Tutorial notes

2024-2025 Optimization theory

1 Basics

Definition 1 (Gradient). *Given a function $f : \mathbb{R}^n \to \mathbb{R}$, and assuming f is well-defined in a neighborhood of a point x, if there exists a vector $g \in \mathbb{R}^n$ such that*

$$\lim_{p \to 0} \frac{f(x+p) - f(x) - g^T p}{\|p\|} = 0, \tag{1}$$

*where $\|\cdot\|$ is any vector norm, then f is said to be **differentiable** (or **Fréchet differentiable**) at x. In this case, g is called the **gradient** of f at x and is denoted as $\nabla f(x)$. If $\nabla f(x)$ exists for every x in a region D, then f is differentiable on D.*

If f has a gradient at x, setting $p = \varepsilon e_i$ in the definition, where e_i is the unit vector with 1 in the i-th component, it follows that the i-th component of $\nabla f(x)$ is given by $\frac{\partial f(x)}{\partial x_i}$. Hence, the gradient can be expressed as:

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \quad \frac{\partial f(x)}{\partial x_2}, \quad \ldots, \quad \frac{\partial f(x)}{\partial x_n} \right]^T. \tag{2}$$

Definition 2 (Hessian Matrix). *If the second-order partial derivatives $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ exist for all $i, j = 1, 2, \ldots, n$ at a point x, then the Hessian matrix of f at x is defined as:*

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}. \tag{3}$$

*If $\nabla^2 f(x)$ exists at every point x in a region D, then f is said to be **twice differentiable** in D. If $\nabla^2 f(x)$ is continuous in D, then f is called **twice continuously differentiable** in D, and it can be shown that the Hessian matrix is symmetric.*

The definition of the gradient for multivariable functions can be extended to cases where the variable is a matrix. For a function $f(X)$ where X is an $m \times n$ matrix, if there exists a matrix $G \in \mathbb{R}^{m \times n}$ such that

$$\lim_{V \to 0} \frac{f(X+V) - f(X) - \langle G, V \rangle}{\|V\|} = 0, \tag{4}$$

where $\|\cdot\|$ denotes any matrix norm, then f is said to be **Fréchet differentiable** at X, and G is called the **gradient** of f at X under the Fréchet derivative definition.

Let $\frac{\partial f}{\partial x_{ij}}$ denote the partial derivative of f with respect to x_{ij}. The gradient of the matrix function

$f(X)$ is given by:

$$\nabla f(X) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}. \tag{5}$$

In practical applications, the definition and use of **Fréchet differentiability** for matrix functions are often complex. Therefore, we introduce another definition—**Gâteaux differentiability**.

Definition 3 (Gâteaux Differentiability)**.** *Let $f(X)$ be a function of a matrix variable. If for any direction $V \in \mathbb{R}^{m\times n}$, there exists a matrix $G \in \mathbb{R}^{m\times n}$ such that*

$$\lim_{t\to 0} \frac{f(X+tV) - f(X) - t\langle G, V\rangle}{t} = 0, \tag{6}$$

*then f is said to be **Gâteaux differentiable** with respect to X. The matrix G satisfying the above equation is called the **Gâteaux derivative** of f at X.*

It can be shown that if f is **Fréchet differentiable**, then it is also **Gâteaux differentiable**, and the two derivatives are equivalent in this case.

Linear Function: $f(X) = \mathrm{tr}(AX^TB)$, where $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{m\times p}$, and $X \in \mathbb{R}^{m\times n}$. For any direction $V \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, we have:

$$\lim_{t\to 0} \frac{f(X+tV) - f(X)}{t} = \lim_{t\to 0} \frac{\mathrm{tr}(A(X+tV)^TB) - \mathrm{tr}(AX^TB)}{t} \tag{7}$$

$$= \mathrm{tr}(AV^TB) = \langle BA, V\rangle. \tag{8}$$

Thus, the gradient is given by:

$$\nabla f(X) = BA. \tag{9}$$

Quadratic Function: $f(X,Y) = \frac{1}{2}\|XY - A\|_F^2$, where $(X,Y) \in \mathbb{R}^{m\times p} \times \mathbb{R}^{p\times n}$.

