Random Matrix Theory

These notes are based on the following sources:

1. *Introduction to the Non-asymptotic Analysis of Random Matrices* by Roman Vershynin.
2. *Error Bounds for Random Matrix Approximation Schemes* by A. Gittens and J. Tropp.

Another excellent source is:
Topics in Random Matrix Theory by Terence Tao.

These can be found online.

Let $X_1, \ldots, X_n \sim P$ where $X_i \in \mathbb{R}^d$. Without loss of generality, assume that $\mu = 0$. Let $\Sigma = \mathrm{Var}(X_i)$ and

$$\widehat{\Sigma} = \frac{1}{n}\sum_i X_i X_i^T = \frac{1}{n} A^T A$$

where A is the $n \times d$ data matrix. Our goal is to show that $\widehat{\Sigma}$ is close to Σ. This implies that the eigenvalues and eigenvectors of of $\widehat{\Sigma}$ are close to eigenvalues and eigenvectors of Σ. Unless we say otherwise, we will assume that $d < n$.

1 Review of Singular Values

Let A be an $n \times d$ matrix with $n \geq d$. Recall that the *singular values* are the square roots of the eigenvalues of $A^T A$. These are denoted by $s_1(A) \geq s_2(A) \geq \cdots \geq s_d(A) \geq 0$. We also let $s_{\max}(A) = s_1(A)$ and $s_{\min}(A) = s_d(A)$. We will also write the singular values as $s_1 \geq \cdots \geq s_d$. The singular value decomposition of A is

$$A = UDV^T \tag{1}$$

where D is $n \times d$ and diagonal with the singular values on the diagonal, U is $n \times n$, V is $d \times d$ and $U^T U = I$ and $V^T V = I$. The columns of U are the eigenvectors of AA^T and the columns of V are the eigenvectors of $A^T A$.

The *spectral norm* (or operator norm) of A is

$$\|A\| = \|A\|_{2\to 2} = \sup_{\|x\|_2=1} \|Ax\|_2 = s_1(A) = \max_{u\in S_d, v\in S_n} \langle Au, v\rangle. \tag{2}$$

If A is a symmetric $d \times d$ matrix, we have

$$||A||^2 = s_1^2(A) = \lambda_1(A^T A) = \max_{||u||=1} u^T A^T A u = \max_{||u||=1} ||Au||^2$$

and so

$$||A|| = \max_{||u||=1} ||Au||.$$

Throughout these notes, $\|x\|$ denotes the L_2 norm if x is a vector and $\|A\|$ denotes the spectral norm if A is a matrix. For any x,

$$s_d(A)\|x\| \leq \|Ax\| \leq s_1(A)\|x\|. \tag{3}$$

(Note: these are all L_2 norms.) Hence, if $s_1(A) \approx s_d(A)$ then A is almost an isometry. Matrices that are close to being isometries are important in many algorithms such as compressed sensing. Thus, it is of interest to bound the singular values. Specifically, we want to show that

$$\sqrt{n} - C\sqrt{d} \leq s_d \leq s_1 \leq \sqrt{n} + C\sqrt{d} \tag{4}$$

or, equivalently,

$$1 - C\sqrt{\frac{d}{n}} \leq s_d(n^{-1/2}A) \leq s_1(n^{-1/2}A) \leq 1 + C\sqrt{\frac{d}{n}}.$$

The following lemma is useful.

Lemma 1 *Let B be a symmetric $d \times d$ matrix. If*

$$\left\|B^T B - I\right\| \leq \max\{\delta, \delta^2\} \tag{5}$$

then

$$1 - \delta \leq s_p(B) \leq s_1(B) \leq 1 + \delta. \tag{6}$$

Conversely, if (6) holds then $\left\|B^T B - I\right\| \leq 3\max\{\delta, \delta^2\}$.

