

Linear algebra/Analysis and Calculus review

Chapter 1

Linear Algebra

1.1 Matrices

1.1.1 Basics

Nullspace The *nullspace* (or, kernel) of a $m \times n$ matrix A is the following subspace of \mathbf{R}^n :

$$\mathcal{N}(A) := \{x \in \mathbf{R}^n : Ax = 0\}.$$

Range and rank The *range* (or, image) of a $m \times n$ matrix A is defined as the following subset of \mathbf{R}^m :

$$\mathcal{R}(A) := \{Ax : x \in \mathbf{R}^n\}.$$

The range is simply the span of the columns of A .

The dimension of the range is called the *rank* of the matrix. As we will see later, the rank cannot exceed any one of the dimensions of the matrix A : $r \leq \min(m, n)$. It is equal to n minus the dimension of its nullspace.

A basic result of linear algebra states that any vector in \mathbf{R}^n can be decomposed as $x = y + z$, with $y \in \mathcal{N}(A)$, $z \in \mathcal{R}(A^T)$, and z, y are orthogonal. (One way to prove this is via the singular value decomposition, seen later.)

Symmetric Matrices : A square matrix $A \in \mathbf{R}^{n \times n}$ is *symmetric* if and only if $A = A^T$. The set of symmetric $n \times n$ matrices is denoted \mathcal{S}^n .

Orthogonal matrices. A square, $n \times n$ matrix $U = [u_1, \dots, u_n]$ is orthogonal if its columns form an orthonormal basis. The condition $u_i^T u_j = 0$ if $i \neq j$, and 1 otherwise, translates in matrix terms as $U^T U = I_n$ with I_n the $n \times n$ identity matrix.

Unitary matrix: An $n \times n$ matrix U is unitary if $UU^* = U^*U = I$ where U^* is the transpose of the conjugate of U .

Normal matrix: An $n \times n$ matrix A is normal if $AA^* = A^*A$

1.1.2 Eigenvalue decomposition

A fundamental result of linear algebra states that any symmetric matrix can be decomposed as a weighted sum of normalized dyads that are orthogonal to each other.

Precisely, for every $A \in \mathcal{S}^n$, there exist numbers $\lambda_1, \dots, \lambda_n$ and an orthonormal basis (u_1, \dots, u_n) , such that

$$A = \sum_{i=1}^n \lambda_i u_i u_i^T.$$

In a more compact matrix notation, we have $A = U\Lambda U^T$, with $\Lambda = \mathbf{diag}(\lambda_1, \dots, \lambda_n)$, and $U = [u_1, \dots, u_n]$.

The numbers $\lambda_1, \dots, \lambda_n$ are called the eigenvalues of A , and are the roots of the characteristic equation

$$\det(\lambda I - A) = 0,$$

where I_n is the $n \times n$ identity matrix. Eigenvalues and eigenvectors satisfies $Au_i = \lambda_i u_i$, some other properties of eigenvalues

- $\det(A) = \prod_{i=1}^n \lambda_i$
- $Tr(A) = \sum_{i=1}^n \lambda_i$

For arbitrary square matrices, eigenvalues can be complex. In the symmetric case, the eigenvalues are always real. There are only n (possibly distinct) solutions to the above equation.

It is interesting to see what the eigenvalue decomposition of a given symmetric matrix A tells us about the corresponding quadratic form, $q_A(x) := x^T A x$. With $A = U\Lambda U^T$, we have

$$q_A(x) = (U^T x)^T \Lambda (U^T x) = \sum_{i=1}^n \lambda_i (u_i^T x)^2.$$

The eigenvalue decomposition thus corresponds to the decomposition of the corresponding quadratic form into a sum of squares.

Spectral radius : The spectral radius $\rho(A)$ of a square matrix A is defined as the maximum of the magnitudes of the eigenvalues of A .

Spectrum : The set of eigenvalues of A is called its spectrum.

Proposition : Two $n \times n$ matrices, A and B are called similar if $B = S^{-1}AS$ for some invertible matrix S . If two matrices are similar then they have same characteristic polynomials.

