COMP4680 Notes

Jeff Li

2023 S2

0 Introduction

A mathematical optimisation problem requires us to minimize an objective function $f_0(x)$, subject to constraint functions $f_i(x) \leq b_i$ for $i = 1, \ldots, m$.

A solution, or optimal point, x^*, has the smallest value of f_0 among all vectors that satisfy the constraints.

There are three main types of problems which can be solved to different extents:

0.1 Least-squares Problems

Least-squares problems attempt to minimize $\|Ax - b\|_2^2$ for some matrix A and vector b.

There exists an analytical solution $x^* = (A^T A)^{-1} A^T b$, there are reliable and efficient algorithms with computation time $O(n^2 k)$ when $A \in \mathbb{R}^{k \times n}$, less if A satisfies certain structures.

0.2 Convex Optimization Problems

Convex optimization problems are optimization problems where both the objective and constraint functions are convex, and is a superset of least-squares problems.

There is no analytical solution, but there are algorithms with computation time $\max(O(n^3), O(n^2 m), F)$ where F is the cost of evaluating f_i's and their second derivatives.

0.3 Nonconvex Optimization Problems

There is no general way to solve nonconvex optimization problems: they all involve some kind of compromise.

We may use local optimization methods (nonlinear programming), which is fast and finds a local minima around an initial guess, but may not be the global minima.

Or we may use global optimization methods, which finds the global solution but requires exponential time complexity.

1 Preliminaries

1.1 Sets

A set, denoted as $S = \{a_1, \ldots, a_n\}$, is a collection of distinct objects.

Some common notations:

- $a \in S$ denotes a is an element of S
- $S \subseteq T$ denotes S is a subset of T, that is, every element of S is also an element of T
- $S \cup T$ denotes the union of S and T, that is, the set of all elements that are in S or T
- $S \cap T$ denotes the intersection of S and T, that is, the set of all elements that are in both S and T
- $S \times T$ denotes the Cartesian product of S and T, that is, the set of all ordered pairs (s, t) where $s \in S$ and $t \in T$
- $S \setminus T$ denotes the set difference of S and T, that is, the set of all elements that are in S but not in T

Some common sets:

- $\mathbb{R}$ is the set of real numbers
- $\mathbb{R}^n$ is the set of n-dimensional real vectors
- $\mathbb{R}^{m \times n}$ is the set of $m \times n$ real matrices
- $\mathbb{C}$ is the set of complex numbers
- $\mathbb{Z}$ is the set of integera
- $\mathbb{R}_+$ is the set of nonnegative real numbers
- $\mathbb{R}_{++}$ is the set of positive real numbers
- $\emptyset$ is the empty set
- $[a, b]$ is the closed interval from a to b (i.e. $\{x \in \mathbb{R} \mid a \leq x \leq b\}$)
- (a, b) is the open interval from a to b (i.e. $\{x \in \mathbb{R} \mid a < x < b\}$)
- $[a, b)$ and $(a, b]$ are half-open intervals, defined similarly

Open and Closed Sets

A subset $S \subseteq \mathbb{R}$ is **open** if for every $x \in S$, there exists $\epsilon > 0$ such that if $\|y - x\|_2 < \epsilon$, then $y \in S$.

A subset $S \subseteq \mathbb{R}$ is **closed** if its complement $\mathbb{R} \setminus S$ is open.

A subset $S \subseteq \mathbb{R}$ is **bounded** if there exists $M > 0$ such that $\|a - b\|_2 \leq M$ for all $a, b \in S$.

Infimum and Supremum

The **infimum** of a set $S \subseteq \mathbb{R}$, written as $\inf(S)$, is the largest $y \in \mathbb{R}$ such that $y \leq x$ for all $x \in S$. If no such y exists, we say $\inf(S) = -\infty$.

The **supremum** of a set $S \subseteq \mathbb{R}$, written as $\sup(S)$, is the smallest $y \in \mathbb{R}$ such that $y \geq x$ for all $x \in S$. If no such y exists, we say $\sup(S) = \infty$.

We define $\inf(\emptyset) = \infty$ and $\sup(\emptyset) = -\infty$.

1.2 Functions

A function $f : A \to B$ is a mapping from its **domain** A to its **codomain** B.

If $U \subseteq A$ and $V \subseteq B$, we define the **image** of U under f as $f(U) = \{f(x) \mid x \in U\} \subseteq B$, and the **preimage** of V under f as $f^{-1}(V) = \{x \in A \mid f(x) \in V\} \subseteq A$.

