Linear and Convex Optimization: Summary

1. Mathematics foundations

1.1 Linear algebra

Matrix space

Given a matrix $A \in \mathbb{R}^{m \times n}$, it also can be viewed as a *linear map* from $\mathbb{R}^n$ to $\mathbb{R}^m$, where $v \in \mathbb{R}^n$ is mapped to $Av \in \mathbb{R}^m$.

Four fundamental (sub)spaces:

1. *column space*, also called *range* or *image* of A, denoted by $\mathcal{R}(A)$ or $\mathrm{im}(A) \triangleq \{Av : v \in \mathbb{R}^n\}$.
2. *row space* of A, also referred to as the column space (or range / image) of A^T.
3. *nullspace*, also known as the *kernel* of A, denoted by $\mathcal{N}(A)$ or $\ker(A) \triangleq \{v \in \mathbb{R}^n : Av = \mathbf{0}\}$.
4. *left nullspace*, or *cokernel* of A, which is the same as $\ker(A^\mathsf{T})$.

The rank of A, denoted by $\mathrm{rank}(A)$, refers the dimension of the image of A, i.e.,

$$\mathrm{rank}(A) = \dim \mathrm{im}(A)\,.$$

Fundamental theorem of linear algebra

We first introduce the *orthogonal subspaces*:

> **Definition** (*Orthogonal subspaces*)
> Two subspaces S_1, S_2 are orthogonal, if $\forall\, v_1 \in S_1, v_2 \in S_2$, $\langle v_1, v_2 \rangle = v_1^\mathsf{T} v_2 = 0$.

> **Theorem** (*Fundamental theorem of linear algebra*)
> Let $A \in \mathbb{R}^{m\times n}$ be a matrix with rank r. Then
>
> $$\dim \mathrm{im}(A) = r, \quad \dim \ker(A) = n - r, \quad \dim \mathrm{coker}(A) = m - r\,.$$
>
> Moreover, $\mathrm{im}(A)$ and $\ker(A)$ are orthogonal subspaces.

Remark: For any linear map $A : V \to W$, the fact that $\dim \mathrm{im}(A) + \dim \ker(A) = \dim V$ is also called the *rank-nullity theorem*.

Matrix decomposition

Let $A \in \mathcal{S}^n$, the set of all real, symmetric, $n \times n$ matrices, then A can be factored:

$$A = U \Lambda U^\mathsf{T}\,,$$

where U is an *orthogonal matrix*, namely, $UU^\mathsf{T} = I$, and $\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is a *diagonal matrix* having all eigenvalues of A as its main diagonal entries.

Remark: If $A \notin \mathcal{S}^n$, then there is also a similar decomposition called the *singular value decomposition*.

Definiteness

Given a real symmetric $A \in \mathbb{R}^{n\times n}$, we sometimes need to study $v^\mathsf{T} A v$ for some vector $v \in \mathbb{R}^n$.

> **Definition** (*Definiteness*)
> Let $A \in \mathbb{R}^{n\times n}$ be a real symmetric matrix.
> - A is *positive definite*, denoted by $A \succ \mathbf{0}$ or $A \in \mathcal{S}^n_{++}$, if $\forall\, \mathbf{0} \neq v \in \mathbb{R}^n$, $v^\mathsf{T} A v > 0$.
> - A is *positive semi-definite*, denoted by $A \succeq \mathbf{0}$ or $A \in \mathcal{S}^n_{+}$ if $\forall\, v \in \mathbb{R}^n$, $v^\mathsf{T} A v \geq 0$.
> - A is *negative definite*, denoted by $A \prec \mathbf{0}$, if $\forall\, \mathbf{0} \neq v \in \mathbb{R}^n$, $v^\mathsf{T} A v < 0$.
> - A is *negative semi-definite*, denoted by $A \preceq \mathbf{0}$, if $\forall\, v \in \mathbb{R}^n$, $v^\mathsf{T} A v \leq 0$.

We can use the following proposition to justify definiteness.

Proposition A symmetric matrix A is positive definite iff all its eigenvalues are positive; A is positive semi-definite iff all its eigenvalues are nonnegative.

Sometimes we also use *Sylvester's criterion* to determine definiteness.

Given a matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix},$$

a $k \times k$ *principal submatrix* of A is a submatrix of A, consisting of k rows and k columns of the same indices $I = \{i_1, \ldots, i_k\}$,

$$A_I = \begin{pmatrix} a_{i_1,i_1} & \cdots & a_{i_1,i_k} \\ \vdots & \ddots & \vdots \\ a_{i_k,i_1} & \cdots & a_{i_k,i_k} \end{pmatrix}.$$

The determinant of A_I $\det(A_I)$ is called the *principal minor* (主子式). In particular, if $I = [k] = \{1, \ldots, k\}$, $\det(A_I)$ is called the *leading principal minor* (顺序主子式).

