36-705: Intermediate Statistics **Fall 2017**

Lecture 29: November 13

Lecturer: Siva Balakrishnan

We first go back and analyze hard-thresholding in the Gaussian sequence model. We will then re-visit the consequences from the end of last class and then talk about regression.

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 29.1 *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 29.2 *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 (recalling that in the sequence model the noise variance is σ^2/n),

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

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. One can use the high-probability bound. For instance note that with probability at least $1 - 1/d^2$ we have that for some big constant $C > 0$,

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

and that we can always trivially upper bound the loss as,

$$\|\widehat{\theta} - \theta\|_2^2 \leq \frac{C\sigma^2 d \log(d)}{n}.$$

Putting these together with the law of total expectation you will obtain the bound that for some constant $C > 0$,

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

29.1 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.

29.2 Low Dimensional Linear Regression – Review

We will now review some basic facts about linear regression. We will not go into much detail here – if you have not seen this all before I recommend reading Chapter 13 of the Wasserman book.

Linear regression is a tool to approximate the conditional expectation of $Y|X$ by a linear function of X. If you take a class on linear regression you will learn in Lecture 1 not to assume the true regression function is linear. We will assume the true regression function is linear, i.e. we assume we observe pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where (x, y) are linked via the linear model:

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i,$$

where $y_i \in \mathbb{R}, x_i \in \mathbb{R}^d$ and $\epsilon_i \sim N(0, \sigma^2)$. We let

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T.$$

Its population counterpart is the covariance matrix of the design, i.e. $\Sigma = \mathbb{E}[xx^T]$.

Least Squares: We will assume throughout that $\widehat{\Sigma}$ is invertible. In the setting where $\widehat{\Sigma}$ is invertible a natural approach to estimating β^* is to use least squares, i.e. we consider the estimator:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle x_i, \beta \rangle)^2.$$

In this setting the least squares estimator can be written in closed form as:

$$\widehat{\beta} = \widehat{\Sigma}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} x_i y_i \right].$$

It should be straightforward to convince yourself that under the model we wrote down (with Gaussian errors) the MLE is the same as the least squares estimator.

In general, we will often assume that the covariates x_i are random, i.e. are drawn from some distribution – this is known as random design regression. Some people alternatively assume that the x_is are fixed, and the only thing that is stochastic is the noise. This is known as fixed design regression. We often write the *design matrix* $X \in \mathbb{R}^{n \times d}$ as the matrix with rows equal to the data samples, then we can write the least squares estimator as:

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

You can verify that,

$$\begin{aligned}\widehat{\beta} &= (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta^* + \epsilon) \\ &= \beta^* + (X^T X)^{-1} X^T \epsilon \\ &\sim N(\beta^*, \sigma^2 (X^T X)^{-1}).\end{aligned}$$

There are several possible quantities of interest in linear regression.

1. The in-sample prediction error, i.e.:

$$\mathbb{E}\left[\frac{\|X\widehat{\beta} - X\beta^*\|_2^2}{n} \right].$$

2. ℓ_2 error in estimating β^*, i.e. $\mathbb{E}[\|\widehat{\beta} - \beta^*\|_2^2]$.
3. The support recovery error (makes most sense when β^* is sparse):

$$\mathbb{P}(\text{supp}(\widehat{\beta}) \neq \text{supp}(\beta^*)).$$

Let us quickly review some of these quantities for low-dimensional regression.

29.2.1 In-sample prediction error

Since under our assumptions:

$$\widehat{\beta} \sim N(\beta^*, \sigma^2 (X^T X)^{-1}).$$

We can see that:

$$X\widehat{\beta} \sim N(X\beta^*, \sigma^2 X(X^T X)^{-1} X^T),$$

(if this is not clear to you, see Wikipedia - multivariate normal distribution, the section on affine transformation). This yields,

$$\begin{aligned}\mathbb{E}\|X\widehat{\beta} - X\beta^*\|_2^2 &= \sigma^2 \mathbb{E}\left[\text{tr}(X(X^T X)^{-1} X^T)\right] \\ &= \sigma^2 d,\end{aligned}$$

since the matrix $P = X(X^T X)^{-1} X^T$ is a (full-rank) projection matrix (i.e. $P^2 = P$) all of its eigenvalues are 1. This gives us that,

$$\mathbb{E}\left[\frac{\|X\widehat{\beta} - X\beta^*\|_2^2}{n}\right] = \frac{\sigma^2 d}{n}.$$

