CS 135

Math Review

1 Vectors and Matrices

We use the notation $x \in \mathbb{R}^n$ to indicate real vectors of size n and $A \in \mathbb{R}^{n \times m}$ to denote matrices of dimensions $n \times m$. Given a vector $x \in \mathbb{R}^n$, we use x_i, $i = 1, \ldots, n$ to denote its i-th coordinate. We treat all vectors in $x \in \mathbb{R}^n$ are *column vectors*, i.e.,:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

1.1 Transposition & Symmetric Matrices

We use the notation $x^\top, A^\top$ to indicate transposition. That is, if $x \in \mathbb{R}^n$, then its transpose $x^\top$ is the row vector:

$$x^\top = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} \in \mathbb{R}^{1 \times n}.$$

Similarly, for

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

its transpose is given by:

$$A^\top = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{12} & a_{22} & \ldots & a_{n2} \\ \vdots & & \ddots & \vdots \\ a_{1m} & a_{2m} & \ldots & a_{nm} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

We say that a square matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if it remains unchanged under transposition, i.e., $A = A^\top$. We denote by

$$\mathbb{S}^n = \{A \in \mathbb{R}^{n \times n} : A = A^\top\}$$

the set of all (real) symmetric matrices.

1.2 Matrix and Vector Multiplication

Given two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$ we use $A \cdot B$ or simply AB to denote the usual *matrix product* between A and B. That is,

$$\begin{aligned} A \cdot B &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1k} \\ b_{21} & b_{22} & \dots & b_{2k} \\ \vdots & & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mk} \end{bmatrix} \\ &= \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \vdots & & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nk} \end{bmatrix} \\ &= C \in \mathbb{R}^{n \times k}, \end{aligned}$$

where[1]

$$c_{ij} = \sum_{\ell=1}^{m} a_{i\ell} b_{\ell j}, \qquad \text{for } i = 1, \dots, n, \text{ and } j = 1, \dots, k.$$

Note that:

$$(AB)^\top = B^\top A^\top.$$

The *inner product* between two vectors $x, y \in \mathbb{R}^n$ can be written as:

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}$$

while the *outer product* is given by:

$$xy^\top = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \dots & x_n y_n \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

2 Multivariate Functions

We use the notation $f : \mathcal{A} \to \mathcal{B}$ to indicate a function that maps elements of set $\mathcal{A}$ to elements in the set $\mathcal{B}$. That is, $f : \mathcal{A} \to \mathcal{B}$ indicates that (a) $f(x)$ is defined over all $x \in \mathcal{A}$, and (b) $f(x) \in \mathcal{B}$. Sets $\mathcal{A}$ and $\mathcal{B}$ are referred to as f's *domain* and *range*, respectively. We list below several examples of real and vector valued functions.

Examples:

- A real function of one variable is denoted by $f : \mathbb{R} \to \mathbb{R}$.

[1] Put differently, the element in the i-th row and j-th column of C is the inner product of the i-th row of A with the j-th column of B.

- A multivariate, real-valued function is denoted by $f : \mathbb{R}^n \to \mathbb{R}$.
- A vector-valued function, mapping vectors of size n to vectors of size m is denoted by $f : \mathbb{R}^n \to \mathbb{R}^m$.

Linear Functions. A *vector-valued* function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *linear* if it satisfies the propery:

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y), \quad \text{for all } x, y \in \mathbb{R}^d, \alpha, \beta \in \mathbb{R}.$$

Equivalently, f is linear if there exists a matrix $A \in \mathbb{R}^{m \times n}$ such that:

$$f(x) = Ax.$$

In particular, function $f : \mathbb{R}^n \to \mathbb{R}$ is linear if there exists a vector $b \in \mathbb{R}^n$ such that:

$$f(x) = b^\top x = \langle b, x \rangle = \sum_{i=1}^{n} b_i x_i.$$

Affine Functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *affine* if is equal to a linear function plus a constant. That is, there exist $b \in \mathbb{R}^n$ and a $c \in \mathbb{R}$ such that:

$$f(x) = b^\top x + c.$$

Similarly, affine vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$ take the form:

$$f(x) = Ax + b, \quad \text{for some } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m.$$

Polynomial and Quadratic Functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a *monomial* if it can be written as the product of integral powers of its arguments times a constant, i.e., it takes the form:

$$f(x) = c \prod_{i=1}^{n} x_i^{k_i},$$

where $c \in \mathbb{R}$ and $k_i \in \mathbb{N}$, for $i = 1, \ldots, n$. The *degree* of the monomial is $k = \sum_{i=1}^{n} k_i$.

