Stats 300b: Theory of Statistics	**Winter 2019**

Lecture 16 – February 28

Lecturer: John Duchi — *Scribe: Zhaonan Qu*

Warning: *these notes may contain factual errors*

Outline:

- Basis Pursuit LP
- Incoherent matrices
- Concentration inequalities for incoherent matrices
- LASSO and High-dimensional regression
  - Basic inequalities
  - Restricted growth

Reading: HDP 2-3

Recap: Recall the setting of the problem to recover a sparse parameter θ^*. We observe $Y = X\theta^*$, as well as design matrix $X \in \mathbb{R}^{n\times d}$, and we want to solve the basis pursuit LP problem

$$\begin{aligned} &\min \|\theta\|_1 \\ &\text{s.t. } Y = X\theta \end{aligned}$$

and to solve this we introduced the restricted null space property for X.

For a set $S \subset [d]$, $v \in \mathbb{R}^d$, let $v_S = [v_j]_{j\in S}$. X satisfies **restricted null space property** if

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d \text{ s.t. } \|\Delta_{S^C}\|_1 \le \|\Delta_S\|_1\}$$

satisfies

$$\text{null}(X) \cap \mathbb{C}(S) = \{0\}$$

Theorem 1. *If X satisfies restricted nullspace property with respect to $S = supp(\theta^*)$, then θ^* uniquely solves the basis pursuit LP problem.*

1 Incoherent Matrices

With the above result demonstrating the usefulness of the restricted nullspace property, the next question is then how we may obtain matrices with the restricted nullspace property. To do this, we use incoherent matrices and concentration inequalities.

Definition 1.1. *Let* $X = \begin{bmatrix} | & \dots & | \\ x_1 & \dots & x_d \\ | & \dots & | \end{bmatrix} \in \mathbb{R}^{n\times d}$. *The pairwise incoherence of X is defined as*

$$\delta_{pw}(X) := \|\frac{1}{n}X^TX - I_{d\times d}\|_\infty = \max_{i,j} |\frac{1}{n}\langle X_i, X_j\rangle - \mathbf{1}(i=j)|$$

Note that as $n \ll d$, $X^T X$ has a large null-space, so the condition number of $X^T X$ is ∞. However, what pairwise incoherence tries to capture is that in some restricted subspaces, $X^T X$ is "well-conditioned". We showed the following result in homework.

Proposition 2. *If X has incoherence $\delta_{pw}(X) < \frac{1}{2k}$, then X satisfies restricted nullspce property for any set S with $|S| \leq k$.*

The next step in the analysis is then to construct incoherent matrices. We do this by showing that random matrices with sub-Gaussian entries are incoherent with high probability.

Definition 1.2. *Let $\psi_q(t) = \exp(|t|^q) - 1$ with $q \in [1, 2]$. The Orlicz norm over random variable X is defined as*

$$\|X\|_{\psi_q} := \inf\{t \in \mathbb{R}_+,\ s.t.\ \mathbb{E}[\psi_q(\frac{X}{t})] \leq 1\}$$

In homework, we showed that

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-\frac{t^q}{\|X\|_{\psi_q}})$$

which subsumes two special classes of random variables we are particularly interested in: when $q = 1$, $\|X\|_{\psi_1} < \infty$ is equivalent to X being sub-exponential, and when $q = 2$, $\|X\|_{\psi_2} < \infty$ is equivalent to X being sub-Gaussian.

Proposition 3. *Let $\|X\|_{\psi_1} < \infty$ and $\mathbb{E}X = 0$. Then*

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \frac{2\lambda^2 \|X\|_{\psi_1}^2}{(1 - \lambda\|X\|_{\psi_1})_+}$$

and if furthermore $|\lambda| < \frac{1}{2\|X\|_{\psi_1}}$,

$$\mathbb{E}[e^{\lambda X}] \leq \exp(4\lambda^2 \|X\|_{\psi_1}^2)$$

Proof Note that by integration by parts (assuming $|X|$ has density which decays faster than t^{k-1})

$$\begin{aligned}
\mathbb{E}|X|^k &= k\int_0^\infty t^{k-1}\mathbb{P}(|X| \geq t)dt \\
&\leq 2k\int_0^\infty t^{k-1}\exp(-\frac{t}{\|X\|_{\psi_1}})dt && \text{(by definition of } \|X\|_{\psi_1}) \\
&= 2k\|X\|_{\psi_1}^k \int_0^\infty u^{k-1}e^{-u}du && (u = \frac{t}{\|X\|_{\psi_1}}) \\
&= 2\|X\|_{\psi_1}^k k!
\end{aligned}$$

On the other hand, noting that $\mathbb{E}X = 0$, we have the expansion

$$\begin{aligned}
\mathbb{E}(e^{\lambda X}) &= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!} \\
&\leq 1 + 2\sum_{k=2}^{\infty} \lambda^k \|X\|_{\psi_1}^k \\
&= 1 + 2\lambda^2 \|X\|_{\psi_1}^2 \cdot \sum_{k=0}^{\infty} \lambda^k \|X\|_{\psi_1}^k \\
&= 1 + \frac{2\lambda^2 \|X\|_{\psi_1}^2}{(1-\lambda\|X\|_{\psi_1})_+}
\end{aligned}$$

