Lecture Notes of Linear Algebra and Convex Optimization (Dr. Jun Qi)

Linear Algebra

1. PSD matrix

A positive semi-definite (PSD) matrix is a symmetric matrix $A \in \mathbb{R}^{n \times n}$ for which the quadratic form $x^T A x$ is non-negative for all vectors x. This means that for any vector x, the expression $x^T A x \geq 0$.

Key properties of PSD:

Symmetry: a PSD A is always symmetric, i.e., $A = A^T$.

Non-negative Eigenvalues: All eigenvalues of a PSD matrix are non-negative.

Eigenvalue Decomposition: $A = Q \Lambda Q^T$. (Q is an orthogonal matrix of eigenvectors; Λ is a diagonal matrix of eigenvalues).

Square Root Construction: The principal square root $A^{1/2}$ can be constructed using the eigenvalue decomposition $A^{1/2} = Q \Lambda^{1/2} Q^T$ ($\Lambda^{1/2}$ is the diagonal matrix obtained by taking the square roots of the non-negative eigenvalues of A).

Lemma 1.1: a PSD matrix A can be expressed as $A = A^{1/2} A^{1/2}$ ($A^{1/2}$ is known as the principal square root of A)

Proposition 1.2: a PSD matrix A has a multiset of eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ and orthonormal eigenvectors $\{a_1, a_2, \cdots, a_n\}$, then $A = \sum_{i=1}^{n} \lambda_i a_i a_i^T$ ($P_i = a_i a_i^T$ are orthogonal projectors and the eigenvectors $\{a_i\}_i$ form an orthonormal basis of the vector space.

2. SVD

Singular Value Decomposition (SVD) decomposes a matrix $A \in \mathbb{R}^{m \times n}$ into three other matrices, capturing essential geometric and algebraic properties of A. the SVD of an $m \times n$ matrix A is given by:

$$A = U \Sigma V^T.$$

U: $n \times s$ orthogonal matrix. The columns of U are called the left singular vectors of A.

Σ: $s \times s$ diagonal matrix with non-negative real numbers on the diagonal. these values are called the singular values of A, typically ordered in a descending order. If A is rank r, then Σ contains r positive singular values, with remaining entries being zero.

V^T: the transpose of an $m \times s$ orthogonal matrix V. The columns of V (or rows of V^T) are called the right singular vectors of A.

Practical Applications: Dimensionality Reduction (low-rank approximation) and Data Compression.

3. Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times p}$ is the matrix.

$$C = AB \in \mathbb{R}^{n \times p}.$$

where $C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}$. the number of Columns in A must equal the number of rows in B.

Vector-Vector Products (inner product / dot product)

$$x^T y \in \mathbb{R} = [x_1, x_2, \cdots, x_n] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

Outer Product of vectors.

$$x y^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} [y_1, y_2, \cdots, y_n] = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & & x_m y_n \end{bmatrix}$$

Matrix–Vector Products

If we write A in column form, we see that

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} x_1 + \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} x_2 + \cdots + \begin{bmatrix} | \\ a_n \\ | \end{bmatrix} x_n .$$

In other words, y is a linear combination of columns of A, where the coefficients of the linear combination are given by the entries of x.

Matrix–Matrix Products

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T \qquad \text{(represent } A \text{ by columns and } B \text{ by rows)}$$

$$= A \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix} \qquad \text{(represent } B \text{ by columns)}$$

$$= \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} B = \begin{bmatrix} - & a_1^T B & - \\ - & a_2^T B & - \\ & \vdots & \\ - & a_m^T B & - \end{bmatrix} \qquad \text{(represent } A \text{ by rows)}$$

Exercise: When $a_i^T B$ represents the sum of elements in the i-row, what's B?

If $a_i^T B = 1$, what's it implied?

Ans: ① $B = \mathbb{1} \in \mathbb{R}^m$

② It implies the one vector is the eigenvector of A, and the eigenvalue is 1.

4. trace

The trace of a square matrix $A \in \mathbb{R}^{n \times n}$. $\mathrm{Tr}(A) = \sum_{i=1}^{n} A_{ii}$.

key properties:

- For $A \in \mathbb{R}^{n \times n}$. $\mathrm{Tr}(A) = \mathrm{Tr}(A^T)$.
- For $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$.
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$. $\mathrm{Tr}(tA) = t\,\mathrm{Tr}(A)$.
- For A, B such that AB is square. $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.
- For A, B, C such that ABC is square, $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB)$.

