Mathematical Review

The purpose of this review is to summarize basic facts on linear algebra and calculus that we are going to use in this course. Theorems that are not proved in this review can be found in most standard textbooks.

Matrices

An $m \times n$ *matrix* $A = (a_{ij})$ is a rectangular array of (real) numbers, which has m rows and n columns, that is

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The transpose of A is the $n \times m$ matrix

$$\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

which is denoted by A^T. If $A = A^T$, then A is called *symmetric*. Clearly, a symmetric matrix must be a *square* matrix, that is, $m = n$. We denote by I the *identity matrix*

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

If α is a number, then

$$\alpha A = (\alpha a_{ij}).$$

If $B = (b_{ij})$ is an $m \times n$ matrix, then

$$A \pm B = (a_{ij} \pm b_{ij}).$$

If $B = (b_{ij})$ is an $n \times p$ matrix, then AB is an $m \times p$ matrix (c_{ij}), where, for each i and j,

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}.$$

Theorem 1. *Matrix product satisfies*
(1) $(AB)C = A(BC)$;
(2) $(AB)^T = B^T A^T$.

Proof. Let $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{n\times p}$, and $C = (c_{ij})_{p\times q}$. Then $AB = (\sum_{\alpha=1}^n a_{i\alpha}b_{\alpha j})_{m\times p}$ and

$$(AB)C = \left(\sum_{\beta=1}^{p} (\sum_{\alpha=1}^{n} a_{i\alpha}b_{\alpha\beta})c_{\beta j} \right)_{m\times q}.$$

Similarly, $BC = (\sum_{\beta=1}^p b_{i\beta}c_{\beta j})_{n\times q}$ and

$$A(BC) = \left(\sum_{\alpha=1}^{n} a_{i\alpha}(\sum_{\beta=1}^{p} b_{\alpha\beta}c_{\beta j}) \right)_{m\times q}.$$

Since

$$\begin{aligned} \sum_{\beta=1}^{p}(\sum_{\alpha=1}^{n} a_{i\alpha}b_{\alpha\beta})c_{\beta j} &= \sum_{\beta=1}^{p}\sum_{\alpha=1}^{n} a_{i\alpha}b_{\alpha\beta}c_{\beta j} \\ = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{p} a_{i\alpha}b_{\alpha\beta}c_{\beta j} &= \sum_{\alpha=1}^{n} a_{i\alpha}(\sum_{\beta=1}^{p} b_{\alpha\beta}c_{\beta j}), \end{aligned}$$

thus (1) is proved. The proof of (2) is even easier. Let $a'_{ij} = a_{ji}$ abd $b'_{ij} = b_{ji}$. Then $A^T = (a'_{ij})$ and $B^T = (b'_{ij})$. Notice that the ij-th entry of $(AB)^T$ is

$$\sum_{\alpha=1}^{n} a_{j\alpha}b_{\alpha i} = \sum_{\alpha=1}^{n} b_{\alpha i}a_{j\alpha} = \sum_{\alpha=1}^{n} b'_{i\alpha}a'_{\alpha j},$$

which is exactly the ij-th entry of $B^T A^T$. □

Linear and quadratic functions

In the following, we express linear and quadratic functions in vector form. First it is clear that the linear function

$$f(x_1, x_2, ..., x_n) = c_0 + c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

can always be expressed as

$$f(x) = c_0 + cx,$$

where $c = [c_1, c_2, ..., c_n]$ and $x = [x_1, x_2, ..., x_n]^T$.

For a general (homogeneous) quadratic function:

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} a_{ii}x_i^2 + \sum_{1\le j<k\le n} 2a_{ij}x_ix_j,$$

let $A = (a_{ij})$, where, for all $n \ge i > j \ge 1$, we define $a_{ij} = a_{ji}$. It is easy to see that A is symmetric and

$$f(x) = x^T Ax.$$

Derivatives

Let f: $\mathbb{R}^n \to \mathbb{R}$. We define

$$\lim_{x\to a} f(x) = \alpha$$

if *for any given* $\epsilon > 0$, *there exists* $\delta > 0$, *such that* $|f(x)-\alpha| < \epsilon$, *for all* x *with* $||x - a|| < \delta$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then we define

$$\nabla f \triangleq \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}\right].$$

For example, if $f(x) = ax$, where a is a row vector, then $\nabla f = a$.

Theorem. Chain Rule. *Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is a function on* n *variables* $x_1, x_2, ..., x_n$ *and each* $x_i : \mathbb{R} \to \mathbb{R}$ *is a function of* t. *Then,*

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dt} + ... + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt} = (\nabla f)\cdot(\nabla x).$$

Let f: $\mathbb{R}^n \to \mathbb{R}$ and let $d \in \mathbb{R}^n$. The *directional derivative* of f in the direction d is

$$\frac{\partial f}{\partial d}(x) \triangleq \lim_{\alpha \to 0} \frac{f(x+\alpha d) - f(x)}{\alpha}.$$

By Chain Rule, it is easy to prove the following.

