Stat 300A Theory of Statistics

Week 6 problem session: information-theoretic lower bounds

Stephen Bates *November 2, 2018*

1 Outline

1.1 Big picture

Recall that we are studying information-theoretic lower bounds. For statistical decision problems, we wish to give a lower bound on the (minimax) risk, and then find a procedure that satisfies the minimax risk. **If we have a lower bound and a procedure that achieves the lower bound, then we understand the problem.** In general, it is difficult to find procedures that are exactly minimax optimal, so we settle for problems that are approximately minimax optimal.

Recall that for any prior Q the minimax risk is larger than the Bayes risk:

$$R_{\mathrm{M}} \geq R_{\mathrm{B}}(Q).$$

Le Cam's method and Fano's method are techniques for lower bounding the Bayes risk for certain choices of the prior Q.

1.2 Today

Today we will look at the following

- An estimator that approximately achieves the minimax lower bound in sparse regression.
- A tail bound for the standard Gaussian
- An application of Stein's unbiased risk estimator (SURE)

2 Sparse regression example

2.1 Minimax lower bound

Recall from section 4.3 of the lecture notes that we derived a following lower bound for the minimax risk of the sparse regression problem.

Theorem 1 (Minimax risk for sparse regression)**.** *Fix a design matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, we let $\mathsf{P}_{\boldsymbol{\theta}} = \mathsf{N}(\boldsymbol{A}\boldsymbol{\theta}, \sigma^2 \mathrm{I})$, where $\boldsymbol{\theta} \in \Theta_0(k)$ belongs to the set of k-sparse vectors (vectors with at most k non-zero entries). If $A_{av}^2 \equiv \sum_{i\leq n, j\leq d} A_{ij}^2/(nd)$, then we have*

$$R_{\mathrm{M}}(\Theta_0(k)) \geq C_0 \frac{k\,\sigma^2}{n\,A_{av}^2} \log\left(d/k\right). \tag{1}$$

2.2 Achieving the lower bound with soft-thresholding

In this section, we will show that we can achieve the minimax lower bound in a simple case [1]. Let $\boldsymbol{A} = I_d \in \mathbb{R}^{d\times d}$, and fix $\sigma^2 = 1$. We will consider the family of soft-thresholding estimators:

$$S_\lambda(x) = \begin{cases} x-\lambda & \text{when } x > \lambda \\ 0 & \text{when } |x| \le \lambda \\ x+\lambda & \text{when } x < -\lambda \end{cases}$$

$$\hat{\theta}_\lambda(x) = \{S_\lambda(x_i)\}_{i=1,\ldots,d} \qquad \text{(i.e. soft-threshold each coordinate)}$$

Notice that this is a nonlinear estimator. It shrinks values toward zero and sets values smaller that λ equal to zero.

We will now compute the risk of this estimator under squared error loss. Since $\hat{\theta}(x)_i$ depends only on x_i. it suffices to consider the case $d = 1$.

$$R(\hat{\theta}_\lambda(\cdot), \theta) = \int (\hat{\theta}_\lambda(\theta + z) - \theta)^2 \phi(z) dz$$

Notice that the first term of the integrand is:

$$(\hat{\theta}_\lambda(\theta + z) - \theta)^2 = \begin{cases} (z+\lambda)^2 & z+\theta < -\lambda \\ \theta^2 & |z+\theta| < \lambda \\ (z-\lambda)^2 & z+\theta > \lambda. \end{cases}$$

Differentiating with respected to θ under the integral sign gives:

$$0 \le \frac{\partial R(\hat{\theta}_\lambda(\cdot), \theta)}{\partial \theta} = 2\theta P(|\theta + Z| \le \lambda) \le 2\theta.$$

We conclude that the the risk is symmetric and increasing for $\theta > 0$.

Next, we apply SURE. Let $g_\lambda(x) = \hat{\theta}_\lambda(x) - x$. Then $\frac{\partial g_\lambda}{\partial x} = -I_{|x|\le\lambda}$. Plugging this into SURE:

$$\mathbb{E}_\theta(\hat{\theta}_\lambda(x) - \theta)^2 = \mathbb{E}_\theta[1 - 2\frac{\partial g_\lambda}{\partial x} + g_\lambda(x)^2] = 1 - 2P_\theta(|x| \le \lambda) + g_\lambda(x)^2) \to 1 + \lambda^2 \text{ as } \theta \to \infty$$

This implies that

$$R(\hat{\theta}_\lambda(\cdot), \theta) \le R(\hat{\theta}_\lambda(\cdot), 0) + \min(\theta^2, 1 + \lambda^2).$$

We now consider the risk at $\theta = 0$

$$R(\hat{\theta}_\lambda(\cdot), 0) = \int (\hat{\theta}_\lambda(z))^2 \phi(z) dz = 2\int_\lambda^\infty (z-\lambda)^2 \phi(z) dz = 2(1 - \Phi(\lambda)) - 2\lambda\phi(\lambda).$$

Proposition 2.1 (Mill's tail bound)**.** *Let $Z \sim \mathsf{N}(0,1)$. Let ϕ be the standard normal density For $t > 0$*

$$(\frac{1}{t} - \frac{1}{t^3})\phi(t) \le P(Z \ge t) \le \frac{1}{t}\phi(t).$$

Proof. The proof is deferred to a future section. ☐

Using this proposition, we have that

$$R(\hat{\theta}_\lambda(\cdot), 0) \le 2\frac{1}{\lambda}\phi(\lambda).$$

[1] See section 2.7 of Iain Johnstone's *Gaussian Estimation*

Choose $\lambda_0 = \sqrt{2\log(d)}$. Then

$$R(\hat{\theta}_{\lambda_0}(\cdot), 0) \leq 2\frac{1}{\lambda_0}\phi(\lambda_0) \leq \frac{1}{d}.$$

$$R(\hat{\theta}_{\lambda_0}(\cdot), \theta) \leq 1/d + (2\log(d) + 1)\min(\theta^2, 1)$$

Now consider the problem with general d. Using the above calculation, we have

$$R(\hat{\theta}_{\lambda_0}(\cdot), \theta) \leq 1 + (2\log(d) + 1)\sum_{i=1}^{n}\min(\theta_i^2, q) \leq 1 + (2\log(d) + 1)k \approx 2k\log(d)$$

We conclude that the soft thresholding estimator is approximately minimax for this problem:

$$\limsup_{d\to\infty}\frac{R(\hat{\theta}_{\lambda_0}(\cdot), \theta)}{R_M(\Theta_k)} = C < \infty.$$

3 Mill's tail bound for a Gaussian

Proof. 2.1 We will only show the upper bound.

$$\begin{aligned}
P(Z \geq t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}dy \\
&= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-(y+t)^2/2}dy \\
&= \frac{e^{-t^2/2}}{\sqrt{2\pi}}\int_0^\infty e^{-y^2/2-yt}dy \\
&\leq \frac{e^{-t^2/2}}{\sqrt{2\pi}}\int_0^\infty e^{-yt}dy \\
&= \frac{e^{-t^2/2}}{t\sqrt{2\pi}}
\end{aligned}$$

□