EE290 Mathematics of Data Science Lecture 20 & 21 - 11/14/2019

Lecture 20 and 21: Wavelet Shrinkage

Lecturer: Jiantao Jiao *Scribe: Yanjun Han and Jiantao Jiao*

This lecture is based partially on EE378A taught at Stanford by Tsachy Weissman and Jiantao Jiao.

1 Recap

In the last lecture, we considered the nonparametric function estimation problem in the regression setting. In particular, we showed that the kernel-based estimator with a suitable-chosen adaptive bandwidth essentially achieves the minimax risk over Hölder and Sobolev balls. Specifically, the estimator is constructed as follows:

1. Fix some kernel (or weight function) with suitable regularity conditions (e.g., keep polynomials up to a prescribed degree), and use this kernel to construct a linear estimator $\hat{f}_h(x)$ for any point $x \in [0,1]$ and bandwidth $h > 0$;
2. For any point $x \in [0,1]$, use some rule (e.g., Lepski's trick) to choose an adaptive bandwidth $\hat{h}(x)$;
3. Finally, for any point $x \in [0,1]$, use $\hat{f}_{\hat{h}(x)}(x)$ as the estimator for $f(x)$.

We also recall the Lepski's trick: for any $x \in [0,1]$, we pick up a set consisting of "admissible" bandwidths:

$$A = \{h \in [0,1] : |\hat{f}_h(x) - \hat{f}_{h'}(x)| \le 4s_{h'}(x), \quad \forall h' \in (0,h)\} \tag{1}$$

and then choose $\hat{h}(x) = \max A$. We validate this choice via two steps: we show that the optimal bandwidth h^* which balances the bias and variance locally belongs to this set, and then we show that $\hat{h}$ also works by relating $\hat{h}$ to the unknown h^*. This idea can also be generalized to high-dimensional cases with caution, and you may look at Homework 6 for details.

The main insights of the previous approach are two-fold:

1. The bias-variance tradeoff should be understood and analyzed carefully: this step determines the choice of the optimal bandwidth;
2. "A little bit" non-linearity needs to be added to the estimator: we have shown in the previous lecture that all linear approaches may fail to give the order-optimal risk in certain models, while an "almost" linear estimator with "a little bit" non-linearity can succeed. In the previous example, we almost use a linear kernel-based estimator, and the only additional non-linearity is to choose the bandwidth differently at different points.

In this lecture, we will attack the same problem using a different approach called *wavelet shrinkage*, where we can see the same phenomena from a different viewpoint.

2 Gaussian White Noise Model and Change of Basis

In the last lecture we have looked at the regression problem in Gaussian noise, and also remarked that the same idea can also be used in the density estimation setting, where the kernel-based estimator is called the KDE (Kernel Density Estimator). In this lecture we will look at the Gaussian White Noise Model, and remark that this is essentially the same as the previous two models.

2.1 Equivalences between Models

We call two statistical models are *equivalent* if they can almost simulate each other. An equivalent formulation is that, for *any* bounded loss function and *any* objective to be estimated, if there is an estimator $\hat{f}_1(x)$ for one model, we can construct another estimator $\hat{f}_2(y)$ for the other model whose estimation performance (i.e., the risk) is almost no worse than that of $\hat{f}_1(x)$ under any true parameter $\theta \in \Theta$, and vice versa. Intuitively, if two models are equivalent, for estimation purposes it suffices to look at any one of them. For a rigorous treatment, we refer interested readers to the Le Cam distance[1].

We will consider four different models in the context of nonparametric estimation:

1. **Regression Model**: we observe n samples $\{y_i\}_{i=1}^n$ with $y_i = f(x_i) + \sigma\xi_i$ for $1 \le i \le n$, where $x_i = \frac{i}{n}$ and $\xi_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$;

2. **Gaussian White Noise Model**: we observe a process $(Y_t)_{t\in[0,1]}$ with

$$Y_t = \int_0^t f(s)ds + \frac{\sigma}{\sqrt{n}}B_t, \qquad t \in [0,1] \tag{2}$$

   where $(B_t)_{t\in[0,1]}$ is the standard Brownion Motion on $[0,1]$. This model can also be written in the stochastic differential equation (SDE) form as $dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}}dB_t$. The reason for the $1/\sqrt{n}$ scaling is that for interval of length $1/n$, the signal in the white noise model is about $\frac{f(t)}{n}$, while the noise in the model is $\frac{\sigma}{\sqrt{n}}\mathcal{N}(0,1/n)$, which after normalization reduces to signal $f(t)$ noise $\mathcal{N}(0,\sigma^2)$.

3. **Density Estimation Model**: we observe n i.i.d samples $y_1, \cdots, y_n \overset{i.i.d}{\sim} g(\cdot)$, where the density g is supported on $[0,1]$;

4. **Poisson Process Model**: we observe a Poisson process $(Y_t)_{t\in[0,1]}$ with a time-varying intensity $ng(\cdot)$, where the density function g is supported on $[0,1]$.

In each model, there is some unknown function/density treated as the unknown parameter, and we observe discrete samples or a stochastic process. Suppose that the parameter space is $f \in \mathcal{F}$, where $\mathcal{F}$ is some function class possessing certain order of smoothness. The main result is that these four models are equivalent:

Theorem 1. [2] [3] *Under certain technical conditions, these four models are asymtotically equivalent as $n \to \infty$, with $g = f^2, \sigma = \frac{1}{2}$ when talking about the last two models.*

2.2 Change of Basis

Due to the model equivalence, in this lecture we consider the Gaussian white noise model

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}}dB_t, \qquad t \in [0,1] \tag{3}$$

and our target is to find an estimator $\hat{f}$ which comes close to minimizing the worst-case squared-error risk

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f \|\hat{f} - f\|_2^2. \tag{4}$$

Now we take a look at how the model and the loss function behave after we transform the problem into a different domain.

