Further background material for Statistical Theory

Comments and corrections to r.samworth@statslab.cam.ac.uk

Matrix norms

There are many different matrix norms that it is useful to consider in Statistics. Let $A = (a_{ij})$ be an $m \times n$ real matrix.

OPERATOR NORMS: For $p \in [1, \infty]$, writing $\|\cdot\|_p$ for the ℓ_p norm of a vector, the pth *operator norm* of A is

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p.$$

Note that $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$ is the maximum absolute column sum of A, and $\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$ is the maximum absolute row sum of A. When $m = n$, we call $\|A\|_2$ the *spectral norm* of A. It is the square root of the largest eigenvalue of $A^T A$ (also called the largest *singular value* of A; see below). Observe that if P is an orthogonal $m \times m$ matrix and Q is an orthogonal $n \times n$ matrix, then $\|PAQ\|_2 = \|A\|_2$.

ENTRYWISE NORMS: For $p \in [1, \infty)$, the pth *entrywise norm* is

$$|||A|||_p = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^p \right)^{1/p}.$$

This is different from the pth operator norm. When $p = 2$, this norm is called the *Frobenius norm*, and $|||A|||_2 = \sqrt{\mathrm{tr}(A^T A)}$. We also write $|||A|||_\infty = \max_{1 \leq i \leq m, 1 \leq j \leq n} |a_{ij}|$.

SCHATTEN NORMS: For $p \in [1, \infty]$, the *Schatten p-norm* of A is the ℓ_p-norm of the vector of singular values of A. When $p = 1$, it is also called the *nuclear norm*, and is equal to $\mathrm{tr}(\sqrt{A^T A})$ (or the sum of the singular values).

Singular value decomposition

We say $\sigma \geq 0$ is a *singular value* of $A \in \mathbb{R}^{m \times n}$ if there exist $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ with $\|u\|_2 = \|v\|_2 = 1$ such that $Av = \sigma u$ and $A^T u = \sigma v$. In this case, u is called a *left-singular vector* of A and v is called a *right-singular vector* of A. A singular value decomposition (SVD) of a matrix is a generalisation of an eigendecomposition of a square matrix.

Theorem 1. *Let $A \in \mathbb{R}^{m \times n}$. Then there exist an $m \times m$ orthogonal matrix U and an $n \times n$ orthogonal matrix V such that*

$$U^T A V = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{m \times n},$$

where $r = \min(m, n)$ and $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$.

Proof. Assume for a contradiction that the result is false, and take r to be minimal such that there is an $m \times n$ matrix A with $r = \min(m, n)$ which cannot be written in this form.

Let $\sigma_1 = \|A\|_2$, so there exist $v_1 \in \mathbb{R}^n$ with $\|v_1\| = 1$ and $u_1 \in \mathbb{R}^m$ with $\|u_1\| = 1$ such that $Av_1 = \sigma_1 u_1$. Let U_1 be an $m \times m$ orthogonal matrix with first column u_1, and let V_1 be an $n \times n$ orthogonal matrix with first column v_1. Then, for some $w \in \mathbb{R}^{n-1}$ and $A_1 \in \mathbb{R}^{(m-1)\times(n-1)}$ we can write

$$U_1^T A V_1 = \begin{pmatrix} \sigma_1 & w^T \\ 0 & A_1 \end{pmatrix} =: B,$$

say. Now

$$\left\| B \begin{pmatrix} \frac{\sigma_1}{\sqrt{\sigma_1^2 + \|w\|_2^2}} \\ \frac{w}{\sqrt{\sigma_1^2 + \|w\|_2^2}} \end{pmatrix} \right\|_2 \geq \sqrt{\sigma_1^2 + \|w\|_2^2}.$$

