HKU MATH3904 Study Notes

Ka Long (Leo) Chiu[*]

Last Updated: 2025-12-12

Contents

[*] email ✉: `leockl@connect.hku.hk`; personal website 🔗: `https://leochiukl.github.io`

1 Basic Properties of Solutions and Algorithms

1.0.1 **Setup.** In MATH3904, we will focus on analyzing the following kind of optimization (minimization) problem:

$$\begin{array}{rl} \text{Minimize} & f(x_1, \dots, x_n) \\ \text{subject to} & (x_1, \dots, x_n) \in \Omega, \end{array}$$

where f is a real-valued function and $\Omega \subseteq \mathbb{R}^n$. For convenience, we often express the optimization problem above simply as follows:

$$\begin{array}{rl} \min & f(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{x} \in \Omega, \end{array}$$

or even just "$\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$".

The optimization problem above is called an **unconstrained optimization problem** if $\Omega = \mathbb{R}^n$, and **constrained optimization problem** otherwise. The following are some commonly used terms for describing the above optimization problem:

- f is the **objective function**.
- Ω is the **feasible region**.
- Every $\boldsymbol{x} \in \Omega$ is a **feasible point** or **feasible solution**.
- A point $\boldsymbol{x}^* \in \Omega$ is an **optimal solution** to the problem if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$, and $f(\boldsymbol{x}^*)$ is called the **optimal value** (or **minimum value** as we are dealing with a minimization problem; for maximization problem this can be called **maximum value**).

The main goal of MATH3904 is to develop *efficient* algorithms for solving an optimization problem of the above form, i.e., finding an *optimal* solution to it (not just a feasible solution!).

[Note: We can focus on minimization problems without loss of generality since the maximization problem $\max_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$ is equivalent to the minimization problem $\min_{\boldsymbol{x}\in\Omega} -f(\boldsymbol{x})$, in the sense that solving one of the problem already gives us enough ingredients to readily solve another problem.]

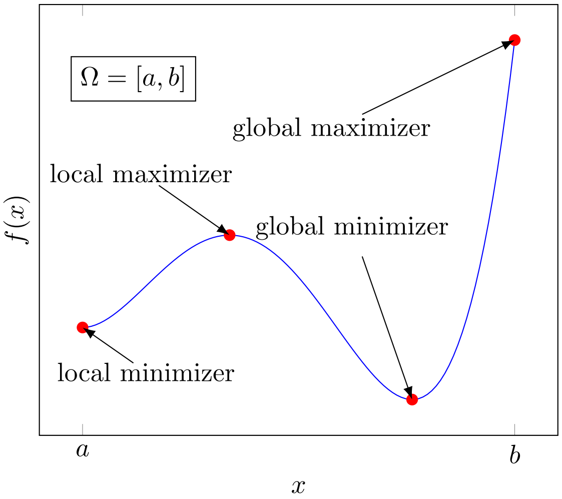
1.1 Necessary and Sufficient Conditions for Local Extremizers

1.1.1 **Why local extremizers?** To solve a minimization problem, we need to explicitly find an optimal solution that minimizes the objective function f over the feasible region Ω. However, in general, this task is *notoriously hard*, if not impossible, from both theoretical and computational viewpoints. Consequently, we often need to resort to a "second best", which is described by the concept of *local* extremizer.

1.1.2 **Definitions.** A point $\boldsymbol{x}^* \in \Omega \subseteq \mathbb{R}^n$ is called a **local minimizer** of f over Ω if there exists $\delta > 0$ such that $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$ with $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta$, where $\|\cdot\|$ denotes the Euclidean norm. In words, this means that the point $\boldsymbol{x}^*$ can minimize f over a *sufficiently small neighbourhood* of itself (so "locally"). In contrast, an optimal solution to the minimization problem is called an **global minimizer** (or simply **minimizer**), since it minimizes f over the whole feasible region Ω (so "globally").

In a similar way, we can define **local maximizer** and **global maximizer** (**maximizer**). Every local minimizer or maximizer is called a **local extremizer**.

These concepts are illustrated in the picture below:



1.1.3 **A sufficient condition for local extremizers of univariate functions.** For univariate functions, the following result is a standard tool for finding local extremizers (you may have already learnt it in your calculus class):

Proposition 1.1.a. Suppose that $f^{(k)}(x^*) = 0$ for all $k = 1, \ldots, n$, $f^{(n+1)}(x^*) \neq 0$, and $f^{(n+1)}$ is continuous in a neighbourhood of x^*. We have:

(a) x^* is a local extremizer iff $n + 1$ is even.

(b) If $n + 1$ is even, then:

i. $f^{(n+1)}(x^*) > 0$ implies that x^* is a local minimizer.

ii. $f^{(n+1)}(x^*) < 0$ implies that x^* is a local maximizer.

Proof. We will prove (a) and (b) at once. Fix any $h \in \mathbb{R}$ with sufficiently small $|h|$ (more details below). By 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 $0 < \theta < 1$. By assumption, we have $f^{(k)}(x^*) = 0$ for all $k = 1, \ldots, n$ and $f^{(n+1)}(x^*) \neq 0$. Thus, we can write

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

By the assumed continuity of $f^{(n+1)}$ on a neighbourhood of x^*, we know that with sufficiently small $|h|$, $f^{(n+1)}(x^* + \theta h)$ has the same sign as $f^{(n+1)}(x^*)$, implying that

$$\frac{f^{(n+1)}(x^* + \theta h)}{(n+1)! f^{(n+1)}(x^*)} > 0. \tag{2}$$

Now, consider:

$$\begin{aligned} x^* \text{ is a local maximizer} &\iff f(x^* + h) - f(x^*) \leq 0 \text{ for all } h \text{ with sufficiently small } |h| \\ &\overset{(1),(2)}{\iff} f^{(n+1)}(x^*) h^{n+1} \leq 0 \text{ for all } h \text{ with sufficiently small } |h| \\ &\iff n + 1 \text{ is even and } f^{(n+1)}(x^*) < 0, \end{aligned}$$

and, similarly,

$$\begin{aligned} x^* \text{ is a local minimizer } &\iff f(x^*+h) - f(x^*) \geq 0 \text{ for all } h \text{ with sufficiently small } |h| \\ &\overset{(1),(2)}{\iff} f^{(n+1)}(x^*)h^{n+1} \geq 0 \text{ for all } h \text{ with sufficiently small } |h| \\ &\iff n+1 \text{ is even and } f^{(n+1)}(x^*) > 0. \end{aligned}$$

This establishes (b) clearly. Also, (a) is established since a local extremizer refers to a local maximizer or minimizer. □

1.1.4 **Notations and terms from multivariable calculus.** We recall some notations and terms learnt in MATH2211, which will be useful for our optimization studies here:

- ∇f: the **gradient** of f, which is given by $(\partial f/\partial x_1\,,\ldots,\,\partial f/\partial x_n)$. [Note: In MATH3904, vectors are all *column* vectors by default; for instance, the gradient here is a column vector.]
- $\nabla^2 f$: the **Hessian matrix** of the second partial derivatives of f, which is the $n \times n$ matrix whose (i,j)th entry is $\partial^2 f/\partial x_i\,\partial x_j$ for all $i,j = 1,\ldots,n$.
- $f \in C^k$: the function f is of **class C^k**, i.e., all its partial derivatives of order at most k exist and are continuous. [Note: This is equivalent to the following definition: f has continuous kth partial derivatives; "⇒" is immediate and "⇐" can be established by standard results in multivariable calculus.]

1.1.5 **Helper results from multivariable calculus.**

Proposition 1.1.b. Let $\boldsymbol{x}(\alpha) = \boldsymbol{x}^* + \alpha\boldsymbol{d}$, where $\boldsymbol{d} = (d_1,\ldots,d_n) \in \mathbb{R}^n$ is fixed and $\alpha \in \mathbb{R}$, and $g(\alpha) = f(\boldsymbol{x}(\alpha))$.

(a) If $f \in C^1$, then $g'(\alpha) = \nabla f(\boldsymbol{x}(\alpha))^T\boldsymbol{d}$.

(b) If $f \in C^2$, then $g''(\alpha) = \boldsymbol{d}^T\nabla^2 f(\boldsymbol{x}(\alpha))\boldsymbol{d}$.

(c) *(Multidimensional Taylor's theorem)* If $f \in C^2$, then

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

for some $0 < \theta < 1$.

Proof.

(a) Letting $y_i := y_i(\alpha) = x_i^* + \alpha d_i$ for all $i = 1,\ldots,n$, we can write $g(\alpha) = f(y_1,\ldots,y_n)$. Then, since $f \in C^1$ we can apply chain rule to get

$$\begin{aligned} g'(\alpha) &= \frac{\partial f}{\partial y_1}\frac{\mathrm{d}y_1}{\mathrm{d}\alpha} + \cdots + \frac{\partial f}{\partial y_n}\frac{\mathrm{d}y_n}{\mathrm{d}\alpha} = \frac{\partial f}{\partial y_1}d_1 + \cdots + \frac{\partial f}{\partial y_n}d_n \\ &= \begin{bmatrix} \dfrac{\partial f}{\partial y_1} & \cdots & \dfrac{\partial f}{\partial y_n} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = \nabla f(y_1,\ldots,y_n)^T\boldsymbol{d} = \nabla f(\boldsymbol{x}(\alpha))^T\boldsymbol{d}. \end{aligned}$$

(b) From the proof of (a), we have

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

Differentiating this with respect to α again and applying chain rule on each of the partial derivatives $\partial f/\partial y_1, \dots, \partial f/\partial y_n$ (allowed since $f \in C^2$), we get

$$\begin{aligned} g''(\alpha) &= \left(\sum_{i=1}^{n} \frac{\partial}{\partial y_i}\left(\frac{\partial f}{\partial y_1}\right)\frac{\mathrm{d}y_i}{\mathrm{d}\alpha}\right)d_1 + \cdots + \left(\sum_{i=1}^{n} \frac{\partial}{\partial y_i}\left(\frac{\partial f}{\partial y_n}\right)\frac{\mathrm{d}y_i}{\mathrm{d}\alpha}\right)d_n \\ &\overset{(f\in C^2)}{=} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_1\,\partial y_i} d_i d_1 + \cdots + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_1\,\partial y_i} d_i d_n \\ &= \sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 f}{\partial y_j\,\partial y_i} d_j d_i = (d_1, \dots, d_n)\left[\frac{\partial^2 f}{\partial y_i\,\partial y_j}\right]\begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \\ &= \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}(\alpha))\boldsymbol{d}. \end{aligned}$$

(c) Applying univariate Taylor's theorem on g, we have that for some $0 < \theta < 1$,

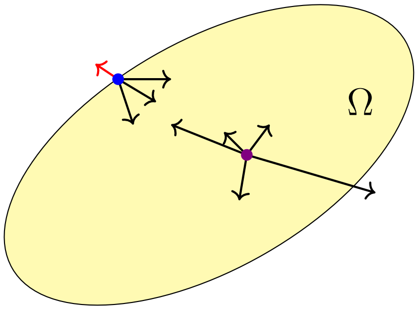
$$\begin{aligned} f(\boldsymbol{x}^* + \boldsymbol{d}) = g(1) &= g(0) + g'(0)(1-0) + \frac{1}{2}g''(\theta)(1-0)^2 \\ &\overset{\text{(a),(b)}}{=} f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^T\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x}^* + \theta\boldsymbol{d})\boldsymbol{d}.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

□

1.1.6 **Feasible directions.** After spending some time on notations, terms, and results about multivariable calculus, let us go back to the discussion of local extremizers. In [1.1.3], we have studied a *sufficient* condition for local extremizers. Our next task is to derive *necessary* conditions for local extremizers. To do this, it turns out to be helpful to consider directions along which slight movements away from a point are *feasible*, i.e., the destinations remain inside the feasible region. This leads to the concept of *feasible direction*.

Let $\boldsymbol{x} \in \Omega \subseteq \mathbb{R}^n$. A vector $\boldsymbol{d} \in \mathbb{R}^n$ is called a **feasible direction** at $\boldsymbol{x}$ if there exists $\delta > 0$ such that $\boldsymbol{x} + \alpha\boldsymbol{d} \in \Omega$ for all $0 \le \alpha \le \delta$.

The black arrows in the picture below illustrate graphically some feasible directions of the two points below: When we move along that direction a bit, we will never leave the feasible region Ω. The red arrow, however, is not a feasible direction at the blue point since we will leave the feasible region Ω after even a slight movement along that direction.



1.1.7 **A first-order necessary condition for local minimizer.** Using the concept of feasible direction, we can derive a first-order necessary condition for local minimizer:

Proposition 1.1.c. Let $f \in C^1$ be a function on $\Omega \subseteq \mathbb{R}^n$. If $\boldsymbol{x}^*$ is a local minimizer of f over Ω, then $\nabla f(\boldsymbol{x}^*)^T\boldsymbol{d} \ge 0$ for all feasible directions $\boldsymbol{d}$ at $\boldsymbol{x}^*$.

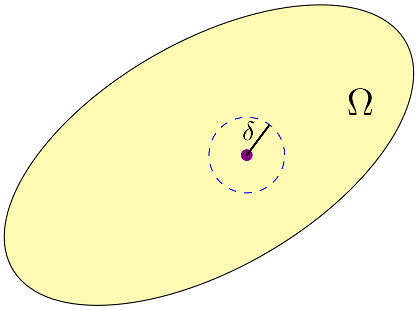
Proof. Fix any feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^*$. By definition, there exists $\delta > 0$ such that $\boldsymbol{x}^* + \alpha\boldsymbol{d} \in \Omega$ for all $0 \le \alpha \le \delta$. Let $\boldsymbol{x}(\alpha) = \boldsymbol{x}^* + \alpha\boldsymbol{d}$, and $g(\alpha) = f(\boldsymbol{x}(\alpha))$ for all $0 \le \alpha \le \delta$ (which is well-defined since

$\boldsymbol{x}(\alpha) \in \Omega$ for all $0 \leq \alpha \leq \delta$). Since $\boldsymbol{x}^*$ is a local minimizer, we have that $g(\alpha) \geq g(0)$ for all sufficiently small $\alpha \geq 0$. Thus,

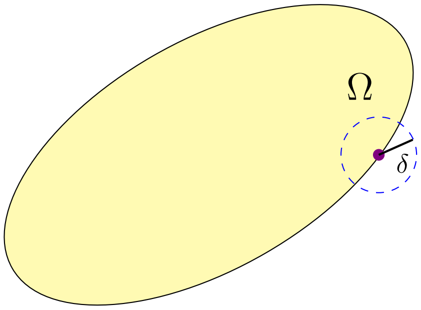
$$\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} \overset{\text{(Proposition 1.1.b)}}{=} g'(0) = \lim_{\alpha \to 0^+} \frac{g(\alpha) - g(0)}{\alpha} \geq 0.$$

□

1.1.8 **Interior points.** In the special case where the local minimizer in consideration is also an *interior point*, it turns out that we can say more. A point $\boldsymbol{x} \in \Omega$ is called an **interior point** of Ω if there exists $\delta > 0$ such that $\boldsymbol{y} \in \Omega$ for every $\boldsymbol{y} \in \mathbb{R}^n$ satisfying that $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$, or in other words, the *open ball centered at $\boldsymbol{x}$ with radius δ*, namely $B(\boldsymbol{x}, \delta) := \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{y}\| < \delta\}$, is a subset of Ω. The set of all interior points of Ω is called the **interior** of Ω.



[Note: A related concept is *boundary point*. A point $\boldsymbol{x} \in \mathbb{R}^n$ is called a **boundary point** of Ω if for all $\delta > 0$, we have $B(\boldsymbol{x}, \delta) \cap \Omega \neq \varnothing$ and $B(\boldsymbol{x}, \delta) \cap (\mathbb{R}^n \setminus \Omega) \neq \varnothing$. Similarly, the set of all boundary points of Ω is called the **boundary** of Ω.



]

1.1.9 **Feasible directions at interior points.** If $\boldsymbol{x}$ is an interior point of $\Omega \subseteq \mathbb{R}^n$, then *every* vector $\boldsymbol{d} \in \mathbb{R}^n$ is a feasible direction at $\boldsymbol{x}$.

Proof. First, $\boldsymbol{d} = \boldsymbol{0}$ is certainly a feasible direction at $\boldsymbol{x}$. So, henceforth assume that $\boldsymbol{d}$ is any nonzero vector in $\mathbb{R}^n$. By the definition of interior point, we have that $B(\boldsymbol{x}, \delta) \subseteq \Omega$ for some $\delta > 0$. The result then follows by noting that $\boldsymbol{x} + \alpha \boldsymbol{d} \in B(\boldsymbol{x}, \delta) \subseteq \Omega$ for all $0 \leq \alpha \leq \delta/(2\|\boldsymbol{d}\|)$. □

1.1.10 **A first-order necessary condition for interior local minimizer.** Using Proposition 1.1.c and [1.1.9], we are able to derive another first-order necessary condition for being a local minimizer that is also an interior point.

Corollary 1.1.d. Let $f \in C^1$ be a function on $\Omega \subseteq \mathbb{R}^n$. If $\boldsymbol{x}^*$ is a local minimizer of f over Ω and $\boldsymbol{x}^*$ is an interior point of Ω, then $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$.

Proof. Since $\boldsymbol{x}^*$ is an interior point of Ω, by [1.1.9] we know that every $\boldsymbol{d} \in \mathbb{R}^n$ is a feasible direction at $\boldsymbol{x}^*$. So, applying Proposition 1.1.c with $\boldsymbol{d} = -\nabla f(\boldsymbol{x}^*)$ gives

$$\nabla f(\boldsymbol{x}^*)^T(-\nabla f(\boldsymbol{x}^*)) \geq 0 \implies -\nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*) \geq 0 \implies \nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*) \leq 0.$$

On the other hand, $\nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*)$ is just the sum of squared entries in the vector $\nabla f(\boldsymbol{x}^*)$, which is always nonnegative. Thus, we must have $\nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*) = 0$, forcing that $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. □

[Note: A point $\boldsymbol{x} \in \Omega$ satisfying $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ is called a **stationary point** of f. So, this result suggests that every interior local minimizer of f must be a stationary point (but not vice versa).]

The condition $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ leads to a system of n equations in n variables, which can be utilized to determine candidates for local minima that are also interior points. In the univariate case, this condition corresponds to the familiar "setting the derivative to zero".

1.1.11 **A second-order necessary condition for local minimizer.** Proposition 1.1.c gives us a *first-order* necessary condition for local minimizer, involving the gradient ∇f only. It turns out that if the function f is of class C^2, then we also have a *second-order* necessary condition, involving both the gradient ∇f and the Hessian matrix $\nabla^2 f$.

Proposition 1.1.e. Let $f \in C^2$ be a function on $\Omega \subseteq \mathbb{R}^n$. If $\boldsymbol{x}^*$ is a local minimizer of f over Ω, then for all feasible directions $\boldsymbol{d}$ at $\boldsymbol{x}^*$, we have:

(a) $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} \geq 0$.

(b) $\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{d} \geq 0$ whenever $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} = 0$.

Proof.

(a) It follows from Proposition 1.1.c directly.

(b) By the definition of feasible direction, there exists $\delta > 0$ such that $\boldsymbol{x}^* + \alpha\boldsymbol{d} \in \Omega$ for all $0 \leq \alpha \leq \delta$. Let $\boldsymbol{x}(\alpha) = \boldsymbol{x}^* + \alpha\boldsymbol{d}$ and $g(\alpha) = f(\boldsymbol{x}(\alpha))$ for all $0 \leq \alpha \leq \delta$. Applying Proposition 1.1.b gives $g'(0) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} \overset{\text{(assumption)}}{=} 0$ and $g''(0) = \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{d}$. Also, since $\boldsymbol{x}^*$ is a local minimizer, we have $g(\alpha) \geq g(0)$ for all sufficiently small $\alpha \geq 0$. Hence,

$$0 \leq \lim_{\alpha \to 0^+} \frac{g(\alpha) - g(0)}{\alpha^2} \overset{\text{(L'Hôpital)}}{=} \lim_{\alpha \to 0^+} \frac{g'(\alpha)}{2\alpha} = \frac{1}{2} \lim_{\alpha \to 0^+} \frac{g'(\alpha) - g'(0)}{\alpha - 0} = \frac{1}{2} g''(0) = \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{d},$$

as desired.

□

1.1.12 **A second-order necessary condition for interior local minimizer.** Using Corollary 1.1.d and Proposition 1.1.e, we can derive another second-order necessary condition for being a local minimizer that is also an interior point.

Corollary 1.1.f. Let $f \in C^2$ be a function on $\Omega \subseteq \mathbb{R}^n$. If $\boldsymbol{x}^*$ is a local minimizer of f over Ω and $\boldsymbol{x}^*$ is an interior point of Ω, then:

(a) $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$.

(b) $\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{d} \geq 0$ for all $\boldsymbol{d} \in \mathbb{R}^n$, i.e., the Hessian matrix $\nabla^2 f(\boldsymbol{x}^*)$ is *positive semidefinite.*

Proof. Part (a) follows directly from Corollary 1.1.d. Having (a), for all $\boldsymbol{d} \in \mathbb{R}^n$, we know that $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} = 0$, which implies that $\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{d} \geq 0$ by (b) of Proposition 1.1.e. This establishes part (b). □

1.1.13 **Positive (semi)definiteness.** In optimization theory, the notion of *positive (semi)definiteness* is frequently encountered. So, let us now spend some time on introducing/reviewing definitions and properties related to this notion.

- *Definitions.*
  - A *symmetric* matrix $A \in \mathbb{R}^{n\times n}$ is **positive definite** if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$.
  - A *symmetric* matrix $A \in \mathbb{R}^{n\times n}$ is **positive semidefinite** if $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
  - Let $A = \begin{bmatrix} a_{ij} \end{bmatrix} \in \mathbb{R}^{n\times n}$. Then:
    - The matrix $A_k = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}$ is the **principal submatrix** made up of the first k rows and k columns, for every $k = 1, \ldots, n$.
    - A submatrix B of A is called a **symmetric submatrix** if it is of the form

      $$B = \begin{bmatrix} a_{i_1 i_1} & \cdots & a_{i_1 i_k} \\ \vdots & \ddots & \vdots \\ a_{i_k i_1} & \cdots & a_{i_k i_k} \end{bmatrix}$$

      where $1 \leq i_1 < \cdots < i_k \leq n$ with $k \in \{1, \ldots, n\}$, or in words, a submatrix of A obtained by selecting the same set of row and column indices (namely $\{i_1, \ldots, i_k\}$ above). [Note: If A is a symmetric matrix, then every symmetric submatrix of A is indeed symmetric. However, the same cannot be said when A is not a symmetric matrix.]
- *Properties.* Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix.
  - The following statements are equivalent.

    (a) A is positive definite.

    (b) The eigenvalues of A are all *positive.*

    (c) The determinant of every *principal submatrix* of A is *positive.*
  - The following statements are equivalent.

    (a) A is positive semidefinite.

    (b) The eigenvalues of A are all *nonnegative.*

    (c) The determinant of every *symmetric submatrix* of A is *nonnegative.*

    [⚠ Warning: It is NOT true that A is positive semidefinite iff the determinant of every principal submatrix is nonnegative. To see this, consider $A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$. The only principal matrices of A, namely $\begin{bmatrix} 0 \end{bmatrix}$ and A itself, both have nonnegative (indeed zero) determinant. However, A is not positive semidefinite since $\begin{bmatrix} 0 & 1 \end{bmatrix} A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = -1 < 0$.]
  - When $n = 2$, the matrix A is positive semidefinite iff $a_{11} \geq 0$, $a_{22} \geq 0$, and $\det A \geq 0$, where a_{ij} denotes the (i,j)-entry of A for every $i, j = 1, 2$.

    Proof. It follows from noting that the only symmetric submatrices of $A \in \mathbb{R}^{2\times 2}$ are $\begin{bmatrix} a_{11} \end{bmatrix}$, $\begin{bmatrix} a_{22} \end{bmatrix}$, and A itself, and applying the previous result that A is positive semidefinite iff the determinant of every symmetric submatrix of A is nonnegative. □

1.1.14 **Strict local minimizer.** We have been discussing *necessary* conditions for *local minimizer*. Using the notion of positive definiteness, it turns out to be possible to derive a *sufficient* condition about a notion that is even stronger than local minimizer, which is known as *strict* local minimizer. A point $\boldsymbol{x}^* \in \Omega \subseteq \mathbb{R}^n$ is a **strict local minimizer** of f over Ω if there exists $\delta > 0$ such that $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$ with $0 < \|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta$. In words, this means that the point $\boldsymbol{x}^*$ is the *unique* one that minimizes f over a sufficiently small neighbourhood of itself.

The differences between this definition and the definition of local minimizer are colored in blue: (i) the inequality between $f(\boldsymbol{x}^*)$ and $f(\boldsymbol{x})$ becomes strict, and (ii) we require $\|\boldsymbol{x} - \boldsymbol{x}^*\|$ to be strictly greater than zero, for excluding the case $\boldsymbol{x} = \boldsymbol{x}^*$.

1.1.15 **A second-order sufficient condition for interior strict local minimizer.** Now we have enough ingredients to derive a second-order sufficient condition for interior strict local minimizer:

Proposition 1.1.g. Let $f \in C^2$ be a function on $\Omega \subseteq \mathbb{R}^n$ and $\boldsymbol{x}^*$ be an interior point of Ω. If (i) $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and (ii) $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, then $\boldsymbol{x}^*$ is a strict local minimizer of f.

Proof. **Claim:** $\nabla^2 f(\boldsymbol{x}^* + \boldsymbol{d})$ is also positive definite whenever $\|\boldsymbol{d}\|$ is sufficiently small.