From the lemma, we see that (4) implies that $A^T A/n$ is approximately the identity matrix. Specifically, if (4) holds, then, with $B = A/\sqrt{n}$,

$$1 - \delta = 1 - C\sqrt{\frac{d}{n}} \leq s_p(B) \leq s_1(B) \leq 1 + C\sqrt{\frac{d}{n}} = 1 + \delta$$

and the lemma implies that

$$\left\|\frac{1}{n}A^T A - I\right\| \leq \epsilon \quad \text{where } \epsilon = \max\{\delta, \delta^2\} \quad \text{and } \delta = O(\sqrt{d/n}). \tag{7}$$

To see why this is useful, suppose that each row X_i of A is a d-dimension sample from a distribution with mean 0 and covariance $\Sigma = I$. Then the above implies that

$$||\widehat{\Sigma} - \Sigma|| \leq \epsilon$$

where $\widehat{\Sigma} = n^{-1}A^T A$. (If the mean is not 0, we can just subtract it off.) More generally, if X_i has covariance Σ, define $W_i = \Sigma^{-1/2}X_i$. Then W_i has covariance I. Let $\mathbb{W}$ be the matrix whose i^{th} row is W_i. Then, $A = \mathbb{W}\Sigma^{1/2}$ and so

$$||\widehat{\Sigma} - \Sigma|| = ||\Sigma^{1/2}(n^{-1}\mathbb{W}^T\mathbb{W} - I)\Sigma^{1/2}|| \leq ||\Sigma||\ ||n^{-1}\mathbb{W}^T\mathbb{W} - I|| \leq \epsilon||\Sigma||.$$

2 Gaussian Matrices

Suppose first that A is an $n \times d$ matrix of independent, standard Normal random variables. We will need the following results:

Lemma 2 *(The Sudakov-Fernique Inequality.) Let X_t and Y_t be Gaussian processes with the same mean. If*

$$\mathbb{E}|X_s - X_t|^2 \leq \mathbb{E}|Y_s - Y_t|^2$$

for all s, t then $\mathbb{E}\sup_t X_t \leq \mathbb{E}\sup_t Y_t$.

Lemma 3 *Let $X \sim N_d(0, I)$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be K-Lipschitz. Then for every $t \geq 0$,*

$$\mathbb{P}\Big(f(X) - \mathbb{E}[f(X)] > t\Big) \leq e^{-t^2/(2K^2)}.$$

The *Sudakov-Fernique inequality* seems quite intuitive but its proof is non-trivial. See *Random Fields and Geometry* by Adler and Taylor (2000).

Theorem 4 *We have*

$$\sqrt{n} - \sqrt{d} \leq \mathbb{E}(s_d) \leq \mathbb{E}(s_1) \leq \sqrt{n} + \sqrt{d}.$$

Furthermore,

$$\mathbb{P}\Big(\sqrt{n} - \sqrt{d} - t \leq s_d \leq s_1 \leq \sqrt{n} + \sqrt{d} + t\Big) \geq 1 - 2e^{-t^2/2}.$$

Proof. Let $S_k = \{x = (x_1, \ldots, x_k) : \ \|x\| = 1\}$. Then

$$s_1 = \max_{u \in S_d, v \in S_n} \langle Au, v\rangle.$$

Let $X_{u,v} = \langle Au, v\rangle$. Note that $X_{u,v}$ is a Gaussian process indexed by u and v. Let $Y_{u,v} = \langle Z, u\rangle + \langle W, v\rangle$ where $Z \in \mathbb{R}^p$ and $W \in \mathbb{R}^N$ are standard normal vectors. Some algebra shows that,

$$\mathbb{E}|X_{u,v} - X_{u',v'}|^2 \leq \mathbb{E}|Y_{u,v} - Y_{u',v'}|^2$$

for any u, v, u', v'. It follows from the Sudakov-Fernique inequality that

$$\mathbb{E}\sup_{u,v} X_{u,v} \leq \mathbb{E}\sup_{u,v} Y_{u,v}.$$

Now

$$Y_{u,v} \leq \|Z\| + \|W\|$$

and so

$$\mathbb{E}\sup_{u,v} Y_{u,v} \leq \sqrt{n} + \sqrt{d}.$$

Thus $\mathbb{E}(s_1) \leq \sqrt{n} + \sqrt{d}$. A similar proof shows that $\sqrt{n} - \sqrt{d} \leq \mathbb{E}(s_d)$. The second result follows from Lemma 3 since s_1 and s_d are 1-Lipschitz (where A is regarded as one long vector). $\square$

