ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak

Lecture 21: Signal Subspaces and Sparsity

1 Signal Subspaces and Sparsity

Recall the classical linear signal model:

$$X = H\theta + w, \quad w \sim N(0, \sigma_w^2 I)$$

where $S = H\theta$, is a linear-parametric model for the signal and w is noise. Here H is a known $n \times k$ matrix, whose columns span the signal subspace, and $\theta \in \mathbb{R}^k$ are the signal parameters. The MLE of θ is:

$$\widehat{\theta}_{MLE} = (H^T H)^{-1} H^T X$$

and the MLE of the signal is:

$$\widehat{S} = H\widehat{\theta} = \underbrace{H(H^T H)^{-1} H^T}_{P_H} X$$

where $P_H = H(H^H H)^{-1} H^T$ is the orthogonal projection operator on to the signal subspace. The Bayes MMSE estimator based on a prior $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$ is the Wiener filter (posterior mean and MAP estimator):

$$\widehat{\theta}_{\text{Wiener}} \;=\; H^T \Big(HH^T + \frac{\sigma_w^2}{\sigma_\theta^2} I \Big)^{-1} X \;=\; \Big(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_w^2} \Big) \widehat{\theta}_{MLE}$$

This follows directly from the Gauss-Markov Theorem. And as the SNR grows

$$H^T \Big(HH^T + \frac{\sigma_w^2}{\sigma_\theta^2} I \Big)^{-1} \longrightarrow (H^T H)^{-1} H^T \qquad \text{as} \qquad \frac{\sigma_w^2}{\sigma_\theta^2} \longrightarrow \infty$$

So in the high SNR situation, the Wiener filter acts essentially the same as the MLE; it projects X onto the signal subspace. At low SNR the Wiener filter "shrinks" the MLE toward zero to balance the tradeoff between bias and variance.

2 Sparsity

In the classic set-up:

$$X = H\theta + w$$

we assume that we know the low-dimensional signal subspace. In many problems we may not have this information, but we might know that the signal lies in one of many subspaces in a certain trasform domain.

Example 1 *Narrowband Communications. The communication signal lies in one of many narrow frequency bands, but we may not know which band it will be in (e.g. frequency hopping communication). If x is the signal and U is the DFT, then $\theta = U^T x$ is a sparse vector (i.e., there are just a few non-zero frequencies), but it is not known which frequencies will have non-zero coefficients.*

Example 2 *Wavelet-based Image Processing. The discrete wavelet trasform (DWT) is very effective at compressing natural images. In fact it is the basis of the JPEG-2000 standard. The DWT of images tends to be "sparse" in the following sense. If x is an image and U denotes the DWT, then the DWT coefficients $\theta = U^T x$ tend to be mostly zero (or very nearly zero). The locations of the relatively few non-zero (or significant) coefficients in the vector θ depend on x in a complicated way. So, while images do approximately lie in a subspaces of the wavelet domain, the subspace is different for each different image.*

3 Sparse Signal Models

Let U be an $n \times n$ matrix whoes columns form an orthobasis for $\mathbb{R}^n$. For example, U could be the DFT or DWT. The signal of interest is represented in this domain as $s = U\theta$. Consider the observation model

$$x = U\theta + w, \; w \sim N(0, \sigma_w^2 I) \; .$$

An equivalent observation model is:

$$\begin{aligned} U^T x &= U^T U\theta + u^T w \\ &= \theta + w' \end{aligned}$$

where $w' \sim \mathcal{N}(0, \sigma_w^2 U^T U)$. Since the columns of U are orthonormal, $U^T U = I_{n\times n}$, and so $w' \sim \mathcal{N}(0, \sigma_w^2 I)$. Thus, after transforming the signal by U^T we have a direct observation of θ plus GWN. The problem of estimating θ is called *denoising*.

If we make no assumption about the θ, then we could use the MLE:

$$\widehat{\theta}_{\text{MLE}} \;\; = \;\; U^T x \; .$$

The MLE of the signal is then $\widehat{s}_{\text{MLE}} = U\widehat{\theta}_{\text{MLE}} = UU^T x = x$, since $UU^T = I_{n\times n}$. If we suppose that the coefficients tend to have a certain energy, then we could use the prior: $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$ and the Wiener filter:

$$\widehat{\theta}_{\text{Wiener}} \; = \; U^T \Big(UU^T + \frac{\sigma_w^2}{\sigma_\theta^2} I \Big)^{-1} x$$

Since U is an orthonormal transform $UU^T = I$ and the wiener filter simplifies to:

$$\begin{aligned} \widehat{\theta}_{\text{Wiener}} = \; & U^T \Big(I + \frac{\sigma_w^2}{\sigma_\theta^2} \Big)^{-1} x \\ = \; & \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_w^2} \right) U^T x \\ = \; & \Big(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_w^2} \Big) \widehat{\theta}_{\text{MLE}} \; , \end{aligned}$$

and we see that the Wiener filter is simply shrinking the MLE according to the SNR.

Now suppose our prior knowledge about θ is that it is sparse; i.e. many or most of the coefficients are zero (or near zero). This is not captured by the Gaussian prior, which models every coefficient as a Gaussian random variable with power σ_θ^2. If many coefficients are zero, then many should have approximately zero power! So we would like to design a prior probability density that reflects our belief that most of the coefficients are zero or near zero in magnitude.