5. Norms (optional)

Definition 5.1: A mapping $\Phi : \mathcal{H} \to \mathbb{R}_+$ is said to define a norm on $\mathcal{H}$ if it verifies the following axioms:

- definiteness: $\forall x \in \mathcal{H}$, $\Phi(x) = 0 \Leftrightarrow x = 0$;
- homogeneity: $\forall x \in \mathcal{H}$, $\forall \alpha \in \mathbb{R}$, $\Phi(\alpha x) = |\alpha| \Phi(x)$;
- triangle inequality: $\forall x, y \in \mathcal{H}$, $\Phi(x+y) \leq \Phi(x) + \Phi(y)$.

A norm is typically denoted by $\|\cdot\|$.

More generally, for any $p \geq 1$, the L_p-norm is defined on $\mathbb{R}^n$ as:

$$\forall x \in \mathbb{R}^n, \quad \|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{\frac{1}{p}}.$$

The L_1, L_2, and L_∞ norms are some of the most commonly used norms, where $\|x\|_\infty = \max_{j \in [N]} |x_j|$. The following general inequalities relating these norms can be proven straightforwardly:

6. Dual Norms (optional)

Definition 6.1: Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, then, the dual norm $\|\cdot\|_*$ associated to $\|\cdot\|$ is the norm defined by

$$\forall y \in \mathbb{R}^n, \quad \|y\|_* = \sup_{\|x\|=1} |\langle y, x \rangle|.$$

For any $p, q \geq 1$ that are conjugate that $\frac{1}{p} + \frac{1}{q} = 1$, the L_p and L_q norms are dual norms of each other. For example, the dual norm of L_2 is the L_2-norm, and the dual norm of the L_1-norm is the L_∞-norm.

Proposition 6.2 (Holder's inequality) Let $p, q \geq 1$ be conjugate: $\frac{1}{p} + \frac{1}{q} = 1$. Then, for all $x, y \in \mathbb{R}^n$.

$$|\langle x, y \rangle| \leq \|x\|_p \cdot \|y\|_q.$$

with equality when $|y_i| = |x_i|^{p-1}$ for all $i \in [n]$.

7. Matrix Norms (optional)

the matrix norm induced by the vector norm $\|\cdot\|_p$ or the operator norm induced by that norm is also denoted by $\|\cdot\|_p$ and denoted by: $\|M\|_p = \sup_{\|x\|_p \leq 1} \|Mx\|_p$.

the norm induced for $p=2$ is known as the spectral norm, which equals the largest singular value of M.

$$\|M\|_2 = \sigma_1(M) = \sqrt{\lambda_{max}(M^T M)}.$$

The Frobenium norm denoted by $\|\cdot\|_F$ is the most notable of such norms and is defined by:

$$\|M\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij}^2\right)^{1/2}$$

Frobenium product: $M, N \in \mathbb{R}^{n \times m}$, $\langle M, N \rangle_F = \mathrm{Tr}(M^T N)$

this relates the Frobenius norm to the singular values of M:

$$\|M\|_F^2 = \mathrm{Tr}(M^T M) = \sum_{i=1}^{r} \sigma_i^2(M), \quad \text{where } r = \mathrm{rank}(M)$$

8. Linear Independence

A set of vectors $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^m$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set can be represented as a linear combination of the remaining vectors, then the vectors are said to be (linearly) dependent. That is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i.$$

for some scalar values $\alpha_1, \alpha_2, \ldots \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors $x_1, x_2, \ldots, x_n$ are linearly dependent.

Exercise: Given $Ax_1 = c_1$, $Ax_2 = c_2$, $Ax_3 = c_3$, c_1, c_2 and c_3 are linearly independent in $\mathbb{R}^3$. then what is $Ax_4 = ?$

Ans: $c_4 = Ax_4 = \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3 = A(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3)$, where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$.

Exercise: Could a randomized always generate a set of linear independent vectors?

Ans: When vectors are initialized randomly, especially over a sufficiently large field of range, the likelihood of picking vectors that are linearly dependent is low because the conditions for linear dependence are very specific.

9. The Inverse

The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1}, and is the unique matrix such that

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

Note that not all matrices have inverses. Non-square matrices, for example, do not have inverse by definition. For some square matrices A, it may still be the case that A^{-1} may not exist. In particular, we say that A is invertible or non-singular if A^{-1} exists and non-invertible or singular otherwise.

In order for a square matrix A to have an inverse A^{-1}, then A must be full rank.

In particular, let's assume $A^{-1} = [\,c_1\ c_2 \cdots c_n\,]$. then $AA^{-1} = [\,Ac_1\ Ac_2 \cdots Ac_n\,] = [\,e_1\ e_2 \cdots e_n\,]$.

