

Lecture 19: Denoising via shrinkage

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

Optional part. This lecture belongs to optional Part VI and is not examined. Its high, mid and low labels organize a suggested reading path within the optional material; they do not add exam requirements. See the Part VI overview.

High priority *Overview*

Shrinkage becomes more effective when it matches the signal structure. A sparse representation suggests thresholding coordinates rather than shrinking every direction equally. We connect smoothness to small transform coefficients, analyze **soft thresholding** in the Gaussian sequence model, and derive its relationship with orthogonal-design LASSO. Wavelets, sharp minimax constants and nonconvex penalties remain optional extensions.

Recap from earlier lectures

From Lectures 1 and 18: earlier ideas used here

Reminder only: this block introduces no new exam requirements.

- **James–Stein shrinkage.** A data-dependent radial correction can lower total Gaussian mean risk. Used to contrast global and coordinatewise shrinkage.
- **Gaussian transformations.** Orthogonal transformations preserve spherical Gaussian noise. Used in the sequence model.
- **Gaussian tails.** Exponential tail bounds and a union bound control the largest noise coordinate. Used to choose a threshold.

An asterisk in a topic heading marks optional material; the priority labels specify the reading scope.

Learning goals

- Distinguish exact sparsity from accurate sparse approximation.
- Explain how orthogonal transforms preserve Gaussian noise and squared error.
- Compare hard and soft thresholding, and interpret an oracle risk bound.
- Derive the LASSO threshold under each stated normalization.

High priority *Introduction*

We have seen that Stein’s shrinkage may help improve estimation. However, the effect of shrinkage is most pronounced if we exploit the *sparsity* structure, which is common in many problems.

1 Sparsity in data analysis

1.1 Motivation

Mid priority *Motivation*

Q: Why is sparsity important?

- (1) Occam's razor. Better interpretability. Less is more.
- (2) Variable selection. We collect more and more covariates or features in "Big Data Age", but important ones are not many.
- (3) Biological mechanism. Nature favors sparsity. Sparse coding in primary visual cortex, genetics, etc.
- (4) Computation and memory. Data compression.

1.2 Quantifying sparsity

High priority *Quantifying sparsity*

For a vector $x \in \mathbb{R}^n$, we say that x is s -sparse if

$$\|x\|_0 := \sum_{j=1}^n \mathbb{1}(x_j \neq 0) \leq s.$$

In other words, the number of nonzeros in x is no larger than s . For this reason, it is often called exact sparsity. Despite its conventional name, $\|x\|_0$ is a count, not a norm.

We can relax this definition by allowing entries of x to have very small nonzero values. For $q \in (0, 2]$, we use the ℓ_q norm for $q \geq 1$ and the ℓ_q quasinorm for $0 < q < 1$

$$\|x\|_q := \left(\sum_{j=1}^n |x_j|^q \right)^{1/q} \leq R.$$

The radius $R > 0$ measures coefficient size; it is not a count of nonzero entries.

First, when $q < 1$, the above definition is only a quasinorm, since it does not satisfy the triangle inequality, but people often call it ℓ_q norm anyway. Second, as q gets smaller, the measurement is more stringent: the region of $\|x\|_q \leq 1$ is smaller (Figure 1). Third, we have $\|x\|_0 = \lim_{q \rightarrow 0} \|x\|_q^q$.

Sparse approximation. Exact sparsity counts nonzero coordinates. Approximate sparsity instead asks whether retaining a few large coordinates gives small error in the chosen norm.

High priority *Practice: A nearly sparse vector*

Exercise prompt. For $x = (1, 1/2, 1/4, 1/8)$, compute its exact sparsity and the squared error from retaining only its two largest coordinates.

See Exercise 1 in Section 8 for the full statement and complete solution.

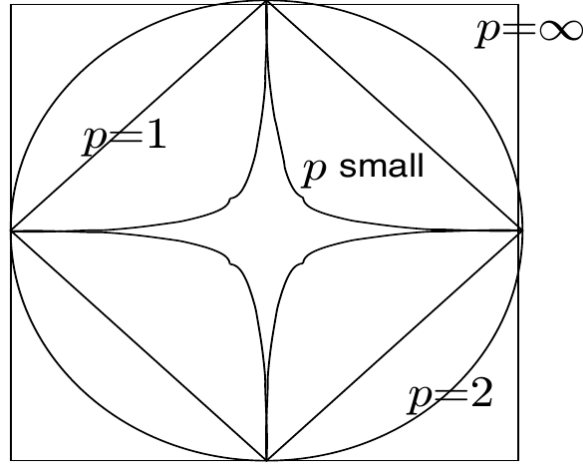


Figure 1: As the exponent decreases, the unit region narrows toward the coordinate axes. The source labels the exponent p ; it is our q , not the ambient dimension. Figure: Johnstone, Figure 1.5.

1.3 Smoothness and sparse representations

Mid priority *Smoothness and sparse representations*

Smoothness is an important property in many fields and arises naturally in physical systems. Let's consider a simple analysis to show that smoothness and sparsity are closely related.

Suppose f is a real, 2π -periodic C^m function, with $m \geq 1$. Its Fourier expansion (in L^2) is

$$f(x) = \frac{A_0}{2} + \sum_{k=1}^{\infty} [A_k \cos(kx) + B_k \sin(kx)]$$

The coefficients are determined by

$$C_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) e^{ikx} dx$$

with $A_0 = C_0$, and A_k, B_k the real and imaginary parts of C_k for $k \geq 1$. Assume the m th derivative is bounded:

$$\sup_{x \in [-\pi, \pi]} \left| \frac{d^m}{dx^m} f(x) \right| \leq C.$$

Assume also that $f^{(j)}(-\pi) = f^{(j)}(\pi)$ for $j = 0, \dots, m-1$, so all boundary terms cancel. Then repeated integration by parts gives

$$C_k = \frac{(-1)^m}{(ik)^m} \cdot \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{d^m}{dx^m} f(x) e^{ikx} dx \quad \implies \quad |C_k| \leq \frac{2C}{k^m}$$

So the high-frequency coefficients decay very fast. (We can measure the sparsity using ℓ_q norm with $mq > 1$, since $\sum k^{-mq} < \infty$).