29.2.2 ℓ_2 error

Again, under our assumptions we know that,

$$\widehat{\beta} \sim N(\beta^*, \sigma^2 (X^T X)^{-1}).$$

So we obtain that,

$$\mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 = \sigma^2 \mathbb{E}\left[\text{tr}((X^T X)^{-1})\right].$$

There are various ways to understand this quantity, and we will just provide some rough heuristics. First, notice that,

$$\mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 = \frac{\sigma^2}{n}\mathbb{E}\left[\text{tr}((X^T X/n)^{-1})\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\text{tr}(\widehat{\Sigma}^{-1})\right].$$

Assuming that $\widehat{\Sigma}$ has eigenvalues that are lower bounded by some small constant $c > 0$, then we will have that,

$$\mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 \le \frac{\sigma^2 d}{cn},$$

which is the usual parametric rate. This result can be as related to our general result on the MLE, $\widehat{\beta}$ is the MLE, and the Fisher information is the expected Hessian of the log-likelihood, and is just,

$$I_n(\beta) = n\mathbb{E}\left[\frac{X^TX}{n\sigma^2}\right] = \frac{n\Sigma}{\sigma^2}.$$

independent of β. So we can conclude that,

$$\sqrt{n}(\widehat{\beta} - \beta^*) \xrightarrow{d} N\left(0, \sigma^2\Sigma^{-1}\right).$$

29.3 High-dimensional Regression

In high-dimensional regression, we are interested in the setting where the covariate distribution has dimension $d \gg n$. The first thing to observe is that even if our old analysis worked (it does not) the prediction error and ℓ_2 error both scale as $\sigma^2 d/n$ which does not go to 0 as we increase the sample-size, which would mean that our methods are inconsistent. From a minimax perspective, it turns out that this is unavoidable, i.e. it is impossible to consistently estimate the regression vector β^*, when $d \gg n$, and we need to turn to structural assumptions to make progress.

A perhaps even more alarming aspect of high-dimensional regression is that the least-squares estimator is no longer well-defined. To see this, observe that the assumption that $\widehat{\Sigma}$ is invertible (which is completely benign in low-dimensions) can never hold in high-dimensions. In particular the matrix,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T,$$

has rank at most n (it is a sum of rank 1 matrices) and is a $(d \times d)$ matrix, so is clearly not invertible if $d > n$. The way to picture this is that in high-dimensions there will be many vectors β such that, $y = X\beta$ which have least squares error of 0 (i.e. exactly pass through all the samples).

This is a form of over-fitting, and one way to avoid this is to use regularization. This is roughly equivalent to imposing some type of structure on the unknown β^* and then attempting to recover β^* by leveraging this structure. We will again focus on versions of sparsity, i.e. settings where β^* is either exactly sparse (i.e. has s non-zero entries) or is approximately sparse (i.e. has bounded ℓ_1 norm).

Analogous to the Gaussian sequence model there are two estimators that one might consider:

1. **Hard-Thresholding type estimator:** The analog of hard thresholding is:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{t^2}{2}\sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0).$$

   This is usually called best-subset regression. The best way to think about the nomenclature is to consider a closely related estimator:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2,$$
$$\text{subject to } \sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0) \leq k,$$

   where now we have a different tuning parameter $k > 0$ (instead of t). You should be able to (with some effort) convince yourself of the fact that these two programs are exactly equivalent, i.e. if you fix any $t > 0$ and solve the first program, then there is some k for which you obtain exactly the same solution. The first form is sometimes called the penalized-form and the second is called the constrained-form.

   The natural way to implement the second estimator would be to enumerate all subsets of size k, fit a regression on this subset and then pick the subset, and estimate β that has lowest mean -squared error. Hence the name, "best-subset regression".

2. **Soft-Thresholding type estimator:** The analog of soft thresholding is known as the LASSO, i.e. the Least Absolute Selection and Shrinkage Operator,

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

   Analogous to the above, one can consider a closely related estimator:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2,$$
$$\text{subject to } \sum_{i=1}^{d} |\beta_i| \leq k,$$

   again there is an equivalence, i.e. every value of t corresponds to some value of k. This program is a convex program, and simple methods (roughly, gradient descent with tweaks) can be used to solve it quite fast. There is typically no closed-form solution but that is not a huge problem.