Theorem. $\dfrac{\partial f}{\partial d} = (\nabla f) \cdot d$.

Corollary. *The gradient ∇f is the direction in which f increases the fastest.*

In general, if $f(x) = [f_1(x), f_2(x), \ldots, f_m(x)]^T : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, we define

$$\nabla f \triangleq \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

For example, let $f(x) = Ax$, where A is an $m \times n$ matrix, and let $a_1, a_2, \ldots, a_m$ be the m rows of A. Then $f = [f_1, f_2, \ldots, f_m]^T$ with $f_i(x) = a_i x$, for all i. Therefore,

$$\nabla f = \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = A.$$

In particular, by taking $A = I$, the identity matrix, we deduce that, if $f(x) = x$, then $\nabla f = I$.
Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then we define Hf to be

$$\nabla(\nabla f)^T = \begin{bmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial^2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{bmatrix},$$

which is aslo called the *Hessian* of f.

Theorem. Product Rule. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are both differentiable. Then*

$$\nabla(f^T g) = g^T(\nabla f) + f^T(\nabla g).$$

As an example, we consider $h(x) = x^T A x$. Let $f(x) = x$ and $g(x) = Ax$. Then $\nabla f = I$ and $\nabla g = A$. From the product rule we deduce that

$$\nabla h = x^T A^T I + x^t A = x^T(A + A^T)$$

and thus

$$Hh = \nabla(\nabla h)^T = \nabla[(A + A^T)x] = A + A^T.$$

If A is symmetric, then we have

$$\nabla(x^T A x) = 2x^T A \quad \text{and} \quad H(x^T A x) = 2A.$$

In particular,

$$\nabla(x^T x) = 2x^T \quad \text{and} \quad H(x^T x) = 2I.$$

Determinant

Next, we define the *determinant* of an $n \times n$ matrix $A = (a_{ij})$, which is denoted by $det(A)$ or $|A|$. If $n = 1$, then $|A| \triangleq a_{11}$. When $n > 1$, for each i, let A_i be the $(n-1) \times (n-1)$ matrix obtained from A by deleting the first row and the i-th column. Then

$$|A| \triangleq a_{11}|A_1| - a_{12}|A_2| + a_{13}|A_3| + \cdots + (-1)^{n-1}|A_n|.$$

In particular,

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - gec - hfa - idb.$$

In general, the following is true.

Theorem 2. *(***Laplace's Expansion Theorem.***)*
$$|A| = \sum_{\pi} sign(\pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)},$$
where the sum is taken over all permutations π of $\{1, 2, ..., n\}$, and $sign(\pi) = \pm 1$.

Matrix $A = (a_{ij})$ is *lower triangular* if $a_{ij} = 0$, for all $i > j$.

Proposition 3. *If $A = (a_{ij})_{n \times n}$ is lower triangular, then $|A| = a_{11}a_{22} \cdots a_{nn}$.*

Proof. We prove the result by induction on n. If $n = 1$, the result holds trivially. Next, we assume that $n > 1$. Let A_1 be the matrix obtained from A by deleting the first row and the first column. Then it is clear that A_1 is an $(n-1) \times (n-1)$ lower triangular matrix, with a_{22}, a_{33}, ..., a_{nn} be its diagonal entries. By induction, $|A_1| = a_{22}a_{33}...a_{nn}$. Then, from the above definition, as A is lower triangular, we deduce that $|A| = a_{11}|A_1| = a_{11}a_{22}...a_{nn}$. □

Theorem 4. $|A| = |A^T|$.

A matrix A is called *triangular* if either A or A^T is lower triangular. The last two results imply the following immediately.

Proposition 5. *If $A = (a_{ij})_{n \times n}$ is triangular, then $|A| = a_{11}a_{22} \cdots a_{nn}$.*

Theorem 6. *(***Binet-Cauchy***) If both A and B are $n \times n$ matrices, then $|AB| = |A||B|$.*

Theorem 7. *If $|A| \neq 0$, then there exists B with $AB = BA = I$.*

Matrix B in the last theorem is called the *inverse* of A and is denoted by A^{-1}. The following result is very useful.

Proposition 8. $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \dfrac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

For any two vectors x, y, since $x \cdot y = ||x|| ||y|| \cos\theta$, where θ is the angle between the two vectors, it is natural to say that x, y are *orthogonal* if $x \cdot y = 0$. A square matrix D is *orthogonal* if $D^T D = I$. Equivalently, D is orthogonal if every column of D is a unit vector (a vector whose Euclidean norm is equal to one), and any two distinct columns of D are orthogonal.