[1] see, e.g., F. Liese and K. J. Miescke, *Statistical Decision Theory: Estimation, Testing and Selection.* Springer, 2008.

[2] L. D. Brown, and M. G. Low. *Asymptotic equivalence of nonparametric regression and white noise.* The Annals of Statistics, 24(6), pp. 2384–2398, 1996.

[3] L. D. Brown, A. V. Carter, M. G. Low, and C. H. Zhang, *Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift.* The Annals of Statistics, 32(5), pp. 2074–2097, 2004.

Let $(\phi_j)_{j=1}^{\infty}$ be an orthonormal basis of $L^2[0,1]$, then we can represent the function f via its coefficients $\theta = (\theta_j)_{j=1}^{\infty}$, where

$$\theta_j \triangleq \int_0^1 \phi_j(t) f(t) dt. \tag{5}$$

Note that the restriction $f \in \mathcal{F}$ will be transformed into the condition $\theta \in \Theta$ for some proper parameter set Θ. Also, for any estimator $\hat{f}$, we can also represent it by

$$\hat{\theta}_j \triangleq \int_0^1 \phi_j(t) \hat{f}(t) dt. \tag{6}$$

As for the observation, we may define

$$\begin{aligned} y_j &\triangleq \int_0^1 \phi_j(t) dY_t && (7)\\ &= \int_0^1 \phi_j(t) \left(f(t)dt + \frac{\sigma}{\sqrt{n}} dB_t \right) && (8)\\ &= \int_0^1 \phi_j(t) f(t) dt + \frac{\sigma}{\sqrt{n}} \int_0^1 \phi_j(t) dB_t && (9)\\ &\equiv \theta_j + \epsilon \cdot \xi_j && (10) \end{aligned}$$

where $\epsilon \triangleq \frac{\sigma}{\sqrt{n}}$ denotes the noise level, and

$$\xi_j \triangleq \int_0^1 \phi_j(t) dB_t. \tag{11}$$

Exercise 2. Use the orthonormality of $(\phi_j)_{j=1}^{\infty}$ to conclude that $\xi_j \overset{i.i.d}{\sim} \mathcal{N}(0,1)$.

By the previous exercise, we know that the Gaussian white noise model reduces to the following *Gaussian sequence model* under the basis $(\phi_j)_{j=1}^{\infty}$:

$$y_j = \theta_j + \epsilon \xi_j, \qquad \theta = (\theta_1, \theta_2, \cdots) \in \Theta, \quad \epsilon = \frac{\sigma}{\sqrt{n}}, \quad \xi_j \overset{i.i.d}{\sim} \mathcal{N}(0,1). \tag{12}$$

Also, $\|\hat{\theta} - \theta\|_2 = \|\hat{f} - f\|_2$ by Parseval's identity, our target becomes to find some $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \cdots)$ which comes close to minimize

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = \sup_{\theta \in \Theta} \sum_{j=1}^{\infty} \mathbb{E}_\theta (\hat{\theta}_j - \theta_j)^2. \tag{13}$$

As a result, under the basis change, we operate in a Gaussian sequence model and would like to estimate the mean vector simultaneously.

3 Besov Ball and Wavelet Basis

Before introducing how to add the non-linearity in the transformed domain, first we specify the choice of the function class $\mathcal{F}$ and the orthonormal basis $(\phi_j)_{j=1}^{\infty}$. Specifically, we will choose $\mathcal{F}$ to be the Besov ball $\mathcal{B}_{p,q}^s(L)$, and $(\phi_j)_{j=1}^{\infty}$ to be the wavelet basis. To avoid technicality, we will treat them informally and only talk about the insights behind these concepts, and refer interested readers to the following reference:

- W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, approximation, and statistical applications*. Springer Science & Business Media, Vol. 129, 2012.

3.1 Introduction to Besov Ball

Like the Hölder and Sobolev balls, the Besov ball is another ball which characterizes the smoothness of a function in a more delicate and complicated way. Specifically, for any $s > 0, p, q \in [1, \infty]$, we may define a norm $\|\cdot\|_{\mathcal{B}^s_{p,q}}$ which is somehow (informally) close to

$$\|f\|_{\mathcal{B}^s_{p,q}} \approx \|f^{(s)}\|_p \tag{14}$$

where $f^{(s)}$ is the order-s derivative of f. Note that s may not be an integer, but in the above definition we have a proper definition of a fractional derivative, and the parameter q also affects the definition of the derivative slightly. Intuitively, $\|\cdot\|_{\mathcal{B}^s_{p,q}}$ is another norm which characterizes the order-s smoothness.

Naturally, the Besov ball $\mathcal{B}^s_{p,q}(L)$ is defined by

$$\mathcal{B}^s_{p,q}(L) \triangleq \{f : \|f\|_{\mathcal{B}^s_{p,q}} \le L\}. \tag{15}$$

3.2 Introduction to Wavelet Basis

The wavelet basis is an orthonormal basis which exploits the idea of *multi-resolution analysis*: any function is viewed from multiple resolutions. Specifically:

1. There is a father wavelet $\phi(x)$ and a mother wavelet $\psi(x)$ on $[0, 1]$;
2. At level j and location k, we define

$$\phi_{jk}(x) \triangleq 2^{\frac{j}{2}}\phi(2^j x - k), \tag{16}$$

$$\psi_{jk}(x) \triangleq 2^{\frac{j}{2}}\psi(2^j x - k), \tag{17}$$

with $j \in \mathbb{N}, 0 \le k \le 2^j - 1$.

Note that $\text{supp}(\phi_{jk}) = \text{supp}(\psi_{jk}) = [\frac{k}{2^j}, \frac{k+1}{2^j}]$, the level j characterizes the resolution 2^{-j}, and the parameter k characterizes the spatial location to look at.