10. Orthogonal matrices

Two vectors $x, y \in \mathbb{R}^n$ are orthogonal if $x^Ty = 0$. A vector $x \in \mathbb{R}^n$ is normalized if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized. (the columns are referred to as being orthonormal).

$$U^TU = I = UU^T$$

key property: Operating on a vector with an orthogonal matrix does not change its Euclidean norm, i.e.,

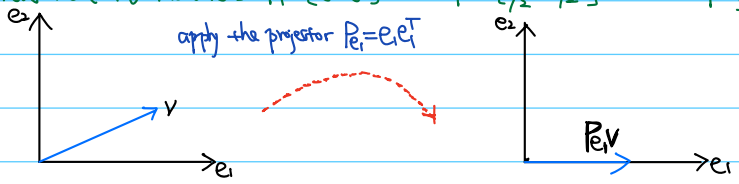
$$\|Ux\|_2 = \|x\|_2, \quad \forall x \in \mathbb{R}^n, \ U \in \mathbb{R}^{n \times n} \text{ orthogonal}.$$

11. Projection Matrix

A matrix P is a projection matrix if $P^2 = P$. Given a set of orthonormal vectors $e_1, e_2, \cdots, e_n$, a projection matrix onto the subspace $W = \text{span}(e_1, e_2, \cdots, e_n)$ is given by

$$P_W = \sum_{i=1}^{n} e_i e_i^T.$$

Exercise: Show that the matrices $P_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $P_2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ are both projection matrices.



12. Eigenvalues and Eigenvectors.

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of A and $x \in \mathbb{C}^n$ is the corresponding eigenvector if

$$Ax = \lambda x, \quad x \neq 0.$$

We can rewrite the equation above to state that (λ, x) is an eigenvalue–eigenvector pair of A, if,

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

But $(\lambda I - A)x = 0$ has a non-zero solution to x if and only if $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e.,

$$|\lambda I - A| = 0.$$

Key properties:

- The trace of a A is equal to the sum of its eigenvalues. $\text{Tr}(A) = \sum_{i=1}^{n} \lambda_i$
- The determinant of A is equal to the product of its eigenvalues, $|A| = \prod_{i=1}^{n} \lambda_i$
- The rank of A is equal to the number of non-zero eigenvalues of A.
- If A is non-singular, then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector x_i, i.e., $A^{-1}x_i = (1/\lambda_i)x_i$.
- We can write all the eigenvector equations simultaneously as $AX = X\Lambda$.

Q: Given A, can you compute the eigenvalues λ_i?

Convex Optimization

13. the Gradient

Suppose that $f: \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $n\times m$ and return a real value. Then, the gradient of f (with respect to $A \in \mathbb{R}^{m\times n}$) is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1m}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{n1}} & \frac{\partial f(A)}{\partial A_{n2}} & \cdots & \frac{\partial f(A)}{\partial A_{nm}} \end{bmatrix}$$

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. If A is a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

the 1st-order gradient of f is a n-dimensional vector.

14. Hessian matrix

the Hessian matrix wrt. x, written $\nabla_x^2 f(x)$ or simply as H is the $n\times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

the 2nd-order gradient of f is a $n\times n$ matrix

Exercise: Calculate the Hessian matrix of $f(x_1, x_2, \cdots, x_n) = \sum_{i=1}^n e^{x_i}$ and verify the obtained Hessian is positive semi-definite.

Ans: $\nabla f = [e^{x_1}, e^{x_2}, \cdots, e^{x_n}]^T$. $\nabla^2 f = \operatorname{diag}(e^{x_1}, e^{x_2}, \cdots, e^{x_n}) \in \mathbb{R}^{n\times n}$ (Hessian Matrix)

thus, $\forall z \in \mathbb{R}^n$, $z^T \nabla^2 f z = \sum_{i=1}^n e^{x_i} z^T z \geq 0 \Rightarrow \nabla^2 f$ is PSD $\Rightarrow f$ is a convex function.