3 Independent Rows

Suppose now that the rows of A are independent vectors. This is more general than assuming that all the entries are independent. We will also relax the Gaussian assumption.

Recall that a mean 0 random variable X is *sub-Gaussian* if, for all $t \geq 0$,

$$\mathbb{P}(|X| > t) \leq e^{-t^2/C}$$

for some C. This is equivalent to

$$\mathbb{E}e^{tX} \leq e^{ct^2}.$$

A random vector X is sub-Gaussian if

$$\mathbb{E}e^{t^T X} \leq e^{ct^2}.$$

Theorem 5 *Let A be $n \times d$ and suppose that the rows A_i are independent, sub-Gaussian random vectors with identity covariance. Then there are constants $c, C > 0$ such that, for all $t \geq 0$,*

$$\mathbb{P}\Big(\sqrt{n} - C\sqrt{d} - t \leq s_d \leq s_1 \leq \sqrt{n} + C\sqrt{d} + t\Big) \geq 1 - 2e^{-ct^2}.$$

Also,

$$\left\| \frac{1}{n} A^T A - I \right\| \leq \epsilon$$

where $\epsilon = \max\{\delta, \delta^2\}$ and

$$\delta = C\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}.$$

Proof. From (7) it suffices to show the second statement. The covering number of S_d is bounded by

$$\left(1 + \frac{2}{a}\right)^d.$$

Let $a = 1/2$. We can thus cover S_d with a a-net $\mathcal{N}$ using 5^d points. Recall that, for any symmetric $d \times d$ matrix B we have

$$||B|| = \sup_{||u||=1} ||Bu||.$$

Every $u \in S_d$ can be written as $u = v + w$ where $v \in \mathcal{N}$ and $||w|| \leq a$. So

$$\|B\| = \sup_{u \in S_d} ||Bu|| \leq \sup_{v \in \mathcal{N}} ||Bv|| + \sup_{||w|| \leq 1/2} ||Bw||.$$

But $\sup_{||w||\leq 1/2} ||Bw|| = (1/2)||B||$. So

$$||B|| \leq \sup_{v\in\mathcal{N}} ||Bv|| + (1/2)||B||$$

which implies that

$$||B|| \leq 2 \sup_{v\in\mathcal{N}} ||Bv||.$$

Hence,

$$||\widehat{\Sigma} - \Sigma|| = \left\| \frac{1}{n} A^T A - I \right\| \leq 2 \max_{u\in\mathcal{N}} \left| \left\langle \left(\frac{1}{n} A^T A - I \right) u, u \right\rangle \right| = 2 \max_{u\in\mathcal{N}} \left| \frac{1}{n} \|Au\|^2 - 1 \right|. \quad (8)$$

Let $Z_i = \langle A_i, u\rangle$. Then $Z_1, \ldots, Z_n$ are independent, sub-Gaussian with $\mathbb{E}(Z_i^2) = 1$. Furthermore,

$$\|Au\|^2 = \sum_{i=1}^n \langle A_i, u\rangle^2 = \sum_{i=1}^n Z_i^2.$$

Since the Z_i are sub-Gaussian, we can use a Chernoff-like concentration argument to get that

$$\mathbb{P}\left(\left| \frac{1}{n} \|Au\|^2 - 1 \right| \geq \frac{\epsilon}{2} \right) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n Z_i^2 - 1 \right| \geq \frac{\epsilon}{2} \right) \leq 2 \exp\left(-c(Cd + t^2) \right).$$