A function $f : \mathbb{R}^n \to \mathbb{R}$ that can be written the sum of monomials is called a *polynomial*. The degree of a polynomial is the highest degree among all its monomials. Hence, a polynomial of degree k can be written as

$$f(x) = \sum_{k_1, k_2, \ldots, k_n : \sum_{i=1}^{n} k_i \leq k} c_{k_1, k_2, \ldots, k_n} \prod_{i=1}^{n} x_i^{k_i}.$$

Linear and affine functions are polynomials of degree 1. A polynomial of degree 2 is called a *quadratic function*. Every quadratic function can be written in the following form:

$$f(x) = \frac{1}{2} x^\top Q x + b^\top x + c$$

where $Q \in \mathbb{R}^{n \times n}$ is symmetric matrix, $b \in \mathbb{R}^n$ is a vector, and $c \in \mathbb{R}$ is a scalar constant.

Proof. Suppose that $f(x) = x^\top Ax + b^\top x + c$ where A is not necesarily symmetric Then

$$x^\top Ax = (x^\top Ax)^\top = x^\top A^\top x$$

This implies that

$$x^\top Ax = \frac{1}{2}x^\top(A + A^\top)x$$

In turn, this implies that $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$, for $Q = A + A^\top \in \mathbb{S}^n$. $\square$

3 Vector Norms

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a *norm* if it satisfies the following properties:

- f is *non-negative*: $f(x) \geq 0$ for all $x \in \mathbb{R}^n$.
- f is *definite*: $f(x) = 0$ implies that $x = 0$.
- f is *homogeneous*: $f(tx) = |t|f(x)$, for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$.
- f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$, for all $x, y \in \mathbb{R}^n$.

We use the notation $f(x) = \|x\|$, which is meant to suggest that a norm is a generalization of the absolute value on $\mathbb{R}$. A norm can be thought of as a measure of the length of a vector $x \in \mathbb{R}^n$: if $\|\cdot\|$ is a norm, the distance between two vectors $x, y \in \mathbb{R}^n$ can be measured through

$$\|x - y\|.$$

Examples. The *Euclidian* or ℓ_2-norm is defined as:

$$\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Similarly, the *sum-absolute-value* or ℓ_1-norm is defined as:

$$\|x\|_1 = \sum_{i=1}^n |x_i| = |x_1| + |x_2| + \ldots + |x_n|$$

and the *Chebyshev* or ℓ_∞-norm is defined as:

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_k|\}.$$

More generally, the *Minkowski* or ℓ_p-norm of a vector, for $p \geq 1$, is defined as:

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}.$$

For $p = 1$ and $p = 2$, the Minkowski norm is precisely the ℓ_1 and ℓ_2 norm defined above. The Minkowski norm can be defined for $p \in (0, 1]$ as well; however, for $p \in$

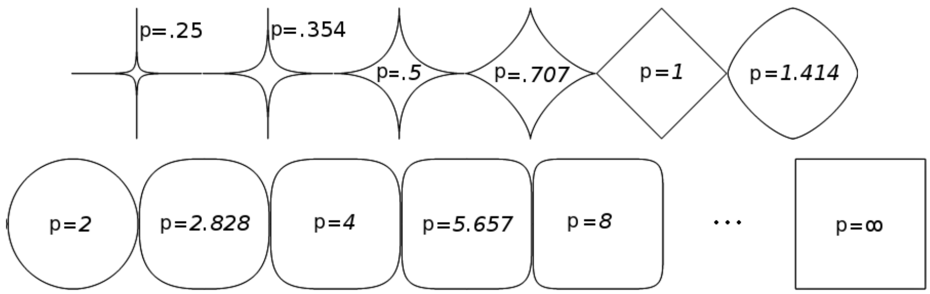


Figure 1: Unit balls in $\mathbb{R}^2$ induced by different Minkowski norms. Source: WikiMedia Commons.

$(0, 1]$, it is strictly speaking *not a norm*, as it *does not satisfy the triangle inequality*. The *unit ball* for a given a norm $\|\cdot\|$ is the set:

$$\{x : \|x\| \leq 1\}$$

An illustration of the unit ball on $\mathbb{R}^2$ induced by different norms can be found in Figure 1. For $p = 2$, the unit ball is a circle (or a sphere, for $n = 3$), while for $p = \infty$ the ball is a square (or a cube, for $n = 3$). The figure also illustrates that, as p tends to ∞, the ℓ_p tends to the ℓ_∞ norm.

