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:

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

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.

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 \( 1 \leq i \leq n \), where \( x_i = \frac{i}{n} \) and \( \xi_i \sim \mathcal{N}(0,1) \);

2. **Gaussian White Noise Model**: we observe a process \( (Y_t)_{t \in [0,1]} \) with
   \[
   Y_t = \int_0^t f(s) \, ds + \sigma \sqrt{n} B_t, \quad t \in [0,1] \tag{2}
   \]
   where \( (B_t)_{t \in [0,1]} \) is the standard Brownian Motion on \([0,1]\). This model can also be written in the stochastic differential equation (SDE) form as \( dY_t = f(t) \, dt + \frac{\sigma}{\sqrt{n}} dB_t \);

3. **Density Estimation Model**: we observe \( n \) i.i.d samples \( y_1, \ldots, y_n \sim g(\cdot) \), 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 \( n g(\cdot) \), 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.** Under certain technical conditions, these four models are asymptotically 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

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

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.

---

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 
\[
    \theta_j \triangleq \int_0^1 \phi_j(t) f(t) dt.
\]
Note that the restriction \(f \in \mathcal{F}\) will be transformed into the condition \(\theta \in \Theta\) for some proper parameter set \(\Theta\). Also, for any estimator \(\hat{f}\), we can also represent it by
\[
    \hat{\theta}_j \triangleq \int_0^1 \phi_j(t) \hat{f}(t) dt.
\]

As for the observation, we may define
\[
    y_j \triangleq \int_0^1 \phi_j(t) dY_t
\]
\[
    = \int_0^1 \phi_j(t) \left( f(t) dt + \frac{\sigma}{\sqrt{n}} dB_t \right)
\]
\[
    = \int_0^1 \phi_j(t) f(t) dt + \frac{\sigma}{\sqrt{n}} \int_0^1 \phi_j(t) dB_t
\]
\[
    \equiv \theta_j + \epsilon \cdot \xi_j
\]
where \(\epsilon \triangleq \frac{\sigma}{\sqrt{n}}\) denotes the noise level, and
\[
    \xi_j \triangleq \int_0^1 \phi_j(t) dB_t.
\]

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

By the previous exercise, we know that the Gaussian white noise model reduces to the following Gaussian sequence model under the basis \((\phi_j)_{j=1}^{\infty}\):
\[
    y_j = \theta_j + \epsilon \xi_j, \quad \theta = (\theta_1, \theta_2, \cdots) \in \Theta, \quad \epsilon = \frac{\sigma}{\sqrt{n}}, \quad \xi_j \overset{i.i.d.}{\sim} N(0,1).
\]
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
\[
    \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = \sup_{\theta \in \Theta} \sum_{j=1}^{\infty} \mathbb{E}_\theta (\hat{\theta}_j - \theta_j)^2.
\]
As a result, under the basis change, we operate in a Gaussian sequence model and would like to estimate the mean vector simultaneously.

### 2.3 Projection Estimator and Bias-Variance Tradeoff

First let’s see how the bias-variance tradeoff will look at after the change of basis. Since the observation itself is a natural estimator for the mean in the Gaussian model, a natural candidate to solve the Gaussian sequence model (12) is the projection estimator:
\[
    \hat{\theta}_j = \begin{cases} 
        y_j, & \text{if } 1 \leq j \leq m, \\
        0, & \text{if } j > m.
    \end{cases}
\]
It’s easy to see that
\[ E(\hat{\theta}_j - \theta_j)^2 = \begin{cases} 
\epsilon^2, & \text{if } 1 \leq j \leq m, \\
\theta_j^2, & \text{if } j > m.
\end{cases} \]  

(15)

As a result, we have
\[ E\|\hat{\theta} - \theta\|^2 = m\epsilon^2 + \sum_{j>m} \theta_j^2. \]  

(16)

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;

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.

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 \( \{\varphi_j\}_{j=1}^\infty \). Specifically, we will choose \( \mathcal{F} \) to be the Besov ball \( \mathcal{B}^{s}_{p,q}(L) \), and \( \{\varphi_j\}_{j=1}^\infty \) 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:


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
\[ \|f\|_{\mathcal{B}^{s}_{p,q}} \approx \|f^{(s)}\|_p \]  

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
\[ \mathcal{B}^{s}_{p,q}(L) \equiv \{ f : \|f\|_{\mathcal{B}^{s}_{p,q}} \leq L \}. \]  