Proposition 9. *If D is orthogonal, then $|D| = \pm 1$.*

PROOF. By Theorem 4, Theorem 6, and Proposition 5, we have $|D|^2 = |D^T||D| = |I| = 1$ and thus the result follows. $\square$

Proposition 10. *If D is orthogonal, then so is D^T. That is, $DD^T = I$.*

PROOF. By Proposition 9, the inverse D^{-1} of D exists. Thus, by Theorem 1, $D^T = D^T DD^{-1} = D^{-1}$. Therefore, we have $DD^T = DD^{-1} = I$. $\square$

Symmetric matrices

Proposition 11. *Let A be an $n \times n$ matrix. Then $p(\lambda) = |\lambda I - A|$ is a polynomial on λ of degree n.*

PROOF. Let $A = (a_{ij})$ and $B = (b_{ij}) = \lambda I - A$. Then we have $b_{ii} = \lambda - a_{ii}$, for all i, and $b_{ij} = -a_{ij}$, for all $i \neq j$. By Laplace's Expansion Theorem,

$$p(\lambda) = \sum_{\pi} sign(\pi) b_{1\pi(1)} b_{2\pi(2)} \cdots b_{n\pi(n)}. \tag{1}$$

Since each b_{ij} is a polynomial of λ, every term in (1) is a polynomial of λ, and thus so is $p(\lambda)$. Notice that each b_{ij} has degree either 0 or 1, and each term in (1) is the product of exactly n b_{ij}'s, so the degree of $p(\lambda)$ is at most n. To prove the degree of $p(\lambda)$ is n, it remains to show that, in (1), the coefficient of λ^n is not zero. Observe that the only term in (1) that has degree n is

$$b_{11} b_{22} \cdots b_{nn} = (\lambda - a_{11})(\lambda - a_{22}) \cdots (\lambda - a_n),$$

which has a non-zero coefficient. Therefore, the proof of the proposition is complete. $\square$

We call $p(\lambda)$ the *characteristic* polynomial of A, and we call its n roots the *eigenvalues* of A.

Theorem 12. *If A is a symmetric matrix, then there exists an orthogonal matrix D such that*

$$DAD^T = diag(\lambda_1, \lambda_2, ..., \lambda_n),$$

where

$$diag(\lambda_1, \lambda_2, ..., \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

and $\lambda_1, \lambda_2, ..., \lambda_n$ are the n eigenvalues of A.

Let A be a symmetric matrix. It is called *positive definite* if $x^T Ax > 0$ for all $x \neq 0$. It is *positive semidefinite* if $x^T Ax \geq 0$ for all x. We write $A > 0$ and $A \geq 0$ for these two cases.

Proposition 13. *Let A and B be $n \times n$ matrices and let A be symmetric. If $A > 0$ and $|B| \neq 0$, then $B^T AB > 0$.*

PROOF. We need to show that $x^T B^T ABx > 0$, for all $x \neq 0$. Let $y = Bx$. Since $x \neq 0$, we must $y \neq 0$, because $y = 0$ would imply, by Theorem 7, $x = B^{-1}y = 0$, which is not the case. Therefore, by Theorem 1, $x^T B^T ABx = y^T Ay > 0$. $\square$

Theorem 14. *A symmetric matrix A is positive definite (or semidefinite) if and only if all its eigenvalues are positive (or nonnegative).*

PROOF. We only consider the case for positive definite. The other case can be proved similarly.

We first prove the "if" part by assuming that all eigenvalues of A are positive. We need to show $A > 0$. Let D and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ be determined as in Theorem 12. Clearly, $\Lambda > 0$, as $\lambda_i > 0$, for all i. By Theorem 1(1), $D^T \Lambda D = D^T DAD^T D = A$. Therefore, by Proposition 9 and Proposition 13, $A > 0$.

Next, we prove the "only if" part by assume that $x^T Ax > 0$, for all $x \neq 0$. We need to show that each eigenvalue λ_i of A are positive. Let D and Λ be as before. Let x be the i-th column of D^T. By Proposition 10, D^T is orthogonal, so x is a unit vector and thus $x \neq 0$. Again, as D^T is orthogonal, $DD^T = I$, which implies $Dx = e_i$, where e_i is the i-th column of I. Now

$$\lambda_i = e_i^T \Lambda e_i = x^T D^T \Lambda Dx = x^T Ax > 0,$$

and the theorem is proved. $\square$

Theorem 15. Sylvester's Criterion. *A symmetric matrix A is positive definite if and only if its leading principal minors are positive, that is, $|A_i| > 0$, for all i, where*

$$A_i = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1i} \\ a_{21} & a_{22} & \cdots & a_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ii} \end{bmatrix}.$$

