Lecture 8: Sparsity

Isabella Zhu

4 March 2025

Definition 0.1. We call θ^* **sparse** if it contains at most s non-zero entries. This implies

$$\theta^* \in B_0(s) = \{\theta \in \mathbb{R}^d \mid \sum 1(\theta_i \neq 0) \leq s\}$$

Today we are under the assumption that θ^* is sparse.

§1 Sparse Linear Regression

Assume once again that $Y = \mathbb{X}\theta^* + \epsilon$ and that each element of ϵ is independent and subgaussian with variance proxy σ^2. We also assume sparsity, i.e. $\theta^* \in B_0(s)$.

Our constrained least squares problem looks like

$$\hat{\theta} = \hat{\theta}_{B_0(s)} \in \underset{\theta \in B_0(s)}{\operatorname{argmin}} |Y - \mathbb{X}\theta|_2^2$$

which is not an easy problem to solve. We will show properties instead.

§1.1 Review from Last Lecture

Note that we have

$$|Y - \mathbb{X}\hat{\theta}|^2 \leq |Y - \mathbb{X}\theta^*|^2$$

which rearranges to

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2 \leq 2\langle \epsilon, \mathbb{X}(\hat{\theta} - \theta^*)\rangle$$

Let $u = \frac{\mathbb{X}(\hat{\theta}-\theta^*)}{|\mathbb{X}(\hat{\theta}-\theta^*)|}$ as before. All the arguments are so far the same. We know that

$$u \in \operatorname{span}(\mathbb{X}_j, j \in S), \ \ |S| \leq 2s.$$

So we can bound by $2s$ instead of by d.

Definition 1.1. The **support** of v, denoted $\operatorname{supp}(v)$, is the set of all indices where the elements are nonzero.

Thus, we get

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2 \leq 2\langle \epsilon, \mathbb{X}(\hat{\theta} - \theta^*)\rangle \leq 2 \max_{S \subset [d], |S| \leq 2s} \sup_{v \in \mathbb{R}^d, \operatorname{supp}(v) \leq 2s} \langle \epsilon, \frac{\mathbb{X}v}{|\mathbb{X}v|}\rangle$$

Let $\Phi_s = \begin{bmatrix}\phi_1 & \phi_2 & \dots & \phi_{2s}\end{bmatrix}$ be an orthogonal basis for span. Then, we get

$$\langle \epsilon, \frac{\mathbb{X}v}{|\mathbb{X}v|^2}\rangle = \langle \epsilon, \frac{\Phi_s v}{|v|}\rangle = \langle \Phi_s^T \epsilon, \frac{v}{|v|}\rangle$$

which we already know how to bound from the previous lecture. Basically we replace occurrences of d with $2s$ instead.

We can also compute tail bounds, which we bound with ϵ-nets.

$$P \le \sum_{S\subset[d],|S|\le 2s} \mathbb{P}\left(\sup_{u\in B_2}\langle \epsilon, u\rangle > \sqrt{t}\right) \le \sum_{S\subset[d],|S|\le 2s} \mathbb{P}\left(2\max_{u\in N}\langle \epsilon, u\rangle > \sqrt{t}\right)$$

where N is a $\frac{1}{2}$-net of B_2. We have by union bound,

$$\sum_{S\subset[d],|S|\le 2s} \mathbb{P}\left(2\max_{u\in N}\langle \epsilon, u\rangle > \sqrt{t}\right) \le \sum_{s\in[d],|S|\le 2s}\sum_{u\in N} \mathbb{P}\left(\langle \epsilon, u\rangle > \sqrt{t}/2\right) \le \binom{d}{\le 2s} 5^{2s} e^{-t/8\sigma^2}$$

Set this equal to δ and solve for t to get the tail bounds.

Lemma 1.2

We have

$$\binom{d}{\le 2s} \le \left(\frac{ed}{2s}\right)^{2s}$$

Therefore, we can write

$$P \le \exp\left(2s\log\left(\frac{ed}{2s}\right) + 2s\log 5 - \frac{t}{8\sigma^2}\right) = \delta$$

Solving, we get

$$t = 8\sigma^2 2s\log\left(\frac{ed}{2s}\right) + 8\sigma^2 2s\log 5 + 8\sigma^2\log(1/\delta).$$

Theorem 1.3

The MSE of a sparse estimator is

$$\begin{aligned}\mathrm{MSE}(\mathbb{X}\hat{\theta}_{B_0(2s)}) &\le \frac{64\sigma^2}{n} s\log\left(\frac{ed}{2s}\right) + \frac{64\sigma^2}{n} s\log 5 + \frac{32\sigma^2}{n}\log\frac{1}{\delta} \\ &\lesssim \frac{\sigma^2 s}{n}\log\left(\frac{ed}{2s}\right) + \frac{32\sigma^2}{n}\log\frac{1}{\delta}\end{aligned}$$

Remark 1.4. In comparison to the bound from last class, we have an extra $\log\left(\frac{ed}{2s}\right)$, which we can think of as the "price to pay" for not knowing which indices are nonzero.