(18)
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 \([0,1]\):

   \[
   \phi_{jk}(x) \triangleq 2^j \phi(2^j x - k), \\
   \psi_{jk}(x) \triangleq 2^j \psi(2^j x - k),
   \]
   \( j, k \in \mathbb{N}, 0 \leq k \leq 2^j - 1. \)

   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 \( \phi(x) = 1(x \in [0,1]) \), \( \psi(x) = 1(x \in [0, \frac{1}{2}]) - 1(x \in [\frac{1}{2}, 1]) \). 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

   \[
   V_j \triangleq \text{span}\{\phi_{jk}(x), 0 \leq k \leq 2^j - 1\} \\
   W_j \triangleq \text{span}\{\psi_{jk}(x), 0 \leq k \leq 2^j - 1\}
   \] for any \( j_0 \in \mathbb{N} \) we have

   \[ L^2[0,1] = V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots. \] (23)

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

   \[
   f(x) = \sum_{k=0}^{2^0-1} \alpha_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k} \psi_{j,k}(x)
   \]
   \( \text{Gross Information} \quad \text{Detail Information at level } j \)
   \[ 0 \leq j_0 \leq j \quad \text{for some coefficients } (\alpha_{j_0,k}), (\beta_{j,k}). \]

   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 \( \| \cdot \|_{B_p^s} \) for function is equivalent to \( \| \cdot \|_{b_p,q} \) for its wavelet coefficients, where

   \[
   \| f \|_{b_p,q} = \left( \sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0,k}|^p \right)^{\frac{1}{p}} + \left[ \sum_{j=j_0}^{\infty} \left( \sum_{k=0}^{2^j-1} |\beta_{j,k}|^p \right)^{\frac{1}{p}} \right]^{\frac{1}{q}}. \] (25)

   We can also write it in a more compact form as \( \| f \|_{b_p,q} = \|\alpha_{j_0}\|_p + \|2^{j(s+\frac{1}{2}-\frac{1}{p})}\|\beta_j\|_p\|q, \) where the \( \ell_p \) norm is taken with respect to \( k \), and the \( \ell_q \) 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 \(c_1 \| f \|_2 \leq \| f \|_1 \leq c_2 \| f \|_2\) for any \(f \in X\). Then the following corollary is immediate:

**Corollary 5.** Defining

\[
\Theta_{p,q}^s \triangleq \{ \theta = ((\alpha_{j_0k}), (\beta_{jk})) : \| \theta \|_{b_{p,q}} \leq 1 \},
\]

then there exists some constants \(c_1, c_2 > 0\) such that

\[
c_1 \Theta_{p,q}^s \subset \mathcal{B}_{p,q}^s (L) \subset c_2 \Theta_{p,q}^s.
\]

We remark that although the form of \(\Theta_{p,q}^s\) is still quite complicated, we will make use of some crucial properties of \(\Theta_{p,q}^s\) 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:


#### 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:

\[
y = \theta + \epsilon \xi, \quad \xi \sim \mathcal{N}(0, 1)
\]

where \(\epsilon > 0\) is known, and the only assumption we impose on \(\theta\) is that \(|\theta| \leq \tau\) for some known \(\tau\). 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 \(\tau\) is large, there is essentially no restriction on the parameter \(\theta\), 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 \(\tau\) is small (e.g., much smaller than \(\epsilon\)), the signal \(\theta\) is almost completely obscured by the noise. In this case, \(\hat{\theta} = 0\) is expected to be a good estimate, since \(\mathbb{E}_\theta (\hat{\theta} - \theta)^2 = \theta^2 \leq \tau^2\) 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 \(\theta\) but is restricted to use \(\hat{\theta} = y\) or \(\hat{\theta} = 0\), it is easy to see that the optimal estimator is

\[
\hat{\theta}_{ITE} = y 1(|\theta| \geq \epsilon)
\]

whose mean squared error is given by

\[
\mathbb{E}_\theta (\hat{\theta}_{ITE} - \theta)^2 = \min \{\theta^2, \epsilon^2\}.
\]

The following theorem shows that the ideal truncated estimator essentially attains the minimax risk for any \(\tau \geq 0\):
Theorem 6. For any $\tau \geq 0$, the ideal truncated estimator attains the minimax risk over $|\theta| \leq \tau$ within a multiplicative factor of 2.22:

$$\min\{\tau^2, \epsilon^2\} = \sup_{|\theta| \leq \tau} \mathbb{E}_\theta(\hat{\theta}_{ITE} - \theta)^2 \leq 2.22 \cdot \inf_{\theta} \sup_{|\theta| \leq \tau} \mathbb{E}_\theta(\hat{\theta} - \theta)^2. \quad (31)$$