From the union bound, and choosing C large enough,

$$\mathbb{P}\left(\max_{u\in\mathcal{N}} \left| \frac{1}{n} \|Au\|^2 - 1 \right| \geq \frac{\epsilon}{2} \right) \leq 5^d \; 2 \exp\left(-c(Cd + t^2) \right) \leq 2e^{-c't^2}.$$

□

4 Back to Covariance Matrices

Let X_i be the i^{th} row of A. The last theorem said that if $X_i \sim N(0, I)$ then, with high probability,

$$\left\| \frac{1}{n} A^T A - I \right\| \leq \epsilon$$

where $\epsilon = \max\{\delta, \delta^2\}$ and

$$\delta = C\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}.$$

If instead, $X_i \sim N(0, \Sigma)$, then we get that $\widehat{\Sigma} = n^{-1} A^T A$ is close to Σ instead of being close to I, as we saw earlier. Using Weyl's theorem and the Dvais-Kahan theorem, this means that the eigenvalues and eigenvectors of $\widehat{\Sigma}$ are close to those of Σ.

If $d > n$, these results fail and different techniques (and stronger assumptions) are needed.

5 Restricted Isometries and Compressed Sensing

Recall that *compressed sensing* is basically just the *lasso* except that we get to generate the design matrix. That is, we observe

$$Y = X\beta + \epsilon$$

where X is $n \times p$ with $n < p$. We want to recover β from Y. We take advantage of two facts: β is assumed to be sparse and we get to generate X randomly. Recovering a sparse β is possible if X satisfies a *restricted isometry* condition.

We say that a matrix A satisfies the restricted isometry property (of order k) if there exists $\delta_k \geq 0$ such that

$$(1 - \delta_k)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta_k)\|x\|^2 \tag{9}$$

for all $x \in \mathbb{R}^p$ for which $\|x\|_0 \leq k$. Let $\delta_k = \delta_k(A)$ be the smallest such number. It can be shown that

$$\delta_k = \max_{T:\ |T|=k} \|A_T^T A_T - I\|.$$

Lemma 6 *If $\delta_k \leq \max\{\delta, \delta^2\}$ then*

$$1 - \delta \leq s_{\min}(A_T) \leq s_{\max}(A_T) \leq 1 + \delta \quad \text{for all } |T| \leq k. \tag{10}$$

Conversely, if (10) holds then, $\delta_k \leq 3\max\{\delta, \delta^2\}$.

Theorem 7 *Let A be an $n \times p$ matrix with independent, sub-Gaussian rows. Let $X = n^{-1/2}A$. For every $0 < \delta < 1$, and $1 \leq k \leq n$, if*

$$n \succeq \frac{k \log(ep/k)}{\delta^2}$$

then

$$\mathbb{P}\Big(\delta_k(X) \leq \delta\Big) \geq 1 - 2e^{-c\delta^2 n}.$$

Proof. Fix $T \subset \{1, \ldots, p\}$ such that $|T| = k$. From the earlier results,

$$\sqrt{n} - \sqrt{k} - s \leq s_{\min}(A_T) \leq s_{\max}(A_T) \leq \sqrt{n} + \sqrt{k} + s$$

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

$$\left|\left|\frac{1}{n}A_T^T A_T - I\right|\right| = \|X_T^T X_T - I\| \leq 3\max\{\delta_0, \delta_0^2\} \quad \text{where } \delta_0 = \sqrt{\frac{k}{n}} + \frac{s}{\sqrt{n}}.$$

The number of such subsets T is $\binom{p}{k} \leq (ep/k)^k$. By the union bound,

$$\max_{|T|=k} \|X_T^T X_T - I\| \leq 3\max\{\delta_0, \delta_0^2\}$$

except on a set of probability at most

$$2\left(\frac{ep}{k}\right)e^{-cs^2} = 2\exp\left(k\log\left(\frac{ep}{k}\right) - cs^2\right).$$