This brings us to an important distinction between the Gaussian sequence model and regression. In the Gaussian sequence model (no X) both of these programs had simple closed-form solutions, whereas now this is no longer the case. More importantly, best-subset is computationally intractable but the LASSO is not.

With this motivation in place, let us study the prediction error of the LASSO. We begin with some assumptions, for simplicity we will study the constrained form of the LASSO, and further we will just assume that the tuning parameter k is chosen to be exactly $\|\beta^*\|_1$. In practice, one might choose this tuning parameter by cross-validation or some other method.

To simplify our calculations we will also assume the design matrix X is column-normalized, i.e. for each column j of the matrix:

$$\sum_{i=1}^{n} X_{ij}^2 \leq n.$$

You can ensure this by re-normalizing every column of X. This does change β^* (and its ℓ_1 norm).

Theorem 29.3 *Suppose we consider the constrained-LASSO with $k = \|\beta^*\|_1$, then the prediction error of our estimator, with probability at least $1 - \delta$, satisfies:*

$$\frac{1}{n}\|X\widehat{\beta} - X\beta^*\|_2^2 \leq 4\sigma\|\beta^*\|_1\sqrt{\frac{2\log(2d/\delta)}{n}}.$$

This bound is exactly analogous to the bound on the error of the hard/soft-thresholding estimator in the Gaussian sequence model when we assumed that the ℓ_1 norm of the mean vector θ^* was bounded. Notice again, that the prediction error goes to 0 with n, even in settings where $d \gg n$.

This result is due to Greenshtein and Ritov and really kicked off the wave of high-dimensional statistics. It showed that high-dimensional prediction was possible (at least in the linear model). Several later works showed that under stronger assumptions, one could achieve small ℓ_2 error and even exactly identify the non-zero components of β^* (i.e. do feature selection) in the high-dimensional setting. Furthermore, most of these phenomena generalize to general parametric models (for instance, high-dimensional logistic regression, high-dimensional graphical model estimation and so on).

Proof (optional): To prove this we note that, since we selected the tuning parameter to be $\|\beta^*\|_1$, the vector β^* is feasible for the program and $\widehat{\beta}$ is optimal, so we have the so-called "basic inequality":

$$\frac{1}{2n}\|y - X\widehat{\beta}\|_2^2 \leq \frac{1}{2n}\|y - X\beta^*\|_2^2,$$

where we divided both sides by n for convenience. Re-arranging this inequality we obtain that,

$$\frac{1}{2n}\|X\widehat{\beta} - X\beta^*\|_2^2 \leq \frac{1}{n}\langle \epsilon, X\widehat{\beta} - X\beta^*\rangle = \langle \frac{X^T\epsilon}{n}, \widehat{\beta} - \beta^*\rangle,$$

where ϵ is the noise in the linear model. Holder's inequality tells us that for any two vectors $a, b \in \mathbb{R}^d$,

$$\langle a,\, b\rangle \le \left(\max_{i=1}^{d} a_i\right)\left(\sum_{i=1}^{d} |b_i|\right).$$

Applying this inequality we obtain,

$$\frac{1}{n}\|X\widehat{\beta} - X\beta^*\|_2^2 \le 2\|\widehat{\beta} - \beta^*\|_1 \max_{i=1}^{d} \frac{X_i^T\epsilon}{n}$$

where X_i denotes the i-th column of the design. Now, by the triangle inequality, $\|\widehat{\beta} - \beta^*\|_1 \le 2\|\beta^*\|_1$ (recall that we constrained our optimal solution to have ℓ_1 norm at most $\|\beta^*\|_1$), so it only remains to bound $\max_{i=1}^{d} \frac{X_i^T\epsilon}{n}$.

Each entry here, has a Gaussian distribution with mean 0 and variance $\sigma^2\|X_i\|_2^2/n \le \sigma^2/n$, using our column normalization assumption. So with probability at least $1-\delta$, we have that,

$$\max_{i=1}^{d} \frac{X_i^T\epsilon}{n} \le \sigma\sqrt{\frac{2\log(2d/\delta)}{n}},$$

and combining these facts we obtain the desired bound.