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 $\tau$:

$$R(\tau) \triangleq \{\theta = (\tau_1, \tau_2, \cdots) : |\theta_i| \leq \tau_i, \ \forall i = 1, 2, \cdots\}. \quad (32)$$

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

$$\sum_{i=1}^{\infty} \min\{\tau_i^2, \epsilon^2\} = \sup_{\theta \in R(\tau)} \mathbb{E}_\theta||\hat{\theta}_{ITE} - \theta||_2^2 \leq 2.22 \cdot \inf_{\hat{\theta}} \sup_{\theta \in R(\tau)} \mathbb{E}_\theta||\hat{\theta} - \theta||_2^2. \quad (33)$$

4.2 Soft- and Hard-Thresholding Estimator

Note that the ideal truncated estimator is not indeed an estimator since it requires the unavailable information $\theta$. However, the previous corollary motivates us to find another estimator $\hat{\theta}$ whose performance is comparable to that of $\hat{\theta}_{ITE}$, which has been shown to be nearly minimax. 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

$$\mathbb{E}_\theta||\hat{\theta} - \theta||_2^2 \leq \text{something} \times \left(\epsilon^2 + \sum_{i=1}^{m} \min\{\theta_i^2, \epsilon^2\}\right) \quad (34)$$

holds for any $\theta \in R^m$. Note that for minor technical reasons we need to have an additional $\epsilon^2$ term in the RHS.

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 $\theta$ is the zero vector, our 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 $\mathbb{P}(|\max_{1 \leq i \leq m} X_i | \geq \epsilon \sqrt{2\log m}) \to 0$ as $m \to \infty$.

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

$$\eta^b_t(y) = \text{sign}(y) \cdot (|y| - t)_+ \quad (35)$$

$$\eta^h_t(y) = y \cdot 1(|y| \geq t). \quad (36)$$

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 coordinate wisely, we may also define the soft-thresholding estimator $\hat{\theta}^s_t(y) = (\eta^s_t(y_1), \cdots, \eta^s_t(y_m))$ and the hard-thresholding estimator $\hat{\theta}^h_t(y) = (\eta^h_t(y_1), \cdots, \eta^h_t(y_m))$, 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 (34) holds for thresholding estimators:

---

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

$$E_\theta \|\hat{\theta}_i^s(y) - \theta\|_2^2 \leq (2 \log m + 1) \cdot \left( \epsilon^2 + \sum_{i=1}^{m} \min\{\theta_i^2, \epsilon^2\} \right).$$

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

4.3. VisuShrink Estimator

Now we are about to describe the VisuShrink estimator for the Gaussian sequence model (12) with $\Theta = \Theta_{p,q}^s$.

The parameters in this model are $\theta = (\alpha_{j,k}, \beta_{j,k}) : j \geq j_0, 0 \leq k \leq 2^j - 1$, and we rewrite this vector as $\theta = (\theta_1, \theta_2, \cdots)$. The VisuShrink estimator does the following:

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

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 $\eta_t(y_i)$). 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

$$E_\theta \|\hat{\theta}_i^{VISU} - \theta\|_2^2 \leq (2 \log m + 1) \left( \epsilon^2 + \sum_{i=1}^{m} \min\{\theta_i, \epsilon^2\} \right) + \sum_{i=m}^\infty |\theta_i|^2$$

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

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

$$E_\theta \|\hat{\theta}_i^{VISU} - \theta\|_2^2 \lesssim \log \frac{1}{\epsilon} \cdot \sum_{i=1}^\infty \min\{\theta_i, \epsilon^2\} + \epsilon^2 \log \frac{1}{\epsilon}. \quad (41)$$

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

$$E_\theta \|\hat{\theta}_i^{VISU} - \theta\|_2^2 \lesssim \log \frac{1}{\epsilon} \cdot \sum_{i=1}^\infty \min\{\theta_i, \epsilon^2\} + \epsilon^2 \log \frac{1}{\epsilon} \quad (42)$$

$$\leq 2.22 \log \frac{1}{\epsilon} \cdot \inf_{\theta' \in R(\{\theta\})} \sup_{\theta \in \Theta_{p,q}^s} E_\theta \|\hat{\theta} - \theta'\|_2^2 + \epsilon^2 \log \frac{1}{\epsilon} \quad (43)$$

$$\leq 2.22 \log \frac{1}{\epsilon} \cdot \inf_{\theta' \in \Theta_{p,q}^s} E_\theta \|\hat{\theta} - \theta'\|_2^2 + \epsilon^2 \log \frac{1}{\epsilon} \quad (44)$$

where in the last inequality we have used the property $R(\{\theta\}) \subset \Theta_{p,q}^s$. As a result, the VisuShrink estimator attains the minimax risk within a logarithmic factor.