All of the above norms over $\mathbb{R}^n$ are *equivalent*; that is, for any two norms $\|\cdot\|_a$, $\|\cdot\|_b$, there exist positive constants $\alpha, \beta \in\in \mathbb{R}_+$ such that:

$$\alpha\|x\|_a \leq \|x\|_b \leq \beta\|x\|_a.$$

This implies that definitions of convergence, function continuity, etc., we present below are not norm-dependent: for example, if a sequence converges to a fixed point with respect to one norm, convergence is indeed implied for all of the above norms.

4 Continuous and Differentiable Functions

4.1 Limits in $\mathbb{R}^n$ and continuity.

A sequence $\{x_k\}_{k=1}^{\infty}$ of vectors in $\mathbb{R}^n$

$$x_1, x_2, x_3, x_4, \ldots$$

converges to a fixed point $x \in \mathbb{R}^n$ w.r.t. a norm $\|\cdot\|_2$ if:

$$\lim_{k\to\infty} \|x_k - x\|_2 = 0.$$

If this is the case, we write

$$\lim_{k\to\infty} x_k = x, \qquad \text{or, simply} \qquad x_k \to x.$$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $x \in \mathbb{R}^n$ if, for any sequence $\{x_k\}_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} x_k = x,$$

we have that:

$$\lim_{k\to\infty} f(x_k) = f(x).$$

We say that a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous if and only if it is continuous at all $x \in \mathbb{R}^n$.

4.2 Gradient

Given $f : \mathbb{R}^n \to \mathbb{R}$, we define the i-th *partial derivative* of f at x is

$$\frac{\partial f(x)}{\partial x_i} \equiv \lim_{\delta\to 0} \frac{f(x+\delta e_i) - f(x)}{\delta},$$

where $e_i \in \mathbb{R}^n$ is a vector with a 1 at coordinate i and zero everywhere else. Note that this naturally generalizes derivatives of functions of one coordinate.

If the limits defining all partial derivatives $\frac{\partial f(x)}{\partial x_i}$ exist, we say that function f is differentiable at $x \in \mathbb{R}^n$. In this case, the *gradient* ∇F of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is the vector of partial derivatives, i.e.:

$$\nabla F(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_i} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Example 1. Show that the gradient of an affine function $f(x) = b^\top x + c$ is the constant function $\nabla F(x) = b \in \mathbb{R}^n$.
Example 2. Show that the gradient of a quadratic function $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ is $\nabla F(x) = Qx + b \in \mathbb{R}^n$.

The Taylor expansion of f at a point x_0 is given by:

$$f(x) = f(x_0) + (\nabla F(x))^\top (x - x_0) + o(\|x - x_0\|_2)$$

Hence, the affine function

$$\hat{f}(x) = f(x_0) + (\nabla F(x))^\top (x - x_0) \tag{1}$$

above approximates the function f near x. Setting $z = \hat{f}(x)$, (1) can be written as the following vector inner product:

$$\begin{bmatrix} z - f(x_0) & ; & (x - x_0)^\top \end{bmatrix} \begin{bmatrix} -1 \\ \\ \nabla f(x_0) \end{bmatrix} = 0$$

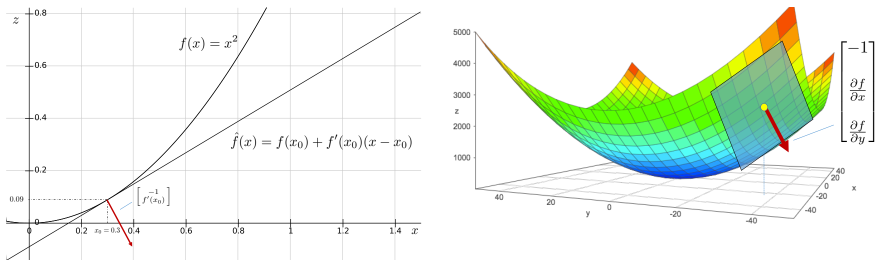


Figure 2: First order Taylor approximation of function of 1 variable and 2 variables. For $f : \mathbb{R} \to \mathbb{R}$, the approximation is forms a line; in higher dimensions, it forms a *hyperplane*.

In other words, Eq. (1) defines a hyperplane of points

$$\begin{bmatrix} z \\ x \end{bmatrix} \in \mathbb{R}^{n+1}$$

that passes through point

$$\begin{bmatrix} f(x_0) \\ x_0 \end{bmatrix} \in \mathbb{R}^{n+1}$$

and whose normal is given by

$$\begin{bmatrix} -1 \\ \nabla f(x_0) \end{bmatrix} \in \mathbb{R}^{n+1}.$$

This is illustrated in Figure 2.