Theorem (*Sylvester's criterion*)
Suppose A is a symmetric matrix, then

- $A \succ 0$ iff $D_k(A) \triangleq \det(A_{[k]}) > 0$ for all $k = 1, \ldots, n$,
- $A \succeq 0$ iff $D_I(A) \triangleq \det(A_I) \geq 0$ for all $I \subseteq [n]$,
- $A \succeq 0$ if $D_k(A) > 0$ for $k \in [n-1]$, and $D_n(A) \geq 0$.

Remark: We cannot get a criterion for semidefiniteness similar to the first criterion for positive definiteness. Consider the following matrix, all of its principal minor are non-negative. Consider the following example:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}.$$

It is easy to see that $D_k(A) \geq 0$ for all k. However, A is not positive semidefinite.

1.2 Analysis

Norm

To discuss the properties of functions, we should define norm first, which is somehow the generalization of absolute value in high dimensions.

Definition (*Norm*)
Given a vector space V over a field F, a norm $\|\cdot\| : V \to \mathbb{R}$ is a function having the following properties:
1. (*Nonnegativity*) $\forall\, v \in V$, $\|v\| \geq 0$.
2. (*Positive definiteness*) $\|v\| = 0$ iff $v = \mathbf{0}$.
3. (*Absolute homogeneity*) $\forall\, r \in \mathbb{R}$ and $v \in V$, $\|r \cdot v\| = |r| \cdot \|v\|$.
4. (*Triangle inequality*) $\forall\, u, v \in V$, $\|u + v\| \leq \|u\| + \|v\|$.

This definition is not constructive. We now see some specific examples.

Example (ℓ_p *norm and operator norm*)

- ℓ_p norm defined on $\mathbb{R}^n$: $\|x\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}$, where $p \geq 1$. In particular,
  - ℓ_1 norm: $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$.
  - ℓ_2 norm: $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$, which is the most common norm in $\mathbb{R}^n$.
  - ℓ_∞ norm: $\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$.
- Operator norm defined on $\mathbb{R}^{m\times n}$: a matrix $A \in \mathbb{R}^{m\times n}$ is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the operator norm is defined by

$$\|A\|_{a,b} = \max_{v\neq \mathbf{0}} \frac{\|Av\|_b}{\|v\|_a} = \max_{\|v\|_a = 1} \|Av\|_b \,.$$

  In particular,
  - Spectral norm: $\|A\|_2 = \|A\|_{2,2} = \sigma_{\max}(A)$, the largest singular value of A. If $A \succeq 0$, $\sigma_{\max}(A) = \lambda_{\max}(A)$, the largest eigenvalue of A.

Properties of functions

We now study some important properties of functions.

- **Limit** The notion of limit is the foundation of our course. Informally, $\lim_{x\to\alpha} f(x)$ is the value that f approaches as x approaches the value α. Distances are measured by norms. In fact it is not necessary to specify a norm in a finite dimensional linear space, as we've showed that all norms are somehow "equivalent" in finite dimensional spaces.
- **Continuity** A function f is said to be continuous at some point $\alpha \in \mathrm{dom}(f)$ if $\lim_{x\to\alpha} f(x) = f(\alpha)$. f is continuous if it is continuous at all points $\alpha \in \mathrm{dom}(f)$.
- **Lipschitz** A function f is Lipschitz with constant L (or L-Lipschitz) if $\|f(x) - f(y)\| \leq L\,\|x - y\|$ for all $x, y \in \mathrm{dom}(f)$. If we refer to a function f as Lipschitz, we are making a stronger statement about the continuity of f. A Lipschitz function is not only continuous, but it does not change value very rapidly, either.
- **Differentiability** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is considered differentiable at $\alpha \in \mathrm{int\,dom}(f)$ if there exists an linear approximation of f locally at α. Formally, $\exists\, A \in \mathbb{R}^{m\times n}$ a linear transform from $\mathbb{R}^n$ to $\mathbb{R}^m$ such that

$$\lim_{x\to\alpha} \frac{\|f(x) - f(\alpha) - A(x-\alpha)\|}{\|x - \alpha\|} = 0\,.$$

Then $Df(\alpha) \triangleq A$ is the differential of f at α. The matrix A is sometimes known as the *Jacobian matrix*. In particular, if $m = 1$, $\nabla f(\alpha) = A^{\mathsf{T}}$ is called the *gradient* of f. Suppose $f : (x_1, \ldots, x_n)^{\mathsf{T}} \to (f_1, \ldots, f_m)^{\mathsf{T}}$. Then the Jacobian matrix is defined by

$$Df = \begin{pmatrix} \nabla f_1^{\mathsf{T}} \\ \vdots \\ \nabla f_m^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Taylor expansion

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function. If all second partial derivatives of f exist and are continuous then the *Hessian* matrix of f is defined by the Jacobian of the gradient of f:

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

By the continuity of the second derivatives, Hessian is always a symmetric matrix.