Finally, we note the relationship between $\Theta_{p,q}^s$ and $B_{p,q}^s(L)$, 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 \( \alpha \)-coefficients and \( \beta \)-coefficients empirically;

3. Use the following procedure to obtain new coefficient estimates:

   (a) For \( \alpha \)-coefficients (which are only at level \( j_0 \)), keep them all;

   (b) For \( \beta \)-coefficients, for \( j > j_\epsilon \) discard them all, and for \( j_0 \leq j \leq j_\epsilon \) 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 10.**  
For any \( s > 0, p, q \in [1, \infty], \) the VisuShrink estimator \( \hat{f}^\text{VISU} \) attains the minimax risk over Besov balls \( \mathcal{B}_{p,q}^s(L) \) within logarithmic factors:

\[
\sup_{f \in \mathcal{B}_{p,q}^s(L)} \mathbb{E}_f \| \hat{f}^\text{VISU} - f \|_2^2 \lesssim \log(\frac{1}{\epsilon}) \cdot \inf_{\hat{f}} \sup_{f \in \mathcal{B}_{p,q}^s(L)} \mathbb{E}_f \| \hat{f} - f \|_2^2.
\]  \hspace{1cm} (45)

### 4.4 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 \( \epsilon^2 \) to \( \min\{\epsilon^2, \theta_0^2\} \) per symbol). We remark that the variance is a main issue in the nonparametric function estimation, while we will see in the next lecture that in functional estimation problems, bias becomes dominant!

Thirdly, by inspecting the proof we find that we do not use any specific properties of \( \Theta_{p,q}^s \) (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 \( \langle \phi_i \rangle_{i \in I} \), the VisuShrink idea still works as long as the associated parameter set \( \Theta \) is solid and orthosymmetric in the transformed space. In fact, this property requires that the basis be an unconditional basis:

---

Definition 11 (Unconditional Basis). An orthonormal basis \( (\varphi_i)_{i \in I} \) 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

\[
\left\| \sum_{i \in J} \epsilon_i \varphi_i \right\| \leq C \left\| \sum_{i \in J} \varphi_i \right\|
\]  

(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 \( (L^2[0,1], \| \cdot \|_{B^s_{p,q}}) \) (and Fourier basis is not).

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 \( j_\epsilon \) 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 \( \Theta_{p,q}^s \) 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 \( \ell_p \) Balls with \( \ell_q \) Error

Consider the Gaussian sequence estimation problem (12) with a simpler parameter set: \( \Theta = \{ \theta : \| \theta \|_p \leq R \} \) is the \( \ell_p \) ball. Also, instead of the mean squared error loss, we consider the \( \ell_q \) 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 \( (p, q, R, \epsilon) \)?

As a result, if we treat the natural estimator \( \hat{\theta} = y \) as the thresholding estimator \( \hat{\theta} = \eta_s^{\theta}(y) \) with threshold \( t = 0 \), and the zero estimator \( \hat{\theta} = 0 \) as the thresholding estimator \( \hat{\theta} = \eta_s^0(y) \) 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 \( (p, q, R, \epsilon) \), there exists a universal constant \( C_{p,q} > 0 \) such that for the Gaussian sequence model (12) with \( \Theta = \{ \theta : \| \theta \|_p \leq R \} \),

\[
\inf_{\theta \in \Theta} \sup_{\theta \in \Theta} \mathbb{E}_\theta \| \eta_s^{\theta}(y) - \theta \|_q^q \leq C_{p,q} \cdot \inf_{\hat{\theta} \in \Theta} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \| \hat{\theta} - \theta \|_q^q.
\]  

(47)

The same result also holds for the hard-thresholding estimator \( \eta_h^{\theta}(y) \).

