Lecture 2: Math Review

Learning goals:

- Review some basic math concepts the course is built upon
- Topics:
  - Vector spaces, norms, inner products
  - Analysis + functions
  - Vector calculus
  - Linear algebra

- In our course, optimization variables will be elements of finite-dimensional vector spaces
  - canonical example: $\mathbb{R}^n$
    - n-dimensional Euclidean vector space
    - $x \in \mathbb{R}^n \longleftrightarrow x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$
    - well-defined operations for adding vectors and multiply them by scalars (usually real numbers)

- another important example: $\mathbb{R}^{n \times m}$
    - vector space of $n \times m$ real-valued matrices
    - can be identified with a Euclidean vector space of dimension nm
    - but get distinct and interesting spaces using matrix norms (more soon)

- we will not consider <u>infinite-dimensional</u> vector spaces such as $C = \{ f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous} \}$
    - cannot represent any element as a linear combination of a finite set of basis vectors
    - more mathematically intricate since certain properties of norms, inner products, etc. not the same as in finite-dim spaces
    - can pose interesting optimization problems in these spaces, but they are mostly theoretically rather than computationally interesting

---

- We can use <u>norms</u> to endow our vector space with notions of <u>magnitude</u> and <u>distance</u>
    - there are <u>many</u> different norms on vector spaces

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

1. $f(x) \geq 0 \quad \forall x \in \mathbb{R}^n$ (nonnegativity)
2. $f(x) = 0 \iff x = 0$ (definiteness)
3. $f(tx) = |t| f(x) \quad \forall x \in \mathbb{R}, \ \forall t \in \mathbb{R}$ (homogeneity)
4. $f(x+y) \leq f(x) + f(y) \quad \forall x, y \in \mathbb{R}^n$ (triangle inequality)

Notation: $f(x) = \|x\|$ or $\|x\|_{\text{symbol}}$

to indicate a specific norm

Can use norm to define distance between vectors:

$$\text{dist}(x, y) = \|x - y\|$$

Examples:

- 2-norm or Euclidean norm: $\|x\|_2 = \sqrt{\sum_i x_i^2}$
- 1-norm: $\|x\|_1 = \sum_i |x_i|$
- ∞-norm: $\|x\|_\infty = \max_i |x_i| = \max\{|x_1|, \ldots, |x_n|\}$
- p-norm: $\|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}, \quad p \geq 1$
- Quadratic norm: $\|x\|_Q = \sqrt{x^T Q x} = \|Q^{\frac{1}{2}} x\|_2, \quad Q = Q^T \succ 0 \ (Q \in S^n_{++})$

matrix square root

Matrix norms on $\mathbb{R}^{m\times n}$: $X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & x_{22} & \\ x_{m1} & & x_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$

- Frobenius norm : $\|X\|_F = \sqrt{tr(X^T X)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2}$
    - i.e. the Euclidean norm of the vector of all entries
- Sum-absolute value norm : $\|X\|_{sav} = \sum_{i=1}^{m}\sum_{j=1}^{n} |X_{ij}|$
- Max-absolute value norm : $\|X\|_{mav} = \max_{ij} |X_{ij}|$
- Induced matrix norms (aka operator norms):

> **Definition:** An operator norm of $X \in \mathbb{R}^{m\times n}$, induced by the vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$ is
>
> $$\|X\|_{a,b} = \max_u \{ \|Xu\|_a \mid \|u\|_b \le 1 \}$$

- When the same vector norm is used in both spaces we write $\|X\|_c$
- Examples:
    - $\|X\|_2 = \underbrace{\sigma_{max}}_{\text{max singular value}}(X) = \sqrt{\underbrace{\lambda_{max}}_{\text{max eigenvalue}}(X^T X)}$
    - $\|X\|_1 = \max_j \sum_{i=1}^{n} |X_{ij}|$ = max column sum
    - $\|X\|_\infty = \max_i \sum_{j=1}^{m} |X_{ij}|$ = max row sum

- Nuclear norm : $\|X\|_* = \text{trace}\left(\sqrt{X^T X}\right) = \sum_{i=1}^{\min(m,n)} \sigma_i(X)$

    = sum of singular values

- The set of vectors with norm less than or equal to 1

$$B = \{ x \in \mathbb{R}^n \mid \|x\| \leq 1 \}$$

is called the <u>unit ball</u> of the norm $\|\cdot\|$.

- <u>Dual norms</u>: Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$.