Example 3 *Gaussian mixture. Let $\theta_1, \ldots, \theta_n$ denote the coefficients and model them as follows:*

$$\theta_i \;\; \overset{\text{iid}}{\sim} \;\; p\mathcal{N}(0, \sigma_0^2) \; + \; (1-p)\mathcal{N}(0, \sigma_1^2) \; , \; \textit{for } i = 1, \ldots, n$$

with $\sigma_0^2 << \sigma_1^2$ and $p \approx 1$. In words this prior is saying that a large fraction p of the coefficients tend to be very small in magnitude (i.e. $|\theta_i| \sim \sigma_0$) and $1-p$ tend to be large. An example is depicted in Figure 1.

Example 4 *Laplacian prior. Let $\theta_1, \ldots, \theta_n$ denote the coefficients and model them as follows:*

$$\theta_i \;\; \sim \;\; \frac{\lambda}{2} e^{-\lambda|\theta_i|} \;\; , \qquad i = 1, \ldots, n$$

We will focus on the Laplacian prior because it leads to very simple and intuitive solutions to the denoising problem and it is log-concave, which makes it computationablly tractable when used in inverse problems such as deconvolution. An example is depicted in Figure 2.

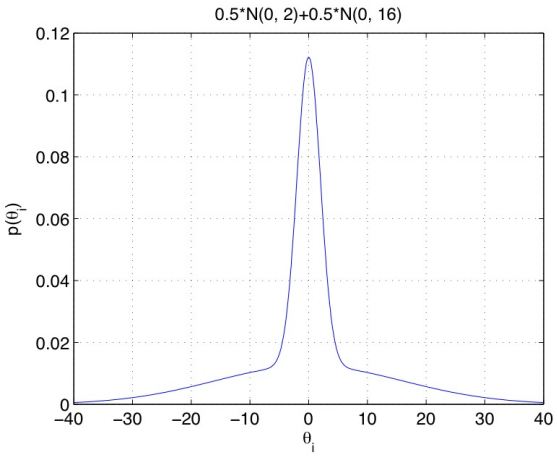

Figure 1: Example of Gaussian-mixture prior

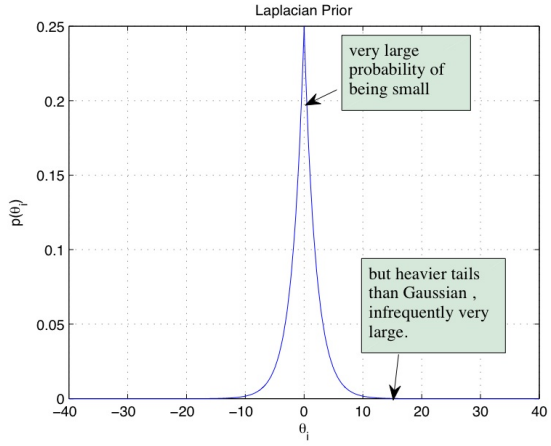

Figure 2: Example of Laplacian prior

4 Laplacian priors for sparsity

Assume the prior

$$p(\theta) \;=\; \prod_{i=1}^{n} p(\theta_i) \;=\; \prod_{i=1}^{n} \frac{\lambda}{2}\, e^{-\lambda|\theta_i|}$$

and observation model

$$x = U\theta + w \quad , \quad w \sim \mathcal{N}((0, \sigma^2 I)$$

or equivalently

$$U^T x = \theta + U^T w$$

Recall that $U^T w \sim \mathcal{N}(0, \sigma^2 I)$. Defining $y = U^T x$, we have the model

$$y = \theta + w \quad , \quad w \sim \mathcal{N}(0, \sigma^2 I)$$

The likelihood of θ given y is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}}$$

The posterior distribution of θ is

$$\begin{aligned} p(\theta|y) \propto\ & p(y|\theta)p(\theta) \\ =\ & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}} \frac{\lambda}{2} e^{-\lambda|\theta_i|} \end{aligned}$$

Consider the MAP estimator

$$\begin{aligned} \widehat{\theta} &= \arg\max_{\theta} p(\theta|x) \\ &= \arg\max_{\theta} \log(p(\theta|x)) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \Big[-\frac{(y_i-\theta_i)^2}{2\sigma^2} - \lambda|\theta_i| \Big] + \text{constant} \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \Big[\frac{(y_i-\theta_i)^2}{2\sigma^2} + \lambda|\theta_i| \Big] \end{aligned}$$

If $\theta_i \neq 0$, then we can differentiate to obtain

$$\begin{aligned} & -\frac{(y_i-\theta_i)}{\sigma^2} + \lambda \text{sign}(\theta_i) = 0 \\ & \Rightarrow\ \theta_i = y_i - \lambda\sigma^2 \text{sign}(\theta_i) \end{aligned}$$

and clearly the minimizer must have the same sign as y_i, and so

$$\widehat{\theta}_i = y_i - \lambda\sigma^2 \text{sign}(y_i)$$

Plugging this into the argument of the minimization yields

$$\begin{aligned} \frac{(y_i-\widehat{\theta}_i)^2}{2\sigma^2} + \lambda|\widehat{\theta}_i| =\ & \frac{\lambda^2\sigma^4}{2\sigma^4} + \lambda|y_i - \lambda\sigma^2\text{sign}(y_i)| \\ =\ & \frac{\lambda^2\sigma^2}{2} + \lambda|y_i - \lambda\sigma^2\text{sign}(y_i)| \end{aligned} \tag{1}$$

