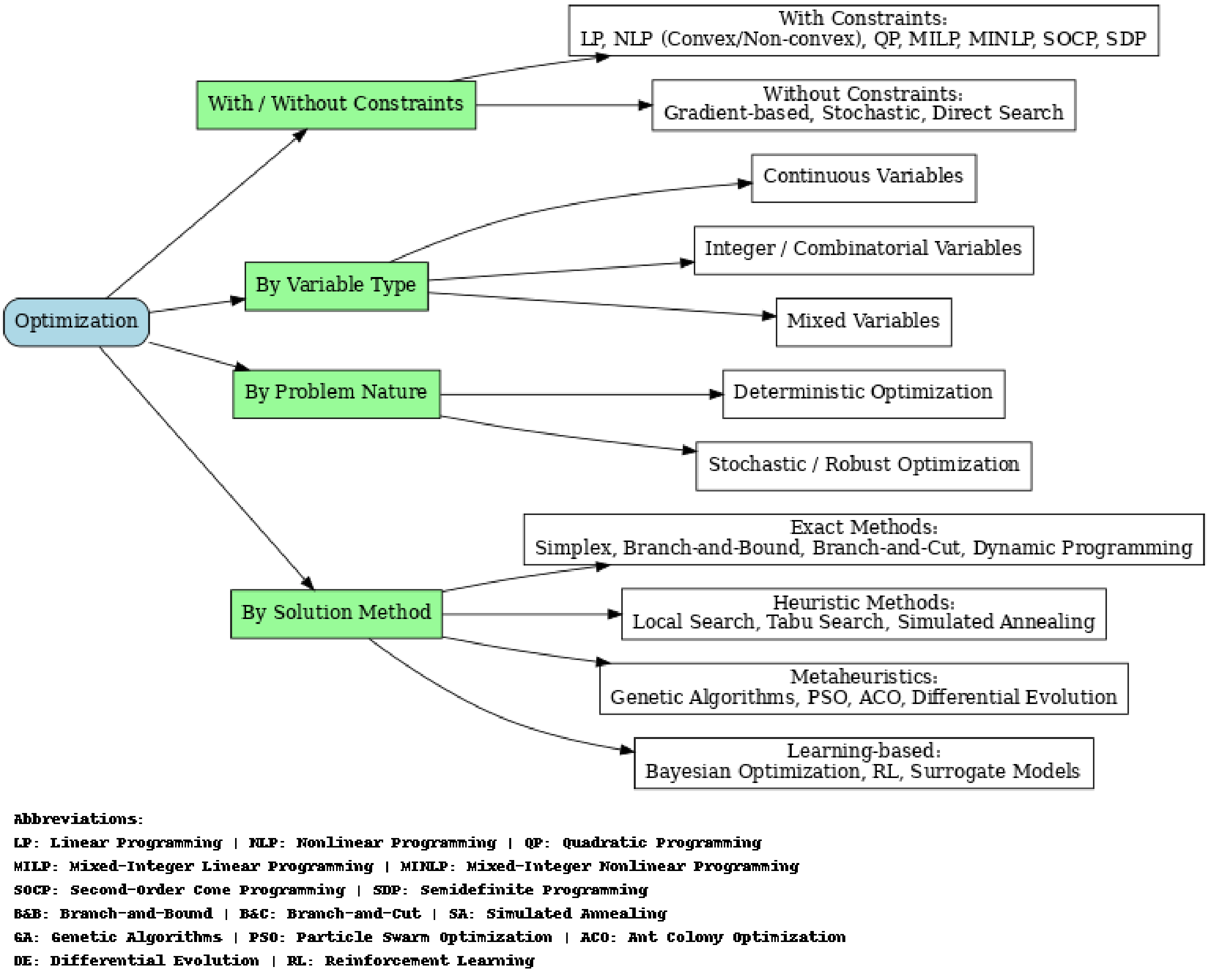

Figure 1: Comprehensive diagram of the main types of optimization and their relationships, with abbreviations explained below the figure.

Chapter 1

Differential Calculus and Convexity

1.1 Review of Differential Calculus

1.1.1 Normed Vector Spaces

Let V be a vector space over the field $\mathbb{R}$.

Definition 1.1.1. *An* ***inner product*** *on V is a mapping*

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

that is bilinear, symmetric, and positive definite, meaning it satisfies:

1. *$\langle u, \cdot \rangle : V \to \mathbb{R}$ is linear for all $u \in V$,*
2. *$\langle \cdot, v \rangle : V \to \mathbb{R}$ is linear for all $v \in V$,*
3. *$\langle u, v \rangle = \langle v, u \rangle$ for all $u, v \in V$,*
4. *$\langle v, v \rangle = 0 \Leftrightarrow v = 0$, and $\langle v, v \rangle \geq 0$ for all $v \in V$.*

1.1.2 Directional Derivative

Definition 1.1.2 (Directional Derivative [18, Section 2.1])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $x \in \mathbb{R}^n$. The directional derivative of f at x in the direction $d \in \mathbb{R}^n$ is defined as:*

$$D_f(x; d) = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t}.$$

If f is differentiable at x, then $D_f(x; d) = \nabla f(x)^T d$.

Example 1.1.1. *Let $f(x, y) = x^2 + y^2$. Compute the directional derivative at point* $(1, 2)$ *in the direction $d =$* $(3, 4)$*. Solution: $\nabla f(x, y) =$* $(2x, 2y)$*. At* $(1, 2)$*, $\nabla f(1, 2) =$* $(2, 4)$*. Thus:*

$$D_f((1, 2); (3, 4)) = (2, 4) \cdot (3, 4) = 6 + 16 = 22.$$

—

1.1.3 Canonical Basis and Norms

1.1.3.1 Canonical Basis

Let $e_1, e_2, \dots, e_n$ denote the elements of the canonical basis of $\mathbb{R}^n$, where e_i is the vector in $\mathbb{R}^n$ given by:

$$(e_i)_j = \delta_{ij} = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i, \end{cases} \quad \text{for all } i, j = 1, 2, \dots, n,$$

where δ_{ij} is the Kronecker delta symbol.

—

1.1.3.2 Dot Product

Definition 1.1.3 (Dot Product [18, Section 2.3]). *For any $x, y \in \mathbb{R}^n$, the dot product $\langle x, y \rangle$ is defined as:*

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i.$$

Two vectors $x, y \in \mathbb{R}^n$ are orthogonal (denoted $x \perp y$) if $\langle x, y \rangle = 0$.

—

1.1.3.3 Euclidean Norm

Definition 1.1.4 (Euclidean Norm [18, Section 2.3]). *For any $x \in \mathbb{R}^n$, the Euclidean norm of x, denoted $\|x\|$, is defined as:*

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Properties of a norm (and thus of the Euclidean norm):

(i) $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$.

(ii) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

(iii) $\|0\| = 0$ and $\|x\| > 0$ if $x \neq 0$.

—

Definition 1.1.5 (Open Ball [18, Section 2.3]). *For every $x \in \mathbb{R}^n$ and $r > 0$, the open ball centered at x with radius r is defined as:*

$$B(x; r) = \{y \in \mathbb{R}^n : \|y - x\| < r\}.$$

—

Definition 1.1.6 (Convergence). *If $(x^{(k)})$ is a sequence in $\mathbb{R}^n$ and x is an element of $\mathbb{R}^n$, we say that $x^{(k)}$ converges to x (denoted $x^{(k)} \to x$) if $\|x^{(k)} - x\| \to 0$ as $k \to \infty$. Equivalently, $x^{(k)} \to x$ if and only if $x_i^{(k)} \to x_i$ in $\mathbb{R}$ for each component i.*

—

Definition 1.1.7 (Interior, Open and Closed Sets [18, Section 2.3]). *Let $U \subset \mathbb{R}^n$.*

1. *The **interior** of U is the set of all $x \in U$ such that there exists $r > 0$ with $B(x; r) \subset U$.*
2. *The set U is **open** if for every $x \in U$, there exists $r > 0$ with $B(x; r) \subset U$.*
3. *The set U is **closed** if for every sequence $\{x^{(k)}\} \subset U$ such that $x^{(k)} \to x \in \mathbb{R}^n$, we have $x \in U$.*

—

Definition 1.1.8 (Segment). *If $a, b \in \mathbb{R}^n$, we denote by $[a, b]$ the subset of $\mathbb{R}^n$ given by:*

$$[a, b] = \{a + t(b - a) : t \in [0, 1]\}.$$

The set $[a, b]$ is called the segment connecting a to b.

—

Definition 1.1.9 (Cauchy-Schwarz Inequality [18, Section 2.3]). *For all $x, y \in \mathbb{R}^n$:*

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

—

Definition 1.1.10 (Directional Derivative). *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ where $f(x_0)$ is defined. The directional derivative of f at x_0 in the direction $d \in \mathbb{R}^n$ is defined by:*

$$f'_d(x_0) = \lim_{t \to 0^+} \frac{f(x_0 + td) - f(x_0)}{t},$$

if it exists. This derivative gives the rate of change of f at x_0 in the direction d.