Now if $|\lambda| < \frac{1}{2\|X\|_{\psi_1}}$, $\frac{2}{(1-\lambda\|X\|_{\psi_1})_+} \leq 4$, so

$$\begin{aligned}
\exp(4\lambda^2 \|X\|_{\psi_1}^2) &\geq 1 + 4\lambda^2 \|X\|_{\psi_1}^2 \\
&\geq 1 + \frac{2\lambda^2 \|X\|_{\psi_1}^2}{(1-\lambda\|X\|_{\psi_1})_+}
\end{aligned}$$

□

So we have shown that random variables with bounded Orlicz 1 norm are $\|X\|_{\psi_1}^2$-sub-Gaussian when λ is small, and sub-exponential when λ is large. Next we show a Bernstein-type tail bound for sums of variables with bounded Orlicz 1 norms that also makes this transition between sub-Gaussian and sub-exponential behavior explicit.

Proposition 4. *Let X_i be independent, $\mathbb{E}X_i = 0$, $a_i \in \mathbb{R}$. Then*

$$\mathbb{P}(\sum_i a_i X_i \geq t) \leq \exp(-C \min\{\frac{t^2}{\sum_i a_i^2 \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i |a_i| \|X_i\|_{\psi_1}}\})$$

Proof First note that if $\lambda \leq \min_i \frac{1}{2|a_i|\|X_i\|_{\psi_1}}$, by the previous proposition, $\mathbb{E}[e^{\lambda a_i X_i}] \leq \exp(4\lambda^2 a_i^2 \|X_i\|_{\psi_1}^2)$ for all i, so that

$$\mathbb{E}[\exp(\lambda a^T X)] \leq \exp(4\lambda^2 \sum_i a_i^2 \|X_i\|_{\psi_1}^2)$$

Now apply Chernoff bound to conclude

$$\begin{aligned}
\mathbb{P}(a^T X \geq t) &\leq \mathbb{E}[\exp(\lambda a^T X - \lambda t)] \\
&\leq \exp(4\lambda^2 \sum_i a_i^2 \|X_i\|_{\psi_1}^2 - \lambda t)
\end{aligned}$$

and choosing $\lambda = \min\{\frac{t}{8\sum_i a_i^2 \|X_i\|_{\psi_i}^2}, \frac{1}{2\max_i |a_i| \|X_i\|_{\psi_1}}\}$ gives the claimed bound. □

Corollary 5. *Let $\sigma = \max_i \|X_i\|_{\psi_1}$ and assume $\mathbb{E}X_i = 0$. Then*

$$\mathbb{P}(|a^T X| \geq t) \leq 2\exp(-C \min\{\frac{t^2}{\|a\|^2\sigma^2}, \frac{t}{\|a\|_\infty \sigma}\}$$

The above result enables us to bound the tail of the diagonal terms of $X^T X$.

Corollary 6. *[Sum] Let X_i be σ^2-sub Gaussian, and $\mathbb{E}X_i^2 = 1$. Then*

$$\mathbb{P}(|\frac{1}{n}\sum_i (X_i^2 - 1)| \geq t) \leq 2\exp(-Cn\min\{\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\})$$

Proof Note that $\|X_i^2\|_{\psi_1} = \|X_i\|_{\psi_2}^2$, since $\mathbb{E}\exp(\frac{X_i^2}{\|X_i\|_{\psi_2}^2}) = 2$. This implies $\|X_i^2\|_{\psi_1} \leq \sigma^2$. Letting $a_i = \frac{1}{n}$ in the previous corollary gives the result. $\square$

Now we provide another result that controls the off-diagonal entries of $X^T X$.

Proposition 7. *[Product] Let $X_1, X_2 \in \mathbb{R}^n$ be independent vectors with σ^2-sub-Gaussian entries. Then $\|\langle X_1, X_2\rangle\|_{\psi_1} \leq C\sigma^2\sqrt{n}$.*

Proof We compute moment generating functions of $\langle X_1, X_2\rangle$.

$$\begin{aligned}
\mathbb{E}[\exp(\lambda\langle X_1, X_2\rangle)] &\leq \mathbb{E}[\exp(\frac{\lambda^2\sigma^2}{2}\|X_2\|_2^2)] && (\text{by } \mathbb{E} \text{ over } X_1) \\
&= \mathbb{E}[\exp(\lambda\sigma\langle Z, X_2\rangle)] && (Z \sim \mathcal{N}(0, I)) \\
&\leq \mathbb{E}[\exp(\frac{\lambda^2\sigma^4}{2}\|Z\|_2^2)] && (\text{by } \mathbb{E} \text{ over } X_2) \\
&= (\frac{1}{(1-\lambda^2\sigma^4)_+})^{n/2} \\
&= \exp(-\frac{n}{2}\log(1-\lambda^2\sigma^4)_+)
\end{aligned}$$

