Lecture 16–17: Sparse Regression (4/7,4/9)

Readings:

- Chap 7.1, 7.4 and 7.6 in Wainwright book.
- The following slides by Wainwright, which complements the above materials:

  `www.cs.berkeley.edu/~wainwrig/winedale/Wainwright_Winedale_060312.pdf`
- Chap 10.6 of Vershynin book.
- Chap 10.1–10.3 and 10.5 in Vershynin book, which presents a different view of sparse regression.

1 Sparse Regression in High Dimension

Suppose that

$$y = X\beta^* + e,$$

where $X \in \mathbb{R}^{n\times d}$ is the covariate matrix, $y \in \mathbb{R}^n$ the response, $e \in \mathbb{R}^n$ the noise, and $\beta^* \in \mathbb{R}^d$ the unknown parameter.

When $n < d$, we cannot hope to recover β^*, because there are less equations than variables. Need additional assumption/structure for β^*. We assume that β^* is sparse.

1.1 Lasso

To find a sparse solution, ideally we would like to solve the ℓ_0-regularized least square problem

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{2n}\left\|X\beta - y\right\|_2^2 + \lambda\left\|\beta\right\|_0, \tag{1}$$

where $\left\|\beta\right\|_0 := |\text{support}(\beta)|$ is the number of non-zero coordinates of β. The ℓ_0 "norm" $\left\|\cdot\right\|_0$ is in fact not a norm, and in particular not convex. The non-convex program (1) is in general hard to solve.

For computational tractability, we replace the ℓ_0 "norm" by its convex approximation, the ℓ_1 norm $\left\|\cdot\right\|_1$. This gives an ℓ_1-regularized least-squares problem known as *Lasso*:

$$\hat{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\left\|X\beta - y\right\|_2^2 + \lambda\left\|\beta\right\|_1. \tag{2}$$

Remark 1. It can be shown that the ℓ_1 norm is the *best* convex approximation of the ℓ_0 "norm" in a precise sense; see Lemma 5 below.

2 Bounds on Estimation Error of Lasso

We consider the following statistical setting:

- Sparsity: $\|\beta^*\|_0 \leq k$.
- Gaussian design: $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
- Gaussian noise: $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$.
- Independence: $X \perp e$.

Theorem 2. *Under the above setting, suppose that the sample size satisfies*

$$n \gtrsim k \log\left(\frac{d}{k}\right), \tag{3}$$

and the regularization parameter in Lasso (2) satisfies

$$\lambda \gtrsim \sigma\sqrt{\frac{\log d}{n}}.$$

Then with probability at least $1 - cd^{-2}$, the Lasso solution $\hat{\beta}$ satisfies the error bound

$$\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \sqrt{k}\lambda. \tag{4}$$

Consequences

- If we take $\lambda \asymp \sigma\sqrt{\frac{\log d}{n}}$, then we get the error bound $\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \sigma\sqrt{\frac{k\log d}{n}}$, which is (almost) minimax optimal. (We are very close though: the optimal bound is $\sigma\sqrt{\frac{k\log(d/k)}{n}}$.[1])
- In particular, if $\sigma \to 0$, then taking $\lambda \to 0$ gives $\left\|\hat{\beta} - \beta^*\right\|_2 \to 0$. That is, we have exact recovery in the noiseless setting.
- If we take λ larger than the optimal value $\sigma\sqrt{\frac{\log d}{n}}$, then Lasso still works, albeit with a larger, sub-optimal error (underfitting).
- If we take λ too small ($\ll \sigma\sqrt{\frac{\log d}{n}}$), then Theorem 2 is not applicable (overfitting).
- In practice, we may choose λ by estimating the noise variance σ^2, or by cross-validation.[2]

Remark 3. Theorem 2 bounds the *estimation error* $\left\|\hat{\beta} - \beta^*\right\|_2$. In the last lecture we derived bounds on the *prediction error* $\left\|X(\hat{\beta} - \beta^*)\right\|$, which in general requires weaker assumptions on the covariates/features X.

