Nonparametric Estimation: Part I

Regression in Function Space

Yanjun Han

Department of Electrical Engineering
Stanford University

yjhan@stanford.edu

November 9, 2015

The statistical model

- Given a family of distributions $\{P_f(\cdot)\}_{f \in \mathcal{F}}$ and an observation $Y \sim P_f$, we aim to:
  1. estimate f;
  2. estimate a functional $F(f)$ of f;
  3. do hypothesis testing: given a partition $\mathcal{F} = \cup_{j=1}^{N} \mathcal{F}_j$, decide which $\mathcal{F}_j$ the true function f belongs to.

The risk

- The risk of using $\hat{f}$ to estimate $\theta(f)$ is

$$R(\hat{f}, f) = \Psi^{-1}\left(\mathbb{E}_f \Psi(d(\hat{f}(Y), \theta(f))\right)$$

  where Ψ is a nondecreasing function on $[0, \infty)$ with $\Psi(0) = 0$, and $d(\cdot, \cdot)$ is some metric

- The minimax approach

$$R(\hat{f}, \mathcal{F}) = \sup_{f \in \mathcal{F}} R(\hat{f}, f)$$
$$R^*(\mathcal{F}) = \inf_{\hat{f}} R(\hat{f}, \mathcal{F})$$

Nonparametric regression

Problem: recover a function $f : [0,1]^d \to \mathbb{R}$ in a given set $\mathcal{F} \subset L_2([0,1]^d)$ via noisy observations

$$y_i = f(x_i) + \sigma\xi_i, \qquad i = 1, 2, \cdots, n$$

Some options:

- Grid $\{x_i\}_{i=1}^n$: deterministic design (equidistant grid, general case), random design
- Noise $\{\xi_i\}_{i=1}^n$: iid $\mathcal{N}(0,1)$, general iid case, with dependence
- Noise level σ: known, unknown
- Function space: Hölder ball, Sobolev space, Besov space
- Risk function: risk at a point, integrated risk (L_q risk, $1 \le q \le \infty$, with normalization)

$$R_q(\hat{f}, f) = \begin{cases} \left(\mathbb{E}_f \int_{[0,1]^d} |\hat{f}(x) - f(x)|^q dx\right)^{\frac{1}{q}}, & 1 \le q < \infty \\ \mathbb{E}_f \left(\operatorname{ess\,sup}_{x\in[0,1]^d} |\hat{f}(x) - f(x)|\right), & q = \infty \end{cases}.$$

Equivalence between models

Under mild smoothness conditions, Brown et al. proved the asymptotic equivalence between the following models:

- Regression model: for iid $\mathcal{N}(0,1)$ noise $\{\xi_i\}_{i=1}^n$,

$$y_i = f\left(i/n\right) + \sigma\xi_i, \qquad i = 1, 2, \cdots, n$$

- Gaussian white noise model:

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}}dB_t, \qquad t \in [0, 1]$$

- Poisson process: generate $N = \text{Poi}(n)$ iid samples from common density g $(g = f^2, \sigma = 1/2)$
- Density estimation model: generate n iid samples from common density g $(g = f^2, \sigma = 1/2)$

Bias-variance decomposition

Deterministic error (bias) and stochastic error of an estimator $\hat{f}(\cdot)$:

$$b(x) = \mathbb{E}_f \hat{f}(x) - f(x)$$
$$s(x) = \hat{f}(x) - \mathbb{E}_f \hat{f}(x)$$

Analysis of the L_q risk:

- For $1 \leq q < \infty$:

$$R_q(\hat{f}, f) = \left(\mathbb{E}_f \|\hat{f} - f\|_q^q\right)^{\frac{1}{q}} = \left(\mathbb{E}_f \|b + s\|_q^q\right)^{\frac{1}{q}}$$
$$\lesssim \|b\|_q + \left(\mathbb{E}\|s\|_q^q\right)^{\frac{1}{q}}$$

- For $q = \infty$:

$$R_\infty(\hat{f}, f) \leq \|b\|_\infty + \mathbb{E}\|s\|_\infty$$

The first example

Consider the following regression model:

$$y_i = f\left(i/n\right) + \sigma\xi_i, \qquad i = 1, 2, \cdots, n$$

where $\{\xi_i\}_{i=1}^n$ iid $\mathcal{N}(0,1)$, and $f \in \mathcal{H}_1^s(L)$ for some known $s \in (0,1]$ and $L > 0$, and the Hölder ball is defined as

$$\mathcal{H}_1^s(L) = \left\{ f \in C[0,1] : |f(x) - f(y)| \leq L|x-y|^s, \forall x, y \in [0,1] \right\}.$$

A window estimate

To estimate the value of f at x, consider the window $B_x = [x - h/2, x + h/2]$, then a natural estimator takes the form

$$\hat{f}(x) = \frac{1}{n(B_x)} \sum_{i:x_i \in B_x} y_i,$$

where $n(B_x)$ denotes the number of point x_i in B_x.

- Bias:

$$|b(x)| = |\mathbb{E}_f \hat{f}(x) - f(x)| = \left| \frac{1}{n(B_x)} \sum_{i:x_i \in B_x} f(x_i) - f(x) \right|$$
$$\leq \frac{1}{n(B_x)} \sum_{i:x_i \in B_x} |f(x_i) - f(x)| \leq \frac{1}{n(B_x)} \sum_{i:x_i \in B_x} L|x_i - x|^s \leq Lh^s$$

- Stochastic term:

$$s(x) = \hat{f}(x) - \mathbb{E}_f \hat{f}(x) = \frac{\sigma}{n(B_x)} \sum_{i:x_i \in B_x} \xi_i$$

Optimal window size: $1 \le q < \infty$

$$|b(x)| \le Lh^s, \qquad s(x) = \frac{\sigma}{n(B_x)} \sum_{i:x_i \in B_x} \xi_i$$

Bounding the integrated risk:

$$\begin{aligned} R_q(\hat{f}, f) &\lesssim \|b\|_q + (\mathbb{E}\|s\|_q^q)^{\frac{1}{q}} \\ &\lesssim Lh^s + \frac{\sigma}{\sqrt{nh}} \end{aligned}$$

The optimal window size h^* should satisfy $L(h^*)^s = \frac{\sigma}{\sqrt{nh^*}}$, i.e., $h^* = \left(\frac{\sigma^2}{L^2 n}\right)^{\frac{1}{2s+1}}$, and the resulting risk is

$$R_q(\hat{f}^*, \mathcal{H}_1^s(L)) \lesssim L \left(\frac{\sigma^2}{L^2 n}\right)^{\frac{s}{2s+1}} \asymp n^{-\frac{s}{2s+1}}$$

Optimal window size: $q = \infty$

$$|b(x)| \le Lh^s, \qquad s(x) = \frac{\sigma}{n(B_x)} \sum_{i: x_i \in B_x} \xi_i$$

Fact: for $\mathcal{N}(0,1)$ rv $\{\xi_i\}_{i=1}^M$ (possibly correlated), there exists a constant C such that for any $w \ge 1$,

$$\mathbb{P}\left(\max_{1 \le i \le M} |\xi_i| \ge Cw\sqrt{\ln M} \right) \le \exp\left(-\frac{w^2 \ln M}{2} \right)$$

- Proof: apply the union bound and $\mathbb{P}(|\mathcal{N}(0,1)| > x) \le 2\exp(-x^2/2)$.
- Corollary: $\mathbb{E}\|s\|_\infty \lesssim \sigma\sqrt{\frac{\ln n}{nh}} \quad (M = \mathcal{O}(n^2))$.