1.3 Vector Spaces

A vector space V is a set with two operations, vector addition and scalar multiplication, that satisfy the following axioms:

- $x + y = y + x$ (commutativity of vector addition)
- $(x + y) + z = x + (y + z)$ (associativity of vector addition)
- $x + \mathbf{0} = x$ (additive identity)
- $\forall x \in V, \exists y \in V$ such that $x + y = 0$, we write y as $-x$ (additive inverse)
- $\alpha(x + y) = \alpha x + \alpha y$ (right distributivity)
- $(\alpha + \beta)x = \alpha x + \beta x$ (left distributivity)
- $1x = x$ (multiplicative identity)

We define the **zero vector** as a vector with all elements equal to 0, and the **ones vector** as a vector with all elements equal to 1.

Euclidean Norm

The Euclidean norm of a vector $\mathbf{v} = (v_1, \ldots, v_n)$ is

$$\|\mathbf{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + x_n^2}$$

$\|\mathbf{v}\|_2$ measures the length of $\mathbf{v}$.

The norm satisfies:

- $\|\alpha \mathbf{v}\| = |\alpha| \|\mathbf{v}\|$

- $\|\mathbf{u}+\mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ (triangle inequality)
- $\|\mathbf{v}\| \geq 0$ and $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$ (positive definiteness)

There are other norms such as $\| \quad \|_1$ and $\| \quad \|_\infty$.

Inner Products

The inner product of two vectors $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ is defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n.$$

The inner product satisfies:

- $\langle \alpha \mathbf{u}, \mathbf{v} \rangle = \alpha \langle \mathbf{u}, \mathbf{v} \rangle$
- $\langle \mathbf{u}_1 + \mathbf{u}_2, \mathbf{v} \rangle = \langle \mathbf{u}_1, \mathbf{v} \rangle + \langle \mathbf{u}_2, \mathbf{v} \rangle$
- $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
- $\langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|_2^2$

Subspaces

A subspace of a vector space is a subset of the vector space that is also a vector space.

Independence

A set of vectors $v_1, \ldots, v_n$ is (linearly) independent if and only if $\alpha_1 v_1 + \cdots + \alpha_n v_n = 0$ implies $\alpha_1 = \cdots = \alpha_n = 0$.

Conversely, if a set of vectors is linearly dependent, we can write one of the vectors as a linear combination of the others.

Bases

The set of vectors $\{v_1, \ldots, v_n\}$ form a basis of a vector space V if

- they are linearly independent
- they span V, that is, every vector in V can be written as a linear combination of the vectors in the set

Equivalently, $\{v_1, \ldots, v_n\}$ form a basis for V if every $v \in V$ can be uniquely expressed as $v = \alpha_1 v_1 + \cdots + \alpha_n v_n$.

We define the **dimension** of a vector space V to be the number of vectors in any basis of V.

The standard basis of $\mathbb{R}^n$ is the set of vectors $\{e_1, \ldots, e_n\}$ where e_i is the vector with a 1 in the i^{th} position and 0 elsewhere.

1.4 Matrices

A matrix $A \in \mathbb{R}^{m \times n}$ is a rectangular array of real numbers with m rows and n columns.

We write A_{ij} for the entry in the i^{th} row and j^{th} column of A.

A $n \times 1$ matrix is called a (column) **vector**, and a $1 \times n$ matrix is called a row **vector**.

We say a matrix is **diagonal** if its nonzero entries are all on the main diagonal (top left to bottom right).

The **zero matrix**, denoted $\mathbf{0}_{m \times n}$, is the matrix with all entries equal to zero.

The **identity matrix**, denoted $\boldsymbol{I}_n$, is the $n \times n$ matrix with ones on the main diagonal and zeros elsewhere.

Special Types of Matrices

A matrix is **triangular** if all its entries above or below the main diagonal are zero. In particular, we refer to a matrix as **upper triangular** if all its entries below the main diagonal are zero, and **lower triangular** if all its entries above the main diagonal are zero.

A matrix is **block diagonal** if it is diagonal and each diagonal entry is itself a matrix.

A matrix is **tri-diagonal** if it has nonzero entries only on the main diagonal and the diagonals immediately above and below the main diagonal.