For the variable Y, taking any direction V and a sufficiently small $t \in \mathbb{R}$, we get:

$$f(X, Y+tV) - f(X,Y) = \frac{1}{2}\|X(Y+tV) - A\|_F^2 - \frac{1}{2}\|XY - A\|_F^2 \tag{10}$$

$$= \langle XV, XY - A\rangle + \frac{1}{2}t^2\|XV\|_F^2. \tag{11}$$

$$= t\langle V, X^T(XY - A)\rangle + \mathcal{O}(t^2). \tag{12}$$

By definition, we obtain:

$$\frac{\partial f}{\partial Y} = X^T(XY - A). \tag{13}$$

Similarly, for the variable X, we obtain:

$$\frac{\partial f}{\partial X} = (XY - A)Y^T. \tag{14}$$

***ln-det* Function:** $f(X) = \ln(\det(X))$, where $X \in S^n_{++}$ and X is positive definite. For any direction $V \in S^n$ and $t \in \mathbb{R}$, we have:

$$f(X + tV) - f(X) = \ln(\det(X + tV)) - \ln(\det(X)) \tag{15}$$

$$= \ln(\det(X^{1/2}(I + tX^{-1/2}VX^{-1/2})X^{1/2})) - \ln(\det(X)) \tag{16}$$

$$= \ln(\det(I + tX^{-1/2}VX^{-1/2})). \tag{17}$$

Since $X^{-1/2}VX^{-1/2}$ is a symmetric matrix, it can be diagonalized orthogonally. Let its eigenvalues be $\lambda_1, \lambda_2, \ldots, \lambda_n$, then:

$$\ln(\det(I + tX^{-1/2}VX^{-1/2})) = \ln \prod_{i=1}^{n}(1 + t\lambda_i). \tag{18}$$

Expanding the logarithm:

$$= \sum_{i=1}^{n} \ln(1 + t\lambda_i) = \sum_{i=1}^{n} t\lambda_i + \mathcal{O}(t^2) \tag{19}$$

$$= t\,\mathrm{tr}(X^{-1/2}VX^{-1/2}) + \mathcal{O}(t^2). \tag{20}$$

Thus, we conclude:

$$\nabla f(X) = (X^{-1})^T. \tag{21}$$

Definition 4 (Extended Real-Valued Function)**.** *Let $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{\pm\infty\}$ be the extended real number space. A mapping $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called an* ***extended real-valued function****.*

As in mathematical analysis, we define:

$$-\infty < a < +\infty, \quad \forall a \in \mathbb{R}. \tag{22}$$

$$(+\infty) + (+\infty) = +\infty, \quad +\infty + a = +\infty, \quad \forall a \in \mathbb{R}. \tag{23}$$

Definition 5 (Proper Function)**.** *Given an extended real-valued function f and a non-empty set $\mathcal{X}$, if there exists $x \in \mathcal{X}$ such that $f(x) < +\infty$, and for any $x \in \mathcal{X}$, we have $f(x) > -\infty$, then f is called* ***proper*** *with respect to $\mathcal{X}$.*

In summary, a proper function has the properties that it "takes at least one finite value" and "never takes the value negative infinity."

Definition 6 (α-Sublevel Set)**.** *For an extended real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the set*

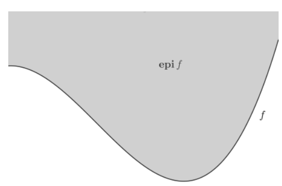
$$C_\alpha = \{x \mid f(x) \leq \alpha\} \tag{24}$$

*is called the α-**sublevel set** of f.*

Definition 7 (Epigraph)**.** *For an extended real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the set*

$$\operatorname{epi} f = \{(x, t) \in \mathbb{R}^{n+1} \mid f(x) \leq t\} \tag{25}$$

is called the ***epigraph*** *of f.*

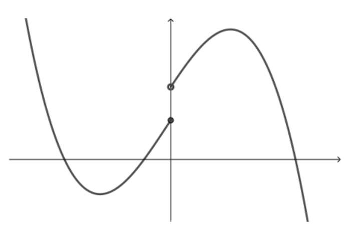


Definition 8 (Closed Function)**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be an extended real-valued function. If* epi f *is a closed set, then f is called a* ***closed function****.*

Definition 9 (Lower Semicontinuous Function)**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be an extended real-valued function. If for any $x \in \mathbb{R}^n$, we have*

$$\liminf_{y \to x} f(y) \geq f(x), \tag{26}$$

then $f(x)$ is called a ***lower semicontinuous function****.*



Although the definitions of these two types of functions appear different on the surface, closed functions and lower semicontinuous functions are equivalent.

Theorem 1. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be an extended real-valued function. Then the following statements are equivalent:*

1. *Any α-sublevel set of $f(x)$ is a closed set;*
2. *$f(x)$ is lower semicontinuous;*
3. *$f(x)$ is a closed function.*

Basic Operations Preserve Closure (Lower Semicontinuity)

Simple operations between closed (lower semicontinuous) functions preserve their properties:

- **Addition:** If f and g are both closed (lower semicontinuous) functions, and $\operatorname{dom} f \cap \operatorname{dom} g \neq \emptyset$, then $f + g$ is also a closed (lower semicontinuous) function. The domain condition ensures that the undefined expression $(-\infty) + (+\infty)$ does not occur.
- **Composition with an Affine Mapping:** If f is a closed (lower semicontinuous) function, then $f(Ax + b)$ is also a closed (lower semicontinuous) function.
- **Pointwise Supremum:** If each function f_α is closed (lower semicontinuous), then $\sup_\alpha f_\alpha(x)$ is also a closed (lower semicontinuous) function.

2 Convex function

Definition 10 (Convex Function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain $\operatorname{dom} f$ is a convex set and it satisfies:*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y), \tag{27}$$

for all $x, y \in \operatorname{dom} f$ and $0 \leq \theta \leq 1$.

- If f is a convex function, then $-f$ is a **concave function**.
- If for all $x, y \in \operatorname{dom} f$ with $x \neq y$ and $0 < \theta < 1$, we have

  $$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y), \tag{28}$$

  then f is called a **strictly convex function**.

Convex Functions:

- **Affine Function:** For any $a, b \in \mathbb{R}$, the function $ax + b$ is convex on $\mathbb{R}$.
- **Exponential Function:** For any $a \in \mathbb{R}$, the function e^{ax} is convex on $\mathbb{R}$.
- **Power Function:** For $\alpha \geq 1$ or $\alpha \leq 0$, the function x^α is convex on $\mathbb{R}_{++}$.
- **Absolute Power:** For $p \geq 1$, the function $|x|^p$ is convex on $\mathbb{R}$.
- **Negative Entropy:** The function $x \log x$ is convex on $\mathbb{R}_{++}$.

Concave Functions:

- **Affine Function:** For any $a, b \in \mathbb{R}$, the function $ax + b$ is concave on $\mathbb{R}$.
- **Power Function:** For $0 \leq \alpha \leq 1$, the function x^α is concave on $\mathbb{R}_{++}$.
- **Logarithm Function:** The function $\log x$ is concave on $\mathbb{R}_{++}$.

All affine functions are both convex and concave. All norms are convex functions.

Examples in Euclidean Space $\mathbb{R}^n$

- **Affine Function:** $f(x) = a^T x + b$
- **Norm:**
$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad (p \geq 1) \tag{29}$$
In particular, the ∞-norm is given by:
$$\|x\|_\infty = \max_k |x_k| \tag{30}$$

Examples in Matrix Space $\mathbb{R}^{m \times n}$

- **Affine Function:**
$$f(X) = \mathrm{tr}(A^T X) + b = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} X_{ij} + b \tag{31}$$
- **Spectral Norm:**
$$f(X) = \|X\|_2 = \sigma_{\max}(X) = \left(\lambda_{\max}(X^T X) \right)^{1/2} \tag{32}$$

Definition 11 (Strongly Convex Function - Definition 1)**.** *If there exists a constant $m > 0$ such that*
$$g(x) = f(x) - \frac{m}{2} \|x\|^2 \tag{33}$$
*is a convex function, then $f(x)$ is called a **strongly convex function**, where m is the **strong convexity parameter**. For convenience, we also refer to $f(x)$ as an m-strongly convex function.*

Definition 12 (Strongly Convex Function - Definition 2)**.** *If there exists a constant $m > 0$ such that for any $x, y \in \operatorname{dom} f$ and $\theta \in (0,1)$, we have*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|^2, \tag{34}$$

*then $f(x)$ is called a **strongly convex function**, where m is the **strong convexity parameter**.*

If f is a strongly convex function and attains a minimum value, then the minimizer of f is unique.

Restrict the function to any line and then determine whether the corresponding one-dimensional function is convex.

Theorem 2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for each $x \in \operatorname{dom} f$ and direction $v \in \mathbb{R}^n$, the function $g : \mathbb{R} \to \mathbb{R}$, defined as*

$$g(t) = f(x+tv), \quad \operatorname{dom} g = \{t \mid x + tv \in \operatorname{dom} f\}, \tag{35}$$

is convex with respect to t.

Example: Consider the function $f(X) = -\log\det X$, which is convex, where the domain is $\operatorname{dom} f = S^n_{++}$. For any $X \succ 0$ and direction $V \in S^n$, restricting f to the line $X + tV$ (where t satisfies $X + tV \succ 0$) gives:

$$g(t) = -\log\det(X+tV) = -\log\det X - \log\det(I + tX^{-1/2}VX^{-1/2}) \tag{36}$$

$$= -\log\det X - \sum_{i=1}^{n} \log(1 + t\lambda_i), \tag{37}$$

where λ_i are the eigenvalues of $X^{-1/2}VX^{-1/2}$. For every $X \succ 0$ and direction V, $g(t)$ is convex in t, and thus f is convex.

Proof. **Necessity:** Assume $f(x)$ is convex. We need to show that $g(t) = f(x+tv)$ is convex. First, we prove that $\operatorname{dom} g$ is a convex set. For any $t_1, t_2 \in \operatorname{dom} g$ and $\theta \in (0,1)$, we have:

$$x + t_1 v \in \operatorname{dom} f, \quad x + t_2 v \in \operatorname{dom} f. \tag{38}$$

Since $\operatorname{dom} f$ is convex, it follows that:

$$x + (\theta t_1 + (1-\theta)t_2)v \in \operatorname{dom} f. \tag{39}$$

This implies that $\theta t_1 + (1-\theta)t_2 \in \operatorname{dom} g$, meaning $\operatorname{dom} g$ is convex.

Moreover, we have:

$$\begin{aligned} g(\theta t_1 + (1-\theta)t_2) &= f(x + (\theta t_1 + (1-\theta)t_2)v) \\ &= f(\theta(x+t_1v) + (1-\theta)(x+t_2v)) \\ &\leq \theta f(x+t_1v) + (1-\theta)f(x+t_2v) \\ &= \theta g(t_1) + (1-\theta)g(t_2). \end{aligned}$$

Combining the above results, we conclude that $g(t)$ is convex.

Sufficiency: We now show that $\operatorname{dom} f$ is convex. Take $v = y - x$, and let $t_1 = 0, t_2 = 1$. Since $\operatorname{dom} g$ is convex, we know:

$$\theta \cdot 0 + (1-\theta) \cdot 1 \in \operatorname{dom} g. \tag{40}$$

This implies that $x + (1-\theta)y \in \operatorname{dom} f$, proving that $\operatorname{dom} f$ is convex.

Now, using the convexity of $g(t) = f(x + tv)$, we obtain:

$$\begin{aligned} g(1-\theta) &= g(\theta t_1 + (1-\theta)t_2) \\ &\le \theta g(t_1) + (1-\theta)g(t_2) \\ &= \theta g(0) + (1-\theta)g(1) \\ &= \theta f(x) + (1-\theta)f(y). \end{aligned}$$

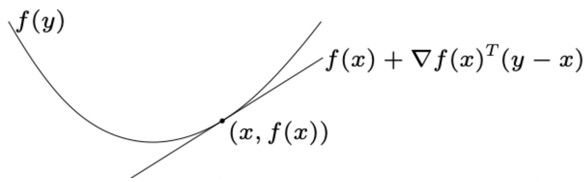
On the other hand, the left-hand side satisfies:

$$g(1-\theta) = f(x + (1-\theta)(y-x)) = f(\theta x + (1-\theta)y). \tag{41}$$

This proves that $f(x)$ is convex. □

Theorem 3 (First-Order Condition)**.** *For a differentiable function f defined on a convex set, f is convex if and only if*

$$f(y) \ge f(x) + \nabla f(x)^T (y - x), \quad \forall x, y \in \operatorname{dom} f. \tag{42}$$



Proof. **Necessity:** Suppose f is convex. Then for any $x, y \in \operatorname{dom} f$ and $t \in (0, 1)$, we have:

$$tf(y) + (1-t)f(x) \ge f(x + t(y-x)). \tag{43}$$

Rearranging and dividing both sides by t (noting that $t > 0$), we obtain:

$$f(y) - f(x) \ge \frac{f(x + t(y-x)) - f(x)}{t}. \tag{44}$$

Taking the limit as $t \to 0$ and using the definition of the directional derivative, we get:

$$f(y) - f(x) \ge \lim_{t \to 0} \frac{f(x + t(y-x)) - f(x)}{t} = \nabla f(x)^T (y - x). \tag{45}$$

The last equality follows from the definition of the directional derivative.

Sufficiency: For any $x, y \in \operatorname{dom} f$ and $t \in (0, 1)$, define:

$$z = tx + (1-t)y. \tag{46}$$

Applying the first-order condition at z, we obtain:

$$f(x) \geq f(z) + \nabla f(z)^T (x - z), \tag{47}$$

$$f(y) \geq f(z) + \nabla f(z)^T (y - z). \tag{48}$$

Multiplying the first inequality by t and the second inequality by $1 - t$, then summing both sides, we get:

$$tf(x) + (1 - t)f(y) \geq f(z) + 0. \tag{49}$$

This is exactly the definition of a convex function, proving the sufficiency. □

Theorem 4. *Suppose f is a differentiable function. Then f is convex if and only if* $\operatorname{dom} f$ *is a convex set and ∇f is a* ***monotone mapping****, i.e.,*

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0, \quad \forall x, y \in \operatorname{dom} f. \tag{50}$$

Proof. **Necessity:** If f is differentiable and convex, then by the first-order condition, we have:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \tag{51}$$

$$f(x) \geq f(y) + \nabla f(y)^T (x - y). \tag{52}$$

Adding these two inequalities together, we obtain the desired result:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0. \tag{53}$$

Sufficiency: If ∇f is a monotone mapping, we construct an auxiliary function:

$$g(t) = f(x + t(y - x)), \quad g'(t) = \nabla f(x + t(y - x))^T (y - x). \tag{54}$$

By the monotonicity of ∇f, we know that $g'(t) \geq g'(0)$ for all $t \geq 0$. Thus,

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)dt \\ \geq g(0) + g'(0) = f(x) + \nabla f(x)^T (y - x). \end{aligned}$$

This is precisely the first-order condition for convexity, proving the sufficiency. □

Theorem 5. *A function $f(x)$ is convex if and only if its epigraph* $\operatorname{epi} f$ *is a convex set.*

Proof. **Necessity:** Suppose f is a convex function. Then for any $(x_1, y_1), (x_2, y_2) \in \operatorname{epi} f$ and any $t \in [0, 1]$, we have:

$$ty_1 + (1 - t)y_2 \geq tf(x_1) + (1 - t)f(x_2) \geq f(tx_1 + (1 - t)x_2). \tag{55}$$

Therefore,

$$(tx_1 + (1 - t)x_2, ty_1 + (1 - t)y_2) \in \operatorname{epi} f, \quad t \in [0, 1]. \tag{56}$$

This shows that $\operatorname{epi} f$ is convex.

Sufficiency: Suppose epi f is a convex set. Then for any $x_1, x_2 \in \mathrm{dom}\, f$ and $t \in [0,1]$, we have:

$$(tx_1 + (1-t)x_2, tf(x_1) + (1-t)f(x_2)) \in \mathrm{epi}\, f. \tag{57}$$

By the definition of the epigraph, this implies:

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2). \tag{58}$$

Thus, $f(x)$ is convex. □

Theorem 6 (Second-Order Condition)**.** *Suppose f is a twice continuously differentiable function defined on a convex set. Then:*

- *f is convex if and only if*
$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \mathrm{dom}\, f. \tag{59}$$
- *If $\nabla^2 f(x) \succ 0$ for all $x \in \mathrm{dom}\, f$, then f is strictly convex.*

Example: A quadratic function

$$f(x) = \frac{1}{2}x^T P x + q^T x + r, \quad \text{where } P \in S^n \tag{60}$$

has

$$\nabla f(x) = Px + q, \quad \nabla^2 f(x) = P. \tag{61}$$

Thus, f is convex if and only if $P \succeq 0$.