  The associated <u>dual norm</u>, denoted $\|z\|_*$, is defined as

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

- <u>Examples</u>:
  - $\|x\|_{1*} = \|x\|_\infty$
  - $\|x\|_{2*} = \|x\|_2$
  - $\|x\|_{\infty *} = \|x\|_1$
  - More generally, $\|x\|_{p*} = \|x\|_q$ where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$

- We can give further structure to a vector space by using an <u>inner product</u>, which gives notions of <u>angle</u> and <u>orthogonality</u>

Definition: A function $\langle \cdot , \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an <u>inner product</u> if

① $\langle x, x \rangle \geq 0$, $\langle x, x \rangle \Leftrightarrow x = 0$ (positivity)

② $\langle x, y \rangle = \langle y, x \rangle$ (symmetry)

③ $\langle x+y, z \rangle = \langle x, y \rangle + \langle y, z \rangle$ (additivity)

④ $\langle tx, y \rangle = t \langle x, y \rangle \quad \forall t \in \mathbb{R}$ (homogeneity)

- standard inner product on $\mathbb{R}^n$:

$$\langle x, y \rangle = x^T y = \sum x_i y_i$$

- standard inner product on $\mathbb{R}^{m \times n}$:

$$\langle X, Y \rangle = \text{trace}(X^T Y) = \sum_i \sum_j X_{ij} Y_{ij}$$

- the <u>angle</u> between nonzero vectors $x, y \in \mathbb{R}^n$ is

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

we say x and y are <u>orthogonal</u> if $\langle x, y \rangle = 0$

- Any inner product defines a norm given by $f(x) = \sqrt{\langle x, x \rangle}$

Theorem (Cauchy-Schwarz) For any $x, y \in \mathbb{R}^n$

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|$$

- Analysis / Topology

  - Open and closed sets

Definitions:

- an element x of a subset $C \subseteq \mathbb{R}^n$ is called an interior point of C if $\exists\, \varepsilon > 0$ for which

$$\{ y \in \mathbb{R}^n \mid \|y - x\| \leq \varepsilon \} \subseteq C$$

- the set of all interior points of C is called the interior of C, denoted $\operatorname{int} C$
- a set $C \subseteq \mathbb{R}^n$ is called open if $C = \operatorname{int} C$
- a set $C \subseteq \mathbb{R}^n$ is called closed if its complement $\overline{C} = \mathbb{R}^n \setminus C = \{ x \in \mathbb{R}^n \mid x \notin C \}$ is open
- the closure of a set C is defined as

$$\operatorname{cl} C = \mathbb{R}^n \setminus \operatorname{int}(\mathbb{R}^n \setminus C)$$

  - a set C is closed iff it contains the limit point of every convergent sequence in it, and the closure is the set of all limit points of convergent sequences in C

- the boundary of a set C is defined as

$$\operatorname{bd} C = \operatorname{cl} C \setminus \operatorname{int} C$$

- The <u>supremum</u> of $C \subseteq \mathbb{R}$ is its least upper bound
  - We take $\sup \phi = -\infty$ and $\sup C = \infty$ if C is unbounded above
- The <u>infimum</u> of $C \subseteq \mathbb{R}$ is its greatest lower bound
  - We take $\inf \phi = \infty$ and $\inf C = -\infty$ if C is unbounded below

- Functions

- The notation $f : A \rightarrow B$ means that f is a function on a subset of A into the set B
- A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is <u>continuous</u> at $x \in \mathbb{R}^n$ if $\forall \varepsilon > 0$, $\exists \delta > 0$ such that for $y \in \mathbb{R}^n$

$$\|y - x\| \leq \delta \implies \|f(y) - f(x)\|_2 \leq \varepsilon$$

- A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is called <u>closed</u> if $\forall \alpha \in \mathbb{R}$ the <u>sublevel set</u> $\{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ is closed

Equivalently, f is closed iff the <u>epigraph</u> of f

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

- The <u>level set</u> of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the set

$$\{x \in \mathbb{R}^n \mid f(x) = \alpha\}$$

Derivatives

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

- The partial derivative with respect to x_i is

$$\frac{\partial f}{\partial x_i} = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}$$

- The gradient of f is the (column) vector of partial derivatives

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

Note: the derivative of f is the row vector

$$Df(x) = \nabla f(x)^T$$

- The Hessian of f, denoted $\nabla^2 f(x)$, is the $n \times n$ symmetric matrix of second derivatives

$$(\nabla^2 f)_{ij} = \frac{\partial f}{\partial x_i \partial x_j}$$

- The Jacobian of $f : \mathbb{R}^n \to \mathbb{R}^m$, where $f = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}$ is the $m \times n$ matrix of (first) derivatives

$$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}$$

- The first and second order Taylor expansions of a function $f: \mathbb{R}^n \to \mathbb{R}$ at a point x_0 are

$$f(x) \cong f(x_0) + \nabla f(x_0)^T (x - x_0) + o(\|x - x_0\|)$$

$$f(x) \cong f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0) + o(\|x - x_0\|^2)$$