2 Example: Wavelet transform*

Low priority *Example: Wavelet transform*

We show an example of the NMR (nuclear magnetic resonance) data taken from [1].

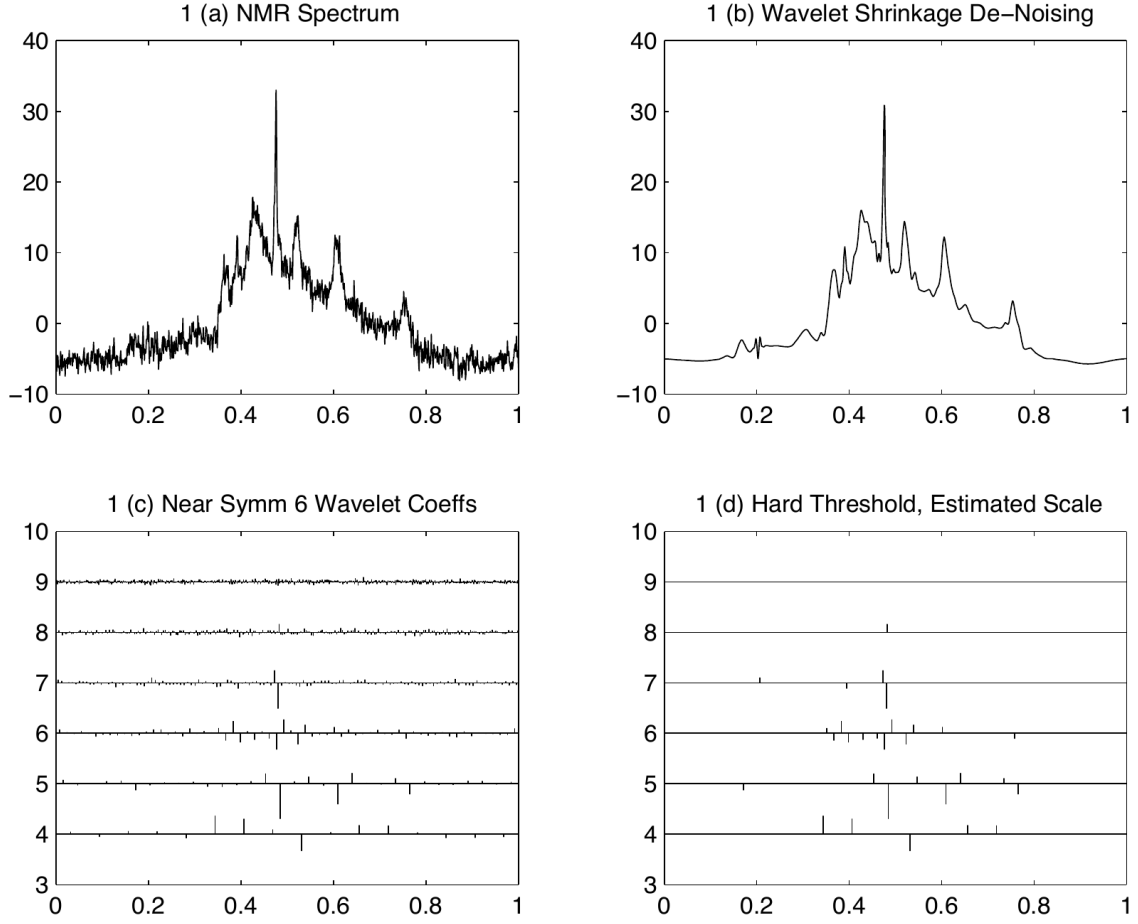


Figure 2: Noisy and denoised NMR signals above; wavelet coefficients before and after thresholding below. Figure: Johnstone, Figure 1.2; data credited there to Chris Raphael and Andrew Maudsley's laboratory.

2.1 Discrete Fourier transform

Low priority *Discrete Fourier transform*

Given $X \in \mathbb{R}^N$, define

$$\hat{X}_k = \sum_{n=0}^{N-1} X_n \exp\left(-\frac{i2\pi kn}{N}\right), \quad k = 0, 1, \dots, N-1.$$

Let

$$F = \left(\exp\left(-\frac{i2\pi kn}{N}\right) \right)_{0 \leq k, n < N} \in \mathbb{C}^{N \times N},$$

satisfy $F^*F = NI_N$; the normalized matrix $F/\sqrt{N}$ is unitary (the complex analogue of orthogonal). Then,

$$\hat{X} = FX,$$

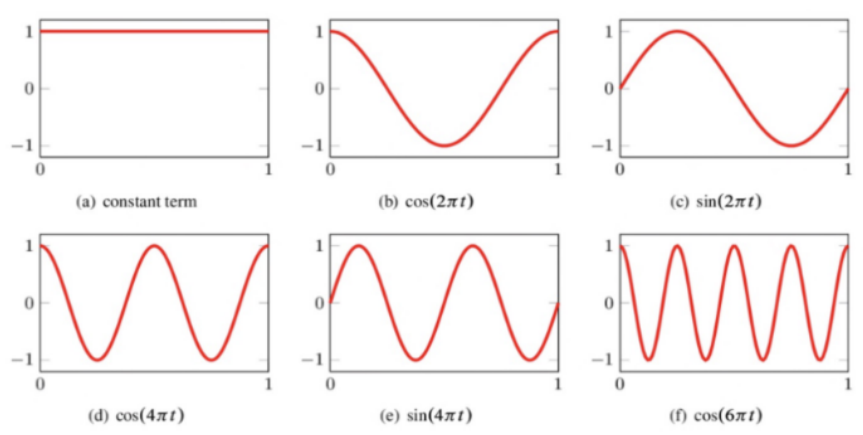


Figure 3: Sines and cosines oscillate across the whole interval. Here the horizontal coordinate $t \in [0, 1]$ gives the unit-period functions $\cos(2\pi kt)$ and $\sin(2\pi kt)$; rescaling $x = 2\pi t$ gives the earlier 2π -periodic convention. Wavelets add localization, which is useful for abrupt features. Diagram retained from the original course notes; for the Fourier–wavelet comparison see Mallat, *A Wavelet Tour of Signal Processing*, Chapter 1.