Optimal window size and risk

$$h^* \asymp \left(\frac{\ln n}{n} \right)^{-\frac{1}{2s+1}}, \qquad R_\infty(\hat{f}^*, \mathcal{H}_1^s(L)) \lesssim \left(\frac{\ln n}{n} \right)^{-\frac{s}{2s+1}}$$

Bias-variance tradeoff

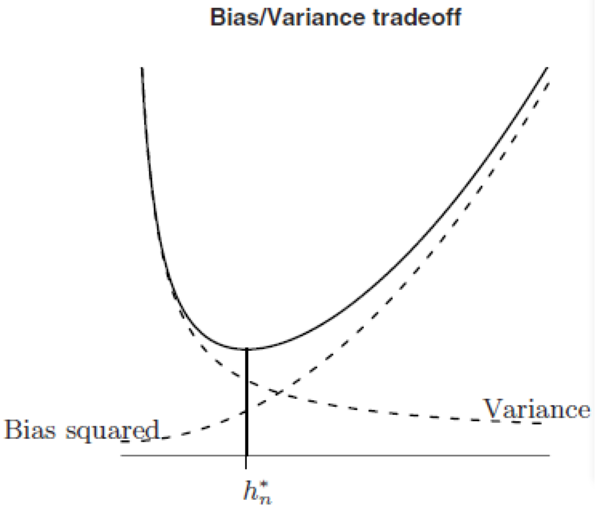

Figure 1: Bias-variance tradeoff

General Hölder ball

Consider the following regression problem:

$$y_\iota = f(x_\iota) + \sigma\xi_\iota, \quad f \in \mathcal{H}_d^s(L), \iota = (i_1, i_2, \cdots, i_d) \in \{1, 2, \cdots, m\}^d$$

where

- $x_{(i_1, i_2, \cdots, i_d)} = (i_1/m, i_2/m, \cdots, i_d/m)$, $n = m^d$
- $\{\xi_\iota\}$ are iid $\mathcal{N}(0,1)$ noises
- The general Hölder ball $\mathcal{H}_d^s(L)$ is defined as

  $$\mathcal{H}_d^s(L) = \left\{ f : [0,1]^d \to \mathbb{R}, |D^k(f)(x) - D^k(f)(x')| \le L|x - x'|^\alpha, \forall x, x' \right\}$$

  where $s = k + \alpha, k \in \mathbb{N}, \alpha \in (0,1]$, and $D^k(f)(\cdot)$ is the vector function comprised of all partial derivatives of f with order k.

  - $s > 0$: modulus of smoothness
  - d: dimensionality

Local polynomial approximation

As before, consider the cube B_x centered at x with edge size h, where $h \to 0, nh^d \to \infty$.

- Simple average no longer works!
- Local polynomial approximation: for $x_\iota \in B_x$, design weights $w_\iota(x)$ such that if the true $f \in \mathcal{P}_d^k$, i.e., polynomial of full degree k and of d variables, we have

$$f(x) = \sum_{\iota : x_\iota \in B_x} w_\iota(x) f(x_\iota)$$

Lemma

There exists weights $\{w_\iota(x)\}_{\iota : x_\iota \in B_x}$ which depends continuously on x and

$$\|w_\iota(x)\|_1 \le C_1, \qquad \|w_\iota(x)\|_2 \le \frac{C_2}{\sqrt{n(B_x)}} = \frac{C_2}{\sqrt{nh^d}}$$

where C_1, C_2 are two universal constants depending only on k and d.

The estimator

Based on the weights $\{w_\iota(x)\}$, construct a linear estimator

$$\hat{f}(x) = \sum_{\iota : x_\iota \in B_x} w_\iota(x) y_\iota$$

- Bias:

$$\begin{aligned} |b(x)| &= \left| \sum_{\iota : x_\iota \in B_x} w_\iota(x) f(x_\iota) - f(x) \right| \\ &= \inf_{p \in \mathcal{P}_d^k} \left| \sum_{\iota : x_\iota \in B_x} w_\iota(x)(f(x_\iota) - p(x_\iota)) - (f(x) - p(x)) \right| \\ &\leq (1 + \|w_\iota(x)\|_1) \inf_{p \in \mathcal{P}_k^d} \|f - p\|_{\infty, B_x} \end{aligned}$$

- Stochastic error:

$$|s(x)| = \sigma \left| \sum_{\iota : x_\iota \in B_x} w_\iota(x) \xi_\iota \right| \lesssim \frac{\sigma}{\sqrt{nh^d}} \cdot \left| \frac{1}{\|w_\iota(x)\|_2} \sum_{\iota : x_\iota \in B_x} w_\iota(x) \xi_\iota \right|$$

Rate of convergence

- Bounding the bias: by Taylor expansion,

$$\inf_{p\in\mathcal{P}_k^d}\|f-p\|_{\infty,B_x}\le\max_{z\in B_x}\left|\frac{D^kf(\eta_z)-D^kf(x)}{k!}(z-x)^k\right|\lesssim Lh^s$$

  Hence, $\|b\|_q\lesssim Lh^s$ for $1\le q\le\infty$
- Bounding the stochastic error: as before,

$$\left(\mathbb{E}\|s\|_q^q\right)^{\frac{1}{q}}\lesssim\frac{\sigma}{\sqrt{nh^d}}\quad(1\le q<\infty),\quad\mathbb{E}\|s\|_\infty\lesssim\sigma\sqrt{\frac{\ln n}{nh^d}}$$

Theorem

The optimal window size h^* and the corresponding risk is given by

$$h^*\asymp\begin{cases}n^{-\frac{1}{2s+d}}\\ \left(\frac{n}{\ln n}\right)^{-\frac{1}{2s+d}}\end{cases},\qquad R_q(\hat{f}^*,\mathcal{H}_d^s(L))\asymp\begin{cases}n^{-\frac{s}{2s+d}}, & 1\le q<\infty\\ \left(\frac{n}{\ln n}\right)^{-\frac{s}{2s+d}}, & q=\infty\end{cases}$$

and the resulting estimator is minimax rate-optimal (see later).

Why approximation?