Proof of claim. By [1.1.13], it suffices to prove that there is $\delta > 0$ such that the determinant of each principal submatrix of $\nabla^2 f(\boldsymbol{x}^* + \boldsymbol{d})$ is positive whenever $\|\boldsymbol{d}\| < \delta$. Fix any $k = 1, \dots, n$, and let H_k ($H_k(\boldsymbol{d})$) denote the principal submatrix made up of the first k rows and k columns of $\nabla^2 f(\boldsymbol{x}^*)$ ($\nabla^2 f(\boldsymbol{x}^* + \boldsymbol{d})$).

Expressing the determinant $\det H_k(\boldsymbol{d})$ in the permutation form, we get

$$\det H_k(\boldsymbol{d}) = \sum_{\sigma} (-1)^{|\sigma|} \underbrace{\left.\frac{\partial^2 f}{\partial x_1\, \partial x_{\sigma(1)}}\right|_{\boldsymbol{x}^* + \boldsymbol{d}}}_{\text{continuous in } \boldsymbol{d}} \cdots \underbrace{\left.\frac{\partial^2 f}{\partial x_k\, \partial x_{\sigma(k)}}\right|_{\boldsymbol{x}^* + \boldsymbol{d}}}_{\text{continuous in } \boldsymbol{d}}.$$

where the sum is taken over all the $k!$ permutation $\sigma = (\sigma(1), \dots, \sigma(k))$ of $\{1, \dots, k\}$, $|\sigma|$ is the number of inversions of the permutation σ, i.e., the number of times where $i < j$ while $\sigma(i) > \sigma(j)$, and the continuity follows from the assumption that $f \in C^2$. Hence, by the continuity we have $\lim_{\boldsymbol{d} \to \mathbf{0}} \det H_k(\boldsymbol{d}) = \det H_k > 0$. This means that there is $\delta_k > 0$ such that $\det H_k(\boldsymbol{d}) > 0$ whenever $\|\boldsymbol{d}\| < \delta_k$.

Finally, by choosing $\delta = \min\{\delta_1, \dots, \delta_n\} > 0$, we have $\det H_k(\boldsymbol{d}) > 0$ for every $k = 1, \dots, n$, whenever $\|\boldsymbol{d}\| < \delta$, as desired. □

Now, we apply Proposition 1.1.b to get

$$f(\boldsymbol{x}^* + \boldsymbol{d}) - f(\boldsymbol{x}^*) = \underbrace{\nabla f(\boldsymbol{x}^*)}_{\mathbf{0} \text{ by assumption}}{}^T \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^* + \theta \boldsymbol{d}) \boldsymbol{d} = \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^* + \theta \boldsymbol{d}) \boldsymbol{d}$$

for some $0 < \theta < 1$. For all nonzero $\boldsymbol{d}$ with sufficiently small $\|\boldsymbol{d}\|$, by the claim above we have that $\nabla^2 f(\boldsymbol{x}^* + \theta \boldsymbol{d})$ is positive definite, and hence $\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^* + \theta \boldsymbol{d}) \boldsymbol{d} > 0$, or $f(\boldsymbol{x}^*) < f(\boldsymbol{x}^* + \boldsymbol{d})$. Therefore, $\boldsymbol{x}^*$ is a strict local minimizer of f. □

[⚠ Warning: We do NOT have an analogous result where (ii) is replaced by "$\nabla^2 f(\boldsymbol{x}^*)$ is positive *semidefinite*" and "strict local minimizer" is replaced by "local minimizer". To see this, let $f(x) = x^3$. While we have $\nabla f(0) = f'(0) = 0$ and $\nabla^2 f(0) = f''(0) = 0$ (positive semidefinite), 0 is not a local minimizer of f.]

1.2 From Local to Global

1.2.1 **A general procedure for finding local minimizers.** Utilizing the sufficient and necessary conditions for local minimizer from Section 1.1, we can follow the general procedure below to find local minimizers:

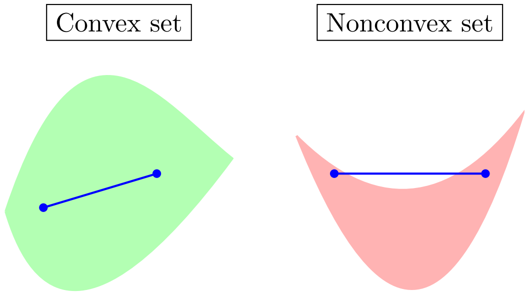
(1) *(Identifying candidates of local minimizer via necessary conditions)* Using first-order and/or second-order necessary conditions, namely Propositions 1.1.c and 1.1.e for general feasible points and Corollaries 1.1.d and 1.1.f for interior feasible points, we exclude all feasible points that are impossible to be local minimizers, and the remaining ones (hopefully only finitely many!) are the candidates of local minimizer.

(2) *(Checking whether each candidate is a local minimizer via sufficient conditions)* For each candidate of local minimizer that is an interior point, we can examine the second-order sufficient condition (Proposition 1.1.g) to check whether it is fulfilled. If it is the case, then we have shown that the candidate is a local minimizer. If not, then we still do not know whether the candidate is a local minimizer, and other techniques are needed to determine whether it is a local minimizer or not (e.g., by definition); the same should also be done for all candidates of local minimizer that are *not* interior points.

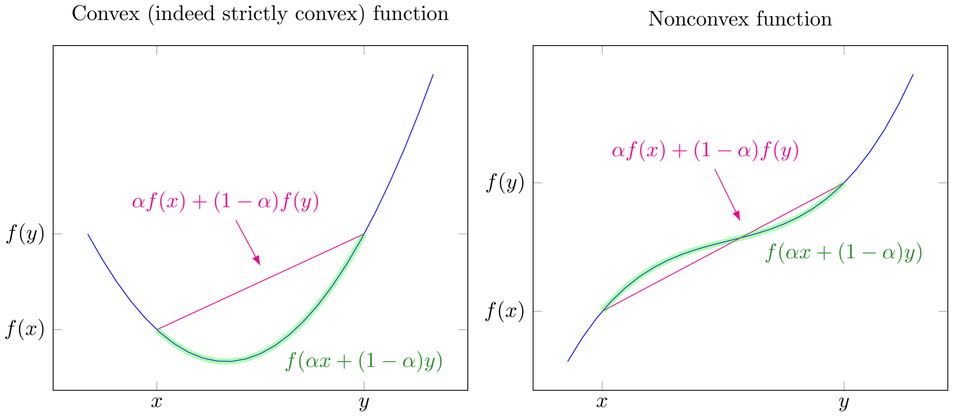
This kind of general procedure is fundamental for optimization theory and serves as an important guideline for designing various algorithms for solving optimization problems to be discussed later. Nonetheless, the procedure described above only applies for finding *local* minimizers, and our ultimate goal, namely solving optimization problems efficiently, requires us to find out *global* minimizers. In order to move from "local" to "global", it turns out that some kind of *convexity* assumptions is important, and thus we are going to have some discussions on the notion of convexity in the following.

1.2.2 **Convexity.** The concept of convexity applies to both *sets* and *functions*:

- *Sets.* A set $\Omega \subseteq \mathbb{R}^n$ is called **convex** if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and all $0 \le \alpha \le 1$, we have $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in \Omega$. Geometrically, a convex set Ω is a set that contains *every* line segment linking two points in Ω. The following pictures illustrate a convex set and a nonconvex set:



- *Functions.* A function f on a *convex* set $\Omega \subseteq \mathbb{R}^n$ is called **convex** if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and all $0 \le \alpha \le 1$, we have $f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \le \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$. Also, f is called **strictly convex** if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ with $\boldsymbol{x} \neq \boldsymbol{y}$ and all $0 < \alpha < 1$, we have $f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$. Geometrically, a function is convex if the line segment joining two points on its graph lies nowhere below the graph, and is strictly convex if the line segment with the two end-points excluded always lies strictly above the graph. More colloquially, a function is convex if its graph is "bowl-shaped".



1.2.3 **Characterizations of (strictly) convex functions.** While checking the convexity of a function by definition is always a valid approach, it can be sometimes cumbersome. So, the following provide some characterizations of (strictly) convex functions, giving us alternative and sometimes much more convenient methods for examining convexity.

Theorem 1.2.a (Characterizations of (strictly) convex functions)**.**

(a) A function f on a convex set Ω is convex iff for all integers $m \ge 2$, all $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \Omega$, and all $\alpha_1, \dots, \alpha_m \ge 0$ with $\sum_{i=1}^m \alpha_i = 1$, we have $f(\sum_{i=1}^m \alpha_i \boldsymbol{x}_i) \le \sum_{i=1}^m \alpha_i f(\boldsymbol{x}_i)$.

(b) Let $f \in C^1$ be a function on a convex set Ω. The function f is convex iff $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$.

(c) Let $f \in C^1$ be a function on a convex set Ω. The function f is strictly convex iff $f(\boldsymbol{y}) > f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$ for all distinct $\boldsymbol{x}, \boldsymbol{y} \in \Omega$.

(d) Let $f \in C^2$ be a function on a convex set Ω *containing an interior point*. The function f is convex iff the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ of f is positive semidefinite for all $\boldsymbol{x} \in \Omega$.

Proof.

(a) "$\Leftarrow$" is immediate since the definition of convexity just corresponds to the case of $m = 2$. So henceforth we focus on proving "$\Rightarrow$", and we will do this by induction. The base case $m = 2$ holds by the definition of convexity. Now, suppose the case $m = k$ holds, where $k \geq 2$ is an integer. Then,

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} \alpha_i \boldsymbol{x}_i\right) &= f\left(\alpha_1 \boldsymbol{x}_1 + \sum_{i=2}^{k+1} \alpha_i \boldsymbol{x}_i\right) = f\left(\alpha_1 \boldsymbol{x}_1 + \left(\sum_{j=2}^{k+1} \alpha_j\right) \sum_{i=2}^{k+1} \frac{\alpha_i}{\sum_{j=2}^{k+1} \alpha_j} \boldsymbol{x}_i\right) \\
&\overset{\text{(convexity)}}{\leq} \alpha_1 f(\boldsymbol{x}_1) + \sum_{j=2}^{k+1} \alpha_j f\left(\sum_{i=2}^{k+1} \frac{\alpha_i}{\sum_{j=2}^{k+1} \alpha_j} \boldsymbol{x}_i\right) \\
&\overset{\text{(inductive hypothesis)}}{=} \alpha_1 f(\boldsymbol{x}_1) + \cancel{\sum_{j=2}^{k+1} \alpha_j} \sum_{i=2}^{k+1} \frac{\alpha_i}{\cancel{\sum_{j=2}^{k+1} \alpha_j}} f(\boldsymbol{x}_i) = \sum_{i=1}^{k+1} \alpha_i f(\boldsymbol{x}_i),
\end{aligned}$$

establishing the case $m = k + 1$. Hence the result follows by induction.

(b) "$\Rightarrow$": Assume that f is convex. For all $0 < \alpha < 1$, we have

$$f(\alpha \boldsymbol{y} + (1 - \alpha)\boldsymbol{x}) \leq \alpha f(\boldsymbol{y}) + (1 - \alpha) f(\boldsymbol{x}),$$

which implies that

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

By Proposition 1.1.b with $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{x}$, letting $\alpha \to 0^+$ yields $\nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$, as desired.

"$\Leftarrow$": Assume that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$. Fix any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega$ and any $0 \leq \alpha \leq 1$. Letting $\boldsymbol{x} = \alpha \boldsymbol{x}_1 + (1 - \alpha)\boldsymbol{x}_2$ and $\boldsymbol{y} = \boldsymbol{x}_1, \boldsymbol{x}_2$ alternatively, we get $f(\boldsymbol{x}_1) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x}_1 - \boldsymbol{x})$ and $f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x}_2 - \boldsymbol{x})$. Hence, we have

$$\begin{aligned}
\alpha f(\boldsymbol{x}_1) + (1 - \alpha) f(\boldsymbol{x}_2) &\geq \alpha\big(f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x}_1 - \boldsymbol{x})\big) + (1 - \alpha)\big(f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x}_2 - \boldsymbol{x})\big) \\
&= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \underbrace{(\alpha \boldsymbol{x}_1 + (1 - \alpha)\boldsymbol{x}_2 - \boldsymbol{x})}_{\boldsymbol{0}} = f(\alpha \boldsymbol{x}_1 + (1 - \alpha)\boldsymbol{x}_2).
\end{aligned}$$

(c) Exercise.

(d) "$\Leftarrow$": For all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, by Taylor's theorem we have

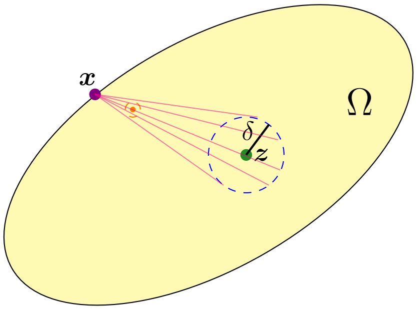
$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})$$

for some $0 < \alpha < 1$. By assumption, $\nabla^2 f(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}))$ is positive semidefinite. Therefore, we have $(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) \geq 0$, which implies that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$. Hence, f is convex by (b).

"$\Rightarrow$": Assume to the contrary that f is convex while $\nabla^2 f(\boldsymbol{x})$ is not positive semidefinite for some $\boldsymbol{x} \in \Omega$, i.e., $\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{d} < 0$ for some $\boldsymbol{d} = (d_1, \dots, d_n) \in \mathbb{R}^n$. Let $g(\boldsymbol{y}) = \boldsymbol{d}^T \nabla^2 f(\boldsymbol{y}) \boldsymbol{d}$ ($\boldsymbol{y}$ is varying while $\boldsymbol{d}$ is fixed). Note that $g(\boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial f}{\partial x_i x_j}(\boldsymbol{y}) d_i d_j$. Since $f \in C^2$, all the second partial derivatives are continuous, and thus g is continuous also.

Claim: We can assume without loss of generality that $\boldsymbol{x}$ is an interior point of Ω.

Proof. Suppose $\boldsymbol{x}$ is not an interior point of Ω (so it is a boundary point). By assumption, there is an interior point $\boldsymbol{z}$ of Ω, which means that there exists sufficiently small $\delta > 0$ such that the ball $B(\boldsymbol{z}, \delta)$, as well as its boundary, is contained in Ω. Connecting $\boldsymbol{x}$ with every point in the boundary of the ball $B(\boldsymbol{z}, \delta)$ forms a "cone", which is contained in Ω due to the convexity of Ω, as illustrated in the picture below:



In the "cone" formed, we can always identify an interior point of Ω that is arbitrarily close to the point $\boldsymbol{x}$. Hence, by the continuity of g, there exists an interior point $\boldsymbol{y}$ of Ω, that is sufficiently close to $\boldsymbol{x}$, such that $g(\boldsymbol{y}) = \boldsymbol{d}^T \nabla^2 f(\boldsymbol{y}) \boldsymbol{d} < 0$. Thus, we can just use the interior point $\boldsymbol{y}$ as our new "$\boldsymbol{x}$" in the subsequent argument. □

Since $\boldsymbol{x}$ is an interior point of Ω and g is continuous, there exists a sufficiently small $r > 0$ such that $B(\boldsymbol{x}, r) \subseteq \Omega$ and $\boldsymbol{d}^T \nabla^2 f(\boldsymbol{y}) \boldsymbol{d} = g(\boldsymbol{y}) < 0$ for all $\boldsymbol{y} \in B(\boldsymbol{x}, r)$. Now, pick a sufficiently large $\lambda > 0$ such that $\boldsymbol{x} + \alpha(\boldsymbol{d}/\lambda) \in B(\boldsymbol{x}, r)$ for all $0 \le \alpha \le 1$. Then, we have

$$\boldsymbol{d}^T f\left(\boldsymbol{x} + \alpha \frac{\boldsymbol{d}}{\lambda}\right) \boldsymbol{d} < 0 \implies \frac{\boldsymbol{d}^T}{\lambda} f\left(\boldsymbol{x} + \alpha \frac{\boldsymbol{d}}{\lambda}\right) \frac{\boldsymbol{d}}{\lambda} < 0.$$

By Taylor's theorem, there exists $0 < \alpha < 1$ such that

$$f\left(\boldsymbol{x} + \frac{\boldsymbol{d}}{\lambda}\right) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \frac{\boldsymbol{d}}{\lambda} + \frac{1}{2} \underbrace{\left(\frac{\boldsymbol{d}}{\lambda}\right)^T \nabla^2 f\left(\boldsymbol{x} + \alpha \frac{\boldsymbol{d}}{\lambda}\right) \frac{\boldsymbol{d}}{\lambda}}_{<0} < f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \frac{\boldsymbol{d}}{\lambda},$$

which implies by (b) that f is not convex, contradiction.

□

[Note: The condition that Ω *contains an interior point* is necessary for (d). To see this, consider $f(x_1, x_2) = x_1 x_2$ with $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = x_2\}$. Then, Ω is a convex set (not containing an interior point) and f is convex on Ω, but

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is not positive semidefinite, for every $\boldsymbol{x} \in \Omega$.]

Corollary 1.2.b. Let $f \in C^2$ be a single variable function on a nonempty interval $\Omega = (a, b)$. Then, f is convex on Ω iff $f''(x) \ge 0$ for all $x \in \Omega$.

Proof. It follows from (d) of Theorem 1.2.a by noting that Ω is a convex set containing an interior point, and the Hessian matrix of f is just $f''(x)$ is this case (so its positive semidefiniteness is equivalent to $f''(x) \ge 0$). □

1.2.4 **A sufficient condition for strict convexity.** Apart from the characterization of strict convexity from Theorem 1.2.a, we can also establish strict convexity via the following sufficient condition:

Theorem 1.2.c. Let $f \in C^2$ be a function on a convex set Ω. If the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ of f is positive definite for all $\boldsymbol{x} \in \Omega$, then f is strictly convex on Ω.

Proof. Fix any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega$ with $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, and $\alpha \in (0,1)$. Let $\boldsymbol{x}_\alpha = \alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2$. By Taylor's theorem (Proposition 1.1.b), for each $i = 1, 2$, there exists $\theta_i \in (0,1)$ such that

$$f(\boldsymbol{x}_i) = f(\boldsymbol{x}_\alpha) + \nabla f(\boldsymbol{x}_\alpha)^T(\boldsymbol{x}_i - \boldsymbol{x}_\alpha) + \underbrace{\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{x}_\alpha)^T \nabla^2 f(\boldsymbol{x}_\alpha + \theta_i(\boldsymbol{x}_i - \boldsymbol{x}_\alpha))(\boldsymbol{x}_i - \boldsymbol{x}_\alpha)}_{>0 \text{ by positive definiteness}} > f(\boldsymbol{x}_\alpha) + \nabla f(\boldsymbol{x}_\alpha)^T(\boldsymbol{x}_i - \boldsymbol{x}_\alpha).$$

Multiplying the inequality for $i = 1$ by α and the inequality for $i = 2$ by $1 - \alpha$ gives

$$\alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2) > f(\boldsymbol{x}_\alpha) + \nabla f(\boldsymbol{x}_\alpha)^T \underbrace{(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 - \boldsymbol{x}_\alpha)}_{\boldsymbol{0}} = f(\boldsymbol{x}_\alpha).$$

□

Corollary 1.2.d. Let $f \in C^2$ be a single variable function on a nonempty interval $\Omega = (a, b)$. If $f''(x) > 0$ for all $x \in \Omega$, then f is strictly convex on Ω.

Proof. Similar to the proof of Corollary 1.2.b. □

[Note: The converse does not hold. To see this, consider $f(x) = x^4$ with $\Omega = \mathbb{R}$. Then, f is strictly convex on Ω but $f''(0) = 0$.]

1.2.5 **Properties of convex functions.** After studying results that help us to examine (strict) convexity, let us turn to analyzing the properties that convex functions enjoy.

Proposition 1.2.e. Let f and g be two convex functions on a convex set Ω. Then:

(a) f is continuous at every interior point of Ω.

(b) $f + g$ is convex on Ω.

(c) cf is convex for all $c \geq 0$.

(d) $\max\{f, g\}$ is convex on Ω.

(e) f^2 is convex on Ω, if $f \geq 0$.

(f) $\Gamma_c := \{\boldsymbol{x} \in \Omega : f(\boldsymbol{x}) \leq c\}$ is a convex set for all $c \in \mathbb{R}$.

Proof.

(a) Omitted.

(b) Fix any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and any $\alpha \in [0, 1]$. Then, we have:

$$\begin{aligned}(f+g)(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) &= f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + g(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \\ &\leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) + \alpha g(\boldsymbol{x}) + (1-\alpha) g(\boldsymbol{y}) \\ &= \alpha (f+g)(\boldsymbol{x}) + (1-\alpha)(f+g)(\boldsymbol{y}).\end{aligned}$$

(c) Fix any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and any $\alpha \in [0, 1]$. Then, we have:

$$(cf)(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) = c \cdot f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \overset{(c \geq 0)}{\leq} c(\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})) = \alpha (cf)(\boldsymbol{x}) + (1-\alpha)(cf)(\boldsymbol{y}).$$

(d) Fix any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and any $\alpha \in [0, 1]$. Then, we have:

$$\begin{aligned}\max\{f, g\}(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) &= \max\{f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}), g(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y})\} \\ &\leq \max\{\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}), \alpha g(\boldsymbol{x}) + (1-\alpha) g(\boldsymbol{y})\} \\ &\leq \max\{\alpha f(\boldsymbol{x}), \alpha g(\boldsymbol{x})\} + \max\{(1-\alpha) f(\boldsymbol{y}), (1-\alpha) g(\boldsymbol{y})\} \\ &= \alpha \max\{f(\boldsymbol{x}), g(\boldsymbol{x})\} + (1-\alpha) \max\{f(\boldsymbol{y}), g(\boldsymbol{y})\} \\ &= \alpha \max\{f, g\}(\boldsymbol{x}) + (1-\alpha) \max\{f, g\}(\boldsymbol{y}).\end{aligned}$$

(e) Fix any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and any $\alpha \in [0,1]$. Define $\phi(x) = x^2$ on $[0,\infty)$. Then, ϕ is increasing and convex. Hence, with $f \geq 0$, we have

$$\begin{aligned} f^2(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) &= \phi\big(f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y})\big) \\ &\overset{(\phi \text{ is increasing})}{\leq} \phi\big(\alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y})\big) \\ &\overset{(\phi \text{ is convex})}{\leq} \alpha\phi(f(\boldsymbol{x})) + (1-\alpha)\phi(f(\boldsymbol{y})) = \alpha f^2(\boldsymbol{x}) + (1-\alpha)f^2(\boldsymbol{y}). \end{aligned}$$

(f) Fix any $\boldsymbol{x}, \boldsymbol{y} \in \Gamma_c$, and any $\alpha \in [0,1]$. By convexity of Ω, we have $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in \Omega$. Also, since $f(\boldsymbol{x}) \leq c$ and $f(\boldsymbol{y}) \leq c$, we have

$$f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y}) \leq \alpha c + (1-\alpha)c \leq c,$$

so $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in \Gamma_c$.

□

1.2.6 **Results about minimizations of convex functions.** Now, let us proceed to discuss some results that relate optimization and convex functions (they are the reasons why we introduce the concept of convexity here!). We shall start with results about minimizations.

Theorem 1.2.f. Let f be a convex function on a convex set Ω. Then:

(a) The set of global minimizers of f over Ω is convex.

(b) *(Important!)* Every local minimizer of f over Ω is also a global minimizer.

(c) If $f \in C^1$ and there exists $\boldsymbol{x}^* \in \Omega$ such that $\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0$ for all $\boldsymbol{x} \in \Omega$, then $\boldsymbol{x}^*$ is a global minimizer of f over Ω.

[Note: Geometrically, (a) suggests that the global minimizers are "located together".]

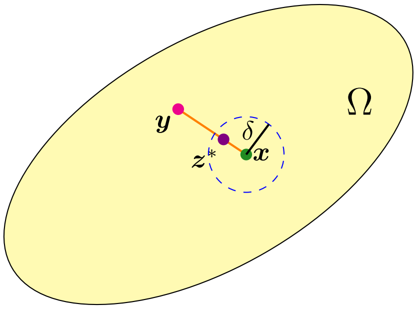
Proof.

(a) Let c be the minimum value of f over Ω. By Proposition 1.2.e, we know that the set $\{\boldsymbol{x} \in \Omega : f(\boldsymbol{x}) \leq c\}$ is convex. But since c is the minimum value, this set is just equal to $\{\boldsymbol{x} \in \Omega : f(\boldsymbol{x}) = c\}$, namely the set of global minimizers of f over Ω. So the result follows.

(b) Assume to the contrary that a local minimizer $\boldsymbol{x}$ is not a global minimizer. Then, there exists $\boldsymbol{y} \in \Omega$ such that $f(\boldsymbol{y}) < f(\boldsymbol{x})$. As $\boldsymbol{x}$ is a local minimizer, there exists an open ball $B(\boldsymbol{x}, \delta)$ centered at $\boldsymbol{x}$ with radius $\delta > 0$ such that $f(\boldsymbol{z}) \geq f(\boldsymbol{x})$ for all $\boldsymbol{z} \in B(\boldsymbol{x}, \delta)$. Note that we can choose some $\alpha \in (0,1)$ such that $\boldsymbol{z}^* := \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in B(\boldsymbol{x}, \delta)$ (as the picture below illustrates). But then by convexity of f, we have

$$f(\boldsymbol{z}^*) = f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y}) \overset{(f(\boldsymbol{y}) < f(\boldsymbol{x}))}{<} f(\boldsymbol{x}),$$

contradiction.