[1] This optimal bound *cannot* be achieved by Lasso but rather requires a different approach such as SLOPE, which uses *sorted-ℓ_1*-norm regularization.)

[2] There is a variant of Lasso called *Square-Root Lasso:*

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{2n}\|X\beta - y\|_2 + \lambda\|\beta\|_1,$$

which does not require estimation σ^2 but is a bit harder to solve than Lasso (SOCP vs QP).

3 Proof of Theorem 2

Let $T := \text{support}(\beta^*) = \{j : \beta^*_j \neq 0\}$ and $\Delta := \hat{\beta} - \beta^*$ be the error vector. The proof follows from the following four claims:

1. Basic inequality: $\frac{1}{2n} \|X\Delta\|_2^2 \leq \frac{1}{n} \left\|X^\top e\right\|_\infty \|\Delta\|_1 + \lambda \left(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1 \right)$.
2. Cone inequality: $\|\Delta_{T^c}\|_1 \leq 3 \|\Delta_T\|_1$.
3. Noise bound: With probability at least$1 - cd^{-2}$,
$$\frac{1}{n} \left\|X^\top e\right\|_\infty \leq C\sigma \sqrt{\frac{\log d}{n}} \leq \frac{\lambda}{2}.$$
4. Restricted Strong Convexity (RSC): With probability at least $1 - 2e^{-cn}$,
$$\frac{1}{n} \|Xu\|_2^2 \geq \frac{1}{2} \|u\|_2^2 , \quad \forall u \in \mathbb{R}^d : \|u_{T^c}\|_1 \leq 3 \|u_T\|_1 . \tag{5}$$

We prove these results later.

Assuming Claims 1-4 hold, the theorem follows easily. We have

$$\begin{aligned}
\frac{1}{4} \|\Delta\|_2^2 &\leq \frac{1}{2n} \|X\Delta\|_2^2 && \text{RSC applied to } \Delta \text{ satisfying cone inequality} \\
&\leq \frac{1}{n} \left\|X^\top e\right\|_\infty \|\Delta\|_1 + \lambda \left(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1 \right) && \text{basic inequality} \\
&\lesssim \lambda \|\Delta\|_1 + \lambda \left(\|\beta^*\|_1 - \|\beta^* + \Delta\|_1 \right) && \text{noise bound} \\
&\leq 2\lambda \|\Delta\|_1 . && \text{triangle inequality}
\end{aligned}$$

We bound the RHS by observing that

$$\begin{aligned}
\|\Delta\|_1 &= \|\Delta_T\|_1 + \|\Delta_{T^c}\|_1 && \text{decomposability of the } \ell_1 \text{ norm} \\
&\leq 4 \|\Delta_T\|_1 && \text{cone inequality} \\
&\leq 4\sqrt{k} \|\Delta_T\|_2 && |T| \leq k \\
&\leq 4\sqrt{k} \|\Delta\|_2 .
\end{aligned}$$

Combining the last two equations, we obtain $\|\Delta\|_2^2 \lesssim \lambda\sqrt{k} \|\Delta\|_2$, which implies the desired error bound.

3.1 Proof of Claim 1: Basic inequality

Basic inequality means that the optimal solution $\hat{\beta}$ has an objective value at least as good as that of β^*, which is a feasible solution. Analysis of many optimization-based algorithms involve such an inequality. We have encountered it before in non-parameteric regression.