The general linear estimator takes the form

$$\hat{f}_{\text{lin}}(x) = \sum_{\iota \in \{1,2,\cdots,m\}^d} w_\iota(x) y_\iota$$

Observations:

- The estimator $\hat{f}_{\text{lin}}$ is unbiased for f if and only if

$$f(x) = \sum_\iota w_\iota(x) f(x_\iota), \qquad \forall x \in [0,1]^d$$

- Plugging in $x = x_\iota$ yields that $\mathbf{z} = \{f(x_\iota)\}$ is a solution to $(w_\iota(x_\kappa) - \delta_{\iota\kappa})_{\iota,\kappa} \mathbf{z} = \mathbf{0}$
- Denote by $\{z_\iota^{(k)}\}_{k=1}^M$ all linearly independent solutions to the previous equation, then

$$f_k(x) = \sum_\iota w_\iota(x) z_\iota^{(k)}, \qquad k = 1, \cdots, M$$

constitutes an approximation basis for $\mathcal{H}_d^s(L)$.

Why polynomial?

Definition (Kolmogorov n-width)

For a linear normed space $(X, \|\cdot\|)$ and subset $K \subset X$, the Kolmogorov n-width of K is defined as

$$d_n(K) \equiv d_n(K, X) = \inf_{V_n} \sup_{x \in K} \inf_{y \in V_n} \|x - y\|$$

where $V_n \subset X$ has dimension $\leq n$.

Theorem (Kolmogorov n-width for $\mathcal{H}_d^s(L)$)

$$d_n(\mathcal{H}_d^s(L), C[0,1]^d) \asymp n^{-\frac{s}{d}}.$$

The piecewise polynomial basis in each cube with edge size h achieves the optimal rate (so other basis does not help):

$$d_{\Theta(1)h^{-d}}(\mathcal{H}_d^s(L), C[0,1]^d) \asymp (h^{-d})^{-\frac{s}{d}} = h^s$$

Methods in nonparametric regression

Function space (this lecture):

- Kernel estimates: $\hat{f}(x) = \sum_\iota K_h(x - x_\iota) y_\iota$ (Nadaraya, Watson, ...)
- Local polynomial kernel estimates: $\hat{f}(x) = \sum_{k=1}^{M} \phi_k(x) c_k(x)$, where $(c_1(x), \cdots, c_k(x)) = \arg\min_c \sum_\iota (y_\iota - \sum_{k=1}^{M} c_k(x)\phi_k(x_\iota))^2 K_h(x - x_\iota)$ (Stone, ...)
- Penalized spline estimates: $\hat{f} = \arg\min_g \|y_\iota - g\|_2^2 + \|g^{(s)}\|_2^2$ (Speckman, Ibragimov, Khas'minski, ...)
- Nonlinear estimates (Lepski, Nemirovski, ...)

Transformed space (next lecture):

- Fourier transform: projection estimates (Pinsker, Efromovich, ...)
- Wavelet transform: shrinkage estimates (Donoho, Johnstone, ...)

Regression in Sobolev space

Consider the following regression problem:

$$y_\iota = f(x_\iota) + \sigma\xi_\iota, \quad f \in \mathcal{S}_d^{k,p}(L), \iota = (i_1, i_2, \cdots, i_d) \in \{1, 2, \cdots, m\}^d$$

where

- $x_{(i_1, i_2, \cdots, i_d)} = (i_1/m, i_2/m, \cdots, i_d/m)$, $n = m^d$
- $\{\xi_\iota\}$ are iid $\mathcal{N}(0,1)$ noises
- The Sobolev ball $\mathcal{S}_d^{k,p}(L)$ is defined as

$$\mathcal{S}_d^{k,p}(L) = \left\{ f : [0,1]^d \to \mathbb{R}, \|D^k(f)\|_p \le L \right\}$$

  where $D^k(f)(\cdot)$ is the vector function comprised of all partial derivatives (in terms of distributions) of f with order k.

Parameters:

- d: dimensionality
- k: order of differentiation
- p: $p \ge d$ to ensure continuous embedding $\mathcal{S}_d^{k,p}(L) \subset C[0,1]^d$
- q: norm of the risk

Minimax lower bound

Theorem (Minimax lower bound)

The minimax risk in Sobolev ball regression problem over all estimators is

$$R_q(\mathcal{S}_d^{k,p}(L), n) \gtrsim \begin{cases} n^{-\frac{k}{2k+d}}, & q < (1 + \frac{2k}{d})p \\ \left(\frac{\ln n}{n}\right)^{\frac{k-d/p+d/q}{2(k-d/p)+d}}, & q \geq (1 + \frac{2k}{d})p \end{cases}$$

Theorem (Linear minimax lower bound)

The minimax risk in Sobolev ball regression problem over all linear estimators is

$$R_q^{lin}(\mathcal{S}_d^{k,p}(L), n) \gtrsim \begin{cases} n^{-\frac{k}{2k+d}}, & q \leq p \\ n^{-\frac{k-d/p+d/q}{2(k-d/p+d/q)+d}}, & p < q < \infty \\ \left(\frac{\ln n}{n}\right)^{\frac{k-d/p+d/q}{2(k-d/p+d/q)+d}}, & q = \infty \end{cases}$$

Start from linear estimates

Consider the linear estimator given by local polynomial approximation with window size h as before:

- Stochastic error:

$$\left(\mathbb{E}\|s\|_q^q\right)^{\frac{1}{q}} \lesssim \frac{\sigma}{\sqrt{nh^d}} \quad (1 \le q < \infty), \quad \mathbb{E}\|s\|_\infty \lesssim \sigma\sqrt{\frac{\ln n}{nh^d}}$$

- Bias: corresponds to polynomial approximation error

Fact

For $f \in \mathcal{S}_d^{k,p}(L)$, there exists constant $C > 0$ such that

$$|D^{k-1}f(x) - D^{k-1}f(y)| \le C|x-y|^{1-d/p}\left(\int_B |D^k f(z)|^p dz\right)^{\frac{1}{p}}, \quad x, y \in B$$

Linear estimator: bias

Upper bound of the bias: Taylor polynomial yields

$$|b(x)| \lesssim h^{k-d/p} \left(\int_{B_x} |D^k f(z)|^p dz \right)^{\frac{1}{p}}$$

$$\Longrightarrow \|b\|_q^q \lesssim h^{(k-d/p)q} \int_{[0,1]^d} \left(\int_{B_x} |D^k f(z)|^p dz \right)^{\frac{q}{p}} dx$$

Note that

$$\int_{[0,1]^d} \int_{B_x} |D^k f(z)|^p dz dx = h^d \int_{[0,1]^d} |D^k f(x)|^p dx \leq h^d L^p$$

- Case $q/p \leq 1$: $\|b\|_q^q \lesssim h^{(k-d/p)q} \cdot L^q h^{d \cdot \frac{q}{p}} = L^q h^{kq}$ (regular case)
- Case $q/p > 1$: $\|b\|_q^q \lesssim h^{(k-d/p)q} \cdot L^q h^d = L^q h^{(k-d/p+d/q)q}$ (sparse case)