The implication of Theorem 12 to our case is that: although the parameter space $\Theta_{p,q}^*$ is of a very complicated form, it only involves the the combination of $\ell_p$ and $\ell_q$ norms! Hence, we expect that the thresholding idea also works in our case over $\Theta_{p,q}^*$. 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.** For the nonparametric function estimation over Besov balls, the thresholding estimator with appropriate thresholds attains the minimax risk within a multiplicative factor:

$$\inf_{t=(t_{j_0}, t_{j_0+1}, \cdots)} \sup_{f \in \mathcal{B}_{p,q}(L)} \mathbb{E}_f \| \hat{f}_t - f \|^2_2 \leq \inf_{f} \sup_{t \in \mathcal{B}_{p,q}(L)} \mathbb{E}_f \| \hat{f} - f \|^2_2. \quad (48)$$

### 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 naive 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:

$$t^* = \arg\min_t \mathbb{E}_f \| \hat{f}_t - f \|^2_2. \quad (49)$$

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 $\tilde{\theta}(y)$ of $\theta$. Note that $g(y) = \tilde{\theta}(y) - y$ only depends on $y$ but not on $\theta$, and

$$\mathbb{E}_\theta \| \tilde{\theta} - \theta \|^2_2 = \mathbb{E}_\theta \| g(y) + y - \theta \|^2_2 \quad (50)$$

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

**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

$$\mathbb{E}_\theta \| \tilde{\theta} - \theta \|^2_2 = \mathbb{E}_\theta \| g(y) \|^2_2 + \mathbb{E}_\theta \| y - \theta \|^2_2 + 2\mathbb{E}_\theta (y - \theta)^T g(y) \quad (52)$$

$$= \mathbb{E}_\theta \| g(y) \|^2_2 + m\epsilon^2 + 2\epsilon^2 \mathbb{E}_\theta (\nabla \cdot g(y)) \quad (53)$$

$$= \mathbb{E}_\theta \left( (m + 2\nabla \cdot g(y))\epsilon^2 + \| g(y) \|^2_2 \right). \quad (54)$$

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

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

$$r^{\text{SURE}}(y) \triangleq (m + 2\nabla \cdot g(y))\epsilon^2 + \| g(y) \|^2_2. \quad (55)$$

The SURE satisfies $\mathbb{E}_\theta [r^{\text{SURE}}(y)] = \mathbb{E}_\theta \| \tilde{\theta} - \theta \|^2_2$ 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 \( \{0, 1, \ldots, 2^j - 1\} \) into two halves \( I, I' \), and for the soft-thresholding we define

\[
t_I = \arg \min_{t \geq 0} \sum_{k \in I} (1 - 2 \cdot \mathbb{1}(|y_{j,k}| \leq t)) t^2 + (\min\{t, |y_{j,k}|\})^2
\]

\[
t_{I'} = \arg \min_{t \geq 0} \sum_{k \in I'} (1 - 2 \cdot \mathbb{1}(|y_{j,k}| \leq t)) t^2 + (\min\{t, |y_{j,k}|\})^2
\]

where \((y_{j,k})_{0 \leq k < 2^j - 1}\) 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 \( \hat{f}_I \) 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 \( s > \frac{1}{2} - \frac{1}{p} \), the SureShrink estimator \( \hat{f} \) essentially performs as well as the optimal thresholding estimator, and thus attains the minimax risk over Besov balls \( B^{p,q}_{s,r}(L) \) within multiplicative constants:

\[
\sup_{f \in B^{p,q}_{s,r}(L)} \mathbb{E}_f \|\hat{f} - f\|_2^2 \leq (1 + o(1)) \cdot \inf_{(t_j)_{j \geq 0}} \sup_{f \in B^{p,q}_{s,r}(L)} \mathbb{E}_f \|\hat{f}_I - f\|_2^2
\]

\[
\lesssim \inf_{\hat{f}} \sup_{f \in B^{p,q}_{s,r}(L)} \mathbb{E}_f \|\hat{f} - f\|_2^2.
\]

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 \textit{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 \( B^{p,q}_{s,r}(L) \) is given by

\[
\left( \inf_{\hat{f}} \sup_{f \in B^{p,q}_{s,r}(L)} \mathbb{E}_f \|\hat{f} - f\|_r^r \right)^{\frac{1}{r}} \approx \begin{cases} 
(2^2 \log(1/e))^{\frac{\sigma}{2(p-r)\pi}} \left( \frac{\log(1/e)}{\sigma^{\frac{1}{2(p-r)\pi} + (2s + 1)/p}} \right)^{\frac{1}{2}\frac{\sigma}{p} + (2s + 1)/p} & \text{if } r < (2s + 1)p \\
(2^2 \log(1/e))^{\frac{\sigma}{2(p-r)\pi}} & \text{if } r = (2s + 1)p \\
(2^2 \log(1/e))^{\frac{\sigma}{2(p-r)\pi}} & \text{if } r > (2s + 1)p
\end{cases}
\]


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

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

(61)

where the infimum is taken over all possible linear estimators.