IEOR 265 – Lecture 3
Sparse Linear Regression

1 M^* Bound

Recall from last lecture that the reason we are interested in complexity measures of sets is because of the following result, which is known as the M^* Bound. Suppose (i) $\mathcal{T} \subset \mathbb{R}^p$ is bounded and symmetric (i.e., $\mathcal{T} = -\mathcal{T}$), and (ii) $\mathcal{E}$ is a random subspace of $\mathbb{R}^p$ with fixed codimension n and drawn according to an appropriate distribution. More specifically, suppose $\mathcal{E} = \text{nullspace}(A) = \{x : Ax = 0\}$, where $A \in \mathbb{R}^{n\times p}$ has iid $\mathcal{N}(0,1)$ entries, meaning they are iid Gaussian entries with zero-mean and unit variance. Then we have

$$\mathbb{E}\Big(\text{diam}(\mathcal{T} \cap \mathcal{E})\Big) \leq \frac{Cw(\mathcal{T})}{\sqrt{n}}.$$

1.1 Three Useful Results

Before we can prove the M^* Bound, we need three results. The *first result* is that (i) if U is a symmetric random variable (e.g., its pdf is even $f_U(-u) = f_U(u)$), and (ii) ϵ is a Rademacher random variable (i.e., $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$) that is independent of U; then ϵU has the same distribution as U. (To see why this is the case, note that if we condition $\epsilon \cdot U$ on the two possible values $\epsilon \pm 1$, then the conditional distribution is the same as the distribution of U because of the symmetry of U.)

The *second result* is that if $\varphi(\cdot)$ is a Lipschitz continuous function (with Lipschitz constant L) with $\varphi(0) = 0$, then we have

$$\mathbb{E}\sup_{t\in T}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \psi(t_i)\Big| \leq 2L \cdot \mathbb{E}\sup_{t\in T}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot t_i\Big|.$$

If $L = 1$, then the function $\varphi(\cdot)$ is also known as a *contraction*, and so this result is sometimes known as a contraction comparison theorem.

The *third result* is a concentration of measure result, and these types of results can be thought of as finite sample versions of the central limit theorem. In particular, suppose $x_1, \ldots, x_n$ are iid $\mathcal{N}(0,1)$, and let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant L. Then we have

$$\mathbb{P}(|\varphi(x_1,\ldots,x_n) - \mathbb{E}\varphi(x_1,\ldots,x_n)| > t) \leq \exp(-t^2/2L^2).$$

1.2 Proof of M^* Bound

Let a_i' be the i-th row of A. We begin by noting that $a_i't$, where $t \in \mathbb{R}^p$ is an arbitrary vector, is a Gaussian random variable with $\mathbb{E}(a_i't) = 0$ and $\text{var}(a_i't) = \|t\|_2^2$, since the entries of a_i are iid $\mathcal{N}(0,1)$. Thus, $|a_i't|$ is a Folded Normal Distribution and has mean

$$\mathbb{E}(|a_i't|) = \sqrt{\frac{2}{\pi}} \cdot \|t\|_2.$$

Next, we use a common trick known as *symmetrization*. In particular, observe that

$$\begin{aligned}
\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(|a_i't| - \mathbb{E}|a_i't| \Big) \Big| &\leq \mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(|a_i't| - |\tilde{a}_i't| \Big) \Big| \\
&\leq \mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(\epsilon_i \cdot (|a_i't| - |\tilde{a}_i't|) \Big) \Big| \\
&\leq 2\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(\epsilon_i \cdot |a_i't| \Big) \Big|,
\end{aligned}$$

where (i) the first line follows by Jensen's inequality, (ii) the second line follows by the first useful result from above, and (iii) the third line follows from the triangle inequality. Since the absolute value function is Lipschitz continuous with $L = 1$ and has $|0| = 0$, we can apply the second useful result from above. This gives the bound

$$\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(|a_i't| - \mathbb{E}|a_i't| \Big) \Big| \leq 4\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(\epsilon_i \cdot a_i't \Big) \Big| = 4\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \Big(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot a_i \Big)' t \Big|.$$

However, note that conditioned on ϵ_i the random variable $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \cdot a_i$ has distribution $\mathcal{N}(0,1)$ since the entries of a_i are iid $\mathcal{N}(0,1)$. Since this holds conditioned on all possible values of ϵ_i, we have that unconditionally the random variable $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \cdot a_i$ has distribution $\mathcal{N}(0,1)$. Thus, we can make the variable substitution $g = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \cdot a_i$, which gives

$$\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(|a_i't| - \mathbb{E}|a_i't| \Big) \Big| \leq \frac{4}{\sqrt{n}} \mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big| g't \Big| = \frac{4}{\sqrt{n}} \mathbb{E} \Big(\sup_{t \in \mathcal{T} \cap \mathcal{E}} g't \Big),$$

where the last equality follows since $\mathcal{T} \cap \mathcal{E}$ is symmetric (since $\mathcal{T}$ is assumed to be symmetric, and since $\mathcal{E}$ is symmetric because it is defined as a nullspace).