Linear estimator: optimal risk

In summary, we have

$$\|b\|_q \lesssim \begin{cases} Lh^k, & q \le p \\ Lh^{k-d/p+d/q}, & p < q \le \infty \end{cases}, \quad \|s\|_q \lesssim \begin{cases} \frac{\sigma}{\sqrt{nh^d}}, & q < \infty \\ \sigma\sqrt{\frac{\ln n}{n}}, & q = \infty \end{cases}$$

Theorem (Optimal linear risk)

$$R_q(\hat{f}^*_{lin}, \mathcal{S}_d^{k,p}(L)) \lesssim \begin{cases} n^{-\frac{k}{2k+d}}, & q \le p \\ n^{-\frac{k-d/p+d/q}{2(k-d/p+d/q)+d}}, & p < q < \infty \\ \left(\frac{\ln n}{n}\right)^{\frac{k-d/p}{2(k-d/p)+d}}, & q = \infty \end{cases}$$

Alternative proof:

Theorem (Sobolev embedding)

For $d \le p < q$, we have $\mathcal{S}_d^{k,p}(L) \subset \mathcal{S}_d^{k-d/p+d/q,q}(L')$.

Minimax lower bound: tool

Theorem (Fano's inequality)

Suppose $H_1, \cdots, H_N$ are probability distributions (hypotheses) on sample space $(\Omega, \mathcal{F})$, and there exists decision rule $D : \Omega \to \{1, 2, \cdots, N\}$ such that $H_i(\omega : D(\omega) = i) \geq \delta_i, 1 \leq i \leq N$. Then

$$\max_{1 \leq i,j \leq N} D_{KL}(F_i \| F_j) \geq \left(\frac{1}{N} \sum_{i=1}^{N} \delta_i \right) \ln(N-1) - \ln 2$$

Apply to nonparametric regression:

- Suppose convergence rate r_n is attainable, construct N functions (hypotheses) $f_1, \cdots, f_N \in \mathcal{F}$ such that $\|f_i - f_j\|_q > 4r_n$ for any $i \neq j$
- Decision rule: after obtaining $\hat{f}$, choose j such that $\|\hat{f} - f_j\|_q \leq 2r_n$
- As a result, $\delta_i \geq 1/2$, and Fano's inequality gives

$$\frac{1}{2\sigma^2} \max_{1 \leq i,j \leq N} \sum_{\iota} |f_i(x_\iota) - f_j(x_\iota)|^2 \geq \frac{1}{2} \ln(N-1) - \ln 2$$

Minimax lower bound: sparse case

Suppose $q \geq (1 + \frac{2k}{d})p$.

- Fix a smooth function g supported on $[0,1]^d$
- Divide $[0,1]^d$ into h^{-d} disjoint cubes with size h, and construct $N = h^{-d}$ hypotheses: f_j supported on j-th cube, and equals $h^s g(x/h)$ on that cube (with translation)
- To ensure $f \in \mathcal{S}_d^{k,p}(L)$, set $s = k - d/p$
- For $i \neq j$, $r_n \asymp \|f_i - f_j\|_q \asymp h^{k-d/p+d/q}$, and

$$\sum_{\iota} |f_i(x_\iota) - f_j(x_\iota)|^2 \asymp h^{2(k-d/p)} \cdot nh^d = nh^{2(k-d/p)+d}$$

- Fano's inequality gives

$$nh^{2(k-d/p)+d} \gtrsim \ln N \asymp \ln h^{-1} \Longrightarrow r_n \gtrsim \left(\frac{\ln n}{n}\right)^{\frac{k-d/p+d/q}{2(k-d/p)+d}}$$

Minimax lower bound: regular case

Suppose $q < (1 + \frac{2k}{d})p$.

- Fix smooth function g supported on $[0, 1]^d$, and set $g_h(x) = h^s g(x/h)$
- Divide $[0, 1]^d$ into h^{-d} disjoint cubes with size h, and construct $N = 2^{h^{-d}}$ hypotheses: $f_j = \sum_{i=1}^{h^{-d}} \epsilon_i g_{h,i}$, where $g_{h,i}$ is the translation of g_h to i-th cube
- Can choose $M = 2^{\Theta(h^{-d})}$ hypotheses such that for $i \neq j$, f_i differs f_j on at least $\Theta(h^{-d})$ cubes
- To ensure $f \in \mathcal{S}_d^{k,p}(L)$, set $s = k$
- For $i \neq j$, $r_n \asymp \|f_i - f_j\|_q \asymp h^k$, and

$$\sum_{\iota} |f_i(x_\iota) - f_j(x_\iota)|^2 \asymp h^{2k} \cdot n = nh^{2k}$$

- Fano's inequality gives

$$nh^{2k} \gtrsim \ln M \asymp h^{-d} \Longrightarrow r_n \gtrsim \left(\frac{1}{n}\right)^{\frac{k}{2k+d}}$$

Construct minimax estimator

Why linear estimator fails in the sparse case?

- Too much variance in the flat region!
- Suppose we know that the true function f is supported at a cube with size h, we have

$$\|b\|_q \lesssim h^{k-d/p+d/q}, \qquad \left(\mathbb{E}\|s\|_q^q\right)^{1/q} \lesssim \frac{1}{\sqrt{nh^d}} \cdot h^{d/q}$$

then $h \asymp n^{-\frac{1}{2(k-d/p)+d}}$ yields the optimal risk $\Theta(n^{-\frac{k-d/p+d/q}{2(k-d/p)+d}})$

Nemirovski's construction

In both cases, construct $\hat{f}$ as the solution to the following optimization problem

$$\hat{f} = \arg\min_{g\in\mathcal{S}_d^{k,p}(L)} \|g-y\|_{\mathcal{B}} = \arg\min_{g\in\mathcal{S}_d^{k,p}(L)} \max_{B\in\mathcal{B}} \frac{1}{\sqrt{n(B)}} \Big|\sum_{\iota:x_\iota\in B}(g(x_\iota)-y_\iota)\Big|$$

where $\mathcal{B}$ is the set of all cubes in $[0,1]^d$ with nodes belonging to $\{x_\iota\}$.

Estimator analysis

Question 1: given a linear estimator $\hat{f}$ at x of window size h, which size h' of $B \in \mathcal{B}$ achieves the maximum of $|\sum_{\iota: x_\iota \in B} \hat{f}(x_\iota) - y_\iota| / \sqrt{n(B)}$?

- If $h' \ll h$, the value $\asymp \max\{\sqrt{n(B)}|b_h(x)|, |\mathcal{N}(0,1)|\}$ increases with h'
- If $h' \gg h$, the value is close to zero
- Hence, $h' \asymp h$ achieves the maximum

Question 2: what window size h^* at x achieves min $\|\hat{f}_h - y\|_{\mathcal{B}}$?