Matrix Transpose

Transpose, denotes as ^{T}, flips a matrix over is main diagonal, i.e. if A is an $m \times n$ matrix then A^T is an $n \times m$ matrix. It satisfies the following properties:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

If a matrix A satisfies $A = A^T$ we say A is **symmetric**.

If a matrix A satisfies $A = -A^T$ we say A is **anti-symmetric**.

Every square matrix A can be written as the sum of a symmetric part and an anti-symmetric part:

$$A = \underbrace{\frac{1}{2}\left(A + A^T\right)}_{\text{symmetric}} + \underbrace{\frac{1}{2}\left(A - A^T\right)}_{\text{anti-symmetric}}$$

Matrix Addition

Two matrices of the same size can be added together: we simply add the corresponding elements in each matrix.

Matrix Multiplication

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

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Matrix multiplication satisfies:

- $(AB)C = A(BC)$ (associativity)
- $A(B + C) = AB + AC$ (left distributivity)
- $(A + B)C = AC + BC$ (right distributivity)

but matrix multiplication is not commutative: $AB \neq BA$ generally.

Null Space

The null space of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as

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

$\mathcal{N}(A)$ can be interpreted as

- the set of all vectors mapped to zero by $y = Ax$
- the set of all vectors orthogonal to the rows of A

Range Space

The range space of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m.$$

$\mathcal{R}(A)$ can be interpreted as

- the set of all vectors that can be "hit" by $y = Ax$
- the span of the columns of A
- the set of all vectors y such that $Ax = y$ has a solution

Orthogonal Complement

The orthogonal complement of $V \subseteq \mathbb{R}^n$ is defined as

$$V^{\perp} = \left\{x \mid z^T x = 0 \text{ for all } z \in V\right\}.$$

We have $V \oplus V^{\perp} = \mathbb{R}^n$.

A result from the Fundamental Theorem of Linear Algebra states that $\mathcal{N}(A) = \mathcal{R}(A^T)^{\perp}$.

Rank

The rank of a matrix $A \in \mathbb{R}^{m \times n}$ is

$$\operatorname{rank}(A) = \dim \mathcal{R}(A).$$

- $\operatorname{rank}(A) = \operatorname{rank}(A^T)$
- $\operatorname{rank}(A)$ is the maximum number of independent columns (or rows) of A. Hence $\operatorname{rank}(A) \leq \min\{m, n\}$.
- $\operatorname{rank}(A) + \dim \mathcal{N}(A) = n$ (rank-nullity)

We say a matrix A is **full rank** if $\operatorname{rank}(A) = \min\{m, n\}$.

The rank of the product of two matrices satisfies

$$\operatorname{rank}(AB) \leq \min \left\{ \operatorname{rank}(A), \operatorname{rank}(B) \right\}.$$

If $A \in \mathbb{R}^{m \times n}$ has rank r then A can be factored as BC with $B \in \mathbb{R}^{m \times r}$ and $C \in \mathbb{R}^{r \times n}$.

Trace

The trace of a square matrix $A \in \mathbb{R}^{n \times n}$ is the sum of its diagonal entries, i.e.

$$\operatorname{tr}(A) = \sum_{j=1}^{n} A_{jj}.$$

Trace satisfies the following properties:

- $\operatorname{tr}(A) = \operatorname{tr}(A^T)$
- $\operatorname{tr}(\alpha A + \beta B) = \alpha \operatorname{tr}(A) + \beta \operatorname{tr}(B)$
- if AB is square then $\operatorname{tr}(AB) = \operatorname{tr}(BA)$

Determinant

The determinant of a square matrix $A \in \mathbb{R}^{n \times n}$ is a function $\det : \mathbb{R}^{n \times n} \to \mathbb{R}$ that satisfies the following properties:

- $\det \boldsymbol{I} = 1$
- $\det \alpha A = \alpha^n \det A$
- swapping any two rows/columns changes the sign of the determinant
- $\det AB = \det A \det B$

We can interpret the determinant as the volume of the parallelepiped spanned by the rows (or columns) of A.