Figure 3 gives further intuition on the physical meaning of the gradient. The gradient at $x_0 \in \mathbb{R}^d$ perpendicular to the contour defined by

$$\{x \in \mathbb{R}^d : f(x) = f(x_0)\}$$

Moreover, $\nabla f(x_0)$ indicates the direction of *steepest ascent*: following the gradient leads to the largest possible increase of f in the vicinity of x_0.

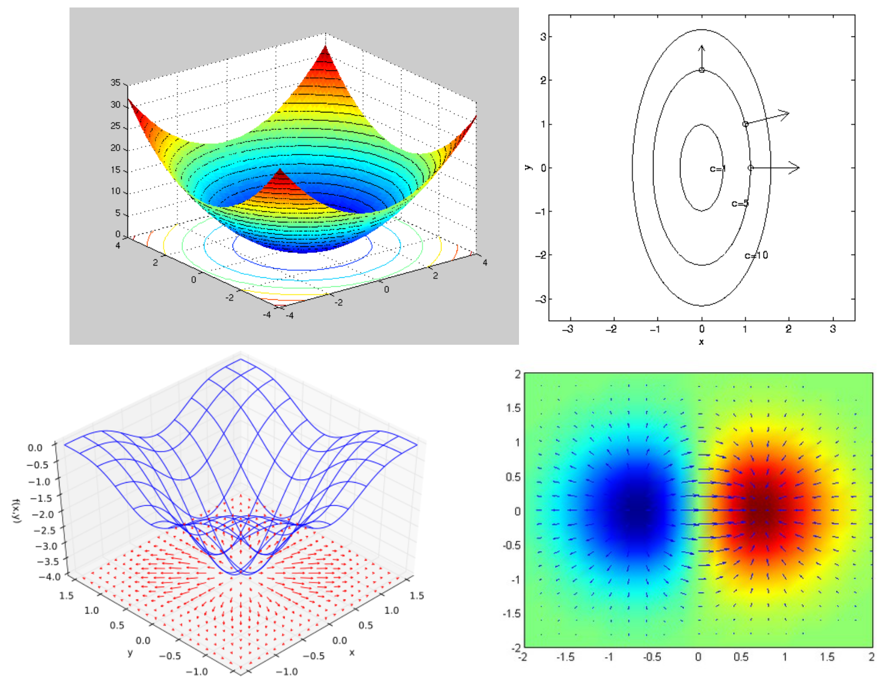
4.3 Hessian

The *Hessian* $\nabla^2 f$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ at point $x \in \mathbb{R}^n$is defined as the $n \times n$ symmetric matrix whose elements are:

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad \text{for } i, j \in \{1, \ldots, n\}$$

The second order Taylor approximation of f at $x_0 \in \mathbb{R}^n$ is then given by:

$$\hat{f}(x) = f(x_0) + (x - x_0)^\top \nabla f(x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0)$$



$f(x, y) = -(\cos^2 x + \cos^2 y)^2$ $\qquad$ $f(x, y) = xe^{-(x^2+y^2)}$

Figure 3: Drawing the *level* or *contour curves* of f on $\mathbb{R}^n$ gives further intuition into what ∇f means. Projecting the normal of the hyperplane tangent to f on $\mathbb{R}^n$, we see that $\nabla f(x_0)$ is always perpendicular to the corresponding level curve that passes through x_0, and points to a direction in which f increases; if fact, it is the direction of steepest ascent. Sources for pictures on the top: 1,2, bottom figures from WikiMedia Commons.

5 Linear Algebra

5.1 Matrix Inverse

The inverse of a square matrix $A \in \mathbb{R}^{n\times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n\times n}$, s.t.:

$$AA^{-1} = A^{-1}A = I$$

If such a matrix exists, then A is called invertible. A matrix is invertible if and only if its determinant $\det(A)$ is non-zero.