But then $\sigma_1 = \|A\|_2 = \|U_1^T A V_1\|_2 = \|B\|_2 \geq \sqrt{\sigma_1^2 + \|w\|_2^2}$, so we must have $w = 0$. Now since $A_1 \in \mathbb{R}^{(m-1)\times(n-1)}$ and r was minimal, we can find an $(m-1) \times (m-1)$ orthogonal matrix $\tilde{U}_2$ and an $(n-1)\times(n-1)$ orthogonal matrix $\tilde{V}_2$ such that $\tilde{U}_2^T A_1 \tilde{V}_2 = \mathrm{diag}(\sigma_2, \ldots, \sigma_r)$, say with $\sigma_2 \geq \ldots \geq \sigma_r \geq 0$. Now let $U_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_2 \end{pmatrix}$ and $V_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{V}_2 \end{pmatrix}$. Then U_2 and V_2 are orthogonal matrices, and if we let $U = U_1 U_2$ and $V = V_1 V_2$ (so U and V are orthogonal), then

$$U^T A V = U_2^T U_1^T A V_1 V_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_2^T \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{V}_2 \end{pmatrix} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r).$$

Since $\sigma_1 = \|A\|_2 = \|U^T A V\|_2 = \|\mathrm{diag}(\sigma_1, \ldots, \sigma_r)\|_2$, we must have $\sigma_1 \geq \sigma_2$. But we therefore have our required SVD of A, which establishes our contradiction. □

Exercises: Show that

(i) the columns of U are left-singular vectors of A and eigenvectors of AA^T;

(ii) the columns of V are right-singular vectors of A and eigenvectors of $A^T A$;

(iii) the diagonal entries of Σ are the singular values of A and are the square roots of the eigenvalues of both $A^T A$ and AA^T.

Corollary 2. *Let $A \in \mathbb{R}^{m\times n}$. Then there exist $r \leq \min(m, n)$, an $m \times r$ matrix U with orthonormal columns, an $n \times r$ matrix V with orthonormal columns and an $r \times r$ diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ such that*

$$A = U \Sigma V^T, \tag{1}$$

with $\sigma_1 \geq \ldots \geq \sigma_r > 0$.

In this representation, Σ is uniquely defined, and if its diagonal entries are distinct, then U and V are uniquely defined up to multiplication of any column of U, and the same column of V by -1. Note that if we write $U = (u_1, \ldots, u_r)$ and $V = (v_1, \ldots, v_r)$ then $A = \sum_{i=1}^r \sigma_i u_i v_i^T$.

The SVD is related to Principal Component Analysis (PCA). Suppose X is an $n \times p$ data matrix, representing n realisations of a random vector in $\mathbb{R}^p$. Writing $\bar{X}$ for the p-vector of column means of X, we have the SVD $n^{-1/2}(X - \mathbf{1}_n \bar{X}^T) = U\Sigma V^T$. The columns of V are the eigenvectors of the covariance matrix $n^{-1}(X - \mathbf{1}_n \bar{X}^T)^T(X - \mathbf{1}_n \bar{X}^T)$, while the diagonal entries of Σ are the square roots of the eigenvalues of this covariance matrix. The PCA transformation is

$$Y = \frac{1}{\sqrt{n}}(X - \mathbf{1}_n \bar{X}^T)V = U\Sigma.$$

The jth column of V, i.e. the jth eigenvector of the covariance matrix, is called the jth *principal component* of X, and Y_{ij} represents the *score* of the ith observation with respect to the jth principal component.

As shown below, truncated versions of the SVD can be used to approximate a matrix with one of low rank.

Theorem 3. *Suppose $A \in \mathbb{R}^{m\times n}$ has SVD (1) with $U = (u_1, \ldots, u_r)$ and $V = (v_1, \ldots, v_r)$, and for $k < r$, let $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$. Then*

$$\inf\{\|A - B\|_2 : B \in \mathbb{R}^{m\times n}, \operatorname{rank}(B) = k\} = \|A - A_k\|_2 = \sigma_{k+1}.$$

Proof. First note that $\operatorname{rank}(A_k) = \operatorname{rank}(U^T A_k V) = \operatorname{rank}(\operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)) = k$. Moreover,

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\operatorname{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_r)\|_2 = \sigma_{k+1}.$$

Now suppose $B \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(B) = k$. Then we can find linearly independent vectors $x_1, \ldots, x_{n-k} \in \mathbb{R}^n$ such that $Bx_j = 0$ for $j = 1, \ldots, n-k$. By considering dimensions, we must have

$$\operatorname{span}(v_1, \ldots, v_{k+1}) \cap \operatorname{span}(x_1, \ldots, x_{n-k}) \neq \{0\},$$

so let $z \in \mathbb{R}^n$ with $\|z\|_2 = 1$ be an element of this intersection. Since $Bz = 0$ and we can write $z = \sum_{i=1}^{k+1} \alpha_i v_i$ with $\alpha_1^2 + \ldots + \alpha_{k+1}^2 = 1$, we have

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \alpha_i^2 \sigma_i^2 \geq \sigma_{k+1}^2.$$

□

The Eckart–Young theorem states that A_k also minimises $|||A - B|||_2$ over all $B \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(B) = k$.