On the other hand if $\widehat{\theta}_i = 0$, then the objective function's value is

$$\frac{(y_i-\widehat{\theta}_i)^2}{2\sigma^2} + \lambda|\widehat{\theta}_i| \ =\ \frac{y_i^2}{2\sigma^2} \tag{2}$$

Observe that

(1) < (2), when $|y_i| > \lambda\sigma^2$
(1) > (2), when $|y_i| \leq \lambda\sigma^2$

Therefore, the optimal solution is

$$\widehat{\theta}_i \ = \ \begin{cases} 0 & \text{if} \quad |y_i| \leq \lambda\sigma^2 \\ \\ y_i - \lambda\sigma^2\text{sign}(y_i) & \text{if} \quad |y_i| > \lambda\sigma^2 \end{cases}$$

This is called a "soft-threshold" function. It can be written compactly as

$$\widehat{\theta}_i = \text{sign}(y_i)\ \max(|y_i| - \lambda\sigma^2, 0)$$

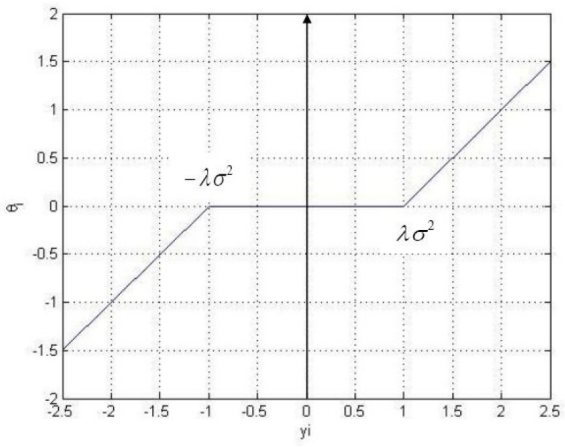

Figure 3: Plot of "soft-threshold."

It is depicted graphically in Figure 3. The "soft-threshold" estimator is:

$$\widehat{\theta} = \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_n \end{bmatrix} \qquad , \qquad \widehat{s} \;=\; U\widehat{\theta} \;=\; \sum_{i:|\widehat{\theta}_i|\neq 0} \widehat{\theta}_i u_i$$

where u_i is the ith column (basis vector) of U. Note that the soft-threshold estimator automatically selects a signal subspace based on the magnitude/energy of the observed data in each 1-dimensional subspace.

5 Summary

We studied the signal plus noise model

$$y \;=\; U^T x + w, \qquad w \sim \mathcal{N}(0, \sigma^2 I)$$

The MLE is

$$\begin{aligned} \widehat{\theta}_{\text{MLE}} &= \arg\min_{\theta} \frac{||x - u\theta||_2^2}{2\sigma^2} \\ &= \arg\min_{\theta} \frac{||x - u\theta||^2}{2\sigma^2} \\ &= U^T x \\ &= y \;. \end{aligned}$$

The Wiener filter (based on a Gaussian prior) is given by

$$\widehat{\theta}_{\text{Wiener}} = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma^2}\right) y \quad , \qquad \theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$$

$$\text{or} \qquad \widehat{\theta}_{\text{Wiener},i} = \left(\frac{\sigma_{\theta i}^2}{\sigma_{\theta i}^2 + \sigma^2}\right) y_i \quad , \qquad \theta_i \sim \mathcal{N}(0, \sigma_{\theta i}^2 I)$$

The soft-thresholding estimator based on the Laplacian prior $\theta_i \overset{\text{iid}}{\sim} \frac{\lambda}{2} e^{-\lambda|\theta_i|}$ has the form

$$\widehat{\theta} = \arg\min_{\theta} \frac{||y-\theta||_2^2}{2\sigma^2} + \lambda||\theta||_1$$
$$\widehat{\theta}_i = \text{sign}(y_i)\ \max(|y_i| - \lambda\sigma^2, 0)$$

$\longrightarrow$ Data-adaptive shrinkage to trade off bias and variance.

Example 5 *Consider the following observation*

$$\sigma^2 = 1, \qquad y = \begin{bmatrix} 10 \\ 1 \end{bmatrix} \swarrow \textit{ probably just noise}$$

MLE:

$$\widehat{\theta}_{MLE} = y = \begin{bmatrix} 10 \\ 1 \end{bmatrix} \qquad \textit{full dimension}$$

Wiener filter:

$$\widehat{\theta}_{Wiener} = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + 1}\right) \begin{bmatrix} 10 \\ 1 \end{bmatrix} \quad \propto \quad \begin{bmatrix} 10 \\ 1 \end{bmatrix} \qquad \textit{full dimension}$$

Soft-threshod:

$$\widehat{\theta}_{ST} = \begin{bmatrix} \max(10-\lambda, 0) \\ \max(1-\lambda, 0) \end{bmatrix} \overset{\lambda=1}{=} \begin{bmatrix} 9 \\ 0 \end{bmatrix} \qquad \textit{shrink to 1-dimension}$$

6 Inverse problems

Suppose we observe a distorted signal s in noise:

$$\begin{aligned} x &= As + w \\ &= AU\theta + w \quad , \quad w \sim \mathcal{N}(0, \sigma^2 I) \end{aligned} \tag{1}$$

A is a known matrix, suppose s is sparse in basis U, and write $s = U\theta$.