Matrix Inverse

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

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

A matrix is **invertible** (i.e. has an inverse) if and only if $\det A \neq 0$. This is equivalent to:

- the columns/rows of A form a basis for $\mathbb{R}^n$
- $y = Ax$ has a unique solution for all $x \in \mathbb{R}^n$
- A is full-rank (i.e. $\mathcal{N}(A) = \{0\}$ and $\mathcal{R}(A) = \mathbb{R}^n$)
- $\det A^T A = \det AA^T \neq 0$

Cauchy-Schwarz Inequality

For any vectors $x, y \in \mathbb{R}^n$, we have that

$$\left|x^T y\right| \leq \|x\|_2 \|y\|_2.$$

The angle between vectors in $\mathbb{R}^n$ is given by

$$\theta = \cos^{-1}\left(\frac{x^T y}{\|x\|_2\|y\|_2}\right).$$

- If x and y are aligned then $x^T y =$

Eigenvalues and Eigenvectors

$\lambda \in \mathbb{C}$ is an eigenvalue of $A \in \mathbb{R}^{n\times n}$ if

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

Equivalently, there exists a non-zero $v \in \mathbb{C}^n$ such that $(\lambda I - A)v = 0$, or $Av = \lambda v$. Any such v here is called an eigenvector of A, associated with eigenvalue λ.

The eigenvalues of a symmetric matrix $A \in \mathbb{R}^{n\times n}$ are real. Moreover, there exists a set of orthogonal eigenvectors $q_1, \ldots, q_n$ such that $Aq_i = \lambda_i q_i$ and $q_i^T q_j = 0$ if $i \neq j$.

In matrix form, there is an orthonormal Q such that $A = Q\Lambda Q^T$.

Norm Matrices

A matrix norm is a function $\|\ \ \| : \mathbb{R}^{m\times n} \to \mathbb{R}$ that, similar to vector norms, satisfy linearity, positive definiteness, and the triangle inequality.

- Induced norms: $\|A\| = \sup\{\|Ax\| \mid x \in \mathbb{R}^n, \|x\| \leq 1\}$

- Frobenius norm: $\|A\|_F = \sqrt{\left\{\sum_{ij} a_{ij}^2\right\}}$
- Nuclear norm: $\|A\|_* = \sum_i \sigma_i(A) = \mathrm{tr}(\sqrt{A^T A})$

Square matrices also satisfy the sub-multiplicative property:

$$\|AB\| \leq \|A\|\|B\|.$$

1.5 Matrix Factorization

LU Factorization

Every nonsingular matrix $A \in \mathbb{R}^{n \times n}$ can be factored as

$$A = PLU$$

where P is a permutation matrix, L is unit lower triangular, and U is upper triangular and non-singular.

Cholesky Factorization

Every symetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ can be factored as

$$A = LL^T$$

where L is lower triangular and non-singular with positive diagonal elements.

Singular Value Decomposition

Any matrix A can be decomposed as

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

where $A \in \mathbb{R}^{m \times n}$ has rank r, $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ which satisfy $U^T U = I$ and $V^T V = I$, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

Since $A^T A = V\Sigma^2 V^T$ we have v_i are the eigenvectors of $A^T A$. Similarly, u_i are the eigenvectors of AA^T.

We can use SVD to interpret a linear map $y = Ax$ as follows:

- we compute coefficients of x along the input directions $v_1, \ldots, v_r$
- scale the coefficients by σ_i
- re-constitute along the output directions $u_1, \ldots, u_r$

Here, v_1 is the most sensitive input direction, and u_1 is the highest gain output direction.

Matrix Calculus

We can compute partial derivatives of a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ as

$$\frac{\partial f(x)}{\partial x_{ij}} = \lim_{\alpha\to 0} \frac{f(x + \alpha e_i e_j^T) - f(x)}{\alpha}.$$

We can also compute the gradient (Jacobian) of f as

$$\nabla_A f(A) = \begin{pmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} & \cdots & \frac{\partial f}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \frac{\partial f}{\partial A_{m2}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{pmatrix}$$

Partial derivatives are linear:

- $\nabla_A(f + g) = \nabla_A f + \nabla_A g$
- $\nabla_A(tf) = t\nabla_A f$

Chain rule and product rule also extend to matrix calculus.

In vector calculus, the **Hessian** of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the matrix of second-order partial derivatives of f, i.e.

$$\nabla_x^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

1.6 Probability Theory

A probability distribution is a function that maps outcomes of an experiment to probabilities:

- for discrete variables we have **probability mass functions**
- for continuous variables we have **probability density functions**

The **mean** or **expected value** of a random variable is the sum of possible values weighted by their probabilities:

$$\mathbb{E}[X] = \int_x x P(X = x)\,\mathrm{d}x$$

The **variance** of a random variable X is $\mathbb{E}[(X - \mathbb{E}[X])^2]$.