"little o" notation : $f = o(g(x))$ if $\lim_{x \to 0} \frac{|f(x)|}{|g(x)|} = 0$

Differentiation Rules

- Product Rule. Let $f, g: \mathbb{R}^n \to \mathbb{R}^m$, $h(x) = f^T(x) g(x)$

$$J_h(x) = f^T(x) J_g(x) + g^T(x) J_f(x), \quad \nabla h(x) = J_h(x)^T$$

- Chain Rule. Let $f: \mathbb{R}^n \to \mathbb{R}^m$ and $g: \mathbb{R}^m \to \mathbb{R}^p$. Define the composition $h: \mathbb{R}^n \to \mathbb{R}^p$ by $h(x) = g(f(x))$

Then $J_h(x) = J_g(f(x)) \, J_f(x)$

Common functions

- Linear functions: $f(x) = c^T x$, $c \in \mathbb{R}^n$, $c \neq 0$

- Affine functions: $f(x) = c^T x + b$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}$

$$\nabla f(x) = c, \quad \nabla^2 f(x) = 0$$

- Quadratic functions: $f(x) = x^T Q x + c^T x + b$, $Q = Q^T \in \mathbb{R}^{n \times n}$

$$\nabla f(x) = 2Qx + c$$

$$\nabla^2 f(x) = 2Q$$

- Important composition example: $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be affine: $f(x) = Ax + b$

Let $g : \mathbb{R}^m \to \mathbb{R}^p$ and define $h(x) = g(f(x)) = g(Ax+b)$

Then applying the chain rule gives

$$J_h(x) = J_g(Ax+b)\,A$$

When $m=1$, we get the gradient formula

$$\nabla g(x) = A^T \nabla f(Ax+b)$$

Linear Algebra

Let $A \in \mathbb{R}^{m \times n}$

- The range of A is $\mathcal{R}(A) = \{ Ax \mid x \in \mathbb{R}^n \} \subseteq \mathbb{R}^m$
- The nullspace (or kernel) of A is $\mathcal{N}(A) = \{ x \in \mathbb{R}^n \mid Ax = 0 \} \subseteq \mathbb{R}^n$
  - These are subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, resp.

- If V is a subspace of $\mathbb{R}^n$, its orthogonal complement is

$$V^\perp = \{ x \in \mathbb{R}^n \mid z^T x = 0 \quad \forall z \in V \}$$

- A fundamental result of linear algebra:

$$\mathcal{N}(A) = \mathcal{R}(A^T)^\perp, \quad \mathcal{R}(A) = \mathcal{N}(A^T)^\perp$$

Symmetric Eigenvalue Decomposition

Let $A \in S^n$ (i.e. $A = A^T \in \mathbb{R}^{n \times n}$).

Then A can be factored/decomposed as

$$A = Q \Lambda Q^T$$

where Q is orthogonal (i.e. $Q^T Q = I$)

$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$

The eigenvalues of $A \in S^n$ are real and can be ordered as

$$\lambda_{max}(A) = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n = \lambda_{min}(A)$$

- We have

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

$$\|A\|_2 = \max_i |\lambda_i| \quad , \quad \|A\|_F = \sqrt{\sum_{i=1}^{n} \lambda_i^2}$$

$$= \max(\lambda_1, -\lambda_n)$$

- A matrix $A \in S^n$ is called <u>positive</u> <u>semidefinite</u> $(A \succeq 0)$ if

$$x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n$$

- " " " " <u>positive</u> <u>definite</u> $(A \succ 0)$ if

$$x^T A x > 0 \quad \forall x \in \mathbb{R}^n, \; x \neq 0 \qquad (A \succeq 0)$$

- " " " " negative semidefinite if $-A \succeq 0$
- " " " " negative definite if $-A \succ 0$ $(A \prec 0)$

- There is a <u>partial</u> <u>ordering</u> of symmetric matrices
  - we write $A \succeq B$ when $A - B \succeq 0$
  - in general, for $A, B \in S^n$ it's possible that $A \not\succeq B$ and $B \not\succeq A$

<u>Singular</u> <u>Value</u> <u>Decomposition</u>

Let $A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank} A = r$. Then A can be factored as

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

where $U \in \mathbb{R}^{m \times r}$ and $U^T U = I$, $V \in \mathbb{R}^{n \times r}$ and $V^T V = I$, and

$\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ with $\underbrace{\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n}_{\text{singular values}}$