Given the definition of the Hessian, we now discuss the Taylor expansion and Taylor's theorem for multivariate functions.

Theorem (*Taylor's theorem*)
Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function at the point $\alpha \in \mathbb{R}^n$, then there exists $h_\alpha(x) : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = f(\alpha) + \nabla f(\alpha)^{\mathsf{T}}(x - \alpha) + \frac{1}{2}(x - \alpha)^{\mathsf{T}} \nabla^2 f(\alpha)(x - \alpha) + h_\alpha(x)\|x - \alpha\|^2,$$

and $\lim_{x \to \alpha} h_\alpha(x) = 0$. Namely,

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

2. Convexity

2.1 Convex sets

We now discuss convex sets. First we define *lines* and *segments*.

Definition (*Lines and segments*)
Given two points $x_1 \neq x_2 \in \mathbb{R}^n$, the *line* containing x_1 and x_2 can be given by

$$\lambda x_1 + (1-\lambda)x_2$$

where $\lambda \in \mathbb{R}$; the *segment* whose endpoints are x_1 and x_2 can be written as

$$\theta x_1 + (1-\theta)x_2$$

where $\theta \in [0,1]$.

For convenience, we use $\bar{\theta}$ to denote $1-\theta$ from now on.

A set is an *affine set* if it contains the line through any two distinct points in the set; a set is a *convex* set if it contains all segments whose endpoints are in the set. Formally,

Definition (*Affine sets*)
A set C is *affine* if the line containing any two points in C lies in C, i.e., $\forall\, x_1, x_2 \in C$, $\forall\, \lambda \in \mathbb{R}$,

$$\lambda x_1 + \bar{\lambda} x_2 \in C\,.$$

Definition (*Convex sets*)
A set C is *convex* if the line segment between any two points in C lies in C, i.e., $\forall\, x_1, x_2 \in C$, $\forall\, \theta \in [0,1]$,

$$\theta x_1 + \bar{\theta} x_2 \in C\,.$$

Now we introduce some simple examples of convex sets.

Example (*Convex sets*)

- The empty set $\varnothing$, the singleton set $\{x_0\}$ and the complete space $\mathbb{R}^n$ are convex.
- Lines $\{w^\mathsf{T} x = b\}$, line segments, hyperplanes $\{Ax = b\}$ and halfspaces $\{Ax \leq b\}$ are convex.
- Eculidean balls $\mathcal{B}(x_0, \varepsilon) \triangleq \{x : \|x - x_0\|_2 \leq \varepsilon\}$, ellipsoids $\{x : (x-x_0)^\mathsf{T} Q (x-x_0) \leq 1\}$ where $Q \in \mathcal{S}^n_{++}$ is a real positive definite matrix are convex.

Convex hull

Given n points $x_1, x_2, \ldots, x_n$ and n reals $\theta_1, \theta_2, \ldots, \theta_n$,

$$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

is a *convex combination* of $\{x_1, x_2, \ldots, x_n\}$ if $\theta_1 + \theta_2 + \cdots + \theta_n = 1$ and $\theta_i \geq 0$ for all $i = 1, 2, \ldots, n$.

Definition (*Convex hull*)
The *convex hull* of a set S is the set of all convex combinations of points in S:

$$\mathbf{conv}\, S = \{\theta_1 x_1 + \cdots + \theta_k x_k : x_i \in S, \theta_i \geq 0, \sum_{j=1}^{k} \theta_j = 1\}\,.$$

Remark: The convex hull of S is always a convex set.

Convexity-preserving operations

We now present some convexity-preserving operations.

- **Intersection** If C and D are both convex, then $C \cap D$ is convex.
- **Affine transform** If $C \subseteq \mathbb{R}^n$ is convex, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, then

  $$AC + b \triangleq \{Ax + b : x \in C\}$$

  is convex.
- **Cartesian product** If C and D are convex, then $C \times D \triangleq \{(x_1, x_2) : x_1 \in C, x_2 \in D\}$ is convex.
- **Set sum (Minkowski addition)** If C and D are both convex, then

  $$C + D \triangleq \{x_1 + x_2 : x_1 \in C, x_2 \in D\}$$

  is convex.

Separating hyperplane and supporting hyperplane

An important property of convex sets is that disjoint convex sets can be separated by hyperplanes.

Theorem (*Separating hyperplane theorem*)
Let C, D be two disjoint convex sets, i.e., $C \cap D = \varnothing$. Then there exists $\mathbf{w} \neq \mathbf{0}$ and b such that

$$\forall\, x \in C, \quad \mathbf{w}^\mathsf{T} x + b \geq 0 \quad \text{and} \quad \forall\, y \in D, \quad \mathbf{w}^\mathsf{T} y + b \leq 0\,.$$

Remark: Moreover, we can define *strictly separation*, where the separating hyperplane does not intersect C or D, namely,

$$\forall\, x \in C, \quad \mathbf{w}^\mathsf{T} x + b > 0 \quad \text{and} \quad \forall\, y \in D, \quad \mathbf{w}^\mathsf{T} y + b < 0\,.$$

In general, not any two convex sets C and D have a strictly separating hyperplane. A sufficient (but not necessary) condition for this to hold is that: C and D are closed and at least one of them is bounded.