—

Definition 1.1.11 (Fréchet Differentiability [2, Section 9.1]). *A function f is said to be **Fréchet differentiable** at $x_0 \in \mathbb{R}^n$ if there exists a continuous linear map $L(x_0) : \mathbb{R}^n \to \mathbb{R}$ such that:*

$$\lim_{d \to 0} \frac{f(x_0 + d) - f(x_0) - L(x_0) \cdot d}{\|d\|} = 0.$$

*The map $L(x_0)$ is called the **derivative** of f at x_0.*

Example 1.1.2 (Example 1.1.1)**.** *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by:*

$$f(x_1, x_2) = x_1 - x_2^2.$$

For any $d \in \mathbb{R}^2$, we have:

$$\begin{aligned}
\lim_{t \to 0^+} \frac{f(x+td) - f(x)}{t} &= \lim_{t \to 0^+} \frac{f(x_1 + td_1, x_2 + td_2) - f(x_1, x_2)}{t} \\
&= (1, -2x_2) \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = f'(x) \cdot d.
\end{aligned}$$

—

Example 1.1.3 (Example 1.1.2)**.** *Consider the function $f : \mathbb{R}^2 \setminus \{(0,0)\} \to \mathbb{R}$ defined by:*

$$f(x_1, x_2) = \begin{cases} x_1^2 x_2 / (x_1^2 + x_2^2), & (x_1, x_2) \neq (0,0), \\ 0, & (x_1, x_2) = (0,0). \end{cases}$$

Obviously, the function f is continuous at $(0,0)$. Indeed, for $x_1 = r\cos\theta$, $x_2 = r\sin\theta$, with $r > 0$ and $\theta \in (0, 2\pi)$, we have:

$$\lim_{(x_1,x_2) \to (0,0)} f(x_1, x_2) = \lim_{r \to 0^+} \frac{r^3 \cos^2\theta \sin\theta}{r^2} = \lim_{r \to 0^+} r\cos^2\theta \sin\theta = 0 = f(0,0).$$

Furthermore, f admits partial derivatives at $(0,0)$ since:

$$\frac{\partial f}{\partial x_1}(0,0) = \lim_{d_1 \to 0} \frac{f(d_1, 0) - f(0,0)}{d_1} = 0,$$

and

$$\frac{\partial f}{\partial x_2}(0,0) = \lim_{d_2 \to 0} \frac{f(0, d_2) - f(0,0)}{d_2} = 0.$$

However, f is not Fréchet differentiable (and thus not differentiable) at $(0,0)$ because:

$$\lim_{(d_1,d_2) \to (0,0)} \frac{f(d_1, d_2) - f(0,0) - \left(\frac{\partial f}{\partial x_1}(0,0), \frac{\partial f}{\partial x_2}(0,0)\right) \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}}{\|(d_1, d_2)\|} = \lim_{r \to 0} \cos^2\theta \sin\theta$$

does not exist.

Definition 1.1.12 (Gâteaux Differentiability)**.** *A function f is said to be* ***Gâteaux differentiable*** *(G-differentiable) at $x_0 \in \mathbb{R}^n$ if it admits a directional derivative at x_0 in every direction $d \in \mathbb{R}^n$.*

—

Definition 1.1.13 (Fréchet Differentiability). *A function f is said to be* ***Fréchet differentiable*** *(F-differentiable) at $x_0 \in \mathbb{R}^n$ if there exists a continuous linear map $L(x_0) : \mathbb{R}^n \to \mathbb{R}$ such that:*

$$\lim_{d \to 0} \frac{f(x_0 + d) - f(x_0) - L(x_0) \cdot d}{\|d\|} = 0.$$

The map $L(x_0)$ is called the ***derivative*** *of f at x_0.*
In other words, f is F-differentiable at $x_0 \in \mathbb{R}^n$ if there exists a continuous linear map $L(x_0) : \mathbb{R}^n \to \mathbb{R}$ such that:

$$f(x_0 + d) = f(x_0) + L(x_0) \cdot d + \|d\| \cdot \omega(d),$$

where $\lim_{d \to 0} \omega(d) = 0$.

—

Remark 1.1.1.

1. *A function f can be G-differentiable at $x_0 \in \mathbb{R}^n$ without being continuous at that point.*

2. *If f is F-differentiable at $x_0 \in \mathbb{R}^n$, then it is continuous at that point.*

3. *The notion of F-differentiability is stronger than that of G-differentiability.*

4. *If f is F-differentiable at $x_0 \in \mathbb{R}^n$ with derivative $L(x_0)$, then f is G-differentiable at x_0 and $L(x_0) = f'(x_0)$. The converse is false.*

5. *If f is F-differentiable, then f is differentiable.*

6. *A differentiable function at a point admits partial derivatives at that point. The converse is generally false.*

7. *A function differentiable on an open set $V \subset \mathbb{R}^n$ whose partial derivatives are all continuous on V is said to be of class C^1 on V.*

—

1.1.4 Gradient and Hessian Matrix

Definition 1.1.14 (Gradient [18, Section 2.1]). *The gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at point x is the vector of partial derivatives:*

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)^T.$$

Example 1.1.4 (Example 1.2.1). *Consider the function $f(x_1, x_2, x_3) = e^{x_1} + x_1^2 x_3 - x_1 x_2 x_3$. Therefore, the gradient of f is given by:*

$$\nabla f(x_1, x_2, x_3) = \begin{pmatrix} e^{x_1} + 2x_1x_3 - x_2x_3 \\ -x_1x_3 \\ x_1^2 - x_1x_2 \end{pmatrix}.$$

Definition 1.1.15 (Hessian Matrix [2, Chapter 2])**.** *If f is twice continuously differentiable, the Hessian matrix is the matrix of second derivatives:*

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j=1}^n.$$

Example 1.1.5. *Let $f(x, y) = x^2 + xy + y^2$. Then:*

$$\nabla f(x, y) = \begin{pmatrix} 2x + y \\ x + 2y \end{pmatrix}, \quad \nabla^2 f(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Example 1.1.6 (Example 1.2.2)**.** *Consider the function $f(x_1, x_2, x_3) = e^{x_1} + x_2^2x_3 - x_1x_2x_3$. The Hessian of f is given by:*

$$H(x) = \begin{pmatrix} e^{x_1} & -x_3 & -x_2 \\ -x_3 & 2x_3 & -x_1 \\ -x_2 & -x_1 & 0 \end{pmatrix}.$$

—

Example 1.1.7 (Example 1.2.3)**.** *Define*

$$f(x_1, x_2, x_3) = e^{x_1} + x_2^2x_3 - x_1x_2x_3.$$

The Hessian of f is given by:

$$H(x) = \begin{pmatrix} e^{x_1} & -x_3 & -x_2 \\ -x_3 & 2x_3 & -x_1 \\ -x_2 & -x_1 & 0 \end{pmatrix}.$$

— —

In the following proposition, we provide the relationship between the gradient and Hessian matrix.

Proposition 1.1.1 ([18, Appendix A.4.3, pp. 656–657])**.**

1. *The i-th row of $\nabla^2 f(x)$ is the Jacobian of the i-th component of ∇f.*
2. *We have*
$$\nabla^2 f(x)h = \nabla_h^T \nabla f(x), \quad \forall x \in \mathbb{R}^n, \forall h \in \mathbb{R}^n.$$

Proof.

1. This is obvious.

2. We have:

$$\frac{\partial}{\partial x_i}\langle \nabla f(x), h\rangle = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x) h_j\right) = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) h_j = (\nabla^2 f(x) h)_i.$$

□

Proposition 1.1.2 (Relation between ∇ and ∇^2 [18, Appendix A.4.2–A.4.3, pp. 655–657])**.**

1. *The i-th row of $\nabla^2 f(x)$ is the Jacobian of the i-th component of ∇f.*
2. *We have:*

$$\nabla^2 f(x) h = \nabla \langle \nabla f(x), h\rangle, \quad \forall x \in K, \forall h \in \mathbb{R}^n.$$

—

Proof.

1. This is obvious.
2. We have:

$$\frac{\partial}{\partial x_i}\langle \nabla f(x), h\rangle = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x) h_j\right) = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) h_j = (\nabla^2 f(x) h)_i.$$

□

—

Example 1.1.8 (Example 1.3.7)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is a constant function, then $\nabla f = \nabla^2 f = 0$.*

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

$$f(x) = \langle a, x\rangle, \quad \forall x \in \mathbb{R}^n,$$

where $a \in \mathbb{R}^n$ is a given vector (i.e. f is a linear function). Then we easily compute:

$$\frac{\partial f}{\partial x_k} = a_k, \quad so$$

$$\nabla f = a$$

(the gradient is constant).

