Lec 12: Wavelet Thresholding

Yanjun Han

Dec. 5, 2023

Last lecture: for nonparametric regression, we may find basis functions $\{\phi_i(x)\}_{i=1}^m$ such that $f(x) \approx \sum_{i=1}^m \theta_i \phi_i(x)$

(choice of basis: polynomials, splines, $\cdots$)

This lecture: how about orthonormal basis?

$$\left(\int \phi_i(x)\phi_j(x)\,dx = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \right)$$

Gaussian sequence model: throughout this lecture we assume that

$$x_i = \frac{i}{n}, \qquad y_i \sim N(f(x_i), \sigma_0^2).$$

Let $\{\phi_i(x)\}_{i=1}^\infty$ be a complete orthonormal basis of $L_2[0,1]$, i.e.

$$f(x) = \sum_{i=1}^\infty \theta_i \phi_i(x),$$

where

$$\boxed{\theta_i = \int_0^1 f(x)\phi_i(x)\,dx.}$$

(Pf:

$$\begin{aligned}\int_0^1 f(x)\phi_i(x)\,dx &= \int_0^1 \Big(\sum_{j=1}^\infty \theta_j \phi_j(x) \Big) \phi_i(x)\,dx \\ &= \sum_{j=1}^\infty \theta_j \int_0^1 \phi_i(x)\phi_j(x)\,dx \\ &= \sum_{j=1}^\infty \begin{cases} \theta_j & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} = \theta_i \qquad \square \;)\end{aligned}$$

How to estimate θ_i? Try

$$\begin{aligned} z_i = \frac{1}{n}\sum_{j=1}^n \phi_i(x_j) y_j &= \frac{1}{n}\sum_{j=1}^n \phi_i(x_j)\big(f(x_j) + \sigma_0 \xi_j\big) \quad (\xi_j \sim N(0,1)) \\ &= \frac{1}{n}\sum_{j=1}^n \phi_i(x_j) f(x_j) + \sigma_0 \cdot \frac{1}{n}\sum_{j=1}^n \phi_i(x_j)\xi_j \\ &\approx \int_0^1 \phi_i(x) f(x)\,dx + \sigma_0 \cdot N\Big(0, \frac{1}{n}\int \phi_i(x)^2\,dx\Big) \\ &= \theta_i + N\Big(0, \frac{\sigma_0^2}{n}\Big) \end{aligned}$$

This approx. error is often negligible; we do not consider it in this lecture.

Therefore, instead of observing $(x_i, y_i)_{i=1}^n$, we can equivalently assume that we observe $(z_1, z_2, \cdots)$ s.t.

$$\boxed{z_i \overset{\text{ind.}}{\sim} N\left(\theta_i, \frac{\sigma_0^2}{n}\right), \quad i=1,2,\cdots}$$

Remark: 1) z_i's are (approx.) independent as

$$\begin{aligned} Cov(z_i, z_j) &= Cov\left(\frac{1}{n}\sum_{k=1}^n \phi_i(x_k) y_k, \frac{1}{n}\sum_{k=1}^n \phi_j(x_k) y_k\right) \\ &= \frac{\sigma_0^2}{n^2}\sum_{k=1}^n \phi_i(x_k)\phi_j(x_k) \\ &\approx \frac{\sigma_0^2}{n}\int_0^1 \phi_i(x)\phi_j(x)\,dx = 0 \quad \text{for } i \neq j ; \end{aligned}$$

2) if we find an estimator $\hat\theta = (\hat\theta_1, \hat\theta_2, \cdots)$ for θ, then we should estimate f by

$$\hat f(x) = \sum_{i=1}^\infty \hat\theta_i \phi_i(x),$$

and

$$\begin{aligned} \|\hat f - f\|_2^2 &= \int_0^1 (\hat f(x) - f(x))^2 dx \\ &= \int_0^1 \left(\sum_{i=1}^\infty (\hat\theta_i - \theta_i)\phi_i(x)\right)^2 dx \\ &= \sum_{i=1}^\infty \sum_{j=1}^\infty (\hat\theta_i - \theta_i)(\hat\theta_j - \theta_j) \underbrace{\int_0^1 \phi_i(x)\phi_j(x)\,dx}_{=\mathbb{1}(i=j)} \\ &= \sum_{i=1}^\infty (\hat\theta_i - \theta_i)^2 \\ &= \|\hat\theta - \theta\|_2^2 \quad \text{(Plancherel / Parseval identity)} \end{aligned}$$

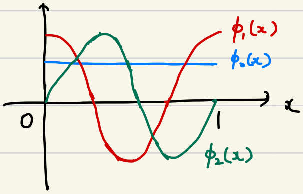
(estimation of f $\iff$ estimation of θ)

Choice I of $\{\phi_i(x)\}_{i=1}^\infty$: Fourier basis

Fourier basis of $L_2[0,1]$:

$\phi_0(x) = 1$, $\phi_{2j-1}(x) = \sqrt{2}\cos(2\pi j t)$, $\phi_{2j}(x) = \sqrt{2}\sin(2\pi j t)$

One can check $\{\phi_i(x)\}_{i=0}^\infty$ are indeed orthonormal.