Moore–Penrose pseudoinverses

The Moore–Penrose pseudoinverse is a generalisation of the inverse of a matrix to arbitrary $m \times n$ real matrices.

Theorem 4. *For each $A \in \mathbb{R}^{m\times n}$, there exists a unique matrix $A^+ \in \mathbb{R}^{n\times m}$, called the* Moore–Penrose pseudoinverse *of A, such that A^+A and AA^+ are symmetric, and the following two properties hold:*

1. *$AA^+A = A$ (i.e. AA^+ need not be the identity matrix, but it maps each column of A to itself);*
2. *$A^+AA^+ = A^+$.*

Proof. For uniqueness, suppose that $B, C \in \mathbb{R}^{n\times m}$ satisfy the required properties. Then

$$AB = (AB)^T = B^TA^T = B^T(ACA)^T = (AB)^T(AC)^T = ABAC = AC,$$

and analogously, $BA = CA$. It follows that $B = BAB = BAC = CAC = C$.

For existence, first suppose that $A = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{m\times n}$, where $r = \min(m, n)$. Then it is a straightforward exercise to check that $A^+ = \operatorname{diag}(\sigma_1^+, \ldots, \sigma_r^+) \in \mathbb{R}^{n\times m}$, where

$$\sigma^+ = \begin{cases} 1/\sigma & \text{if } \sigma \neq 0 \\ 0 & \text{if } \sigma = 0. \end{cases}$$

For general $A \in \mathbb{R}^{m\times n}$, suppose it has SVD $A = U\Sigma V^T$. Then it is a straightforward exercise to show that $A^+ = V\Sigma^+U^T$. □

Exercise: Show that $A^+ = (A^TA)^+A^T$ (which means that it suffices to understand Moore–Penrose pseudoinverses for symmetric matrices). In particular, if the columns of A are linearly independent (so $m \geq n$), then A^TA is positive definite, and $A^+ = (A^TA)^{-1}A^T$. Hence AA^+ represents an orthogonal projection onto the subspace spanned by the columns of A.

Similarly, $A^+ = A^T(AA^T)^+$, so if the rows of A are linearly independent (so $m \leq n$), then $A^+ = A^T(AA^T)^{-1}$.

Exercise: Show that pseudoinversion commutes with transposition, i.e. $(A^T)^+ = (A^+)^T$.

Exercise: Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Suppose that one of the following three conditions hold:

1. the columns of A are orthonormal (so $A^TA = I_n$);
2. the rows of B are orthonormal;
3. A has linearly independent columns and B has linearly independent rows.

Show that $(AB)^+ = B^+A^+$.

Here are two of the most useful properties of the Moore–Penrose pseudoinverse:

Theorem 5. *For any $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, we have $\|Ax - b\|_2 \geq \|Az - b\|_2$, where $z = A^+b$.*

Proof. First note that

$$A^T(Az - b) = A^T(AA^+b - b) = A^T(AA^+)^Tb - A^Tb = (AA^+A)^Tb - A^Tb = 0.$$

It follows that

$$\begin{aligned}\|Ax - b\|_2^2 = \|A(x - z) + Az - b\|_2^2 &= \|Az - b\|_2^2 + 2(x - z)^TA^T(Az - b) + \|A(x - z)\|_2^2 \\ &= \|Az - b\|_2^2 + \|A(x - z)\|_2^2 \geq \|Az - b\|_2^2.\end{aligned}$$

□

Theorem 6. *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. If $z = A^+b$ satisfies $Az = b$, then among all solutions $x \in \mathbb{R}^n$ of $Ax = b$, z is the unique solution with minimal Euclidean norm.*