Thus,

$$\nabla^2 f = 0.$$

—

Definition 1.1.16. *We say that x^* is a* ***stationary point*** *of f if $\nabla f(x^*) = 0$.*

—

1.1.4.1 Taylor Expansion

Theorem 1.1.1 (Taylor's Theorem [2, Section 9.1])**.** *If f is twice continuously differentiable, then:*

$$f(x+d) \approx f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d.$$

In particular, if one uses the Hessian at x one obtains the second-order approximation

$$f(x+d) = f(x) + \nabla f(x)^\top d + \tfrac{1}{2}\, d^\top \nabla^2 f(x)\, d + o(\|d\|^2) \quad (\|d\| \to 0). \tag{1.1}$$

Proof. The proof proceeds by reducing the multivariate statement to the classical one-variable Taylor theorem applied to the function obtained by restricting f to the line through x in direction d.

Define the one-variable function

$$g : [0,1] \to \mathbb{R}, \qquad g(t) := f(x+td).$$

Because $f \in C^2$ on a neighborhood of the segment $\{x + td : t \in [0,1]\}$, the function g is C^2 on $[0,1]$. By the chain rule we have for every $t \in [0,1]$:

$$g'(t) = \nabla f(x+td)^\top d, \tag{1.2}$$

$$g''(t) = d^\top \nabla^2 f(x+td)\, d. \tag{1.3}$$

Apply the one-variable Taylor theorem with Lagrange remainder to g about $t = 0$ evaluated at $t = 1$. There exists $\xi \in (0,1)$ such that

$$g(1) = g(0) + g'(0) \cdot 1 + \tfrac{1}{2} g''(\xi) \cdot 1^2.$$

Substituting $g(0) = f(x)$ and using (1.6)–(1.7) yields

$$f(x+d) = f(x) + \nabla f(x)^\top d + \tfrac{1}{2}\, d^\top \nabla^2 f(x + \xi d)\, d,$$

which is exactly (**??**) with $\theta = \xi$. This proves the theorem.

To obtain the approximation (1.1), note that continuity of $\nabla^2 f$ implies $\nabla^2 f(x + \theta d) = \nabla^2 f(x) + o(1)$ as $\|d\| \to 0$, and therefore the remainder term in (**??**) equals $\frac{1}{2} d^\top \nabla^2 f(x) d + o(\|d\|^2)$. □

Remark 1.1.2 (Integral form of the remainder)**.** *A convenient and often-used alternative is the integral form of the remainder, obtained by integrating $g''(t)$ twice:*

$$g(1) = g(0) + g'(0) + \int_0^1 (1-t)\, g''(t)\, dt.$$

Replacing $g''(t) = d^\top \nabla^2 f(x+td)\, d$ gives the identity

$$f(x+d) = f(x) + \nabla f(x)^\top d + \int_0^1 (1-t)\, d^\top \nabla^2 f(x+td)\, d\, dt. \tag{1.4}$$

Subtracting $\frac{1}{2} d^\top \nabla^2 f(x) d$ from both sides yields an explicit expression for the error of the quadratic approximation:

$$R(d) := f(x+d) - \Big(f(x) + \nabla f(x)^\top d + \tfrac{1}{2} d^\top \nabla^2 f(x) d\Big) = \int_0^1 (1-t)\, d^\top \big(\nabla^2 f(x+td) - \nabla^2 f(x)\big) d\, dt.$$

Corollary 1.1.1 (Remainder bound under Lipschitz Hessian)**.** *If, in addition, the Hessian is Lipschitz continuous on the segment* $\{x + td : t \in [0, 1]\}$ *with Lipschitz constant* $L > 0$*, i.e.*

$$\|\nabla^2 f(y) - \nabla^2 f(z)\| \le L\|y - z\| \quad \text{for } y, z \text{ on the segment},$$

then the remainder satisfies the cubic bound

$$|R(d)| \le \frac{L}{6}\,\|d\|^3, \tag{1.5}$$

so the quadratic model approximates f *to order* $\mathcal{O}(\|d\|^3)$*.*

Sketch of proof of the bound. From the integral expression for $R(d)$ and the inequality $|d^\top Ad| \le \|A\|\,\|d\|^2$ (operator norm), we obtain

$$|R(d)| \le \int_0^1 (1-t)\,\big\|\nabla^2 f(x+td) - \nabla^2 f(x)\big\|\,\|d\|^2\,dt.$$

By the Lipschitz condition $\|\nabla^2 f(x+td) - \nabla^2 f(x)\| \le Lt\|d\|$, hence

$$|R(d)| \le \int_0^1 (1-t)\,(Lt\|d\|)\,\|d\|^2\,dt = L\|d\|^3 \int_0^1 t(1-t)\,dt = \frac{L}{6}\,\|d\|^3,$$

which proves (1.5). □

Remark 1.1.3 (Exactness for quadratic functions)**.** *If* f *is a quadratic function whose Hessian is constant (i.e.* $\nabla^2 f$ *does not depend on the point), then the Lagrange remainder vanishes and* (**??**) *is exact for all* d*:*

$$f(x+d) = f(x) + \nabla f(x)^\top d + \tfrac{1}{2} d^\top \nabla^2 f\, d.$$

Suggested use in a handout: include Theorem **??**, the compact proof above, and the integral remainder (1.8). Add Corollary 1.1.1 when you discuss error estimates used in optimization (trust-region / Newton methods).

Example 1.1.9. *Using* $f(x,y) = x^2 + xy + y^2$ *at point* $(0,0)$ *with* $d = (1,1)$*:*

$$f(0,0) = 0, \quad \nabla f(0,0) = (0,0)^T, \quad \nabla^2 f(0,0) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Thus,

$$f(1,1) \approx 0 + 0 + \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2}(2+1+1+2) = \frac{1}{2} \times 6 = 3.$$

1.1.4.2 Different forms of Taylor's formulas

The Taylor formula is an important tool in convex analysis. We recall it here in the general case.

Let $\Omega \subset \mathbb{R}^n$ be open, $f : \Omega \to \mathbb{R}$, $a \in \Omega$, and $h \in \mathbb{R}^n$ such that $[a, a+h] \subset \Omega$. Then: —

1. If $f \in C^1(\Omega)$, then:

(a) **First-order Taylor formula with integral remainder**:

$$f(a+h) = f(a) + \int_0^1 \langle \nabla f(a+th), h \rangle \, dt.$$

(b) **First-order Taylor-Maclaurin formula**:

$$f(a+h) = f(a) + \langle \nabla f(a+th), h \rangle.$$

(c) **First-order Taylor-Young formula**:

$$f(a+h) = f(a) + \langle \nabla f(a+th), h \rangle + o(\|h\|).$$

—

2. If $f \in C^2(\Omega)$, then:

(a) **Second-order Taylor formula with integral remainder**:

$$f(a+h) = f(a) + \langle \nabla f(a), h \rangle + \int_0^1 (1-t) \langle \nabla^2 f(a+th)h, h \rangle \, dt.$$

(b) **Second-order Taylor-Maclaurin formula**:

$$f(a+h) = f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2} \langle \nabla^2 f(a+\theta h)h, h \rangle, \quad \text{with } 0 < \theta < 1.$$

(c) **Second-order Taylor-Young formula**:

$$f(a+h) = f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2} \langle \nabla^2 f(a)h, h \rangle + o(\|h\|^2).$$

Theorem 1.1.2 (Second-order Taylor's Theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function on an open set $V \subset \mathbb{R}^n$. Then:*

1. *For any $d \in \mathbb{R}^n$ such that $x + d \in V$, we have:*

$$f(x+d) = f(x) + d^T \nabla f(x) + \frac{1}{2} d^T \nabla^2 f(x) d + o(\|d\|^2).$$

2. *For any $d \in \mathbb{R}^n$ such that $x + d \in V$, there exists $t \in [0,1]$ such that:*

$$f(x+d) = f(x) + d^T \nabla f(x) + \frac{1}{2} d^T \nabla^2 f(x+td) d.$$

Proof. We prove both items by reducing to a one-variable Taylor expansion along the line segment from x to $x + d$.

Define the one-variable function

$$g : [0,1] \to \mathbb{R}, \qquad g(t) := f(x+td).$$

Because $f \in C^2$ on an open set containing the segment $\{x + td : t \in [0,1]\}$, the function g is of class C^2 on $[0,1]$. By the chain rule we obtain for every $t \in [0,1]$:

$$g'(t) = \nabla f(x+td)^\top d, \tag{1.6}$$
$$g''(t) = d^\top \nabla^2 f(x+td)\, d. \tag{1.7}$$

Integral form (useful intermediate identity). Integrate g'' twice using the fundamental theorem of calculus. First,

$$g'(t) - g'(0) = \int_0^t g''(s)\, ds,$$

and integrating this identity from $t = 0$ to $t = 1$ with weight $(1-t)$ (or directly performing the standard two-fold integration) yields the well-known integral remainder formula

$$g(1) = g(0) + g'(0) + \int_0^1 (1-t)\, g''(t)\, dt. \tag{1.8}$$

Substituting (1.6)–(1.7) into (1.8) gives

$$f(x+d) = f(x) + \nabla f(x)^\top d + \int_0^1 (1-t)\, d^\top \nabla^2 f(x+td)\, d\, dt. \tag{1.9}$$

Equation (1.9) is exact and will be the basis for proving (1) and (2).