Example 3. The *Haar wavelet* is defined by $\phi(x) = \mathbb{1}(x \in [0, 1]), \psi(x) = \mathbb{1}(x \in [0, \frac{1}{2}]) - \mathbb{1}(x \in [\frac{1}{2}, 1])$. This is the first wavelet basis proposed in 1909.

Not all functions can be the father and wavelet wavelets. A crucial property (besides the orthonormality) for wavelets is that, defining

$$V_j \triangleq \text{span}\{\phi_{jk}(x), 0 \le k \le 2^j - 1\} \tag{18}$$

$$W_j \triangleq \text{span}\{\psi_{jk}(x), 0 \le k \le 2^j - 1\} \tag{19}$$

then for any $j_0 \in \mathbb{N}$ we have

$$L^2[0, 1] = \overline{V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots}. \tag{20}$$

As a result, any $f \in L^2[0, 1]$ can be written as

$$f(x) = \underbrace{\sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(x)}_{\text{Gross Information}} + \sum_{j=j_0}^{\infty} \underbrace{\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x)}_{\text{Detail Information at level } j} \tag{21}$$

for some coefficients $(\alpha_{j_0 k}), (\beta_{jk})$. The first term corresponds to the "gross information", i.e., some information in the average sense, e.g., the average magnitude in a small interval. The second term corresponds to

the "detail information", i.e., some information related to the local change, e.g., how function oscillates in a small interval. We can view the detail in different scale/resolution, which is characterized by different levels $j = j_0, j_0 + 1, \cdots$.

The reason why we introduce the wavelet basis is that it is the right basis for the Besov ball, as is shown in the following theorem.

Theorem 4. *Under certain regularity conditions on the wavelet basis, the Besov norm $\|\cdot\|_{\mathcal{B}^s_{p,q}}$ for function is equivalent to $\|\cdot\|_{b^s_{p,q}}$ for its wavelet coefficients, where*

$$\|f\|_{b^s_{p,q}} \triangleq \left(\sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0 k}|^p\right)^{\frac{1}{p}} + \left[\sum_{j=j_0}^{\infty} \left(2^{j(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_{k=0}^{2^j-1} |\beta_{jk}|^p\right)^{\frac{1}{p}}\right)^q\right]^{\frac{1}{q}}. \tag{22}$$

We can also write it in a more compact form as $\|f\|_{b^s_{p,q}} = \|\alpha_{j_0}\|_p + \|2^{j(s+\frac{1}{2}-\frac{1}{p})}\|\beta_j\|_p\|_q$, where the ℓ_p norm is taken with respect to k, and the ℓ_q norm is taken with respect to j.

Recall that we call two norms $(X, \|\cdot\|_1), (X, \|\cdot\|_2)$ are equivalent if there exists universal constants $c_1, c_2 > 0$ such that $c_1\|f\|_2 \le \|f\|_1 \le c_2\|f\|_2$ for any $f \in X$. Then the following corollary is immediate:

Corollary 5. *Defining*

$$\Theta^s_{p,q} \triangleq \{\theta = ((\alpha_{j_0 k}), (\beta_{jk})) : \|\theta\|_{b^s_{p,q}} \le 1\}, \tag{23}$$

there exists some constants $c_1, c_2 > 0$ such that

$$c_1\Theta^s_{p,q} \subset \mathcal{B}^s_{p,q}(L) \subset c_2\Theta^s_{p,q}. \tag{24}$$

We remark that although the form of $\Theta^s_{p,q}$ is still quite complicated, we will make use of some crucial properties of $\Theta^s_{p,q}$ to propose sound estimators and validate the fact that the wavelet basis is the right basis for the Besov space.

4 Thresholding and VisuShrink Estimator

In this section we present the thresholding idea to add the correct non-linearity, and thereby motivates the VisuShrink estimator. The reference for this section and the following ones is:

- I. Johnstone, *Gaussian Estimation: Sequence and Wavelet Models*. Online manuscript (`http://statweb.stanford.edu/~imj/GE09-08-15.pdf`), September 2015.

4.1 Ideal Truncated Estimator

Before we look into the Gaussian sequence estimation problem, first we gain some insights from the Gaussian mean estimation in the scalar case. Consider estimating the mean $\theta \in \mathbb{R}$ in the following scalar model:

$$y = \theta + \epsilon\xi, \qquad \xi \sim \mathcal{N}(0, 1) \tag{25}$$

where $\epsilon > 0$ is known, and the only assumption we impose on θ is that $|\theta| \le \tau$ for some known τ. The target is to find some estimator $\hat{\theta}$ such that the mean squared error $\mathbb{E}_\theta(\hat{\theta} - \theta)^2$ is small. The following insights are straightforward:

1. When τ is large, there is essentially no restriction on the parameter θ, and it is expected that the observation itself $\hat{\theta} = y$ should be a near-optimal estimator. In fact, if $\tau = \infty$, the natural estimator $\hat{\theta} = y$ is the Uniformly Minimum Variance Unbiased Estimator (UMVUE) and also minimax.

2. When τ is small (e.g., much smaller than ϵ), the signal θ is almost completely obscured by the noise. In this case, $\hat{\theta} = 0$ is expected to be a good estimate, since $\mathbb{E}(\hat{\theta} - \theta)^2 = \theta^2 \le \tau^2$ is really small.

Based on these insights, we expect that either $\hat{\theta} = y$ or $\hat{\theta} = 0$ can do a good job. As a result, we introduce the following concept of the *ideal truncated estimator*: suppose that there is a genie who knows the true parameter θ but is restricted to use $\hat{\theta} = y$ or $\hat{\theta} = 0$, it is easy to see that the optimal estimator is

$$\hat{\theta}_{\text{ITE}} = y \mathbb{1}(|\theta| \ge \epsilon) \tag{26}$$

whose mean squared error is given by

$$\mathbb{E}_\theta(\hat{\theta}_{\text{ITE}} - \theta)^2 = \min\{\theta^2, \epsilon^2\}. \tag{27}$$

The following theorem shows that the ideal truncated estimator essentially attains the minimax risk for any $\tau \ge 0$. Of course, the ideal truncated estimator cannot be computed since it knows the signal θ, but it is definitely non-trivial to show that the minimax risk in fact cannot do more than a constant better than the ideal truncated estimator: indeed, the ideal truncated estimator is only allowed to use either y or 0 to estimate, but our estimator can produce anything!