Proof. If $x \in \mathbb{R}^n$ satisfies $Ax = b$, then since A^+A is symmetric,

$$z^T(x - z) = (A^+Az)^T(x - z) = z^T(A^+A)(x - z) = z^T(A^+b - z) = 0.$$

Thus

$$\|x\|_2^2 = \|z\|_2^2 + 2z^T(x - z) + \|x - z\|^2 = \|z\|_2^2 + \|x - z\|_2^2 \geq \|z\|_2^2,$$

with equality if and only if $x = z$. □

Basic convex analysis

In this course, we will only need to consider convex functions taking finite values. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if

$$f\big((1 - t)x + ty\big) \leq (1 - t)f(x) + tf(y)$$

for all $x, y \in \mathbb{R}^n$ and $t \in (0, 1)$. It is *strictly convex* if the inequality is strict for all $x, y \in \mathbb{R}^n$ and $t \in (0, 1)$. A vector $v \in \mathbb{R}^n$ is a *subgradient* of f at x if

$$f(y) \geq f(x) + v^T(y - x)$$

for all $y \in \mathbb{R}^n$. The set of subgradients of f at x is denoted $\partial f(x)$. For finite convex functions, $\partial f(x)$ is a non-empty, compact, convex set. The following easy exercise, which gives a criterion for determining minimisers of a convex function, is often referred to in the statistical literature as the Karush–Kuhn–Tucker (KKT) conditions, though really we are only using a much simplified version of them.

Exercise: $x^* \in \operatorname{argmin}_{x\in\mathbb{R}^n} f(x)$ if and only if $0 \in \partial f(x^*)$.

We will also require the following two results:

Proposition 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, and suppose f is differentiable at x. Then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Suppose $v \in \mathbb{R}^n$ is a subgradient of f at x. Then, for any $z \in \mathbb{R}^n$, we have

$$\nabla f(x)^T z = \lim_{\lambda \searrow 0} \frac{f(x + \lambda z) - f(x)}{\lambda} \geq v^T z,$$

so we must have $v = \nabla f(x)$. Conversely, for any $z \in \mathbb{R}^n$ with $z \neq x$, note that $t \mapsto \{f(x + tz) - f(x)\}/t$ decreases as $t \searrow 0$ (check). Thus, for any $\lambda > 0$, we have

$$\frac{f(x + \lambda z) - f(x)}{\lambda} \geq \lim_{t \searrow 0} \frac{f(x + tz) - f(x)}{t} = \nabla f(x)^T z,$$

which is equivalent to the statement that $\nabla f(x)$ is a subgradient of f at x. □

Proposition 8. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be convex, and suppose that f is differentiable at x. Then*

$$\partial(f + g)(x) = \{\nabla f(x)\} + \partial g(x).$$

Remark: In fact, for finite convex functions defined on $\mathbb{R}^n$, the more general statement that $\partial(f + g)(x) = \partial f(x) + \partial g(x)$ is true, though we will not require this, and the proof is more complicated.

Proof. Suppose $v \in \mathbb{R}^n$ is a subgradient of g at x. Then, for any $y \in \mathbb{R}^n$, and by Proposition 7, we have

$$f(y) + g(y) \geq f(x) + g(x) + \{\nabla f(x) + v\}^T (y - x),$$

so $\{\nabla f(x)\} + \partial g(x) \subseteq \partial(f + g)(x)$.

Conversely, let $v \in \partial(f + g)(x)$. Then, for any $y \in \mathbb{R}^n$ with $y \neq x$, writing $z_t = x + t(y - x)$,

$$\begin{aligned}
0 &\leq \liminf_{t \searrow 0} \frac{f(z_t) + g(z_t) - \{f(x) + g(x) + v^T(z_t - x)\}}{\|z_t - x\|} - \lim_{t \searrow 0} \frac{f(z_t) - f(x) - \nabla f(x)^T (z_t - x)}{\|z_t - x\|} \\
&= \liminf_{t \searrow 0} \frac{g\big(x + t(y - x)\big) - g(x) - t(v - \nabla f(x))^T (y - x)}{t\|y - x\|} \\
&\leq \frac{g(y) - g(x) - (v - \nabla f(x))^T (y - x)}{\|y - x\|},
\end{aligned}$$

so $v - \nabla f(x) \in \partial g(x)$, as required. □