- Answer: h^* achieves the bias-variance balance locally at x
- The $1/\sqrt{n(B)}$ term helps to achieve the spatial homogeneity

Theorem (Optimality of the estimator)

$$R_q(\hat{f}, \mathcal{S}_d^{k,p}(L)) \lesssim \left(\frac{\ln n}{n}\right)^{\min\{\frac{k}{2k+d}, \frac{k-d/p+d/q}{2(k-d/p)+d}\}}$$

Hence $\hat{f}$ is minimax rate-optimal (with a logarithmic gap in regular case).

Proof of the optimality

First observation:

$$\|\hat{f} - f\|_{\mathcal{B}} \leq \|\hat{f} - y\|_{\mathcal{B}} + \|f - y\|_{\mathcal{B}} \leq 2\|\xi\|_{\mathcal{B}} \asymp \sqrt{\ln n}$$

and $e \triangleq \hat{f} - f \in \mathcal{S}_d^{k,p}(2L)$.

Definition (Regular cube)

A cube $B \subset [0,1]^d$ is called a regular cube if

$$e(B) \geq C[h(B)]^{k-d/p}\Omega(e, B)$$

where $C > 0$ is a suitably chosen constant, $e(B) = \max_{x \in B} |e(x)|$, $h(B)$ is the edge size of B, and

$$\Omega(e, B) = \left(\int_B |D^k f(x)|^p dx \right)^{\frac{1}{p}}.$$

One can show that $[0,1]^d$ can be (roughly) partitioned into maximal regular cubes with $\geq$ replaced by $=$ (i.e., balanced bias and variance).

Property of regular cubes

Lemma

If cube B is regular, we have

$$\sup_{B' \in \mathcal{B}, B' \subset B} \frac{1}{\sqrt{n(B')}} |\sum_{\iota : x_\iota \in B'} e(x_\iota)| \lesssim \sqrt{\ln n} \Longrightarrow e(B) \lesssim \sqrt{\frac{\ln n}{n(B)}}$$

Proof:

- Since B is regular, there exists a polynomial $e_k(x)$ in B of d variables and degree no more than k such that $e(B) \geq 4\|e - e_k\|_{\infty, B}$
- On one hand, $|e_k(x)| \leq 5e(B)/4$, and Markov's inequality for polynomial implies that $\|De_k\|_\infty \lesssim e(B)/h(B)$
- On the other hand, $|e_k(x_0)| \geq 3e(B)/4$ for some $x_0 \in B$, the derivative bound implies that there exists $B' \subset B, h(B') \asymp h(B)$ such that $|e_k(x)| \geq e(B)/2$ on B'
- Choosing this B' in the assumption completes the proof

Upper bound for the L_q risk

Since $[0,1]^d$ can be (roughly) partitioned into regular cubes $\{B_i\}_{i=1}^{\infty}$ such that $e(B_i) \asymp [h(B_i)]^{k-p/d}\Omega(e, B_i)$, we have

$$\|e\|_q^q = \sum_{i=1}^{\infty} \int_{B_i} |e(x)|^q dx \le \sum_{i=1}^{\infty} [h(B_i)]^d e^q(B_i)$$

The previous lemma asserts that $e(B_i) \le \sqrt{\frac{\ln n}{n[h(B_i)]^d}}$, and we cancel out $h(B_i), e(B_i)$ and get

$$\begin{aligned}
\|e\|_q^q &\lesssim \left(\frac{\ln n}{n}\right)^{\frac{q(k-d/p+d/q)}{2(k-d/p)+d}} \sum_{i=1}^{\infty} \Omega(e, B_i)^{\frac{d(q-2)}{2(k-d/p)+d}} \\
&\le \left(\frac{\ln n}{n}\right)^{\frac{q(k-d/p+d/q)}{2(k-d/p)+d}} \left(\sum_{i=1}^{\infty} \Omega(e, B_i)^p\right)^{\frac{d(q-2)}{p(2(k-d/p)+d)}}
\end{aligned}$$

if $q \ge q^* = (1 + \frac{2k}{d})p$. For $1 \le q < q^*$ we use $\|e\|_q \le \|e\|_{q^*}$. Q.E.D.

Data-driven window size

Consider again a linear estimate with window size h locally at x, recall that

$$|\hat{f}(x) - f(x)| \lesssim \inf_{p \in \mathcal{P}_d^k} \|f - p\|_{\infty, B_h(x)} + \frac{\sigma \mathcal{N}(0,1)}{\sqrt{nh^d}}$$

- The optimal window size $h(x)$ should balance these two terms
- The stochastic term can be upper bounded (with overwhelming probability) by $s_n(h) = w\sigma\sqrt{\frac{\ln n}{nh^d}}$ depending only on known parameters and h, where the constant $w > 0$ is large enough
- But the bias term depends on the unknown local property of f on $B_h(x)$!

Bias-variance tradeoff: revisit

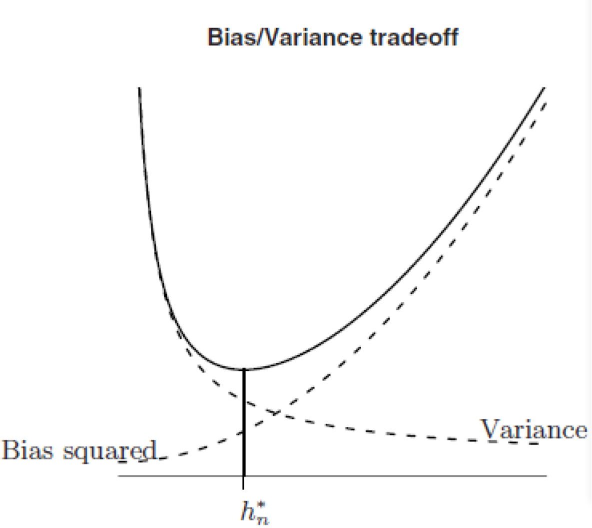

Figure 2: Bias-variance tradeoff

Lepski's trick

Lepski's adaptive scheme

Construct a family of local polynomial approximation estimators $\{\hat{f}_h\}$ with all window size $h \in (0,1)$. Then use $\hat{f}_{\hat{h}(x)}(x)$ as the estimate of $f(x)$, where

$$\hat{h}(x) \triangleq \sup\{h \in (0,1) : |\hat{f}_h(x) - \hat{f}_{h'}(x)| \le 4s_n(h'), \forall h' \in (0,h)\}.$$

Denote by $h^*(x)$ the optimal window size where two errors are equal: $b_n(h^*) = s_n(h^*)$

- Existence of $\hat{h}(x)$: clearly $h^*(x)$ satisfies the condition, for $h' \in (0, h^*)$

$$\begin{aligned}|\hat{f}_{h^*}(x) - \hat{f}_{h'}(x)| &\le |\hat{f}_{h^*}(x) - f| + |f - \hat{f}_{h'}(x)| \\ &\le b_n(h^*) + s_n(h^*) + b_n(h') + s_n(h') \le 4s_n(h'),\end{aligned}$$

- Performance of $\hat{h}(x)$:

$$\begin{aligned}|\hat{f}_{\hat{h}(x)}(x) - f| &\le |\hat{f}_{h^*(x)}(x) - f| + |\hat{f}_{\hat{h}(x)}(x) - \hat{f}_{h^*(x)}(x)| \\ &\le b_n(h^*) + s_n(h^*) + 4s_n(h^*) = 6s_n(h^*)\end{aligned}$$

Lepski's estimator is adaptive!