(c) By Theorem 1.2.a, since f is convex, for all $\boldsymbol{x} \in \Omega$ we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \overset{\text{(assumption)}}{\geq} f(\boldsymbol{x}^*).$$

Hence, $\boldsymbol{x}^*$ is a global minimizer of f over Ω by definition.

□

An important implication of Theorem 1.2.f is that for a convex function $f \in C^1$ on a convex set Ω, the first-order necessary conditions (Proposition 1.1.c for general points and Corollary 1.1.d for interior points) are *both necessary and sufficient* for a point to be a global minimizer. More explicitly, we have:

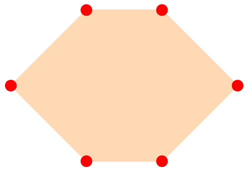
(a) A point $\boldsymbol{x}^*$ is a global minimizer of f over Ω iff $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} \geq 0$ for all feasible directions $\boldsymbol{d}$ at $\boldsymbol{x}^*$.

(b) An interior point $\boldsymbol{x}^*$ of Ω is a global minimizer of f over Ω iff $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

Part (b) justifies why in "nice" cases (with $f \in C^1$ being convex and Ω being open, i.e., every point in Ω is an interior point), we can simply solve the system of equations $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ to determine all global minimizers (which is a rather well-known and commonly used method for solving optimization problems).

1.2.7 **A result about maximizations of convex functions.** Next, we discuss a result about maximizations of convex functions, which involves the following preliminary concept. A point $\boldsymbol{x}$ in a convex set Ω is said to be an **extreme point** (or **vertex**) of Ω if there do not exist two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega$ such that $\boldsymbol{x} = \alpha \boldsymbol{x}_1 + (1 - \alpha)\boldsymbol{x}_2$ for some $0 < \alpha < 1$.

In words, a point $\boldsymbol{x}$ is an extreme point if it is *not* contained in the line segment between any two points that are both different from $\boldsymbol{x}$. For instance, the red points in the convex set illustrated below are extreme points (which coincide with our graphical intuition about vertices).



Proposition 1.2.g. Let f be a convex function on a closed, bounded, and convex set Ω. If there exists a maximizer of f over Ω, then it must be an extreme point of Ω.

Proof. Omitted. □

As a simple application of Proposition 1.2.g, consider a convex function f on $[a, b]$ (with $-\infty < a < b < \infty$), which is convex, closed, and bounded. By Proposition 1.2.e, f is continuous on (a, b). We can then show that f has a maximizer over $[a, b]$, and hence conclude that the maximum value of f over $[a, b]$ is $\max\{f(a), f(b)\}$, since the only candidates of maximizers are a and b.

1.2.8 **Alternative ways for finding global minimizers.** While convexity provides a neat way to find global minimizers, in practice the objective function is *not necessarily* convex. This calls for alternative ways to solve optimization problems where the objective function is non-convex. In general, this kind of optimization problems is quite difficult to solve, and indeed non-convex optimization is still an active field of research. Nevertheless, in some special cases we can still solve non-convex problems in a relatively straightforward manner.

Consider the following minimization problem

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^n, \end{aligned}$$

where f is not necessarily convex. Then we can follow the steps below to try solving this problem, provided that some assumptions are met (specified below):

(1) *(Showing the existence of global minimizer)*

- *Strategy 1: Extreme value theorem.* If f is continuous and Ω is closed and bounded, then by the *extreme value theorem*, a global minimizer exists.
- *Strategy 2: Coercive function.* If $\Omega = \mathbb{R}^n$ and f is **coercive**, i.e., (i) f is continuous on $\mathbb{R}^n$, and (ii) $f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$, then a global minimizer exists.
  Proof. By (ii), we know that there exists $M > 0$ such that $f(\boldsymbol{x}) > f(\boldsymbol{0})$ for all $\boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\| > M$. Now consider the set $A = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq M\}$, which is closed and bounded. Therefore, by the continuity of f and the extreme value theorem, f has a minimum value over A, achieved at some $\boldsymbol{x}^* \in A$.
  For all $\boldsymbol{x} \in \mathbb{R}^n \setminus A$, we have $\|\boldsymbol{x}\| > M$ and thus

$$f(\boldsymbol{x}) > f(\boldsymbol{0}) \overset{(\boldsymbol{0} \in A)}{\geq} f(\boldsymbol{x}^*),$$

  implying that $\boldsymbol{x}^*$ is a global minimizer of f over $\mathbb{R}^n$. □

(2) *(Finding all candidates of local minimizers)* Follow the procedure in [1.2.1] to find out *all* candidates of local minimizers (don't miss the ones in the boundary, in particular ⚠). [Note: This is perhaps easier when $\Omega = \mathbb{R}^n$, i.e., the problem is unconstrained. For constrained problems, things can get more complicated and there are specialized tools to deal with them, to be discussed in Sections 3 and 4.]

(3) *(Comparing the objective function values of the candidates, if feasible)* If comparison between the objective function values of the candidates is feasible (e.g., there are only finitely many candidates), then the one with the *smallest* objective function value is the global minimizer. Otherwise, we need to consider other approaches for solving this problem.

1.3 Properties of Algorithms

1.3.1 As mentioned in [1.0.1], the main goal of MATH3904 is to develop efficient algorithms for solving optimization problems. The algorithms discussed in MATH3904 take the following general form. A sequence $\{\boldsymbol{x}_k\}_{k \geq 0}$ of points is generated, where $\boldsymbol{x}_0$ is the starting point ("initial guess"), and each later point $\boldsymbol{x}_{k+1}$ is computed based on the previously obtained points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k$ according to a certain principle such that

$$\lim_{k \to \infty} f(\boldsymbol{x}_k) = \text{the optimal value of the optimization problem.}$$

While the principle that drives the generation of the sequence $\{\boldsymbol{x}_k\}$ can differ substantially for these algorithms, they can indeed all be seen as *iterative descent* algorithms:

- *Iterative:* The algorithm generates a series of points with each point being computed based on the points preceding it.
- *Descent:* As each new point is generated by the algorithm, the corresponding value of *some* function (not necessarily the objective function!) decreases.

1.3.2 **Correctness.** In the analysis of these iterative descent algorithms, there are two key aspects to consider: correctness and efficiency. For correctness, it means that the sequence $\{\boldsymbol{x}_k\}$ converges to an optimal solution. We call an algorithm **globally convergent** if for every starting point $\boldsymbol{x}_0$, the algorithm is guaranteed to generate a sequence of points that converges to an optimal solution.

1.3.3 **Efficiency.** To describe the efficiency of an algorithm, the notion of *order of convergence* is needed.

Let $\{\boldsymbol{x}_k\}$ be a sequence converging to $\boldsymbol{x}^*$. The **order of convergence** (or **rate of convergence** or **speed of convergence**) of $\{\boldsymbol{x}_k\}$ is the *supremum* of nonnegative values p satisfying that

$$\limsup_{k \to \infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p} = \beta < \infty. \tag{3}$$

The sequence $\{\boldsymbol{x}_k\}$ is said to **converge with order i** if (3) holds for $p = i$. It is also said to have **linear convergence** if (3) holds for $p = 1$ and $\beta < 1$, and **superlinear convergence** if (3) holds for $p > 1$, or for $p = 1$ and $\beta = 0$.

Remarks:

- *(Limit superior vs. limit)* Here we consider limit superior "$\limsup$" rather than limit "$\lim$" since the former always exist while the latter may not exist. But if the limit exists, then the limit superior is always equal to that.
- If a sequence $\{\boldsymbol{x}_k\}$ converges with order p, then it also converges with order p' for all $0 \le p' < p$, because for all $0 \le p' < p$ we have

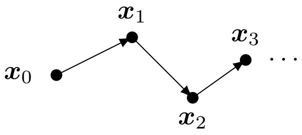
$$\begin{aligned}\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p'}} &= \limsup_{k\to\infty}\left(\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p}} \cdot \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p-p'}\right) \\ &\le \left(\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p}}\right)\left(\limsup_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p-p'}\right) \\ &= \beta \cdot 0 = 0 < \infty.\end{aligned}$$

  This explains why we are interested in the supremum of such values of p (the "largest one", loosely). Note that if the order of convergence of a sequence is p^*, then we know that the sequence converges with order p for all $0 \le p < p^*$ (and possibly also for $p = p^*$).

2 Basic Optimization Algorithms

2.0.1 To illustrate the general idea underlying algorithms for solving optimization algorithms mentioned in [1.3.1] more concretely, in Section 2 we will discuss several basic algorithms for solving unconstrained optimization problems. Since they often serve as simple and direct methods for obtaining optimal solutions numerically, they are of high practical importance.

2.0.2 **Structure of descent methods.** Descent methods work in the following way. We start at an initial point, determine a movement direction based on *a certain principle*, and then move in this direction to a *local minimizer* of the objective function *on that line*. At the new point, a new movement direction is determined and the process is repeated.



Descent methods are all in this form, and their differences lie in the *principle* by which successive movement directions are chosen.

2.0.3 **Line search.** The process of determining a local minimizer on a given line is called **line search**, or **one-dimensional search** (since it is related to one-dimensional minimization problems). As high-dimensional minimization problems are often solved (numerically) by executing a sequence of successive line searches, let us begin with analyzing some methods for line search.

2.1 Line Search Algorithms

2.1.1 **Setup.** Consider the line search problem

$$\begin{array}{ll} \text{Minimize} & f(x) \\ \text{subject to} & x \in [a,b] \end{array}$$

where the interval $[a,b]$ may be identified as a portion of the line that contains a local minimizer (so we can focus on minimizing the objective function f over this interval).

Assume that there exists a minimizer of f over $[a,b]$. While we do not know its exact location at this stage, we know that it must lie in the interval $[a,b]$ (trivially). This interval is an example of *interval of uncertainty*. In general, an interval I is called an **interval of uncertainty** if it contains a minimizer of f.

The key idea for algorithms solving this kind of line search problem is to shrink the length of the interval of uncertainty, so that the "uncertainty" about the minimizer can be reduced and we can deduce its location more precisely.

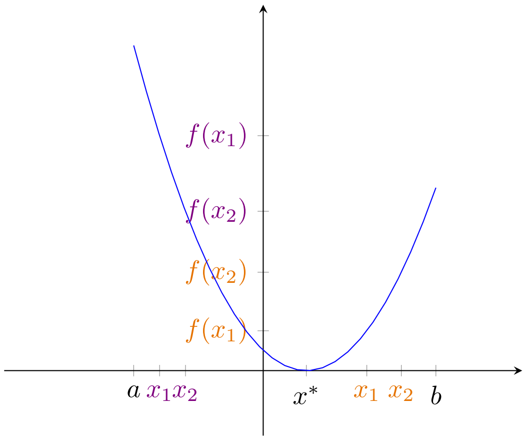
In practice, we often impose a limit $\ell > 0$ on the length of the interval of uncertainty when designing an algorithm: If the length of a certain interval of uncertainty obtained in the iteration is not greater than ℓ, then we terminate the algorithm and output any point in the interval (e.g., its mid-point) as an estimated minimizer. The limit ℓ can be seen as the *allowable final length of uncertainty*, or the *tolerance*, which is the maximum distance between the estimated minimizer and the true minimizer allowed.

2.1.2 **Unimodality.** Let us first discuss line search algorithms *without using derivatives* (zeroth order), which are applied to *unimodal* functions. A function f on $[a,b]$ is said to be **unimodal** if there exists $x^* \in [a,b]$ such that:

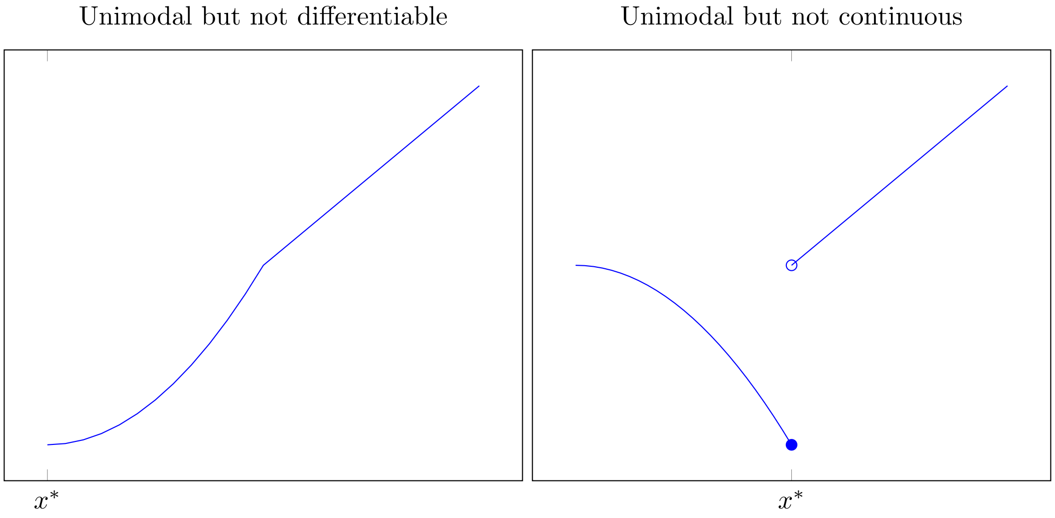
(a) x^* minimizes f over $[a,b]$.

(b) For all $x_1, x_2 \in [a,b]$ with $x_1 < x_2$, we have:

i. $x_2 \leq x^* \implies f(x_1) > f(x_2)$.
ii. $x^* \leq x_1 \implies f(x_2) > f(x_1)$.



While it is instructive to think of a unimodal function as a function being U-shaped like above, it is not necessarily the case. Indeed, unimodal functions are not necessarily continuous or differentiable, as the following pictures illustrate:



2.1.3 **A sufficient condition for unimodality.** Although the concept of unimodality plays an important role for line search algorithms without using derivatives, we can see that it is generally rather complicated to show that a given function is unimodal by definition directly. Instead of doing so, we can use the following result, which provides a convenient sufficient condition for unimodality.

Proposition 2.1.a. Let f be a continuous and strictly convex function on $[a, b]$. Then, f is unimodal.

Proof. Since f is continuous on a closed and bounded interval $[a, b]$, f attains minimum at some point $x^* \in [a, b]$ by extreme value theorem. This verifies the condition (a) for unimodality.

Claim: x^* is the unique minimizer of f over $[a, b]$.

Proof. Assume to the contrary that $f(y^*) = f(x^*)$ for some $y^* \neq x^*$. Let z be a point strictly between x^* and y^*. Then, we can write $y = \alpha x^* + (1-\alpha)y^*$ for some $\alpha \in (0, 1)$. By the strict convexity, we then have $f(z) < \alpha f(x^*) + (1-\alpha)f(y^*) = \alpha f(x^*) + (1-\alpha)f(x^*) = f(x^*)$, contradicting to the fact that x^* minimizes f. □

Now we are ready to justify the condition (b) for unimodality. Fix any $x_1, x_2 \in [a,b]$ with $x_1 < x_2$. Consider the case with $x_2 \leq x^*$. By the claim above, we have $f(x_1) > f(x^*)$. So if $x_2 = x^*$, then we readily have $f(x_1) > f(x_2)$, as desired. Thus, we assume henceforth that $x_2 \neq x^*$. With $x_1 < x_2 < x^*$, we can write $x_2 = \lambda x_1 + (1-\lambda)x^*$ for some $\lambda \in (0,1)$. It then follows by strict convexity that $f(x_2) < \lambda f(x_1) + (1-\lambda)f(x^*) \leq \lambda f(x_1) + (1-\lambda)f(x_1) = f(x_1)$, as desired. The argument is similar for the other case with $x^* \leq x_1$. □

2.1.4 **Properties of unimodal functions.** To develop the line search algorithms without using derivatives for unimodal functions, the following properties are important.

Theorem 2.1.b. Let f be a unimodal function on $[a,b]$, x^* be a minimizer of f over $[a,b]$, and λ, μ be two points in $[a,b]$ with $\lambda < \mu$. Then:

(a) x^* is the unique minimizer of f over $[a,b]$.

(b) If $f(\lambda) \leq f(\mu)$, then $x^* \in [a,\mu]$.

(c) If $f(\lambda) \geq f(\mu)$, then $x^* \in [\lambda,b]$.

(d) If $f(\lambda) = f(\mu)$, then $x^* \in [\lambda,\mu]$.

Proof.

(a) Assume to the contrary that $f(y^*) = f(x^*)$ for some $y^* \neq x^*$. Without loss of generality, assume that $x^* < y^*$. As y^* is a minimizer, there exists $z \in (x^*, y^*)$ such that $f(z) \geq f(y^*)$. Now take $x_1 = z$ and $x_2 = y^*$. Then, we have $x^* < x_1$, and thus $f(x_2) > f(x_1)$ by the definition of unimodality. This means that $f(y^*) > f(z)$, contradiction.

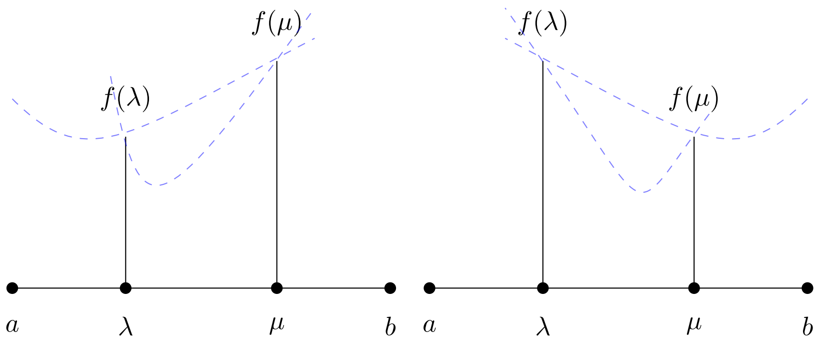
(b) Take $x_1 = \lambda$ and $x_2 = \mu$. If we had $x^* \in (\mu, b]$ (so $x^* \leq x_2$), then by the definition of unimodality we would have $f(x_1) > f(x_2)$, i.e., $f(\lambda) > f(\mu)$, contradiction.

(c) Take $x_1 = \lambda$ and $x_2 = \mu$. If we had $x^* \in [a, \lambda)$ (so $x_1 \leq x^*$) , then by the definition of unimodality we would have $f(x_1) < f(x_2)$, i.e., $f(\lambda) < f(\mu)$, contradiction.

(d) By (b) and (c), $f(\lambda) = f(\mu)$ implies that $x^* \in [a,\mu] \cap [\lambda, b] = [\lambda,\mu]$, as desired.

□

[Note: We can indeed replace the closed intervals $[a,\mu]$, $[\lambda,b]$, and $[\lambda,\mu]$ by $[a,\mu)$, $(\lambda,b]$, and (λ,μ), respectively. But for the purpose of developing the upcoming line search algorithms, those closed intervals are enough.]



2.1.5 **Golden section method.** The first line search algorithm for unimodal functions is **golden section method**, which is given below.

(1) *(Initialization)* Choose a tolerance $\ell > 0$. Set $a_1 = a$, $b_1 = b$, $\lambda_1 = a_1 + (1-\alpha)(b_1 - a_1)$, $\mu_1 = a_1 + \alpha(b_1 - a_1)$, and $k = 1$, where $\alpha := (\sqrt{5}-1)/2 = 0.618$ is the **golden section ratio**.

(2) *(Checking for termination)* If $b_k - a_k \leq \ell$, stop and output any point in $[a_k, b_k]$ as an estimated minimizer.

(3) *(Updating the interval of uncertainty)*

i. If $f(\lambda_k) > f(\mu_k)$, set $a_{k+1} = \lambda_k$, $b_{k+1} = b_k$, $\lambda_{k+1} = \mu_k$, and $\mu_{k+1} = a_{k+1} + \alpha(b_{k+1} - a_{k+1})$.

ii. If $f(\lambda_k) \le f(\mu_k)$, set $a_{k+1} = a_k$, $b_{k+1} = \mu_k$, $\mu_{k+1} = \lambda_k$, and $\lambda_{k+1} = a_{k+1}+(1-\alpha)(b_{k+1}-a_{k+1})$.

(4) *(Repetition)* Replace k by $k+1$ and go to (2).

2.1.6 **Idea behind the golden section method.** The idea behind the development of the golden section method is as follows. Consider any iteration k, where $[a_k, b_k]$ is the starting interval of uncertainty. Suppose that $b_k - a_k > \ell$ so that the algorithm is not yet terminated. The values λ_k and μ_k are two points in $[a_k, b_k]$, with $\lambda_k < \mu_k$, at which functional evaluations are to be made. In view of Theorem 2.1.b, we can update the interval of uncertainty to $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$ if $f(\lambda_k) > f(\mu_k)$, and $[a_{k+1}, b_{k+1}] = [a_k, \mu_k]$ otherwise.

The key characteristic of golden section method is the principles under which the values of λ_k and μ_k are chosen:

(1) For each iteration k, the reduction ratio $(b_{k+1} - a_{k+1})/(b_k - a_k)$ is always a fixed constant c; the length of the new interval of uncertainty $b_{k+1} - a_{k+1}$ does not depend on the outcome of the kth iteration, i.e., whether $f(\lambda_k) > f(\mu_k)$ or $f(\lambda_k) \le f(\mu_k)$.

(2) As λ_{k+1} and μ_{k+1} are selected for a new iteration, either λ_{k+1} coincides with μ_k or μ_{k+1} coincides with λ_k (so that only *one* new functional evaluation is needed, improving the efficiency of the algorithm).

Based on these two principles, we can derive the way we choose λ_k and μ_k above. By principle (1), we have

$$b_k - \lambda_k = \mu_k - a_k = c(b_k - a_k),$$

so that the length of the new interval $[a_{k+1}, b_{k+1}]$ is always $c(b_k - a_k)$, regardless of the outcome of the kth iteration.

It follows that

$$\lambda_k = a_k + (1-c)(b_k - a_k) \quad \text{and} \quad \mu_k = a_k + c(b_k - a_k).$$

Now, we apply principle (2). Consider first the case with $f(\lambda_k) > f(\mu_k)$. Then, we have $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$. Since λ_k is the left end-point, it is impossible to have $\mu_{k+1} = \lambda_k$, which necessitates that $\lambda_{k+1} = \mu_k$. Hence, we get

$$\begin{aligned}
\mu_k = \lambda_{k+1} &\overset{\text{(above)}}{=} a_{k+1} + (1-c)(b_{k+1} - a_{k+1}) = \lambda_k + (1-c)(b_k - \lambda_k) \\
\implies \quad \mu_k - \lambda_k &= (1-c)(b_k - \lambda_k) \\
\implies \quad (2c-1)(b_k - a_k) &= (1-c)(b_k - a_k - (1-c)(b_k - a_k)) \\
\implies \quad 2c - 1 &= c - c^2 \\
\implies \quad c^2 + c - 1 &= 0,
\end{aligned}$$

which implies that $c = \alpha = (\sqrt{5} - 1)/2 = 0.618$, the golden section ratio, since $c = (b_{k+1} - a_{k+1})/(b_k - a_k) \in [0, 1]$. The idea is similar for the other case with $f(\lambda_k) \le f(\mu_k)$, and the resulting c is also the golden section ratio.

2.1.7 **Properties of the golden section method.**

(a) The reduction ratio $(b_{k+1} - a_{k+1})/(b_k - a_k)$ is always the golden section ratio $\alpha = 0.618$.

(b) The number of iterations needed for golden section method is the smallest positive integer k such that $(0.618)^k \le \ell/(b_1 - a_1)$.

Proof.

(a) It follows from the principle (2) in [2.1.5].

(b) This is because the length of the interval of uncertainty after k iterations is $b_{k+1} - a_{k+1} = \alpha^k(b_1 - a_1) = 0.618^k(b_1 - a_1)$. So, the golden section method terminates after k iterations iff $b_{k+1} - a_{k+1} \leq \ell$ iff $(0.618)^k \leq \ell/(b_1 - a_1)$.

□

2.1.8 **Fibonacci search method.** The second line search algorithm for unimodal functions presented here is *Fibonacci search method*, which makes two functional evaluations at the first iteration and then only one evaluation at each subsequent iteration, like the golden section method. However, the reduction ratio of the Fibonacci search method varies from one iteration to another, unlike the golden section method. The **Fibonacci search method** is given below.

Let $\{F_n\}_{n\geq 0}$ be the **Fibonacci sequence**, defined recursively as follows:

$$F_0 = F_1 = 1 \quad \text{and} \quad F_{n+1} = F_n + F_{n-1} \text{ for all } n \in \mathbb{N};$$

the first few terms of this sequence are $1, 1, 2, 3, 5$.

(1) *(Initialization)* Choose a tolerance $\ell > 0$ and a distinguishability constant $0 < \varepsilon \ll \ell$. Let n be the smallest integer $m \geq 3$ such that $F_m \geq (b-a)/\ell$. Set $a_1 = a$, $b_1 = b$, $\lambda_1 = a_1 + (F_{n-2}/F_n)(b_1 - a_1)$, $\mu_1 = a_1 + (F_{n-1}/F_n)(b_1 - a_1)$, and $k = 1$.

(2) *(Updating the interval of uncertainty)*

i. If $f(\lambda_k) > f(\mu_k)$, set $a_{k+1} = \lambda_k$, $b_{k+1} = b_k$, $\lambda_{k+1} = \mu_k$, and $\mu_{k+1} = a_{k+1} + (F_{n-k-1}/F_{n-k})(b_{k+1} - a_{k+1})$.

ii. If $f(\lambda_k) \leq f(\mu_k)$, set $a_{k+1} = a_k$, $b_{k+1} = \mu_k$, $\mu_{k+1} = \lambda_k$, and $\lambda_{k+1} = a_{k+1} + (F_{n-k-2}/F_{n-k})(b_{k+1} - a_{k+1})$.