Here, $\theta_i = \int_0^1 f(x)\phi_i(x)\,dx$: Fourier coefficients of f

$z_i = \frac{1}{n}\sum_{j=1}^{n} f(x_j)\phi_i(x_j)$: discrete Fourier transform of $(y_i)_{i=1}^{n}$

Estimation in Sobolev space.

$$H^k(L) = \left\{ f \in L^2[0,1] : \int_0^1 |f^{(k)}(x)|^2\,dx \le L^2 \right\}$$

average notion of smoothness

Theorem. $f \in H^k(L)$ if and only if its Fourier coefficients $(\theta_0, \theta_1, \cdots)$ satisfies

$$\sum_{j=1}^{\infty} (2\pi j)^{2k} \left(\theta_{2j-1}^2 + \theta_{2j}^2 \right) \le L^2$$

Intuition: smoothness in time domain $\Longleftrightarrow$ tail in frequency domain

Estimator I: Fourier projection estimator.

$$\hat{\theta}_i = \begin{cases} z_i & \text{if } i \le m \\ 0 & \text{if } i > m \end{cases}$$

Analysis:

$$\mathbb{E}\|\hat{\theta} - \theta\|_2^2 = \sum_{i=0}^{m} \mathbb{E}(z_i - \theta_i)^2 + \sum_{i>m} (0 - \theta_i)^2$$

$$= (m+1)\frac{\sigma_0^2}{n} + \sum_{i>m} \theta_i^2$$

$$\le \frac{(m+1)\sigma_0^2}{n} + \frac{1}{(\pi m)^{2k}} \underbrace{\sum_{i>m} (\pi i)^{2k} \theta_i^2}_{\le L^2}$$

$$= O\left(\frac{m}{n} + \frac{1}{m^{2k}}\right)$$

Choosing $m \asymp n^{\frac{1}{2k+1}}$ gives ($m \nearrow$, bias $\searrow$, var $\nearrow$)

$$\mathbb{E}\|\hat{f} - f\|_2^2 = \mathbb{E}\|\hat{\theta} - \theta\|_2^2 = O\left(n^{-\frac{2k}{2k+1}}\right).$$

Estimator II: optimal linear estimator (optional)

Set $\hat{\theta}_i = c_i z_i$ with $c_i \in [0,1]$

Then
$$\mathbb{E}\|\hat{\theta}-\theta\|_2^2 = \sum_{i=0}^{\infty} \mathbb{E}(c_i z_i - \theta_i)^2 = \sum_{i=0}^{\infty}\left[(1-c_i)^2\theta_i^2 + c_i^2 \cdot \frac{\sigma_0^2}{n}\right]$$

Choose $\{c_i\}_{i=0}^{\infty}$ to solve the following min-max program:

$$\min_{\{c_i\}_{i=0}^{\infty}} \quad \max_{\{\theta_i\}_{i=0}^{\infty} : \sum_{j=1}^{\infty}(2\pi j)^{2k}(\theta_{2j-1}^2+\theta_{2j}^2)\le L^2} \quad \sum_{i=0}^{\infty}\left[(1-c_i)^2\theta_i^2 + c_i^2\frac{\sigma_0^2}{n}\right]$$

Pinsker's Theorem: the above min-max program exhibits an explicit solution, and the resulting estimator attains $(1+o(1))\cdot$ minimax risk.

Problem with Fourier:

1) estimators become suboptimal for other Sobolev balls
$$W^{k,p}(L) = \left\{ f \in L^p[0,1] : \int_0^1 |f^{(k)}(x)|^p dx \le L^p \right\}$$
for large p, or general Besov balls;

2) estimators become suboptimal when f has spatial inhomogeneity;

3) any linear estimator suffers from the same problem.

Solution: wavelets!

Choice II of $\{\phi_i(x)\}_{i=1}^{\infty}$: Wavelets

Definition: Idea: multiresolutional analysis

A wavelet basis consists of a father wavelet $\phi(x)$ and a mother wavelet $\psi(x)$ on $[0,1]$, s.t. if

$$V_j = \text{span}\{\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k) : 0 \le k \le 2^j - 1\}$$

$$W_j = \text{span}\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k) : 0 \le k \le 2^j - 1\}$$

then:

1) $V_{j+1} = V_j \oplus W_j$ $(\Rightarrow V_{j+s} = V_j \oplus W_{j+1} \oplus W_{j+2} \oplus \cdots \oplus W_{j+s-1})$

2) $L^2[0,1] = \overline{V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots}$ (spans all function on $[0,1]$)

3) $\{\phi_{j_0k} : 0 \le k \le 2^{j_0}-1\}$ & $\{\psi_{jk} : j \ge j_0, 0 \le k \le 2^j - 1\}$ are orthonormal.