Proof of (2) (Lagrange form). Because $g \in C^2([0,1])$, the one-variable Taylor theorem with Lagrange remainder guarantees existence of some $\xi \in (0,1)$ such that

$$g(1) = g(0) + g'(0) + \tfrac{1}{2} g''(\xi).$$

Using (1.6)–(1.7) we obtain

$$f(x+d) = f(x) + \nabla f(x)^\top d + \tfrac{1}{2}\, d^\top \nabla^2 f(x+\xi d)\, d.$$

Setting $t = \xi$ yields exactly the claimed Lagrange form of the remainder. This proves (2).

Proof of (1) (the $o(\|d\|^2)$ statement). We start from the integral identity (1.9) and split the integrand into the value at x plus the deviation:

$$\begin{aligned} f(x+d) &= f(x) + \nabla f(x)^\top d + \int_0^1 (1-t)\, d^\top \big(\nabla^2 f(x+td)\big) d\, dt \\ &= f(x) + \nabla f(x)^\top d + \tfrac{1}{2}\, d^\top \nabla^2 f(x) d + \underbrace{\int_0^1 (1-t)\, d^\top \big(\nabla^2 f(x+td) - \nabla^2 f(x)\big) d\, dt}_{=:R(d)}. \end{aligned}$$

Thus the remainder beyond the quadratic term is

$$R(d) = \int_0^1 (1-t)\, d^\top \big(\nabla^2 f(x+td) - \nabla^2 f(x)\big) d\, dt.$$

Now use the continuity of $\nabla^2 f$ at x (since $f \in C^2$). For any $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\|d\| < \delta$ and $t \in [0,1]$ we have

$$\|\nabla^2 f(x+td) - \nabla^2 f(x)\| \le \varepsilon,$$

where $\|\cdot\|$ denotes the operator norm (or any matrix norm equivalent to it). Using the quadratic form bound $|d^\top Ad| \le \|A\|\,\|d\|^2$, we obtain for $\|d\| < \delta$:

$$|R(d)| \le \int_0^1 (1-t)\,\|\nabla^2 f(x+td) - \nabla^2 f(x)\|\,\|d\|^2\,dt \le \varepsilon\|d\|^2 \int_0^1 (1-t)\,dt = \frac{\varepsilon}{2}\,\|d\|^2.$$

Since $\varepsilon > 0$ was arbitrary, this shows

$$\lim_{\|d\|\to 0} \frac{|R(d)|}{\|d\|^2} = 0,$$

i.e. $R(d) = o(\|d\|^2)$. Substituting back gives the expansion

$$f(x+d) = f(x) + d^\top \nabla f(x) + \frac{1}{2} d^\top \nabla^2 f(x) d + o(\|d\|^2),$$

which proves (1). □

Remark 1.1.4.
- *If the Hessian $\nabla^2 f$ is* Lipschitz *on the segment $\{x+td : t\in[0,1]\}$ with constant L, one can strengthen the preceding estimate to the cubic bound*

$$|R(d)| \le \frac{L}{6}\,\|d\|^3,$$

showing the quadratic model approximates f to order $\mathcal{O}(\|d\|^3)$.

- *If f is a quadratic polynomial (Hessian constant), the remainder vanishes identically and the formula is exact for all d.*

—

1.1.5 Convexity

1.1.5.1 Convex Sets

Definition 1.1.17 (Convex Set [18, Section 2.1])**.** *A set $C \subseteq \mathbb{R}^n$ is convex if for any $x, y \in C$ and $\theta \in [0,1]$:*

$$\theta x + (1-\theta)y \in C.$$

Example 1.1.10. *The set $C = \{x \in \mathbb{R}^n : Ax \le b\}$ is convex since it is an intersection of halfspaces.*

—

1.1.5.2 Convex Hull

Definition 1.1.18 (Convex Hull [18, Section 2.3])**.** *The convex hull of a set $S \subseteq \mathbb{R}^n$ is the smallest convex set containing S, defined as:*

$$conv(S) = \left\{ \sum_{i=1}^k \theta_i x_i : x_i \in S, \theta_i \ge 0, \sum_{i=1}^k \theta_i = 1, k \in \mathbb{N} \right\}.$$

Definition 1.1.19. *Given two points $a, b \in \mathbb{R}^n$, the segment $[a, b]$ is defined as:*

$$[a, b] = \{x \in \mathbb{R}^n : x = (1-t)a + tb \text{ for some } t \in [0, 1]\}.$$

A set $K \subset \mathbb{R}^n$ is called ***convex*** *if, for all points $a, b \in K$, the segment $[a, b]$ is contained in K.*

—

Properties of Convex Sets.

1. The definition of a convex set can be interpreted by saying that the segment connecting x and y must be contained in C.
2. Let $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ and $t_j \geq 0$ such that $\sum_{j=1}^{k} t_j = 1$. Any expression of the form:
$$\sum_{j=1}^{k} t_j x_j$$
is called a **convex combination** of the points x_j or a **barycenter**.
3. If C_1 and C_2 are two convex sets in $\mathbb{R}^n$, then $K = C_1 \cap C_2$ is convex, and
$$K = \{x : x = x_1 + x_2, x_1 \in C_1, x_2 \in C_2\}$$
is convex.

—

Example 1.1.11 (Examples 1.4.1)**.**

1. *The intersection of convex sets is convex. This follows directly from the definition.*
2. *An affine subspace K of $\mathbb{R}^n$ is convex. Recall that $K \subset \mathbb{R}^n$ is an affine subspace if there exists a vector subspace $V \subset \mathbb{R}^n$ and a vector $v \in \mathbb{R}^n$ such that $x \in K$ if and only if $x - v \in V$. To show that K is convex, let $x, y \in K$, and set $z = (1-t)x + ty$ with $t \in [0, 1]$. Then $x - v, y - v \in V$, and thus $z - v = (1-t)(x-v) + t(y-v) \in V$, hence $z \in K$.*

—

1.1.6 Convex Functions

1.1.6.1 Convexity and Differentiability

Definition 1.1.20 (Convex Function [18, Section 3.1])**.** *A function $f : C \to \mathbb{R}$ defined on a convex set C is convex if:*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y), \quad \forall x, y \in C, \theta \in [0, 1].$$

Proposition 1.1.3 (First Order Condition). *If f is differentiable, then f is convex if and only if:*
$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \forall x, y \in C.$$

Proof of equivalence. ($\Rightarrow$) Assume f is convex. Fix $x, y \in C$ and $t \in (0, 1]$. By convexity,

$$f\big((1-t)x + ty\big) \leq (1-t)f(x) + tf(y).$$

Rearranging gives
$$\frac{f(x + t(y-x)) - f(x)}{t} \leq f(y) - f(x).$$
Letting $t \to 0^+$, the left-hand side converges to the directional derivative of f at x in the direction $y - x$, that is $\nabla f(x)^\top (y - x)$. Hence,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

($\Leftarrow$) Conversely, assume that

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \qquad \forall x, y \in C.$$

Fix $x, y \in C$ and $\theta \in [0, 1]$, and set $z = (1-\theta)x + \theta y$. Applying the inequality with base point z yields

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z), \qquad f(y) \geq f(z) + \nabla f(z)^\top (y - z).$$

Multiplying the first inequality by $(1 - \theta)$ and the second by θ, then adding, we obtain

$$(1-\theta)f(x) + \theta f(y) \geq f(z) + \nabla f(z)^\top \big((1-\theta)(x - z) + \theta(y - z)\big).$$

But $(1-\theta)(x - z) + \theta(y - z) = 0$ by construction of z. Hence

$$(1-\theta)f(x) + \theta f(y) \geq f(z),$$

which is exactly the convexity inequality. Therefore f is convex. □

Proposition 1.1.4 (Second Order Condition). *If f is twice differentiable, then f is convex if and only if its Hessian is positive semidefinite everywhere on C.*

Proof. We prove both directions.

$\Rightarrow$ Assume f is convex on C and $f \in C^2(C)$. Fix $x \in C$ and an arbitrary vector $v \in \mathbb{R}^n$ such that the segment $\{x + tv : t \in (-\varepsilon, \varepsilon)\}$ is contained in C for some $\varepsilon > 0$ (this holds at least for small ε). Define the one-variable function

$$\varphi(t) := f(x + tv), \qquad t \in (-\varepsilon, \varepsilon).$$

Since f is convex on C and $t \mapsto x + tv$ is affine, φ is a convex function of t. Because $f \in C^2(C)$, φ is twice continuously differentiable and by the chain rule

$$\varphi'(t) = \nabla f(x + tv)^\top v, \qquad \varphi''(t) = v^\top \nabla^2 f(x + tv)\, v.$$

Convexity of φ implies $\varphi''(t) \geq 0$ for all t in the interval; evaluating at $t = 0$ gives

$$v^\top \nabla^2 f(x)\, v = \varphi''(0) \geq 0.$$

Since v was arbitrary, $\nabla^2 f(x)$ is positive semidefinite. As x was arbitrary in C, the Hessian is positive semidefinite everywhere on C.