Specifically to Lasso, we have

$$\frac{1}{2n} \left\|X\hat{\beta} - y\right\|_2^2 + \lambda \left\|\hat{\beta}\right\|_1 \leq \frac{1}{2n} \|X\beta^* - y\|_2^2 + \lambda \|\beta^*\| .$$

Using $\hat{\beta} = \beta^* + \Delta$ and $y = X\beta^* + e$, we rewrite the above inequality as

$$\frac{1}{2n} \|X(\beta^* + \Delta) - X\beta^* - e\|_2^2 + \lambda \left\|\hat{\beta}\right\|_1 \leq \frac{1}{2n} \|X\beta^* - X\beta^* - e\|_2^2 + \lambda \|\beta^*\|_1$$
$$\Downarrow$$
$$\frac{1}{2n} \|X\Delta - e\|_2^2 + \lambda \left\|\hat{\beta}\right\|_1 \leq \frac{1}{2n} \|e\|_2^2 + \lambda \|\beta^*\|_1 .$$

Expanding the ℓ_2 norm on the LHS

$$\|X\Delta - e\|_2^2 = (X\Delta - e)^\top (X\Delta - e) = \|X\Delta\|_2^2 + \|e\|_2^2 - 2e^\top X\Delta,$$

we obtain

$$\begin{aligned}\frac{1}{2n} \|X\Delta\|_2^2 &\leq \frac{1}{n} e^\top X\Delta + \lambda \left(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1\right) \\ &\leq \frac{1}{n} \left\|X^\top e\right\|_\infty \|\Delta\|_1 + \lambda \left(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1\right).\end{aligned}$$

3.2 Proof of Claim 2: Cone inequality

Lower bounding the last LHS by 0, we obtain

$$\begin{aligned}0 &\leq \frac{1}{n} \left\|X^\top e\right\|_\infty \|\Delta\|_1 + \lambda \left(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1\right) \\ &\leq \frac{1}{2}\lambda \|\Delta\|_1 + \lambda \left(\|\beta^*\|_1 - \|\beta^* + \Delta\|_1\right) && \text{noise bound (proved later)} \\ &\leq \frac{1}{2}\lambda \left(\|\Delta_T\|_1 + \|\Delta_{T^c}\|_1\right) + \lambda \left(\|\beta^*\|_1 - \|(\beta^* + \Delta)_T\|_1 - \|(\beta^* + \Delta)_{T^c}\|_1\right) && \text{decomposability of the } \ell_1 \text{ norm} \\ &\leq \frac{1}{2}\lambda \left(\|\Delta_T\|_1 + \|\Delta_{T^c}\|_1\right) + \lambda \left(\|\Delta_T\|_1 - \|\Delta_{T^c}\|_1\right) && \text{triangle inequlaity and } \beta^*_{T^c} = 0 \\ &= \frac{3}{2}\lambda \|\Delta_T\|_1 - \frac{1}{2} \|\Delta_{T^c}\|_1 .\end{aligned}$$

Rearranging gives the cone inequality $\|\Delta_{T^c}\|_1 \leq 3 \|\Delta_T\|_1$.

3.3 Proof of Claim 3: Noise bound

Let X_j denote the j-th column of X. We want to upper bound

$$\frac{1}{n} \left\|X^\top e\right\|_\infty = \frac{1}{n} \max_{j=1,\ldots,p} \left|X_j^\top e\right|.$$

For each j we can bound $X_j^\top e$ using concentration inequalities, and then use the union bound. (Left as exercise. Hint: condition on X_j.) Doing so gives

$$\frac{1}{n} \max_{j=1,\ldots,p} \left|X_j^\top e\right| \leq C\sigma \sqrt{\frac{\log d}{n}}.$$

By assumption, $\frac{\lambda}{2}$ is larger than the last RHS.

3.4 Proof of Claim 4: Restricted Strong Convexity

This is the most technical part of the proof. The proof is broken into two steps: we first show that X satisfies a property called *Restricted Isometry Property* (RIP), and then prove that RIP implies RSC.

3.4.1 Step 1: Gaussian design satisfies RIP

We show that X is an approximate isometric mapping when restricted to sparse vectors.

Lemma 4 (RIP). *If $n \geq Ck \log\left(\frac{d}{k}\right)$ for a sufficiently large positive constant C, then the matrix X satisfies*

$$\left(1 - \frac{1}{8}\right) \|v\|_2^2 \leq \frac{1}{n} \|Xv\|_2^2 \leq \left(1 + \frac{1}{8}\right) \|v\|_2^2, \quad \forall v \in \mathbb{R}^d : \|v\|_0 \leq 32k$$

with probability at least $1 - 2e^{-cn}$.