Theorem 6. [4] *For any $\tau \ge 0$, the ideal truncated estimator attains the minimax risk over $|\theta| \le \tau$ within a multiplicative factor of 2.22:*

$$\min\{\tau^2, \epsilon^2\} = \sup_{|\theta| \le \tau} \mathbb{E}_\theta(\hat{\theta}_{\text{ITE}} - \theta)^2 \le 2.22 \cdot \inf_{\hat{\theta}} \sup_{|\theta| \le \tau} \mathbb{E}_\theta(\hat{\theta} - \theta)^2. \tag{28}$$

It's straightforward to generalize this result to the sequence case. Consider the Gaussian sequence model (12) with $\theta \in R(\tau)$, where $\tau = (\tau_1, \tau_2, \cdots)$ is a non-negative sequence and $R(\tau)$ is a orthosymmetric rectangle with one vertex τ:

$$R(\tau) \triangleq \{\theta = (\theta_1, \theta_2, \cdots) : |\theta_i| \le \tau_i, \quad \forall i = 1, 2, \cdots\}. \tag{29}$$

The following corollary is immediate:

Corollary 7. *For any non-negative vector $\tau = (\tau_1, \tau_2, \cdots)$, for Gaussian sequence model (12) we have*

$$\sum_{i=1}^{\infty} \min\{\tau_i^2, \epsilon^2\} = \sup_{\theta \in R(\tau)} \mathbb{E}_\theta \|\hat{\theta}_{\text{ITE}} - \theta\|_2^2 \le 2.22 \cdot \inf_{\hat{\theta}} \sup_{\theta \in R(\tau)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2. \tag{30}$$

4.2 Soft- and Hard-Thresholding Estimator

We showed that the minimax risk for constrained set $\{\theta : \|\theta\| \le \tau\}$ is lower bounded by $\sum_i \min\{\theta_i^2, \epsilon^2\}$, but is it achievable in any sense? Specifically, we want to find an estimator $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_m)$ such that in the Gaussian sequence model (12) of length m, the inequality

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \le \text{something} \times \left(\epsilon^2 + \sum_{i=1}^{m} \min\{\theta_i^2, \epsilon^2\}\right) \tag{31}$$

holds for any $\theta \in \mathbb{R}^m$. Note that for technical reasons we need to have an additional ϵ^2 term in the RHS: indeed, our estimator does not know θ, and it in general cannot tell the difference between the case of $\theta = 0$ and $\theta \approx \epsilon$.

Now we take a careful look at our requirement. As a sanity check, the RHS is really small when $\theta = 0$, which forces the LHS to be small as well. In other words, when the true parameter θ is the zero vector, our

[4] D. L. Donoho, R. C. Liu, and B. MacGibbon. *Minimax risk over hyperrectangles, and implications.* The Annals of Statistics, pp. 1416–1437, 1990.

estimator $\hat{\theta}$ must be close to the zero vector as well. Prove the following result:

Exercise 8. For $X_1, \cdots, X_m \overset{i.i.d}{\sim} \mathcal{N}(0, \epsilon^2)$, we have $\mathbb{P}(\max_{1\le i\le m} |X_i| \ge \epsilon\sqrt{2\log m}) \to 0$ as $m \to \infty$.

Using this exercise, a natural constraint on the estimator $\hat{\theta}$ can be that, $\hat{\theta} = 0$ whenever $\max_{1\le i\le m} |y_i| \le \epsilon\sqrt{2\log m}$. This observation motivates us to do some type of thresholding: specifically, we can define the *soft-thresholding* and *hard-thresholding* functions as follows:

$$\eta_t^s(y) = \text{sign}(y) \cdot (|y| - t)_+ \tag{32}$$

$$\eta_t^h(y) = y \cdot \mathbb{1}(|y| \ge t). \tag{33}$$

Note that these thresholding functions are close to truncation: when $|y|$ is small the functions return zero, and when $|y|$ is large the functions return something close to y.

By acting coordinatewisely, we may also define the soft-thresholding estimator $\hat{\theta}_t^s(y) = (\eta_t^s(y_1), \cdots, \eta_t^s(y_m))$ and the hard-thresholding estimator $\hat{\theta}_t^h(y) = (\eta_t^h(y_1), \cdots, \eta_t^h(y_m))$, and the previous analysis motivates us to choose the threshold t to be roughly $\epsilon\sqrt{2\log m}$. The following theorem shows that the oracle inequality (31) holds for thresholding estimators:

Theorem 9. [5] *For $t = \epsilon\sqrt{2\log m}$, the soft-thresholding estimator $\hat{\theta}_t^s(y)$ satisfies the following oracle inequality in the Gaussian sequence model* (12)*:*

$$\mathbb{E}_\theta \|\hat{\theta}_t^s(y) - \theta\|_2^2 \le (2\log m + 1) \cdot \left(\epsilon^2 + \sum_{i=1}^m \min\{\theta_i^2, \epsilon^2\}\right). \tag{34}$$

The same result holds for the hard-thresholding estimator $\hat{\theta}_t^h(y)$ with $t = \epsilon\sqrt{2\log m + \log\log m}$.

4.3 Projection Estimator and Bias-Variance Tradeoff

Now we use the previous high-dimensional result to derive results for the infinite-dimensional case. We first consider the projection estimator, but the bias-variance decomposition shown below can be generalized to arbitrary estimators that cutoff at a certain threshold.