is essentially a change of basis, where

$$F : \text{time domain} \longleftrightarrow \text{frequency domain}.$$

2.2 Wavelet transform

Low priority *Wavelet transform*

$$\begin{cases} \text{Fourier: localized in frequency.} \\ \text{Wavelet: localized in both time and frequency.} \end{cases}$$

A generic form:

$$\psi_{jk}(t) = 2^{j/2} \psi(2^j t - k),$$

where in the RHS, ψ is the mother wavelet, 2^j is the scale, and k is the shift. For a suitable orthonormal mother wavelet, $\{\psi_{jk}\}_{j,k \in \mathbb{Z}}$ forms an orthonormal basis for $L_2(\mathbb{R})$. Pair (j, k) indicates wavelet ψ_{jk} , at resolution j and approximate location $k2^{-j}$. For discrete data $X = (X_l)_{l=1}^N$, define $\hat{X} = WX$, where $W \in \mathbb{R}^{N \times N}$ satisfies $W^\top W = I_N$ and is a discrete wavelet transform¹. Often $\hat{X}$ is sparse or approximately sparse; $X = W^\top \hat{X}$ reconstructs the signal from those coefficients.

$$\begin{cases} \text{Smoothness of functions} \implies \text{Fourier coefficients decay fast.} \\ \text{(Piecewise) smoothness of functions} \implies \text{Wavelet coefficients sparse / decay fast.} \end{cases}$$

¹For standard compactly supported wavelets, applying the transform can take $O(N)$ operations using a filter algorithm. Storing an arbitrary dense $N \times N$ matrix would instead require N^2 entries.

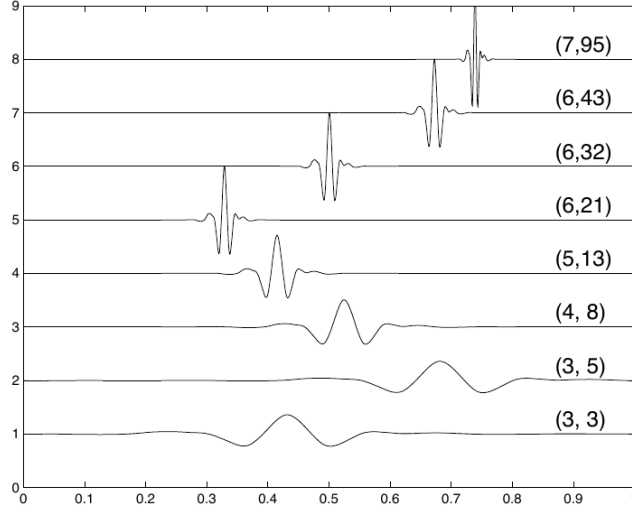


Figure 4: Changing j changes resolution; changing k shifts location. Curves are vertically offset for display. Figure: Johnstone, Figure 7.1, left panel (Symmlet 8 wavelets).

3 Thresholding

3.1 Gaussian sequence model

High priority *Gaussian sequence model*

We assume that our observations follow the model:

$$Y_k = \theta_k + \sigma Z_k, \quad \text{where } Z_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \quad (1)$$

The observations (Y_k) can be regarded as transformed or preprocessed measurements. We have raw measurements $X \in \mathbb{R}^n$, target (unknown) discretized function $(f(t_k))_{k \leq n}$,

$$X_k = f(t_k) + \sigma \omega_k, \quad \omega_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

Apply orthogonal transform $Y = W X$ where W is an orthogonal matrix (constructed from a real orthonormal Fourier basis or an orthogonal wavelet transform). Then we introduce

$$(\theta_1, \dots, \theta_n)^\top = W(f(t_1), \dots, f(t_n))^\top.$$

Since a standard Gaussian random vector remains standard Gaussian after orthogonal transformation, we obtain (1). In vector form,

$$Y \sim \mathcal{N}(\theta, \sigma^2 I_n). \quad (2)$$

Assumption. We assume that $\theta \in \mathbb{R}^n$ is sparse, such that

$$(\text{exact sparsity}) \quad \|\theta\|_0 = \sum_{j=1}^n \mathbb{1}(\theta_j \neq 0) \leq S.$$

Orthogonal denoising. A real orthogonal transform preserves independent spherical Gaussian noise and squared Euclidean error. This lets us estimate sparse coefficients and transform the estimate back without changing the risk.

High priority *Practice: Risk is preserved by an orthogonal transform*

Exercise prompt. If $Y = WX$ and $\theta = Wf$ for a real orthogonal W , show that $\hat{f} = W^\top \hat{\theta}$ has the same squared error as $\hat{\theta}$.

See Exercise 2 in Section 8 for the full statement and complete solution.

3.2 Thresholding functions

High priority *Thresholding functions*

Without the sparsity assumption, we may estimate $\theta \in \mathbb{R}^n$ via either the unbiased natural estimator Y or James–Stein estimator $(1 - \frac{(n-2)\sigma^2}{\|Y\|^2})Y$ (assuming $n \geq 3$ and σ^2 is known).

We exploit the sparsity assumption by estimation via two typical thresholding functions $\hat{\theta}_j = T_\tau(Y_j)$.