$\Leftarrow$ Conversely, assume $\nabla^2 f(x) \succeq 0$ for every $x \in C$. To prove convexity of f, fix arbitrary $x, y \in C$ and consider the function

$$g(t) := f\big(x + t(y - x)\big), \qquad t \in [0, 1].$$

Then $g \in C^2([0, 1])$ and by the chain rule

$$g''(t) = (y - x)^\top \nabla^2 f\big(x + t(y - x)\big)(y - x) \geq 0$$

for all $t \in [0, 1]$, because each Hessian along the segment is positive semidefinite. Hence g is a convex function on the interval $[0, 1]$.

Convexity of g implies, for every $\theta \in [0, 1]$,

$$g(\theta) \leq (1 - \theta)g(0) + \theta g(1).$$

Translating back to f (noting $g(0) = f(x)$, $g(1) = f(y)$ and $g(\theta) = f((1 - \theta)x + \theta y)$) yields

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

for all $x, y \in C$ and all $\theta \in [0, 1]$. This is exactly the definition of convexity of f on C; therefore f is convex.

Integral (Taylor) representation. For completeness, we derive the integral identity that explicitly shows how the Hessian appears in the difference $f(y) - f(x) - \nabla f(x)^\top (y - x)$. With $v := y - x$ and $g(t) = f(x + tv)$ as above, integrate g'' twice on $[0, 1]$:

$$g'(t) - g'(0) = \int_0^t g''(s)\, ds,$$

and then

$$g(1) - g(0) - g'(0) = \int_0^1 \big(g'(t) - g'(0)\big)\, dt = \int_0^1 \left(\int_0^t g''(s)\, ds\right) dt.$$

By Fubini (or direct computation of the triangular integral),

$$\int_0^1 \left(\int_0^t g''(s)\, ds\right) dt = \int_0^1 (1 - s)g''(s)\, ds.$$

Therefore

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \int_0^1 (1 - s)\, (y - x)^\top \nabla^2 f(x + s(y - x))(y - x)\, ds.$$

If $\nabla^2 f(\cdot) \succeq 0$ on C, the integrand is nonnegative for all $s \in [0, 1]$, hence the integral is ≥ 0, which immediately yields the first-order inequality

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x),$$

and therefore convexity (by the first-order characterization). This integral identity also quantifies how the curvature (encoded by the Hessian) contributes to the deviation from linearity.

$\square$

—

1.1.6.2 Monotone Functions

Definition 1.1.21 (Monotone Function [18, Section 3.1])**.** *A function $f : \mathbb{R} \to \mathbb{R}$ is monotone increasing if $x \leq y$ implies $f(x) \leq f(y)$.*

Example 1.1.12. *The exponential function $f(x) = e^x$ is monotone increasing.*

Definition 1.1.22. *A real-valued function $J : K \subset V \to \mathbb{R}$ defined on a convex set K of a vector space V is said to be **convex** on K if for all points $u, v \in K$ and for all real numbers t with $0 \leq t \leq 1$, the following inequality holds:*

$$J(tu + (1-t)v) \leq tJ(u) + (1-t)J(v).$$

*It is said to be **strictly convex** on K if:*

$$u, v \in K, u \neq v, t \in (0,1) \quad \Longrightarrow \quad J(tu + (1-t)v) < tJ(u) + (1-t)J(v).$$

—

Definition 1.1.23. *A function f is said to be **strongly convex** with constant $\alpha > 0$ if:*

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) - \frac{\alpha}{2}t(1-t)\|x_1 - x_2\|^2.$$

It can be easily shown that a strongly convex function is strictly convex.

—

Remark 1.1.5. *Strong convexity $\Longrightarrow$ strict convexity $\Longrightarrow$ convexity.*

Definition 1.1.24. *A function $f : K \subset V \to \mathbb{R}$ defined on a convex subset K of a vector space V is said to be **(strictly) concave** if the function $(-f)$ is **(strictly) convex**.*

—

Example 1.1.13 (Examples 1.4.2)**.**

1. *An affine function is convex on $\mathbb{R}^n$, but not strictly convex. An affine function is given by $f(x) = \langle a, x \rangle + b$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.*
2. *A norm is convex on $\mathbb{R}^n$ but not strictly convex. This follows from the triangle inequality: for all $x, y \in \mathbb{R}^n$ and $t \in (0,1)$,*

   $$\|(1-t)x + ty\| \leq (1-t)\|x\| + t\|y\|.$$

   The inequality is not strict because if $y = \lambda x$ for some $\lambda \geq 0$, equality holds.

—

Operations on Convex Functions. Let $K \subset \mathbb{R}^n$ be convex.

1. The sum of two convex functions on K is convex.
2. If $f : K \to \mathbb{R}$ is convex and $\lambda \geq 0$, then λf is convex on K.
3. The maximum of two convex functions on K is convex.
4. If $f : K \to \mathbb{R}$ is convex and $L : \mathbb{R}^m \to \mathbb{R}^n$ is linear, then $f \circ L$ is convex on

$$V = \{x \in \mathbb{R}^m : L(x) \in K\}.$$

—

Definition 1.1.25 (Epigraph of a Convex Function)**.** *Let K be a non-empty subset of $\mathbb{R}^n$ and $f : K \to \mathbb{R}$. The **epigraph** of f, denoted epi(f), is defined as:*

$$epi(f) = \{(x, y) \in \mathbb{R}^{n+1} : x \in K, y \geq f(x)\}.$$

Equivalently,

$$epi(f) = \{(x, y) \in K \times \mathbb{R} : y \geq f(x)\}.$$

It is the set of points above the graph of f.

—

Theorem 1.1.3. *Let K be a non-empty convex set in $\mathbb{R}^n$ and $f : K \to \mathbb{R}$. Then*

$$f \text{ is convex} \iff epi(f) \text{ is convex.}$$

Proof. We prove the two implications separately.

$\Rightarrow$ Assume f is convex on K. We show that epi(f) is convex. Fix two arbitrary points $(x_1, t_1), (x_2, t_2) \in \text{epi}(f)$ and let $\theta \in [0, 1]$. By membership in the epigraph we have $t_1 \geq f(x_1)$ and $t_2 \geq f(x_2)$. Because K is convex and $x_1, x_2 \in K$, the point

$$x_\theta := (1 - \theta)x_1 + \theta x_2$$

belongs to K. Consider the convex combination in $\mathbb{R}^n \times \mathbb{R}$

$$(x_\theta, t_\theta) := (1 - \theta)(x_1, t_1) + \theta(x_2, t_2) = \big((1 - \theta)x_1 + \theta x_2,\ (1 - \theta)t_1 + \theta t_2\big).$$

Using convexity of f and the inequalities $t_i \geq f(x_i)$, we obtain

$$(1 - \theta)t_1 + \theta t_2 \geq (1 - \theta)f(x_1) + \theta f(x_2) \geq f\big((1 - \theta)x_1 + \theta x_2\big) = f(x_\theta).$$

Thus $t_\theta \geq f(x_\theta)$, so $(x_\theta, t_\theta) \in \text{epi}(f)$. Since the two arbitrary points and arbitrary $\theta \in [0, 1]$ were arbitrary, epi(f) is convex.

$\Leftarrow$ Conversely, assume epi(f) is convex. We prove f is convex on K. Fix arbitrary $x_1, x_2 \in K$ and $\theta \in [0, 1]$. By definition of f we have

$$(x_1, f(x_1)),\ (x_2, f(x_2)) \in \text{epi}(f).$$

Convexity of epi(f) implies that the point

$$(x_\theta, t_\theta) := (1-\theta)(x_1, f(x_1)) + \theta(x_2, f(x_2))$$

belongs to epi(f). As before $x_\theta = (1-\theta)x_1 + \theta x_2$ and $t_\theta = (1-\theta)f(x_1) + \theta f(x_2)$. Membership $(x_\theta, t_\theta) \in \mathrm{epi}(f)$ means $t_\theta \geq f(x_\theta)$, i.e.

$$(1-\theta)f(x_1) + \theta f(x_2) \geq f\big((1-\theta)x_1 + \theta x_2\big).$$

Rewriting yields the usual convexity inequality

$$f\big((1-\theta)x_1 + \theta x_2\big) \leq (1-\theta)f(x_1) + \theta f(x_2).$$

Because $x_1, x_2 \in K$ and $\theta \in [0,1]$ were arbitrary, f is convex on K.