1.1.3 Singular value decomposition

The singular value decomposition states that any matrix $A \in \mathbf{R}^{m \times n}$ can be expressed as

$$A = U \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T = \sum_{i=1}^r \sigma_i u_i v_i^T,$$

with $U \in \mathbf{R}^{m \times m}$, $V \in \mathbf{R}^{n \times n}$, U, V orthogonal, and $\Sigma = \mathbf{diag}(\sigma_1, \dots, \sigma_r)$ contain the *singular values* of A . The number $r \leq \min(m, n)$ is the rank of A , and equals the dimension of its range.

The singular values of A is the square roots of the positive eigenvalues of $A^T A$

Moore Penrose inverse of a matrix $A \in \mathbf{R}^{m \times n}$, denoted by A^+ is defined as

$$A^+ := V \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T$$

Thus $A^+ y$ is the solution for minimizing $|Ax - y|$.

1.1.4 Positive semi-definite matrices

Definition. A matrix $A \in \mathcal{S}^n$ is said to be *positive-definite* (resp. *positive semi-definite*) if and only if all the eigenvalues are positive (resp. non-negative). We use the acronyms PD and PSD for these properties. The set of $n \times n$ PSD matrices is denoted \mathcal{S}_+^n , while that of PD matrices is written \mathcal{S}_{++}^n . Often, we use the notation $A \succeq 0$ (resp. $A \succ$) for the PSD (resp. PD) property.

In terms of the associated quadratic form $q_A(x) = x^T A x$, the interpretation is as follows. A matrix A is PD if and only if q_A is a positive-definite function, that is, $q_A(x) = 0$ if and only if $x = 0$. Indeed, when $\lambda_i > 0$ for every i , then the condition

$$q_A(x) = \sum_{i=1}^n \lambda_i (u_i^T x)^2 = 0$$

trivially implies $u_i^T x = 0$ for every i , which can be written as $Ux = 0$. Since U is orthogonal, it is invertible, and we conclude that $x = 0$.

Example A well-known example of a PSD matrix is the covariance matrix associated with a random variable in \mathbf{R}^n . This matrix is defined as

$$\Sigma = \mathbf{E}(x - \hat{x})(x - \hat{x})^T,$$

where $\hat{x} := \mathbf{E} x$, and \mathbf{E} denotes the expectation operator associated with the distribution of the random variable x .

1.1.5 Schur complement

Schur complement of a block matrix $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ is defined as

$$M/A = D - CA^{-1}B$$

Some properties of Schur complement

- $\det(M) = \det(A)\det(M/A)$
- $\text{rank}(M) = \text{rank}(A) + \text{rank}(M/A)$

1.2 Norms

1.2.1 Norms

Definition A mapping $(\cdot, \cdot) : V \times V \rightarrow F$ is called an inner product if

- $(x, y) = \overline{(y, x)}$
- $(x, x) \geq 0, \forall x \in V, '= ' \text{ iff } x = 0$
- $(ax + by, z) = a(x, z) + b(y, z), \forall a, b \in F$

The inner product on $\mathbb{R}^n : \langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i$

The inner product on $\mathbb{R}^{m \times n} : \langle X, Y \rangle = \text{tr}(X^T Y) = \sum_{i,j} X_{ij} Y_{ij}$

Lemma $\langle x, y \rangle$ is an inner product on \mathbb{R}^n iff there exists a positive definite matrix A such that $\langle x, y \rangle = x^T A y$

Definition A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is called a norm if

- f is nonnegative: $f(x) \geq 0, \forall x \in \mathbb{R}^n$
- f is definite: $f(x) = 0$ only if $x = 0$
- f is homogenous: $f(tx) = |t|f(x), \forall x \in \mathbb{R}^n, t \in \mathbb{R}$
- f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y), \forall x, y \in \mathbb{R}^n$

1.2.2 Examples

The Euclidean norm is $\|x\|_2 = \sqrt{x^T x}$

The l_1 -norm or Manhattan distance: $\|x\|_1 = \sum_{i=1}^n |x_i|$

Chebyshev or l_∞ -norm: $\|x\|_\infty = \max_i \{|x_i|\}$

The l_p -norm ($p \geq 1$): $\|x\|_p = \left(\sum_{i=1}^n \|x_i\|^p \right)^{1/p}$

Quadratic norms: $P \in \mathcal{S}_{++}^n$, $\|x\|_P = \sqrt{x^T P x} = \|P^{1/2} x\|_2$

The Frobenius norm of a matrix $X \in \mathbb{R}^{m \times n}$ is

$$\|X\|_F = \sqrt{\text{tr}(X^T X)} = \sqrt{\sum_{i=1}^n \lambda_i^2}$$

There are many other norms that are important or interesting in some applications. For example, for $k \in \{1, \dots, n\}$ we can define

$$\|x\|_{1,k} := \sum_{i=1}^k |x|_{[i]}$$

where for every i , $|x|_{[i]}$ is the i -th largest absolute value of elements of x . The norm is a kind of mixture between the l_1 - and l_∞ -norms, respectively obtained upon setting $k = n$ and $k = 1$.

