Math Preliminaries and Optimality Conditions

Ruoyu Sun

Outline

Overview

1. Notations: Sets, functions, derivatives, gradients
2. Vectors, matrices
3. Norms, sequences, limits, continuity
4. Mean value theorems
5. Implicit function theorem
6. Contraction mappings
7. Reference Appendix A, B of the textbook
8. Get yourself familiar with them

Notations

1. **Sets**: $X, \; x \in X, \; X_1 \cup X_2, \; X_1 \cap X_2$
2. **Inf and Sup**:
   The supremum of a nonempty set $X \subset \mathbb{R}$ is the smallest scalar y such that:
   $$y \geq x, \; \forall \, x \in X$$
   The infimum of a nonempty set $X \subset \mathbb{R}$ is the largest scalar y such that:
   $$y \leq x, \; \forall \, x \in X$$
   If $\sup X \in X$ (or, $\inf \in X$), then we say $\sup X = \max X$ (or, $\inf X = \min X$).
   $$\sup\{1/n \mid n \geq 1\} =?, \quad \inf\{x \in \mathbb{R} \mid 0 < x < 1\} =?$$
3. **Function**:
   $$f : X \to \mathbb{R}^{d_y}, \; X \text{ is called the domain}$$
   - If $d_y = 1$, we say f is a scalar-valued function; otherwise, a vector-valued function

Vectors

1. **Vector**: a vector $\mathbf{x} = [x_1; \cdots; x_n] \in \mathbb{R}^{n\times 1}$ is a column of scalars

   a vector $\mathbf{x} = [x_1, \cdots, x_n] \in \mathbb{R}^{1\times n}$ is a row of scalars
2. **Linear combination**: if $\mathbf{x} = [x_1, \cdots, x_n]$ and $\mathbf{y} = [y_1, \cdots, y_n]$, then the linear combination is given by

$$\alpha\mathbf{x} + \beta\mathbf{y} = (\alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \cdots, \alpha x_n + \beta y_n)$$

3. **Inner product**: $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{y} = \sum_{i=1}^n x_i y_i$

   Question: when is inner product positive, negative, zero?

   Orthogonality: $\mathbf{x} \perp \mathbf{y}$ iff (if and only if) $\langle x, y \rangle = 0$.
4. **Linearly Independent**: A set of vectors $\{\mathbf{x}^1, \cdots, \mathbf{x}^r\}$ are linearly independent if there does not exist a $(\alpha_1, \cdots, \alpha_r) \neq 0$ s.t.

$$\alpha_1\mathbf{x}^1 + \alpha_2\mathbf{x}^2 + \cdots + \alpha_r\mathbf{x}^r = 0.$$

Vectors

1. Basis and dimension of a linear space
2. Orthogonal complement of a subspace S:

$$S^{\perp} := \{\mathbf{x} \mid \langle \mathbf{x}, \mathbf{y} \rangle = 0, \ \forall \ \mathbf{y} \in S\}$$

3. **Vector norms**: A norm $\|\mathbf{x}\|$ on $\mathbb{R}^n$ that assigns a scalar $\|\mathbf{x}\|$ to every $\mathbf{x} \in \mathbb{R}^n$ that satisfying
   3.1 $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x}$ (non-negativity)
   3.2 $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$ for all $c \in \mathbb{R}$ and all $\mathbf{x}$ (homogeneous)
   3.3 $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = 0$
   3.4 $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y}$ (triangular inequality)

Vectors

- **Common norms**

$$\text{Euclidean norm : } \|\mathbf{x}\|_2 = (\mathbf{x}^\top \mathbf{x})^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

$$\ell_p \text{ norm : } \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \text{ for some } p \geq 1$$

$$\ell_1 \text{ norm : } \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\ell_\infty \text{ norm : } \|\mathbf{x}\|_\infty = \max_i |x_i|$$

Cauchy-Schwartz inequality

An important inequality about the inner product of two vectors is the Cauchy-Swartz inequality

1. Bound the inner product of two vectors with their norms
2. Given two vectors $\mathbf{x}$ and $\mathbf{y}$ of the same size, we have

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

3. Useful fact about inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \cos(\theta) \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$

Matrices

1. For any matrix $\mathbf{A}$, we use a_{ij} (or A_{ij}) to denote its (i,j)th entry.
2. Matrix addition, multiplication, transpose, symmetirx matrices $\mathbf{A} = \mathbf{A}^\top$. We use both A' and $A^\top$ to denote the transpose of A.

$$[\mathbf{AB}]' = \mathbf{B}'\mathbf{A}', \ \mathbf{AB} \neq \mathbf{BA}$$

3. Let $\mathbf{A}$ be a $m \times n$ matrix.
   - Range of $\mathbf{A}$: $R(\mathbf{A}) = \{\mathbf{Ax}, \mid \mathbf{x} \in \mathbb{R}^n\}$;
   - Null space of $\mathbf{A}$: $N(\mathbf{A}) = \{\mathbf{x} \mid \mathbf{Ax} = 0\}$
   - Rank of $\mathbf{A}$ $\mathrm{Rank}(\mathbf{A})$. Full rank matrix $\mathbf{A}$: $\mathrm{Rank}(\mathbf{A}) = \min\{m, n\}$.
4. **Inner product**:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{AB}') = \sum_{i,j} A_{ij} B_{ij}$$

   where the trace operate is given by

$$\mathrm{Tr}[\mathbf{A}] = \sum_{i=1}^{n} A_{ii}$$

5. Property: $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{AB}') = \mathrm{Tr}(\mathbf{B}'\mathbf{A})$.

Square Matrices

1. **Square matrix** $(m = n)$; Identity matrix $\mathbf{I}$
2. Determinant $\det(\mathbf{A})$, inverse $\mathbf{A}^{-1}$.

   $\mathbf{A}^{-1}$ exists iff $\det(\mathbf{A}) \neq 0$
3. Useful identities: $\det(\mathbf{A}) = \det(\mathbf{A}')$
4. Orthogonal matrices: $\mathbf{A}\mathbf{A}' = \mathbf{I}$
5. (Complex) Eigenvalue λ: $\mathbf{Ax} = \lambda\mathbf{x}$ for some $\mathbf{x} \neq 0$
6. Spectral radius: $\rho(\mathbf{A}) = \max_i\{|\lambda_i|\}$, λ_i is an eigenvalue of $\mathbf{A}$. Here the modulus $|z| = \sqrt{a^2 + b^2}$ for a complex number $z = a + b\sqrt{-1}$.

Square Matrices

1. Eigen-decomposition of a (real) symmetric matrix:

$$\mathbf{A} = \mathbf{P}'\mathbf{\Lambda}\mathbf{P}$$

   where $\mathbf{P}$ is an orthogonal matrix ($\mathbf{P}'\mathbf{P} = \mathbf{I}$), $\mathbf{\Lambda}$ is diagonal and real.

2. Positive semi-definite (PSD) matrix: $\mathbf{A} \succeq 0$ iff

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0, \ \forall \ \mathbf{x}. \tag{1}$$

   A is a positive definite (PD) matrix iff $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \ \forall \ \mathbf{x} \neq 0$; denoted as $A \succ 0$.

3. **Property**: $\mathbf{A} \succeq 0, \ \mathbf{B} \succeq 0 \Rightarrow \mathbf{A} + \mathbf{B} \succeq 0$; $\mathbf{A} \succeq \mathbf{B} \Rightarrow \mathbf{A} - \mathbf{B} \succeq 0$
   $\mathbf{A} \succeq 0 \Rightarrow$ All eigenvalues of $\mathbf{A}$ are non-negative

4. Condition number (for PD matrix): $\kappa(\mathbf{A}) = \lambda_{\max}/\lambda_{\min} > 0$
   important for optimization!!

Single Value Decomposition: Definition

Definition (SVD and singular values): For any matrix $M \in R^{m \times n}$, we say $M = USV^\top$ is a singular value decomposition if the following is satisfied: let $q = \min\{m, n\}$.

- $U \in R^{m \times m}$ and $V \in R^{n \times n}$ are orthogonal matrices, i.e., $UU^\top = I_m, VV^\top = I_n$
- There exists a square diagonal matrix $S_q = diag(\sigma_1, \ldots, \sigma_q)$, where $\sigma_1 \geq \cdots \geq \sigma_q \geq 0$ such that

$$S = \begin{cases} S_q & m = n \\ [S_q, 0] & m < n \\ \begin{bmatrix} S_q \\ 0 \end{bmatrix} & m > n \end{cases} \tag{2}$$

The singular values of M are $\sigma_1, \ldots, \sigma_q$.

Single Value Decomposition

1. Relationship of SVD and ED: σ_i^2 is an eigenvalue of $\mathbf{AA}'$ (Why?)

$$\mathbf{AA}' = (\mathbf{U\Sigma V}')(\mathbf{V\Sigma U}') = \mathbf{U\Sigma}^2\mathbf{U}'$$

2. Difference of SVD and ED:
   - SVD applies to all rectangular matrices;
   - ED applies to some square matrices (including symmetric matrics).

   Uncommon definition: Some books define singular values of M as the square root of the eigenvalues of $M^\top M$.
   - This definition is NOT equivalent to our definition.
   - This definition is clean, but less common.

Practice Questions

Q1: For an $m \times n$ matrix M, how many singular values does it have (counting multiplicity)?

A: m   B: n   C: $\min(m, n)$ D: $rank(M)$

Answer: C.

Q2: Does M and $M^\top$ have the same singular values?

Answer: Yes.

- Assume $M = (\mathbf{a})$ is an $m \times 1$ matrix. Then $M^T M$ has one eigenvalue $\|\mathbf{a}\|^2$ and MM^T has m eigenvalues $\|\mathbf{a}\|^2, 0, 0, \ldots, 0$.
- A common misconception is: M has one singular value $\|\mathbf{a}\|$ and M^T has m singular values $\|\mathbf{a}\|, 0, 0, \ldots, 0$.
- By our definition in the last slide, the correct answer is: both M and M^T have only one singular value $\|\mathbf{a}\|^2$.

Matrices and Norms

1. Norms:

   $$\text{Frobenious Norm : } \|\mathbf{A}\|_F = \left(\sum_{i,j} |A_{ij}|^2\right)^{1/2} = \left(\sum_i \sigma_i^2\right)^{1/2}$$

   $$\text{Nuclear Norm : } \|\mathbf{A}\|_* = \sum_i \sigma_i$$

   $$\text{Matrix 2-norm (spectral norm) : } \|\mathbf{A}\|_2 = \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max_i \sigma_i$$

2. **Question**: difference and relation between spectral radius and specral norm?
   - Relation: For real symmetric matrices, $\|A\|_2 = \rho(A)$.
   - Difference: For general matrices, $\|A\|_2 \geq \rho(A)$.
3. Important property: $\|\mathbf{A}\|_F^2 = \langle \mathbf{A}, \mathbf{A} \rangle = trace(\mathbf{A}'\mathbf{A})$
4. Useful inequality:

   $$\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2.$$

Big O Notation

For two scalar functions $f(x) \in \mathbb{R}, g(x) \in \mathbb{R}_+$, where $x \in \mathbb{R}$, we write:

- $f(x) = \mathcal{O}(g(x))$ if $\limsup_{x\to\infty} \frac{|f(x)|}{g(x)} < \infty$; we say f is dominated by g asymptotically
- $f(x) = \Omega(g(x))$ if $\liminf_{x\to\infty} \frac{|f(x)|}{g(x)} > 0$;
- $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$ both hold
- $f(x) = o(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$.

Example:

- $n^3 + n + 2 = \Omega(1)$, $n^3 + n + 2 = \Omega(n^2)$.
  $n^3 + n + 2 = \mathcal{O}(n^3)$, $n^3 + n + 2 = \Theta(n^3)$, $n^3 + n + 2 = \Omega(n^3)$,
  $n^3 + n + 2 = \mathcal{O}(n^4)$, $n^3 + n + 2 = o(n^4)$,
- $\frac{1+n}{n^3} = \mathcal{O}(\frac{1}{n^2})$, $\frac{1+n}{n^3} = \Theta(\frac{1}{n^2})$, $\frac{1+n}{n^3} = \Omega(\frac{1}{n^2})$,
  $\frac{1+n}{n^3} = o(\frac{1}{n})$, $\frac{1+n}{n^3} = \Omega(\frac{1}{n^3})$

For two scalar functions $f(x) \in \mathbb{R}, g(x) \in \mathbb{R}_+$, where $x \in \mathbb{R}$, we write:

- $f(x) = \mathcal{O}(g(x))$ as $x \to a$ if $\limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty$;
- Example: $\epsilon^2 + \epsilon^3 = \mathcal{O}(\epsilon^2)$ as $\epsilon \to 0$

Reference: `https://en.wikipedia.org/wiki/Big_O_notation`

Gradient

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously twice differentiable function.

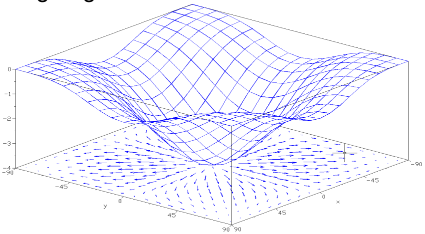
1. Partial derivative (where $\mathbf{e}_i$ is the ith unit vector of $\mathbb{R}^n$)

$$\frac{\partial f(\mathbf{x})}{\partial x_i} := \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t}$$

2. Gradient vector (a column vector)

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}; \cdots ; \frac{\partial f(\mathbf{x})}{\partial x_n} \right) \in \mathbb{R}^{n \times 1}$$

Physical meaning of gradient?



Wikipedia: “the gradient points in the direction of the greatest rate of increase of the function, and its magnitude is the slope of the graph in that direction”

Hessian and Taylor Expansion

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously twice differentiable function.

1. Hessian matrix:

$$\nabla^2 f = \left[\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j} \right] \in \mathbb{R}^{n \times n}$$

2. Taylor expansion:

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla f(\mathbf{x})'(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})' \nabla^2 f(\mathbf{x})(\mathbf{x} - \mathbf{y}) + o(\|\mathbf{x} - \mathbf{y}\|^2)$$

Derivatives: Chain Rule

1. Scalar case: suppose $f, g : \mathbb{R} \to \mathbb{R}$ are functions, and $f'(x)$ and $g'(f(x))$ exist, then $h(x) \triangleq g(f(x))$ satisfies

$$h'(x) = g'(f(x))f'(x).$$

Example: $f(x) = \sin x$, $g(y) = y^2$, $h(x) = (\sin x)^2$, then

$$h'(x) = 2\sin x \;\cos x.$$

Lipschitz Continuous

1. **Lipschitz continuous**: if a function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ satisfies

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|, \ \forall\ \mathbf{x}, \mathbf{y}$$

   the function is called γ-Lipschitz continuous;
   - If f is γ-Lipschitz continuous, then it is also $(\gamma + 1)$-Lipschitz continuous
   - The minimal such γ is called a Lipschitz constant of function f

2. Remark: Here $\|\cdot\|$ can be any given norm of the space $\mathbb{R}^n$ and $\mathbb{R}^m$, such as Euclidean norm, ℓ_1-norm, etc.

   When not specified, we assume it is Euclidean norm.

3. Example 1: $f(x) = 2x$ is 2-Lipschitz continuous;

   Example 2: What about $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a matrix?
   Spectral norm $\|\mathbf{A}\|_2$ (for Euclidean norm).

   Example 3: What about $f(x) = x^2$?
   Not Lipschitz continuous, or the Lipschitz constant is ∞.

Contraction Mapping

1. If the Lipschitz constant $\gamma \leq 1$, then f is called a non-expansive mapping
2. If $\gamma < 1$, then f is called a contraction mapping

   Example 1: $f(x) = 2x$ is not a contraction mapping; $f(x) = 0.5x$ is.

   Example 2: $f(x) = Ax$ is a contraction mapping (with respect to Euclidean norm) iff $\|A\|_2 < 1$.

Fixed Point Theorem

1. **Fixed point theorem**: If f is a contraction mapping that maps $\mathbb{R}^n$ to itself, then the following two results hold:
   1) There exists a unique fixed point $\mathbf{x}^*$ satisfying

   $$\mathbf{x}^* = f(\mathbf{x}^*).$$

   2) In addition, the iterated function sequence

   $$\mathbf{x}, f(\mathbf{x}), f(f(\mathbf{x})), \cdots ,$$

   converges to this unique fixed point $\mathbf{x}^*$ (independent of the initial point $\mathbf{x}$) .

2. Remark: This is a special case of “Banach fixed point theorem” (which applies to any complete metric space).

   Reference: e.g., `https://www.clear.rice.edu/comp360/lectures/old/FixedPtThmText.pdf`

Outline

Unconstrained Optimization

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

- **Objective function** $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function
- **Optimization variable** $x \in \mathbb{R}^n$
- **(Unconstrained) local minimum** $\hat{x}$: $\exists \epsilon > 0$ s.t. $f(x) \geq f(\hat{x})$, for all $\|x - \hat{x}\| \leq \epsilon$;
  i.e., x^* is the best in a small enough neighborhood
- **(Unconstrained) global minimum** x^*: $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$
- Strict global minimum: change $f(x) \geq f(x^*)$ to $f(x) > f(x^*)$ in the above definition. Similar for “strict local minimum”
- Switching the direction of inequalities, we obtain “local maximum”, “global maximum”, etc.

Discussion of Terminology

Plural form: "five minima", NOT "five minimums"; similarly, "maxima"

Minimizer v.s. minimal value:

- Sometimes, we call x^* "global minimizer" or "global minimum point" ; call $\hat{x}$ "local minimizer" or "local minimum point"
- We call $f(x^*)$ the "minimal value" or "minimum value"

Possible confusion of "minimum":

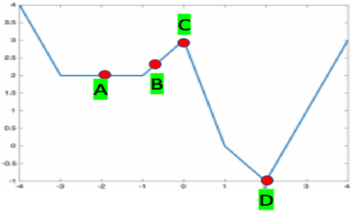
- In this class (and most optimization textbooks), "global minimum" refers to the argument $\hat{x}$, not the function value
- but outside class, some people may think it refers to the value (e.g. the first sentence of [R1] below may give you this impression)

Abbreviation: local-min = "local minimizer", global-min = "global minimizer"

[R1] Reference:
https://en.wikipedia.org/wiki/Maxima_and_minima

Local-min and global-min

Question: Which are local minima, which are global minima, and which are strict local minima?



Possible confusion of "local minimum": some people may think A is NOT local-min, and only D is local-min.

- This is because their notion of "local-min" should actually called "strict local-min"

Checkable Conditions for Local Min

- Given a point x, how to decide whether it is a local/global min (for a twice continuously differentiable function f)?

- First answer: exhaustive check
  - **Global-min**: verify $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$; or find one x s.t. $f(x) < f(x^*)$
  - **Local-min**: check $f(x) \geq f(x^*)$ for all $\|x - x^*\| \leq \epsilon$; or find a sequence $x_n \to x^*$ such that $f(x_n) < f(x^*)$

- **Challenge**: need to check infinitely many points (rigorously speaking).

- We need easily checkable conditions

- **Idea for local-min checking:** Use Taylor expansion to analyze local behavior around x

Checkable Conditions for Local Min

- **Question**: If x^* is local min, what's its property?
- Necessary conditions for local-min:

$$\nabla f(x^*) = 0, \quad \text{(first-order condition)},$$
$$\nabla^2 f(x^*) \succeq 0, \quad \text{(second-order condition)}.$$

- **Definition** (**stationary point**) We call the solutions that satisfy $\nabla f(x^*) = 0$ as stationary solutions, or stationary points
- **Definition** (**saddle point**) If a stationary point is neither a local minimum nor a local maximum, then it is a saddle point.

1D Case: Optimality Conditions

First, let us look at the simple one dimensional case (x is scalar)
Claim 1: Suppose $f : \mathbb{R} \to \mathbb{R}$ is a twice-differentiable function. If x^* is a local minimum of $f(x)$, then

$$f'(x) = 0, \quad f''(x) \geq 0 \tag{3}$$

Proof idea: **Step 1**: By first order Taylor expansion,

$$f(x) \approx f(x^*) + f'(x^*)(x - x^*) \geq f(x^*), \tag{4}$$

thus $f'(x^*)(x - x^*) \geq 0$ for any x close enough to x^*.

Step 2: Pick $x = x^* + \epsilon$, we have $f'(x^*) \geq 0$;
pick $x = x^* - \epsilon$, we have $f'(x^*) \leq 0$.
Thus $f'(x^*) = 0$.
Think: How to extend this proof to **constrained problems**?

Disclaimer on Proofs

Do we need to learn the proofs?

- Helpful for understanding (when applying to different scenarios)
- Some students complain about proofs; some complain about too few proofs

I know many people do NOT like proofs on slides.

- They like seeing writing the proof by hand (on board, or Pad)

A few reasons I don't don't write proofs this time:

- Tech issue (internet not good enough for recording);
- The best way to learn this proof is to DO IT YOURSELF;
- The proofs are NOT that critical for understanding the contents (though still improtant)

So I will be a bit quick on showing the proofs (if videos, you can pause).

Formal Proof of Claim 1

Proof of Claim 1: (using definition of derivatives; NOT Taylor expansion)

- Consider a sequence $\{x^r\} \to x^*$ where $x^r > x^*, \forall r$.
  Since x^* is a local min, $f(x^r) \geq f(x^*)$ for large enough r. Then

$$0 \overset{(i)}{\leq} \lim_{x^r \downarrow x^*} \frac{f(x^r) - f(x^*)}{x^r - x^*} = f'(x^*) \tag{5}$$

- Similarly, consider a sequence that approaches x^* from below, we have

$$f'(x^*) = \lim_{x^r \uparrow x^*} \frac{f(x^r) - f(x^*)}{x^r - x^*} \overset{(ii)}{\leq} 0 \tag{6}$$

- Together, we have $f'(x^*) = 0$.
- Consider a sequence $x^r \to x^*$, we have

$$f''(x^*) = \lim_{x^r \to x^*} \frac{f(x^r) - f(x^*) - f'(x^*)(x^r - x^*)}{(x^r - x^*)^2} \geq 0. \tag{7}$$

Checkable Conditions for Local Min

- For higher dimensions, derivation is similar (Prop 1.1.1)
- **Proof sketch**: Consider the one dimensional function $g(\alpha) = f(\mathbf{x}^* + \alpha\mathbf{d})$, where $\mathbf{d} \in \mathbb{R}^n$ is a direction.
- This function of α has a local minimizer $\alpha = 0$ (why?)
- Apply the previous theorem, we have

$$g'(0) = \langle \nabla f(\mathbf{x}^*), \mathbf{d} \rangle = 0, \; g''(0) = \langle \mathbf{d}, \nabla^2 f(\mathbf{x}^*)\mathbf{d} \rangle \geq 0 \tag{8}$$

- Note $\mathbf{d}$ is an arbitrary direction:
  - the first equation means $\nabla f(\mathbf{x}^*) = 0$,
  - the second equation means $\nabla^2 f(\mathbf{x}^*) \succeq 0$
- For detailed proof, please read Section 1.1 of the text book

Necessary v.s. Sufficient

- What have we done so far?
- For a given solution x^*, I can check whether it is local minimum?
- No, necessary conditions only help identify "a point is NOT a local-min"
  - If a point does NOT satisfy necessary conditions, then it is NOT a local-min
- **Question**: Can we have some simple conditions that guarantee that a point is a local optimum (sufficient condition)?

Sufficient Conditions for Local Min

- We have the following sufficient conditions

$$
\begin{aligned}
\nabla f(x^*) &= 0, \quad \text{(first-order condition)}, \\
\nabla^2 f(x^*) &\succ 0, \quad \text{(second-order condition)}.
\end{aligned}
$$

Proposition 2: Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Suppose $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$, then x^* is a local minimum of f.

- Why?
- Prove by Taylor expansion (see next page)

Proof of Prop. 2

Proof of Prop. 2:

- By the assumption, we have $\mu \triangleq \lambda_{\min}(\nabla^2 f(x^*)) > 0$.
- Write the Taylor expansion:

$$
\begin{aligned}
f(x^* + \delta) - f(x^*) &= \langle \nabla f(x^*), \delta \rangle + \frac{1}{2}\delta^T \nabla^2 f(x^*)\delta + O(\|\delta\|^3), \\
&= \frac{1}{2}\delta^T \nabla^2 f(x^*)\delta + O(\|\delta\|^3) \\
&\geq \|\delta\|^2 \lambda_{\min}(\nabla^2 f(x^*)) + O(\|\delta\|^3).
\end{aligned}
$$

- Write $\delta = \epsilon v$ where $\|v\| = 1$, and $\epsilon > 0$, then for any $\|v\| = 1$,

$$
f(x^* + \epsilon v) - f(x^*) \geq \epsilon^2 \mu + O(\epsilon^3).
$$

- This implies

$$
\lim_{\epsilon \to 0} \frac{1}{\epsilon^2}[f(x^* + \epsilon v) - f(x^*)] \geq \mu + \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} O(\epsilon^3) = \mu. \tag{9}
$$

- Thus for ϵ small enough, $\frac{1}{\epsilon^2}[f(x^* + \epsilon v) - f(x^*)] \geq \mu/2,$ which implies $f(x^* + \epsilon v) \geq f(x^*) + \mu\epsilon^2/2 > f(x^*)$.
- This holds for any $\|v\| = 1$ and small enough ϵ, thus x^* is a (strict) local-min by definition.

Why Optimality Conditions?

- Optimality conditions are useful because:
  1. **sufficient condition**: provide guarantees for a candidate solution to be optimal
  2. **necessary condition**: indicate when a point is NOT optimal
- Guide algorithm design
  1. Algorithm design: algorithms should look for points achieving the optimality conditions
  2. Stopping criterion: algorithm should stop when the optimality condition is approximately satisfied

Outline

Mathematical Review

Local-Min and Optimality Conditions

Application of Optimality Conditions

Use of Optimality Condition: Finding Optimal Solutions

- How to find a global minimum?
- **Tentative-Method 1**: among all stationary points, find the minimal-value one.
- **Tentative-Method 2**: among all points satisfying 1st order and 2nd order necessary conditions, find the minimal-value one.
- More detailed steps:

  Step 1: Find all stationary points (candidates) by solving $\nabla f(\mathbf{x}) = 0$;

  Step 2 (optional): Find all candidates s.t. $\nabla^2 f(\mathbf{x}) \succeq 0$.

  Step 3: Among all candidates, find the one with minimal value.

Tentative Use of Optimality Condition (cont.)

- **Example 1:** $\min \ \frac{1}{2}(x-b)^2$
- Set gradient $x - b = 0$, get $x = b$, so this is the minimal solution.

- **Example 2:** $\min \ x^2 + 2y^2 + 3xy.$
- First order condition: $2x + 3y = 0, 4y + 3x = 0$, which implies $x = y = 0$.

- The only stationary point $(x, y) = (0, 0)$ is the global minimum, with value 0
  $\rightarrow$ WRONG CONCLUSION!

Tentative-Method 1 and 2 are Flawed!

- Tentative-Method 1 and 2 in the last page are FLAWED.
- **Fact:** A global minimum:
  - is a stationary point;
  - has the smallest function value among all stationary points

  Tentative-method 1: check all stationary points, and find the one with the minimal function value.

  Question: is it always a global minimum?
- **Logic**: "A is B", does not mean "B is A".
  - Analogy: A president is an official with the most power in a country.
  - But the official with the most power in a country may or may not be a president (could be a prime minster...)

Correct Use of Optimality Condition

- **Example 1:** $\min \; \frac{1}{2}(x-b)^2$
- **Step 1**: Set gradient $x - b = 0$, get $x = b$.
  Step 2: Since $f(b) = 0 \leq f(x), \forall x$, so b is the minimizer.
- **Example 2:** $\min \; x^2 + 2y^2 + 3xy.$
- **Step 1**: First order condition: $2x + 3y = 0, 4y + 3x = 0$, which implies $x = y = 0$.
  Step 2: Let $x = -1.5, y = 1$, then
  $f(x,y) = x^2 + 2y^2 + 3xy = 2.25 + 2 - 4.5 = -0.25 < 0 = f(0,0)$.
  Thus $(0,0)$ is not a global minimizer.
  Thus the function has no global minimizer.
- **Alternative method**:
  - $f(x,y) = x^2 + 2y^2 + 3xy = (x + 1.5y)^2 - 0.25y^2$.
  - Let $y = M, x = -1.5M$, then $f(x,y) = -0.25M^2$.
  - As $M \to \infty$, $f(x,y) \to -\infty$, so there is no global minimizer!

Review Questions of Math Preliminaries

- **Q1:** Is $\sum_{n=1}^{\infty} \frac{1}{n}$ finite? Is $\sum_{n=1}^{\infty} \frac{1}{n^2}$ finite? Is $\sum_{n=1}^{\infty} \frac{1}{n^{0.5}}$ finite?
- **Q2:**(derivative) What are $\frac{de^x}{dx}$, $\frac{d \log x}{dx}$ and $\frac{d \log(1+e^x)}{dx}$?
- **Q3:**(Taylor series) Suppose $f : \mathbb{R} \to \mathbb{R}$ is a scalar function. Write down the Taylor series of f at a point a.
- **Q4:** If a real symmetric matrix A has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigen-vectors $v_1, \ldots, v_n$, how to express A as λ_i's and v_i's?
- **Q5:** If $\lambda_1 \neq \lambda_2$ are two distinct eigenvalues of a real symmetric matrix A, and $Av_1 = \lambda_1 v_1$ and $Av_2 = \lambda_2 v_2$. Prove that $v_1 \perp v_2$.
- **Q6:** Which of the following notions are only defined for square matrices? (i) Inverse; (ii) Rank; (iii) Eigenvalues; (iv) Singular values; (v) PSD.

Summary

Math preliminaries: calculus; linear algebra (eigenvalues; PSD; etc.); fixed point theorem.

Definitions of **Local minimizer** (minimum) and **global minimizer** (minimum).

Necessary conditions for local minimizers: gradient equals 0 and PSD Hessian.

- **Stationary point**: satisfy first order condition
- **Saddle point**: stationary point, but not local-min or local-max

Sufficient condition for local minimizers : gradient equals 0 and PD Hessian.

Finding stationary points and global-min: for simple functions, can directly solve equations and compare values.

- **Caveat:** It is a common mistake to assume “stationary point with the minimal value among stationary points is a global-min”.
- Need extra conditions (check definition, or see next lecture)