The last step is to make substitutions. First, recall that $w(\mathcal{T} \cap \mathcal{E}) = \mathbb{E}\Big(\sup_{t \in \mathcal{T} \cap \mathcal{E}} g't \Big)$ by definition. Second, recall we showed above that $\mathbb{E}(|a_i't|) = \sqrt{\frac{2}{\pi}} \cdot \|t\|_2$. Third, note that $a_i't = 0$ whenever $t \in \mathcal{E}$ by definition of $\mathcal{E}$ as the nullspace of the matrix A. Making these substitutions into the last inequality above, we have

$$\mathbb{E} \sup_{t \in \mathcal{T} \cap \mathcal{E}} \Big(\sqrt{\tfrac{2}{\pi}} \cdot \|t\|_2 \Big) \leq \frac{4 \cdot w(\mathcal{T} \cap \mathcal{E})}{\sqrt{n}}.$$

Since (i) diam$(\mathcal{T} \cap \mathcal{E}) = \sup_{t \in \mathcal{T} \cap \mathcal{E}} \|t\|_2$, and (ii) $w(\mathcal{T} \cap \mathcal{E}) \leq w(\mathcal{T})$ because $\mathcal{T} \cap \mathcal{E} \subseteq \mathcal{T}$, this proves the M^* Bound.

1.3 High Probability Version

The M^* Bound has a version that holds with high probability, which can be proved using the third useful result from above. In particular, assume that (i) $\mathcal{T} \subset \mathbb{R}^p$ is bounded and symmetric (i.e., $\mathcal{T} = -\mathcal{T}$), and (ii) $\mathcal{E}$ is a random subspace of $\mathbb{R}^p$ with fixed codimension n and drawn according to an appropriate distribution. More specifically, suppose $\mathcal{E} = \text{nullspace}(A) = \{x : Ax = 0\}$, where $A \in \mathbb{R}^{n \times p}$ has iid $\mathcal{N}(0, 1)$ entries. Then with probability at least $1 - 2\exp(-nt^2/2\text{diam}(\mathcal{T})^2)$ we have

$$\text{diam}(\mathcal{T} \cap \mathcal{E}) \leq \frac{Cw(\mathcal{T})}{\sqrt{n}} + Ct.$$

The benefit of this version is that allows us to state that the diameter of $\mathcal{T} \cap \mathcal{E}$ is small with high probability, as opposed to merely stating that the expected diameter is small.

2 High-Dimensional Linear Regression

Recall the setting of a linear model with noiseless measurements: We have pairs of independent measurements (x_i, y_i) for $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, and the system is described by a linear model

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i = x_i' \beta.$$

In the case where p is large relative to n, the OLS estimator has high estimation error. We now return to the original question posed in the first lecture:

How can we estimate β when p is large relative to n? In general, this problem is ill-posed. However, suppose we know that the majority of entries of β are exactly equal to zero. We can imagine that our chances of estimating β are improved in this specific setting. Intuitively, the reason is that if we knew which coefficients were zero beforehand, then we could have used a simple OLS estimator to estimate the nonzero coefficients of β. However, the situation is more complicated in our present case. Specifically, we do not know *a priori* which coefficients of β are nonzero. Fortunately, it turns out that this is not a major impediment.

2.1 Estimator for Sparse Linear Models

Assume that the true value of β has at most s nonzero components and has $|\beta_j| \leq \mu$. Now suppose we solve the feasibility problem

$$\hat{\beta} = \text{find} \Big\{ \beta : Y = X\beta, \beta \in \{t \in \mathbb{R}^p : \|t\|_0 \leq s, |t_j| \leq \mu\} \Big\},$$

where $\|t\|_0 = \sum_{j=1}^p \mathbf{1}(t_j \neq 0)$ counts the number of nonzero coefficients of t. By the M^* bound we know that if the entries of x_i are iid $\mathcal{N}(0,1)$, then we have

$$\mathbb{E}\Big(\|\hat{\beta} - \beta\|_2\Big) \leq \frac{C \cdot w(\mathcal{T})}{\sqrt{n}},$$

where $\mathcal{T} = \{t \in \mathbb{R}^p : \|t\|_0 \leq s, |t_j| \leq \mu\}$. The final question is: What is the value of $w(\mathcal{T})$?