Proof. Let $S \subset \{1, 2, \ldots d\}$ be a *fixed* subset of column indices with $|S| \leq m := 32k$. Let $X_S \in \mathbb{R}^{n \times m}$ be the submatrix of X with column indices in S. Then *for all* vectors $v \in \mathbb{R}^d$ supported on S, we have

$$\begin{aligned}
\left| \frac{1}{n} \|Xv\|_2^2 - \|v\|_2^2 \right| &= \left| \frac{1}{n} \|X_S v\|_2^2 - \|v\|_2^2 \right| \\
&= \left| v_S^\top \left(\frac{1}{n} X_S^\top X_S \right) v_S - v_S^\top v_S \right| \\
&= \left| v_S^\top \left(\frac{1}{n} X_S^\top X_S - I_m \right) v_S \right| \\
&\leq \|v\|_2^2 \cdot \left\| \frac{1}{n} X_S^\top X_S - I_m \right\|_{\text{op}}.
\end{aligned}$$

We can bound the RHS using previous results on (sub-)Gaussian random matrices. In particular, we use Theorem 1 in Lecture 8, but take $t = \frac{1}{8}$ in the proof. Doing so shows that with probability at least $1 - 2e^{-c_0 n}$,

$$\left\| \frac{1}{n} X_S^\top X_S - I_m \right\|_{\text{op}} \leq \frac{1}{8}.$$

Combining the last two equations gives

$$\left| \frac{1}{n} \|Xv\|_2^2 - \|v\|_2^2 \right| \leq \frac{1}{8} \cdot \|v\|_2^2 \qquad \forall v \in \mathbb{R}^d : \text{support}(v) = S.$$

Taking a union bound over $\binom{d}{m}$ possible subsets S with $|S| = m$, we find that

$$\left| \frac{1}{n} \|Xv\|_2^2 - \|v\|_2^2 \right| \leq \frac{1}{8} \cdot \|v\|_2^2 \qquad \forall v \in \mathbb{R}^d : \|v\|_0 \leq m$$

with probability at least

$$\begin{aligned}
1 - 2e^{-c_0 n} \cdot \binom{d}{m} &\geq 1 - 2e^{-c_0 n + m \log(ed/m)} && \binom{d}{m} \leq \left(\frac{ed}{m}\right)^m = e^{m \log(ed/m)} \\
&\geq 1 - 2e^{-cn}. && n \gtrsim k \log(d/k) \asymp m \log(ed/m) \text{ by assumption}
\end{aligned}$$

□

3.4.2 Step 2: RIP implies RSC

Lemma 4 shows that X satisfies RIP with high probability. We next show that this implies RSC. This part of the proof is completely deterministic and geometrical. In particular, we make use of the following geometric result:

Lemma 5 (Convex hull of sparse vectors = ℓ_1 ball)**.** *We have*

$$\underbrace{\left\{x \in \mathbb{R}^p : \|x\|_1 \leq 4\sqrt{k}, \|x\|_2 \leq 1\right\}}_{A} \subseteq \operatorname{conv}\left(\underbrace{\{x \in \mathbb{R}^p : \|x\|_0 \leq 16k, \|x\|_2 \leq 2\}}_{B} \right).$$

Remark 6. It is easy to show the reverse inclusion $B \subseteq 2A$.

Proof. Consider an arbitrary $x \in A$. We want to write x as

$$x = \sum_i \alpha_i v^i$$

for some $v^i \in B$, $\alpha_i \geq 1$ and $\sum_i \alpha_i = 1$.