The *projection estimator* is defined as follows for the Gaussian sequence model (12)

$$\hat{\theta}_j = \begin{cases} y_j, & \text{if } 1 \le j \le m, \\ 0, & \text{if } j > m. \end{cases} \tag{35}$$

It's easy to see that

$$\mathbb{E}_\theta(\hat{\theta}_j - \theta_j)^2 = \begin{cases} \epsilon^2, & \text{if } 1 \le j \le m, \\ \theta_j^2, & \text{if } j > m. \end{cases} \tag{36}$$

As a result, we have

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = m\epsilon^2 + \sum_{j>m} \theta_j^2. \tag{37}$$

Now we have the bias-variance tradeoff in the Gaussian sequence model:

1. Bias: the second term $\sum_{j>m} \theta_j^2$, which originates from throwing away the data in the tail. Note that when m increases, the bias will decrease;

[5] D. L. Donoho, and I. M. Johnstone. *Ideal spatial adaptation by wavelet shrinkage.* biometrika, pp. 425–455, 1994.

2. Variance: the first term $m\epsilon^2$, which originates from the noise in the observation sequence $(y_j)_{j=1}^{\infty}$. Note that when m increases, the variance will increase.

We can think of the threshold m as the reciprocal of the bandwidth h in the kernel-based method: as we know from Fourier analysis, the bandwidth in the time domain and that in the frequency domain satisfies the uncertainty principle. Also, in this case, the relationship of the bias/variance in m corresponds to that of the bias/variance in h in the time domain. As a result, in the transformed domain, the bias-variance tradeoff depends on the cutting position of the observed sequence.

4.4 VisuShrink Estimator

Now we are about to describe the VisuShrink estimator for the Gaussian sequence model (12) with $\Theta = \Theta^s_{p,q}$. The parameters in this model are $\theta = ((\alpha_{j_0k}), (\beta_{jk}) : j \geq j_0, 0 \leq k \leq 2^j - 1)$, and we rewrite this vector as $\theta = (\theta_1, \theta_2, \cdots)$. The VisuShrink estimator does the following:

$$\hat{\theta}_i^{\text{VISU}} = \begin{cases} \eta_t^s(y_i) & \text{if } i \leq m \\ 0 & \text{if } i > m \end{cases} \tag{38}$$

where m is a parameter to be chosen later, and the threshold t is chosen to be $t = \epsilon\sqrt{2\log m}$. Note that compared with the projection estimator, the only difference is that when $i \leq m$ we replaced the raw data y_i by its soft-thresholding $\eta_t^s(y_i)$. Similarly, we can do hard-thresholding as well with $t = \epsilon\sqrt{2\log m + \log\log m}$ according to Theorem 9.

Now we're about to analyze the performance of the VisuShrink estimator. According to Theorem 9, we know that

$$\mathbb{E}_\theta\|\hat{\theta}^{\text{VISU}} - \theta\|_2^2 \leq (2\log m + 1)\left(\epsilon^2 + \sum_{i=1}^{m}\min\{\theta_i^2, \epsilon^2\}\right) + \sum_{i>m}|\theta_i|^2 \tag{39}$$

$$\leq (2\log m + 1)\left(\epsilon^2 + \sum_{i=1}^{\infty}\min\{\theta_i^2, \epsilon^2\}\right) + \sum_{i>m}|\theta_i|^2. \tag{40}$$

By the definition of $\Theta^s_{p,q}$, by choosing $m = \epsilon^{-A}$ for some large enough constant $A > 0$, we may have the tail bound: $\sup_{\theta\in\Theta^s_{p,q}}\sum_{i>m}|\theta_i|^2 \ll \epsilon^2$. As a result, for this choice of m, we have

$$\mathbb{E}_\theta\|\hat{\theta}^{\text{VISU}} - \theta\|_2^2 \lesssim \log(\frac{1}{\epsilon}) \cdot \sum_{i=1}^{\infty}\min\{\theta_i^2, \epsilon^2\} + \epsilon^2\log(\frac{1}{\epsilon}). \tag{41}$$

Now we come to the crucial property of the parameter set $\Theta^s_{p,q}$: it is *solid* and *orthosymmetric*. Equivalently, this means that for any $\theta \in \Theta^s_{p,q}$, the orthosymmetric hyperrectangle $R(|\theta|)$ with a vertex θ is contained again in the parameter set $\Theta^s_{p,q}$. As a result, by Corollary 7 we know that

$$\mathbb{E}_\theta\|\hat{\theta}^{\text{VISU}} - \theta\|_2^2 \lesssim \log(\frac{1}{\epsilon}) \cdot \sum_{i=1}^{\infty}\min\{\theta_i^2, \epsilon^2\} + \epsilon^2\log(\frac{1}{\epsilon}) \tag{42}$$

$$\leq 2.22\log(\frac{1}{\epsilon}) \cdot \inf_{\hat{\theta}}\sup_{\theta'\in R(|\theta|)}\mathbb{E}_{\theta'}\|\hat{\theta} - \theta'\|_2^2 + \epsilon^2\log(\frac{1}{\epsilon}) \tag{43}$$

$$\leq 2.22\log(\frac{1}{\epsilon}) \cdot \inf_{\hat{\theta}}\sup_{\theta'\in\Theta^s_{p,q}}\mathbb{E}_{\theta'}\|\hat{\theta} - \theta'\|_2^2 + \epsilon^2\log(\frac{1}{\epsilon}) \tag{44}$$

where in the last inequality we have used the property

$$R(|\theta|) \subset \Theta^s_{p,q}, \tag{45}$$

where $R(\cdot)$ is defined in (29).

As a result, the VisuShrink estimator attains the minimax risk within a logarithmic factor.