Wiener Filter (with Gaussian Prior): $\theta \sim \mathcal{N}(0, \lambda I)$

$$\widehat{\theta}_{\text{Wiener}} \ = \ \arg\min_{\theta} \left(\frac{||x - AU\theta||_2^2}{2\sigma^2} + \lambda||\theta||_2^2 \right)$$

$\Rightarrow$ linear, non-adaptive.

Sparse Solution (Laplacian Prior): "Lasso"

$$\widehat{\theta}_L \ = \ \arg\min_{\theta} \underbrace{\left(\frac{||x - AU\theta||_2^2}{2\sigma^2} + \lambda||\theta||_1 \right)}_{\text{"Lasso"}}$$

$\Rightarrow$ non-linear, adaptive.

Both are convex optimizations. The Wiener filter, which is linear, has a linear-algebraic solution. The Lasso (Least absolute shrinkage and selection operator) is nonlinear, and does not have a simple closed-form solution (except when $A = I$). The EM algorithm can be used to solve the Lasso optimization. Recall Example 6 from Lecture 16. That shows how to compute the function $Q(\theta, \theta^{(t)})$ for the observation model $x = H\theta + w$. Above we are simply using A instead of H. The EM algorithm can be used to maximize the log likelihood plus log prior simply by changing the M-step to

$$\theta^{(t+1)} \ = \ \arg\max_{\theta} Q(\theta, \theta^{(t)}) + \log p(\theta) \ .$$

In the Lasso problem, $\log p(\theta) = -\lambda||\theta||_1 = -\lambda \sum_i |\theta_i|$. In this case, the M-step involves a simple coordinate-wise soft-thresholding operation.

7 The Lasso

Consider the observation model

$$x \;=\; A\theta + w \;,\; w \sim \mathcal{N}(0, \sigma^2 I) \;,$$

and the following estimator of θ

$$\widehat{\theta} \;=\; \arg\min_{\theta} \left\{ ||x - A\theta||_2^2 + \lambda ||\theta||_1 \right\} \;.$$

This is called the *lasso* estimator, standing for 'least absolute shrinkage and selection operator,' originally proposed in

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).

As discussed above, the lasso can be derived as a MAP estimator for a Gaussian likelihood and Laplacian prior. Based on this observation, the lasso estimator can be computed using an EM algorithm. The EM algorithm is a 'bound optimization' algorithm; at each step we optimize a bound on the objective function that is easier to solve than the original problem. Here we derive a bound optimization algorithm for the lasso.

The bound optimization algorithm will produce a sequence of iterates $\theta^{(1)}$, $\theta^{(2)}$, ..., and the initial iterate $\theta^{(1)}$ can be arbitrary (e.g., random, equal to zero, equal to x). To derive the algorithm, suppose we have a current iterate $\theta^{(t)}$. We can write the objective function as follows.

$$\begin{aligned} L(\theta) &= \|x - A\theta\|_2^2 + \lambda\|\theta\|_1 \\ &= \|x - A\theta^{(t)} + A\theta^{(t)} - A\theta\|_2^2 + \lambda\|\theta\|_1 \\ &= \|x - A\theta^{(t)}\|_2^2 + 2[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + \|A\theta^{(t)} - A\theta\|_2^2 + \lambda\|\theta\|_1 \\ &= C + 2[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + \|A\theta^{(t)} - A\theta\|_2^2 + \lambda\|\theta\|_1 \end{aligned}$$

where $C = \|x - A\theta^{(t)}\|_2^2$ is a constant independent of the variable θ. Continuing, we have

$$\begin{aligned} L(\theta) &= C + 2[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + (\theta^{(t)} - \theta)^T A^T A(\theta^{(t)} - \theta) + \lambda\|\theta\|_1 \\ &\leq C + 2[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + \alpha^{-1}\|\theta^{(t)} - \theta\|_2^2 + \lambda\|\theta\|_1 \end{aligned}$$

where α^{-1} any value greater than or equal to the largest eigenvalue of $A^T A$. Our next iterate will be the value of θ that minimizes this upper bound. The minimizer is unaffected by multiplying or adding constants, so we have

$$\begin{aligned} \theta^{(t+1)} &= \arg\min_{\theta} \left\{ C + 2[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + \alpha^{-1}\|\theta^{(t)} - \theta\|_2^2 + \lambda\|\theta\|_1 \right\} \\ &= \arg\min_{\theta} \left\{ \alpha C + 2\alpha[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + \|\theta^{(t)} - \theta\|_2^2 + \alpha\lambda\|\theta\|_1 \right\} \\ &= \arg\min_{\theta} \left\{ C' + 2\alpha[A^T(x - A\theta^{(t)})]^T(\theta^{(t)} - \theta) + \|\theta^{(t)} - \theta\|_2^2 + \lambda'\|\theta\|_1 \right\} \end{aligned}$$