Partition the support of x into subsets $T_1, T_2, \ldots$, so that T_1 indexes the largest (in absolute value) $16k$ elements of x, T_2 the next $16k$ largest elements, and so on. For $i \geq 2$, each coordinate of x_{T_i} is smaller in magnitude than the average of the coordinates of $x_{T_{i-1}}$, so

$$\begin{aligned}\|x_{T_i}\|_2 \leq \sqrt{16k} \cdot \max_{j \in T_i} |x_j| &\leq \sqrt{16k} \cdot \min_{j \in T_{i-1}} |x_j| \\ &\leq \sqrt{16k} \cdot \frac{1}{16k} \sum_{j \in T_{i-1}} |x_j| = \frac{1}{\sqrt{16k}} \left\|x_{T_{i-1}}\right\|_1 .\end{aligned}$$

Summing up, we get

$$\begin{aligned}\gamma := \sum_{i \geq 1} \|x_{T_i}\|_2 &= \|x_{T_1}\|_2 + \sum_{i \geq 2} \|x_{T_i}\|_2 \\ &\leq \|x_{T_1}\|_2 + \frac{1}{\sqrt{16k}} \sum_{i \geq 2} \left\|x_{T_{i-1}}\right\|_1 \\ &\leq 1 + \frac{1}{\sqrt{16k}} \|x\|_1 \\ &\leq 1 + \frac{1}{\sqrt{16k}} \cdot 4\sqrt{k} = 2.\end{aligned}$$

Therefore, we have the desired decomposition

$$x = \sum_{i=1}^m x_{T_i} = \underbrace{\frac{\|x_{T_i}\|_2}{\gamma}}_{\alpha_i} \times \underbrace{\gamma \frac{x_{T_i}}{\|x_{T_i}\|_2}}_{v^i},$$

where $\sum_i \alpha_i = 1$ by definition, $\left\|v^i\right\|_2 = \gamma \leq 2$ and thus $v^i \in B$. Since $x \in A$ is arbitrary, the lemma follows. □

We return to our goal of establishing RSC:

$$\frac{1}{n}\left\|Xu\right\|_2^2 \geq \frac{1}{2}\left\|u\right\|_2^2, \qquad \text{when } \left\|u_{T^c}\right\|_1 \leq 3\left\|u_T\right\|_1.$$

With out loss of generality, we may assume that $\|u\|_2 = 1$. Note that the cone condition means $\|u\|_1 \leq 4\|u_T\| \leq 4\sqrt{k}\|u_T\|_2 \leq 4\sqrt{k}$. So $u \in A$. Using Lemma 5, we can write such a u as

$$u = \sum_i \alpha_i v^i$$

with $v^i \in B$, $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$. It follows that

$$\begin{aligned}\frac{1}{n}\left\|Xu\right\|_2^2 &= \frac{1}{n}u^\top X^\top X u \\ &= \frac{1}{n}\left(\sum_i \alpha_i v^i\right)^\top X^\top X\left(\sum_i \alpha_i v^i\right) \\ &= \sum_{i,j}\alpha_i\alpha_j\left(\frac{1}{n}(v^i)^T X^\top X v^j\right).\end{aligned}$$

We focus on each term in the parenthesis: one has

$$\begin{aligned}\frac{1}{n}\left\langle Xv^i, Xv^j\right\rangle &= \frac{1}{n}\cdot\frac{1}{4}\left(\left\|X(v^i+v^j)\right\|_2^2 - \left\|X(v^i-v^j)\right\|_2^2\right) && \text{"polarization identity"} \\ &\geq \frac{1}{4}\left(\frac{7}{8}\left\|v^i+v^j\right\|^2 - \frac{9}{8}\left\|v^i-v^j\right\|^2\right) && \left\|v^i \pm v^j\right\|_0 \leq 32k;\ \text{RIP} \\ &= \frac{1}{4}\left(4\left\langle v^i, v^j\right\rangle - \frac{1}{4}(\left\|v^i\right\|_2^2 + \left\|v^j\right\|_2^2)\right) \\ &\geq \left\langle v^i, v^j\right\rangle - \frac{1}{2}. && \left\|v^i\right\|_2^2 \leq 4\end{aligned}$$