(3) *(Repetition)* If $k = n - 2$, go to (4). Otherwise, replace k by $k + 1$ and go to (2).

(4) *(Last iteration)* Set $\lambda_n = \lambda_{n-1}$ and $\mu_n = \lambda_{n-1} + \varepsilon$.

i. If $f(\lambda_n) > f(\mu_n)$, set $a_n = \lambda_n$ and $b_n = b_{n-1}$.

ii. If $f(\lambda_n) \leq f(\mu_n)$, set $a_n = a_{n-1}$ and $b_n = \lambda_n$.

Stop and output the mid-point of $[a_n, b_n]$ as an estimated minimizer.

2.1.9 **Idea behind the golden section method.** The idea behind the development of the Fibonacci search method is as follows. Recall from the golden section method that $\lambda_k = a_k + (1 - \alpha)(b_k - a_k)$ and $\mu_k = a_k + \alpha(b_k - a_k)$ for all $k \geq 1$. Instead of using the pair of constants $1 - \alpha$ and α (which sum to 1), the Fibonacci search method uses the pair F_{n-k-1}/F_{n-k+1} and F_{n-k}/F_{n-k+1} (which still sum to 1), for all $k = 1, \ldots, n - 1$:

$$\lambda_k = a_k + \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) \quad \text{and} \quad \mu_k = a_k + \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k).$$

Now consider any iteration $k \in \{1, \ldots, n - 2\}$ where the starting interval of uncertainty is $[a_k, b_k]$. Like before, we set $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$ if $f(\lambda_k) > f(\mu_k)$, and $[a_{k+1}, b_{k+1}] = [a_k, \mu_k]$ if $f(\lambda_k) \leq f(\mu_k)$.

[Note: For all $k = 1, \ldots, n - 2$, we have $F_{n-k-1} < F_{n-k}$, and thus $\lambda_k < \mu_k$.]

Let us consider the reduction ratio $(b_{k+1} - a_{k+1})/(b_k - a_k)$. As mentioned above, it is no longer fixed for the Fibonacci search method; it is indeed related to some terms in the Fibonacci sequence:

Claim:

$$b_{k+1} - a_{k+1} = \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k).$$

Proof. For the case with $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$, we have

$$b_{k+1} - a_{k+1} = (b_k - a_k) - \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) = \frac{F_{n-k+1} - F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) = \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k).$$

It is similar for the case with $[a_{k+1}, b_{k+1}] = [a_k, \mu_k]$. □

Apart from the reduction ratio, we are also interested in the number of functional evaluations needed at each iteration. As mentioned above, only one functional evaluation is needed at each iteration except the first, like the golden search method. This is justified below:

Claim: If $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$, then $\lambda_{k+1} = \mu_k$. If $[a_{k+1}, b_{k+1}] = [a_k, \mu_k]$, then $\mu_{k+1} = \lambda_k$.

Proof. Consider the case with $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$. Then, we have

$$\begin{aligned}
\lambda_{k+1} &= a_{k+1} + \frac{F_{n-k-2}}{F_{n-k}}(b_{k+1} - a_{k+1}) \\
&= \lambda_k + \frac{F_{n-k-2}}{F_{n-k}}(b_k - \lambda_k). \\
&= a_k + \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) + \frac{F_{n-k-2}}{F_{n-k}}\left(b_k - a_k - \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k)\right) \\
&= a_k + \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) + \frac{F_{n-k-2}}{\cancel{F_{n-k}}}\left(\frac{\cancel{F_{n-k+1} - F_{n-k-1}}}{F_{n-k+1}}\right)(b_k - a_k) \\
&= a_k + \frac{F_{n-k-1} + F_{n-k-2}}{F_{n-k+1}}(b_k - a_k) \\
&= a_k + \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k) = \mu_k.
\end{aligned}$$

The argument is similar for the case with $[a_{k+1}, b_{k+1}] = [a_k, \mu_k]$. □

So far we have considered the iterations $1, \ldots, n-2$. However, the Fibonacci search method described above actually has $n-1$ iterations, and the last iteration to get from $[a_{n-1}, b_{n-1}]$ to $[a_n, b_n]$ is not yet considered. This iteration actually has some subtleties and deserves careful analysis.

For $k = n-1$, from the above formulas of λ_k and μ_k we have

$$\lambda_{n-1} = \mu_{n-1} = a_{n-1} + \frac{1}{2}(b_{n-1} - a_{n-1}) = \frac{1}{2}(a_{n-1} + b_{n-1})$$

since $F_0 = F_1 = 1$ and $F_2 = 2$. This means that both λ_{n-1} and μ_{n-1} are the mid-point of the interval $[a_{n-1}, b_{n-1}]$. Thus, we are unable to apply Theorem 2.1.b on λ_{n-1} and μ_{n-1} right away. To proceed further, we retain the "λ" and choose a "μ" that is slightly to the right of the mid-point $\lambda_{n-1} = \mu_{n-1}$. More precisely, we set $\lambda_n = \lambda_{n-1}$ and $\mu_n = \lambda_{n-1} + \varepsilon$, and apply Theorem 2.1.b on this pair of values: setting $[a_n, b_n] = [\lambda_n, b_{n-1}]$ if $f(\lambda_n) > f(\mu_n)$, and $[a_n, b_n] = [a_{n-1}, \mu_n]$ if $f(\lambda_n) \le f(\mu_n)$.

Since the length of the interval $[a_{n-1}, b_{n-1}]$ is

$$b_{n-1} - a_{n-1} = \frac{F_2}{F_3}\frac{F_3}{F_4}\cdots\frac{F_{n-1}}{F_n}(b_1 - a_1) = \frac{2}{F_n}(b_1 - a_1),$$

and $F_n \ge (b_1 - a_1)/\ell$, we know that the length of the interval $[a_n, b_n]$ is $(b_1 - a_1)/F_n \le \ell$ if $f(\lambda_n) > f(\mu_n)$), and $(b_1 - a_1)/F_n + \varepsilon \le \ell + \varepsilon$ if $f(\lambda_n) \le f(\mu_n)$. Due to the appearance of $\ell + \varepsilon$, we cannot be sure that the precision requirement from the tolerance ℓ is *always* satisfied if we just output any point in the interval $[a_n, b_n]$ as an estimated minimizer. To ensure that the precision requirement is met, we can output the *mid-point* of $[a_n, b_n]$ as an estimated minimizer, so that even for the case with $f(\lambda_n) \le f(\mu_n)$, the distance between the estimated minimizer and the true minimizer is at most

$$\frac{1}{2}\left(\frac{b_1 - a_1}{F_n} + \varepsilon\right) \le \frac{\ell + \varepsilon}{2} \le \ell$$

provided that ε is sufficiently small relative to ℓ ($\varepsilon \ll \ell$).

2.1.10 **Newton's method.** After discussing line search methods without using derivatives, let us proceed to some methods that use derivatives, applicable for functions possessing some (continuous) derivatives. The first method to be discussed is **Newton's method**, which works as follows.

Suppose that the objective function f to be minimized is convex and twice differentiable.

(1) *(Initialization)* Choose a termination scalar $\varepsilon > 0$ and an initial guess x_0 of minimizer (starting point). Set $k = 0$.

(2) *(Quadratic approximation)* Set $x_{k+1} = x_k - f'(x_k)/f''(x_k)$, assuming that $f''(x_k) \neq 0$.

(3) *(Checking for termination)* If $|x_{k+1} - x_k| < \varepsilon$ or $|f'(x_{k+1})| < \varepsilon$, stop and output x_{k+1} as an estimated minimizer. Otherwise, replace k by $k+1$ and go to (2).

2.1.11 **Idea behind the Newton's method.** At each iteration with a point x_k, Newton's method exploits the *quadratic approximation* of the function f at x_k:

$$q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2,$$

which agrees with f at x_k up to the second-order derivative, i.e., $f(x_k) = q(x_k)$, $f'(x_k) = q'(x_k)$, and $f''(x_k) = q''(x_k)$. Then, the next point x_{k+1} is chosen to be the solution to $q'(x) = 0$, i.e., the stationary point of the quadratic approximation $q(x)$, which is also its minimizer as q is convex (we have $q''(x) = f''(x_k) \geq 0$ since f is convex). Since $q'(x) = f'(x_k) + f''(x_k)(x - x_k)$, we have

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$

The terminating condition $|x_{k+1} - x_k| < \varepsilon$ suggests that the estimated minimizer x_{k+1} should be already rather close to the true minimizer, provided that $\{x_k\}$ converges to it. Another condition $|f'(x_{k+1})| < \varepsilon$ suggests that the point x_{k+1} is already "nearly" a stationary point (hence a minimizer) of f. In either case, it seems plausible to terminate the algorithm.

2.1.12 **Another perspective of the Newton's method.** Since determining a minimizer of the objective function f is equivalent to finding a root of f', we can actually view Newton's method as a *root-finding* technique, for iteratively solving equations of the form $g(x) = 0$; in the context above, we have $g(x) = f'(x)$. Then, we can rephrase the **Newton's method** as follows:

(1) *(Initialization)* Choose a termination scalar $\varepsilon > 0$ and an initial guess x_0 of minimizer (starting point). Set $k = 0$.

(2) *(Moving to the next point)* Set $x_{k+1} = x_k - g(x_k)/g'(x_k)$, assuming that $g'(x_k) \neq 0$.

(3) *(Checking for termination)* If $|x_{k+1} - x_k| < \varepsilon$ or $|g(x_{k+1})| < \varepsilon$, stop and output x_{k+1} as an estimated root. Otherwise, replace k by $k+1$ and go to (2).

2.1.13 **Property of the Newton's method.**

Proposition 2.1.c. Let $g \in C^2$ and let x^* be the root of g with $g'(x^*) \neq 0$. Then, the sequence $\{x_k\}$ generated by the iterative formula $x_{k+1} = x_k - g(x_k)/g'(x_k)$ of the Newton's method converges to x^* with an order of convergence at least two, provided that x_0 is sufficiently close to x^*.

Proof. With $g \in C^2$, we know that g' and g'' are continuous. Therefore, for all x sufficiently close to x^*, we have $|g''(x)| < c_1$ (by extreme value theorem) and $|g'(x)| > c_2$ (by the definition of continuity, as $g'(x^*) \neq 0$) for some $c_1, c_2 \in \mathbb{R}$. Hence, we have

$$x_{k+1} - x^* \overset{(g(x^*)=0)}{=} x_k - x^* - \frac{g(x_k) - g(x^*)}{g'(x_k)} = \frac{1}{g'(x_k)}(g(x^*) - g(x_k) - g'(x_k)(x^* - x_k)).$$

By Taylor's theorem, we have $g(x^*) - g(x_k) - g'(x_k)(x^* - x_k) = (1/2)g''(\xi_k)(x^* - x_k)^2$ for some ξ_k between x^* and x_k. It follows that

$$|x_{k+1} - x^*| = \frac{1}{2}|g''(\xi_k)|(x_k - x^*)^2/|g'(x_k)| \leq \frac{c_1}{2c_2}|x_k - x^*|^2$$

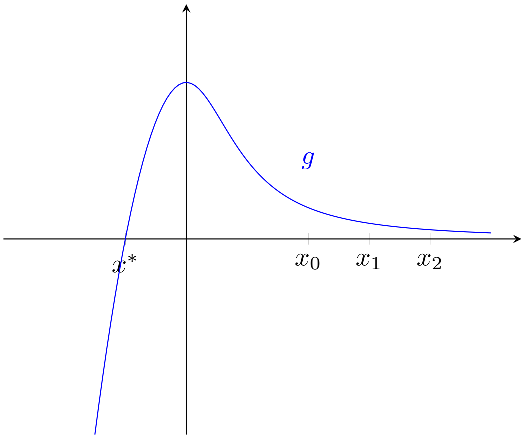
whenever x_k and ξ_k are sufficiently close to x^*.

Provided that x_0 is sufficiently close to x^* such that $|g''(x_0)| < c_1$, $|g'(x_0)| > c_2$, and $c := c_1/(2c_2)|x_0 - x^*| < 1$, we can ensure that x_k and ξ_k are sufficiently close to x^*, and $|x_k - x^*| \leq c^k|x_0 - x^*|$ for all $k \geq 0$, by induction. This implies that $\{x_k\}$ converges to x^*. Since

$$\limsup_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} \leq \frac{c_1}{2c_2} < \infty,$$

we conclude that Newton's method converges to x^* with an order at least two. □

An important assumption in Proposition 2.1.c is that x_0 *is sufficiently close to* x^*. If the initial guess x_0 is too far from x^*, then Newton's method may fail to converge, meaning that it is not *globally* convergent. To see this more concretely, consider the following picture:



With such an initial guess x_0, the sequence $\{x_k\}$ generated by the iterative formula of the Newton's method does not converge to x^*.

2.2 Steepest Descent Method

2.2.1 An important type of descent method is the *steepest descent method*, which is one of the most fundamental algorithms for minimizing a differentiable multivariate function (without constraints). Many other algorithms for minimization are developed by modifying this method for achieving more superior convergence properties.

2.2.2 **Direction of descent.** To describe the steepest descent method, we need to use the notion of *direction of descent*. A vector $\boldsymbol{d}$ is called a **direction of descent** of a function f at $\boldsymbol{x}$ if there exists $\delta > 0$ such that $f(\boldsymbol{x} + \alpha\boldsymbol{d}) < f(\boldsymbol{x})$ for all $0 < \alpha < \delta$. Intuitively, this means that moving along the direction $\boldsymbol{d}$ slightly yields reductions in the values taken by f ("descent").

2.2.3 **Sufficient condition for direction of descent.** If

$$f'(\boldsymbol{x}, \boldsymbol{d}) := \lim_{\alpha\to 0^+} \frac{f(\boldsymbol{x} + \alpha\boldsymbol{d}) - f(\boldsymbol{x})}{\alpha} < 0,$$

then $\boldsymbol{d}$ is a direction of descent at $\boldsymbol{x}$.

Proof. It follows from the definition of limit. □

2.2.4 **Direction of steepest descent.** The key idea of *steepest descent method* is to move along the direction $\boldsymbol{d}$ that *minimizes* $f'(\boldsymbol{x}, \boldsymbol{d})$ with $\|\boldsymbol{d}\| = 1$ (normalization), which is the **direction of steepest descent**. Here, we can view $f'(\boldsymbol{x}, \boldsymbol{d})$ as a measure of "steepness of descent" of a direction $\boldsymbol{d}$; a more negative value indicates a "steeper" descent.

The following result gives us the formula of the direction of steepest descent.

Proposition 2.2.a. Let f be differentiable at $\boldsymbol{x}$ with $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$. The optimal solution to the problem

$$\begin{aligned} \text{Minimize} \quad & f'(\boldsymbol{x}, \boldsymbol{d}) \\ \text{subject to} \quad & \|\boldsymbol{d}\| = 1, \end{aligned}$$

is given by $\boldsymbol{d}^* = -\nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|$.

Proof. Using a similar proof as Proposition 1.1.b, from the differentiability of f at $\boldsymbol{x}$ we have $f'(\boldsymbol{x}, \boldsymbol{d}) = \nabla f(\boldsymbol{x})^T \boldsymbol{d}$. Then, by the Cauchy-Schwartz inequality, we have

$$\nabla f(\boldsymbol{x})^T \boldsymbol{d} \geq -\|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{d}\| = -\|\nabla f(\boldsymbol{x})\|,$$

where the equality holds iff $\boldsymbol{d} = -\nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|$, as desired. □

2.2.5 **Steepest descent method.** In view of Proposition 2.2.a, we know that the movement *direction* at a point $\boldsymbol{x}_k$ for the steepest descent method should be $-\nabla f(\boldsymbol{x}_k) / \|\nabla f(\boldsymbol{x}_k)\|$, or equivalently, $-\nabla f(\boldsymbol{x})$. To determine *how far* we should move along that direction, we perform line search. More precisely, the **steepest descent method** works as follows. Suppose that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

(1) *(Initialization)* Choose a termination scalar $\varepsilon > 0$ and a starting point $\boldsymbol{x}_0$. Set $k = 0$.

(2) *(Checking for termination)* If $\|\nabla f(\boldsymbol{x}_k)\| < \varepsilon$, stop and output $\boldsymbol{x}_k$ as an estimated minimizer.

(3) *(Moving along the direction of steepest descent)* Set $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$, α_k be an optimal solution to the line search problem

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \\ \text{subject to} \quad & \alpha \geq 0 \end{aligned}$$

(assuming an optimal solution exists), and $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$. Replace k by $k + 1$ and go to (2).

[⚠ Warning: While the steepest descent method often works well during early stages of the optimization process, as a stationary point is approached, it usually takes smaller steps and has a poorer performance.]

2.2.6 **Basic properties of the steepest descent method.**

(a) We always have $\alpha_k > 0$ in (3).

(b) We always have $\boldsymbol{d}_k^T \boldsymbol{d}_{k+1} = 0$, i.e., any two consecutive movement directions are orthogonal.

Proof. Let $g(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$.

(a) By the differentiability of f we have

$$g'(\alpha) = \nabla f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)^T \boldsymbol{d}_k$$

(using a similar proof as Proposition 1.1.b). Hence, $g'(0) = \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)^T \nabla f(\boldsymbol{x}_k) = -\|\nabla f(\boldsymbol{x}_k)\|^2 < 0$, where the last inequality follows from noting that arriving at step (3) implies that $\|\nabla f(\boldsymbol{x}_k\| \geq \varepsilon > 0$ *(not terminated)*. So, 0 is not a stationary point of g and thus cannot possibly be a minimizer of g. Therefore, $\alpha_k > 0$.

(b) Since α_k is a minimizer of g, we have $g'(\alpha_k) = 0$. It follows that

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k = 0 \implies \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_{k+1}) = 0 \implies \boldsymbol{d}_k^T \boldsymbol{d}_{k+1} = 0.$$

□

2.2.7 **The quadratic case.** To analyze the convergence properties of the steepest descent method (or other optimization algorithms), we usually first consider the *quadratic case*, since properties derived for the quadratic case can often be translated into a similar one for nonquadratic case, and thus investigations of the method applied to quadratic problems yield important insights about its convergence behaviour.

Here, we focus on the following (unconstrained) quadratic problem

$$\text{Minimize} \quad f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$$

where Q is a *positive definite* $n \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^n$ is a vector.

2.2.8 **Preliminary results about the quadratic problem.** For the quadratic problem in [2.2.7], we have:

(a) $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} - \boldsymbol{b}$ and $\nabla^2 f(\boldsymbol{x}) = Q$.

(b) f is strictly convex on $\mathbb{R}^n$.

(c) $\boldsymbol{x}^* = Q^{-1}\boldsymbol{b}$ is the unique optimal solution.

Before proceeding to the proof, we first introduce an useful strategy for computing $\nabla f(\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x})$ for a **function f in $\boldsymbol{x}$** (i.e., it can be expressed using only the vector $\boldsymbol{x}$ and not its entries explicitly):

(1) Treat the function as a univariate function $f(x)$.

(2) Determine $f'(x)$ and $f''(x)$, with matrices and vectors involved in the expression treated as "scalars" in the differentiation process.

(3) *(Trickier part)* Transform $f'(x)$ and $f''(x)$ into $\nabla f(\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x})$, respectively, by replacing x by $\boldsymbol{x}$ with some suitable adjustments to have suitable dimensions of matrices and vectors, such that the resulting expressions match with the dimensions of $\nabla f(\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x})$, and the products involved are well-defined.

Proof.

(a) We follow the strategy introduced above.
First, we have

$$\nabla f(\boldsymbol{x}) = \frac{1}{2}\underbrace{(Q\boldsymbol{x} + \overbrace{(\boldsymbol{x}^T Q)^T}^{\text{adjust}})}_{\text{“product rule”}} - \overbrace{(\boldsymbol{b}^T)^T}^{\text{adjust}} = \frac{1}{2}(Q\boldsymbol{x} + Q^T\boldsymbol{x}) - \boldsymbol{b} \overset{(Q=Q^T)}{=} Q\boldsymbol{x} - \boldsymbol{b}$$

where we perform the adjustments to let those row vectors become column vectors, matching with the dimension of $\nabla f(\boldsymbol{x})$.
Second, we have

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{2}(Q + Q) = Q.$$

(b) Since Q is positive definite, we know that $\nabla^2 f(\boldsymbol{x})$ is always positive definite, and thus by Theorem 1.2.c we conclude that f is strictly convex on $\mathbb{R}^n$.

(c) Since $\boldsymbol{x}^* = Q^{-1}\boldsymbol{b}$ is the unique solution to the equation $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}$, by Theorem 1.2.f we conclude that $\boldsymbol{x}^*$ is the unique optimal solution.

□

2.2.9 **The steepest descent method for the quadratic problem.**

Lemma 2.2.b. In step (3) of the steepest descent method for the quadratic problem in [2.2.7], we have

$$\boldsymbol{d}_k = \boldsymbol{b} - Q\boldsymbol{x}_k, \quad \alpha_k = \frac{\boldsymbol{d}_k^T\boldsymbol{d}_k}{\boldsymbol{d}_k^T Q\boldsymbol{d}_k}, \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \frac{\boldsymbol{d}_k^T\boldsymbol{d}_k}{\boldsymbol{d}_k^T Q\boldsymbol{d}_k}\boldsymbol{d}_k.$$

Proof. Since $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} - \boldsymbol{b}$, we have $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k) = \boldsymbol{b} - Q\boldsymbol{x}_k$.

Next, we set $g(\alpha) = f(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$. Since $g''(\alpha) = \boldsymbol{d}_k^T\nabla f^2(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)\boldsymbol{d}_k = \boldsymbol{d}_k^T Q\boldsymbol{d}_k \geq 0$, g is convex on $\mathbb{R}$. Now consider:

$$\begin{aligned} g'(\alpha) = \nabla f(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)^T\boldsymbol{d}_k &= 0 \\ \implies [Q(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k) - \boldsymbol{b}]^T\boldsymbol{d}_k &= 0 \\ \implies [-\boldsymbol{d}_k + \alpha Q\boldsymbol{d}_k]^T\boldsymbol{d}_k &= 0 \\ \implies \alpha &= \frac{\boldsymbol{d}_k^T\boldsymbol{d}_k}{\boldsymbol{d}_k^T Q\boldsymbol{d}_k} > 0. \end{aligned}$$

By the convexity of g, being a stationary point of g is equivalent to being a minimizer of g. Hence, we have $\alpha_k = (\boldsymbol{d}_k^T\boldsymbol{d}_k)/(\boldsymbol{d}_k^T Q\boldsymbol{d}_k)$. Thus, $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + (\boldsymbol{d}_k^T\boldsymbol{d}_k)/(\boldsymbol{d}_k^T Q\boldsymbol{d}_k)\boldsymbol{d}_k$. □

2.2.10 **Behaviour of error function during the iteration.** Recall from [2.2.8] that the optimal solution to the quadratic problem in [2.2.7] is $\boldsymbol{x}^* = Q^{-1}\boldsymbol{b}$. Then, we define the *error function*

$$E(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T Q(\boldsymbol{x} - \boldsymbol{x}^*).$$

The function E indeed captures the "error" $f(\boldsymbol{x}) - f(\boldsymbol{x}^*)$ of the objective function value $f(\boldsymbol{x})$ from the minimum value $f(\boldsymbol{x}^*)$, because we can write:

$$\begin{aligned} \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T Q(\boldsymbol{x} - \boldsymbol{x}^*) &= \frac{1}{2}\left(\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{x}^T Q\boldsymbol{x}^* - (\boldsymbol{x}^*)^T Q\boldsymbol{x} + (\boldsymbol{x}^*)^T Q(\boldsymbol{x}^*)\right) \\ &= \frac{1}{2}\left(\boldsymbol{x}^T Q\boldsymbol{x} - 2\boldsymbol{x}^T Q\boldsymbol{x}^* + (\boldsymbol{x}^*)^T Q\boldsymbol{x}^*\right) \\ &= \frac{1}{2}\left(\boldsymbol{x}^T Q\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{b} + (\boldsymbol{x}^*)^T Q\boldsymbol{x}^*\right) \\ &= \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}(\boldsymbol{x}^*)^T Q\boldsymbol{x}^* \\ &= \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} - \left(\frac{1}{2}(\boldsymbol{x}^*)^T Q\boldsymbol{x}^* - \boldsymbol{b}^T\boldsymbol{x}^*\right) \\ &= f(\boldsymbol{x}) - f(\boldsymbol{x}^*). \end{aligned}$$

Since the functions E and f only differ by a constant $f(\boldsymbol{x}^*)$ (which does not depend on $\boldsymbol{x}$), minimizing E is just equivalent to minimizing f. Specifically, performing a change of variable $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{x}^*$, we can turn the quadratic problem in [2.2.7] into the following equivalent problem

$$\text{Minimize} \quad E(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^T Q\boldsymbol{y}.$$

In other words, we may assume that $\boldsymbol{b} = \boldsymbol{0}$ without loss of generality, to make our analysis more convenient.