(1) **Hard Thresholding.**

$$\hat{\theta}_j = \begin{cases} Y_j, & \text{if } |Y_j| > \tau \\ 0, & \text{if } |Y_j| \leq \tau. \end{cases}$$

(2) **Soft Thresholding.**

$$T_\tau(u) = \text{sgn}(u)(|u| - \tau)_+ = \arg \min_{v \in \mathbb{R}} \left\{ \frac{1}{2}(v - u)^2 + \tau|v| \right\}.$$

Determining thresholding level τ . From lectures on convergence modes, we know that

$$\frac{\|Z\|_\infty}{\sqrt{2 \log n}} \xrightarrow{p} 1 \quad \text{as } n \rightarrow \infty.$$

For $n \geq 2$ and known $\sigma > 0$, the universal threshold is

$$\tau = \sqrt{2\sigma^2 \log n}.$$

A finite-sample choice makes the confidence level explicit. A Gaussian tail bound and union bound give

$$\mathbb{P} \left(\max_j |\sigma Z_j| > \sigma \sqrt{2 \log(2n/\alpha)} \right) \leq \alpha, \quad 0 < \alpha < 1.$$

At the universal threshold, Mills' bound gives the sharper estimate

$$\mathbb{P} \left(\max_j |Z_j| > \sqrt{2 \log n} \right) \leq \frac{1}{\sqrt{\pi \log n}} \rightarrow 0.$$

Thus even the uninflated universal threshold removes all pure-noise coordinates with probability tending to one. Choosing $1.01\sqrt{2\sigma^2 \log n}$ also works, directly from the coarser union bound. Nonzero signal coordinates still incur estimation error.

Summary. The transform-shrinkage approach is

$$(X_k) \xrightarrow{W} Y \xrightarrow[\text{Shrinkage}]{T_\tau} \hat{\theta} \xrightarrow{W^{-1}} (\hat{f}(t_k)).$$

Instructor's Note Spend error where the signal is

If most true coordinates are zero, carrying one unit of noise in every coordinate is wasteful. Thresholding accepts some bias near the cutoff to silence many empty coordinates. Large surviving coordinates still pay the soft-threshold offset, which is why the choice of cutoff matters.

Picture: hard and soft thresholding. For fixed observations $u \in [-4, 4]$, compare $H_\tau(u) = uI\{|u| > \tau\}$ with $T_\tau(u) = \text{sgn}(u)(|u| - \tau)_+$. At $\tau = 1$ and $u = (-2, -0.5, 0, 0.5, 2)$, **hard thresholding** gives $(-2, 0, 0, 0, 2)$ and soft thresholding gives $(-1, 0, 0, 0, 1)$. The hard rule jumps at the cutoff; the soft rule is continuous and shrinks every surviving value.

Thresholding and normalization. Hard thresholding keeps large observed coordinates unchanged; soft thresholding also shrinks them toward zero. The threshold in orthogonal LASSO depends on both the loss normalization and the design normalization.

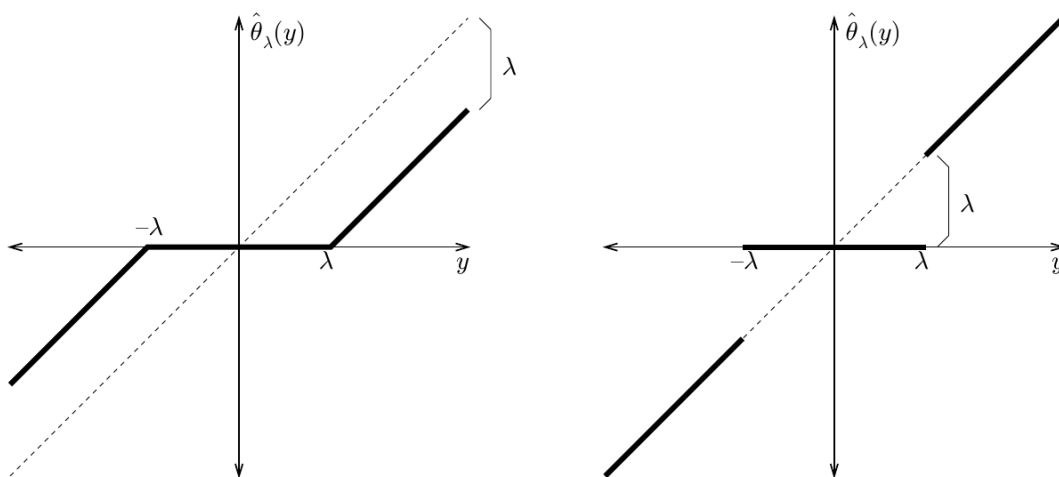


Figure 5: Soft thresholding (left) is continuous; hard thresholding (right) jumps. The source uses $y = u$ and $\lambda = \tau$. At equality, our hard rule returns zero. Figure: Johnstone, Figure 2.2.

High priority Practice: Threshold a new observation vector

Exercise prompt. For $y = (-3, -1, 0, 1, 3)$ and $\tau = 2$, calculate hard and soft threshold estimates. See Exercise 3 in Section 8 for the full statement and complete solution.

4 Denoising theory

4.1 General oracle inequalities

Mid priority *General oracle inequalities*

Let us evaluate how well thresholding estimators perform. Consider the MSE as risk function $R(\theta) = \mathbb{E} \|\hat{\theta} - \theta\|_2^2$ for an estimator $\hat{\theta}$. In this section, our results hold *without any sparsity assumption*.

Suppose there is an oracle, who tells us the locations (indices) of large signals θ_j : we estimate θ_j by Y_j and incur MSE σ^2 , and by 0 and incur MSE θ_j^2 otherwise.