It follows that

$$\begin{aligned}\frac{1}{n}\left\|Xu\right\|_2^2 &\geq \sum_{i.j}\alpha_i\alpha_j\left(\left\langle v^i, v^j\right\rangle - \frac{1}{2}\right) \\ &= \left\langle \sum_i \alpha_i v^i, \sum_j \alpha_j v^j\right\rangle - \frac{1}{2}\left(\sum_i \alpha_i\right)\left(\sum_j \alpha_j\right) \\ &= \left\|u\right\|_2^2 - \frac{1}{2} \\ &= \frac{1}{2}\left\|u\right\|_2^2 && \|u\| = 1 \text{ by assumption}\end{aligned}$$

As u is arbitrary in the cone, we conclude that RSC holds.

4 Generalization to Matrices

The sparse regression problem has a natural generalization to the matrix setting. Let $\langle A, B\rangle = \operatorname{trace}\left(A^\top B\right)$ denote the trace inner product between two matrices. Consider the model

$$y_i = \left\langle X^i, \Theta^*\right\rangle + e_i, \qquad i = 1, \ldots, n,$$

where we assume

- Low-rank: $\Theta^* \in \mathbb{R}^{d\times d}$ has rank-r;
- Gaussian design: the sensing/covariate matrices $X^i \in \mathbb{R}^{d\times d}$ satisfies $X^i_{ab} \sim N(0,1)$, independently across $i \in [n]$ and $(a,b) \in [d] \times [d]$;
- Gaussian noise: the noise e_i satisfies $e_i \sim N(0,\sigma^2)$, independently across i;
- Independence: $\{X_i\}$ and $\{e_i\}$ are independent.

This problem is sometimes called *matrix sensing*, as it is the matrix counterpart of the *compressed sensing* problem.

To recovery Θ^*, we consider the following nuclear norm regularized least-squares problem

$$\widehat{\Theta} \in \arg\min_{\Theta\in\mathbb{R}^{d\times d}} \frac{1}{2n}\sum_{i=1}^{n}\left(\left\langle X^i, \Theta\right\rangle - y_i\right)^2 + \lambda\left\|\Theta\right\|_*, \tag{6}$$

where $\|\Theta\|_*$ is the nuclear norm of Θ, defined as the sum of the singular values of Θ. The nuclear norm is a convex approximation to the (non-convex) rank function, which equals the number of non-zero singular values. The program (6) is sometiems called *Matrix Lasso*, for obvious reasons. It can be written as an SDP and is hence a convex program.

Much of the analysis for Lasso can be generalized to the above setting. Let us first introduce a useful piece of notation. Define the linear opeartor $\mathcal{X} : \mathbb{R}^{d\times d} \to \mathbb{R}^n$ with elements $[\mathcal{X}(\Theta)]_i = \left\langle X^i, \Theta\right\rangle$. Thent the matrix sensing model and Matrix Lasso can be written compactly as

$$y = \mathcal{X}(\Theta^*) + e$$

and

$$\widehat{\Theta} \in \arg\min_{\Theta\in\mathbb{R}^{d\times d}} \frac{1}{2n}\left\|\mathcal{X}(\Theta) - y\right\|_2^2 + \lambda\left\|\Theta\right\|_*.$$

With the above notation, one can show that a form of RIP holds: if $n \gtrsim rd\log d$, then with high probability,

$$\left(1-\frac{1}{8}\right)\|D\|_F^2 \le \frac{1}{n}\|\mathcal{X}(D)\|_2^2 \le \left(1+\frac{1}{8}\right)\|D\|_F^2, \qquad \forall D \in \mathbb{R}^{d\times d} : \operatorname{rank}(D) \le r.$$

We also have a form of RSC:

$$\frac{1}{n}\|\mathcal{X}(D)\|_2^2 \ge \frac{1}{4}\|D\|_F^2, \qquad \forall D \in \mathbb{R}^{d\times d} : \|D\|_* \le 10\sqrt{r}\,\|D\|_F.$$

In homework, we will work through some of these results.