To bound $w(\mathcal{T})$, observe that that $w(\mathcal{T}) \leq w(\mathcal{S})$ if $\mathcal{T} \subseteq \mathcal{S}$. Define the ℓ_1-ball with radius $s \cdot \mu$ to be $\mathcal{S} = \{t : \|t\|_1 \leq s \cdot \mu\}$, and note that by construction $\mathcal{T} \subseteq \mathcal{S}$. Thus, we have that

$$w(\mathcal{T}) \leq s \cdot \mu\sqrt{4 \log p},$$

where we have used the Gaussian average of an ℓ_1-ball from last lecture. This allows us to conclude that

$$\mathbb{E}\Big(\|\hat{\beta} - \beta\|_2\Big) \leq C\mu \cdot s\sqrt{\frac{\log p}{n}}.$$

Thus, solving the above feasibility problem is able to estimate the β vector in high dimensions because the dependence on dimensionality is logarithmic. However, this feasibility problem is nonconvex, and so an alternative solution approach is needed.

2.2 Convex Estimators

Instead, consider the convex feasibility problem

$$\hat{\beta} = \text{find } \Big\{\beta : Y = X\beta, \beta \in \{t \in \mathbb{R}^p : \|t\|_1 \leq s \cdot \mu\}\Big\}.$$

By the same reasoning as above, when x_i has iid $\mathcal{N}(0,1)$ entries we have that

$$\mathbb{E}\Big(\|\hat{\beta} - \beta\|_2\Big) \leq C\mu \cdot s\sqrt{\frac{\log p}{n}}.$$

Note that we can reformulate the convex feasibility problem as

$$\hat{\beta} = \arg\min_{\beta}\{\|\beta\|_1 \mid X\beta = Y\}.$$

It turns out that a solution of this optimization problem gives a solution to the convex feasibility problem. The reason is that the true value of β and the estimated value $\hat{\beta}$ are both feasible (to $X\beta = Y$) when there is no noise, and so we have $\|\hat{\beta}\|_1 \leq \|\beta\|_1 \leq s \cdot \mu$. Thus, $\hat{\beta}$ is also a solution to the convex feasibility problem.

Another reformulation of the convex feasibility problem is

$$\hat{\beta} = \arg\min_{\beta}\{\|Y - X\beta\|_2^2 \mid \|\beta\|_1 \leq s \cdot \mu\}.$$

Note that the true value of β is feasible (to this convex optimization problem) by assumption, and provides a minimum of $\|Y - X\beta\|_2^2 = 0$. Hence, any minimizer to this second convex optimization problem provides a solution with $\|Y - X\hat{\beta}\|_2^2 = 0$, which implies that $Y = X\hat{\beta}$. As a result, any solution to this optimization problem is also a solution to the convex feasibility problem.

3 Lasso Regression

We have so far developed an estimator for high-dimensional linear models when our measurements y_i have no noise. However, suppose we have noisy measurements

$$y_i = x_i'\beta + \epsilon_i,$$

where ϵ_i are iid zero-mean and bounded random variables. A common approach to estimating the parameters is known as *lasso regression*, and is given by

$$\hat{\beta} = \arg\min_{\beta}\{\|Y - X\beta\|_2^2 \mid \|\beta\|_1 \leq \lambda\}.$$

This estimator is often written as

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \mu\|\beta\|_1,$$

and the reason is that the minimizer to the two above optimization problems are identical for an appropriate choice of the value μ.

To see why, consider the first optimization problem for $\lambda > 0$. Slater's condition holds, and so the Langrange dual problem has zero optimality gap. This dual problem is given by

$$\begin{aligned} &\max_{\nu\geq 0}\min_{\beta} \|Y - X\beta\|_2^2 + \nu(\|\beta\|_1 - \lambda) \\ \Rightarrow &\max_{\nu}\{\|Y - X\hat{\beta}^{\nu}\|_2^2 + \nu\|\hat{\beta}^{\nu}\|_1 - \nu\lambda : \nu \geq 0\}. \end{aligned}$$

Let the optimizer be ν^*, and set $\mu = \nu^*$. Then the solution $\hat{\beta}^{\mu}$ minimizes $\|Y - X\beta\|_2^2 + \mu\|\beta\|_1$ and is identical to $\hat{\beta} = \arg\min_{\beta}\{\|Y - X\beta\|_2^2 \mid \|\beta\|_1 \leq \lambda\}$.

This result is useful because it has a graphical interpretation that provides additional insight. Visualizing the constrained form of the estimator provides intuition into why using the ℓ_1-norm provide sparsity in the estimated coefficients $\hat{\beta}$.

4 More Reading

The material in these sections follows that of

R. Vershynin (2014) *Estimation in high dimensions: a geometric perspective*, in Sampling Theory, a Renaissance. Springer. To appear.

R. Vershynin (2009) *Lectures in geometric functional analysis*. Available online:
http://www-personal.umich.edu/~romanv/papers/GFA-book/GFA-book.pdf.

M. Ledoux and M. Talagrand (1991) *Probability in Banach Spaces*. Springer.

More details about these sections can be found in the above references.