We remark that we do not use any specific properties of $\Theta^s_{p,q}$ (which is of a complicated form) other than that it is *solid* and *orthosymmetric*. Also, we prove that the VisuShrink estimator is nearly minimax *without* even figuring out what the minimax risk is. These observations indicate that the *geometry* of the parameter set is really important, and the reason why we choose the wavelet basis for the Besov ball is that the Besov ball becomes solid and orthosymmetric in the wavelet domain! In other words, for any orthonormal basis $(\phi_i)_{i\in I}$, the VisuShrink idea still works as long as the associated parameter set Θ is solid and orthosymmetric in the transformed space (45). In fact, this property requires that the basis be an *unconditional basis*:

Definition 10 (Unconditional Basis)**.** *An orthonormal basis $(\phi_i)_{i\in I}$ is an* unconditional basis *of the real normed vector space $(X, \|\cdot\|)$ if and only if there exists a universal constant $C > 0$ such that*

$$\left\|\sum_{i\in J} \epsilon_i \phi_i\right\| \le C \left\|\sum_{i\in J} \phi_i\right\| \tag{46}$$

holds for any finite $J \subset I$.

The main messages are that:

1. Unconditional basis is the optimal basis in nonparametric function estimation;
2. Wavelet basis is an unconditional basis for the Besov norm $(L^2[0,1], \|\cdot\|_{\mathcal{B}^s_{p,q}})$ (and Fourier basis is not in general).

Finally, we note the relationship between $\Theta^s_{p,q}$ and $\mathcal{B}^s_{p,q}(L)$, and summarize the VisuShrink estimator as follows:

1. Fix some initial level j_0 (which is of the constant order) and termination level $j_\epsilon \asymp \log(\frac{1}{\epsilon})$;
2. Transform the observation process $(Y_t)_{t\in[0,1]}$ to the wavelet domain with initial level j_0, and obtain the corresponding α-coefficients and β-coefficients empirically;
3. Use the following procedure to obtain new coefficient estimates:
   (a) For α-coefficients (which are only at level j_0), keep them all;
   (b) For β-coefficients, for $j > j_\epsilon$ discard them all, and for $j_0 \le j \le j_\epsilon$ apply the thresholding estimator (either soft or hard one) with suitable threshold given in Theorem 9 to the observation vector.
4. Transform the estimated wavelet coefficients back to the function space, and obtain $\hat{f}^{\text{VISU}}$.

The property of the VisuShrink estimator $\hat{f}^{\text{VISU}}$ is summarized in the following theorem:

Theorem 11. [6] *For any $s > 0, p, q \in [1, \infty]$, the VisuShrink estimator $\hat{f}^{\text{VISU}}$ attains the minimax risk over Besov balls $\mathcal{B}^s_{p,q}(L)$ within logarithmic factors:*

$$\sup_{f\in\mathcal{B}^s_{p,q}(L)} \mathbb{E}_f \|\hat{f}^{\text{VISU}} - f\|_2^2 \lesssim \log(\frac{1}{\epsilon}) \cdot \inf_{\hat{f}} \sup_{f\in\mathcal{B}^s_{p,q}(L)} \mathbb{E}_f \|\hat{f} - f\|_2^2. \tag{47}$$

[6] D. L. Donoho, and I. M. Johnstone. *Ideal spatial adaptation by wavelet shrinkage.* biometrika, pp. 425–455, 1994.

4.5 Discussions

We make some discussions on the previous VisuShrink estimator.

First of all, the VisuShrink estimator is almost a linear estimator (i.e., close to the projection estimator), while there is also a little bit non-linearity here, i.e., the *thresholding* idea. We can think of the thresholding approach as a selector which selects the coefficients to keep in a data-dependent manner: when the empirical coefficient is large, we expect it to be useful signal and keep it; when the empirical coefficient is small, we expect it to be the noise and discard it. We can compare this idea with Lepski's trick to deal with the *sparse regime* we defined in the last lecture: in the sparse regime, the signal is supported on a small interval and all others are noise. In this case, Lepski's trick selects a large bandwidth in the noise regime to essentially neglect all noise, and the VisuShrink estimator simply selects the peak and neglects all others of the transformed signal in the wavelet domain.

Secondly, the VisuShrink estimator employs the *shrinkage* idea, which means that *reduce the variance significantly with a little bit increase on the bias* in statistics. Actually, this is where the term "shrink" in the name "VisuShrink" comes from. Specifically, compared with the projection estimator which keeps the raw observation, the thresholding idea incurs a larger bias (note that the previous one is indeed *unbiased*!), while reduces the variance significantly (e.g., from ϵ^2 to $\min\{\epsilon^2, \theta_i^2\}$ per symbol).

Finally, we remark that by construction, the VisuShrink estimator does not require the knowledge of parameters s, p, q, L and is thus an adaptive estimator. Similar to the Lepski's estimator, the only knowledge the VisuShrink requires is an *upper bound* of the smoothness parameter s, for the termination level j_ϵ depends on this upper bound.

5 Thresholding and SureShrink Estimator

In the previous section we have validated the thresholding idea by proving an oracle inequality (Theorem 9), and use the geometry of the parameter set $\Theta^s_{p,q}$ to relate the risk of the ideal truncated estimator to the minimax risk. In this section we will validate the thresholding idea from a different viewpoint, and introduce the resulting SureShrink estimator.

5.1 Gaussian Mean Estimation over ℓ_p Balls with ℓ_q Error

Consider the Gaussian sequence estimation problem (12) with a simpler parameter set: $\Theta = \{\theta : \|\theta\|_p \le R\}$ is the ℓ_p ball. Also, instead of the mean squared error loss, we consider the ℓ_q loss as the general loss functions where $p \in (0, \infty], q \in [1, \infty)$. The question is that: for this simple example, which estimator is nearly minimax under different parameter configurations (p, q, R, ϵ)?

