36-705: Intermediate Statistics **Fall 2019**

Lecture 30: November 13

Lecturer: Siva Balakrishnan

Our goal for today and the next two lectures will be to discuss some basic results in *high-dimensional statistics* and in *non-parametric statistics* but we will need some background before we can get there.

30.1 The Gaussian Sequence Model

As a warm-up on sparse estimation, let us consider the Gaussian sequence model. Suppose that we observed $\{y_1, \ldots, y_d\}$ where:

$$y_i = \theta_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2/n)$. To understand why we divided the variance by n in the model, you should observe that this corresponds to taking n i.i.d. observations and averaging them. To think about this as a high-dimensional problem, we just assume that $d \to \infty$ as $n \to \infty$, i.e. d is not assumed to be a constant so the number of parameters we want to estimate grows with n.

We have already derived the minimax estimator (and its ℓ_2 risk) for this problem in our lecture on minimax estimation. The minimax estimator is:

$$\widehat{\theta} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix},$$

and its ℓ_2 risk is:

$$R(\widehat{\theta}, \theta) = \mathbb{E}\left[\sum_{i=1}^{d} \epsilon_i^2\right] = \frac{\sigma^2 d}{n},$$

so if $d \gg n$ then we cannot consistently estimate θ. This is a form of the curse of dimensionality but it is much milder than in non-parametric problems. You can see the rate is the "usual parametric rate".

Random Comment: There is a sense in which the Gaussian sequence model is an extremely rich model, even though it seems somewhat trivial on the surface. In particular, due to something known as Le Cam's equivalence, one can "reduce" many parametric and non-parametric problems (including things like density estimation and non-parametric regression) to sequence model problems, with constraints on the vector θ. Many things we understand about rates of convergence are seen most clearly in this model.

In order to recover θ in the high-dimensional setting (when $d \gg n$) we need some sort of structural assumptions on θ. A natural assumption from a practical standpoint is that of sparsity, i.e. we assume that the true underlying θ has many entries which are 0 or nearly zero.

What are natural estimators in this case? A couple of popular ones are based on thresholding:

1. **Hard Thresholding:** Here we use the estimator:
$$\widehat{\theta}_i = y_i \mathbb{I}(|y_i| \geq t), \quad \forall\, i \in \{1, \ldots, d\},$$
where $t > 0$ is some threshold that we need to select.

2. **Soft Thresholding:** One that is closer in spirit to the LASSO (its regression counterpart) is based on soft thresholding, i.e.
$$\widehat{\theta}_i = \mathrm{sign}(y_i) \max\{|y_i| - t, 0\}, \quad \forall\, i \in \{1, \ldots, d\},$$
where $t > 0$ is some threshold that we need to select. Soft thresholding sets any entry to zero if its absolute value is smaller that t (same as hard thresholding) but shrinks other values by t.

There is a different way to motivate these estimators as solutions to (regularized) least-squares problems.

1. **Classical Estimator:** The classical estimator is the solution to the least-squares problem:
$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2} \|y - \theta\|_2^2.$$

2. **Hard Thresholding Estimator:** The hard-thresholding estimator is the solution to the problem:
$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2} \|y - \theta\|_2^2 + \frac{t^2}{2} \sum_{i=1}^{d} \mathbb{I}(\theta_i \neq 0).$$
The penalty here is known as the ℓ_0 penalty, it penalizes solutions that are non-sparse. You should convince yourself that for each coordinate, we only decide to use a non-zero estimate if $y_i^2 \geq t^2$ and if we use a non-zero estimate we should just match $\theta_i = y_i$ to minimize the penalty, this is just the hard-thresholding estimator.

3. **Soft Thresholding Estimator:** The soft-thresholding estimator is the solution to the problem:

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2}\|y - \theta\|_2^2 + t\sum_{i=1}^{d} |\theta_i|.$$

   Showing that this is equivalent to the soft-thresholding estimator is a little bit more work (and requires some basic sub-gradient calculus) so we'll skip it.

A basic question is then: what is the risk of the hard/soft thresholding estimators? They will turn out to be nearly identical for appropriate choices of the penalty so we will analyze the hard-thresholding estimator here.

Maximum of Gaussians: Before we continue we take another detour to study the maximum of Gaussian RVs. Here is a lemma:

Lemma 30.1 *Suppose that, $\epsilon_1, \ldots, \epsilon_d \sim N(0, \sigma^2)$ then with probability at least $1 - \delta$,*

$$\max_{i=1}^{d} |\epsilon_i| \leq \sigma\sqrt{2\log(2d/\delta)}.$$

Proof: One can slightly improve constants by a more refined proof. Recall, our Gaussian tail bound, if $\epsilon \sim N(0, \sigma^2)$:

$$\mathbb{P}(|\epsilon| \geq t) \leq 2\exp(-t^2/(2\sigma^2)),$$

so by the union bound we obtain that,

$$\mathbb{P}(\max_{i} |\epsilon_i| \geq t) \leq 2d\exp(-t^2/(2\sigma^2)),$$

which implies the desired lemma. ■

With this lemma we can analyze the hard-thresholding estimator, and obtain the following theorem. Once again one can improve the constant factors (and some other minor things) by a more careful analysis.

Theorem 30.2 *Suppose we choose the threshold:*

$$t = 2\sigma\sqrt{\frac{2\log(2d/\delta)}{n}},$$

then with probability at least $1 - \delta$,

$$\|\widehat{\theta} - \theta\|_2^2 \leq 9\sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{t^2}{4}\right\}.$$

Proof: We condition on the event from the previous lemma, i.e. that

$$\max_{i=1}^{d} |\epsilon_i| \leq \sigma\sqrt{2\log(2d/\delta)} \leq \frac{t}{2}.$$

Now, observe that,

$$\|\widehat{\theta} - \theta\|_2^2 = \sum_{i=1}^{d} (\widehat{\theta}_i - \theta_i)^2,$$

so we can consider each co-ordinate separately. Let us consider some cases:

1. If for any co-ordinate $|\theta_i| \leq \frac{t}{2}$ our estimate is 0, so our risk for that coordinate is simply θ_i^2.
2. If $|\theta_i| \geq \frac{3t}{2}$ our estimate is simply $\widehat{\theta}_i = y_i$ so our risk is simply $\epsilon_i^2 \leq \frac{t^2}{4}$.
3. If $\frac{t}{2} \leq |\theta_i| \leq \frac{3t}{2}$, then our risk,

$$(\widehat{\theta}_i - \theta_i)^2 = (y_i \mathbb{I}(|y_i| \geq t) - \theta_i)^2 = \theta_i^2 \mathbb{I}(|y_i| < t) + \epsilon_i^2 \mathbb{I}(|y_i| \geq t) \leq \max\{\epsilon_i^2, \theta_i^2\} \leq \frac{9t^2}{4}.$$

Putting these together we see that,

$$\|\widehat{\theta} - \theta\|_2^2 \leq 9 \sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{t^2}{4}\right\}.$$

■

Corollary (optional): To bound the actual risk we need the expected loss. We need to slightly modify the argument above. Let us denote by

$$\alpha = \max_i |\epsilon_i|.$$

We pick

$$t = 2\sigma\sqrt{\frac{6\log(2d)}{n}},$$

and note that when $\alpha \leq t/2$, our previous analysis shows that $\|\widehat{\theta} - \theta\|_2^2 \leq 9\sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{t^2}{4}\right\}$, and when $\alpha \geq t/2$ we note that,

$$\|\widehat{\theta} - \theta\|_2^2 \leq C d \alpha^2.$$

Now, we note that,

$$\begin{aligned} \mathbb{E}\|\widehat{\theta} - \theta^*\|_2^2 &= \int_0^\infty \mathbb{E}(\|\widehat{\theta} - \theta^*\|_2^2 | \alpha = u)\mathbb{P}(\alpha = u)du \\ &\leq \text{err} + \int_{\text{err}}^\infty \mathbb{P}(\|\widehat{\theta} - \theta^*\|_2^2 \geq u)du \end{aligned}$$

30.2 Interpreting the bound

We have seen that the risk of the hard-thresholding estimator is upper bounded by,

$$R(\widehat{\theta},\theta) \lesssim \sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{\sigma^2 \log(d)}{n}\right\}.$$

In the worst case, all of the θ_is are non-zero or large, and we obtain that the risk is upper bounded by $\sigma^2 d \log d/n$, which is almost the same as that of the classical estimator (except for the log-factor which you can eliminate by a more careful analysis).

On the other hand if θ is s-sparse, i.e. only s of its entries are non-zero then you observe that the risk looks like:

$$R(\widehat{\theta},\theta) \lesssim \frac{\sigma^2 s \log(d)}{n},$$

which means that the hard-thresholding estimator is consistent even if $d \gg n$, so long as $s\log(d)/n \to 0$. In fact you can obtain non-trivial estimates even when d is exponentially larger than n. This is quite miraculous: we can avoid the curse of dimensionality in a parametric problem if the target parameter θ is sufficiently structured.

Perhaps one might not expect the vector θ to be exactly sparse but only approximately so, i.e. in some meaningful sense most of its entries are small. There are various ways to measure sparsity and these will all lead to different, interesting bounds on the risk. Just to get a flavor of this idea, suppose we considered ℓ_1 sparsity, i.e.

$$\sum_{i=1}^{d} |\theta_i| \leq R,$$

for some radius R. Then we can see that, the number of entries of θ larger than R/k is at most k, for any k. So for any k, we can use the previous risk bound to obtain:

$$\begin{aligned} R(\widehat{\theta},\theta) &\lesssim \sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{\sigma^2 \log(d)}{n}\right\} \\ &\lesssim \sum_{i:\theta_i^2 \geq \sigma^2\log(d)/n} \frac{\sigma^2 \log(d)}{n} + \sum_{i:\theta_i^2 \leq \sigma^2\log(d)/n} \theta_i^2. \end{aligned}$$

Since the number of entries of the vector θ that can exceed $\sigma\sqrt{\log(d)/n}$ is at most $\sqrt{n}R/\sigma\sqrt{\log(d)}$, we obtain that bound that,

$$\begin{aligned} R(\widehat{\theta}, \theta) &\lesssim R\sigma\sqrt{\frac{\log(d)}{n}} + \sum_{i:\theta_i^2 \leq \sigma^2 \log(d)/n} \theta_i^2 \\ &\lesssim R\sigma\sqrt{\frac{\log(d)}{n}} + \sigma\sqrt{\frac{\log(d)}{n}} \sum_{i:\theta_i^2 \leq \sigma^2 \log(d)/n} |\theta_i| \\ &\lesssim 2R\sigma\sqrt{\frac{\log(d)}{n}}. \end{aligned}$$

Notice that the rate of convergence is different from the s-sparse case, roughly behaving as $1/\sqrt{n}$ instead of $1/n$. Ignoring this distinction however, the result should again surprise you – we are not even assuming that the unknown vector θ is sparse, just that is has ℓ_1-norm that is controlled, and once again we can obtain consistent estimators when $d \gg n$. More generally, there are many ways in which we can measure sparsity or impose structure on the unknown parameter, and depending on the structural assumption we might obtain improved rates of convergence.

While all of this might seem extremely contrived, we will see in the next lecture that similar things happen in high-dimensional regression (under appropriate assumptions), and are well-understood now to happen in many other interesting models. Roughly, this is the area of high-dimensional statistics: the main features are we do not assume the dimension of the model, i.e. the number of parameters is fixed as $n \to \infty$, and often we use structural assumptions of various kinds (typically variants of sparsity) to obtain fast rates of convergence.