5.2 Spectral Decomposition of Symmetric Matrices

A vector $e \in \mathbb{R}^n$, where $\|e\|_2 = 1$, and a scalar λ are called an *eigenvector* and *eigenvalue* of a symmetric matrix A, respectively, if

$$Ae = \lambda e.$$

Any symmetric matrix $A \in \mathbb{S}^n$ can be written as:

$$A = Q\Lambda Q^\top \tag{2}$$

where $Q \in \mathbb{R}^{n\times n}$ is *orthogonal*, i.e., it satisfies:

$$Q^\top Q = QQ^\top = I,$$

and

$$\Lambda = \texttt{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

is a diagonal matrix, in which $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ The columns of Q constitute the eigenvectors of A, i.e.,

$$Q = [e_1, e_2, \ldots, e_n] \in \mathbb{R}^{n\times n},$$

while the values λ_i, $i = 1, \ldots, n$ are the corresponding eigenvalues. Eq. (2) is known as the *spectral* or *eigen* decomposition of matrix A. It also implies that

$$A = \sum_{i=1}^{n} \lambda_i e_i e_i^\top,$$

i.e., A can be written as the weighted sum of the outer products of its eigenvectors. We usually denote the maximum eigenvalue of A as $\lambda_{\max}(A) = \lambda_1$, and its minimum eigenvalue as $\lambda_{\min}(A) = \lambda_n$.

The determinant of A and the trace of A relate to its eigenvalues as follows:

$$\det(A) = \prod_{i=1}^{n} \lambda_i, \quad \mathrm{trace}(A) = \sum_{i=1}^{n} \lambda_i.$$

Hence, a symmetric matrix A is invertible if and only if none of its eigenvalues is zero. Such a matrix is also known as *full-rank*: all its rows are linearly independent. If a matrix A is invertible, then the eigenvalues of its inverse A^{-1} are

$$\lambda_i' = \frac{1}{\lambda_i}, \quad i = 1, \ldots, n$$

Example 1. Given $A \in \mathbb{S}^n$ with eigenvalues λ_i, $i = 1, \ldots, n,$, the matrix $\lambda I + A$ has the same eigenvectors as A, and its corresponding eigenvalues are $\lambda_i + \lambda$, $i = 1, \ldots, n$. To see this, note that if e_i is an eigenvector of A, then

$$(\lambda I + A)e_i = \lambda I e_i + A e_i = \lambda e_i + \lambda_i e_i = (\lambda + \lambda_i)e_i.$$

5.3 Positive Definite and Positive Semi-Definite Matrices

A symmetric matrix $A \in \mathbb{S}^n$ is called *positive semi-definite* (PSD) if:

$$x^\top A x \geq 0, \qquad \text{for all } x \in \mathbb{R}^n.$$

A symmetric matrix is called *positive-definite* (PD) if:

$$x^\top A x > 0, \qquad \text{for all } x \in \mathbb{R}^n \setminus \{0\}.$$

Equivalently, a matrix is PSD if and only if all its eigenvalues are *non-negative*, i.e.,

$$\lambda_{\min}(A) \geq 0.$$

Similarly, a matrix is PD if and only if all its eigenvalues are *positive*, i.e.,

$$\lambda_{\min}(A) > 0.$$

We write $A \succeq 0$ and $A \succ 0$ to indicate that A is PSD or PD, respectively. We also use the notation

$$\mathbb{S}_+^n = \{A \in \mathbb{S}^n, A \succeq 0\}, \qquad \mathbb{S}_{++}^n = \{A \in \mathbb{S}^n, A \succ 0\},$$

to indicate the sets of PSD and PD matrices, respectively.

Example 1. Given any vector $z \in \mathbb{R}^d$, the matrix $A = zz^\top$ defined by the outer product of z with itself is positive semidefinite. Indeed, for any $x \in \mathbb{R}^n$,

$$x^\top A x = x^\top (zz^\top) x = (x^\top z)(z^\top x) = (x^\top z)^2 \geq 0.$$

Example 2. Given two PSD matrices $A, B \succeq 0$, and two non-negative scalars $\alpha, \beta \geq 0$, $\alpha A + \beta B \succeq 0$. Indeed, for any $x \in \mathbb{R}^n$,

$$x^\top (\alpha A + \beta B) x = \alpha x^\top A x + \beta x^\top B x \geq 0.$$

Example 3. For any matrix $Y \in \mathbb{R}^{n \times m}$, the matrix $A = Y^\top Y \in \mathbb{S}^m$ is PSD. To see this, note that

$$A = Y^\top Y = \sum_{i=1}^m y_i y_i^T,$$

where y_i is the i-th row of Y. Positive semidefiniteness therefore follows from Examples 1 and 2.

Example 4. If $A \in \mathbb{S}^n$, and $\lambda_{\min}(A) < 0$, then $\lambda I + A \succeq 0$ for $\lambda = |\lambda_{\min}(A)|$. This follows from Example 1 in Sec. **??**.

6 Further Reading

See Boyd and Vandenberghe [1], Appendix A, pp. 633–652.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.
https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf