Mathematical Foundations of Data Science

Review of optimization basics

Yue Xie

Nov 4, 2025

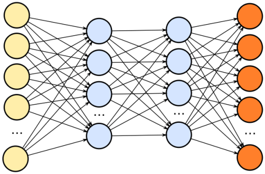
Optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \Omega. \end{aligned} \tag{1}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is called the objective function.
- $\Omega \subseteq \mathbb{R}^n$ is called the constraint set or feasible region.
- Every x in Ω is called a feasible point (or feasible solution).
- A point $x^* \in \Omega$ is called an optimal solution if $f(x^*) \leq f(x)$ for any $x \in \Omega$. $f(x^*)$ is called the optimal value.
- (1) is called an unconstrained optimization problem if $\Omega = \mathbb{R}^n$.
- (1) is called a constrained optimization problem if $\Omega \neq \mathbb{R}^n$.

Examples

1. *Deep neural networks training.*

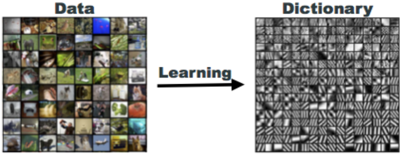


$$\min_{W_1,\dots,W_N} \quad \sum_{i=1}^{n} l(h(x_i), y_i).$$

- $h(x) = W_N(\sigma(W_{N-1}\dots(\sigma(W_1 x + b_1))\dots + b_{N-1})) + b_N$.
- An unconstrained optimization problem.

Examples

2. *Dictionary learning.*

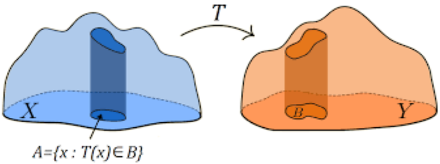


$$
\begin{aligned}
\min_{D \in \mathbb{R}^{p \times p}} \quad & \|Y^T D\|_1 \\
\text{subject to} \quad & D^T D = I_p.
\end{aligned}
$$

- A constrained optimization problem.

Examples

3. *Optimal transport problem (Wasserstein distance).*



$$
\begin{aligned}
\min_{X \in \mathbb{R}^{m\times n}} \quad & \langle C, X \rangle \\
\text{subject to} \quad & X\mathbf{1} = r, X^\top \mathbf{1} = c, X_{i,j} \geq 0, \forall i, j. \\
& \mathbf{1} = (1, ..., 1)_n^\top.
\end{aligned}
$$

- A linear programming problem.

Tasks in optimization

- Problem formulation
  - Formulate the correct optimization problem that captures the applications.
  - Sometimes formulate the dual problem.
- Optimality conditions
  - Find out the equations that describe the optimal solutions of the optimization problems.
- Algorithm design
  - Iterative algorithms.
  - Convergence - correctness.
  - Speed - efficiency.
  - Scalable algorithms/methods - first-order methods (computation does not increase drastically as the problem dimension grows)

Syllabus (optimization part)

- Week 9 (Xie): Optimization: Review (general formulation, examples in data science); Convexity
- Week 10 (Zou): Optimization: Descent methods and stochastic gradient descent (SGD); Examples: Gradient descent for linear models, backpropagation for network training, SGD for loss minimization
- Week 11 (Xie): Optimization: Lagrange multipliers/duality, focusing on conceptual/computational aspects
- Week 12 (Zou): Optimization: Advanced topics (Variance Reduction; Momentum methods), focusing on conceptual/computational aspects

Example

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Solve the problem

$$\begin{array}{ll} \min & x^T A x \\ \text{s.t.} & x^T x = 1. \end{array}$$

Question 1. What is the optimal value of the corresponding maximization problem?

Question 2. Suppose A is an arbitrary square matrix. What is the solution?

Solution

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be all the eigenvalues of A. Since A is symmetric, by a theorem in linear algebra, A can be diagonalized by using an orthogonal matrix Q, that is

$$A = Q^T \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} Q.$$

Let $y = Qx$. Then $y^T y = x^T Q^T Q x = x^T x = 1$. Thus

$$\begin{aligned} x^T A x = x^T Q^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} Q x = y^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} y \\ = \sum_{i=1}^{n} \lambda_i y_i^2 \overset{(1)}{\geq} \lambda_n \sum_{i=1}^{n} y_i^2 = \lambda_n y^T y = \lambda_n. \end{aligned}$$

Note that (1) holds equality if $y = (0, 0, \ldots, 0, 1)^T$. Hence the optimal value is λ_n which is achieved when $y = (0, 0, \ldots, 0, 1)^T$ or $x = Q^T y = Q^T \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$. $\square$

Example

Find an optimal solution to the problem

$$
\begin{array}{lll}
\min & f(x) = \sum_{i=1}^{n} \frac{c_i}{x_i} & \\
\text{s.t.} & \sum_{i=1}^{n} a_i x_i = b & \\
 & x_i \geq 0 \quad \text{for} & i = 1, 2, \ldots, n,
\end{array}
$$

where a_i, b, c_i are all positive constants for $i = 1, 2, \ldots, n$.

Solution

Our approach relies on the famous Schwartz inequality: Let $g = (g_1, g_2, \ldots, g_n)^T$ and $h = (h_1, h_2, \ldots, h_n)^T$ be two nonzero vectors in $\mathbb{R}^{n\times n}$. Then $(g^T h)^2 \leq \|g\|^2\|h\|^2$ and equality holds iff g is a constant multiple of h.

Clearly $f(x) = \frac{1}{b}(\sum_{i=1}^{n} a_i x_i)(\sum_{i=1}^{n} \frac{c_i}{x_i}) = \frac{1}{b}\|g\|^2\|h\|^2$, where $g = (\sqrt{a_1 x_1}, \ldots, \sqrt{a_n x_n})^T$ and $h = (\sqrt{\frac{c_1}{x_1}}, \ldots, \sqrt{\frac{a_n}{x_n}})^T$. By the Schwartz inequality, $\|g\|^2\|h\|^2 \geq (g^T h)^2 = (\sum_{i=1}^{n} \sqrt{a_i c_i})^2$ and with equality iff $g = kh$ for some constant k. It follows that $f(x) \geq \frac{1}{b}(\sum_{i=1}^{n} \sqrt{a_i c_i})^2$ and equality holds iff $\sqrt{a_i x_i} = k\sqrt{\frac{c_i}{x_i}}$ or $x_i = k\sqrt{\frac{c_i}{a_i}}$ since $x_i \geq 0$ for $i = 1, 2, \ldots, n$. Now plug x_i into the equality constraint, we have $k\sum_{i=1}^{n} \sqrt{a_i c_i} = b$ and thus $k = \frac{b}{\sum_{i=1}^{n} \sqrt{a_i c_i}}$. From the above arguments, we can conclude that the optimal value $\frac{1}{b}(\sum_{i=1}^{n} \sqrt{a_i c_i})^2$ is achieved at $x = k(\sqrt{\frac{c_1}{a_1}}, \ldots, \sqrt{\frac{c_n}{a_n}})^T = \frac{b}{\sum_{i=1}^{n} \sqrt{a_i c_i}}(\sqrt{\frac{c_1}{a_1}}, \ldots, \sqrt{\frac{c_n}{a_n}})^T$. $\square$

Local and global extrema

Definition 1

A point $x^* \in \Omega$ is called a local minimum point of f over Ω if there exists $\delta > 0$ such that $f(x^*) \leq f(x) \quad \forall\, x \in \Omega$ with $\|x - x^*\| < \delta$.

An optimal solution of (1) is also called a global minimum point. Similarly, we can define local maximum point and global maximum point. Every local minimum or maximum point is called a local extremum.

Sufficient condition for local extrema (1d)

Theorem 1

If $f^{(k)}(x^*) = 0$ *for* $k = 1, 2, \ldots, n$, $f^{(n+1)}(x^*) \neq 0$, *and* $f^{(n+1)}(x)$ *is continuous in a neighborhood of* x^*, *then* x^* *is a local extremum iff* $n + 1$ *is even. Furthermore, for even* $n + 1$, *if* $f^{(n+1)}(x^*) > 0$ *then* x^* *is a local minimum; if* $f^{(n+1)}(x^*) < 0$ *then* x^* *is a local maximum.*

Proof.

To justify the statements, let h be an arbitrary number with sufficiently small $|h|$ (keep in mind that h can be both positive and negative). According to Taylor's theorem, we have

$$f(x^* + h) - f(x^*) = \sum_{k=1}^{n} \frac{f^{(k)}(x^*)}{k!} h^k + \frac{f^{(n+1)}(x^* + \theta h)}{(n+1)!} h^{n+1}$$

for some $0 < \theta < 1$. It follows from the hypothesis that

$$\begin{aligned} f(x^* + h) - f(x^*) &= \frac{f^{(n+1)}(x^* + \theta h)}{(n+1)!} h^{n+1} \\ &= \frac{f^{(n+1)}(x^* + \theta h)}{(n+1)! f^{(n+1)}(x^*)} \left(f^{(n+1)}(x^*) h^{n+1}\right). \qquad (*) \end{aligned}$$

■

Proof.

(Cont.) By the continuity of $f^{(n+1)}(x)$, $f^{(n+1)}(x^* + \theta h)$ has the same sign as $f^{(n+1)}(x^*)$ for all h with sufficiently small $|h|$. Thus

$$\frac{f^{(n+1)}(x^* + \theta h)}{(n+1)! f^{(n+1)}(x^*)} > 0 \tag{**}$$

for all h with sufficiently small $|h|$. Now we are ready to complete the proof. Clearly, x^* is a local maximum iff $f(x^* + h) - f(x^*) \leq 0$ for all h with sufficiently small $|h|$ iff, by (*) and (**), $f^{(n+1)}(x^*)h^{n+1} \leq 0$ for all h with sufficiently small $|h|$ iff $n+1$ is even (why?) and $f^{(n+1)}(x^*) < 0$. Similarly, x^* is a local minimum iff $f(x^* + h) - f(x^*) \geq 0$ for all h with sufficiently small $|h|$ iff, by (*) and (**), $f^{(n+1)}(x^*)h^{n+1} \geq 0$ for all h with sufficiently small $|h|$ iff $n+1$ is even and $f^{(n+1)}(x^*) > 0$. ∎

Examples

1. Find all the local extrema of the function

$$f(x) = x^3 - 9x^2 + 27x - 27.$$

2. Find all the local extrema of the function

$$f(x) = x^4 - 8x^3 + 24x^2 - 32x + 16.$$

Multivariable calculus basics

- ∇f — The gradient of f, i.e., $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^T$.

- $\nabla^2 f$ — The Hessian Matrix of the 2^{nd} partial derivatives of f, i.e., $\nabla^2 f$ is the $n \times n$ matrix whose (i, j) entry is $\frac{\partial^2 f}{\partial x_i \partial x_j}$.

- $f \in C^k \iff f$ has continuous k^{th} partial derivatives.

Multivariable calculus basics

Theorem 2

Let $f \in C^1$ and let $x(\alpha) = x^* + \alpha d$ and $g(\alpha) = f(x(\alpha))$. Then

$$g'(\alpha) = \nabla f(x(\alpha))^T d.$$

Multivariable calculus basics

Theorem 2

Let $f \in C^1$ *and let* $x(\alpha) = x^* + \alpha d$ *and* $g(\alpha) = f(x(\alpha))$. *Then*

$$g'(\alpha) = \nabla f(x(\alpha))^T d.$$

Theorem 3

Let $f \in C^2$ *and let* $x(\alpha) = x^* + \alpha d$ *and* $g(\alpha) = f(x(\alpha))$. *Then*

$$g''(\alpha) = d^T \nabla^2 f(x(\alpha)) d.$$

Multivariable calculus basics

Theorem 4 (Taylor's Theorem: Multidimensional Case)

Let $f \in C^2$. Then

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

for some $0 \leq \theta \leq 1$.

Proof.

(Theorem 2) By definition

$$\begin{aligned} g(\alpha) &= f(x_1 + \alpha d_1, x_2 + \alpha d_2, \ldots, x_n + \alpha d_n) \\ &\triangleq f(y_1, y_2, \ldots, y_n), \end{aligned}$$

where $y_i = x_i + \alpha d_i,\ i = 1, 2, \ldots, n$. It follows from the Chain Rule that

$$\begin{aligned} g'(\alpha) &= \frac{\partial f}{\partial y_1} \cdot \frac{dy_1}{d\alpha} + \frac{\partial f}{\partial y_2} \cdot \frac{dy_2}{d\alpha} + \cdots + \frac{\partial f}{\partial y_n} \cdot \frac{dy_n}{d\alpha} \\ &= \frac{\partial f}{\partial y_1} \cdot d_1 + \frac{\partial f}{\partial y_2} \cdot d_2 + \cdots + \frac{\partial f}{\partial y_n} \cdot d_n \\ &= \left(\frac{\partial f}{\partial y_1}, \ldots, \frac{\partial f}{\partial y_n} \right) \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix} \\ &= \nabla f(x(\alpha))^T d. \end{aligned}$$

$\blacksquare$

Proof.

(Theorem 3) Recall the proof of the preceding theorem, we have

$$g'(\alpha) = \frac{\partial f}{\partial y_1} \cdot d_1 + \frac{\partial f}{\partial y_2} \cdot d_2 + \cdots + \frac{\partial f}{\partial y_n} \cdot d_n.$$

From the Chain Rule, it can be seen that

$$\begin{aligned}
g''(\alpha) &= \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_1 \partial y_i} \, \frac{dy_i}{d\alpha} \cdot d_1 + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_2 \partial y_i} \, \frac{dy_i}{d\alpha} \cdot d_2 + \cdots + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_n \partial y_i} \, \frac{dy_i}{d\alpha} \cdot d_n \\
&= \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_1 \partial y_i} \, d_i d_1 + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_2 \partial y_i} \, d_i d_2 + \cdots + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_n \partial y_i} \, d_i d_n \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_j \partial y_i} \, d_j d_i = (d_1, \ldots, d_n) \left[\frac{\partial^2 f}{\partial y_i \partial y_j} \right] \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} \\
&= d^T \nabla^2 f(x(\alpha)) d.
\end{aligned}$$

∎

Proof.

(Theorem 4) Let $g(\alpha) = f(x^* + \alpha d)$. Then by Theorems 2 and 3, we have $g'(\alpha) = \nabla f(x^* + \alpha d)^T d$ and $g''(\alpha) = d^T \nabla^2 f(x^* + \alpha d) d$. Thus from Taylor's theorem, it follows that

$$\begin{aligned} g(1) &= g(0) + g'(0)(1-0) + \frac{1}{2} g''(\theta)(1-0)^2 \\ &= f(x^*) + \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^* + \theta d) d. \end{aligned}$$

Or

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

for some $0 \leq \theta \leq 1$. ∎

First-order Necessary Conditions

Definition 2

Let $x \in \Omega$. A vector d is called a <u>feasible direction</u> at x if there exists $\delta > 0$ such that $x + \alpha d \in \Omega$ for any $0 \leq \alpha \leq \delta$.

First-order Necessary Conditions

Definition 2

Let $x \in \Omega$. A vector d is called a <u>feasible direction</u> at x if there exists $\delta > 0$ such that $x + \alpha d \in \Omega$ for any $0 \leq \alpha \leq \delta$.

Theorem 5 (First-Order Necessary Conditions)

Let $f \in C^1$ be a function on $\Omega \subseteq \mathbb{R}^n$. If x^* is a local minimum point of f over Ω, then $\nabla f(x^*)^T d \geq 0$ for any feasible direction $d \in \mathbb{R}^n$ at x^*.

Proof.

By definition, $\exists\ \delta > 0$ such that $x^* + \alpha d \in \Omega \quad \forall\ 0 \leq \alpha \leq \delta$. Set $x(\alpha) = x^* + \alpha d$ and define $g(\alpha) = f(x(\alpha))$ for $0 \leq \alpha \leq \delta$. Then $g(\alpha) \geq g(0)$ for all sufficiently small α as x^* is a local minimum point. Hence

$$g'(0) = \lim_{\alpha \to 0^+} \frac{g(\alpha) - g(0)}{\alpha} \geq 0.$$

The desired conclusion follows instantly from Theorem 2 as $g'(0) = \nabla f(x(0))^T d = \nabla f(x^*)^T d$. ∎

First-order Necessary Conditions

Definition 3

A point x is called an <u>interior point</u> *of $\Omega \subseteq \mathbb{R}^n$ if there exists $\delta > 0$, such that $y \in \Omega$ for all y in $\mathbb{R}^n$ satisfying $\|x - y\| < \delta$.*

Remark. If x is an interior point of Ω, then any nonzero vector d in $\mathbb{R}^n$ is a feasible direction at x.

First-order Necessary Conditions

Definition 3

A point x is called an interior point of $\Omega \subseteq \mathbb{R}^n$ if there exists $\delta > 0$, such that $y \in \Omega$ for all y in $\mathbb{R}^n$ satisfying $\|x - y\| < \delta$.

Remark. If x is an interior point of Ω, then any nonzero vector d in $\mathbb{R}^n$ is a feasible direction at x.

Corollary 1 (Unconstrained Case)

Let $f \in C^1$ be a function on $\Omega \subseteq \mathbb{R}^n$. If x^* is a local minimum point of f over Ω and x^* is an interior point of Ω, then $\nabla f(x^*) = 0$.

Remark. This condition leads to n equations in n variables, which can be used to determine a solution in many cases.

Convexity

Definition 4

A set $\Omega \subseteq \mathbb{R}^n$ is called convex if for any two points $x, y \in \Omega$ and any $0 \leq \alpha \leq 1$, there holds $\alpha x + (1-\alpha)y \in \Omega$.

Convexity

Definition 4

A set $\Omega \subseteq \mathbb{R}^n$ is called <u>convex</u> if for any two points $x, y \in \Omega$ and any $0 \leq \alpha \leq 1$, there holds $\alpha x + (1-\alpha)y \in \Omega$.

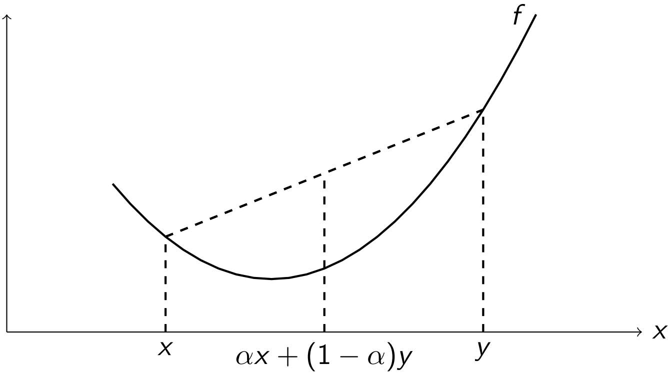
Definition 5

A function f defined on a convex set Ω is said to be <u>convex</u> if, for every $x,\ y \in \Omega$ and every $\alpha,\ 0 \leq \alpha \leq 1$, there holds

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y).$$

Function f is said to be <u>strictly convex</u> if, for every $x \neq y$ in Ω and every $\alpha,\ 0 < \alpha < 1$, there holds

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



Geometrically, a function is convex if and only if the line joining two points on its graph lies nowhere below the graph. Thinking of a function in two dimensions, it is convex if its graph is bowl shaped.

Property of convex function

Theorem 6

Let f and g be two convex functions on the convex set Ω. Then the following statements hold:

- (i) $f + g$ *is convex on* Ω;
- (ii) cf *is convex for any constant* $c > 0$;
- (iii) $\Gamma_c = \{x : \ x \in \Omega \quad and \quad f(x) \leq c\}$ *is a convex set for any* c.

Proof.

(i) and (ii) are immediate.
(iii) Let $x,\ y \in \Gamma_c$. Then $f(x) \le c$, $f(y) \le c$, and $x + (1-\alpha)y \in \Omega$ for any $0 \le \alpha \le 1$ since Ω is convex. By convexity, we have

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y) \le c.$$

So $\alpha x + (1-\alpha)y \in \Gamma_c$. ∎

Theorem 7

Let $f \in C^1$. Then f is convex over a convex set Ω iff

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

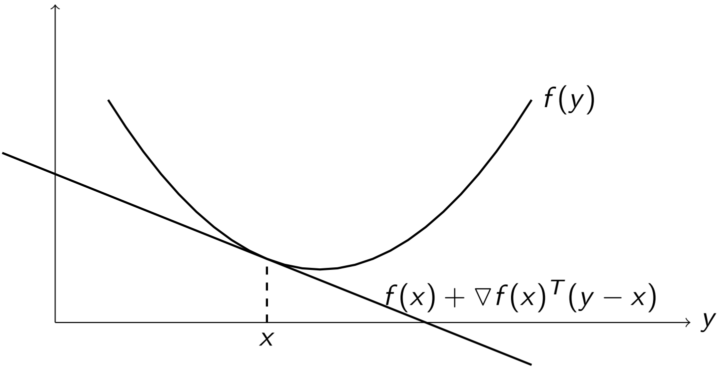
for all $x, y \in \Omega$.

Theorem 7

Let $f \in C^1$. Then f is convex over a convex set Ω iff

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

for all $x, y \in \Omega$.



Proof.

($\Longrightarrow$) If f is convex, then $\forall\ 0 < \alpha < 1$, we have

$$f(\alpha y + (1-\alpha)x) \leq \alpha f(y) + (1-\alpha) f(x).$$

Thus

$$\frac{f(x + \alpha(y-x)) - f(x)}{\alpha} \leq f(y) - f(x).$$

Letting $\alpha \to 0$ we obtain

$$\nabla f(x)^T (y-x) \leq f(y) - f(x).$$

∎

Proof.

(Cont.)
($\Longleftarrow$) Assume $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \Omega$, we aim to prove that for any $x_1, x_2 \in \Omega$ and $0 \leq \alpha \leq 1$, there holds

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

Setting $x = \alpha x_1 + (1-\alpha)x_2$ and $y = x_1$ or $y = x_2$ alternatively, we have

$$f(x_1) \geq f(x) + \nabla f(x)^T(x_1 - x) \tag{*}$$

$$f(x_2) \geq f(x) + \nabla f(x)^T(x_2 - x). \tag{**}$$

Multiplying (*) by α and (**) by $(1-\alpha)$ and adding them, we have

$$\begin{aligned} \alpha f(x_1) + (1-\alpha)f(x_2) &\geq f(x) + \nabla f(x)^T[\alpha x_1 + (1-\alpha)x_2 - x] \\ &= f(x) \end{aligned}$$

as desired. ∎

Theorem 8

Let $f \in C^2$. Then f is a convex function over a convex set Ω containing an interior point iff the Hessian matrix $\nabla^2 f(x)$ of f is positive semi-definite throughout Ω.

Corollary 2

Let $f \in C^2$ be a single variable function defined over an interval Ω. If $f''(x) \geq 0$ for all $x \in \Omega$, then f is convex over Ω.

Proof.

($\Longleftarrow$) For any $x, y \in \Omega$, by Taylor's Theorem we have

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x+\alpha(y-x))(y-x)$$

for some $0 < \alpha < 1$. Since $\nabla^2 f$ is positive semi-definite everywhere over Ω, $(y-x)^T \ \nabla^2 f(x+\alpha(y-x))(y-x) \geq 0$. Hence

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

If follows from Theorem 7 that f is convex. ∎

Proof.