where $C' = \alpha C$ and $\lambda' = \alpha\lambda$. Let $z := \alpha[A^T(x - A\theta^{(t)})]$ a constant vector independent of θ. Now write

$$\begin{aligned} \theta^{(t+1)} &= \arg\min_{\theta} \left\{ C' + 2z^T(\theta^{(t)} - \theta) + \|\theta^{(t)} - \theta\|_2^2 + \lambda'\|\theta\|_1 \right\} \\ &= \arg\min_{\theta} \left\{ C'' + \|z + \theta^{(t)} - \theta\|_2^2 + \lambda'\|\theta\|_1 \right\} \\ &= \arg\min_{\theta} \left\{ \|z + \theta^{(t)} - \theta\|_2^2 + \lambda'\|\theta\|_1 \right\} \end{aligned}$$

where $C'' = C' - \|z\|_2^2$, another constant independent of θ. Notice that the final optimization is equivelent to

$$\theta^{(t+1)} = \arg\min_{\theta} \left\{ \|x^{(t)} - \theta\|_2^2 + \lambda' \|\theta\|_1 \right\}$$

where $x^{(t)} := \theta^{(t)} + \alpha[A^T(x - A\theta^{(t)})]$. The solution to this optimization is the soft-thresholding estimator, easily computed as follows. $\theta_i^{(t+1)} = \text{sign}(x_i^{(t)}) \max(|x_i^{(t)} - \lambda', 0)$.

To sum up, the iterative algorithm starts with an initial guess $\theta^{(1)}$ and computes for $t = 1, 2, \ldots$

$$\begin{aligned} x^{(t)} &= \theta^{(t)} + \alpha[A^T(x - A\theta^{(t)})] \\ \theta_i^{(t+1)} &= \text{sign}(x_i^{(t)}) \max(|x_i^{(t)} - \alpha\lambda, 0) \ , \ i = 1, \ldots, n \end{aligned}$$

where $0 < \alpha \leq \frac{1}{\lambda}$, λ the largest eigenvalue of $A^T A$.

8 Statistical Analysis of Soft-Thresholding

Consider the "direct" observation model where $y \in R^n$ is given by

$$y = \theta + w \ , \ w \sim \mathcal{N}(0, \sigma^2 I) \ .$$

Suppose that many of the coefficients in θ are equal to zero. The MLE of θ is simply x, and its MSE is $n\sigma^2$. The soft-thresholding estimator

$$\widehat{\theta}_i = \text{sign}(y_i) \ \max(|y_i| - t, 0) \ , \ t > 0$$

can perform much better, especially if θ is sparse.

Before we analyze the soft-thresholding estimator, let us consider an ideal thresholding estimator. Suppose that an oracale tells us the magnitude of each θ_i. The *oracle* estimator is

$$\widehat{\theta}_i^O = \begin{cases} y_i & \text{if } |\theta_i|^2 \geq \sigma^2 \\ 0 & \text{if } |\theta_i|^2 < \sigma^2 \end{cases}$$

In other words, we estimate a coefficient if and only if the signal power is at least as large as the noise power. The MSE of this estimator is

$$\mathbb{E} \sum_{i=1}^n (\widehat{\theta}_i^O - \theta_i)^2 = \sum_{i=1}^n \min(|\theta_i|^2, \sigma^2)$$

Notice that the MSE of the oracle estimator is always less than or equal to the MSE of the MLE. If θ is sparse, then the MSE of the oracle estimator can be much smaller. If all but $k < n$ coefficients are zero, then the MSE of the oracle estimator is at most $k\sigma^2$. Remarkably, the soft-thresholding estimator comes very close to achieving the performance of the oracle, and shown by the following theorem (Theorem 1 in "Ideal Spatial Adaptation by Wavelet Thresholding," by Donoho and Johnstone).

Theorem 1 *Assume the direct observation model above and let*

$$\widehat{\theta}_i = \text{sign}(y_i) \ \max(|y_i| - t, 0)$$

with $t = \sqrt{2\sigma^2 \log n}$. *Then*

$$\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \ \leq \ (2 \log n + 1) \left\{ \sigma^2 + \sum_{i=1}^n \min(|\theta_i|^2, \sigma^2) \right\}$$

The theorem shows that the soft-thresholding estimator mimics the MSE performance of the oracle estimator to within a factor of roughly $2\log n$. For example, if θ is k-sparse (with non-zero coefficients larger than σ in magnitude), then the MSE of the oracle is $k\sigma^2$ and the MSE of the soft-thresholding estimator is at most $(2\log n + 1)(k+1)\sigma^2 \approx 2k\log n\,\sigma^2$ when n is large. This also corresponds to a huge improvement over the MLE if $2k\log n \ll n$.