Lemma 2.2.c. In the steepest descent method for the quadratic problem in [2.2.7], the error function E satisfies:

$$E(\boldsymbol{x}_{k+1}) = \left(1 - \frac{(\boldsymbol{d}_k^T\boldsymbol{d}_k)^2}{(\boldsymbol{d}_k^T Q\boldsymbol{d}_k)(\boldsymbol{d}_k^T Q^{-1}\boldsymbol{d}_k)}\right)E(\boldsymbol{x}_k).$$

Proof. Let $\boldsymbol{y}_k = \boldsymbol{x}_k - \boldsymbol{x}^*$. Then, we have $Q\boldsymbol{y}_k = Q\boldsymbol{x}_k - Q\boldsymbol{x}^* = Q\boldsymbol{x}_k - \boldsymbol{b} = -\boldsymbol{d}_k$ and $\boldsymbol{x}_{k+1} - \boldsymbol{x}^* =$

$\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k - \boldsymbol{x}^* = \boldsymbol{y}_k + \alpha_k \boldsymbol{d}_k$. Hence,

$$\begin{aligned}\frac{E(\boldsymbol{x}_{k+1}) - E(\boldsymbol{x}_k)}{E(\boldsymbol{x}_k)} &= \frac{(1/2)(\boldsymbol{y}_k + \alpha_k \boldsymbol{d}_k)^T Q (\boldsymbol{y}_k + \alpha_k \boldsymbol{d}_k) - (1/2)\boldsymbol{y}_k^T Q \boldsymbol{y}_k}{(1/2)\boldsymbol{y}_k^T Q \boldsymbol{y}_k} \\ &= \frac{2\alpha_k \boldsymbol{d}_k^T Q \boldsymbol{y}_k + \alpha_k^2 \boldsymbol{d}_k^T Q \boldsymbol{d}_k}{\boldsymbol{y}_k^T Q Q^{-1} Q \boldsymbol{y}_k} \\ &= \frac{-2(\boldsymbol{d}_k^T \boldsymbol{d}_k)^2/(\boldsymbol{d}_k^T Q \boldsymbol{d}_k) + (\boldsymbol{d}_k^T \boldsymbol{d}_k)^2/(\boldsymbol{d}_k^T Q \boldsymbol{d}_k)}{\boldsymbol{d}_k Q^{-1} \boldsymbol{d}_k} \\ &= -\frac{(\boldsymbol{d}_k^T \boldsymbol{d}_k)^2}{(\boldsymbol{d}_k^T Q \boldsymbol{d}_k)(\boldsymbol{d}_k^T Q^{-1} \boldsymbol{d}_k)}.\end{aligned}$$

It follows that

$$E(\boldsymbol{x}_{k+1}) = \left(1 - \frac{(\boldsymbol{d}_k^T \boldsymbol{d}_k)^2}{(\boldsymbol{d}_k^T Q \boldsymbol{d}_k)(\boldsymbol{d}_k^T Q^{-1} \boldsymbol{d}_k)}\right) E(\boldsymbol{x}_k).$$

□

2.2.11 **Kantorovich inequality.** The last lemma we need for establishing convergence properties of the steepest descent method is the *Kantorovich inequality*, which is a technical result that gives a lower bound on a certain expression.

To prove the Kantorovich inequality, we need to use the following important theorem in linear algebra: *principal axis theorem* or *spectral theorem*.

Theorem 2.2.d (Principal axis theorem)**.** Let A be an $n \times n$ symmetric matrix with eigenvalues $\lambda_1, \dots, \lambda_n$. There exists an orthogonal matrix P such that $P^{-1}AP = \Lambda$ where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, which refers to the $n \times n$ diagonal matrix whose (i, i)th entry is λ_i for all $i = 1, \dots, n$.

Proof. Omitted. □

[Note: Since P is an orthogonal matrix, we have $P^{-1} = P^T$. Hence, we can also write:

$$P^T AP = \Lambda, \quad A = P\Lambda P^{-1}, \quad A = P\Lambda P^T.$$

Indeed, since P^T is also orthogonal, by taking $Q = P^T$, we can also write:

$$QAQ^{-1} = \Lambda, \quad QAQ^T = \Lambda, \quad A = Q^{-1}\Lambda Q, \quad A = Q^T \Lambda Q,$$

for some orthogonal matrix Q.]

Lemma 2.2.e (Kantorovich inequality)**.** Let λ_1 and λ_n be the smallest and largest eigenvalues of an $n \times n$ positive definite matrix Q, respectively. For all $\boldsymbol{x} \neq \boldsymbol{0}$, we have

$$\frac{(\boldsymbol{x}^T \boldsymbol{x})^2}{(\boldsymbol{x}^T Q \boldsymbol{x})(\boldsymbol{x}^T Q^{-1} \boldsymbol{x})} \geq \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}.$$

Proof. Since Q is positive definite (hence symmetric), by Theorem 2.2.d there exists an orthogonal matrix P such that $Q = P^T \Lambda P$ where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ with $0 < \lambda_1 \leq \dots \leq \lambda_n$ being the eigenvalues of Q, sorted in ascending order. Now let $\boldsymbol{y} = P\boldsymbol{x}$. Then, we have:

- $\boldsymbol{y}^T \boldsymbol{y} = \boldsymbol{x}^T P^T P \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{x}$.
- $\boldsymbol{x}^T Q \boldsymbol{x} = \boldsymbol{x}^T P^T \Lambda P \boldsymbol{x} = (P\boldsymbol{x})^T \Lambda (P\boldsymbol{x}) = \boldsymbol{y}^T \Lambda \boldsymbol{y}$.
- $\boldsymbol{x}^T Q^{-1} \boldsymbol{x} = \boldsymbol{x}(P^T \Lambda P)^{-1} \boldsymbol{x} = \boldsymbol{x}^T P^{-1} \Lambda^{-1} (P^T)^{-1} \boldsymbol{x} = (P\boldsymbol{x})^T \Lambda^{-1} (P\boldsymbol{x}) = \boldsymbol{y}^T \Lambda^{-1} \boldsymbol{y}$.

Hence, writing $\boldsymbol{y} = (y_1, \dots, y_n)$, we have

$$\frac{(\boldsymbol{x}^T \boldsymbol{x})^2}{(\boldsymbol{x}^T Q \boldsymbol{x})(\boldsymbol{x}^T Q^{-1} \boldsymbol{x})} = \frac{(\boldsymbol{y}^T \boldsymbol{y})^2}{(\boldsymbol{y}^T \Lambda \boldsymbol{y})(\boldsymbol{y}^T \Lambda^{-1} \boldsymbol{y})} = \frac{(\sum_{j=1}^n y_j^2)^2}{(\sum_{i=1}^n \lambda_i y_i^2)(\sum_{i=1}^n \lambda_i^{-1} y_i^2)}.$$

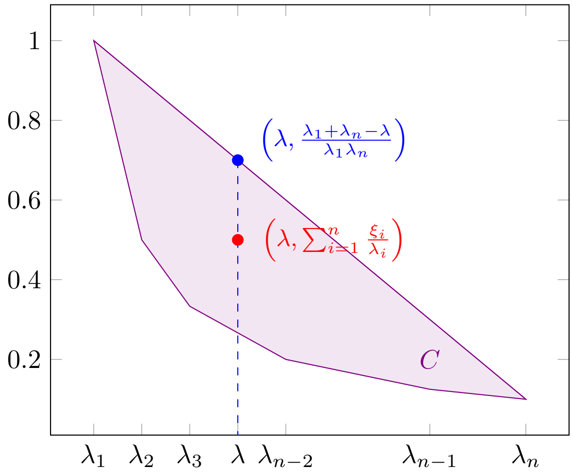
Taking reciprocal, it remains to show that

$$\frac{(\sum_{i=1}^n \lambda_i y_i^2)(\sum_{i=1}^n \lambda_i^{-1} y_i^2)}{(\sum_{j=1}^n y_j^2)^2} \leq \frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n}.$$

To this end, let $\xi_i = y_i^2 / \sum_{j=1}^n y_j^2$ for all $i = 1, \dots, n$. Then, $\xi_i \geq 0$ for all $i = 1, \dots, n$ with $\sum_{i=1}^n \xi_i = 1$. Hence, the inequality above can be rewritten as

$$\left(\sum_{i=1}^n \xi_i \lambda_i\right)\left(\sum_{i=1}^n \frac{\xi_i}{\lambda_i}\right) \leq \frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n}.$$

To establish this inequality, consider the *convex polygon generated by the points* $(\lambda_1, 1/\lambda_1), \dots, (\lambda_n, 1/\lambda_n)$, which is given by $C = \left\{(x,y) \in \mathbb{R}^2 : (x,y) = \sum_{i=1}^n \alpha_i(\lambda_i, 1/\lambda_i) \text{ with } \alpha_i \geq 0\ \forall i = 1,\dots,n \text{ and } \sum_{i=1}^n \alpha_i = 1\right\}$, namely the set of all convex combinations of those points, and can be illustrated graphically as follows:


1
0.8
0.6
0.4
0.2
$\left(\lambda, \frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n}\right)$
$\left(\lambda, \sum_{i=1}^n \frac{\xi_i}{\lambda_i}\right)$
C
λ_1
λ_2
λ_3
λ
λ_{n-2}
λ_{n-1}
λ_n

Let $\lambda = \sum_{i=1}^n \xi_i \lambda_i$. Then, we have $(\lambda, \sum_{i=1}^n \xi_i/\lambda_i) \in C$ since the ξ_i's are nonnegative and sum to 1. From the picture above, we can observe the following inequality:

$$\sum_{i=1}^n \frac{\xi_i}{\lambda_i} \leq \frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n}.$$

(We shall omit the formal proof of this inequality here.)

Therefore, we have

$$\begin{aligned}\left(\sum_{i=1}^n \xi_i \lambda_i\right)\left(\sum_{i=1}^n \frac{\xi_i}{\lambda_i}\right) &\leq \lambda \cdot \frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n} \\ &\leq \max_{\lambda_1 \leq \lambda \leq \lambda_n} \left(\lambda \cdot \frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n}\right) \\ &\leq \frac{1}{\lambda_1\lambda_n} \underbrace{\max_{\lambda_1 \leq \lambda \leq \lambda_n} \left[\left(\frac{\lambda_1+\lambda_n}{2}\right)^2 - \left(\lambda - \frac{\lambda_1+\lambda_n}{2}\right)^2\right]}_{\text{maximum achieved iff } \lambda = (\lambda_1+\lambda_n)/2} \\ &= \frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n},\end{aligned}$$

as desired. □

2.2.12 Convergence properties of the steepest descent method for the quadratic problem.

Theorem 2.2.f. Consider the quadratic problem in [2.2.7]. For every starting point $\boldsymbol{x}_0 \in \mathbb{R}^n$, the sequence $\{\boldsymbol{x}_k\}$ generated by the iterative formula of the steepest descent method converges to the unique minimizer $\boldsymbol{x}^*$ of f. Furthermore, the error function E satisfies:

$$E(\boldsymbol{x}_{k+1}) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 E(\boldsymbol{x}_k),$$

where λ_1 and λ_n are the smallest and largest eigenvalues of the matrix Q, respectively.

Proof. By Lemmas 2.2.c and 2.2.e, we readily have

$$E(\boldsymbol{x}_{k+1}) \leq \left(1 - \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}\right) E(\boldsymbol{x}_k) = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 E(\boldsymbol{x}_k).$$

This establishes the inequality between error functions.

Now, we proceed to show the convergence of the steepest descent method. By induction, we have

$$0 \leq E(\boldsymbol{x}_k) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^{2k} E(\boldsymbol{x}_0) \underset{k\to\infty}{\longrightarrow} 0,$$

since $0 \leq (\lambda_n - \lambda_1)/(\lambda_n + \lambda_1) < 1$. This implies that $E(\boldsymbol{x}_k) \underset{k\to\infty}{\longrightarrow} 0$.

Claim: We have $\lambda_1 \boldsymbol{x}^T\boldsymbol{x} \leq \boldsymbol{x}^T Q\boldsymbol{x} \leq \lambda_n \boldsymbol{x}^T\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

Proof. By the principal axis theorem, there exists an orthogonal matrix P such that $Q = P^T\Lambda P$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Let $\boldsymbol{y} = P\boldsymbol{x}$. From the proof of Lemma 2.2.e, we know that $\boldsymbol{y}^T\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{x}$ and $\boldsymbol{x}^T Q\boldsymbol{x} = \boldsymbol{y}^T\Lambda\boldsymbol{y}$. Thus, it suffices to show that $\lambda_1\boldsymbol{y}^T\boldsymbol{y} \leq \boldsymbol{y}^T\Lambda\boldsymbol{y} \leq \lambda_n\boldsymbol{y}^T\boldsymbol{y}$, which follows from noting that

$$\boldsymbol{y}^T\Lambda\boldsymbol{y} = \sum_{i=1}^n \lambda_i y_i^2 \begin{cases} \geq \lambda_1 \sum_{i=1}^n y_i^2 = \lambda_1\boldsymbol{y}^T\boldsymbol{y}, \\ \leq \lambda_n \sum_{i=1}^n y_i^2 = \lambda_n\boldsymbol{y}^T\boldsymbol{y}. \end{cases}$$

□

Using the claim, we then have

$$0 \leq \frac{1}{2}\lambda_1(\boldsymbol{x}_k - \boldsymbol{x}^*)^T(\boldsymbol{x}_k - \boldsymbol{x}^*) \leq \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}^*)^T Q(\boldsymbol{x}_k - \boldsymbol{x}^*) = E(\boldsymbol{x}_k) \underset{k\to\infty}{\longrightarrow} 0,$$

which implies that $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 = (\boldsymbol{x}_k - \boldsymbol{x}^*)^T(\boldsymbol{x}_k - \boldsymbol{x}^*) \underset{k\to\infty}{\longrightarrow} 0$, and hence $\boldsymbol{x}_k \underset{k\to\infty}{\longrightarrow} \boldsymbol{x}^*$. □

2.3 Newton's Method

2.3.1 In [2.1.10], we have already encountered the *Newton's method* in the context of line search algorithms. However, we actually have a multivariate version of Newton's method, which is designed in a similar way, also using quadratic approximations.

2.3.2 **Newton's method.** Consider the following unconstrained minimization problem

$$\begin{aligned} &\text{Minimize} && f(\boldsymbol{x}) \\ &\text{subject to} && \boldsymbol{x} \in \mathbb{R}^n. \end{aligned}$$

Suppose that the objective function $f \in C^2$, and that the Hessian matrix $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite at a local minimizer $\boldsymbol{x}^*$, which is to be estimated by the Newton's method. Then, we know that $\nabla^2 f(\boldsymbol{x})$ is positive definite (hence invertible) whenever $\boldsymbol{x}$ is sufficiently close to $\boldsymbol{x}^*$.

The **Newton's method** then works as follows.

(1) *(Initialization)* Choose a termination scalar $\varepsilon > 0$ and a starting point $\boldsymbol{x}_0$. Set $k = 0$.

(2) *(Quadratic approximation)* Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\nabla^2 f(\boldsymbol{x}_k)]^{-1}\nabla f(\boldsymbol{x}_k)$, assuming that $\nabla^2 f(\boldsymbol{x}_k)$ is invertible.

[Note: We have the invertibility provided that $\boldsymbol{x}_k$ is sufficiently close to $\boldsymbol{x}^*$.]

(3) *(Checking for termination)* If $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \varepsilon$ or $\|\nabla f(\boldsymbol{x}_{k+1})\| < \varepsilon$, stop and output $\boldsymbol{x}_{k+1}$ as an estimated minimizer. Otherwise, replace k by $k+1$ and go to (2).

2.3.3 **Idea behind the Newton's method.** At each iteration with a point $\boldsymbol{x}_k$, Newton's method exploits the *quadratic approximation* of the function f at $\boldsymbol{x}_k$:

$$q(\boldsymbol{x}) = f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^T\nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k)$$

which agrees with f at $\boldsymbol{x}_k$ up to the Hessian matrix, i.e., $f(\boldsymbol{x}_k) = q(\boldsymbol{x}_k)$, $\nabla f(\boldsymbol{x}_k) = \nabla q(\boldsymbol{x}_k)$, and $\nabla^2 f(\boldsymbol{x}_k) = \nabla^2 q(\boldsymbol{x}_k)$. Then, the next point $\boldsymbol{x}_{k+1}$ is chosen to be the solution to $\nabla q(\boldsymbol{x}) = \mathbf{0}$, i.e., the stationary point of the quadratic approximation $q(\boldsymbol{x})$; provided that $\boldsymbol{x}_k$ is sufficiently close to $\boldsymbol{x}^*$ such that $\nabla^2 f(\boldsymbol{x}_k)$ is positive definite, we know that the stationary point is also a minimizer of q since q is convex, which follows from the positive definiteness of $\nabla^2 q(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}_k)$.

Since $\nabla q(\boldsymbol{x}) = \nabla f(\boldsymbol{x}_k) + \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k)$, we have

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\nabla^2 f(\boldsymbol{x}_k)]^{-1}\nabla f(\boldsymbol{x}_k).$$

The two terminating conditions are analogous to the ones for the Newton's method in [2.1.10], with a similar rationale.

2.3.4 **Property of the Newton's method.**

Proposition 2.3.a. Let $f \in C^3$ be a function on $\mathbb{R}^n$. Suppose that the Hessian matrix $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite at a local minimizer $\boldsymbol{x}^*$. Then, the sequence $\{\boldsymbol{x}_k\}$ generated by the iterative formula $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\nabla^2 f(\boldsymbol{x}_k)]^{-1}\nabla f(\boldsymbol{x}_k)$ of the Newton's method in [2.3.2] converges to $\boldsymbol{x}^*$ with order at least two, provided that $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{x}^*$.

Proof. Omitted. □

[⚠ Warning: While this result suggests that the Newton's method has good convergence performance if the starting point is sufficiently close to $\boldsymbol{x}^*$, a major drawback of the Newton's method is its high computational complexity, since we need to obtain the *inverse of a Hessian matrix* at each iteration, which requires substantial computations.]

2.4 Conjugate Direction Methods

2.4.1 From the discussions of the steepest descent method and the Newton's method in Sections 2.2 and 2.3, we can observe that each method has its own merits and drawbacks. In particular, the *steepest descent method* has a relatively low computational complexity since matrix inversion is not needed, but it tends to have poor performance as a stationary point is approached. On the other hand, the *Newton's method* has a good convergence performance (if the starting point is sufficiently close to the minimizer), but has a high demand on computations.

These observations motivate the development of *conjugate direction methods*, which aim to be "intermediate" between these two methods, thereby achieving "the best of both worlds" hopefully. Conjugate direction methods are considered among the *best general purpose methods* available currently.

2.4.2 **Q-orthogonality.** Before stating the conjugate direction methods, we first introduce some preliminary concepts. Let Q be a $n \times n$ symmetric matrix. Two vectors $\boldsymbol{d}_1, \boldsymbol{d}_2 \in \mathbb{R}^n$ are said to be **Q-orthogonal** (or **Q-conjugate**), if $\boldsymbol{d}_1^T Q \boldsymbol{d}_2 = 0$. [Note: This reduces the usual concept of orthogonality if Q is the identity matrix.]

A finite collection of vectors $\boldsymbol{d}_0, \boldsymbol{d}_1, \ldots, \boldsymbol{d}_k$ is said to be **Q-orthogonal** (or **Q-conjugate**) if $\boldsymbol{d}_i^T Q \boldsymbol{d}_j = 0$ for all $i, j = 0, 1, \ldots, k$ with $i \neq j$.

2.4.3 **Properties of Q-orthogonality.**

Proposition 2.4.a. If Q is positive definite and $\boldsymbol{d}_0, \boldsymbol{d}_1, \ldots, \boldsymbol{d}_k$ are *nonzero* vectors that are Q-orthogonal, then these vectors are linearly independent.

Proof. We have

$$\begin{aligned} \alpha_0\boldsymbol{d}_0 + \alpha_1\boldsymbol{d}_1 + \cdots + \alpha_k\boldsymbol{d}_k &= \boldsymbol{0} \\ \implies \boldsymbol{d}_i^T Q(\alpha_0\boldsymbol{d}_0 + \alpha_1\boldsymbol{d}_1 + \cdots + \alpha_k\boldsymbol{d}_k) &= \boldsymbol{d}_i^T Q\boldsymbol{0} \\ \overset{(Q\text{-orthogonality})}{\implies} \alpha_i \underbrace{\boldsymbol{d}_i^T Q\boldsymbol{d}_i}_{>0 \text{ as } Q \text{ is positive definite}} &= 0 \\ \implies \alpha_i &= 0, \end{aligned}$$

for all $i = 1, \ldots, n$. Thus, the vectors are linearly independent by definition. □

Proposition 2.4.b. Let Q be a symmetric matrix. Then, any two eigenvectors of Q corresponding to distinct eigenvalues of Q are Q-orthogonal.

Proof. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two such eigenvectors, with eigenvalues λ and μ, respectively, with $\lambda \neq \mu$. By definition, $Q\boldsymbol{x} = \lambda\boldsymbol{x}$ and $Q\boldsymbol{y} = \mu\boldsymbol{y}$. Note that:

- $\boldsymbol{x}^T Q\boldsymbol{y} = \boldsymbol{x}^T(\mu\boldsymbol{y}) = \mu\boldsymbol{x}^T\boldsymbol{y}$.
- $\boldsymbol{y}^T Q\boldsymbol{x} = \boldsymbol{y}^T(\lambda\boldsymbol{x}) = \lambda\boldsymbol{x}^T\boldsymbol{y}$.

As Q is symmetric, we have $\boldsymbol{x}^T Q\boldsymbol{y} = \boldsymbol{y}^T Q\boldsymbol{x}$, and thus

$$\mu\boldsymbol{x}^T\boldsymbol{y} = \lambda\boldsymbol{x}^T\boldsymbol{y} \implies (\mu - \lambda)\boldsymbol{x}^T\boldsymbol{y} = 0 \overset{(\mu\neq\lambda)}{\implies} \boldsymbol{x}^T\boldsymbol{y} = 0.$$

Therefore, $\boldsymbol{x}^T Q\boldsymbol{y} = \mu\boldsymbol{x}^T\boldsymbol{y} = \mu(0) = 0$, as desired. □

Remarks:

- The proof also reveals that any two such eigenvectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal (i.e., $\boldsymbol{x}^T\boldsymbol{y} = 0$).
- Repeated application of Proposition 2.4.b suggests that a collection of eigenvectors of Q whose eigenvalues are all distinct is Q-orthogonal.

2.4.4 **Conjugate direction methods for the quadratic problem.** Again, we start by analyzing the quadratic problem in [2.2.7]:

$$\text{Minimize} \quad f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$$

where Q is a positive definite $n \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^n$ is a vector.

For the quadratic problem, a **conjugate direction method** works as follows. Let $\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{n-1}$ be n nonzero Q-orthogonal vectors.

(1) *(Initialization)* Choose a starting point $\boldsymbol{x}_0$. Set $k = 0$.

(2) *(Iterative step)* Set $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k) = Q\boldsymbol{x}_k - \boldsymbol{b}$, $\alpha_k = -\boldsymbol{g}_k^T\boldsymbol{d}_k/(\boldsymbol{d}_k^T Q\boldsymbol{d}_k)$, and $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$.

(3) *(Checking for termination)* If $k = n - 1$, then stop and $\boldsymbol{x}_n$ is outputted as the *exact* minimizer $\boldsymbol{x}^*$. Otherwise, replace k by $k + 1$ and go to (2).

2.4.5 **Conjugate direction theorem.** The property that $\boldsymbol{x}_n$ resulting from a conjugate direction method is the *exact* minimizer $\boldsymbol{x}^*$ is justified by the *conjugate direction theorem*.

Theorem 2.4.c (Conjugate direction theorem)**.** For a conjugate direction method on the quadratic problem in [2.2.7], we have $\boldsymbol{x}_n = \boldsymbol{x}^*$, where $\boldsymbol{x}^*$ is the unique minimizer, namely $Q^{-1}\boldsymbol{b}$.

Proof. By Proposition 2.4.a, the n vectors $\boldsymbol{d}_0, \dots, \boldsymbol{d}_{n-1}$ are linearly independent, and thus they form a basis for $\mathbb{R}^n$. This implies that we can write

$$\boldsymbol{x}^* - \boldsymbol{x}_0 = c_0\boldsymbol{d}_0 + \cdots + c_{n-1}\boldsymbol{d}_{n-1} \tag{4}$$

for some $c_0, \dots, c_{n-1} \in \mathbb{R}$. Note that we have

$$\begin{aligned}
\boldsymbol{x}_n &= \boldsymbol{x}_{n-1} + \alpha_{n-1}\boldsymbol{d}_{n-1} \\
&= \boldsymbol{x}_{n-2} + \alpha_{n-2}\boldsymbol{d}_{n-2} + \alpha_{n-1}\boldsymbol{d}_{n-1} \\
&= \cdots \\
&= \boldsymbol{x}_0 + \alpha_0\boldsymbol{d}_0 + \cdots + \alpha_{n-1}\boldsymbol{d}_{n-1},
\end{aligned}$$

and hence $\boldsymbol{x}_n - \boldsymbol{x}_0 = \alpha_0\boldsymbol{d}_0 + \cdots + \alpha_{n-1}\boldsymbol{d}_{n-1}$.
So, it remains to show that $c_k = \alpha_k = -\boldsymbol{g}_k^T\boldsymbol{d}_k/(\boldsymbol{d}_k^T Q\boldsymbol{d}_k)$ for all $k = 0, \dots, n-1$.
Fix any $k \in \{0, \dots, n-1\}$. From Equation (4), we get

$$\boldsymbol{d}_k^T Q(\boldsymbol{x}^* - \boldsymbol{x}_0) = \boldsymbol{d}_k^T Q(c_0\boldsymbol{d}_0 + \cdots + c_{n-1}\boldsymbol{d}_{n-1}) \overset{(Q\text{-orthogonality})}{=} c_k\boldsymbol{d}_k^T Q\boldsymbol{d}_k,$$

which implies that

$$c_k = \frac{\boldsymbol{d}_k^T Q(\boldsymbol{x}^* - \boldsymbol{x}_0)}{\boldsymbol{d}_k^T Q\boldsymbol{d}_k}.$$

Note that

$$\boldsymbol{d}_k^T(\boldsymbol{x}_k - \boldsymbol{x}_0) = \boldsymbol{d}_k^T Q(\alpha_0\boldsymbol{d}_0 + \cdots + \alpha_{k-1}\boldsymbol{d}_{k-1}) \overset{(Q\text{-orthogonality})}{=} 0,$$

if $k \geq 1$ and $\boldsymbol{d}_k^T(\boldsymbol{x}_k - \boldsymbol{x}_0) = 0$ also if $k = 0$. Thus, we have

$$c_k = \frac{\boldsymbol{d}_k^T \overbrace{Q(\boldsymbol{x}^* - \boldsymbol{x}_k)}^{\boldsymbol{b} - Q\boldsymbol{x}_k}}{\boldsymbol{d}_k^T Q\boldsymbol{d}_k} = -\frac{\boldsymbol{g}_k^T\boldsymbol{d}_k}{\boldsymbol{d}_k^T Q\boldsymbol{d}_k},$$

as desired. □

2.4.6 To apply a conjugate direction method, we need to have n nonzero Q-orthogonal vectors. However, so far we do not have any guide on how to choose them. To provide a systematic way to choose those vectors, the *conjugate gradient method* is developed, which is a conjugate direction method with those vectors determined sequentially at each iteration, based on the gradient $\nabla f(\boldsymbol{x}_k)$ at the point $\boldsymbol{x}_k$.