(Cont.)
($\Longrightarrow$) Suppose the contrary: $\nabla^2 f$ is not positive semi-definite at some $x \in \Omega$, that is, $d^T \nabla^2 f(x) d < 0$ for some $d \in \mathbb{R}^n$. The continuity of Hessian allows us to assume that x is an interior point of Ω (why?). This assumption together with $f \in C^2$ guarantee the existence of a ball $\delta(x)$ centered at x such that $\delta(x) \subseteq \Omega$ and $d^T \nabla^2 f(y) d < 0$ for any $y \in \delta(x)$. Now let us scale d by a positive number λ so that $x + \frac{\alpha}{\lambda} d \in \delta(x)$ for all $0 \leq \alpha \leq 1$. Then we have $\frac{d^T}{\lambda} \nabla^2 f\left(x + \frac{\alpha d}{\lambda}\right) \frac{d}{\lambda} < 0$ for all $0 \leq \alpha \leq 1$. Thus

$$
\begin{aligned}
f\left(x + \frac{d}{\lambda}\right) &= f(x) + \nabla f(x)^T \frac{d}{\lambda} + \frac{1}{2}\left(\frac{d}{\lambda}\right)^T \nabla^2 f\left(x + \alpha \frac{d}{\lambda}\right) \frac{d}{\lambda} \quad (\exists\, 0 \leq \alpha \leq 1) \\
&< f(x) + \nabla f(x)^T \frac{d}{\lambda},
\end{aligned}
$$

contradicting Theorem 7. ∎

Example

Determine if the following functions are convex.

(a) $f(x_1, x_2) = (x_1 - x_2)^2 + 4x_1x_2 + e^{x_1+x_2}$;

(b) $f(x_1, x_2) = x_1 e^{-(x_1+x_2)}$.

Property of convex functions

Theorem 9

Let f be a convex function defined on a convex set Ω. Then

(i) *the set where f achieves the minimum is convex;*

(ii) *any local minimum point of f is also a global minimum point.*

Geometrically, (i) implies that all the minimum points are located together.

Proof.

(i) Let c be the minimum value of f over Ω. Then by Theorem 6 (iii), $\{x : x \in \Omega$ and $f(x) \le c\}$ is convex.

(ii) Suppose the contrary: some local minimum point x is not global minimum. Then there exists $y \in \Omega$ such that $f(y) < f(x)$. By local minimality of x, there is a neighborhood $\delta(x)$ of x such that $f(z) \ge f(x)$ for any $z \in \delta(x)$. Now choose $0 < \alpha < 1$ so that $z = \alpha x + (1-\alpha)y \in \delta(x)$. Then $f(z) = f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y) < f(x)$, a contradiction. ∎

Optimality conditions

Theorem 10

Let $f \in C^1$ be convex on the convex set Ω. If there exists $x^* \in \Omega$ such that

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall \quad x \in \Omega$$

then x^* is a global minimum of f over Ω.

Remark. For convex functions, the first-order necessary conditions are both necessary and sufficient for a point to be a global minimum.

Proof.

By Theorem 7, for any $x \in \Omega$, we have

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*) \geq f(x^*).$$

∎

Example

Solve the following problem

$$\min \quad f(x) = \frac{1}{2}x^T Ax - b^T x,$$

where A is an $n \times n$ positive definite matrix.

Solution. Clearly $\nabla f(x) = Ax - b$ and $\nabla^2 f(x) = A$. Since A is positive definite, by Theorem 8, $f(x)$ is a convex function. Notice that $x^* = A^{-1}b$ is the unique solution to $\nabla f(x) = 0$. By Theorem 10 and Corollary 1, x^* is the unique optimal solution. □

Question 4. Let A be an $m \times n$ matrix of rank n and let $b \in \mathbb{R}^m$. What is the solution to the following problem

$$\min \quad \|Ax - b\|?$$