Proof: To simplify the analysis, assume that $\sigma^2 = 1$. The general result follows directly. It suffice to show that

$$\mathbb{E}[(\widehat{\theta}_i - \theta_i)^2] \;\le\; (2\log n + 1)\left\{\frac{1}{n} + \min(\theta_i^2, 1)\right\}$$

for each i. So let $x \sim \mathcal{N}(\mu, 1)$ and let $\eta_t(x) = \text{sign}(x)\max(|x| - t, 0)$. We will show that with $t = \sqrt{2\log n}$

$$\mathbb{E}[(\eta_t(x) - \mu)^2] \;\le\; (2\log n + 1)\left\{\frac{1}{n} + \min(\mu^2, 1)\right\}\;.$$

First note that $\eta_t(x) = x - \text{sign(x)}(|x| \wedge t)$, where $a \wedge b$ is shorthand notation for $\min(a, b)$. It follows that

$$\begin{aligned}\mathbb{E}[(\eta_t(x) - \mu)^2] &= \mathbb{E}[(x-\mu)^2] - 2\mathbb{E}[\text{sign(x)}(|x| \wedge t)(x-\mu)] + \mathbb{E}[x^2 \wedge t^2] \\ &= 1 - 2\mathbb{E}[\text{sign}(x)(|x| \wedge t)(x-\mu)] + \mathbb{E}[x^2 \wedge t^2]\end{aligned}$$

The expected value in the second term is equal to $\mathbb{P}(|x| < t)$, which is verified as follows.

The expectation can be split into integrals over four intervals, $(\infty, -t]$, $(-t, 0]$, $(0, t]$, and (t, ∞). Each integrand is a linear or quadratic function of x times the Gaussian density function. Let $\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi(x) := \int_{-\infty}^{x} \phi(y)dy$, the cumulative distribution function of $\phi(x)$, and consider the following indefinite Gaussian integral forms:

$$\begin{aligned}\int \phi(x)\,dx &= \Phi(x)\;,\text{ by definition of }\Phi, \\ \int x\phi(x)\,dx &= \frac{1}{\sqrt{2\pi}}\int x e^{-x^2/2}\,dx \;=\; \underbrace{-\frac{1}{\sqrt{2\pi}}\int e^u\,du}_{u=-x^2/2} \;=\; -\frac{1}{\sqrt{2\pi}}e^u \;=\; -\phi(x)\;, \\ \int x^2\phi(x)\,dx &= \Phi(x) - x\phi(x)\;.\end{aligned}$$

The last form is verified as follows. Let $u = x$ and $dv = x\phi(x)dx$. Then integration by parts $\int udv = uv - \int vdu$ and $\int x\phi(x)dx = -\phi(x)$ show that

$$\int x^2\phi(x)\,dx = x\int x\phi(x)dx - \int\int x\phi(x)dx = -x\phi(x) + \int \phi(x) = \Phi(x) - x\phi(x)\;.$$

The Gaussian distribution we are considering has mean μ so the shifted integral forms below, which follow immediately from the derviations above by variable substitution, will be used in our analysis:

$$\begin{aligned}(i) \quad \textstyle\int \phi(x-\mu)dx &= \Phi(x-\mu) \\ (ii) \quad \textstyle\int x\phi(x-\mu)dx &= \mu\Phi(x-\mu) - \phi(x-\mu) \\ (iii) \quad \textstyle\int x^2\phi(x-\mu)dx &= (1+\mu^2)\Phi(x-\mu) - (x+\mu)\phi(x-\mu)\end{aligned}$$

Using these forms we compute

$$\begin{aligned}
\mathbb{E}[\text{sign(x)}(|x| \wedge t)(x-\mu)] &= \int_{-\infty}^{\infty} \text{sign(x)}(|x| \wedge t)(x-\mu)\phi(x-\mu)\,dx \\
&= \underbrace{\int_{-\infty}^{-t} -t(x-\mu)\phi(x-\mu)\,dx}_{t\phi(-t-\mu)} \underbrace{-\int_{-t}^{0} x(x-\mu)\phi(x-\mu)dx}_{\Phi(-\mu)-\Phi(-t-\mu)-t\phi(-t-\mu)} \\
&\quad + \underbrace{\int_{0}^{t} x(x-\mu)\phi(x-\mu)dx}_{\Phi(t-\mu)-\Phi(-\mu)-t\phi(t-\mu)} + \underbrace{\int_{t}^{\infty} t(x-\mu)\phi(x-\mu)dx}_{t\phi(t-\mu)} \\
&= \Phi(t-\mu) - \Phi(-t-\mu) = \mathbb{P}(|x| < t)
\end{aligned}$$

So we have shown that

$$\mathbb{E}[(\eta_t(x)-\mu)^2] = 1 - 2\mathbb{P}(|x| < t) + \mathbb{E}[x^2 \wedge t^2]$$

Note first that since $x^2 \wedge t^2 \leq t^2$ we have

$$\mathbb{E}[(\eta_t(x)-\mu)^2] \leq 1 + t^2 = 1 + 2\log n < (2\log n + 1)(1/n + 1) \ .$$

On the other hand, since $x^2 \wedge t^2 \leq x^2$ we also have

$$\mathbb{E}[(\eta_t(x)-\mu)^2] \ \leq \ 1 - 2\mathbb{P}(|x| < t) + \mu^2 + 1 \ = \ 2(1 - \mathbb{P}(|x| < t)) + \mu^2 \ = \ 2\mathbb{P}(|x| \geq t) + \mu^2 \ .$$