2.4.7 **Conjugate gradient method for the quadratic problem.** As usual, we start by analyzing the quadratic problem in [2.2.7]. For the quadratic problem, the **conjugate gradient method** works as follows.

(1) *(Initialization)* Choose a starting point $\boldsymbol{x}_0$. Set $k = 0$.

(2) *(Determining the first movement direction)* Set $\boldsymbol{g}_0 = \nabla f(\boldsymbol{x}_0) = Q\boldsymbol{x}_0 - \boldsymbol{b}$ and $\boldsymbol{d}_0 = -\boldsymbol{g}_0$.

(3) *(Moving to the next point)* Set $\alpha_k = -\boldsymbol{g}_k^T\boldsymbol{d}_k/(\boldsymbol{d}_k^T Q\boldsymbol{d}_k)$ and $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$.

(4) *(Checking for termination)* Set $\boldsymbol{g}_{k+1} = \nabla f(\boldsymbol{x}_{k+1}) = Q\boldsymbol{x}_{k+1} - \boldsymbol{b}$. If $\boldsymbol{g}_{k+1} = \boldsymbol{0}$, stop and $\boldsymbol{x}_{k+1}$ is outputted as the *exact* minimizer $\boldsymbol{x}^*$.

(5) *(Determining the next movement direction)* Set $\beta_k = \boldsymbol{g}_{k+1}^T Q\boldsymbol{d}_k/\boldsymbol{d}_k^T Q\boldsymbol{d}_k$ and $\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_k\boldsymbol{d}_k$.

(6) *(Repetition)* Replace k by $k+1$ and go to (3).

[Note: It can be guaranteed that for some $k \leq n$, we have $\boldsymbol{g}_k = \boldsymbol{0}$ and the point $\boldsymbol{x}_k$ is the exact minimizer, so this algorithm can output the exact minimizer $\boldsymbol{x}^*$ in finite number of steps.]

2.4.8 **Conjugate gradient theorem.** The property that the conjugate gradient method is a kind of conjugate direction method is justified by the *conjugate gradient theorem*.

Theorem 2.4.d (Conjugate gradient theorem)**.** The conjugate gradient method on the quadratic problem in [2.2.7] is a conjugate direction method, i.e., the vectors $\boldsymbol{d}_k$'s obtained during the iteration are Q-orthogonal.

Proof. Omitted. □

2.4.9 **Quadratic approximations for nonquadratic problems.** To attack a general unconstrained minimization problem

$$\text{Minimize} \quad f(\boldsymbol{x})$$

by the conjugate gradient method, we can make a suitable *quadratic approximation* to the problem. Specifically, we make the following associations at each point $\boldsymbol{x}_k$: $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$ and $Q = \nabla^2 f(\boldsymbol{x}_k)$, assuming that the gradient and Hessian matrix are defined. This yields the following **conjugate gradient method** for general unconstrained minimization problem:

(1) *(Initialization)* Choose a termination scalar $\varepsilon > 0$ and a starting point $\boldsymbol{x}_0$. Set $k = 0$.

(2) *(Determining the first movement direction)* Set $\boldsymbol{g}_0 = \nabla f(\boldsymbol{x}_0)$ and $\boldsymbol{d}_0 = -\boldsymbol{g}_0$.

(3) *(Iterative steps)* For $k = 0, \dots, n-1$, do:

   i. Set $\alpha_k = -\boldsymbol{g}_k^T \boldsymbol{d}_k / \boldsymbol{d}_k^T \nabla^2 f(\boldsymbol{x}_k) \boldsymbol{d}_k$ and $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, assuming that $\boldsymbol{d}_k^T \nabla^2 f(\boldsymbol{x}_k) \boldsymbol{d}_k \neq 0$.

   ii. Set $\boldsymbol{g}_{k+1} = \nabla f(\boldsymbol{x}_{k+1})$. If $\|\boldsymbol{g}_k\| < \varepsilon$, then stop and output $\boldsymbol{x}_{k+1}$ as an estimated minimizer.

   iii. Unless $k = n-1$, set $\beta_k = \boldsymbol{g}_{k+1}^T \nabla^2 f(\boldsymbol{x}_k) \boldsymbol{d}_k / \boldsymbol{d}_k^T \nabla^2 f(\boldsymbol{x}_k) \boldsymbol{d}_k$, $\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_k \boldsymbol{d}_k$, and go to (i).

(4) *(Repetition)* Replace $\boldsymbol{x}_0$ by $\boldsymbol{x}_n$, $k = n$ by $k = 0$, and go to (2).

[⚠ Warning: Beyond the quadratic case, this conjugate gradient method suffers from a similar issue as the Newton's method: high computational complexity; here a Hessian matrix needs to be computed at each iteration!]

3 Constrained Optimization

3.0.1 In Sections 1 and 2, we have primarily focused on developing methods for solving *unconstrained* optimization problems. However, many optimization problems of practical interest are constrained, with the constraints often coming from real-life considerations and limitations (e.g., resource availability). Hence, in Section 3, we will develop methods for solving *constrained* optimization (minimization) problems, using a similar strategy as before, namely deriving necessary and sufficient conditions for optimality.

3.0.2 **Constrained minimization problems in standard form.** For tractability and consistency, throughout Section 3 we will focus on analyzing constrained minimization problems in the following **standard form**:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_1(\boldsymbol{x}) = 0, \quad h_2(\boldsymbol{x}) = 0, \quad \ldots, \quad h_m(\boldsymbol{x}) = 0, \\ & g_1(\boldsymbol{x}) \le 0, \quad g_2(\boldsymbol{x}) \le 0, \quad \ldots, \quad g_q(\boldsymbol{x}) \le 0, \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and all the functions f, h_i's and g_j's are continuous.

Remarks:

- We can change this to a constrained maximization problem via replacing f by $-f$; such a maximization problem is also in the standard form.
- This encapsulates many possible constrained optimization problems. For example, taking $q = 0$ leads to a problem with equality constraints only (here we interpret $q = 0$ as meaning that there is no such g_j's), and for "≥ 0" constraint, we can simply take negative on the left-hand side to convert it into the "≤ 0" form.

To express this kind of problem in a more compact way, we let $\boldsymbol{h} = (h_1, \ldots, h_m)$ and $\boldsymbol{g} = (g_1, \ldots, g_q)$ be *row*-vector-valued functions, and then the problem above can be rewritten as:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}_{1\times m}, \\ & \boldsymbol{g}(\boldsymbol{x}) \le \mathbf{0}_{1\times q}, \end{aligned}$$

where $\mathbf{0}_{1\times m}$ and $\mathbf{0}_{1\times q}$ denote the $1 \times m$ and $1 \times q$ row vectors, respectively.

Remarks:

- Henceforth we just use "$\mathbf{0}$" to stand for either row vector, for convenience.
- When using "$\le$" and "$\ge$" to compare vectors, they are understood to carry the componentwise meaning.

3.0.3 **Basic terms about constrained optimization.**

- The constraints $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \le \mathbf{0}$ are called **functional constraints**, with the former being called **equality constraints** and the latter being called **inequality constraints**.
- An inequality constraint $g_j(\boldsymbol{x}) \le 0$ is said to be **active** at a feasible point $\boldsymbol{x}$ if $g_j(\boldsymbol{x}) = 0$ (attaining equality), and **inactive** at $\boldsymbol{x}$ if $g_j(\boldsymbol{x}) < 0$ (not attaining equality). By convention, every equality constraint $h_i(\boldsymbol{x}) = 0$ is said to be **active** at every feasible point $\boldsymbol{x}$. [Note: A general idea for finding local minimizers is to focus on *active* constraints.]
- A **surface** $S \subseteq \mathbb{R}^n$ is defined by equality constraints: $S = \{\boldsymbol{x} \in \mathbb{R}^n : h_i(\boldsymbol{x}) = 0 \ \forall i = 1, \ldots, m\}$. It is said to be **smooth** if each $h_i \in C^1$.
- A **curve** on a surface S is a family of points $\boldsymbol{x}(t) \in S$ continuously parameterized by t for all $t \in [a, b]$, i.e., the set $\{\boldsymbol{x}(t) : t \in [a, b]\}$. [Note: This plays a similar role as the *feasible direction* introduced previously in the context of constrained optimization, as we will see in some of the upcoming proofs.]

- A curve is **differentiable** if $\dot{\boldsymbol{x}}(t) := \mathrm{d}/\mathrm{d}t\ \boldsymbol{x}(t)$ exists for all $t \in (a,b)$, and **twice-differentiable** if $\ddot{\boldsymbol{x}}(t) := \mathrm{d}^2/\mathrm{d}t^2\ \boldsymbol{x}(t)$ exists for all $t \in (a,b)$.
- A curve is said to **pass through a point $\boldsymbol{x}^*$** if $\boldsymbol{x}^* = \boldsymbol{x}(t^*)$ for some $t^* \in (a,b)$, and the **derivative of the curve at the point $\boldsymbol{x}^*$** is $\dot{\boldsymbol{x}}(t^*)$.
- The **tangent plane** at $\boldsymbol{x}^*$ is the set of the derivatives of all differentiable curves on a smooth surface S at the point $\boldsymbol{x}^*$.

3.0.4 **An explicit representation of the tangent plane.** The concept of *tangent plane* turns out to play an important role in the theory of constrained optimization. However, it is difficult to determine what the tangent plane is by its definition directly. Fortunately, there is indeed an explicit representation of the tangent plane at a *regular point*, overcoming this difficulty.

Let $S = \{\boldsymbol{x} \in \mathbb{R}^n : h_i(\boldsymbol{x}) = 0\ \forall i = 1, \dots, m\}$ be a smooth surface. A point $\boldsymbol{x}^* \in S$ is said to be a **regular point** of the constraints $h_i(\boldsymbol{x}) = 0\ \forall i = 1, \dots, m$ if $\nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_m(\boldsymbol{x}^*)$ are linearly independent.

Theorem 3.0.a. Let $h_1, \dots, h_m \in C^1$. At a regular point $\boldsymbol{x}^*$ of the surface defined by $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, the tangent plane equals

$$M = \{\boldsymbol{y} \in \mathbb{R}^n : \nabla \boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{y} = \boldsymbol{0}\},$$

where

$$\nabla \boldsymbol{h}(\boldsymbol{x}) = \begin{bmatrix} \nabla h_1(\boldsymbol{x}) & \cdots & \nabla h_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \partial h_1/\partial x_1 & \cdots & \partial h_m/\partial x_1 \\ \vdots & \ddots & \vdots \\ \partial h_1/\partial x_n & \cdots & \partial h_m/\partial x_n \end{bmatrix}$$

is an $n \times m$ matrix.

Proof. Omitted. □

3.1 Necessary and Sufficient Conditions for Local Extremizers

3.1.1 Now, we have enough ingredient to derive necessary and sufficient conditions for local extremizers for a constrained optimization problem of the standard form. We start by proving the following lemma.

Lemma 3.1.a. Let $f, h_1, \dots, h_m \in C^1$ and $\boldsymbol{x}^*$ be a regular point of the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ and a local extremizer of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$. Then, every $\boldsymbol{y} \in \mathbb{R}^n$ satisfying $\nabla \boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{y} = \boldsymbol{0}$ *(lying in the tangent plane at $\boldsymbol{x}^*$)* must also satisfy $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{y} = 0$.

Proof. Fix any $\boldsymbol{y} \in \mathbb{R}^n$ with $\nabla \boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{y} = 0$. By Theorem 3.0.a, $\boldsymbol{y}$ lies in the tangent plane at $\boldsymbol{x}^*$. Hence, there exists a differentiable curve $\boldsymbol{x}(t)$ on the surface defined by $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ such that $\boldsymbol{x}(0) = \boldsymbol{x}^*$, $\dot{\boldsymbol{x}}(0) = \boldsymbol{y}$, and $\boldsymbol{h}(\boldsymbol{x}(t)) = \boldsymbol{0}$ for all $-a \le t \le a$, with $a > 0$. [Note: By reparameterization if necessary, we can assume that $\dot{\boldsymbol{x}}(0) = \boldsymbol{y}$ rather than just $\dot{\boldsymbol{x}}(t^*) = \boldsymbol{y}$ for some t^*.]

Since $\boldsymbol{x}^*$ is a local extremizer of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, it is also a local extremizer of f along the curve $\boldsymbol{x}(t)$. Therefore, 0 is a local extremum point of $f(\boldsymbol{x}(t))$ over $[-a, a]$, which implies by first-order necessary condition that

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}(t))\bigg|_{t=0} = 0 \implies \nabla f(\boldsymbol{x}(t))^T \dot{\boldsymbol{x}}(t)\big|_{t=0} = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{y} = 0.$$

□

3.1.2 **First-order necessary conditions for problems with equality constraints only.** With the help of Lemma 3.1.a, we can obtain the following first-order necessary conditions for constrained optimization problem with *only equality constraints*.

Proposition 3.1.b. Let $f, h_1, \dots, h_m \in C^1$. Assume that $\boldsymbol{x}^*$ is a local extremizer of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, and is also a regular point of these constraints. Then, there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ (known as **Lagrange multiplier**) such that $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} = \boldsymbol{0}$.

Proof. By Lemma 3.1.a, we have $\nabla f(\boldsymbol{x}^*)^T\boldsymbol{y} = 0$ whenever $\nabla \boldsymbol{h}(\boldsymbol{x}^*)^T\boldsymbol{y} = \boldsymbol{0}$. Therefore, the two systems $\nabla \boldsymbol{h}(\boldsymbol{x}^*)^T\boldsymbol{y} = \boldsymbol{0}$ and $[\nabla \boldsymbol{h}(\boldsymbol{x}^*)|\nabla f(\boldsymbol{x}^*)]^T\boldsymbol{y} = \boldsymbol{0}$ have the same solution set. Hence, from linear algebra we know that the matrices $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$ and $[\nabla \boldsymbol{h}(\boldsymbol{x}^*|\nabla f(\boldsymbol{x}^*)]$ share the same rank. It follows that $\nabla f(\boldsymbol{x}^*)$ must be a linear combination of columns in $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$ (otherwise, the matrix $[\nabla \boldsymbol{h}(\boldsymbol{x}^*|\nabla f(\boldsymbol{x}^*)]$ would have more linearly independent columns and hence a higher rank than the matrix $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$). This implies that $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} = \boldsymbol{0}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^m$. □

[Note: The systems of equations $\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} = \boldsymbol{0}$ and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ together give rise to $n+m$ equations in $n+m$ variables (including $\boldsymbol{x}$ and $\boldsymbol{\lambda}$), which often yield a unique solution. If *every* feasible point is a regular point of the constraints, then we can conclude that the "$\boldsymbol{x}$" in such a unique solution is the only *candidate* of local extremizer. This is how the *method of Lagrange multipliers* works in MATH2211, right? 😉]

3.1.3 Second-order necessary conditions for problems with equality constraints only.

Proposition 3.1.c. Let $f, h_1, \dots, h_m \in C^2$. Assume that $\boldsymbol{x}^*$ is a local minimizer of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, and is also a regular point of these constraints. Let $M = \{\boldsymbol{y} \in \mathbb{R}^n : \nabla \boldsymbol{h}(\boldsymbol{x}^*)^T\boldsymbol{y} = \boldsymbol{0}\}$ be the tangent plane at $\boldsymbol{x}^*$. Then, there exists $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$ such that

(a) $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} = \boldsymbol{0}$.

(b) The **Lagrangian matrix** $L(\boldsymbol{x}^*, \boldsymbol{\lambda}) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\boldsymbol{x}^*)$ is positive semi-definite on M.

[Note: A symmetric matrix A is **positive semi-definite on M** if $\boldsymbol{y}^T A\boldsymbol{y} \geq 0$ for all $\boldsymbol{y} \in M$. Similarly, a symmetric matrix A is **positive definite on M** (**negative definite on M** resp.) if $\boldsymbol{y}^T A\boldsymbol{y} > 0$ ($\boldsymbol{y}^T A\boldsymbol{y} < 0$ resp.) for all $\boldsymbol{y} \in M \setminus \{\boldsymbol{0}\}$.]

Proof. Part (a) follows directly from Proposition 3.1.b, so it remains to prove part (b).

Fix any $\boldsymbol{y} \in M$. Then, there exists a twice-differentiable curve $\boldsymbol{x}(t)$ on the surface defined by $\boldsymbol{h}(\boldsymbol{x}) = 0$ such that $\boldsymbol{x}(0) = \boldsymbol{x}^*$, $\dot{\boldsymbol{x}}(0) = \boldsymbol{y}$, and $\boldsymbol{h}(\boldsymbol{x}(t)) = \boldsymbol{0}$ for all $-a \leq t \leq a$ with $a > 0$ (similar to the proof of Lemma 3.1.a). [Note: The existence of a twice-differentiable curve on the surface can be shown under the assumptions here, but we shall omit the details.]

Let $g(t) = f(\boldsymbol{x}(t))$. Then, we have $g(h) - 2g(0) + g(-h) \geq 0$ for all h sufficiently close to 0, since $\boldsymbol{x}^*$ is a local minimizer. Applying the L'Hôpital rule, we get

$$g''(0) = \lim_{h\to 0} \frac{g'(h) - g'(-h)}{2h} = \lim_{h\to 0} \frac{g(h) - 2g(0) + g(-h)}{h^2} \geq 0.$$

Hence,

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\boldsymbol{x}(t))\right|_{t=0} = g''(0) \geq 0.$$

With $g'(t) = \nabla f(\boldsymbol{x}(t))^T\dot{\boldsymbol{x}}(t)$, by chain rule we have $g''(t) = \nabla f(\boldsymbol{x}(t))^T\ddot{\boldsymbol{x}}(t) + [\nabla^2 f(\boldsymbol{x}(t))\dot{\boldsymbol{x}}(t)]^T\dot{\boldsymbol{x}}(t) = \dot{\boldsymbol{x}}(t)\nabla^2 f(\boldsymbol{x}(t))\dot{\boldsymbol{x}}(t) + \nabla f(\boldsymbol{x}(t))^T\ddot{\boldsymbol{x}}(t)$. Putting $t = 0$, we get

$$\dot{\boldsymbol{x}}(0)^T\nabla^2 f(\boldsymbol{x}^*)\dot{\boldsymbol{x}}(0) + \nabla f(\boldsymbol{x}^*)^T\ddot{\boldsymbol{x}}(0) = g''(0) \geq 0. \tag{5}$$

With $\boldsymbol{h}(\boldsymbol{x}(t)) = \boldsymbol{0}$, we have $\sum_{i=1}^m \lambda_i h_i(\boldsymbol{x}(t)) = 0$ (where the λ_i's are taken as those satisfying (a)). Differentiating this with respect to t twice and evaluating it at $t = 0$ gives

$$\sum_{i=1}^m \lambda_i\dot{\boldsymbol{x}}(0)^T\nabla^2 h_i(\boldsymbol{x}^*)\dot{\boldsymbol{x}}(0) + \sum_{i=1}^m \lambda_i\nabla h_i(\boldsymbol{x}^*)^T\ddot{\boldsymbol{x}}(0) = 0. \tag{6}$$

Adding (5) and (6) then yields

$$\dot{\boldsymbol{x}}(0)^T\nabla^2 f(\boldsymbol{x}^*)\dot{\boldsymbol{x}}(0)+\sum_{i=1}^m \lambda_i\dot{\boldsymbol{x}}(0)^T\nabla^2 h_i(\boldsymbol{x}^*)\dot{\boldsymbol{x}}(0)+\underbrace{\left(\nabla f(\boldsymbol{x}^*)^T+\sum_{i=1}^m \lambda_i\nabla h_i(\boldsymbol{x}^*)^T\right)}_{=\boldsymbol{0}\text{ by (a)}}\ddot{\boldsymbol{x}}(0)\geq 0$$

$$\implies \dot{\boldsymbol{x}}(0)^T\left(\nabla^2 f(\boldsymbol{x}^*)+\sum_{i=1}^m \lambda_i\nabla^2 h_i(\boldsymbol{x}^*)\right)\dot{\boldsymbol{x}}(0)\geq 0$$

$$\implies \boldsymbol{y}^T L(\boldsymbol{x}^*,\boldsymbol{\lambda})\boldsymbol{y}\geq 0,$$

as desired. □

3.1.4 **Second-order sufficient conditions for problems with equality constraints only.**

Proposition 3.1.d. Let $f, h_1, \dots, h_m \in C^2$. Assume that $\boldsymbol{x}^* \in \mathbb{R}^n$ and $\boldsymbol{\lambda} \in \mathbb{R}^m$ satisfy:

(a) $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$.

(b) $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} = \boldsymbol{0}$.

(c) The Lagrangian matrix $L(\boldsymbol{x}^*, \boldsymbol{\lambda}) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\boldsymbol{x}^*)$ is positive definite on M (negative definite on M resp.), where $M = \{\boldsymbol{y} \in \mathbb{R}^n : \nabla \boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{y} = \boldsymbol{0}\}$ is the tangent plane at $\boldsymbol{x}^*$.

Then, $\boldsymbol{x}^*$ is a strict local minimizer (maximizer resp.) of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$.

Proof. Omitted. □

3.1.5 **First-order necessary conditions for general problems: Karush-Kuhn-Tucker (KKT) conditions.** After discussing the case with only equality constraints, we proceed to the general case where both equality and inequality constraints can present. Similar to the case with equality constraints only, we need the notion of *regular point*.

Let $\boldsymbol{x}^*$ be a point satisfying the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$, and let J be the set of all indices j where $g_j(\boldsymbol{x}^*) = 0$ (active constraint). Then, $\boldsymbol{x}^*$ is said to be a **regular point** of these (equality and inequality) constraints if $\nabla h_i(\boldsymbol{x}^*), \nabla g_j(\boldsymbol{x}^*),\ 1 \leq i \leq m,\ j \in J$ are linearly independent.

We are now ready to state arguably the *most important* theorem in optimization theory, giving us the famous *Karush-Kuhn-Tucker (KKT) conditions*, which are first-order necessary conditions for constrained minimization in the standard form:

Theorem 3.1.e (Karush-Kuhn-Tucker (KKT) conditions)**.** Let $f, g_1, \dots, g_q, h_1, \dots, h_m \in C^1$. Assume that $\boldsymbol{x}^*$ is a local minimizer of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$, and is also a regular point of these constraints *(constraint qualification)*. Then, there exist $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_q) \in \mathbb{R}^q$ (known as **Lagrange multipliers**) such that:

(a) *(Stationarity)* $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} + \nabla \boldsymbol{g}(\boldsymbol{x}^*)\boldsymbol{\mu} = \boldsymbol{0}$.

(b) *(Primal feasibility)* $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}$.

(c) *(Dual feasibility)* $\boldsymbol{\mu} \geq \boldsymbol{0}$.

(d) *(Complementary slackness)* $\mu_j g_j(\boldsymbol{x}^*) = 0$ for all $j = 1, \dots, q$.

Remarks:

- The requirement that $\boldsymbol{x}^*$ is a regular point of the equality and inequality constraints is referred to as the *constraint qualification* since it impose conditions on the constraints that qualify $\boldsymbol{x}^*$ as a local minimizer that fulfills the necessary conditions (KKT conditions).
- The *complementary slackness* condition can be interpreted as follows. For each $j = 1, \dots, q$, there is a "slackness" available for *either* μ_j *or* $g_j(\boldsymbol{x}^*)$ (so complementary); here "slackness" means not attaining the equality for their respective inequality constraints ($g_j(\boldsymbol{x}^*) \leq 0$ and $\mu_j \geq 0$), i.e., not being zero. Hence, $\mu_j g_j(\boldsymbol{x}^*) = 0$ ensures that *at least one* of μ_j and $g_j(\boldsymbol{x}^*)$ is zero (does not "slack").

- The conditions $\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}$, $\boldsymbol{g}(\boldsymbol{x}^*) \le \mathbf{0}$ and $\boldsymbol{\mu} \ge \mathbf{0}$ are called *primal feasibility* and *dual feasibility*, respectively, because it turns out that the constrained minimization problem in consideration here can be seen as a "primal" problem, and $\boldsymbol{\mu}$ is a variable for another optimization problem known as the "dual" problem, where the feasibility requirement is $\boldsymbol{\mu} \ge \mathbf{0}$. These will become clearer in Section 5.
- The function $\ell(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}$ is called the **Lagrangian** associated with the constrained minimization problem. Using Lagrangian, we can express the stationarity condition as $\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$ *(Lagrangian derivative condition)*, where $\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ denotes the gradient of ℓ with respect to $\boldsymbol{x}$ (i.e., treating ℓ as a function of $\boldsymbol{x}$): $\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla f(\boldsymbol{x}) + \nabla \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} + \nabla \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}$. In the special case where there are only equality constraints in the constrained minimization problem in consideration, the Lagrangian then becomes $\ell(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda}$.