A related result is the *supporting hyperplane theorem*. We first define some notions for point sets.

Definition (*Interior points and boundary*)
Given a set $C \in \mathbb{R}^n$, a point $x_0 \in C$ is called an *interior point* if $\exists\, \varepsilon > 0, \mathcal{B}(x_0, \varepsilon) \subseteq C$. Then we define

- $\operatorname{int} C$: the set of all interior points of C;
- $\operatorname{cl} C$: the set of all x that $\exists\, x_1, x_2, \ldots$, $\{x_m\}$ converges to x and $x_i \in C$ for all $i \in \mathbb{N}$;
- ∂C: the boundary of C, defined by $\partial C = \operatorname{cl} C \setminus \operatorname{int} C$.

Now we can present the supporting hyperplane theorem.

Theorem (*Supporting hyperplane theorem*)
For any convex set C and any boundary point $x_0 \in \partial C$, there exists a *supporting hyperplane* for C at x_0, that is, $\exists\, \mathbf{w} \neq \mathbf{0}$ and b such that

$$\forall\, x \in C, \quad \mathbf{w}^\mathsf{T} x + b \geq 0 \quad \text{and} \quad \mathbf{w}^\mathsf{T} x_0 + b = 0\,.$$

2.2 Convex functions

Definition (*Convex functions*)
A real-valued function $f : D \to \mathbb{R}$ is called a *convex function* if its domain D is a convex set and f satisfies *Jensen's inequality*:

$$\forall\, x, y \in D, \forall\, \theta \in [0,1], \quad f(\theta x + \bar{\theta} y) \leq \theta f(x) + \bar{\theta} f(y)\,.$$

Moreover, if $f(\theta x + \bar{\theta} y) < \theta f(x) + \bar{\theta} f(y)$ for all $\theta \in (0,1)$, f is called *strictly convex*.

Remark: f is said to be *concave* if $-f$ is convex; f is *strictly concave* if $-f$ is strictly convex.

Epigraph

We now explain why we call these functions *convex*.

Definition (*Epigraph*)
Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued function. The *graph* of f is defined by

$$\{(x, f(x)) : x \in D\}$$

and the *epigraph* of f is given by

$$\operatorname{epi}(f) \triangleq \{(x,t) : x \in D, t \geq f(x)\}\,.$$

Note that both graph and epigraph are subsets of $\mathbb{R}^{n+1}$. The next proposition reveals the connection between convex sets and convex functions.

Proposition A function f is a convex function **iff** its epigraph is a convex set.

Criteria for convexity

We now discuss how to verify convexity.

Theorem (*Zeroth-order condition*)
A function is convex iff it is convex when restricted to any line that intersects its domain. Namely, f is convex iff $\forall\, x \in \mathrm{dom}(f)$, $\forall\, v$, the function $g(t) = f(x+tv)$ is convex, when restricted to its
domain $\{t : x + tv \in \mathrm{dom}(f)\}$.

Theorem (*First-order condition*)
Suppose $f : D \to \mathbb{R}$ is a differentiable function. Then f is a convex function iff its domain D is a convex set, and

$$\forall\, x, y \in D, \quad f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y - x)\,.$$

The function $g(y) = f(x) + \nabla f(x)^{\mathsf{T}}(y-x)$ is the first-order Taylor approximation of f near x. The inequality above states that for a convex function, the first-order Taylor approximation is in fact a global underestimator of the function, and vice versa.

Theorem (*Second-order condition*)
Let $f : D \to \mathbb{R}$ be a twice differentiable function on an open domain $\mathrm{dom}(f) = D$. The f is convex iff D is convex and its Hessian is positive semi-definite:

$$\forall\, x \in D, \quad H(x) = \nabla^2 f(x) \succeq \mathbf{0}\,.$$

Remark: If $\forall\, x \in D$, $H(x) \succ \mathbf{0}$, then f is strictly convex. Note that it is not a necessary condition.

Convexity-preserving operations

- **Nonnegative (weighted) sum** Suppose $f_1, f_2, \ldots, f_m$ are convex functions over D. Then for all $\omega_1, \omega_2, \ldots, \omega_m \geq 0$,

  $$f = \omega_1 f_1 + \omega_2 f_2 + \cdots + \omega_m f_m$$

  is also a convex function over D. Furthermore,

  $$g(x) = \int_{\Omega} \omega(y) f(x, y)\, \mathrm{d}y$$

  is convex, if for any fixed $y \in \Omega$, $\omega(y) \geq 0$, $f(x, y)$ is a convex function w.r.t. x, and the integral exists.
- **Pointwise maximum (supremum)** Suppose $f_1, f_2, \ldots, f_m$ are convex functions over D. Then

  $$g(x) \triangleq \max\{f_1(x), f_2(x), \ldots, f_m(x)\}$$

is also a convex function over D. Furthermore,

$$g(x) = \sup_{y \in \Omega} f(x, y)$$

is convex if for any fixed $y \in \Omega$, $f(x, y)$ is a convex function w.r.t. x.