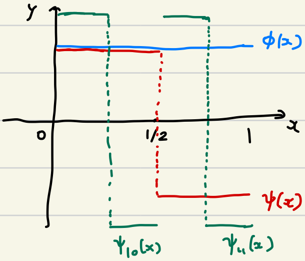
By 2) & 3), any $f \in L^2[0,1]$ can be written as

$$f(x) = \underbrace{\sum_{k=0}^{2^{j_0}-1} \alpha_{j_0k}\phi_{j_0k}(x)}_{\text{Gross information}} + \sum_{j \ge j_0} \underbrace{\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x)}_{\text{Detailed information at level } j}$$

Example of wavelets:

Haar wavelets. $\phi(x) = 1(x \in [0,1])$

$$\psi(x) = \begin{cases} 1, & \text{if } x \in [0, \frac{1}{2}) \\ -1, & \text{if } x \in [\frac{1}{2}, 1] \end{cases}$$



(j: resolution parameter; $|\text{supp}(\phi_{jk})| = |\text{supp}(\psi_{jk})| = 2^{-j}$; $k \in \{0, 1, \cdots, 2^j - 1\}$: spatial location)

Meyer wavelets. All moments vanish, but infinite support in time domain

Daubechies wavelets. Vanishing moments up to desired order + compactly supported (most widely used wavelets)

Cohen-Daubechies-Feauveau wavelet: used in JPEG 2000 standard.

Estimation: wavelet thresholding

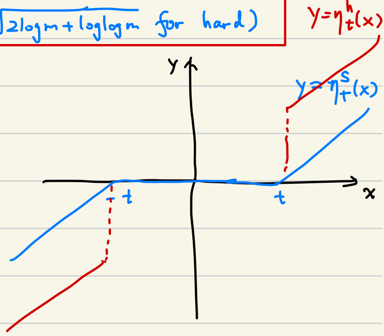
Soft and hard thresholding: consider Gaussian sequence model

$$z_i \sim N(\theta_i, \sigma^2), \quad i=1,\cdots,m.$$

Soft-thresholding estimator: $\hat{\theta}_i^s = \eta_t^s(z_i) = \mathrm{sign}(z_i)\cdot(|z_i|-t)_+$

Hard-thresholding estimator: $\hat{\theta}_i^h = \eta_t^h(z_i) = z_i \cdot \mathbb{1}(|z_i| \geq t)$

(choice of thresholds: $t=\sigma\sqrt{2\log m}$ for soft, $t=\sigma\sqrt{2\log m+\log\log m}$ for hard)



Intuition: when z is small, think of $\theta \approx 0$;
when z is large, think of $\theta \approx z$.

Property (pf omitted): thresholding estimators are optimal when θ has "sparse" structures

Wavelet thresholding: choose $j_0 \sim 1$, $j_1 \sim \log n$, use wavelet transform to obtain

$$\begin{cases} \hat{\alpha}_{j_0k} \sim N(\alpha_{j_0k}, \frac{\sigma_0^2}{n}), & 0 \leq k \leq 2^{j_0}-1 \\ \hat{\beta}_{jk} \sim N(\beta_{jk}, \frac{\sigma_0^2}{n}), & j_0 \leq j < j_1, \ 0 \leq k \leq 2^j-1. \end{cases}$$

Wavelet thresholding estimator:

$$\tilde{\alpha} = \hat{\alpha}, \qquad \tilde{\beta} = \eta_t^s(\hat{\beta}) \text{ or } \eta_t^h(\hat{\beta}),$$

and estimate f by

$$\tilde{f}(x) = \sum_{k=0}^{2^{j_0}-1} \tilde{\alpha}_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{j_1-1}\sum_{k=0}^{2^j-1} \tilde{\beta}_{jk}\psi_{jk}(x)$$

(inverse wavelet transform)

Diagram: $y \xrightarrow{\text{wavelet transform}} z \xrightarrow{\text{soft/hard thresholding}} \hat{\theta} \xrightarrow{\text{inverse wavelet transform}} \hat{f}$

Choice of threshold t:

Option I: estimate noise level σ_0, use the theory prediction (VisuShrink)

Option II: use cross validation or unbiased risk estimate to choose t

(SureShrink and others)

Properties: 1) optimal in all Sobolev and Besov classes;

2) adaptive to smoothness and local inhomogeneity of f;

(non-linearity of estimator plays a key role here)

3) Easy to implement using fast wavelet transforms.

Why wavelet thresholding (option-I)?

1) why wavelets? ⟶ representation power of wavelets

Wavelet is an unconditional basis for Sobolev or Besov norms $\|\cdot\|$,

i.e. for every $\varepsilon_i \in [-1,1]$,

$$\left\| \sum_i \varepsilon_i \phi_i \right\| \le C \left\| \sum_i \phi_i \right\|$$

(Fourier basis does not satisfy this property)

2) why thresholding? ⟶ idea of "shrinkage"

Similar to the James-Stein estimator, thresholding introduces a small bias to significantly reduce the variance

(HW 8: mimic the performance of the "ideal truncation estimator"

$$\hat{\theta}_i^{ITE} = z_i \cdot \mathbb{1}(|\theta_i| \ge \sigma)$$)