Properties of Lepski's scheme:

- Adaptive to parameters: agnostic to p, k, L
- Spatially adaptive: still work when there is spatial inhomogeneity / only estimate a portion of f

Theorem (Adaptive optimality)

Suppose that the modulus of smoothness k is upper bounded by a known hyper-parameter S:

$$R_q(\hat{f}, \mathcal{S}_d^{k,p}(L)) \lesssim \left(\frac{\ln n}{n}\right)^{\min\{\frac{k}{2k+d}, \frac{k-d/p+d/q}{2(k-d/p)+d}\}}$$

and $\hat{f}$ is adaptively optimal in the sense that

$$\inf_{\hat{f}} \sup_{k \le S, p > d, L > 0} \frac{R_q(\hat{f}, \mathcal{S}_d^{k,p}(L))}{\inf_{\hat{f}^*} R_q(\hat{f}^*, \mathcal{S}_d^{k,p}(L))} \gtrsim (\ln n)^{\frac{S}{2S+d}}.$$

Covering of the ideal window

By the property of the data-driven window size $\hat{h}(x)$, it suffices to consider the ideal window size $h^*(x)$ satisfying

$$s_n(h^*(x)) \asymp [h^*(x)]^{k-d/p}\Omega(f, B_{h^*(x)}(x))$$

Lemma

There exists $\{x_i\}_{i=1}^M$ and a partition $\{V_i\}_{i=1}^M$ of $[0,1]^d$ such that

1. Cubes $\{B_{h^*(x_i)}(x_i)\}_{i=1}^M$ are pairwise disjoint
2. For every $x \in V_i$, we have

$$h^*(x) \geq \frac{1}{2}\max\{h^*(x_i), \|x - x_i\|_\infty\}$$

and $B_{h^*(x)}(x) \cap B_{h^*(x_i)}(x_i) \neq \emptyset$

Analysis of the estimator

Using the covering of the local windows, we have

$$\|\hat{f} - f\|_q^q \lesssim \int_{[0,1]^d} |s_n(h^*(x))|^q dx = \sum_{i=1}^{M} \int_{V_i} |s_n(h^*(x))|^q dx$$

Recall: $s_n(h) \asymp \sqrt{\frac{\ln n}{nh^d}}$ and $h^*(x) \geq \max\{h^*(x_i), \|x - x_i\|_\infty\}/2$ for $x \in V_i$,

$$\begin{aligned}
\|\hat{f} - f\|_q^q &\lesssim \left(\frac{\ln n}{n}\right)^q \sum_{i=1}^{M} \int_{V_i} [\max\{h^*(x_i), \|x - x_i\|_\infty\}]^{-dq/2} dx \\
&\lesssim \left(\frac{\ln n}{n}\right)^q \sum_{i=1}^{M} \int_0^\infty r^{d-1} [\max\{h^*(x_i), r\}]^{-dq/2} dr \\
&\lesssim \left(\frac{\ln n}{n}\right)^q \sum_{i=1}^{M} [h^*(x_i)]^{d-dq/2}
\end{aligned}$$

Plugging in $s_n(h^*(x)) \asymp [h^*(x)]^{k-d/p} \Omega(f, B_{h^*(x)}(x))$ and use the disjointness of $\{B_{h^*(x_i)}(x)\}_{i=1}^M$ completes the proof. Q.E.D.

Experiment: Blocks

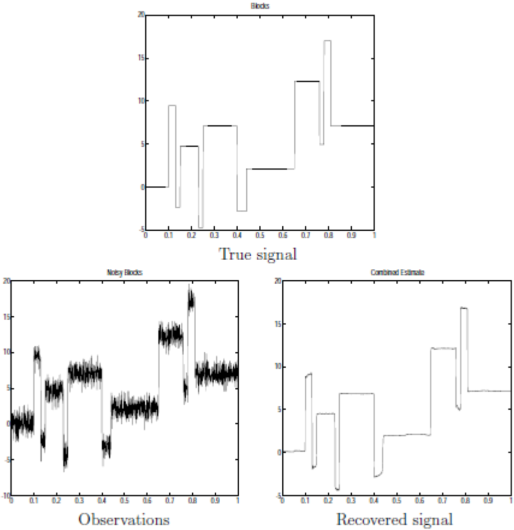

Figure 3: Blocks: original, noisy and reconstructed signal

Experiment: Bumps

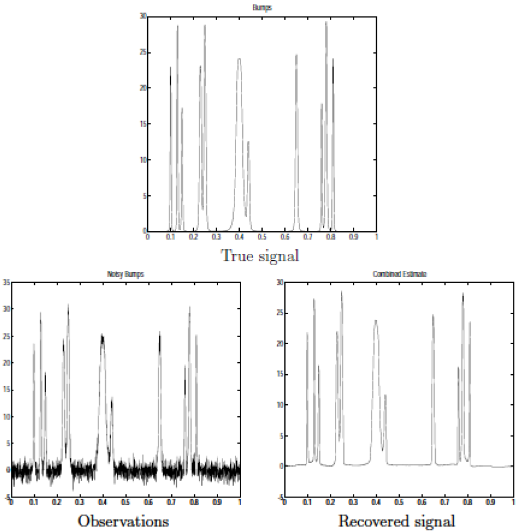

Figure 4: Bumps: original, noisy and reconstructed signal

Experiment: Heavysine

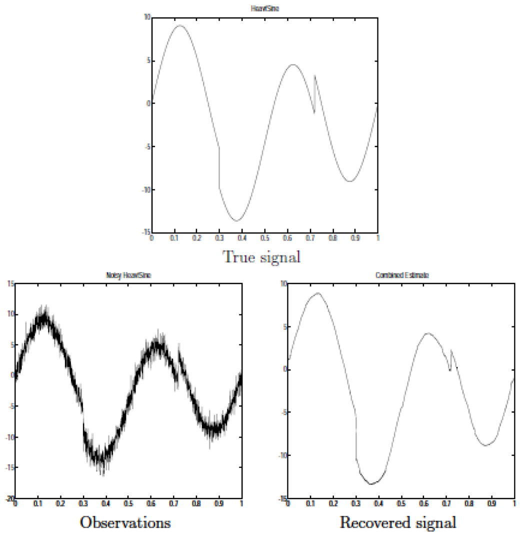

Figure 5: Heavysine: original, noisy and reconstructed signal

Experiment: Doppler

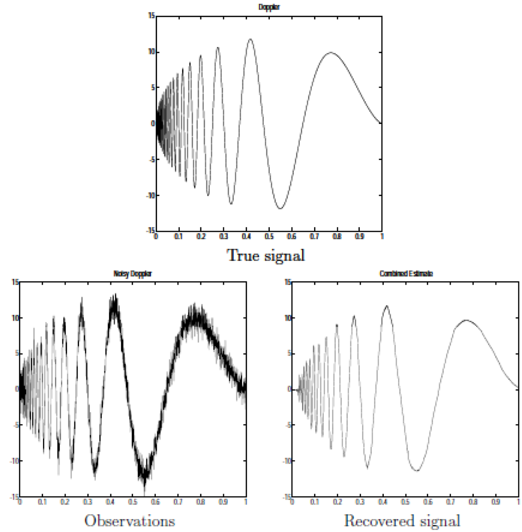

Figure 6: Doppler: original, noisy and reconstructed signal

Further generalization

Is (spatially local) Sobolev ball large enough?

- Do not include the very simple function $f(x) = \sin \omega x$ for ω large!
- Cannot recover "modulated signal", e.g., $f(x) = g(x)\sin(\omega x + \phi)$ for $g \in \mathcal{S}_d^{k,p}(L)$

Some observations:

- For $f(x) = \sin \omega x$, we have $\left(\frac{d^2}{dx^2} + \omega^2\right) f(x) = 0$
- For $f(x) = g(x)\sin(\omega x + \phi)$, we write $f(x) = f_+(x) + f_-(x)$, where

$$f_\pm(x) = \frac{g(x)\exp(\pm i(\omega x + \phi))}{\pm 2i}$$

and by induction we have

$$\left\| \left(\frac{d}{dx} \mp i\omega \right)^k f_\pm(x) \right\|_p = \frac{\|g^{(k)}\|_p}{2} \leq \frac{L}{2}$$

General regression model

Definition (Signal satisfying differential inequalities)

Function $f : [0,1] \to \mathbb{R}$ belongs to the class $\mathcal{W}^{l,k,p}(L)$, if and only if $f = \sum_{i=1}^{l} f_i$, and there exist monic polynomials $r_1, \cdots, r_l$ of degree k such that

$$\sum_{i=1}^{l} \left\| r_i\left(\frac{d}{dx}\right) f_i \right\|_p \le L.$$

- For example, $\mathcal{S}_1^{k,p}(L) \subset \mathcal{W}^{1,k,p}(L)$ with $r_1(z) = z^k$.

General regression problem: recover $f \in \mathcal{W}^{l,k,p}(L)$ from noisy observations

$$y_i = f(i/n) + \sigma\xi_i, \qquad i = 1, 2, \cdots, n, \qquad \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

where we only know:

- Noise level σ
- An upper bound $S \ge kl$

Recovering approach

The risk function

- Note that now we cannot recover the whole function f (e.g., $f \equiv 0$ and $f(x) = \sin(2\pi nx)$ correspond to the same model)
- The discrete q-norm: $\|\hat{f} - f\|_q = (\frac{1}{n}\sum_{i=1}^{n} |\hat{f}(i/n) - f(i/n)|^q)^{\frac{1}{q}}$

Recovering approach

- Discretization: transform differential inequalities to inequalities of finite difference
- Sequence estimation: given a window size, recover the discrete sequence satisfying an unknown difference inequality
- Adaptive window size: apply Lepski's trick to choose the optimal window size adaptively

Discretization: from derivative to difference

Lemma

For any $f \in C^{k-1}[0,1]$ and any monic polynomial $r(z)$ of degree k, there corresponds another monic polynomial $\eta(z)$ of degree k such that

$$\|\eta(\Delta) f_n\|_p \lesssim n^{-k+1/p} \|r(\frac{d}{dx}) f\|_p$$

where $f_n = \{f(i/n)\}_{i=1}^n$, and $(\Delta\phi)_t = \phi_{t-1}$ is the backward shift operator.

Applying to our regression problem:

- The sequence f_n can be written as $f_n = \sum_{i=1}^l f_{n,i}$
- There correspond monic polynomials η_i of degree k such that

$$\sum_{i=1}^{l} \|\eta_i(\Delta) f_{n,i}\|_p \lesssim L n^{-k+1/p}$$

Sequence estimation: window estimate

Given a sequence of observations $\{y_t\}$, consider using $\{y_t\}_{|t|\leq T}$ to estimate $f_n[0]$, where T is the window size

- The estimator: $\hat{f}_n[0] = \sum_{t=-T}^{T} w_{-t} y_t$
- On one hand, the filter $\{w_t\}_{|t|\leq T}$ should have a small L_2 norm to suppress the noise
- On the other hand, if $\eta(\Delta) f_n \equiv 0$ with a known η, the filter should be designed such that the error term only consists of the stochastic error

The approach

The filter $\{w_t\}_{|t|\leq T}$ is the solution to the following optimization problem:

$$\min \left\| \mathcal{F}\left(\left\{ \sum_{t=-T}^{T} w_{-t} y_{t+s} - y_s \right\}_{|s|\leq T} \right) \right\|_\infty \quad \text{s.t.} \quad \|\mathcal{F}(w_T)\|_1 \leq \frac{C}{\sqrt{T}}$$

where $\mathcal{F}(\{\phi_t\}_{|t|\leq T})$ denotes the discrete Fourier transform of $\{\phi_t\}_{|t|\leq T}$.

Frequency domain

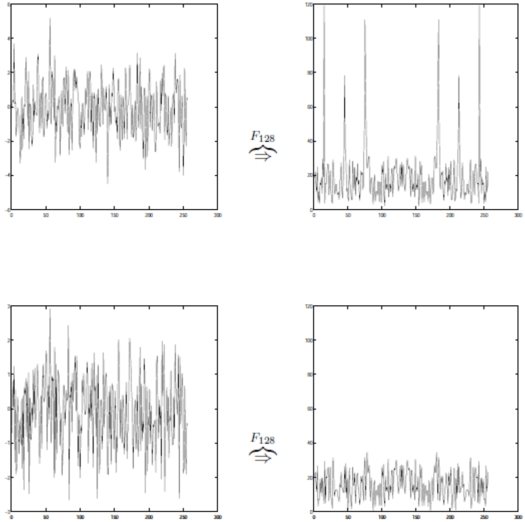

Figure 7: The observations (upper panel) and the noises (lower panel)

Performance analysis

Theorem

If $f_n = \sum_{i=1}^{l} f_{n,i}$ *and there exist monic polynomials* η_i *of degree* k *such that* $\sum_{i=1}^{l} \|\eta_i(\Delta) f_{n,i}\|_{p,T} \le \epsilon$, *we have*

$$|f_n[0] - \hat{f}_n[0]| \lesssim T^{k-1/p}\epsilon + \frac{\Theta^T(\xi)}{\sqrt{T}}$$

where $\Theta^T(\xi)$ *is the supremum of* $\mathcal{O}(T^2)$ $\mathcal{N}(0,1)$ *random variables and is thus of order* $\sqrt{\ln T}$.

- This result is a uniform result (η_i is unknown), and the $\sqrt{\ln T}$ gap is avoidable to achieve the uniformity.
- Plugging in $\epsilon \asymp n^{-k+1/p}\|f\|_{p,B}$ from the discretization step yields

$$|f_n[m] - \hat{f}_n[m]| \lesssim (\frac{T}{n})^{k-1/p}\|f\|_{p,B} + \frac{\Theta^T(\xi)}{\sqrt{T}}$$

where B is a segment with center m/n and length $\asymp T$.

Polynomial multiplication

Begin with the Hölder ball case, where we know that $\|(1-\Delta)^k f_n\|_\infty \le \epsilon$

- Write convolution as polynomial multiplication, we have $1 - w_T(z) = 1 - \sum_{|t|\le T} w_t z^t$ can be divided by $(1-z)^k$

$$(1 - w_T(\Delta))p = \frac{1 - w_T(\Delta)}{(1-\Delta)^k} \cdot (1-\Delta)^k p = 0, \quad \forall p \in \mathcal{P}_1^{k-1}$$

- Moreover, $\|w_T\|_1 \lesssim 1, \|w_T\|_2 \lesssim 1/\sqrt{T}$, and $\|\mathcal{F}(w_T)\|_1 \lesssim 1/\sqrt{T}$

Lemma

There exists $\{\eta_t\}_{|t|\le T}$ such that $\eta_T(z) = \sum_{|t|\le T} \eta_t z^t \equiv 1 - w_T^*(z)$ such that $\|\mathcal{F}(w_T^*)\|_1 \lesssim 1/\sqrt{T}$, and for each $i = 1, 2, \cdots, l$, we have

$$\eta_T(z) = \eta_i(z)\rho_i(z) \qquad \text{with} \qquad \|\rho_i\|_\infty \lesssim T^{k-1}$$

If we knew $\eta_T(z)$, we could just use $\hat{f}_n^*[0] = [w_T^*(\Delta)y]_0$ to estimate $f_n[0]$

Optimization solution

Performance of $\hat{f}_n^*$:

$$|f - \hat{f}_n^*| = |(1 - w_T^*(\Delta))f + w_T^*(\Delta)\xi| \leq |\eta_T(\Delta)f| + |w_T^*(\Delta)\xi|$$
$$\leq \sum_{i=1}^{I} |\eta_T(\Delta)f_i| + |w_T^*(\Delta)\xi| = \sum_{i=1}^{I} |\rho_i(\Delta)(\eta_i(\Delta)f_i)| + |w_T^*(\Delta)\xi|$$

Fact

$$\|\mathcal{F}(f - \hat{f}_n^*)\|_\infty \lesssim \sqrt{T}\left(T^{k-1/p}\epsilon + \frac{\Theta_T(\xi)}{\sqrt{T}}\right), \quad \|\eta_T(\Delta)f\|_\infty \lesssim T^{k-1/p}\epsilon$$

Observation: by definition $\|\mathcal{F}(f - \hat{f}_n)\|_\infty \leq \|\mathcal{F}(f - \hat{f}_n^*)\|_\infty$, thus

$$\|\mathcal{F}((1 - w_T(\Delta))f)\|_\infty \leq \|\mathcal{F}(f - \hat{f}_n)\|_\infty + \|\mathcal{F}(w_T(\Delta)\xi)\|_\infty$$
$$\lesssim \sqrt{T}\left(T^{k-1/p}\epsilon + \frac{\Theta_T(\xi)}{\sqrt{T}}\right)$$

Performance analysis

Stochastic error:

$$|s| = |[w_T(\Delta)\xi]_0| \leq \|\mathcal{F}(w_T)\|_1 \|\mathcal{F}(\xi)\|_\infty \lesssim \frac{\Theta_T(\xi)}{\sqrt{T}}$$

Bias:

$$\begin{aligned}
|b| &= |[(1 - w_T(\Delta))f]_0| \\
&\leq |[(1 - w_T(\Delta))\eta_T(\Delta)f]_0| + |[(1 - w_T(\Delta))w_T^*(\Delta)f]_0| \\
&\leq \|\eta_T(\Delta)f\|_\infty + \|\mathcal{F}(\eta_T(\Delta)f)\|_\infty \| \|\mathcal{F}(w_T)\|_1 \\
&\qquad\qquad + \|\mathcal{F}((1 - w_T(\Delta))f)\|_\infty \|\mathcal{F}(w_T^*)\|_1 \\
&\lesssim T^{k-1/p}\epsilon + \frac{\Theta_T(\xi)}{\sqrt{T}}
\end{aligned}$$

and we're done. Q.E.D.

Adaptive window size

Apply Lepski's trick to select $\hat{T}[m] = n\hat{h}[m]$:

$$\hat{h}[m] = \sup\{h \in (0,1) : |\hat{f}_{n,nh}[m] - \hat{f}_{n,nh'}[m]| \leq C\sqrt{\frac{\ln n}{nh'}}, \forall h' \in (0,h)\}$$

Theorem

Suppose $S \geq kl$, we have

$$R_q(\hat{f}, \mathcal{W}^{l,k,p}(L)) \lesssim \left(\frac{\ln n}{n}\right)^{\min\left\{\frac{k}{2k+1}, \frac{k-1/p+1/q}{2(k-1/p)+1}\right\}}$$

and $\hat{f}$ is adaptively optimal in the sense that

$$\inf_{\hat{f}} \sup_{kl \leq S, p>0, L>0} \frac{R_q(\hat{f}, \mathcal{W}^{l,k,p}(L))}{\inf_{\hat{f}^*} R_q(\hat{f}^*, \mathcal{W}^{l,k,p}(L))} \gtrsim (\ln n)^{\frac{S}{2S+1}}.$$

References

Nonparametric regression:

- A. Nemirovski, *Topics in Non-Parametric Statistics. Lectures on probability theory and statistics, Saint Flour 1998, 1738.* 2000.
- A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.
- L. Györfi, et al. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Approximation theory:

- R. A. DeVore and G. G. Lorentz, *Constructive approximation*. Springer Science & Business Media, 1993.
- G. G. Lorentz, M. von Golitschek, and Y. Makovoz, *Constructive approximation: advanced problems*. Berlin: Springer, 1996.

Function space:

- O. V. Besov, V. P. Il'in, and S. M. Nikol'ski, *Integral representations of functions and embedding theorems*. Moscow: Nauka Publishers, 1975 (in Russian).