- **Composition with affine mapping** If $f(x)$ is a convex function, so is $f(Ax + b)$.
- **Composition with monotone functions** Suppose $h(x) = f(g_1(x), g_2(x), \ldots, g_m(x))$. Then $h(x)$ is convex if f is convex, and
  - f is increasing and g_i is convex for all i, or;
  - f is decreasing and g_i is concave for all i.
- **Mimimum (infimum) over convex sets** Suppose $f(x, y)$ is a convex function, and $C \neq \varnothing$ is a convex set. Then

$$g(x) \triangleq \inf_{y \in C} f(x, y)$$

is convex.

2.3. Convex optimization problems

We now discuss what type of optimization problems we should consider in this course. In general, an optimization problem is to find the minimum value of $f(x)$ where x satisfies $g(x) = 0$ and $h(x) \leq 0$. Namely, it can be written as

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) = 0 \quad 1 \leq i \leq m\,, \\ & h_j(x) \leq 0 \quad 1 \leq j \leq \ell\,. \end{aligned}$$

Here $f(x)$ is called the *objective function*, and $g_i(x), h_j(x)$ are called the *constraint functions*. We say x is a *feasible solution* if $g_i(x) = 0$ and $h_j(x) \leq 0$; and x is the (or an) *optimal solution* if x is feasible and $f(x)$ takes the minimum value over all feasible solutions. The *domain* of the problem is defined by $\mathrm{dom}(f) \cap \mathrm{dom}(g) \cap \mathrm{dom}(h)$. The *feasible set* of the problem is the set of all feasible solutions.
In this course, we only consider *linear* and *convex* optimization.

Linear programming

Linear optimization is also called *linear programming*. A linear program is an optimization problem whose objective function and constraint functions are all affine functions, for example,

$$\begin{aligned} \min_x \quad & c^{\mathsf{T}} x \quad (\text{or } \max_x \, c^{\mathsf{T}} x) \\ \text{subject to} \quad & A_1 x \leq b_1, \\ & A_2 x = b_2, \\ & A_3 x \geq b_3. \end{aligned}$$

General convex optimization

A convex optimization problem is a problem with a *convex objective function*, *affine equality constraints*, and *convex inequality constraints*, namely, f is a convex function, g_i is an affine function for all i, (i.e., $g(x) = Ax - b$), and h_j is a convex function for all j.

In particular, a linear program is a convex optimization.

3. Linear programming

We first study the linear programming.

3.1 Standard forms, slackness

The standard form of a LP is to maximize a linear function, with equality constraints and non-negative variables. Namely, the standard form is written as the form

$$\begin{aligned} \max_x \quad & c^\mathsf{T} x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

We now discuss how to rewrite a general LP as the standard form. Note that $\min c^\mathsf{T} x$ is equivalent to $\max -c^\mathsf{T} x$ and $Ax \geq b$ is equivalent to $-Ax \leq -b$. So we only need to consider how to add non-negativity constraints and replace inequalities by equalities.
Assume that we have a LP

$$\begin{aligned} \max_x \quad & c^\mathsf{T} x \\ \text{subject to} \quad & Ax \leq b \end{aligned}$$

Since any real number can be written as the difference of two non-negative numbers, we could assume that $x_i = y_i^+ - y_i^-$ where $y_i^+, y_i^- \geq 0$. Now the LP can be written as

$$\begin{aligned} \max_x \quad & c^\mathsf{T} y^+ - c^\mathsf{T} y^- \\ \text{subject to} \quad & Ay^+ - Ay^- \leq b \\ & y^+, y^- \geq 0 \end{aligned}$$

Finally, for each inequality constraint, add a *slackness variable* s_i, and rewrite the LP as

$$\begin{aligned} \max_x \quad & c^\mathsf{T} y^+ - c^\mathsf{T} y^- \\ \text{subject to} \quad & Ay^+ - Ay^- + s = b \\ & y^+, y^-, s \geq 0 \end{aligned}$$

Here b, c, x, y^+, y^-, s are vectors, and $\leq, =, \geq$ for vectors mean $\leq, =, \geq$ for all terms of vectors.

3.2 Duality of LP

Primal and dual

Consider a LP

$$\begin{aligned} \max_x \quad & c^\mathsf{T} x \\ \text{subject to} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

If we assign a non-negative multiplier y_i to the i-th constraint such that $\sum_{i=1}^m A_{ij} y_i \geq c_j$ for all j,

then $y^\mathsf{T} b$ gives an upper bound of the optimal value of the primal LP. We would like to minimize the upper bound $y^\mathsf{T} b$, which is to solve another LP:

$$\begin{aligned} \min_{y} \quad & y^\mathsf{T} b \\ \text{subject to} \quad & y^\mathsf{T} A \geq c^\mathsf{T} \\ & y \geq 0 \end{aligned}$$

This is called the *dual* of the primal LP.