We first begin with some insights. When R is large (or equivalently ϵ is small), the constraint on θ is quite loose, and thus we should use an estimator close to the natural one (i.e., the empirical observation). When R is small (ϵ is large), the vector is close to zero and we may directly apply a zero estimator. When $p \in (0, \infty]$ is small, we know that the parameter θ is somehow quite *sparse*, and thus the resulting estimator should have many zero entries. In contrast, when p is large, the parameter θ can be quite dense, and the natural estimator is expected to work here.

As a result, if we treat the natural estimator $\hat{\theta} = y$ as the thresholding estimator $\hat{\theta} = \eta^s_t(y)$ with threshold $t = 0$, and the zero estimator $\hat{\theta} = 0$ as the thresholding estimator $\hat{\theta} = \eta^s_t(y)$ with threshold $t = \infty$, it seems that the thresholding estimator with some suitable threshold should work. This intuition turns out to be correct, which is shown in the following theorem:

Theorem 12. [7] *For most parameter configurations (p, q, R, ϵ), there exists a universal constant $C_{p,q} > 0$*

[7] D. L. Donoho, and I. M. Johnstone. *Minimax risk over ℓ_p-balls for ℓ_q-error.* Probability Theory and Related Fields, 99(2), pp. 277–303, 1994.

such that for the Gaussian sequence model (12) *with* $\Theta = \{\theta : \|\theta\|_p \le R\}$,

$$\inf_t \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\eta_t^s(y) - \theta\|_q^q \le C_{p,q} \cdot \inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat\theta - \theta\|_q^q. \tag{48}$$

The same result also holds for the hard-thresholding estimator $\eta_t^h(y)$.

The implication of Theorem 12 to our case is that: although the parameter space $\Theta_{p,q}^s$ is of a very complicated form, it only involves the the combination of ℓ_p and ℓ_q norms! Hence, we expect that the thresholding idea also works in our case over $\Theta_{p,q}^s$. Specifically, we consider the following estimator:

1. Transform the observation to wavelet coefficients starting from initial level j_0 (of a constant order);
2. Keep all father wavelet coefficients, and for each level j, apply the (soft- or hard-)thresholding estimator to the mother wavelet coefficients with threshold t_j;
3. Transform back the coefficients into functions to yield $\hat f$.

We write the resulting estimator as $\hat f_t$, where $t = (t_{j_0}, t_{j_0+1}, \cdots)$ is the threshold sequence. Based on the previous insights and with the help of Theorem 12, the following result holds:

Theorem 13. [8] *For the nonparametric function estimation over Besov balls, the thresholding estimator with appropriate thresholds attains the minimax risk within a multiplicative factor:*

$$\inf_{t=(t_{j_0}, t_{j_0+1}, \cdots)} \sup_{f\in\mathcal{B}_{p,q}^s(L)} \mathbb{E}_f \|\hat f_t - f\|_2^2 \lesssim \inf_{\hat f} \sup_{f\in\mathcal{B}_{p,q}^s(L)} \mathbb{E}_f \|\hat f - f\|_2^2. \tag{49}$$

5.2 SURE (Stein's Unbiased Risk Estimate)

The previous Theorem ensures that *some* thresholding estimator works, but does not specify which threshold we should choose. A naïve thought is that, if we could compare the performances of $\hat f_t$ with different t's, then we should choose the one with the minimum error:

$$t^* = \arg\min_t \mathbb{E}_f \|\hat f_t - f\|_2^2. \tag{50}$$

However, this approach is infeasible since we do not know the true function f. Despite this difficulty, the good news is that we can still apply this idea and use an unbiased estimator of $\mathbb{E}_f\|\hat f_t - f\|_2^2$ without knowing f.

Now we start to illustrate the idea. Consider the Gaussian sequence model in (12) with length m, and fix any estimator $\hat\theta(y)$ of θ. Note that $g(y) = \hat\theta(y) - y$ only depends on y but not on θ, and

$$\mathbb{E}_\theta\|\hat\theta - \theta\|_2^2 = \mathbb{E}_\theta\|g(y) + y - \theta\|_2^2 \tag{51}$$

$$= \mathbb{E}_\theta\|g(y)\|_2^2 + \mathbb{E}_\theta\|y - \theta\|_2^2 + 2\mathbb{E}_\theta[(y-\theta)^T g(y)]. \tag{52}$$

Exercise 14. Prove Stein's identity: for $X \sim \mathcal{N}(\mu, \sigma^2)$ and any (weakly) differentiable function f, we have $\mathbb{E}[(X-\mu)f(X)] = \sigma^2\mathbb{E}[f'(X)]$.

By Stein's identity, we further have

$$\mathbb{E}_\theta\|\hat\theta - \theta\|_2^2 = \mathbb{E}_\theta\|g(y)\|_2^2 + \mathbb{E}_\theta\|y - \theta\|_2^2 + 2\mathbb{E}_\theta[(y-\theta)^T g(y)] \tag{53}$$

$$= \mathbb{E}_\theta\|g(y)\|_2^2 + m\epsilon^2 + 2\epsilon^2\mathbb{E}_\theta[\nabla\cdot g(y)] \tag{54}$$

$$= \mathbb{E}_\theta\left[(m + 2\nabla\cdot g(y))\epsilon^2 + \|g(y)\|_2^2\right]. \tag{55}$$

As a result, we have the following definition of the Stein's Unbiased Risk Estimator:

[8] D. L. Donoho, and I. M. Johnstone. *Minimax estimation via wavelet shrinkage.* The annals of Statistics, 26(3), pp. 879–921, 1998.