15 Convex set

Definition 15.1: A set $\mathcal{X} \subseteq \mathbb{R}^n$ is said to be convex if for any two points $x, y \in \mathcal{X}$, the segment $[x, y]$ lies in $\mathcal{X}$, that is

$$\{\partial x + (1-\partial) y : 0 \leq \partial \leq 1\} \subseteq \mathcal{X}$$

Definition 15.2 (Convex hull) the convex hull $\operatorname{Conv}(\mathcal{X})$ of a set of points $\mathcal{X} \subseteq \mathbb{R}^n$ is the minimal convex set containing $\mathcal{X}$ and can be equivalently defined as follows:

$$\operatorname{Conv}(\mathcal{X}) = \left\{ \sum_{i=1}^n \partial_i x_i : n \geq 1, \forall i \in [n], x_i \in \mathcal{X}, \partial_i \geq 0, \sum_{i=1}^n \partial_i = 1 \right\}.$$

'6 Convex function

Definition 16.1. Let $\mathcal{X}$ be a convex set. A function $f: \mathcal{X} \to \mathbb{R}$ is said to be convex iff $\text{Epi}(f)$ is a convex set, or equivalently, if for all $x, y \in \mathcal{X}$ and $\theta \in [0,1]$,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

f is said to be concave when $-f$ is convex.

Theorem 16.2. Let f be a differentiable function, then f is convex if and only if $\text{dom}(f)$ is convex and the following inequalities hold:

$$\forall x, y \in \text{dom}(f), \quad f(y) - f(x) \geq \nabla f(x)^T (y - x).$$

Theorem 16.3. Let f be a twice differentiable function, then f is convex iff $\text{dom}(f)$ is convex and its Hessian is positive semidefinite: $\forall x \in \text{dom}(f), \ \nabla^2 f(x) \succeq 0$.

Example (Quadratic function). The function $f(x) = x^2$ defined over $\mathbb{R}$ is convex since it is twice differentiable and for all $x \in \mathbb{R}$, $f''(x) = 2 > 0$.

Example (Norms). Any norm $\|\cdot\|$ defined over a convex set $\mathcal{X}$ is convex since by the triangle inequality and the homogeneity property of the norm, for all $\theta \in [0,1]$, $x, y \in \mathcal{X}$, we can write

$$\|\theta x + (1-\theta)y\| \leq \|\theta x\| + \|(1-\theta)y\| = \theta \|x\| + (1-\theta)\|y\|.$$

Proposition 16.4. For an unconstrained optimization problem $\min_{x \in \mathbb{R}^n} f(x)$, the vector x^* minimizes f over $\mathbb{R}^n$ if and only if

$$\nabla f(x) = 0.$$

'7 Constrained optimization problem.

Definition 17.1. Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $f, g_i : \mathcal{X} \to \mathbb{R}$, for all $i \in [n]$, then, a constrained optimization problem has the form:

$$\min_{x \in \mathcal{X}} f(x)$$
$$\text{s.t.}, \ g_i(x) \leq 0, \ \forall i \in [n].$$

Definition 17.2. The Lagrangian function or Lagrangian associated to the general constrained optimization problem is the function defined over $\mathcal{X} \times \mathbb{R}_+$ by:

$$\forall x \in \mathcal{X}, \ \forall \alpha \geq 0, \quad L(x, \alpha) = f(x) + \sum_{i=1}^{n} \alpha_i g_i(x)$$

where the variables α_i are known as the Lagrangian or dual variables with $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$.

Definition 17.3. The (Lagrangian) dual function associated to the constrained optimization problem is defined by

$$\forall \alpha \geq 0, \quad F(\alpha) = \inf_{x \in \mathcal{X}} L(x, \alpha) = \inf_{x \in \mathcal{X}} \left(f(x) + \sum_{i=1}^{n} \alpha_i g_i(x) \right)$$

Note that F is always concave (the Lagrangian is linear w.r.t. α and the infimum preserves concavity.

We further observe that $\forall \alpha \geq 0, \ F(\alpha) \leq p^* = f(x^*), \ x^* \in \mathcal{X}. \ \left(\forall x \in \mathcal{X}. \ f(x) + \sum_{i=1}^{n} \alpha_i g_i(x) \leq f(x) \right)$.

Definition 7.4 (Dual Problem) The dual optimization problem associated to the constrained optimization problem is

$$\max_{\lambda} F(\lambda)$$
$$\text{s.t: } \lambda \geq 0.$$

The dual problem is always a convex optimization problem (as a maximization of a concave problem). Let d^* denote its optimal value. There is $d^* \leq p^*$ (weak duality).

The difference $(p^* - d^*)$ is called the duality gap. The equality case

$$d^* = p^* \quad \text{(strong duality)}.$$

does not hold in general. However, strong duality holds when the constrained convex optimization problem is conceived.

8. Conjugate function. Definition 8.1. For a given function $f: \mathbb{R}^n \to \mathbb{R}$, its conjugate function $f^*: \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \left(\langle y, x \rangle - f(x) \right)$$

Key property: the conjugate function is always a convex function, regardless of whether the original function is convex.

Proof: the supremum of affine functions $(\langle y, x \rangle - f(x))$ is affine in y when x is fixed) is always convex.