We define the (oracle-aided) ideal risk as follows.

$$R(\theta, \sigma) = \sum_{i=1}^n \min\{\theta_i^2, \sigma^2\}.$$

In particular, if θ is S -sparse, then

$$R(\theta, \sigma) \leq S\sigma^2.$$

Theorem 1. *Suppose $Y \sim \mathcal{N}(\theta, \sigma^2 I_n)$. For the soft-thresholding estimator $\hat{\theta}_{\tau_n} = T_{\tau_n}(Y)$ with the level $\tau_n = \sqrt{2\sigma^2 \log n}$,*

$$\mathbb{E}[\|\hat{\theta}_{\tau_n} - \theta\|_2^2] \leq (2 \log n + 1)[\sigma^2 + R(\theta, \sigma)].$$

The proof is given in [1, Proposition 8.8, p. 227]; the linked text in Further reading develops the coordinatewise oracle bound. Similar bounds hold for hard thresholding. One striking aspect of this result is that we don't know the indices of large signals among $\theta_1, \dots, \theta_n$. Compared with the ideal risk, the risk is increased only by a log factor.

Remark 1. *Theorem 1 holds with no assumption on θ . Without assumption on θ , this is not improvable due to the following minimax result [1, Eq. (8.35), Section 8.6]. As $n \rightarrow \infty$,*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^n} \frac{\mathbb{E}[\|\hat{\theta} - \theta\|_2^2]}{\sigma^2 + \sum_{i=1}^n \min\{\theta_i^2, \sigma^2\}} \geq (2 \log n) \cdot (1 + o(1)).$$

4.2 Oracle inequalities for S -sparse signals

Low priority *Oracle inequalities for S -sparse signals*

Denote the space of S -sparse vectors as

$$\Theta_{n, S_n} = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq S_n\}.$$

Theorem 2. *Suppose the integers $1 \leq S_n < n$ satisfy*

$$\frac{S_n}{n} \rightarrow 0, \quad n \rightarrow \infty.$$

Then, as $n \rightarrow \infty$,

(1) *The soft-thresholding estimator $\hat{\theta}_{\tau_n}$ with $\tau_n = \sqrt{2\sigma^2 \log(n/S_n)}$,*

$$\sup_{\theta \in \Theta_{n, S_n}} \mathbb{E}[\|\hat{\theta}_{\tau_n} - \theta\|_2^2] \leq (1 + o(1)) \cdot 2S_n\sigma^2 \log\left(\frac{n}{S_n}\right).$$

(2) The soft-thresholding estimator matches the minimax lower bound:

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_{n, S_n}} \mathbb{E} \|\hat{\theta} - \theta\|^2 \geq (1 + o(1)) \cdot 2S_n \sigma^2 \log \left(\frac{n}{S_n} \right).$$

Proof of (1). Let us simply write $\hat{\theta}$ instead of $\hat{\theta}_{\tau_n}$ and define

$$\mathcal{S} = \{i \leq n : \theta_i \neq 0\}.$$

The high-level idea is the following: if $i \notin \mathcal{S}$, most noise Y_i will be denoised by $\hat{\theta}$. If $i \in \mathcal{S}$, the estimator incurs additional bias. We will conduct coordinatewise analysis since

$$\|\hat{\theta} - \theta\|^2 = \sum_{j=1}^n (\hat{\theta}_j - \theta_j)^2$$

and divide our analysis into two cases.

Case I: If $j \notin \mathcal{S}$, namely $\theta_j = 0$,

$$\begin{aligned} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] &= \mathbb{E}[(\sigma|Z_j| - \tau)_+^2] \\ &= \sigma^2 \int_0^\infty \mathbb{P}((|Z_j| - \frac{\tau}{\sigma})_+^2 > t) dt \quad (\text{by the tail-integral identity}) \\ &= \sigma^2 \int_0^\infty \mathbb{P}(|Z_j| - \frac{\tau}{\sigma} > \sqrt{t}) dt \\ &\leq 2\sigma^2 \int_{\tau/\sigma}^\infty (u - \frac{\tau}{\sigma}) \mathbb{P}(|Z_j| \geq u) du \quad (\text{change of variables } u := \sqrt{t} + \frac{\tau}{\sigma}) \\ &\leq 4\sigma^2 \int_{\tau/\sigma}^\infty u \exp\left(-\frac{u^2}{2}\right) du \quad (\text{by Gaussian tail probability bound}) \\ &= 4\sigma^2 \exp\left(-\frac{\tau^2}{2\sigma^2}\right). \end{aligned}$$

Case II: If $j \in \mathcal{S}$ namely $\theta_j \neq 0$,

$$\begin{aligned} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] &= \mathbb{E}[(\hat{\theta}_j - T_\tau(\theta_j) + T_\tau(\theta_j) - \theta_j)^2] \\ &\leq \mathbb{E}[2(T_\tau(Y_j) - T_\tau(\theta_j))^2 + 2(T_\tau(\theta_j) - \theta_j)^2] \quad (\text{since } (a+b)^2 \leq 2a^2 + 2b^2) \\ &\leq \underbrace{2\mathbb{E}[(Y_j - \theta_j)^2]}_{\text{estimation error}} + \underbrace{2\tau^2}_{\text{squared threshold offset}} \quad (\text{since } T_\tau \text{ is 1-Lipschitz}) \\ &= 2\sigma^2 + 2\tau^2. \end{aligned} \tag{3}$$

Combining the two cases: there are $n - |\mathcal{S}| \leq n$ indices in Case I and $|\mathcal{S}| \leq S_n$ indices in Case II, so

$$\begin{aligned} \mathbb{E}[\|\hat{\theta} - \theta\|^2] &\leq 2S_n(\sigma^2 + \tau_n^2) + 4n\sigma^2 \exp\left(-\frac{\tau_n^2}{2\sigma^2}\right) \\ &= 2S_n(\sigma^2 + 2\sigma^2 \log\left(\frac{n}{S_n}\right)) + 4S_n\sigma^2 \\ &= 6S_n\sigma^2 + 4S_n\sigma^2 \log\left(\frac{n}{S_n}\right) \end{aligned}$$

For the sharp leading constant, use $(a+b)^2 \leq (1+\eta)a^2 + (1+\eta^{-1})b^2$ with $a = T_\tau(\theta_j) - \theta_j$ and $b = T_\tau(Y_j) - T_\tau(\theta_j)$. The signal-coordinate risk is at most $(1+\eta)\tau^2 + (1+\eta^{-1})\sigma^2$. Put $L = \log(n/S_n) \rightarrow \infty$ and $\eta = L^{-1/2}$. Summing yields $2S_n\sigma^2 L + O(S_n\sigma^2\sqrt{L}) + O(S_n\sigma^2)$, proving the claimed leading constant. Assume $1 \leq S_n < n$. $\square$

Remark 2 (Comparison with James–Stein estimation). *Global James–Stein shrinkage does not exploit sparse coordinate structure. For $\sigma = 1$ and $\|\theta_n\|^2 = n$, its exact ordinary-JS risk is*

$$R_{\text{JS}}(\theta_n) = n - (n - 2)^2 \mathbb{E} \|Y\|^{-2}.$$

Here $Q = \|Y\|^2$ is a sum of squares of independent unit-variance Gaussians with squared means summing to n ; this is the noncentral chi-square distribution with n degrees of freedom and noncentrality n . Its mean is $2n$ and variance $6n$, so $Q/n \rightarrow_p 2$. To control its reciprocal, use the mixture representation $Q \mid J \sim \chi_{n+2J}^2$ with $J \sim \text{Poisson}(n/2)$. Since $\mathbb{E}(\chi_k^2)^{-2} = 1/[(k-2)(k-4)]$ for $k > 4$, we obtain $\mathbb{E} Q^{-2} \leq 1/[(n-2)(n-4)]$ for $n > 4$. Thus n/Q is uniformly integrable.² Hence $\mathbb{E}(n/Q) \rightarrow 1/2$ and $R_{\text{JS}}/n \rightarrow 1/2$. For an S_n -sparse vector with nonzero magnitude $\sqrt{n/S_n}$, the same norm condition holds. If $S_n \log(n/S_n) = o(n)$, soft thresholding has risk $o(n)$ by the preceding theorem. Global shrinkage responds to total signal energy; coordinatewise thresholding also uses where that energy is concentrated.

Remark 3 (Bias-variance tradeoff). *The threshold $\sqrt{2\sigma^2 \log(n/S_n)}$ is smaller than the universal level $\sqrt{2\sigma^2 \log n}$. It has a smaller thresholding level, so many noise coordinates are not completely removed (reaching 0). This smaller thresholding level introduces a bit more variance from the noises, but incurs much smaller bias on large signals, thus leading to the minimax risk.*

5 LASSO with orthogonal design

High priority *LASSO with orthogonal design*

Suppose we have data $(x_i, y_i)_{1 \leq i \leq n}$, where $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. As usual, we write our data in matrix form

$$\begin{cases} X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}, \\ Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n. \end{cases}$$

The celebrated LASSO, an abbreviation for *least absolute shrinkage and selection operator*, solves the following optimization problem. Given $\lambda \geq 0$, solve

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \|X\theta - Y\|_2^2 + \lambda \|\theta\|_1 =: \min_{\theta \in \mathbb{R}^p} \ell_{n,\lambda}(\theta).$$

5.1 Connecting LASSO with denoising

High priority *Connecting LASSO with denoising*

Assume $X^\top X = I_p$, with $p \leq n$, and write y for the observed response vector Y . Write $z = X^\top y \in \mathbb{R}^p$ and $r = (I_n - XX^\top)y$. The residual r is orthogonal to every column of X , so

$$\|X\theta - y\|_2^2 = \|\theta - z\|_2^2 + \|r\|_2^2.$$

The square case $n = p$, $X \in O(n)$ has $r = 0$. In general,

$$\ell_{n,\lambda}(\theta) = \frac{\|r\|_2^2}{2n} + \frac{1}{n} \sum_{j=1}^p \left[\frac{1}{2} (\theta_j - z_j)^2 + n\lambda |\theta_j| \right].$$

²Uniform integrability means that the contribution of large values to the expectation is uniformly small: $\sup_n \mathbb{E}[|Z_n| \mathbf{1}_{\{|Z_n| > M\}}] \rightarrow 0$ as $M \rightarrow \infty$. A uniform bound on $\mathbb{E}|Z_n|^2$ is sufficient, since the tail expectation is at most $\sup_n \mathbb{E}|Z_n|^2/M$. Together with convergence in probability, this condition allows passage to the limit in expectations. Here take $Z_n = n/Q$ for $n \geq 5$.

The constant does not affect minimization, and each remaining term involves only one coordinate. Therefore

$$\hat{\theta} = (T_{n\lambda}(z_1), \dots, T_{n\lambda}(z_p))^\top.$$

The scalar minimization is proved in Exercise 5. We can analyze LASSO with an orthogonal design using our results from denoising.

Under the alternative convention $X^\top X = nI_p$, use $z = X^\top y/n$. Then the objective is a constant plus $\sum_j [\frac{1}{2}(\theta_j - z_j)^2 + \lambda|\theta_j|]$, and the threshold is λ . For a fixed design and noise $\varepsilon \sim N(0, \sigma^2 I_n)$ in $y = X\theta^* + \varepsilon$, the transformed noise variance is σ^2 under $X^\top X = I_p$ and σ^2/n under $X^\top X = nI_p$. Always read the loss and design normalizations together. In particular, Lecture 1 used $(v - u)^2 + \lambda|v|$, whose threshold is $\lambda/2$; the scalar loss here has an extra factor $1/2$ and threshold τ .

High priority *Practice: Check the LASSO normalization*

Exercise prompt. If $n = 10$, $X^\top X = I$, $z = X^\top y = (3, -1)$ and $\lambda = 0.2$, find the LASSO solution for loss $\|X\theta - y\|^2/(2n)$.

See Exercise 4 in Section 8 for the full statement and complete solution.

5.2 Reducing shrinkage bias*

Low priority *Reducing shrinkage bias*

For the statistical interpretation, assume the fixed-design model

$$y = X\theta^* + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \quad X^\top X = I_p.$$

Then $z = X^\top y = \theta^* + X^\top \varepsilon \sim N(\theta^*, \sigma^2 I_p)$. Even when z_j is large (strong signal), we shrink our estimate by $n\lambda$ because $T_{n\lambda}(z_j)$ always applies shrinkage. At a fixed positive threshold, the bias approaches a nonzero offset as the signal magnitude grows.

Two improvements: SCAD (smoothly clipped absolute deviation) and MCP (minimax concave penalty). Consider the scalar case $\ell_{1,\lambda}(\theta)$ where $\theta \in \mathbb{R}$:

$$\ell_{1,\lambda}(\theta) = \frac{1}{2}\|\theta - z\|^2 + P_\lambda(\theta)$$

where the regularizer $P_\lambda(\theta)$ is depicted in Figure 6.

- Instead of the ℓ_1 regularizer, we use a different regularizer P_λ that is flat after a threshold.
- For larger signals, it applies small shrinkage to incur less bias.

Like LASSO, SCAD and MCP are nondifferentiable at $\theta = 0$. At a differentiable local minimizer $\hat{\theta} > 0$, the necessary stationarity condition is

$$\ell'_{1,\lambda}(\theta)|_{\theta=\hat{\theta}} = 0. \quad \Longleftrightarrow \quad \hat{\theta} - z + P'_\lambda(\hat{\theta}) = 0.$$

If $P'_\lambda(\hat{\theta}) = 0$, then the penalized estimate equals z in that region. This removes the local shrinkage offset; it does not claim globally zero statistical bias after averaging over all noisy observations. We note that

$$\begin{cases} P'_\lambda(\hat{\theta}) \equiv \lambda, & \text{in LASSO, inducing a bias regardless} \\ P'_\lambda(\hat{\theta}) = 0, & \text{if } \hat{\theta} \text{ large in SCAD / MCP, giving no shrinkage offset in this region.} \end{cases}$$

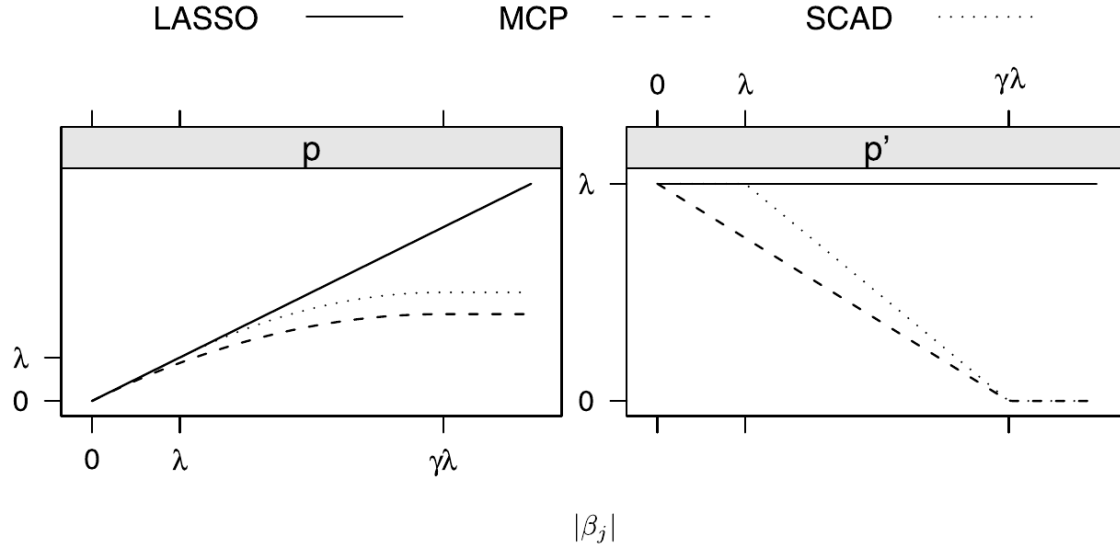


Figure 6: Penalties on the left and their derivatives for positive coefficient magnitudes on the right. SCAD and MCP flatten for large coefficients, reducing the shrinkage of strong signals. Here p denotes the scalar penalty, λ its initial slope, and γ a shape parameter. Figure: Breheny and Huang (2011), Figure 1.

Pros and Cons. Compared with LASSO, MCP and SCAD can suppress small coordinates while reducing shrinkage bias for large signals. Under suitable assumptions this can improve statistical accuracy. However, unlike LASSO, they involve a nonconvex optimization problem! It is harder (though still possible under additional conditions) to solve the following nonconvex problem under a general design matrix.

$$\frac{1}{2} \|y - X\theta\|_2^2 + \sum_{j=1}^p P_\lambda(\theta_j).$$

6 Further reading

Further reading *Further reading*

The author's open book gives the proof of the oracle upper bound on p. 227 and the matching lower bound in Section 8.6. The main reference for shrinkage theory is [1]. The theorem in Section 4.2 can be found in Section 8.4 of [1].

Key ideas

- Sparse approximation depends on the representation: many transform coefficients can be small even when the raw signal is not sparse.
- Orthogonal transformations preserve squared error and spherical Gaussian noise.
- Thresholding suppresses small coordinates; an oracle bound measures the price of not knowing which coordinates matter.
- The soft-threshold level follows from both the loss normalization and the column lengths.

7 Detailed reference

High priority *Sparse approximation*

Exact sparsity counts nonzero coordinates. Approximate sparsity instead asks whether retaining a few large coordinates gives small error in the chosen norm. Return to the earlier explanation.

High priority *Orthogonal denoising*

A real orthogonal transform preserves independent spherical Gaussian noise and squared Euclidean error. This lets us estimate sparse coefficients and transform the estimate back without changing the risk. Return to the earlier explanation.

High priority *Thresholding and normalization*

Hard thresholding keeps large observed coordinates unchanged; soft thresholding also shrinks them toward zero. The threshold in orthogonal LASSO depends on both the loss normalization and the design normalization. Return to the earlier explanation.

Looking ahead

Lecture 20 removes the orthogonal-design simplification. The next question is whether a design preserves enough information about sparse signals to support recovery and estimation. Continue within optional Part VI.

8 Exercises and complete solutions

High priority *A nearly sparse vector*

Exercise 1 (A nearly sparse vector).

For $x = (1, 1/2, 1/4, 1/8)$, compute its exact sparsity and the squared error from retaining only its two largest coordinates.

Solution.

Every coordinate is nonzero, so $\|x\|_0 = 4$. The truncated vector is $(1, 1/2, 0, 0)$ and its squared error is $1/16 + 1/64 = 5/64$. Exact sparsity and accurate sparse approximation are different notions. $\square$

High priority *Risk is preserved by an orthogonal transform*

Exercise 2 (Risk is preserved by an orthogonal transform).

If $Y = WX$ and $\theta = Wf$ for a real orthogonal W , show that $\hat{f} = W^\top \hat{\theta}$ has the same squared error as $\hat{\theta}$.

Solution.

$\hat{f} - f = W^\top (\hat{\theta} - \theta)$, and orthogonal matrices preserve Euclidean norm. Thus $\|\hat{f} - f\|^2 = \|\hat{\theta} - \theta\|^2$ for each realization; taking expectation gives equal risks. $\square$

High priority *Threshold a new observation vector*

Exercise 3 (Threshold a new observation vector).

For $y = (-3, -1, 0, 1, 3)$ and $\tau = 2$, calculate hard and soft threshold estimates.

Solution.

Hard thresholding retains values with magnitude strictly above 2, giving $(-3, 0, 0, 0, 3)$. Soft thresholding subtracts 2 from each retained magnitude, giving $(-1, 0, 0, 0, 1)$. Values at exactly the cutoff would be zero under both stated conventions.

□

High priority *Check the LASSO normalization*

Exercise 4 (Check the LASSO normalization).

If $n = 10$, $X^\top X = I$, $z = X^\top y = (3, -1)$ and $\lambda = 0.2$, find the LASSO solution for loss $\|X\theta - y\|^2/(2n)$.

Solution.

The threshold is $n\lambda = 2$. Thus the solution is $(T_2(3), T_2(-1)) = (1, 0)$. Using threshold 0.2 would correspond to a different loss or design normalization.

□

High priority *Prove the soft-threshold formula*

Exercise 5 (Prove the soft-threshold formula).

Minimize $\frac{1}{2}(v - u)^2 + \tau|v|$ for $\tau \geq 0$ by elementary calculus.

Solution.

For $v > 0$, the derivative is $v - u + \tau$, yielding $v = u - \tau$ when $u > \tau$. For $v < 0$, it is $v - u - \tau$, yielding $v = u + \tau$ when $u < -\tau$. At zero, the left derivative is $-u - \tau$ and right derivative $-u + \tau$; they straddle zero exactly when $|u| \leq \tau$. Strict convexity of the quadratic gives a unique minimizer, the stated soft rule.

□

Mid priority *An ideal coordinate choice*

Exercise 6 (An ideal coordinate choice).

For $\theta = (0, \sigma/2, 2\sigma)$, compute the oracle risk and compare it with the raw-observation risk.

Solution.

The oracle contributions are 0, $\sigma^2/4$, and σ^2 , totaling $5\sigma^2/4$. The raw observation incurs $3\sigma^2$. Although there are two nonzero coordinates, the ideal risk is strictly less than $2\sigma^2$ because the small signal is better set to zero.

□

High priority *A finite-sample universal threshold*

Exercise 7 (A finite-sample universal threshold).

Choose a threshold that removes every coordinate of pure Gaussian noise with probability at least $1 - \alpha$ at a fixed n .

Solution.

A union bound and the two-sided Gaussian bound give $\mathbb{P}(\max_j |\sigma Z_j| > \tau) \leq 2n \exp[-\tau^2/(2\sigma^2)]$. Taking $\tau = \sigma\sqrt{2\log(2n/\alpha)}$ makes this at most α . This finite-confidence threshold differs from a purely asymptotic leading-order formula.

□

Mid priority *Fourier decay and an approximation bound*

Exercise 8 (Fourier decay and an approximation bound).

If $|C_k| \leq Ak^{-m}$, identify $q > 0$ for which $\sum_{k \geq 1} |C_k|^q < \infty$, and bound the squared tail for $m > 1/2$.

Solution.

The q -series is bounded by $A^q \sum k^{-mq}$, which converges when $mq > 1$. For the squared tail,

$$\sum_{k > K} |C_k|^2 \leq A^2 \int_K^\infty x^{-2m} dx = \frac{A^2 K^{1-2m}}{2m-1}.$$

For the real Fourier convention in Section 1.3, let $f_K = A_0/2 + \sum_{k=1}^K [A_k \cos(kx) + B_k \sin(kx)]$. Orthogonality gives the exact normalization

$$\frac{1}{\pi} \int_{-\pi}^{\pi} |f - f_K|^2 dx = \sum_{k > K} (A_k^2 + B_k^2) = \sum_{k > K} |C_k|^2.$$

Thus the bound above is the squared function-approximation error for the normalized integral; for the unnormalized integral multiply it by π .

□

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 16, 2026: Stated the regression noise model and coefficient noise scales, corrected support counting in the sparse-risk proof, and made the Fourier approximation normalization explicit. Added optional-part guidance, a backward recap and concise reading summaries. October 17, 2026: Published in optional Part VI with lecture links, student downloads and issue reporting; this material is not examined.

References

- [1] Iain Johnstone. *Gaussian estimation: Sequence and wavelet models*. 2017.