Proof. **Necessity:** Suppose $f(x)$ is convex, but assume for contradiction that $\nabla^2 f(x) \not\succeq 0$, meaning there exists a nonzero vector $v \in \mathbb{R}^n$ such that:

$$v^T \nabla^2 f(x) v < 0. \tag{62}$$

Expanding $f(x+tv)$ using Peano's remainder theorem, we obtain:

$$f(x+tv) = f(x) + t\nabla f(x)^T v + \frac{t^2}{2} v^T \nabla^2 f(x) v + o(t^2). \tag{63}$$

Dividing both sides by t^2 and rearranging, we get:

$$\frac{f(x+tv) - f(x) - t\nabla f(x)^T v}{t^2} = \frac{1}{2} v^T \nabla^2 f(x) v + o(1). \tag{64}$$

As $t \to 0$, the right-hand side becomes negative, contradicting the first-order convexity condition. Thus, $\nabla^2 f(x) \succeq 0$ must hold.

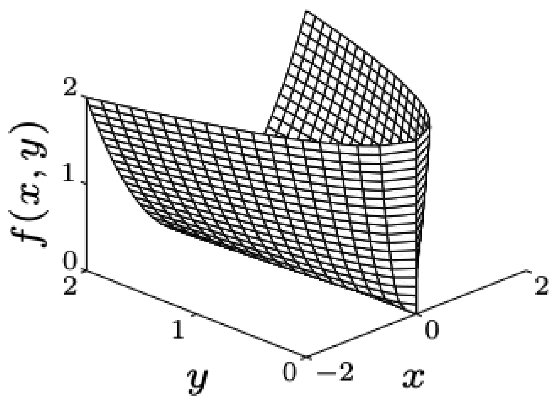
Sufficiency: Suppose $f(x)$ satisfies the second-order condition $\nabla^2 f(x) \succeq 0$. Expanding $f(y)$ using Taylor's theorem,

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x + t(y-x))(y-x), \tag{65}$$

where $t \in (0,1)$ is some constant depending on x and y. Since $\nabla^2 f(x) \succeq 0$, it follows that:

$$f(y) \geq f(x) + \nabla f(x)^T (y-x). \tag{66}$$

By the first-order condition, this implies that $f(x)$ is convex. Moreover, if $\nabla^2 f(x) \succ 0$, then the inequality holds strictly for $x \neq y$, proving that $f(x)$ is strictly convex. □



Least Squares Function:

$$f(x) = \|Ax - b\|_2^2 \tag{67}$$

$$\nabla f(x) = 2A^T(Ax - b), \quad \nabla^2 f(x) = 2A^T A. \tag{68}$$

For any A, $f(x)$ is a convex function.

Quadratic-over-Linear Function:

$$f(x, y) = \frac{x^2}{y} \tag{69}$$

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0. \tag{70}$$

Thus, $f(x, y)$ is convex on the domain $\{(x, y) \mid y > 0\}$.

Log-Sum-Exp Function:

$$f(x) = \log \sum_{k=1}^{n} \exp x_k \tag{71}$$

$$\nabla^2 f(x) = \frac{1}{\mathbf{1}^T z} \operatorname{diag}(z) - \frac{1}{(\mathbf{1}^T z)^2} z z^T, \quad z_k = \exp x_k. \tag{72}$$

To prove $\nabla^2 f(x) \succeq 0$, it suffices to show that for any v,

$$v^T \nabla^2 f(x) v = \frac{(\sum_k z_k v_k^2)(\sum_k z_k) - (\sum_k z_k v_k)^2}{(\sum_k z_k)^2} \geq 0. \tag{73}$$

By the Cauchy-Schwarz inequality, we have $(\sum_k z_k v_k)^2 \leq (\sum_k z_k v_k^2)(\sum_k z_k)$, proving convexity.

Geometric Mean:

$$f(x) = \left(\prod_{k=1}^{n} x_k \right)^{1/n}, \quad x \in \mathbb{R}_{++}^n \tag{74}$$

is a concave function.

Jensen's Inequality: Basic Jensen's Inequality: If f is a convex function, then for $0 \leq \theta \leq 1$,

$$f(\theta x + (1 - \theta) y) \leq \theta f(x) + (1 - \theta) f(y). \tag{75}$$

Probabilistic Jensen's Inequality: If f is convex, then for any random variable z,

$$f(\mathbb{E}[z]) \leq \mathbb{E}[f(z)]. \tag{76}$$

The basic Jensen's inequality can be seen as a special case of the probabilistic version when z follows a two-point distribution:

$$\text{prob}(z = x) = \theta, \quad \text{prob}(z = y) = 1 - \theta. \tag{77}$$

References