Analysis of Soft-Thresholding

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

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

Suppose that many of the weights/coefficients in $\boldsymbol{w}$ are equal to zero. The MLE of $\boldsymbol{w}$ is simply $\boldsymbol{y}$, and its MSE is $n\sigma^2$. The soft-thresholding estimator

$$\widehat{w}_i = \text{sign}(y_i) \; \max(|y_i| - \lambda, 0) \; , \; \lambda > 0$$

can perform much better, especially if $\boldsymbol{w}$ 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 w_i. The *oracle* estimator is

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

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

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

Notice that the MSE of the oracle estimator is always less than or equal to the MSE of the MLE. If w 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).

The theorem uses the threshold $\lambda = \sqrt{2\sigma^2 \log n}$. This choice of threshold is motivated by the following observation. Assume, for the moment, that we observe no signal at all, just noise (i.e., $w_i = 0$ for $i = 1, \dots, n$). In this case, we should set the threshold so that it is larger than the magnitude of any of the y_i (so they are all set to zero). If we take $\lambda = \sqrt{2\sigma^2 \log \frac{n}{\delta}}$, then using the Gaussian tail bound and the union bound we have $\mathbb{P}(\bigcup_{i=1}^n \{|y_i| \geq \lambda\}) \leq \delta$.

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

$$\widehat{w}_i = \text{sign}(y_i) \; \max(|y_i| - \lambda, 0)$$

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

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

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 $\boldsymbol{w}$ 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 $2k\log n \ll n$.

Intution: Consider the case with $\sigma^2 = 1$ (the general case follows by simple rescaling). First recall that if $y \sim \mathcal{N}(0,1)$, then $\mathbb{P}(|y| \geq \lambda) \leq e^{-\lambda^2/2}$. This inequality is easily derived as follows. Since $\mathbb{P}(y \geq \lambda) = \mathbb{P}(y \leq -\lambda)$, we only need to show that $\mathbb{P}(y \geq \lambda) = \frac{1}{2\pi}\int_\lambda^\infty e^{-x^2/2}dx \leq \frac{1}{2}e^{-\lambda^2/2}$. Note that

$$\frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-x^2/2}dx}{\frac{1}{2}e^{-\lambda^2/2}} \;=\; \frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-(x^2-\lambda^2)/2}dx}{\frac{1}{2}} \;=\; \frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-(x-\lambda)(x+\lambda)/2}dx}{\frac{1}{2}}\;.$$

The desired inequality results by making change of variable $t = y + \lambda$ to yield

$$\frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-x^2/2}dx}{\frac{1}{2}e^{-\lambda^2/2}} \;=\; \frac{\frac{1}{2\pi}\int_0^\infty e^{-t(t+2\lambda)/2}dt}{\frac{1}{2}} \;\leq\; \frac{\frac{1}{2\pi}\int_0^\infty e^{-t^2/2}dt}{\frac{1}{2}} \;=\; 1\;.$$

Now observe that if $\lambda = \sqrt{2\sigma^2\log n}$, then $\mathbb{P}(|y_i| \geq \lambda | w_i = 0) \leq e^{-\log n} = \frac{1}{n}$. Using this we have

$$\mathbb{E}\left[\sum_{i:w_i=0} \mathbb{1}\{\widehat{w}_i \neq 0\}\right] \;=\; \sum_{i:w_i=0}\frac{1}{n} \;\leq\; 1\;.$$

In other words, using this threshold we expect that at most one of the $w_i = 0$ will not be estimated as $\widehat{w}_i = 0$. Next consider cases when $w_i \neq 0$. Let's suppose that $|w_i| \gg \lambda$, so that $\widehat{w}_i = y_i - \lambda\text{sign}(y_i)$. In this case,

$$(w_i - \widehat{w}_i)^2 \;=\; (-\epsilon_i + \lambda\text{sign}(y_i))^2 \;\leq\; \epsilon_i^2 + 2|\epsilon_i|\lambda + \lambda^2\;.$$

Taking the expecation of this upper bound yields

$$\mathbb{E}[(w_i - \widehat{w}_i)^2] \;\leq\; 1 + 2\lambda + \lambda^2 \;\leq\; 3\lambda^2 + 1\;,\; \text{assuming } \lambda > 1\;.$$

Thus, if $\boldsymbol{w}$ has only k nonzero weights, then this intution suggests that

$$\sum_{i=1}^n \mathbb{E}[(w_i - \widehat{w}_i)^2] \;=\; O(k\log n)\;.$$

This is formalized in the following proof of Theorem 1.

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

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

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

$\sqrt{2\log n}$

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

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

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

The expected value in the second term is equal to $\mathbb{P}(|x| < \lambda)$, 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:

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

Using these forms we compute

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

So we have shown that

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

Note first that since $x^2 \wedge \lambda^2 \le \lambda^2$ we have

$$\mathbb{E}[(f_\lambda(x)-\mu)^2] \le 1 + \lambda^2 = 1 + 2\log n < (2\log n + 1)(1/n + 1)\;.$$

On the other hand, since $x^2 \wedge \lambda^2 \le x^2$ we also have

$$\mathbb{E}[(f_\lambda(x)-\mu)^2] \;\le\; 1 - 2\mathbb{P}(|x|<\lambda) + \mu^2 + 1 \;=\; 2(1-\mathbb{P}(|x|<\lambda)) + \mu^2 \;=\; 2\mathbb{P}(|x|\ge\lambda) + \mu^2\;.$$

The proof will be finished if we show that

$$2\mathbb{P}(|x|\ge\lambda) \;\le\; (2\log n + 1)/n + (2\log n)\mu^2\;.$$

Define $g(\mu) := 2\mathbb{P}(|x|\ge\lambda)$ and note that g is symmetric about 0. Using a Taylor's series with remainder we have

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

where g'' is the second derivative of g. Note that $g(\mu) = 2\,[1 - \mathbb{P}(z \le \lambda-\mu) + \mathbb{P}(z \le -\lambda-\mu)]$, where $z \sim \mathcal{N}(0,1)$. Using the Gaussian tail bound $\mathbb{P}(z > \lambda) \le \frac{1}{2}e^{-\lambda^2/2}$ and plugging in $\lambda = \sqrt{2\log n}$ we obtain $g(0) \le 2/n$. Note that $g'(\mu) = 2[\phi(\lambda-\mu) - \phi(-\lambda-\mu)]$ and $g'(0) = 0$. The integral (ii) above shows that the derivative of $\phi(\lambda-\mu)$ with respect to μ is equal to $(\lambda-\mu)\phi(\lambda-\mu)$. So we have $g''(\mu) = 2[(\lambda-\mu)\phi(\lambda-\mu) + (-\lambda-\mu)\phi(-\lambda-\mu)]$. It is easy to check that $|g''(\mu)| < 1$ so it follows that $\sup_\mu g''(\mu) \le 4\log n$ for all $n \ge 2$.