Proposition The dual of the dual is the primal.

Remark: We will see later that this proposition is not true for general optimization.

In general, the relation between the primal and the dual can be summarized as the following table:

Primal (max)	**Dual (min)**
i-th constraint $\leq$	i-th variable ≥ 0
i-th constraint $=$	i-th variable unrestricted
j-th variable ≥ 0	j-th constraint $\geq$
j-th variable unrestricted	j-th constraint $=$

Weak and strong duality

Theorem (*Weak duality*)

$$\max_x \ c^\mathsf{T} x \leq \min_y \ y^\mathsf{T} b \,.$$

▼ Proof

$$c^\mathsf{T} x \leq y^\mathsf{T} A x \leq y^\mathsf{T} b \,.$$

In fact, as we will see later, the weak duality theorem holds for general optimization. For LP, we further have the following strong duality theorem.

Theorem (*Strong duality*)
If the primal LP has feasible and bounded optimal solution, so is the dual. Moreover,

$$\max c^\mathsf{T} x = \min \ y^\mathsf{T} b \,.$$

The proof of the strong duality theorem uses the *Farkas' lemma* (https://en.wikipedia.org/wiki/Farkas%27_lemma).

Complementary slackness

We can also use the *complementary slackness* condition to verify optimality.

Theorem (*Complementary slackness*)
Suppose the primal has a feasible solution x^*, and the dual has a solution y^*. Then x^* and y^* are optimal for the primal and the dual respectively, iff

$$y^\mathsf{T}(Ax - b) = 0, \quad \text{and} \quad (y^\mathsf{T}A - c^\mathsf{T})x = 0\,.$$

In other words,

- for all i, either $y_i = 0$, or $(Ax)_i = b_i$; and
- for all j, either $x_j = 0$, or $(y^\mathsf{T}A)_j = c_j$.

4. Unconstrained convex optimization

4.1 Conditions for optimality

We first consider how to justify optimality. Note that if f is a general function, then we have the following conditions for *local* optimal points:

Proposition (*First and second condition for optimality*)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. We say x is a *local* minimum point if there exists $\varepsilon > 0$ such that $f(x) \le f(y)$ for all $y \in \mathcal{B}(x, \varepsilon)$. Then we have the following necessary or sufficient conditions:

- if x^* is a *local* minimum point, then $\nabla f(x^*) = \mathbf{0}$;
- if x^* is a *local* minimum point, then $\nabla^2 f(x^*) \succeq 0$;
- if $\nabla f(x) = \mathbf{0}$ and $\nabla^2 f(x^*) \succ 0$, then x^* is a *local* minimum point.

However, if we only consider *convex* functions, the optimality condition becomes much simpler, due to the first-order condition for convexity.

Theorem (*Opitmality condition for convex functions*)
Suppose $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then x^* is a *global* minimum point of f iff

$$\forall\, y \in D, \quad \nabla f(x^*)^\mathsf{T}(y - x^*) \ge 0\,.$$

In particular, if $D = \mathbb{R}^n$, then x^* is a global minimum point iff $\nabla f(x^*) = \mathbf{0}$.

4.2 Algorithms: descent methods

Suppose we pick a point x. If $\nabla f(x) \neq \mathbf{0}$, then we know that x is not the optimal point. But where can we find the optimal point? A naive idea is to look along the descending direction of the function value.

Gradient descent method and Newton's method

We would like to use the descent algorithms, namely, iteration algorithms that $x_{k+1} = x_k + \eta_k d_k$, where d_k is a descent direction. Use different descent directions we could obtain different algorithms. Note that d_k is a descent direction iff the *directional derivative* $\nabla f(x_k)^\mathsf{T} d_k \leq 0$. We now consider how to choose a proper descending direction d_k.

The first idea is to choose the *negative gradient* $d_k = -\nabla f(x_k)$, since it is the direction where the value of f decreases at the fastest rate. In fact, by the *Cauchy–Schwarz inequality*, we have

$$\left|\nabla f(x_k)^\mathsf{T} d_k\right| \leq \|\nabla f(x_k)\|_2 \cdot \|d_k\|_2 \,,$$

and the equality happens only when $d_k = \alpha \nabla f(x_k)$ for some α. Then we apply the descent algorithm, and call such an algorithm the *gradient descent method*:

$\textbf{given a starting point } x_0$
$\textbf{repeat}$
$\quad \textbf{choose a proper step size } \eta_k$
$\quad x_{k+1} \leftarrow x_k - \eta_k \nabla f(x_k)$
$\quad k \leftarrow k + 1$
$\textbf{until } \|\nabla f(x_k)\| \leq \delta \textbf{ for some sufficiently small } \delta$