The proof will be finished if we show that

$$2\mathbb{P}(|x| \geq t) \ \leq \ (2\log n + 1)/n + (2\log n)\mu^2 \ .$$

Define $g(\mu) := 2\mathbb{P}(|x| \geq t)$ and note that g is symmetric about 0. Using a Taylor's series with remainder we have

$$g(\mu) \leq g(0) + \frac{1}{2}\sup|g''|\mu^2 \ ,$$

where g'' is the second derivative of g. Note that $g(\mu) = 2\left[1 - \mathbb{P}(z \leq t - \mu) - \mathbb{P}(z \leq -t - \mu)\right]$, where $z \sim \mathcal{N}(0,1)$. Using the Gaussian tail bound $\mathbb{P}(z > t) \leq \frac{1}{2}e^{-t^2/2}$ and plugging in $t = \sqrt{2\log n}$ we obtain $g(0) \leq 2/n$. Note that $g'(\mu) = 2[\phi(t-\mu)+\phi(-t-\mu)]$ and $g'(0) = 0$. The integral (ii) above shows that the derivative of $\phi(t-\mu)$ with respect to μ is equal to $(t-\mu)\phi(t-\mu)$. So we have $g''(\mu) = 2[(t-\mu)\phi(t-\mu)+(-t-\mu)\phi(-t-\mu)]$. It is easy to check that $|g''(\mu)| < 1$ so it follows that $\sup_\mu g''(\mu) \leq 4\log n$ for all $n \geq 2$.

8.1 The Lasso

Consider the observation model

$$x \ = \ A\theta + w \ , \ w \sim \mathcal{N}(0, \sigma^2 I) \ ,$$

and the following estimator of θ

$$\widehat{\theta} \ = \ \arg\min_\theta \left\{ \frac{1}{2}||x - A\theta||_2^2 + \lambda||\theta||_1 \right\} \ .$$

This is called the *lasso* estimator, standing for 'least absolute shrinkage and selection operator,' originally proposed in

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol.

58, No. 1, pages 267-288).

The lasso has received an enormous amount of attention in recent years. If $A = I$, then this reduces to the soft-thresholding estimator with threshold λ. More generally if A 'acts' like the identity operator for all sparse vectors, then statistical error bounds similar to those for soft-thresholding can be obtained. For example, if x is k-sparse and A satisfies the so-called 'Restricted Isometry Property'

$$(1-\delta)\|u-v\|_2^2 \;\leq\; \|A(u-v)\|_2^2 \;\leq (1+\delta)\|u-v\|_2^2$$

for all k sparse vectors $u, v \in \mathbb{R}^n$ for a $\delta > 0$, then the MSE of the lasso estimator is nearly equal to that of an oracle estimator that knows the locations of the non-zero components of x. A comprehensive overview of lasso error bounds is given in this paper

van de Geer, S. and Buhlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics 3, 1360-1392.

9 Wavelet Denoising

Wavelets are locally supported, oscillating functions that integrate to zero. The *Daubechies wavelets* are special functions that satisfy the so-called "vanishing moment" conditions. If $\psi(t)$ is the Daubechies wavelet with k vanishing moments, then

$$\int t^\ell \psi(t)\, dt \;=\; 0\;,\; \ell = 0, 1, \ldots, k-1\;.$$

The discrete-time analogs ψ defined on $t \in \{0, 1, \ldots, n-1\}$ satisfy

$$\sum_{t=0}^{n-1} t^\ell \psi(t) \;=\; 0\;,\; \ell = 0, 1, \ldots, k-1\;.$$

The vanishing moments property implies that if we integrate such a wavelet against a polynomial of degree k or less, then the integral (i.e., inner product between the wavelet and the polynomial) is zero.

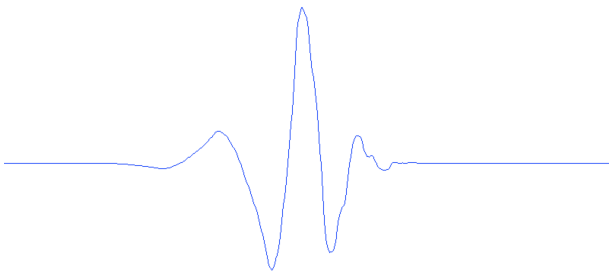

Figure 4: Daubechies wavelet with 4 vanishing moments.

Wavelets also can be used to generate an orthonormal basis for L_2 functions on the unit interval $[0, 1)$ or vectors in $\mathbb{R}^n$ (based respectively on the continuous or discrete wavelets, above). An orthonomal basis is generated by scaling and shifting the argument of the basic wavelet function. This produces compressed and dilated versions of the wavelet at different locations in the interval.

In the case of $\mathbb{R}^n$, the Daubechies wavelet basis consists of wavelet functions at scales $1, 1/2, 1/4, \ldots, 1/n$ plus one constant or low frequency vector to complete the basis. The scale indicates the approximate fraction

of the set $\{0, 1, \ldots, n-1\}$ that each wavelet function is supported on. At scale 2^{-j} each wavelet is supported on $O(n2^{-j})$ points and there are 2^j wavelet functions spaces uniformly over the set $\{0, 1, \ldots, n-1\}$, for $j = 0, 1, \ldots, \log_2 n/2$.