Proof. Let J be the set of indices j where $g_j(\boldsymbol{x}^*) = 0$. Note that $\boldsymbol{x}^*$ is also a local minimizer of the problem

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}, \\ & g_j(\boldsymbol{x}) = 0 \quad \forall j \in J, \end{aligned}$$

because the feasibility of $\boldsymbol{x}^*$ ensures that $g_j(\boldsymbol{x}^*) < 0$ for every $j \notin J$, and the continuity of those g_j's ensure that $g_j(\boldsymbol{x}) < 0$ for every $j \notin J$, for all $\boldsymbol{x}$ sufficiently close to $\boldsymbol{x}^*$. Hence, within that small neighbourhood, the feasible points of the problem above coincide with those of the original problem.

Now, by Proposition 3.1.b, there exist $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\mu_j,\ j \in J$ such that

$$\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} + \sum_{j \in J} \mu_j \nabla g_j(\boldsymbol{x}^*) = \mathbf{0}. \tag{7}$$

Letting $\mu_j = 0$ for all $j \in \{1, \dots, q\} \setminus J$, we can then obtain:

$$\begin{aligned} \nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} + \nabla \boldsymbol{g}(\boldsymbol{x}^*)\boldsymbol{\mu} = \mathbf{0}, \\ \mu_j g_j(\boldsymbol{x}^*) = 0 \text{ for all } j = 1, \dots, q. \end{aligned}$$

This establishes (a) and (d). Part (b) is immediate by the feasibility of $\boldsymbol{x}^*$. Hence, it remains to show part (c). With $\mu_j = 0$ for all $j \notin J$, it suffices to show $\mu_j \ge 0$ for all $j \in J$.

Assume to the contrary that $\mu_k < 0$ for some $k \in J$. Let S be the surface defined by $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$, $g_j(\boldsymbol{x}) = 0$ for all $j \in J \setminus \{k\}$, and let M be the set $\{\boldsymbol{y} \in \mathbb{R}^n : \nabla \boldsymbol{h}(\boldsymbol{x}^*)^T\boldsymbol{y} = \mathbf{0},\ \nabla g_j(\boldsymbol{x}^*)^T\boldsymbol{y} = 0\ \forall j \in J \setminus \{k\}\}$. By the constraint qualification ($\boldsymbol{x}^*$ is a regular point of the constraints), there exists $\boldsymbol{d} \in M$ such that $\nabla g_k(\boldsymbol{x}^*)^T\boldsymbol{d} < 0$. With $\mu_k < 0$, by (7) we then have $\nabla f(\boldsymbol{x}^*)^T\boldsymbol{d} < 0$.

Now, let $\boldsymbol{x}(t)$ be a differentiable curve on the surface S continuously parameterized by $t \in [-a, a]$ with $a > 0$, $\boldsymbol{x}(0) = \boldsymbol{x}^*$, and $\dot{\boldsymbol{x}}(0) = \boldsymbol{d}$. Then, for all t sufficiently close to 0, we know that $g_k(\boldsymbol{x}(t)) \le g_k(\boldsymbol{x}(0)) = g_k(\boldsymbol{x}^*) = 0$ since $\mathrm{d}/\mathrm{d}t\, g_k(\boldsymbol{x}(t))|_{t=0} = \nabla g_k(\boldsymbol{x}^*)^T\boldsymbol{d} < 0$, so $\boldsymbol{x}(t)$ is a feasible point of the original problem. However, we have

$$\left.\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}(t))\right|_{t=0} = \nabla f(\boldsymbol{x}^*)^T\boldsymbol{d} < 0,$$

which contradicts the local minimality of $\boldsymbol{x}(0) = \boldsymbol{x}^*$. □

3.1.6 Second-order necessary conditions for general problems.

Proposition 3.1.f. Let $f, g_1, \dots, g_q, h_1, \dots, h_m \in C^2$. Assume that $\boldsymbol{x}^*$ is a local minimizer of f subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \le \mathbf{0}$, and is also a regular point of these constraints *(constraint qualification)*. Then, there exist $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_q) \in \mathbb{R}^q$ such that:

(a) *(Stationarity)* $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} + \nabla \boldsymbol{g}(\boldsymbol{x}^*)\boldsymbol{\mu} = \mathbf{0}$.

(b) *(Primal feasibility)* $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}$.

(c) *(Dual feasibility)* $\boldsymbol{\mu} \geq \boldsymbol{0}$.

(d) *(Complementary slackness)* $\mu_j g_j(\boldsymbol{x}^*) = 0$ for all $j = 1, \dots, q$.

(e) The **Lagrangian matrix** $L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\boldsymbol{x}^*) + \sum_{j=1}^q \mu_j \nabla^2 g_j(\boldsymbol{x}^*)$ is positive semi-definite on $M = \{\boldsymbol{y} \in \mathbb{R}^n : \nabla \boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{y} = \boldsymbol{0},\ \nabla g_j(\boldsymbol{x}^*)^T \boldsymbol{y} = 0\ \forall j \in J\}$ where J is the set of all indices j where $g_j(\boldsymbol{x}^*) = 0$ *(the tangent subspace of the active constraints at $\boldsymbol{x}^*$).*

Proof. Parts (a)-(d) directly follow from Theorem 3.1.e, so it remains to show part (e).

Let J be the set of indices j where $g_j(\boldsymbol{x}^*) = 0$. Similar to the proof of Theorem 3.1.e, $\boldsymbol{x}^*$ is also a local minimizer of the problem

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \\ & g_j(\boldsymbol{x}) = 0 \quad \forall j \in J. \end{aligned}$$

With $\mu_j = 0$ for all $j \in \{1, \dots, q\} \setminus J$, the Lagrangian matrix can be written as $L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\boldsymbol{x}^*) + \sum_{j \in J} \mu_j \nabla^2 g_j(\boldsymbol{x}^*)$. Part (e) can then be shown by following the same line of argument as the proof of Proposition 3.1.c for this equality constrained problem. □

3.1.7 **Second-order sufficient conditions for general problems.**

Proposition 3.1.g. Let $f, g_1, \dots, g_q, h_1, \dots, h_m \in C^2$. Assume that $\boldsymbol{x}^* \in \mathbb{R}^n$, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$, and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_q) \in \mathbb{R}^q$ satisfy:

(a) *(Stationarity)* $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} + \nabla \boldsymbol{g}(\boldsymbol{x}^*)\boldsymbol{\mu} = \boldsymbol{0}$.

(b) *(Primal feasibility)* $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}$.

(c) *(Dual feasibility)* $\boldsymbol{\mu} \geq \boldsymbol{0}$.

(d) *(Complementary slackness)* $\mu_j g_j(\boldsymbol{x}^*) = 0$ for all $j = 1, \dots, q$.

(e) The Lagrangian matrix $L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\boldsymbol{x}^*) + \sum_{j=1}^q \mu_j \nabla^2 g_j(\boldsymbol{x}^*)$ is positive definite on $M' = \{\boldsymbol{y} \in \mathbb{R}^n : \nabla \boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{y} = \boldsymbol{0},\ \nabla g_j(\boldsymbol{x}^*)^T \boldsymbol{y} = 0\ \forall j \in J'\}$ where J' is the set of all indices j where $g_j(\boldsymbol{x}^*) = 0$ and $\mu_j > 0$.

Then, $\boldsymbol{x}^*$ is a strict local minimizer of $f(\boldsymbol{x})$ subject to the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$.

Proof. Omitted. □

3.2 Lagrange Methods

3.2.1 In Section 3.1, we have discussed numerous necessary and sufficient conditions for local extremizers of constrained optimization problem in the standard form. We can also see that many terms involved in those conditions are associated with *Lagrange* (Lagrange multiplier, Lagrangian, Lagrangian matrix). Hence, solution methods for constrained optimization based on those conditions are also known as **Lagrange methods**. The development of the Lagrange methods follows a similar logic as the procedure mentioned in [1.2.1].

3.2.2 **General procedure for Lagrange methods on constrained minimization problems in the standard form.** We have considered two types of constrained minimization problems in the standard form:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \end{aligned}$$

and

$$\begin{array}{rl}\text{Minimize} & f(\boldsymbol{x})\\ \text{subject to} & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0},\\ & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}.\end{array}$$

The following steps can help us find *local* minimizers:

(1) *(Identifying candidates of local minimizer via necessary conditions)* Using first-order and/or second-order necessary conditions for *regular* feasible points, namely Propositions 3.1.b and 3.1.c for equality constrained problem, or Theorem 3.1.e and Proposition 3.1.f for general problem, we exclude all *regular* feasible points that are impossible to be local minimizers, and the ones obtained from solving the systems are some candidates of local minimizer (no matter whether they are regular points or not; those that are regular are of course candidates, but those that are not regular are also candidates since they are not excluded).

(2) *(Checking whether each candidate is a local minimizer via sufficient conditions)* For each candidate of local minimizer, we can examine the second-order sufficient condition (Proposition 3.1.d for equality constrained problem, or Proposition 3.1.g for general problem) to check whether it is fulfilled (particularly, we do *not* need to check whether the candidate is a regular point of constraints here). If it is the case, then we have shown that the candidate is a local minimizer. If not, then we still do not know whether the candidate is a local minimizer, and other techniques are needed to determine whether it is a local minimizer or not (e.g., by definition).

In order to go from *local* to *global*, there are two common strategies:

- *Strategy 1: Appealing to the convexity.* If the feasible region formed by the constraints is convex and f is convex, then any local minimizer found above is a global minimizer.
- *Strategy 2: Utilizing the extreme value theorem.* If the feasible region formed by the constraints is closed and bounded, then by the extreme value theorem there must exist a global minimizer (recall that f is assumed to be continuous throughout).

  If every feasible point is a regular point of the constraints, then a global minimizer can be found by comparing the objective function values of all the candidates obtained in the above process (if feasible). This is because when every feasible point is regular, we know that all local minimizers must satisfy the necessary conditions, and thus solutions to the systems already cover all candidates of local minimizers.

3.2.3 **An example of applying Lagrange method.** Usually, the major difficulty for following the procedure in [3.2.2] lies in solving the (often complex) system of equations from the necessary conditions to obtain candidates of regular local minimizers. This is illustrated by the following example.

Suppose that we would like to solve the problem

$$\begin{array}{rl}\text{Minimize} & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2\\ \text{subject to} & x_1^2 + x_2^2 \leq 5,\\ & 3x_1 + x_2 \leq 6\end{array}$$

by a Lagrange method.

In accordance with [3.2.2], the steps are as follows.

(1) *(Converting the problem to the standard form)* We first rewrite the problem in the standard form:

$$\begin{array}{rl}\text{Minimize} & f(x_1, x_2) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2\\ \text{subject to} & g_1(x_1, x_2) = x_1^2 + x_2^2 - 5 \leq 0,\\ & g_2(x_1, x_2) = 3x_1 + x_2 - 6 \leq 0.\end{array}$$

[⚠ Warning: Make sure that you don't miss this step!]

(2) *(Stating the KKT conditions)* We compute:

$$\begin{aligned}\nabla f(x_1,x_2) &= \begin{bmatrix}4x_1+2x_2-10\\2x_1+2x_2-10\end{bmatrix},\\ \nabla \boldsymbol{g}(x_1,x_2) &= \begin{bmatrix}\nabla g_1(x_1,x_2) & \nabla g_2(x_1,x_2)\end{bmatrix} = \begin{bmatrix}2x_1 & 3\\2x_2 & 1\end{bmatrix}.\end{aligned}$$

Hence, the KKT conditions are:

$$4x_1+2x_2-10+2\mu_1x_1+3\mu_2=0 \tag{8}$$
$$2x_1+2x_2-10+2\mu_1x_2+\mu_2=0 \tag{9}$$
$$x_1^2+x_2^2-5\le 0 \tag{10}$$
$$3x_1+x_2-6\le 0 \tag{11}$$
$$\mu_1\ge 0, \mu_2\ge 0 \tag{12}$$
$$\mu_1(x_1^2+x_2^2-5)=0 \tag{13}$$
$$\mu_2(3x_1+x_2-6)=0 \tag{14}$$

(3) *(Solving the system from the KKT conditions)* As we can see, the system arising from the KKT conditions are rather complex and hard to solve directly. A *tip* 💡 for dealing with such a complex system is to perform a *case-by-case analysis* by *trying various combination of active constraints*, so that the system can be simplified in each case, and hopefully some solutions can be obtained in this way.

In this context, we consider 4 cases:

Case Constraint	1	2	3	4
$g_1(x_1,x_2)\le 0$	Inactive	Active	Inactive	Active
$g_2(x_1,x_2)\le 0$	Inactive	Inactive	Active	Active

Case 1: Both constraints are inactive. Then, from (13) and (14), we must have $\mu_1=\mu_2=0$. Plugging these into (8) and (9) yields

$$\begin{cases}4x_1+2x_2=10\\2x_1+2x_2=10\end{cases} \implies (x_1,x_2)=(0,5).$$

However, we have $g_1(0,5)=25>5$, violating (10). Hence, we conclude that there is *no solution* in this case.

Case 2: (10) *is active while* (11) *is inactive.* Then, from (14), we must have $\mu_2=0$. Plugging this into (8) and (9), together with the activeness of (10), yields

$$4x_1+2x_2-10+2\mu_1x_1=0 \tag{15}$$
$$2x_1+2x_2-10+2\mu_1x_2=0 \tag{16}$$
$$x_1^2+x_2^2=5 \tag{17}$$

[Note: It is still quite challenging to solve this system, even if it is already simplified. Unfortunately, there is not a fixed "recipe" to deal with this kind of system, apart from the general idea that we should try to relate the variables x_1, x_2, and μ_1, and plug some expressions into some equations.]

From (15) we get $x_2=5-(2+\mu_1)x_1$. Plugging it into (16) yields

$$\begin{aligned}2x_1-10+(2+2\mu_1)(5-(2+\mu_1)x_1)&=0\\ \implies [1-(1+\mu_1)(2+\mu_1)]x_1&=5-5(1+\mu_1)=5\mu_1\\ \implies x_1&=\frac{5\mu_1}{\mu_1^2+3\mu_1+1}.\end{aligned}$$

Hence, from $x_2 = 5 - (2 + \mu_1)x_1$ we get

$$x_2 = \frac{5\mu_1^2 + 15\mu_1 + 5 - 10\mu_1 - 5\mu_1^2}{\mu_1^2 + 3\mu_1 + 1} = \frac{5(1 + \mu_1)}{\mu_1^2 + 3\mu_1 + 1}.$$

With x_1 and x_2 both expressed in terms of a common variable μ_1, we can plug these expressions into (17) to solve for μ_1:

$$\begin{aligned}
\left(\frac{5\mu_1}{\mu_1^2 + 3\mu_1 + 1}\right)^2 + \left(\frac{5(1 + \mu_1)}{\mu_1^2 + 3\mu_1 + 1}\right)^2 &= 5 \\
\implies 25\mu_1^2 + 25(1 + \mu_1)^2 &= 5(\mu_1^2 + 3\mu_1 + 1)^2 \\
\implies 5\mu_1^2 + 5 + 10\mu_1 + 5\mu_1^2 &= \mu_1^4 + 2\mu_1^2(3\mu_1 + 1) + (3\mu_1 + 1)^2 \\
\implies \mu_1^4 + 6\mu_1^3 + \mu_1^2 - 4\mu_1 - 4 &= 0 \\
\implies (\mu_1 - 1)\underbrace{(\mu_1^3 + 7\mu_1^2 + 8\mu_1 + 4)}_{\geq 4 \text{ as } \mu_1 \geq 0 \text{ from (12)}} &= 0 \\
\implies \mu_1 &= 1.
\end{aligned}$$

It follows that

$$\begin{aligned}
x_1 &= \frac{5\mu_1}{\mu_1^2 + 3\mu_1 + 1} = 1, \\
x_2 &= \frac{5(1 + \mu_1)}{\mu_1^2 + 3\mu_1 + 1} = 2.
\end{aligned}$$

It remains to check whether $(x_1, x_2, \mu_1) = (1, 2, 1)$ also satisfies (11). With $3(1) + 2 - 6 = -1 \leq 0$, (11) is indeed satisfied. Hence, $(x_1, x_2, \mu_1) = (1, 2, 1)$ is the solution in this case, meaning that $\boldsymbol{x}^* = (1, 2)$ is a candidate of local minimizer.

After obtaining *a* candidate of local minimizer, we can stop and proceed to later steps to check whether it is indeed a local minimizer. If it turns out that it is not a local minimizer, we can go back to here to consider the remaining cases.

(4) *(Verifying that the candidate is a local minimizer)* Now consider the second-order sufficient conditions (Proposition 3.1.g). For $(x_1, x_2, \mu_1, \mu_2) = (1, 2, 1, 0)$, the process of obtaining it already ensures that the KKT conditions (first 4 conditions) are satisfied. Hence, it remains to check the condition about the Lagrangian matrix.

First compute:

$$\begin{aligned}
\nabla^2 f(x_1, x_2) &= \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}, \\
\nabla^2 g_1(x_1, x_2) &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.
\end{aligned}$$

Thus, at $(x_1, x_2, \mu_1, \mu_2) = (1, 2, 1, 0)$, the Lagrangian matrix is

$$L(x_1 = 1, x_2 = 2, \mu_1 = 1, \mu_2 = 0) = \nabla^2 f(1, 2) + (1)\nabla^2 g_1(1, 2) = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 2 & 4 \end{bmatrix},$$

which is positive definite by [1.1.13], since its $(1, 1)$th entry and determinant are both positive. Hence, by Proposition 3.1.g, $\boldsymbol{x}^* = (1, 2)$ is a local minimizer.

(5) *(Showing that $\boldsymbol{x}^*$ is a global minimizer by appealing to the convexity)* With

$$\nabla^2 g_1(x_1, x_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad \nabla^2 g_2(x_1, x_2) = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

both being positive semi-definite throughout $\mathbb{R}^2$, g_1 and g_2 are both convex on $\mathbb{R}^2$. Hence, by Proposition 1.2.e and the fact that intersection of convex sets is again convex (it is straightforward to show this by definition), we can see that the feasible region $\Omega = \{\boldsymbol{x} \in \mathbb{R}^2 : g_1(\boldsymbol{x}) \leq 0, g_2(\boldsymbol{x}) \leq 0\} = \{\boldsymbol{x} \in \mathbb{R}^2 : g_1(\boldsymbol{x}) \leq 0\} \cap \{\boldsymbol{x} \in \mathbb{R}^2 : g_2(\boldsymbol{x}) \leq 0\}$ is convex. Furthermore, noting that $\nabla^2 f(x_1, x_2)$ is positive semi-definite throughout $\mathbb{R}^2$, f is convex on Ω. Therefore, by Theorem 1.2.f, the local minimizer $\boldsymbol{x}^* = (1, 2)$ is a global minimizer.

4 Penalty and Barrier Methods

4.0.1 To solve constrained optimization problems, apart from using *Lagrange methods* discussed in Section 3.2, there are two more common approaches: *penalty methods* and *barrier methods*. Both work by *approximating* constrained optimization problems by *unconstrained* problems, but the approximation method differs for the two methods. For *penalty methods*, the approximation is achieved by adding to the objective function a term that prescribes a high cost for violation of the constraints *(penalty)*. On the other hand, for *barrier methods*, it is achieved by adding a term that *favors* points in the more "inner" part of the feasible region over those near its boundary.

4.1 Penalty Methods

4.1.1 **Definition.** Consider the following minimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \Omega, \end{aligned} \tag{18}$$

where f is continuous on $\mathbb{R}^n$, and $\Omega \subseteq \mathbb{R}^n$ is the feasible region. A **penalty method** works as follows. We first replace this problem by an unconstrained minimization problem of the form

$$\text{Minimize} \quad \phi_c(\boldsymbol{x}) = f(\boldsymbol{x}) + cp(\boldsymbol{x})$$

where $c > 0$ is a constant, and p is a function on $\mathbb{R}^n$ satisfying:

(a) *(Continuity)* p is continuous.

(b) *(Nonnegativity)* $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

(c) *(No penalty for feasible points only)* $p(\boldsymbol{x}) = 0$ iff $\boldsymbol{x} \in \Omega$.

This unconstrained problem is called a **penalty problem**, p is called a **penalty function**, and c is called a **penalty factor** or **penalty coefficient**.

Now, let $\{c_k\}$ be a sequence tending to ∞ with $c_k > 0$ and $c_k < c_{k+1}$ for all $k \in \mathbb{N}$. Then, for each $k \in \mathbb{N}$, we solve the unconstrained problem

$$\text{Minimize} \quad \phi_{c_k}(\boldsymbol{x})$$

to obtain a minimizer $\boldsymbol{x}_k$ (by methods discussed in Sections 1 and 2, for instance).

Intuitively, as the penalty coefficient c gets larger, the minimizer of the penalty problem should lie in a region where $p(\boldsymbol{x})$ is small and close to 0; as c tends to infinity, the penalty becomes increasingly "severe" and we expect that the corresponding minimizers should approach to a minimizer of the original problem, which falls in the feasible region and incurs zero penalty.

4.1.2 **Theoretical guarantee.** The following is a theoretical result which ensures that any sequence $\{\boldsymbol{x}_k\}$ generated by a penalty method indeed converges to a minimizer $\boldsymbol{x}^*$ of the original problem, meaning that a minimizer $\boldsymbol{x}_c$ obtained from solving a penalty problem with large c can approximate $\boldsymbol{x}^*$ fairly well.

Theorem 4.1.a. Let $\{\boldsymbol{x}_k\}$ be a sequence of minimizers for penalty problems generated by a penalty method. Then, every limit point of the sequence $\{\boldsymbol{x}_k\}$ is a minimizer to the original problem (18).

Proof. Omitted. □

[Note: A point $\boldsymbol{x}^*$ is a **limit point** of a sequence $\{\boldsymbol{x}_k\}$ if there exists a subsequence $\{\boldsymbol{x}_{k_i}\}$ of $\{\boldsymbol{x}_k\}$ that converges to $\boldsymbol{x}^*$ as $i \to \infty$. So, particularly, if the sequence $\{\boldsymbol{x}_k\}$ itself converges to some point $\boldsymbol{x}^*$, then that point is the only limit point of $\{\boldsymbol{x}_k\}$ (since in that case every subsequence of $\{\boldsymbol{x}_k\}$ must converge to $\boldsymbol{x}^*$).]

4.1.3 **The penalty function for constrained minimization problem in the standard form.** While there are many possible choices of penalty function from the definition of penalty method, to solve a constrained minimization problem in the standard form:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_1(\boldsymbol{x}) = 0, \quad h_2(\boldsymbol{x}) = 0, \quad \ldots, \quad h_m(\boldsymbol{x}) = 0, \\ & g_1(\boldsymbol{x}) \leq 0, \quad g_2(\boldsymbol{x}) \leq 0, \quad \ldots, \quad g_q(\boldsymbol{x}) \leq 0, \end{aligned}$$

it is typical to use the following penalty function:

$$\boxed{p(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{m} h_i(\boldsymbol{x})^2 + \frac{1}{2}\sum_{j=1}^{q} \max\{0, g_j(\boldsymbol{x})\}^2}.$$

Remarks:

- In MATH3904, we almost always use this penalty function, with "1/2" dropped sometimes (this does not affect its validity as a penalty function).
- The maximum $\max\{0, g_j(\boldsymbol{x})\}$ is sometimes known as the *positive part* of $g_j(\boldsymbol{x})$ and denoted by $g_j^+(\boldsymbol{x})$.

[Intuition 💡: The size of each $h_i(\boldsymbol{x})^2$ indicates the "severity" of violation of the constraint $h_i(\boldsymbol{x}) = 0$ (the larger the size, the more severe the violation is), and we use $\max\{0, g_j(\boldsymbol{x})\}^2$ to indicate the "severity" of the violation of the constraint $g_j(\boldsymbol{x}) \leq 0$ because only the positive part of $g_j(\boldsymbol{x})$ is relevant to the violation.]

4.1.4 **Some helper results for solving penalty problems.** To solve constrained minimization problem in the standard form by the penalty method with the penalty function in [4.1.3], we need to solve unconstrained problem of the form

$$\text{Minimize} \quad \phi_c(\boldsymbol{x}) = f(\boldsymbol{x}) + c\left(\frac{1}{2}\sum_{i=1}^{m} h_i(\boldsymbol{x})^2 + \frac{1}{2}\sum_{j=1}^{q} \max\{0, g_j(\boldsymbol{x})\}^2\right).$$

To this end, the following results can be helpful.

(a) *(Peressini et al., 1993, Lemma 6.1.3)* Let $g \in C^1$ be a function on $\mathbb{R}^n$. Then, $h(\boldsymbol{x}) = \max\{g(\boldsymbol{x}), 0\}^2$ is also in C^1, and we have

$$\frac{\partial h}{\partial x_i}(\boldsymbol{x}) = 2\max\{g(\boldsymbol{x}), 0\}\frac{\partial g}{\partial x_i}(\boldsymbol{x}) \quad \text{for all } i = 1, \ldots, n,$$

for every $\boldsymbol{x} \in \mathbb{R}^n$ *("like" chain rule).*

(b) If f, h_i's, and g_j's are all convex with h_i's being all nonnegative, then $\phi_c(\boldsymbol{x})$ is convex over $\mathbb{R}^n$ for every $c > 0$.

Proof.

(a) Omitted; see Peressini et al. (1993).

