 4.1 of Brown and Low (1996)

NR model is asymptotically equivalent to WN model under the following conditions:

1. $\sup\{|f(t)| : t \in [0,1], f \in \Theta\} = B < \infty$
2. Define $\bar{f}_n(t) = f(\frac{i}{n})\mathbb{1}\{\frac{i-1}{n} \le t < \frac{i}{n}\}$ for $i = 1, \ldots, n-1$ and $\bar{f}_n(1) = f(1)$, then

$$\lim_{n\to\infty} \sup_{f\in\Theta} n \int_0^1 (f(t) - \bar{f}_n(t))^2 dt = 0.$$

# Remarks

- Condition (2) is satisfied (when $s > 1/2$) for Sobolev type($s \in \mathbb{N}$) condition,

$$\Theta^S_{s,B} := \left\{ f : \int (f^{(s)}(t))^2 dt \le B \text{ and } \sup_{x\in[0,1]} |f(x)| \le B \right\}$$

or Lipshitz condition

$$\Theta^L_{s,B} := \left\{ f : \left| f(x+\delta) - f(x) - \sum_{i=1}^{[s]-1} \frac{f^{(i)}(x)\delta^i}{i!} \right| \right.$$
$$\left. \le B^2|\delta|^s, \text{ and } \sup_{x\in[0,1]} |f(x)| \le B^2 \right\} \text{ if } s \ge 1,$$

$$\Theta^L_s = \{ f : |f(x+\delta) - f(x)| \le B|\delta|^s \} \text{ if } 0 < s < 1.$$

- Equivalence of WN and NR with random design was proved in Brown, et al. (2002).

# Proof ideas of Theorem 4.1

- Construct $\bar{f}_n(t) = f(i/n)\mathbb{1}\{\frac{i-1}{n} \le t < \frac{i}{n}\}$ from NR.
- Let $d\bar{Z}_n(t)$ be WN model with $d\bar{Z}_n(t) = \bar{f}_n(t)dt + n^{-1/2}dW(t)$. Then by Condition (2), $\Delta(\{Z_n(t)\}, \{\bar{Z}_n(t)\}) \to 0$.
- Let $S_n(i) = n\int_{(i-1)/n}^{i/n} d\bar{Z}_n(t)$. Since $\{S_n(i)\}$ are sufficient for $\{\bar{Z}_n(t)\}$, $\Delta(\{\bar{Z}_n(t)\}, \{S_n(i)\}) = 0$.
- Note that $S_n(i)$ are independent normal with $\mathrm{E}S_n(i) = f(i/n)$ and $\mathrm{Var}S_n(i) = 1$, i.e. $\{S_n(i)\}$ has the same distribution with $\{Y_i\}$, which yields $\Delta(\{\bar{Z}_n(t)\}, \{Y_i\}) = 0$. This implies that

$$\Delta(\{Z_n(t)\}, \{Y_i\}) \to 0.$$