Figure 5: Daubechies wavelet basis functions with 2 vanishing moments.

Denote the wavelet functions by $\{\phi_i\}_{i=1}^n$. There are n functions and this set forms an orthonormal basis for $\mathbb{R}^n$. That is, any vector $x \in \mathbb{R}^n$ can be represented exactly as

$$x \;=\; \sum_i (\phi_i' x)\, \phi_i \;.$$

Define the *wavelet coefficients* of x to be the inner products $\theta_i = \phi_i' x$, $i = 1, \ldots, n$ and let W denote the matrix with columns equal to the wavelet basis functions. Then the vector of wavelet coefficients $\theta = W'x$ and $x = W\theta$. The matrix W' is called the *wavelet transform* and W is the *inverse wavelet transform*.

Let us record a few key facts about wavelet transforms. Because of the local support of the wavelet functions, the value of $x(t)$ for any point $t \in \{0, 1, \ldots, n-1\}$ is a linear combination of about $\log_2(n)$ wavelet functions (roughly one at each scale). In other words, all but about $\log_2(n)$ wavelet functions are equal to zero at any given point t. Also note that if the values of x are determined by a polynomial, then most the *wavelet coefficients* (the inner products between x and the wavelet basis functions) will be zero by the vanishing moments property. Putting these two facts together we have the following result. If the values of $x \in \mathbb{R}^n$ are piecewise polynomial with m pieces, then all but about $m \log n$ wavelet coefficients will be zero (i.e., the coefficients will be sparse).

Now suppose that x is a vector of samples of a Hölder α-smooth function on $[0, 1]$. Let $\theta \in \mathbb{R}^n$ denote the wavelet transform of this vector based on a wavelet with at least $k = \lceil \alpha \rceil - 1$ vanishing moments. Since the samples are approximately polynomial in any small subinterval, we conclude less than $Cm \log n$ coefficients will be large in magnitude, where $C > 0$ is a constant. The rest of the coefficients will be less than $Cm^{-\alpha}$ in magnitude.

Suppose that we make noisy measurements of x. Each sample is corrupted with independent standard Gaussian noise. Since the wavelet transform is orthonormal, the transform of the noisy samples is given by

$$y_i \;=\; \theta_i \;+\; w_i \;,\; i = 1, \ldots, n \;,$$

where $w_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. We can denoise these coefficients according to Theorem 1 above to obtain an estimate $\widehat{\theta}$ satisfying

$$\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \;\leq\; (2 \log n + 1) \left\{ 1 + \sum_{i=1}^n \min(|\theta_i|^2, 1) \right\}$$

Since few coefficients are large and most are very small we have that the average MSE is bounded according to

$$\frac{1}{n}\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \;\leq\; C \log n \left(\frac{m}{n}\log n + m^{-2\alpha}\right) \;.$$

If we take $m = \left(\frac{n}{\log n}\right)^{1/(2\alpha+1)}$, then

$$\frac{1}{n}\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \;\leq\; C \log n \left(\frac{\log n}{n}\right)^{2\alpha/(2\alpha+1)} \;.$$

This is within a poly-logarithmic factor of the MSE based on piecewise polynomial fitting.

The wavelet denoising approach also has significant advantages. First of all, notice that wavelet denoising is automatically adaptive to the smoothness of the underlying function. Using a wavelet basis with k vanishing moments, it achieves the optimal rate of convergence (up to a polylog factor) for any Hölder α class with $\lceil \alpha \rceil \leq k+1$. Wavelet denoising requires no prior knowledge of α, unlike classical methods that need to know α in order to choose the number of polynomial pieces to fit to the data. Secondly, wavelet denoising can even handle piecewise smooth functions with discontinuities between pieces. Suppose that the underlying function is only piecewise Hölder smooth (e.g., it may have a few points of discontinuity). Then the wavelet denoising rate stays the same as above, but the piecewise polynomial rate degrades to $n^{-1/2}$.

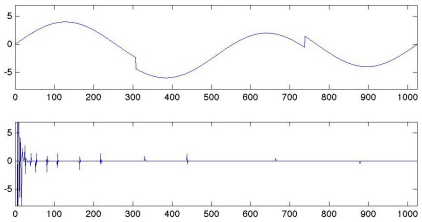

Figure 6: Piecewise smooth function (top). Coefficients of Daubechies wavelet transform with 4 vanishing moments (bottom). Notice that most of the coefficients are zero.

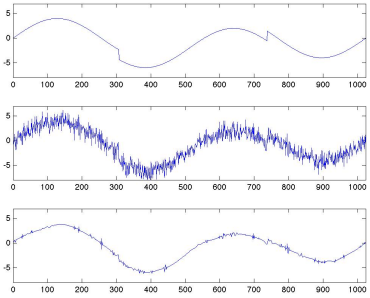

Figure 7: Denoising in action. Original function (top). Noisy version (middle). Denoised estimate obtained by thresholding noisy wavelet coefficients (bottom).