Definition 15 (SURE)**.** *For the Gaussian sequence model in* (12) *with length m, then for any estimator $\hat{\theta}(y)$ with a weakly differentiable $g(y) \triangleq \hat{\theta}(y) - y$, the* Stein's Unbiased Risk Estimator (SURE) *is defined by*

$$r^{\mathrm{SURE}}(y) \triangleq (m + 2\nabla \cdot g(y))\epsilon^2 + \|g(y)\|_2^2. \tag{56}$$

The SURE satisfies $\mathbb{E}_\theta[r^{\mathrm{SURE}}(y)] = \mathbb{E}_\theta\|\hat{\theta} - \theta\|_2^2$ for any $\theta \in \Theta$.

5.3 The SureShrink Estimator

With the SURE in hand, we may introduce the SureShrink estimator, which chooses the threshold sequence $t = (t_j)_{j \geq j_0}$ based on the risk estimate. Specifically, on each level $j \geq j_0$, we randomly divide the index set $\{0, 1, \cdots, 2^j - 1\}$ into two halves I, I', and for the soft-thresholding we define

$$t_I \triangleq \arg\min_{t \geq 0} \sum_{k \in I'} (1 - 2 \cdot \mathbb{1}(|y_{j,k}| \leq t))\epsilon^2 + (\min\{t, |y_{j,k}|\})^2 \tag{57}$$

$$t_{I'} \triangleq \arg\min_{t \geq 0} \sum_{k \in I} (1 - 2 \cdot \mathbb{1}(|y_{j,k}| \leq t))\epsilon^2 + (\min\{t, |y_{j,k}|\})^2 \tag{58}$$

where $(y_{j,k})_{0 \leq k \leq 2^j - 1}$ are the empirical wavelet coefficients on level j. Then we apply the soft-thresholding estimator with threshold t_I to all $y_{j,k}$ for $k \in I$, and apply it with threshold $t_{I'}$ to all $y_{j,k}$ for $k \in I'$. The same idea can also be applied to the hard-thresholding estimator. We remark that the random sample splitting approach is purely for technical purposes (to gain independence), which is not necessary in practice. The SureShrink estimator is defined to be $\hat{f}_{\hat{t}}$ with the thresholds given by the previous equation.

The theoretical performance of the SureShrink estimator is summarized in the following theorem.

Theorem 16. [9] *For $s > \frac{1}{p} - \frac{1}{2}$, the SureShrink estimator $\hat{f}^{\mathrm{SURE}} = \hat{f}_{\hat{t}}$ essentially performs as well as the optimal thresholding estimator, and thus attains the minimax risk over Besov balls $\mathcal{B}^s_{p,q}(L)$ within multiplicative constants:*

$$\sup_{f \in \mathcal{B}^s_{p,q}(L)} \mathbb{E}_f\|\hat{f}^{\mathrm{SURE}} - f\|_2^2 \leq (1 + o(1)) \cdot \inf_{(t_j)_{j \geq j_0}} \sup_{f \in \mathcal{B}^s_{p,q}(L)} \mathbb{E}_f\|\hat{f}_t - f\|_2^2 \tag{59}$$

$$\lesssim \inf_{\hat{f}} \sup_{f \in \mathcal{B}^s_{p,q}(L)} \mathbb{E}_f\|\hat{f} - f\|_2^2. \tag{60}$$

6 Minimax L_r Risk over Besov Balls

In the previous sections we have studied the nonparametric function estimation problem over Besov balls using the mean squared error loss, where we have used the L_2 isometry to establish the transformation from the function domain to the wavelet domain. However, we remark that the L_2 isometry is not essential here: the same thresholding idea also applies to general L_r error, for $r \in [1, \infty]$. Furthermore, note that we have shown the minimax optimality of various estimators *without* specifying the value of the minimax risk, for completeness we give the most general result here.

Theorem 17. [10] *For any $s > \frac{1}{p} - \frac{1}{r}, p, q, r \in [1, \infty]$, the minimax L_r risk in estimating the function f over Besov balls $\mathcal{B}^s_{p,q}(L)$ is given by*

$$\left(\inf_{\hat{f}} \sup_{f \in \mathcal{B}^s_{p,q}(L)} \mathbb{E}_f\|\hat{f} - f\|_r^r\right)^{\frac{1}{r}} \asymp \begin{cases} (\epsilon^2)^{\frac{s}{2s+1}} & \text{if } r < (2s+1)p \\ (\epsilon^2 \log(1/\epsilon))^{\frac{s}{2s+1}} (\log(1/\epsilon))^{(\frac{1}{2} - \frac{p}{qr})_+} & \text{if } r = (2s+1)p \\ (\epsilon^2 \log(1/\epsilon))^{\frac{s - \frac{1}{p} + \frac{1}{r}}{2(s - \frac{1}{p}) + 1}} & \text{if } r > (2s+1)p \end{cases}. \tag{61}$$

[9] D. L. Donoho, and I. M. Johnstone. *Adapting to unknown smoothness via wavelet shrinkage.* Journal of the American statistical association, 90(432), pp. 1200–1224, 1995.

[10] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. *Density estimation by wavelet thresholding.* The Annals of Statistics, pp. 508–539, 1996.

In contrast, the minimax linear *risk in estimating the function f over Besov balls $\mathcal{B}^s_{p,q}(L)$ is given by*

$$\left(\inf_{\hat{f}^{\text{lin}}} \sup_{f\in\mathcal{B}^s_{p,q}(L)} \mathbb{E}_f\|\hat{f}^{\text{lin}} - f\|_r^r\right)^{\frac{1}{r}} \asymp \begin{cases} (\epsilon^2)^{\frac{s}{2s+1}} & \text{if } r \le p \\ (\epsilon^2)^{\frac{s-\frac{1}{p}+\frac{1}{r}}{2(s-\frac{1}{p}+\frac{1}{r})+1}} & \text{if } p < r < \infty \\ (\epsilon^2\log(1/\epsilon))^{\frac{s-\frac{1}{p}}{2(s-\frac{1}{p})+1}} & \text{if } r = \infty \end{cases} \tag{62}$$

where the infimum is taken over all possible linear estimators.