This completes the proof of the equivalence. □

—

1.1.7 Convexity and Continuity

Proposition 1.1.5 (Jensen's Inequality [18, Section 3.1.5, p. 72])**.** *Suppose f is convex on K, and let $p_1, \dots, p_m \geq 0$ with $p_1 + \cdots + p_m = 1$. Then, for all $x_1, \dots, x_m \in K$,*

$$f(p_1x_1 + \cdots + p_mx_m) \leq p_1f(x_1) + \cdots + p_mf(x_m).$$

Proof. Immediate for $m = 1$; equivalent to the definition of convexity for $m = 2$; the result follows by induction for $m > 2$. □

—

Theorem 1.1.4 (Continuity of convex functions [11, Theorem 10.1])**.** *If K is open and f is convex on K, then f is continuous on K.*

Proof. We prove it in the case $K = \mathbb{R}^2$ and at the origin for simplicity, with $\|\cdot\|_\infty$ norm. We proceed in two steps:

Step 1. Show there exists $C \in \mathbb{R}$ such that for all $r > 0$ in a neighborhood of 0, and for all x with $\|x\| \leq r$,

$$f(x) - f(0) \leq Cr.$$

Assuming x is in the first quadrant (other cases are similar), we have:

$$x = p_1 0 + p_2 e_1 + p_3 e_2,$$

where $p_1 = 1 - x_1 - x_2$, $p_2 = x_1$, $p_3 = x_2$, and $0 = (0,0)$, $e_1 = (1,0)$, $e_2 = (0,1)$. For sufficiently small r, $p_1, p_2, p_3 > 0$ and $p_1 + p_2 + p_3 = 1$.

By Jensen's inequality,

$$f(x) \leq p_1 f(0) + p_2 f(e_1) + p_3 f(e_2),$$

thus

$$f(x) - f(0) \leq |p_1 - 1||f(0)| + p_2|f(e_1)| + p_3|f(e_2)|.$$

□

The assertion then follows from the fact that

$$|p_1 - 1| \leq 2r, \quad p_2 \leq r, \quad p_3 \leq r.$$

—

Step 2. We now show that there exists $C' \in \mathbb{R}$ such that for all $r > 0$ in a neighborhood of 0, and for all x with $\|x\| \leq r$, we have:

$$f(0) - f(x) \leq C'r.$$

To see this, we write:

$$0 = p_1 x + p_2(-e_1) + p_3(-e_2),$$

with

$$p_1, p_2, p_3 > 0, \quad p_1 + p_2 + p_3 = 1.$$

This decomposition is made possible by taking:

$$p_1 = \frac{1}{1 + x_1 + x_2}, \quad p_2 = \frac{x_1}{1 + x_1 + x_2}, \quad p_3 = \frac{x_2}{1 + x_1 + x_2},$$

assuming again that x is in the first quadrant. By convexity of f, we find:

$$f(0) - f(x) \leq |p_1 - 1||f(x)| + p_2|f(-e_1)| + p_3|f(-e_2)|.$$

Taking r sufficiently small, and observing again that:

$$|p_1 - 1| \leq 2r, \quad p_2 \leq r, \quad p_3 \leq r,$$

and using Step 1 to bound $|f(x)|$, we obtain the desired result.

Thus, continuity at the origin follows directly from the two assertions above.

—

Remark 1.1.6. *The proposition is false if K is not open. For example, let $K = [0, 1] \subset \mathbb{R}$ and consider $f : K \to \mathbb{R}$ defined by*

$$f(x) = \begin{cases} 0, & \text{if } x \in (0, 1), \\ 1, & \text{if } x \in \{0, 1\}. \end{cases}$$

—

Definition 1.1.26. *Let f be a function defined from K to $\mathbb{R} \cup \{+\infty\}$. The* ***domain*** *of f, denoted $dom(f)$, is defined as:*

$$dom(f) = \{x \in K : f(x) \neq +\infty\}.$$

Definition 1.1.27. *A function f defined on K is said to be* ***coercive*** *if:*

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

—

Definition 1.1.28. *A function f defined from K to $\mathbb{R} \cup \{+\infty\}$ is said to be* ***lower semicontinuous*** *(LSC) if for all $x \in K$:*

$$\liminf_{y \to x} f(y) \geq f(x).$$

—

Remark 1.1.7. *1. Lower semicontinuity is stable under addition.*

2. Every continuous function is LSC.

3. The indicator function of a closed convex set is LSC (i.e., the function that is zero on the convex set and $+\infty$ outside).

4. If f is LSC, then for every $\alpha \in \mathbb{R}$, the sets

$$\{x \in K : f(x) \leq \alpha\} \quad \textit{and} \quad \{(x, \alpha) \in K \times \mathbb{R} : f(x) \leq \alpha\}$$

are closed.

—

Definition 1.1.29. *A function from K to $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is said to be* ***proper*** *if it is not identically equal to $+\infty$ and does not take the value $-\infty$.*

—

1.2 Convexity and First-Order Differentiability

Below is a first result that allows one to recognize the convexity of a function using its first derivatives.

Proposition 1.2.1 (Characterization of convexity via derivative [11, Theorem 24.1])**.** *Let f be a differentiable function on an interval $I \subset \mathbb{R}$. Then:*

$$f \textit{ is convex } \iff f' \textit{ is non-decreasing on } I.$$

Proof. Suppose f is convex and show that f' is non-decreasing. Let $0 \leq \lambda_1 \leq \lambda_2$. Then, by convexity of f:

$$f(x + \lambda_1 d) = f\Big((1 - \frac{\lambda_1}{\lambda_2})x + \frac{\lambda_1}{\lambda_2}(x + \lambda_2 d)\Big) \leq (1 - \frac{\lambda_1}{\lambda_2})f(x) + \frac{\lambda_1}{\lambda_2}f(x + \lambda_2 d).$$

□

Therefore:

$$\frac{f(x + \lambda_1 d) - f(x)}{\lambda_1} \leq \frac{f(x + \lambda_2 d) - f(x)}{\lambda_2}.$$

Hence the result.

—

2. Suppose that f' is increasing, and let us show that f is convex. Let x, y such that $y > x$. Define the function:

$$\varphi(t) = tf(x) + (1-t)f(y) - f(tx + (1-t)y).$$

We want to show that $\varphi(t) \geq 0$ for all $t \in [0,1]$. Note that $\varphi(0) = \varphi(1) = 0$. Moreover:

$$\varphi'(t) = f(x) - f(y) + (y-x)f'((x-y)t + y).$$

Since $y - x > 0$, we have:

$$\varphi'(t) = k + af'(y - at),$$

with $a > 0$. This function is decreasing since f' is increasing.

If $\varphi'(1) > 0$, then for all $t \in [0,1]$,

$$\varphi'(t) > 0,$$

and

$$\varphi(1) = \varphi(0) + \int_0^1 \varphi'(t)dt > \varphi(0) = \varphi(1),$$

which is absurd. Hence $\varphi'(1) \leq 0$. Similarly, one shows that $\varphi'(0) \geq 0$. Therefore, for all $t \in [0,1]$, $\varphi(t) \geq 0$, and thus f is convex.

—

Corollary 1.2.1. *Let f be twice differentiable on an interval $I \subset \mathbb{R}$. Then:*

$$f \text{ is convex} \iff f'' \geq 0 \text{ on } I,$$

$$f \text{ is concave} \iff f'' \leq 0 \text{ on } I.$$

Example 1.2.1.

1. *$f(x) = x^2$ is convex since $f''(x) = 2 > 0$, $\forall x \in \mathbb{R}$.*
2. *$f(x) = e^x$ is convex since $f''(x) = e^x > 0$, $\forall x \in \mathbb{R}$.*
3. *$f(x) = \ln x$ is concave since $f''(x) = -\frac{1}{x^2} < 0$, $\forall x \in (0, +\infty)$.*

—

Theorem 1.2.1 (First-order characterization of convexity [**?**, Theorem 25.1])**.** *Let E be a normed space, K an open convex subset of E, and $f : K \to \mathbb{R}$ a differentiable function. Then the following are equivalent:*

1. *f is convex on K;*
2. *$\forall x, y \in K, f(y) \geq f(x) + f'(x) \cdot (y - x)$;*
3. *$\forall x, y \in K, [f'(y) - f'(x)] \cdot (y - x) \geq 0$.*

Proof. We prove the cycle of implications (1) ⇒ (2) ⇒ (3) ⇒ (2) and then (2) ⇒ (1).

(1) ⇒ **(2)**. Assume f is convex on K. Fix $x, y \in K$ and define the one-variable function

$$\varphi : [0,1] \to \mathbb{R}, \qquad \varphi(t) := f(x + t(y-x)).$$

Since f is convex on K and the map $t \mapsto x + t(y - x)$ is affine, φ is a convex function on $[0, 1]$. Because f is differentiable on K, φ is differentiable on $[0, 1]$ with derivative

$$\varphi'(t) = f'\big(x + t(y - x)\big)\big(y - x\big).$$

A standard property of convex differentiable functions on an interval is that their graph lies above every tangent line; in particular, $\varphi(1) \geq \varphi(0) + \varphi'(0)$. Substituting the definitions gives

$$f(x + y - x) = f(y) \geq f(x) + f'(x)\big(y - x\big),$$

which is exactly (2).