Pick $\epsilon > 0$ and let

$$s = c_1\sqrt{k\log(ep/k)} + \epsilon\sqrt{p}.$$

Then $3\max\{\delta_0, \delta_0^2\} < 2\delta$ and so $\delta_k \leq 2\delta$ with probability at least $1 - 2e^{-c'\epsilon^2 n} = 1 - 2e^{-c\delta^2 n}$. □

6 Sparsification

We have concentrated on the spectral norm but other norms are also of interest. In particular, Gittens and Tropp study the norm

$$\|A\|_{\infty\to p} = \max_{u\neq 0}\frac{\|Au\|_p}{\|u\|_\infty}.$$

To see why this norm might be of interest, consider a graph $G = (V, E)$ and weights w_{jk}. Recall that a *cut* partitions the vertices $V = S \cup S^c$. The cost of a cut is the sum of weights of edges with one vertex in S and one vertex in S^c. Finding the maximum cut corresponds to finding the *cut-norm*

$$\|A\|_C = \max_{S\subset E}\left|\sum_{(j,k)\in E} w_{jk}I(j\in S)I(k\in S^c)\right|.$$

It is very difficult to work with $\|A\|_C$. However,

$$\|A\|_C \leq \|A\|_{\infty\to 1} \leq 4\|A\|_C.$$

Hence, it is useful to bound the norm $\|A\|_{\infty\to 1}$. We will state a bound from Gittens and Tropp without proof and then we'll discuss an application. The application involves finding a random sparse approximating matrix X.

Theorem 8 *(Gittens and Tropp.) Let A be an $m \times n$ matrix. Let X be a random matrix with independent entries such that $\mathbb{E}(X) = A$. Suppose that $|X_{jk}| \leq D/2$ for all j and k. Finally, let $\epsilon_1, \epsilon_2, \ldots$ be Rademacher random variables. Let $Z = A - X$ and define q by $(1/p) + (1/q) = 1$. Then:*

1. *The mean is bounded by:*

$$\mathbb{E}\|Z\|_{\infty\to p} \leq 2\mathbb{E}\left\|\sum_k \epsilon_k z_k\right\|_p + 2\max_{\|u\|_q=1}\mathbb{E}\sum_k\left|\sum_j \epsilon_j Z_{jk}u_j\right|.$$

2. *Concentration around the mean:*

$$\mathbb{P}\left(\|Z\|_{\infty\to p} > \mathbb{E}\|Z\|_{\infty\to p} + t\right) \leq \exp\left(-\frac{t^2}{4D^2nm^2}\right).$$

3. *The special case $p = 1$:*

$$\mathbb{E}\|Z\|_{\infty\to 1} \leq 2\left[\sum_k \left(\sum_j \mathsf{Var}(X_{jk})\right)^{1/2} + \sum_j \left(\sum_k \mathsf{Var}(X_{jk})\right)^{1/2}\right].$$

Let us apply the theorem to the an example. Let A be an $n \times n$ matrix with bounded, positive entries. Let X be a *sparsification* of A defined by

$$X_{jk} \sim A_{jk}B/p$$

where $B \sim \text{Bernoulli}(p)$. Then $\mathsf{Var}(X_{jk}) = A_{jk}^2/p - A_{jk}^2$. It follows that

$$\frac{2\left[\sum_k \left(\sum_j \mathsf{Var}(X_{jk})\right)^{1/2} + \sum_j \left(\sum_k \mathsf{Var}(X_{jk})\right)^{1/2}\right]}{\|A\|_{\infty\to 1}} = O\left(\sqrt{\frac{1-p}{np}}\right).$$

Set $p = (1 + n\gamma^2)^{-1}$. We get that $\mathbb{E}\|A - X\|_{\infty\to 1} < \gamma$ and the expected number of nonzero entries is $O(n/\gamma^2)$.