The largest singular value of A can be characterized as

$$\sigma_{\max}(A) = \max_{1 \leq i \leq r} \sigma_i = \max_x \|Ax\|_2 : \|x\|_2 = 1.$$

The largest singular value is a matrix norm.

1.2.3 Dual norm

Let $\|\cdot\|$ be a norm on \mathbb{R}^n . The associated dual norm, denoted $\|\cdot\|_*$ is defined as

$$\|z\|_* = \sup\{z^T x \mid \|x\| \leq 1\}$$

The dual of the dual norm is the original norm: $\|x\|_{**} = \|x\|, \forall x$ and

$$z^T x \leq \|x\| \|z\|_*, \forall x, z$$

The dual of the Euclidean norm is the Euclidean norm, due to Cauchy-Schwarz inequality

$$\sup\{z^T x \mid \|x\|_2 \leq 1\} = \|z\|_2$$

The dual of l_1 -norm is the l_∞ -norm

Generally, the dual of the l_p norm is the l_q norm where $1/p + 1/q = 1$

1.2.4 Norm equivalence

For any two norms $\|\cdot\|$ and $\|\cdot\|'$ on R^n , there exists some positive constant $c \in R$ such that $\|x\| \leq c\|x\|', \forall x \in R^n$

Weierstrass's Theorem If f is a real-valued continuous function on a non-empty compact set S , then there $\min_x f$ is attained.

Proof Let $a = \min_{\|x\|=1} \|x\|'$, then

$$0 < a = \|x/\|x\|\|'\| = \|x\|'/\|x\|$$

(Using Weierstrass's theorem)

Chapter 2

Analysis and Calculus

2.1 Open and closed sets

An element $x \in C \subseteq \mathbb{R}^n$ is called an interior point of C if there exists an $\epsilon > 0$ such that $\{y \mid \|y - x\|_2 \leq \epsilon\} \subseteq C$

The set of all points interior to C is called the interior of C , denoted as $\mathbf{int}C$.

A set C is open if $\mathbf{int}C = C$

A set C is closed if its complement $\mathbb{R}^n \setminus C$ is open

The closure of C is defined as $\bar{C} = \mathbb{R}^n \setminus \mathbf{int}(\mathbb{R}^n \setminus C)$. A point x is in the closure of C if $\forall \epsilon > 0, \exists y \in C : \|x - y\|_2 \leq \epsilon$

The boundary of C is defined as $C^0 = \bar{C} \setminus \mathbf{int}C$.

2.2 Sequences

A sequence $\{x_k\}$ of scalars is said to converge to a scalar x if

$$\forall \epsilon > 0, \exists K = K(\epsilon) : |x_k - x| < \epsilon, \forall k \geq K$$

We write this as $\lim_{k \rightarrow \infty} x_k = x$

A sequence $\{x_k\}$ of scalars is said to converge to ∞ if

$$\forall b, \exists K : x_k \geq b, \forall k \geq K$$

A sequence $\{x_k\}$ is Cauchy sequence if

$$\forall \epsilon > 0, \exists K = K(\epsilon) : |x_m - x_n| < \epsilon, \forall m, n \geq K$$

Increasing/nonincreasing (monotonic) sequence, bounded above/below sequence

Proposition: Every monotonic sequence converges to a possibly infinite number. If it is bounded, then it converges to a finite real number.

We say that a vector $x \in \mathbb{R}^n$ is a limit point of a sequence $\{x_k\} \subset \mathbb{R}^n$ if there exists a subsequence of $\{x_k\}$ that converges to x .

Proposition:

- (a) A bounded sequence converges iff it has a unique limit point.
- (b) A sequence converges iff it is a Cauchy sequence
- (c) Every bounded sequence has at least one limit point.