Taking $\lambda^2 = \frac{1}{2n\sigma^4}$, we get

$$\mathbb{E}[\exp(\frac{\langle X_1, X_2\rangle}{\sqrt{2n\sigma^4}})] \leq \exp 2$$

and finally

$$\|\langle X_1, X_2\rangle\|_{\psi_1} \leq C\sigma^2\sqrt{n}$$

$\square$

Theorem 8. *Let $X = \begin{bmatrix} | & \dots & | \\ x_1 & \dots & x_d \\ | & \dots & | \end{bmatrix} \in \mathbb{R}^{n\times d}$ have independent O(1)-sub-Gaussian entries, and $\mathbb{E}X_{ij}^2 = 1$ (e.g. X_{ij} are iid $\mathcal{N}(0,1)$). Then*

$$\mathbb{P}(\|\frac{1}{n}X^T X - I_{d\times d}\|_\infty \geq t) \leq 2d^2\exp(-C\sqrt{n}t) + 2d\exp(-Cn\min\{t^2, t\})$$

Proof For $i \neq j$, proposition [Product] implies

$$\mathbb{P}(\frac{1}{n}|\langle X_i, X_j\rangle| \geq t) \leq 2\exp(-C\frac{nt}{\sigma^2\sqrt{n}})$$
$$\leq 2\exp(-C\sqrt{n}t)$$

For $i = j$, by proposition [Sum] we have

$$\mathbb{P}(\frac{1}{n}|\langle X_i, X_j\rangle - 1| \geq t) \leq 2\exp(-Cn\min\{t^2, t\})$$

Applying union bound gives the result. □

The theorem implies that with high probability, matrix X has small pairwise incoherence. More precisely, we have the following corollary.

Corollary 9. *With probability at least* $1-\delta$*,*

$$\delta_{pw}(X) \leq C \cdot \frac{\log d + \log\frac{1}{\delta}}{\sqrt{n}}$$

In other words, to recover k-sparse signals $Y = X\theta^*$ where $\|\theta^*\|_0 \leq k$ with high probability requires at most $n \geq Ck^2\log^2 d$, which is exponentially better than $n \geq d$.

2 LASSO (linear model in high dimensions)

Now we turn to the setting where there is noise in observations, i.e.

$$Y = X\theta^* + \varepsilon$$

In order to recover a sparse θ^* with precision, we again want to penalize the non-zeroes of θ. As X is a fat matrix, there will be lots of null directions in θ space of the loss surface, i.e. those directions of θ which does not change the loss much. In order to recover with precision, we want to penalize the null directions.

First let's consider the constrained form. Suppose we know $\|\theta^*\|_1 = b$. Then we can try to solve the problem

$$\min \frac{1}{2}\|X\theta - Y\|_2^2$$
$$\text{s.t. } \|\theta_1\| \leq b$$

and we want to show that the solution $\hat{\theta} = \theta^* + \Delta$ with Δ small.

Observe first that

$$\Delta \in \mathbb{C}(S) := \{\Delta \text{ s.t. } \|\Delta_{S^C}\|_1 \leq \|\Delta_S\|_1\}$$

where $S = \text{supp}(\theta^*)$. This is because

$$\begin{aligned}\|\theta^*\|_1 = \|\theta^*_S\|_1 \geq \|\hat{\theta}\|_1 &= \|\theta^* + \Delta\|_1 \\ &= \|\theta^*_S + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\ &\geq \|\theta^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^C}\|_1\end{aligned}$$

which implies $\|\Delta_{S^C}\|_1 \leq \|\Delta_S\|_1$.

Also, we have the basic inequality

$$\Delta^T X^T X \Delta \leq 2\Delta^T X \varepsilon$$

To see this, note

$$\frac{1}{2}\|X\Delta - \varepsilon\|_2^2 = \frac{1}{2}\|X\hat{\theta} - Y\|_2^2 \leq \frac{1}{2}\|X\theta^* - Y\|_2^2 = \frac{1}{2}\|\varepsilon\|_2^2$$

and expanding gives the claimed inequality.

If X is "nice" on $\mathbb{C}(S)$, i.e. if $\frac{1}{n}\Delta^T X^T X \Delta \geq \mu\|\Delta\|_2^2$ for $\Delta \in \mathbb{C}(S)$, then

$$\begin{aligned}
n\mu\|\Delta\|_2^2 &\leq 2\Delta^T X^T \varepsilon \\
&\leq 2\|\Delta\|_1 \|X^T\varepsilon\|_\infty \\
&\leq 4\|\Delta_S\|_1 \|X^T\varepsilon\|_\infty \\
&\leq 4\sqrt{k}\|\Delta_S\|_2 \|X^T\varepsilon\|_\infty
\end{aligned}$$

which implies

$$\begin{aligned}
\|\Delta\|_2 = \|\hat{\theta} - \theta^*\|_2 &\leq \frac{4\sqrt{k}\|X^T\varepsilon\|_\infty}{n\mu} \\
&\leq O(1)\sqrt{\frac{k \log d}{n}}
\end{aligned}$$

since $\|X_i^T\varepsilon\|_\infty \lesssim \sqrt{n \log d}$.