§2 SubGaussian Sequence Model

Assume that $\frac{\mathbb{X}^T\mathbb{X}}{n} = I_d$, so clearly $d \leq n$. Again we assume that $Y = \mathbb{X}\theta^* + \epsilon$, so we multiply by $\mathbb{X}$ and redefine Y to get

$$Y = \frac{\mathbb{X}^T Y}{n} = \theta^* + \frac{\mathbb{X}^T\epsilon}{n}$$

If we assume for now that $\epsilon \in N(0, I_n)$, then we can conclude that

$$\frac{\mathbb{X}^T\epsilon}{n} \sim N_d\left(0, \frac{\mathbb{X}^T\mathbb{X}}{n^2}\right) = N_d\left(0, \frac{1}{n}I_d\right).$$

Now let's think about the subGaussian case. Once again, all we say about ϵ is that for any u, we have

$$u^T\epsilon \sim \text{subG}\left(\frac{\sigma^2}{n}|u|^2\right).$$

Then, we have

$$\text{MSE} = \frac{1}{n}|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2 = (\hat{\theta} - \theta^*)^T\frac{\mathbb{X}^T\mathbb{X}}{n}(\hat{\theta} - \theta^*) = |\hat{\theta} - \theta^*|^2.$$

Remember that we redefined Y above! We have

$$\hat{\theta}_{B_0(s)} = \underset{\theta \in B_0(s)}{\text{argmin}}\, |Y - \theta|^2 = \sum_{j=1}^{d}(Y_j - \theta_j)^2$$

where we define

$$\theta_j = \begin{cases} Y_j & \text{if } |Y_j| \text{ is one of } s \text{ largest} \\ 0 & \text{otherwise} \end{cases}$$

We know from the first section that

$$\mathbb{E}[|\hat{\theta}_{B_0(s)} - \theta^*|^2] \lesssim \frac{\sigma^2 s}{n}\log\left(\frac{ed}{2s}\right)$$

However, now we are assuming that s is **unknown**. We want to be able to adapt to s.

§2.1 Sparsity Adaptive Thresholding Estimators

Definition 2.1. The **hard thresholding estimator** is

$$\hat{\theta}_j^{HARD} = \begin{cases} Y_j & \text{if } |Y_j| > 2\tau \\ 0 & \text{otherwise} \end{cases}$$

Theorem 2.2

If $\tau = \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$, then with probability at least $1 - \delta$,

$$|\hat{\theta}^{HARD} - \theta^*|^2 \lesssim \frac{\sigma^2 s}{n}\log\left(\frac{2d}{\delta}\right)$$

Note that this estimator is adaptive to s since τ doesn't depend on s.

Proof. Define the event A as

$$A = \{\max_j |\epsilon_j| \le \tau\}$$

We pick $\tau = \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$ so that $\mathbb{P}(A) \ge 1 - \delta$. On A, we have

1. $|Y_j| \ge 2\tau$ implies $|\theta_j^*| = |Y_j - \epsilon_j| \ge |Y_j| - |\epsilon_j| \ge \tau$.
2. $|Y_j| \le 2\tau$ implies $|\theta_j^*| \le |Y_j| + |\epsilon_j| \le 3\tau$.

We have

$$\begin{aligned}
|\hat{\theta} - \theta^*|^2 &= \sum_{j=1}^{d} |\hat{\theta}_j - \theta_j^*|^2 \\
&\le \sum_{j:|Y_j|>2\tau} \epsilon_j^2 + \sum_{j:|Y_j|<2\tau} |\theta_j^*|^2 \\
&\le \sum_{j:|Y_j|>2\tau} \min(\tau, \theta_j^*)^2 + \sum_{j:|Y_j|<2\tau} \min(3\tau, \theta_j^*)^2 \\
&\le \sum_{j:|Y_j|>2\tau} \min(\tau, \theta_j^*)^2 + \sum_{j:|Y_j|<2\tau} 9\min(\tau, \theta_j^*)^2 \\
&\le 9\sum_{j=1}^{d} \min(\tau, \theta_j^*)^2 \\
&\le 9\sum_{j:\theta_j \neq 0} \tau^2 \\
&= 9s\tau^2 \\
&= \frac{18s\sigma^2}{n} \log\left(\frac{2d}{\delta}\right).
\end{aligned}$$