2.3 Functions

Domain of a function. The *domain* of a function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is the set $\mathbf{dom} f \subseteq \mathbf{R}^n$ over which f is well-defined, in other words:

$$\mathbf{dom} f := \{x \in \mathbf{R}^n : -\infty < f(x) < +\infty\}.$$

Here are some examples:

- The function with values $f(x) = \log(x)$ has domain $\mathbf{dom} f = \mathbf{R}_{++}$.
- The function with values $f(X) = \log \det(X)$ has domain $\mathbf{dom} f = \mathcal{S}_{++}^n$ (the set of positive-definite matrices).

epi graph Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the epigraph of f is the set

$$\mathbf{epi} f = \{(x, t) \in \mathbb{R}^{n+1} \mid x \in \mathbf{dom} f, f(x) \leq t\}$$

Function f is said to be closed if its epigraph is closed.

Continuous function: f is continuous at x if for any sequence $\{x_i\}$ that converges to x , $f(x_i)$ also converges to $f(x)$: $\lim_{n \rightarrow \infty} f(x_i) = f(\lim_{n \rightarrow \infty} x_i)$

Linear function : What kind of function satisfies:

$$f(\lambda x + (1 - \lambda)y) = \lambda f(x) + (1 - \lambda)f(y), \forall x, y, \lambda \in [0, 1]$$

2.4 Derivatives

Definition Let $f : R^n \rightarrow R$ be some function, fix some $x \in R^n$ and consider the expression

$$\lim_{t \rightarrow 0} \frac{f(x + te_i) - f(x)}{t}$$

where e_i is the i^{th} unit vector. If the limit exists, it is called the i^{th} partial derivative of f at x , denoted by $\partial f(x) / \partial x_i$

The gradient of f at x is defined as the column vector

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_i} \right)_i$$

For any $y \in R^n$, we define one-sided directional derivative of f in the direction y to be

$$f'(x; y) = \lim_{t \rightarrow 0^+} \frac{f(x + ty) - f(x)}{t}$$

The function f is called differentiable at x if the gradient $\nabla f(x)$ exists and $\nabla f(x)^T y = f'(x; y)$ for every $y \in R^n$

If $f : R^n \rightarrow R^m$ then the gradient matrix of f is the $n \times m$ whose i^{th} column is the gradient $\nabla f_i(x)$ of f_i

$$\nabla f(x) = [\nabla f_1(x) \dots \nabla f_m(x)]$$

The transpose of ∇f (or Df) is called the Jacobian of f and is the matrix whose ij^{th} entry is $\partial f_i / \partial x_j$

Chain rule Let $f : R^k \rightarrow R^m$ and $g : R^m \rightarrow R^n$ be continuously differentiable functions and h be their composition: $h(x) = g(f(x))$, the the chain rule for differentiation states that

$$\partial h(x) = \partial f(x) \partial g(f(x)), \forall x \in R^k$$

Examples:

(a) $\nabla(f(Ax)) = A^T \nabla f(Ax)$

(b) $\nabla^2(f(Ax)) = A^T \nabla^2 f(Ax) A$

(c) $f(x) = \log \sum_{i=1}^m \exp(a_i^T x + b_i)$ then $\nabla f(x) = \frac{1}{\mathbf{1}^T z} A^T z$ where $z = (\exp(a_i^T x + b_i))_{i=1}^m$

Derivative of matrix trace

$$\frac{\partial \text{Tr}(XA)}{\partial x_{ij}} = a_{ji} \Rightarrow \nabla_X \text{Tr}(XA) = A^T$$

Second derivative Suppose each one of the partial derivatives of a function $f : R^n \rightarrow R$ is continuously differentiable function of x , the Hessian of f is the symmetric matrix $[\frac{\partial^2 f}{\partial x_i \partial x_j}(x)]$

Mean value theorem If $f : R \rightarrow R$ is continuously differentiable over an interval I , then $\forall x, y \in I, \exists \xi \in [x, y]$ such that

$$f(y) - f(x) = f'(\xi)(y - x)$$

Cauchy's mean value theorem If $f, g : R \rightarrow R$ is continuously differentiable over an interval I , then $\forall x, y \in I, \exists \xi \in [x, y]$ such that

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(y) - f(x)}{g(y) - g(x)}$$