PROOF. We first prove "only if" by assuming that A is positive definite. We need to show that $|A_i| > 0$, for every i. By Theorem 12, there exists an orthogonal matrix D such that $DA_iD^T = diag(\lambda_1, \lambda_2, ..., \lambda_i)$, where $\lambda_1, \lambda_2, ..., \lambda_i$ are the eigenvalues of A_i. By Theorem 6,

$$|DA_iD^T| = |D||A_i||D^T| = |D^T D||A_i| = |A_i|,$$

which implies, by Proposition 3, that $|A_i| = \lambda_1 \lambda_2 ... \lambda_i$. Therefore, by Theorem 14, we only need to show that A_i is positive definite. To prove this, take any $x \neq 0$ that belongs to R^i. Let

$$y = \begin{bmatrix} x \\ 0 \end{bmatrix} \in R^n.$$

It is clear that $y \neq 0$. Since $A > 0$, we deduce that $x^T A_i x = y^T Ay > 0$, and thus $A_i > 0$, as we wanted.

Next, we prove the "if" part by assuming that $|A_i| > 0$, for all i. We need to show that $A > 0$. We prove this by induction on n. The result is clear if $n = 1$, so we assume that $n > 1$. We express A as

$$\begin{bmatrix} A_{n-1} & a \\ a^T & a_{nn} \end{bmatrix}.$$

Since $|A_{n-1}| > 0$, by Theorem 7, the inverse A_{n-1}^{-1} of A_{n-1} exists. From

$$A_{n-1}(A_{n-1}^{-1})^T = A_{n-1}^T (A_{n-1}^{-1})^T = (A_{n-1}^{-1} A_{n-1})^T = I$$

we deduce that $(A_{n-1}^{-1})^T = A_{n-1}^{-1}$. Let $b = A_{n-1}^{-1} a$ and $\alpha = a_{nn} - a^T A_{n-1}^{-1} a$. Then it is straightforward to verify that

$$A = \begin{bmatrix} I_{n-1} & 0 \\ b^T & 1 \end{bmatrix} \begin{bmatrix} A_{n-1} & 0 \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} I_{n-1} & b \\ 0 & 1 \end{bmatrix}. \tag{2}$$

It is clear from Proposition 13 and Proposition 5 that we only need to show that $B = \begin{bmatrix} A_{n-1} & 0 \\ 0 & \alpha \end{bmatrix} > 0$.

By taking determinants on both sides of (2), we deduce from Theorem 6 that

$$|A| = \alpha |A_{n-1}|,$$

which implies $\alpha > 0$. Let us express any vector $x \neq 0$ as $\begin{bmatrix} y \\ z \end{bmatrix}$, where $y \in R^{n-1}$ and $z \in R$. Then we have $x^T Bx = y^T A_{n-1} y + \alpha z^2$. By induction, $A_{n-1} > 0$. Therefore, $x^T Bx > 0$, as we wanted. □

It is much harder to use determinants to test if a symmetric matrix $A_{n\times n}$ is positive semidefinite. Let $i_1, i_2, ..., i_k$ be distinct members of $\{1, 2, ..., n\}$. Let B be the matrix obtained from A by deleting rows indexed by $i_1, i_2, ..., i_k$ and also columns indexed by $i_1, i_2, ..., i_k$. Then $det(B)$ is called a *principal minor* of A. Notice that A has only n leading principal minors, but it has $2^n - 1$ principal minors.

Theorem 16. *A symmetric matrix is positive (negative) semidefinate if and only if all its principal minors are nonnegative (nonpositive).*

Theorem 17. Rayleigh Inequality. *If A is an $n \times n$ symmetric matrix and $x \in R^n$, then,*

$$\lambda_{min}||x||^2 \leq x^T Ax \leq \lambda_{max}||x||^2,$$

where λ_{min} and λ_{max}, respectively, are the smallest and largest eigenvalues of A.

Proof. Let D and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ be determined as in Theorem 12. Then $A = D^T \Lambda D$. Let $y = [y_1, y_2, ..., y_n]^T = Dx$. Then

$$\begin{aligned} \lambda_{min}||x||^2 &= \lambda_{min} x^T x = \lambda_{min} x^T D^T Dx = \lambda_{min} y^T y \\ &= \lambda_{min}(y_1^2 + y_2^2 + ... + y_n^2) \\ &\leq \lambda_1 y_1^2 + \lambda_2 y_2^2 + ... + \lambda_n y_n^2 \\ &= y^T \Lambda y = x^T D^T \Lambda Dx = x^T Ax. \end{aligned}$$

The other inequality can be proved similarly. □