However, although $-\nabla f(x)$ is the direction of the largest descending rate, it is usually not the optimal direction globally. Consider $f(x) = \frac{1}{2} x^\mathsf{T} Q x$ for some positive-semi-definite Q. Then $-\nabla f(x) = -Qx$, but the direction to the optimal point $x^* = \mathbf{0}$ is $-x$. If the *condition number* of Q, defined by

$$\kappa(Q) \triangleq \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \,,$$

is large, then so is the difference between two directions.
An improved method is to consider the local approximation of $f(x)$. Assume f is twice continuously differentiable, then we obtain

$$f(x) \approx \tilde{f}\,(x) = f(x_k) + \nabla f(x_k)^\mathsf{T} (x - x_k) + \frac{1}{2} (x - x_k)^\mathsf{T} \nabla^2 f(x_k) (x - x_k) \,,$$

using the Taylor's theorem. Now we let

$$x_{k+1} = \arg\min_x \tilde{f}\,(x) = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \,.$$

This method is called the *Newton's method*. Sometimes the vector $(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ has a large norm so that $\tilde{f}\,(x_{k+1})$ is not a sufficiently good approximation of $f(x_{k+1})$. Then we could add a step size η_k, since $-(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ is always a descending direction of $f(x)$, provided that f is convex.

$\textbf{given a starting point } x_0$
$\textbf{repeat}$
$\quad \textbf{choose a proper step size } \eta_k$
$\quad x_{k+1} \leftarrow x_k - \eta_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$
$\quad k \leftarrow k + 1$
$\textbf{until } \|\nabla f(x_k)\| \leq \delta \textbf{ for some sufficiently small } \delta$

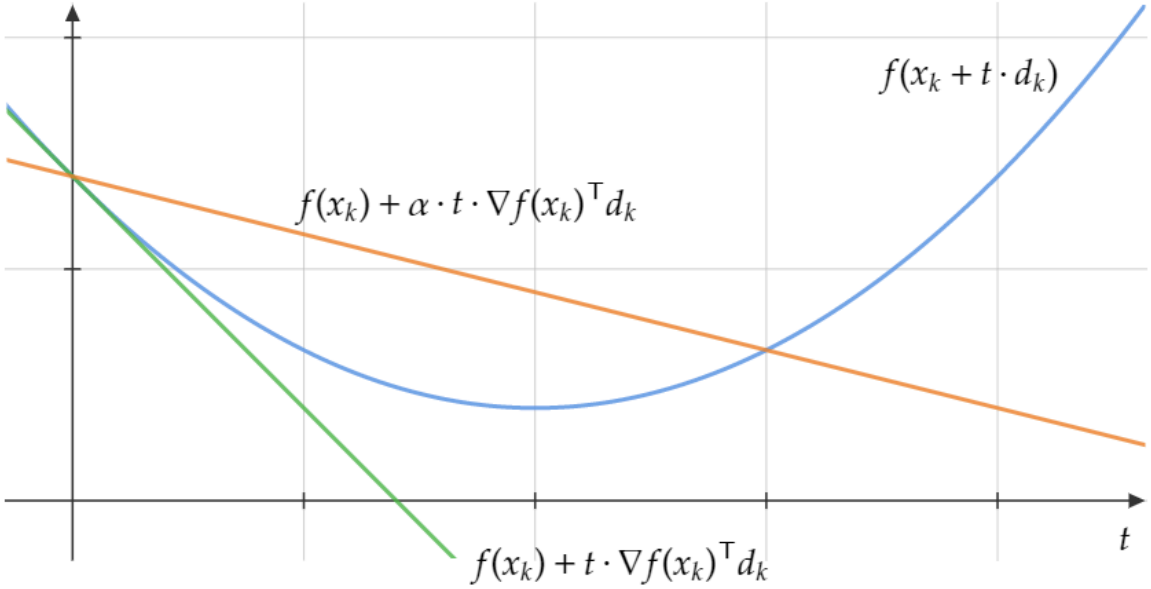
Selecting step size: exact line search and backtracking line search

When we apply the gradient descent method or the Newton's method, we sometimes could set the step size to a constant number (e.g., $\eta = 1/L$ if the function is L-smooth, or $\eta = 1$ in the Newton's method). But sometimes, especially when $\|d_k\|$ is large, setting to a constant is not a good idea. We now introduce two refined methods.