(2) $\Rightarrow$ **(3)**. Assume (2) holds for all $x, y \in K$. Apply (2) with the pair (x, y) and then with the pair (y, x):

$$f(y) \geq f(x) + f'(x)\big(y - x\big),$$
$$f(x) \geq f(y) + f'(y)\big(x - y\big).$$

Add these two inequalities to obtain

$$0 \geq f'(x)\big(y - x\big) + f'(y)\big(x - y\big) = -\big(f'(y) - f'(x)\big)\big(y - x\big).$$

Rearranging yields $(f'(y) - f'(x))\big(y - x\big) \geq 0$, which is (3).

(3) $\Rightarrow$ **(2)**. Assume (3) holds. Fix $x, y \in K$ and consider again the one-variable function $\varphi(t) := f(x + t(y - x))$ for $t \in [0, 1]$. As before, $\varphi'(t) = f'(x + t(y - x))\big(y - x\big)$. The hypothesis (3) implies that for every $t \in [0, 1]$,

$$\varphi'(t) - \varphi'(0) = \big(f'(x + t(y - x)) - f'(x)\big)\big(y - x\big) \geq 0,$$

so $\varphi'(t) \geq \varphi'(0)$ for all $t \in [0, 1]$. Integrating this inequality over $t \in [0, 1]$ yields

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\, dt \geq \int_0^1 \varphi'(0)\, dt = \varphi'(0).$$

Unwinding the definitions gives

$$f(y) - f(x) \geq f'(x)\big(y - x\big),$$

which is exactly (2). Thus (3)$\Rightarrow$(2).

(2) $\Rightarrow$ **(1)**. Assume (2) holds. For each $x \in K$ define the affine function $L_x : K \to \mathbb{R}$ by

$$L_x(y) := f(x) + f'(x)\big(y - x\big).$$

Hypothesis (2) states precisely that $L_x(y) \leq f(y)$ for all $y \in K$, i.e. every tangent affine map L_x is a global underestimator of f. Moreover $L_x(x) = f(x)$ for each x. Hence, for any fixed $y \in K$,

$$\sup_{x \in K} L_x(y) = f(y),$$

because the supremum is bounded above by $f(y)$ (by (2)) and is attained at $x = y$. But the pointwise supremum of an arbitrary family of affine (hence convex) functions is a convex function. Therefore f is convex on K. This proves (2)$\Rightarrow$(1).

Combining the implications above gives the equivalence of (1), (2), and (3). $\square$

Remark 1.2.1. • *The pairing $f'(x)(y-x)$ denotes the action of the bounded linear functional $f'(x)$ on the vector $y-x$. In finite dimensions $f'(x)$ is the usual gradient and the pairing reduces to the dot product $\nabla f(x)^\top (y-x)$.*

- *The representation $f = \sup_{x\in K} L_x$ used in (2)⇒(1) is the standard fact that a differentiable function that lies above all its tangents is the supremum of those tangents; this is the cornerstone of the first-order characterization of convexity.*

—

Theorem 1.2.2 (First-order characterization of convexity [**?**, Theorem 25.1])**.** *Under the same assumptions, the following are equivalent:*

1. *f is strictly convex on K;*
2. *$\forall x, y \in K, x \neq y : f(y) > f(x) + f'(x)\cdot(y-x)$;*
3. *$\forall x, y \in K, x \neq y : [f'(y) - f'(x)]\cdot(y-x) > 0$.*

Proof. We prove the cycle of implications (1) ⇒ (2) ⇒ (3) ⇒ (2) and then (2) ⇒ (1).

(**1**) ⇒ (**2**). Assume f is strictly convex on K. Fix $x, y \in K$ with $x \neq y$, and define the one-variable function

$$\varphi : [0,1] \to \mathbb{R}, \qquad \varphi(t) := f\big(x + t(y-x)\big).$$

Because f is strictly convex on K, the restriction φ is strictly convex on $[0,1]$. Differentiability of f implies φ is differentiable, with

$$\varphi'(t) = f'\big(x + t(y-x)\big)\big(y-x\big).$$

A standard property of strictly convex differentiable functions on an interval is that the graph lies *strictly* above any tangent line except at the tangency point; in particular, $\varphi(1) > \varphi(0) + \varphi'(0)$. Unwinding definitions yields

$$f(y) > f(x) + f'(x)\big(y-x\big),$$

which is (2).

(**2**) ⇒ (**3**). Assume (2) holds for all distinct $x, y \in K$. Apply (2) first to the pair (x, y) and then to (y, x):

$$f(y) > f(x) + f'(x)\big(y-x\big),$$
$$f(x) > f(y) + f'(y)\big(x-y\big).$$

Adding these strict inequalities gives

$$0 > f'(x)\big(y-x\big) + f'(y)\big(x-y\big) = -\big(f'(y) - f'(x)\big)\big(y-x\big).$$

Rearranging yields $(f'(y) - f'(x))(y-x) > 0$, which is (3).

(**3**) ⇒ (**2**). Assume (3) holds. Fix $x, y \in K$ with $x \neq y$ and define again $\varphi(t) = f(x + t(y-x))$ for $t \in [0,1]$. Then $\varphi'(t) = f'(x + t(y-x))(y-x)$. From (3) we have for every $t \in (0,1]$

$$\big(f'(x + t(y-x)) - f'(x)\big)\big(y-x\big) > 0,$$

hence $\varphi'(t) - \varphi'(0) > 0$, so $\varphi'(t) > \varphi'(0)$ for all $t \in (0,1]$. Integrating φ' over $[0,1]$ yields

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\,dt > \int_0^1 \varphi'(0)\,dt = \varphi'(0).$$

Unwinding definitions gives

$$f(y) - f(x) > f'(x)\big(y - x\big),$$

which is (2). Thus (3)⇒(2).

(2) ⇒ **(1)**. Assume (2) holds. We show f is strictly convex. Fix distinct points $x, y \in K$ and take any $\lambda \in (0,1)$. Put $z := (1-\lambda)x + \lambda y \in K$. We must prove

$$f(z) < (1-\lambda)f(x) + \lambda f(y).$$

Apply (2) with base point x and argument z. Since $z \neq x$,

$$f(z) > f(x) + f'(x)(z - x).$$

Rearrange this inequality to express $f'(x)(z-x)$:

$$f'(x)(z - x) < f(z) - f(x).$$

Now multiply the strict inequality from (2) for the pair (x, y) by $\lambda > 0$:

$$\lambda\big(f(y) - f(x) - f'(x)(y - x)\big) > 0.$$

But $z - x = \lambda(y - x)$, so $f'(x)(z - x) = \lambda f'(x)(y - x)$. Substitute this into the previous displayed inequality to obtain

$$\lambda f(y) - \lambda f(x) - f'(x)(z - x) > 0.$$

Rearrange:

$$f'(x)(z - x) < \lambda\big(f(y) - f(x)\big).$$

Combine this with $f(z) > f(x) + f'(x)(z - x)$:

$$f(z) < f(x) + \lambda\big(f(y) - f(x)\big) = (1-\lambda)f(x) + \lambda f(y).$$

This is the strict convexity inequality at z. Since x, y and λ were arbitrary (with $x \neq y$, $\lambda \in (0,1)$), f is strictly convex on K. Thus (2)⇒(1).

Collecting the implications above we obtain the equivalence of (1), (2), and (3). □

Remark 1.2.2.
- *In finite dimensions the Fréchet derivative $f'(x)$ identifies with the gradient $\nabla f(x)$ and the pairing $f'(x)(y-x)$ becomes the dot product $\nabla f(x)^\top (y-x)$.*
- *The arguments above mirror the nonstrict case but replace weak inequalities by strict inequalities; care is taken when integrating or adding inequalities to preserve strictness.*

—

Theorem 1.2.3. *Let E be a Euclidean space, K an open convex subset of E, and $f : K \to \mathbb{R}$ a differentiable function. Then the following are equivalent:*

1. *f is strongly convex on K;*
2. *$\exists \alpha > 0, \forall x, y \in K : f(y) \geq f(x) + f'(x) \cdot (y - x) + \frac{\alpha}{2}\|y - x\|^2$;*
3. *$\exists \alpha > 0, \forall x, y \in K : [f'(y) - f'(x)] \cdot (y - x) \geq \alpha\|y - x\|^2$.*

Proof. We prove the cycle of implications (1) ⇒ (2) ⇒ (3) ⇒ (2) and then (2) ⇒ (1).

