Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 37 — November 13

Instructor: Stas Minsker *Scribe: Mose Wintner*

37.1 Matrix recovery problems and sparse linear regression

$Y_j = \langle X_j, A_0\rangle + \varepsilon_j$ where ε_j is independent of X_j with mean 0. Let's consider

$$\hat{A}_\lambda = \operatorname*{argmin}_{A\in\mathbb{R}^{m_1\times m_2}} \frac{1}{n}\sum_{j=1}^{n}(Y_j - \langle X_j, A\rangle)^2 + \lambda\,||A||_*\,.$$

$$\frac{1}{n}\sum_1^n (Y_j - \langle X_j, A\rangle)^2 = \frac{1}{n}\sum Y_j^2 + \frac{1}{n}\langle A, X_j\rangle^2 - \frac{2}{n}\langle\sum_1^n Y_j X_j, A\rangle.$$

The first term doesn't depend on A so we can ignore it.

$$\mathbb{E}\frac{1}{n}\sum\langle A, X_j\rangle^2 = \mathbb{E}\langle A, X_1\rangle^2 = ||A||^2_{L_2(\Pi)}\,,$$

where Π is the distribution of X_1.

Example 1. $X_1 \sim U(\mathcal{X})$, $\mathcal{X} = \{e_i(m_1)e_j(m_2)^T\}$.

$$\mathbb{E}\langle A, X_1\rangle^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{1}{m_1m_2}A_{ij}^2 = \frac{1}{m_1m_2}\,||A||_F^2\,.$$

Example 2. $(X_1)_{ij} \sim \mathcal{N}(0,1)$ independent.

$$\mathbb{E}\langle A, X_1\rangle^2 = ||A||_F^2\,.$$

We will therefore study the estimator

$$\hat{A}_\lambda = \operatorname*{argmin}_{A\in\mathbb{R}^{m_1\times m_2}}\left(||A||^2_{L_2(\Pi)} - \langle\frac{2}{n}\sum Y_jX_j, A\rangle + \lambda\,||A||_*\right).$$

Example 3. Diagonal matrices: LASSO (least absolute shrinkage and selection operator). $x_1,\dots,x_n$ nonrandom vectors, $A\in\mathbb{R}^p$. Then

$$\hat{A}_\lambda = \operatorname*{argmin}_{A\in\mathbb{R}^{m_1\times m_2}}\left(||A||^2_{L_2(\Pi)} - \langle\frac{2}{n}\sum Y_jX_j, A\rangle + \lambda\,||A||_1\right).$$

Definition 1. Let $||\cdot||'$ be the dual norm, defined via

$$||y||'_* = \sup_{||x||_*\le 1}\langle y, x\rangle.$$

Lemma 1. The dual of the nuclear norm is the operator (spectral) norm. The operator norm is $||A|| = \max_{j=1,\ldots,rank(A)} \sigma_j$, where σ_j are the singular values of A.

Proof: Let $A = U\Sigma V^T$ be the SVD of A and let $Q = UV^T$. Then

$$\langle Q, A\rangle = \mathrm{tr}(VU^TU\Sigma V^T) = \mathrm{tr}(V\Sigma V^T) = \mathrm{tr}(\Sigma) = ||A||_* .$$

Then

$$\begin{aligned}
\sup_{||Q||\le 1} \langle Q, A\rangle &= \sup_{||Q||\le 1} \mathrm{tr}(Q^TU\Sigma V^T) \\
&= \sup_{||Q||\le 1} \mathrm{tr}(V^TQ^TU\Sigma) \\
&= \sup_{||Q||\le 1} \mathrm{tr}(V^T\Sigma Q^TU) \\
&= \sup_{||Q||\le 1} \sum_j \sigma_j(A)(U^TQV)_{jj} \\
&= \sup_{||Q||\le 1} \sum_j \sigma_j(A)\langle Qv_j, u_j\rangle \\
&\le \sup_{||Q||\le 1} \sum_j \sigma_j(A)\, ||Q|| \\
&\le \sum_j \sigma_j(A) = ||A||_* .
\end{aligned}$$

37.1.1 Subdifferential of the nuclear norm

Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex. The subdifferential is

$$\partial f(x) = \{y \in \mathbb{R}^p,\ f(x) \ge f(z) + \langle y, z - x\rangle \text{ for all } z \in \mathbb{R}^p\}.$$