(b) By Proposition 1.2.e, we know that each $h_i(\boldsymbol{x})^2$ is convex since h_i is nonnegative and convex, and each $\max\{0, g_j(\boldsymbol{x})\}^2$ is convex since $\max\{0, g_j(\boldsymbol{x})\}$ is also nonnegative and convex. Since $c > 0$ and f is convex, by Proposition 1.2.e we know that ϕ_c is convex.

□

4.1.5 **Typical steps for penalty method.** Now, let us collect the previous ingredients to put forward typical steps involved when applying penalty method to solve constrained minimization problem in the standard form.

To solve the following constrained minimization problem in the standard form

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_1(\boldsymbol{x}) = 0, \quad h_2(\boldsymbol{x}) = 0, \quad \dots, \quad h_m(\boldsymbol{x}) = 0, \\ & g_1(\boldsymbol{x}) \leq 0, \quad g_2(\boldsymbol{x}) \leq 0, \quad \dots, \quad g_q(\boldsymbol{x}) \leq 0, \end{aligned}$$

by a *penalty method*, we can follow the steps below.

(1) *(Stating the penalty problem)* The penalty problem is

$$\text{Minimize} \quad \phi_c(\boldsymbol{x}) = f(\boldsymbol{x}) + c\left(\frac{1}{2}\sum_{i=1}^{m} h_i(\boldsymbol{x})^2 + \frac{1}{2}\sum_{j=1}^{q} \max\{0, g_j(\boldsymbol{x})\}^2\right).$$

(2) *(Establishing the convexity of ϕ_c)* Assuming that f, h_i's, and g_j's are all convex with h_i's being all nonnegative, we know that ϕ_c is convex on $\mathbb{R}^n$ by [4.1.4]b.

(3) *(Finding global minimizer by solving $\nabla\phi_c(\boldsymbol{x}) = \boldsymbol{0}$)* Due to the convexity of ϕ_c, any solution to the system $\nabla\phi_c(\boldsymbol{x}) = \boldsymbol{0}$ is a global minimizer; to evaluate $\nabla\phi_c(\boldsymbol{x})$, the result in [4.1.4]a is often helpful, and to deal with the maximums $\max\{g_j(\boldsymbol{x}), 0\}$ involved in the system, we may use a case-by-case analysis like what was done in [3.2.3], by considering whether $g_j(\boldsymbol{x}) \leq 0$ or $g_j(\boldsymbol{x}) > 0$ for each $j = 1, \dots, q$.

(4) *(Letting $c \to \infty$ for minimizer of the penalty problem)* After obtaining an expression of minimizer $\boldsymbol{x}_c$ of the penalty problem for all $c > 0$, we let $c \to \infty$ to get a minimizer $\boldsymbol{x}^*$ of the constrained minimization problem (assuming the existence of limit).

4.2 Barrier Methods

4.2.1 **Definition.** Consider the following minimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \Omega, \end{aligned} \tag{19}$$

where f is continuous on $\mathbb{R}^n$, and $\Omega \subseteq \mathbb{R}^n$ is the feasible region. We assume that Ω is **robust**, i.e., it has a nonempty interior that is arbitrarily close to any boundary point of Ω, which means that:

(a) The interior of Ω (the set of all interior points of Ω) is nonempty.

(b) For every boundary point $\boldsymbol{x}$ of Ω, there exists a sequence $\{\boldsymbol{y}_n\}$ of interior points of Ω that converges to $\boldsymbol{x}$.

A **barrier method** works as follows. We first replace the minimization problem (19) by a minimization problem of the form

$$\begin{aligned} \text{Minimize} \quad & \phi_c(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{c}B(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \text{the interior of } \Omega, \end{aligned}$$

where $c > 0$ is a constant and B is a function on the *interior* of Ω satisfying that:

(a) *(Continuity)* B is continuous.

(b) *(Nonnegativity)* $B(\boldsymbol{x}) \geq 0$.

(c) *("Barrier" for being away from the "inner" part)* $B(\boldsymbol{x}) \to \infty$ as $\boldsymbol{x}$ approaches the boundary of Ω (i.e., approaching an arbitrary boundary point of Ω). [Note: The robustness of Ω ensures that it is possible for $\boldsymbol{x}$ to approach the boundary of Ω while being in the interior of Ω (the domain of B).]

This minimization problem is called a **barrier problem**, and B is called a **barrier function**.

Now, let $\{c_k\}$ be a sequence tending to ∞ with $c_k > 0$ and $c_k < c_{k+1}$ for all $k \in \mathbb{N}$. Then, for each $k \in \mathbb{N}$, we solve the minimization problem

$$\begin{aligned} \text{Minimize} \quad & \phi_{c_k}(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \text{the interior of } \Omega \end{aligned}$$

to obtain a minimizer $\boldsymbol{x}_k$.

[Note: Although a barrier problem has a constraint that $\boldsymbol{x} \in$ the interior of Ω, it can indeed be solved using algorithms for *unconstrained* optimization. To find a minimizer, we can start at an initial interior point and then searches from the point using the *steepest descent method* or some other methods. Since the objective function value approaches ∞ near the boundary of Ω, the algorithm will automatically remain within the interior of Ω, so the constraint needs not be accounted for explicitly during the algorithm implementation. Hence, from a *computational* viewpoint a barrier problem is unconstrained, but it is still constrained from a mathematical viewpoint.]

Intuitively, as the value c gets larger, within the interior of Ω, the expression $B(\boldsymbol{x})/c$ becomes small and thus the objective function $\phi_c(\boldsymbol{x})$ becomes close to the original objective function $f(\boldsymbol{x})$. With the continuity of f and the robustness of Ω, we then expect that the minimizers obtained from solving barrier problems with increasing values of c should approach to a minimizer of the original problem, even if it is a boundary point of Ω.

4.2.2 **Theoretical guarantee.** Like the penalty method, any sequence $\{\boldsymbol{x}_k\}$ generated by a barrier method indeed converges to a minimizer $\boldsymbol{x}^*$ of the original problem.

Theorem 4.2.a. Let $\{\boldsymbol{x}_k\}$ be a sequence of minimizers for penalty problems generated by a barrier method. Then, every limit point of the sequence $\{\boldsymbol{x}_k\}$ is a minimizer to the original problem (19).

Proof. Omitted. □

4.2.3 **Barrier functions for minimization problem with inequality constraints only.** In MATH3904, the feasible region Ω here often takes the form $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : g_j(\boldsymbol{x}) \leq 0 \ \forall j = 1, \dots, q\}$, i.e., we are considering the following minimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_1(\boldsymbol{x}) \leq 0, \quad g_2(\boldsymbol{x}) \leq 0, \quad \dots, \quad g_q(\boldsymbol{x}) \leq 0. \end{aligned}$$

To solve this problem by a barrier method, the following two barrier functions are commonly used:

(a) *(Reciprocal barrier function)* $B(\boldsymbol{x}) = -\sum_{j=1}^{q} 1/g_j(\boldsymbol{x})$.

(b) *(Frisch's logarithmic barrier function)* $B(\boldsymbol{x}) = -\sum_{j=1}^{q} \ln(-g_j(\boldsymbol{x}))$.

Remarks:

- It can be verified that the reciprocal barrier function satisfies all three conditions for a being a barrier function, and the Frisch's logarithmic barrier function satisfies two of them, with the nonnegativity potentially being violated when the $g_j(\boldsymbol{x})$'s are "too negative". Although the Frisch's logarithmic barrier function is technically not a barrier function according to the definition above, it can be shown that using such a function B also leads to the convergence property in Theorem 4.2.a (see Bazaraa et al. (2006, p. 502)). Hence, it effectively still functions as a barrier function, and so it is common to still call it "barrier function" despite the potential violation of nonnegativity.
- When both barrier functions are applicable for solving the problem by a barrier method, it is recommended to choose the *Frisch's logarithmic barrier function* since during the process we would need to compute the gradient of B, and the reciprocal barrier function would yield some degree-2 terms while the Frisch's logarithmic barrier function would only yield some degree-1 terms (often simpler to deal with).

The steps needed for solving the minimization problem by a barrier method are similar to the ones for penalty method in [4.1.5]; we just need to remember that there is a constraint "$\boldsymbol{x} \in$ the interior of Ω" when solving the barrier problem ⚠.

[Note: When both *penalty method* and *barrier method* are applicable for solving an optimization problem, it may be more convenient to use the *barrier method* since we can avoid dealing with the "max" which appears only in the penalty method.]

5 Duality Theory

5.0.1 *Duality* has a central role in the development and unification of the optimization theory. Duality methods are based on the viewpoint that for a constrained minimization problem in the standard form, the *Lagrange multipliers* $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ serve as the fundamental unknowns; once we know $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, determining the corresponding solution point $\boldsymbol{x}$ is straightforward (in some cases at least). Based on this idea, duality methods do not attack the original constrained problem directly, but rather attack an *alternate* problem known as the *dual* problem, whose unknowns are the *Lagrange multipliers* $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ of the original problem.

5.1 Equality Constrained Minimization

5.1.1 **Primal and dual problems.** We start with the simpler case where the constrained minimization problem in consideration has only equality constraints (similar to the approach in Section 3). Consider the following constrained minimization problem

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0} \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $f, h_1, \dots, h_m \in C^2$. This is known as the **primal problem** in duality theory, denoted by (P). In the literature, several *dual problems* closely related to this primal problem have been proposed. Among them, the *Lagrangian duality* formulation is considered to be the simplest and have the most "elegant" form, and has perhaps attracted the most attention. So, in Section 5 we will exclusively focus on Lagrangian duality.

The **(Lagrangian) dual problem** of (P), denoted by (D), is given by the following unconstrained *maximization* problem

$$\begin{aligned} \text{Maximize} \quad & \phi(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \in \mathbb{R}^m \end{aligned}$$

where $\phi(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x}\in\mathbb{R}^n} \ell(\boldsymbol{x}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{x}\in\mathbb{R}^n} [f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda}]$ is the **dual function**, i.e., for every fixed $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\phi(\boldsymbol{\lambda})$ equals the minimum value of the following unconstrained minimization problem

$$\begin{aligned} \text{Minimize} \quad & \ell(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} \\ \text{subject to} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

if a minimizer exists, and equals $-\infty$ otherwise (meaning that the objective function is not bounded below on $\mathbb{R}^n$). [Note: Here, $\ell(\boldsymbol{x}, \boldsymbol{\lambda})$ is the *Lagrangian* associated with the primal problem (P).]

5.1.2 **Local duality theorem.** The connection between the primal problem (P) and the Lagrangian dual problem (D) is provided by the following *local duality theorem*, relating *local* minimizer of (P) and *local* maximizer of (D).

Theorem 5.1.a (Local duality theorem (equality constraints only))**.** Assume that $\boldsymbol{x}^* \in \mathbb{R}^n$ satisfies:

(a) $\boldsymbol{x}^*$ is a regular point of the constraints $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ in (P).

(b) $\boldsymbol{x}^*$ is a local minimizer of (P) with associated Lagrange multiplier $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ (which is obtained by solving $\nabla f(\boldsymbol{x}^*) + \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda} = \mathbf{0}$ for $\boldsymbol{\lambda}$; recall Proposition 3.1.b).

(c) The Lagrangian matrix $L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(\boldsymbol{x}^*)$ is positive definite.

Then, $\boldsymbol{\lambda}^*$ is a local maximizer of (D).

Proof. Omitted. □

5.2 Constrained Minimization in the Standard Form

5.2.1 **Primal and dual problems.** We now consider the general case with the following constrained minimization problem in the standard form

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \\ & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $f, h_1, \ldots, h_m, g_1, \ldots, g_q \in C^2$. This is called the **primal problem**, denoted by (P). The **(Lagrangian) dual problem** of (P), denoted by (D), is given by the following constrained *maximization* problem

$$\begin{aligned} \text{Maximize} \quad & \phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \in \mathbb{R}^m, \\ & \boldsymbol{\mu} \geq \boldsymbol{0}, \end{aligned}$$

where $\phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \ell(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} [f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}]$ is the **dual function**, i.e., for every fixed $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \geq \boldsymbol{0}$, $\phi(\boldsymbol{\lambda})$ equals the minimum value of the following unconstrained minimization problem

$$\begin{aligned} \text{Minimize} \quad & \ell(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu} \\ \text{subject to} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

if a minimizer exists, and equals $-\infty$ otherwise.

5.2.2 **Local duality theorem.** We can extend the local duality theorem for equality constrained minimization (Theorem 5.1.a) to the general case here.

Theorem 5.2.a (Local duality theorem)**.** Assume that $\boldsymbol{x}^* \in \mathbb{R}^n$ satisfies:

(a) $\boldsymbol{x}^*$ is a regular point of the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$ in (P).

(b) $\boldsymbol{x}^*$ is a local minimizer of (P) with associated Lagrange multipliers $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ and $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_q^*) \geq \boldsymbol{0}$ (obtained from the KKT conditions).

(c) The Lagrangian matrix $L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(\boldsymbol{x}^*) + \sum_{j=1}^q \mu_j \nabla^2 g_j(\boldsymbol{x}^*)$ is positive definite.

Then, $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a local maximizer of (D).

Proof. Omitted. □

5.2.3 **Concavity of the dual function.**

Proposition 5.2.b. The dual function $\phi(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is **concave**, i.e., $-\phi(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is convex, on the set $\{(\boldsymbol{\lambda}, \boldsymbol{\mu}) : \boldsymbol{\lambda} \in \mathbb{R}^m, \ \boldsymbol{\mu} \geq \boldsymbol{0}\}$ (which is convex).

[Note: Here we extend the notion of convex function slightly to include extended-real-valued functions (possibly taking values $\pm\infty$), by following some basic arithmetic rules about $\pm\infty$.]

Proof. Fix any $(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1)$ and $(\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2)$ in the domain of ϕ, and any $\alpha \in [0, 1]$. Then,

$$\begin{aligned} & \phi(\alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2) \\ = & \inf_{\boldsymbol{x} \in \mathbb{R}^n} [f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})(\alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2) + g(\boldsymbol{x})(\alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2)] \\ = & \inf_{\boldsymbol{x} \in \mathbb{R}^n} [\alpha\big(f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda}_1 + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}_1\big) + (1-\alpha)\big(f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda}_2 + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}_2\big)] \\ \geq & \inf_{\boldsymbol{x} \in \mathbb{R}^n} [\alpha\big(f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda}_1 + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}_1\big)] + \inf_{\boldsymbol{x} \in \mathbb{R}^n} [(1-\alpha)\big(f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda}_2 + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}_2\big)] \\ = & \alpha\phi(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) + (1-\alpha)\phi(\boldsymbol{\lambda}_2, \mu_2), \end{aligned}$$

which implies the concavity of the dual function. □

5.2.4 **Weak duality.** An interesting relationship between primal and dual problems is given by the *weak duality*, which suggests that dual objective function value is always bounded above by primal objective function value, regardless of what primal and dual feasible solutions we are considering.

Proposition 5.2.c (Weak duality)**.** Let $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ be feasible solutions of (P) and (D), respectively. Then, $\phi(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) \leq f(\bar{\boldsymbol{x}})$.

Proof. By the feasibility of $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$, we have $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \mathbf{0}$, $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$, and $\bar{\boldsymbol{\mu}} \geq \mathbf{0}$. Hence, the Lagrangian satisfies

$$\ell(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) = f(\bar{\boldsymbol{x}}) + \underbrace{\boldsymbol{h}(\bar{\boldsymbol{x}})\bar{\boldsymbol{\lambda}}}_{=0} + \underbrace{\boldsymbol{g}(\bar{\boldsymbol{x}})\bar{\boldsymbol{\mu}}}_{\leq 0} \leq f(\bar{\boldsymbol{x}}),$$

and thus

$$\phi(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) = \inf_{\boldsymbol{x}\in\mathbb{R}^n} \ell(\boldsymbol{x}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) \leq \ell(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) \leq f(\bar{\boldsymbol{x}}).$$

□

A helpful corollary of the weak duality is the following, which suggests that primal and dual feasible solutions with the same objective function value are optimal to the primal and dual problems, respectively.

Corollary 5.2.d. Let $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ be feasible solutions of (P) and (D), respectively. If they share the same objective function value, i.e., $f(\bar{\boldsymbol{x}}) = \phi(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$, then $\bar{\boldsymbol{x}}$ is a global minimizer of (P) and $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ is a global maximizer of (D).

Proof. We first show that $\bar{\boldsymbol{x}}$ is a global minimizer of (P). Fix any feasible solution $\boldsymbol{x}$ of (P). Then, we have

$$f(\bar{\boldsymbol{x}}) \overset{\text{(assumption)}}{=} \phi(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) \overset{\text{(weak duality)}}{\leq} f(\boldsymbol{x}),$$

which implies that $\bar{\boldsymbol{x}}$ is a global minimizer of (P).

Next, we show that $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ is a global maximizer of (D). Fix any feasible feasible solution $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ of (D). Then, we have

$$\phi(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) \overset{\text{(assumption)}}{=} f(\bar{\boldsymbol{x}}) \overset{\text{(weak duality)}}{\geq} \phi(\boldsymbol{\lambda}, \boldsymbol{\mu}),$$

which implies that $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ is a global maximizer of (D). □

5.2.5 **Solving primal problem using duality.** So far, we have developed two ways of solving constrained minimization problem in standard form via duality theory.

(1) *(Finding primal and dual feasible solutions with the same objective function value)* Corollary 5.2.d suggests that, if we are able to somehow come up with a pair of primal feasible and dual feasible solutions that share the same objective function value, then both primal and dual problems are solved immediately. While this method of solving optimization problem seems attractive, it is generally hard to identify such a pair of primal and dual feasible solutions, and thus it may not be too "reliable".

(2) *(Using local duality result)* An alternative (and perhaps more "reliable") method is to use *local* duality result. For the purpose of solving primal problem, the *local duality theorem* mainly serves for providing a way to identify a potential local minimizer, functioning like a "necessary condition". Specifically, we can follow the procedure motioned in [3.2.2], but for the first step of identification of local minimizers, we try to find a local maximizer $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of the Lagrangian dual problem, and then via the KKT conditions, we deduce the corresponding $\boldsymbol{x}^*$, which serves as a *candidate* of local minimizer of the primal problem. Then, we can proceed to the later steps in [3.2.2] to examine whether $\boldsymbol{x}^*$ is a local minimizer, or even a global minimizer, of the primal problem.

5.3 Convex Duality

5.3.1 **Primal and dual problems.** With additional convexity assumptions imposed on the primal problem, we can obtain more duality results, including some *global* (rather than local) ones. Hence, *convex duality* is a subset of duality theory that plays an important role, and it is often more helpful for solving optimization problems than the more general duality theory considered before.

Consider the *primal problem* (P):

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \\ & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \end{aligned}$$

where $f, g_1, \dots, g_q$ are convex, and $h_1, \dots, h_m$ are affine. The Lagrangian dual problem (D) of (P) is:

$$\begin{aligned} \text{Maximize} \quad & \phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \in \mathbb{R}^m, \\ & \boldsymbol{\mu} \geq \boldsymbol{0}, \end{aligned}$$

where $\phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \ell(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} [f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\lambda} + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{\mu}]$ for every fixed $\boldsymbol{\lambda}$ and $\boldsymbol{\mu} \geq \boldsymbol{0}$.

Remarks:

- A function h on $\mathbb{R}^n$ is **affine** if we can write $h(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + \boldsymbol{b}$ for some $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$. Note that for an affine function h, both h and $-h$ are convex.
- By Proposition 1.2.e and the fact that intersection of convex sets remains convex, we can see that the feasible region of (P), namely

$$\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}\} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{h}(\boldsymbol{x}) \leq \boldsymbol{0},\ -\boldsymbol{h}(\boldsymbol{x}) \leq \boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}\},$$

  is a convex set.

5.3.2 **Convex duality theorem.** In the special setting here, we have a *convex duality theorem*, which relates *global* minimizer of (P) and *global* maximizer of (D).

Theorem 5.3.a (Convex duality theorem)**.** Assume that $\boldsymbol{x}^* \in \mathbb{R}^n$ satisfies:

(a) $\boldsymbol{x}^*$ is a regular point of the constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$ in (P).

(b) $\boldsymbol{x}^*$ is a global minimizer of (P) with the associated objective function value $f(\boldsymbol{x}^*) = r^*$ and Lagrange multipliers $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ and $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_q^*) \geq \boldsymbol{0}$ (obtained from the KKT conditions).

Then, $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a global maximizer of (D) with the associated objective function value $\phi(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = r^*$.

Proof. Omitted. □

This result gives us some specialized and often useful methods for solving convex primal problems specified in [5.3.1], on top of the general methods introduced in [5.2.5]. For example, we may apply this result in the following way:

(1) By solving the dual problem (D) (which may be easier than solving the primal problem (P), due to the concavity of ϕ and the simpler constraints), show that it has a unique optimal solution $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

(2) Show that there exists a global minimizer $\boldsymbol{x}^*$ of (P) that is a regular point of the constraints (say, by the extreme value theorem).

(3) By the convex duality theorem, the Lagrange multipliers of $\boldsymbol{x}^*$ are the ones contained in the optimal solution $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ to the dual problem. Hence, we can solve for $\boldsymbol{x}^*$ by considering the KKT conditions.

Links to Definitions

Results

Section 1

- Proposition 1.1.a: a sufficient condition for local extremizers of univariate functions
- Proposition 1.1.b: formulas related to gradient and Hessian matrix
- Proposition 1.1.c: a first-order necessary condition for local minimizer
- [1.1.9]: every vector is a feasible direction at an interior point
- Corollary 1.1.d: a first-order necessary condition for interior local minimizer
- Proposition 1.1.e: a second-order necessary condition for local minimizer
- Corollary 1.1.f: a second-order necessary condition for interior local minimizer
- [1.1.13]: properties of positive (semi)definite matrices
- Proposition 1.1.g: a second-order sufficient condition for interior strict local minimizer
- [1.2.1]: a general procedure for finding local minimizers
- Theorem 1.2.a: characterizations of (strictly) convex functions
- Corollary 1.2.b: equivalence of convexity and nonnegative second derivative for a univariate function
- Theorem 1.2.c: a sufficient condition for strict convexity
- Corollary 1.2.d: positive second derivative implies strict convexity for a univariate function
- Proposition 1.2.e: properties of convex function
- Theorem 1.2.f: properties about minimizations of convex functions
- Proposition 1.2.g: property about maximizations of convex functions
- [1.2.8]: alternative ways for finding global minimizers

Section 2

- Proposition 2.1.a: strict convexity and continuity implies unimodality
- Theorem 2.1.b: properties of unimodal functions
- [2.1.7]: properties of the golden section method
- Proposition 2.1.c: convergence property of the univariate Newton's method
- [2.2.3]: sufficient condition for direction of descent
- Proposition 2.2.a: formula of the direction of steepest descent
- [2.2.6]: basic properties of the steepest descent method

- [2.2.8]: preliminary results about the quadratic problem
- Lemma 2.2.b: formulas of $\boldsymbol{d}_k, \alpha_k, \boldsymbol{x}_{k+1}$ for the steepest descent method on the quadratic problem
- Lemma 2.2.c: behaviour of error function during the iteration of the steepest descent method on the quadratic problem
- Theorem 2.2.d: principal axis theorem
- Lemma 2.2.e: Kantorovich inequality
- Theorem 2.2.f: convergence property of the steepest descent method for the quadratic problem
- Proposition 2.3.a: convergence property of the multivariate Newton's method
- Proposition 2.4.a: linear independence of nonzero Q-orthogonal vectors for positive definite matrix Q
- Proposition 2.4.b: any two eigenvectors of a symmetric matrix Q are Q-orthogonal
- Theorem 2.4.c: conjugate direction theorem
- Theorem 2.4.d: conjugate gradient theorem

Section 3

- Lemma 3.1.a: every point in the tangent plane at a regular point $\boldsymbol{x}^*$ is orthogonal to $\nabla f(\boldsymbol{x}^*)$
- Proposition 3.1.b: first-order necessary conditions for problems with equality constraints only
- Proposition 3.1.c: second-order necessary conditions for problems with equality constraints only
- Proposition 3.1.d: second-order sufficient conditions for problems with equality constraints only
- Theorem 3.1.e: Karush-Kuhn-Tucker (KKT) conditions
- Proposition 3.1.f: second-order necessary conditions for problems with equality and inequality constraints
- Proposition 3.1.g: second-order sufficient conditions for problems with equality and inequality constraints
- [3.2.2]: general procedure for Lagrange methods on constrained minimization problems in the standard form

Section 4

- Theorem 4.1.a: convergence of penalty method
- [4.1.3]: typical penalty function for constrained minimization in the standard form
- [4.1.4]: helper results for solving penalty problems
- [4.1.5]: typical steps for penalty method
- Theorem 4.2.a: convergence of barrier method
- [4.2.3]: typical barrier functions for constrained minimization with inequality constraints only

Section 5

- Theorem 5.1.a: local duality theorem for constrained minimization with equality constraints only
- Theorem 5.2.a: local duality theorem for constrained minimization in the standard form
- Proposition 5.2.b: concavity of the dual function
- Proposition 5.2.c: weak duality
- Corollary 5.2.d: optimality of primal and dual feasible solutions with the same objective function value
- [5.2.5]: some methods for solving constrained minimization problem in the standard form using duality
- Theorem 5.3.a: convex duality theorem

References

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: Theory and algorithms* (3rd ed.). Wiley-Interscience.

Bertsekas, D. P. (2016). *Nonlinear programming* (3rd ed.). Athena Scientific.

Luenberger, D. G., & Ye, Y. (2021). *Linear and nonlinear programming* (5th ed.). Springer International Publishing.

Peressini, A. L., Sullivan, F. E., & Uhl, J. J. (1993). *The mathematics of nonlinear programming*. Springer-Verlag.