(1) ⇒ **(2)**. By definition, f is strongly convex with modulus $\alpha > 0$ iff the function $g := f - \frac{\alpha}{2}\| \cdot \|^2$ is convex on K. Since g is convex and differentiable, the first-order condition for convex functions gives for every $x, y \in K$,

$$g(y) \geq g(x) + g'(x) \cdot (y - x).$$

Now $g'(x) = f'(x) - \alpha x$ (the derivative of $\frac{1}{2}\alpha\|x\|^2$ is αx), so substituting g and g' yields

$$f(y) - \tfrac{\alpha}{2}\|y\|^2 \geq f(x) - \tfrac{\alpha}{2}\|x\|^2 + \big(f'(x) - \alpha x\big) \cdot (y - x).$$

Rearrange terms (expand the quadratic terms or move α-terms to the right) to obtain the desired inequality

$$f(y) \geq f(x) + f'(x) \cdot (y - x) + \tfrac{\alpha}{2}\|y - x\|^2.$$

Thus (1)⇒(2).

(2) ⇒ **(3)**. Assume (2) holds for some $\alpha > 0$. Apply (2) with the ordered pair (x, y) and then with (y, x):

$$f(y) \geq f(x) + f'(x) \cdot (y - x) + \tfrac{\alpha}{2}\|y - x\|^2,$$
$$f(x) \geq f(y) + f'(y) \cdot (x - y) + \tfrac{\alpha}{2}\|x - y\|^2.$$

Add the two inequalities (note $\|x - y\| = \|y - x\|$) to get

$$0 \geq f'(x) \cdot (y - x) + f'(y) \cdot (x - y) + \alpha\|y - x\|^2 = -\big(f'(y) - f'(x)\big) \cdot (y - x) + \alpha\|y - x\|^2.$$

Rearranging yields

$$\big(f'(y) - f'(x)\big) \cdot (y - x) \geq \alpha\|y - x\|^2,$$

which is (3).

(3) ⇒ **(2)**. Assume (3) holds for some $\alpha > 0$. Fix $x, y \in K$ and set $d := y - x$. Because f is differentiable, the map $t \mapsto f'(x + td)$ is defined for $t \in [0, 1]$. For each $t \in [0, 1]$ apply (3) to the pair $(x, x + td)$ to obtain

$$\big(f'(x + td) - f'(x)\big) \cdot \big((x + td) - x\big) \geq \alpha\|(x + td) - x\|^2,$$

i.e.

$$\big(f'(x + td) - f'(x)\big) \cdot (td) \geq \alpha t^2\|d\|^2.$$

For $t > 0$ divide by t to get

$$\big(f'(x + td) - f'(x)\big) \cdot d \geq \alpha t\|d\|^2.$$

(When $t = 0$ the left-hand side is 0 by continuity of f', and the inequality then holds in the limiting sense.) Now integrate this inequality with respect to t over $[0, 1]$:

$$\int_0^1 \big(f'(x+td) - f'(x)\big) \cdot d\, dt \geq \int_0^1 \alpha t \|d\|^2\, dt = \tfrac{\alpha}{2}\|d\|^2.$$

But

$$\int_0^1 \big(f'(x+td) - f'(x)\big) \cdot d\, dt = \int_0^1 f'(x+td) \cdot d\, dt - \int_0^1 f'(x) \cdot d\, dt = \int_0^1 f'(x+td) \cdot d\, dt - f'(x) \cdot d.$$

Using the fundamental identity (line integral representation of the increment)

$$f(y) - f(x) = \int_0^1 f'(x+td) \cdot d\, dt,$$

we obtain

$$f(y) - f(x) - f'(x) \cdot d \geq \tfrac{\alpha}{2}\|d\|^2,$$

which is exactly

$$f(y) \geq f(x) + f'(x) \cdot (y - x) + \tfrac{\alpha}{2}\|y - x\|^2.$$

This proves (3)⇒(2).

(2) $\Rightarrow$ **(1)**. Finally, assume (2) holds for some $\alpha > 0$. Define $g := f - \frac{\alpha}{2}\|\cdot\|^2$. Then for every $x, y \in K$ the inequality (2) is equivalent to

$$g(y) \geq g(x) + g'(x) \cdot (y - x),$$

because the gradient of $\frac{\alpha}{2}\|\cdot\|^2$ is αx, so $g'(x) = f'(x) - \alpha x$. The displayed inequality is the first-order condition for convexity of the differentiable function g, hence g is convex on K. Therefore $f = g + \frac{\alpha}{2}\|\cdot\|^2$ is strongly convex with modulus α, which is (1).

Combining the implications above proves the equivalence of (1), (2), and (3). □

Remark 1.2.3.

- *In finite dimensions the derivative $f'(x)$ is identified with the gradient $\nabla f(x)$ and the pairing $f'(x) \cdot (y - x)$ becomes $\nabla f(x)^\top (y - x)$.*
- *If f is twice continuously differentiable, an equivalent and often used pointwise formulation is $\nabla^2 f(x) \succeq \alpha I$ for all $x \in K$.*
- *In the implication (3)⇒(2) we used the line integral identity $f(y) - f(x) = \int_0^1 f'(x+td) \cdot d\, dt$ together with a simple integration of the linear inequalities obtained from (3).*

—

1.3 Convexity and Second-Order Differentiability

Definition 1.3.1. *Let $H(x^*)$ be the Hessian matrix of f at point x^*. Then:*

1. *$H(x^*)$ is said to be* ***positive semi-definite*** *(PSD), denoted $H(x^*) \geq 0$, if*

$$\forall x \in \mathbb{R}^n, \quad x^T H(x^*) x \geq 0.$$

Thus, all eigenvalues of $H(x^*)$ are non-negative.

2. *$H(x^*)$ is said to be* ***positive definite*** *(PD), denoted $H(x^*) > 0$, if*

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \quad x^T H(x^*) x > 0.$$

Thus, all eigenvalues of $H(x^*)$ are strictly positive.

Theorem 1.3.1 (Second-order characterization of convexity [18, Section 3.1.3])**.** *Let $K \subset \mathbb{R}^n$ be an open, non-empty convex set and $f : K \to \mathbb{R}$ be a real-valued function that is twice differentiable in K. Then:*

1. *f is convex in K if and only if the Hessian matrix is positive semi-definite (p.s.d) at every point in K.*

2. *f is strictly convex in K if and only if the Hessian matrix is positive definite (p.d) at every point in K.*

Proof. 1. Assume that f is convex in K. Then:

$$\forall x^* \in K, \quad f(x^* + \lambda x) \geq f(x^*) + \lambda(\nabla f(x^*))^T x, \quad \forall x \in K. \quad (1.1)$$

Since f is twice differentiable in K, by the second-order Taylor-Maclaurin expansion:

$$f(x^* + \lambda x) = f(x^*) + \lambda(\nabla f(x^*))^T x + \frac{1}{2}\lambda^2 x^T H(x^*) x + \lambda^2 \|x\|^2 \alpha(x^*, \lambda x). \quad (1.2)$$

Combining (1.1) and (1.2) gives:

$$\frac{1}{2}\lambda^2 x^T H(x^*) x + \lambda^2 \|x\|^2 \alpha(x^*, \lambda x) \geq 0.$$

Dividing by λ^2 and letting $\lambda \to 0$ yields:

$$x^T H(x^*) x \geq 0.$$

2. Conversely, suppose $H(x^*)$ is p.s.d. for all x^*. Then:

$$f(x) = f(x^*) + (\nabla f(x^*))^T (x - x^*) + \frac{1}{2}(x - x^*)^T H(\zeta)(x - x^*), \quad \zeta = tx + (1-t)x^*,\ t \in (0, 1).$$

So:

$$f(x) - [f(x^*) + (\nabla f(x^*))^T (x - x^*)] = \frac{1}{2}(x - x^*)^T H(\zeta)(x - x^*) \geq 0,$$

hence f is convex.

□

Example 1.3.1. *Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x, y, z) = (x-2)^2 + (y-3)^2 + z^2$.*

- *The gradient is:*

$$\nabla f(x, y, z) = \begin{pmatrix} 2x - 4 \\ 2y - 6 \\ 2z \end{pmatrix}.$$

- *The Hessian is:*

$$H(x, y, z) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} = 2I_3 \geq 0.$$

  Therefore, f is convex.

Remark 1.3.1. *If f is convex, then any local minimum is also global. Moreover, if f is strictly convex, then any local minimum is not only global but unique.*

—

1.3.1 Exercises

1. Compute the directional derivative of $f(x, y) = xy^2$ at point $(1, 2)$ in the direction $d = (2, 1)$.
2. Prove that the set $S = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ is convex.
3. Show that $f(x) = x^4$ is convex on $\mathbb{R}$.

—

1.3.2 Solutions

1. $\nabla f(x, y) = (y^2, 2xy)$. At $(1, 2)$, $\nabla f = (4, 4)$. Thus:

$$D_f((1, 2); (2, 1)) = (4, 4) \cdot (2, 1) = 8 + 4 = 12.$$

2. For any x, y in the set and $\theta \in [0, 1]$:

$$\|\theta x + (1-\theta)y\|_2 \leq \theta\|x\|_2 + (1-\theta)\|y\|_2 \leq \theta + (1-\theta) = 1.$$

   Thus, it is convex.

3. $f''(x) = 12x^2 \geq 0$ for all x, thus f is convex.