- **Exact line search** Note that $f(x)$ is convex, so f restricted to d_k, namely, $f(x_k + td_k)$ is a convex function with respect to t. Then set

$$\eta_k = \arg\min_t f(x_k + td_k)$$

  so that x_{k+1} is the minimal point of $f(x_k + td_k)$.
- **Backtracking line search** The method of exact line search is usually expensive. We now consider another approach, where we only require that $f(x_{k+1})$ is reduced by a sufficient large number of values compared to $f(x_k)$. Note that $f(x_k) + t\nabla f(x_k)^\mathsf{T} d_k$ provides a lower bound of $f(x_k + td_k)$. We now choose $\alpha \in (0, 1)$ (usually $\alpha = 1/2$) and we think that the decrease is large enough if $f(x_k + td_k) \leq f(x_k) + \alpha t \nabla f(x_k)^\mathsf{T} d_k$. This method is called *Armijo's rule* and can be formally described as follows.



choose $\alpha, \beta \in (0, 1)$ (usually set $\alpha = 1/2$, $\beta \in [0.3, 0.7]$)
$\eta_k \leftarrow 1$ (or other proper values)
while $f(x_k + \eta_k d_k) > f(x_k) + \alpha \eta_k \nabla f(x_k)^\mathsf{T} d_k$
    $\eta_k \leftarrow \beta \cdot \eta_k$

Convergence analysis

In order to introduce the results of convergence analysis, we first define some conditions.

Definition (*Smoothness*)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *smooth* with constant L (or simply, L-*smooth*), if its gradient is L-*Lipschitz*, namely,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \,.$$

Definition (*Strong convexity*)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be μ-*strongly convex*, if $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

Note that if f is a convex function, we obtain that

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

Smoothness and *strong convexity* provide more refined upper and lower bounds on the difference. In fact, we have the following properties.

Proposition (*Smoothness and strong convexity*)
Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function.

- If f is L-smooth, then

$$f(y) \leq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{L}{2}\|y - x\|_2^2 \,.$$

- If f is μ-strongly convex, then

$$f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{\mu}{2}\|y - x\|_2^2 \,.$$

Now we present the results of convergence analysis.

Theorem (Convergence analysis for gradient descent)
Suppose f is a convex function and is L-smooth. Let x^* be a minimum point of f. Then for any fixed step size $\eta \leq 1/L$, the sequence $\{x_k\}$ produced by the gradient descent satisfies

$$f(x_T) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2T\eta} \,.$$

Moreover, if f is μ-strongly convex for some constant $\mu > 0$, then

$$\|x_T - x^*\|^2 \leq (1 - \mu\eta)^T \|x_0 - x^*\|^2 \,,$$

which further implies that

$$f(x_T) - f(x^*) \leq \frac{L(1 - \mu\eta)^T}{2}\|x_0 - x^*\|^2 \,.$$

Theorem (Convergence analysis for Newton's method)
Suppose f is an μ-strongly convex function, and $\nabla^2 f$ is M-Lipschitz continuous (with respect to the *spectral norm* of matrices). Let x^* be a minimum point of f. Then the sequence $\{x_k\}$ produced by Newton's method satisfies

$$\|x_{k+1} - x^*\| \leq \frac{M}{2\mu}\|x_k - x^*\|^2 .$$

Remark: To guarantee global convergence in Newton's method, we should apply the backtracking line search method.

Proximal gradient descent

To apply gradient descent method or Newton's method, we require that the objective function is (twice) continuously differentiable. However, the objective function sometimes is not differentiable. So we now present another algorithm, which is called the *proximal gradient descent* method.

Consider an objective function f, which is the sum of two convex functions. Namely,

$$f(x) = g(x) + h(x) ,$$

where both g and h are convex functions, and g is differentiable but h is not necessary differentiable. Then we approximate $g(x)$ with a quadratic function $\hat{g}(x)$ and let the next iteration be

$$x_{k+1} = \arg\min_x \ \hat{g}(x) + h(x)$$

to approximate the minimum point of $f(x) = g(x) + h(x)$. Formally, we define the *proximal operator* or *proximal mapping* as follows:

$$\text{prox}_h(y) = \arg\min_x \frac{1}{2}\|x - y\|^2 + h(x) ,$$

and the iteration of the proximal gradient descent is given by

$$x_{k+1} = \text{prox}_{\eta h}(x_k - \eta \nabla g(x_k)) .$$

5. Equality constrained optimization

5.1 Optimality condition: Lagrange multiplier conditions

We now consider the optimization problem with equality constraints:

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) = 0 \quad 1 \leq i \leq m . \end{aligned}$$

Again, the first problem is to determine whether x^* is the optimal solution for a given x^*.
We will give a first-order condition, which is called the *Lagrange multipliers conditions*, also known as the *method of Lagrange multipliers*. We omit the proof here, but give some intuitions:

- If $g(x)$ is a real-valued function, e.g., $g(x) = \|x\|^2$, then (**not formally**) x^* is optimal only if the *level set* of $f(x) = f(x^*)$ is tangent to the *level set* of $g(x) = 0$. Namely, $\nabla f(x^*) = \lambda \nabla g(x^*)$ for some λ.
- If $g(x)$ is an affine function, i.e., $g(x) = A(x) - b$, then for any feasible x, and any i, we have

  $$\nabla g_i(x^*)^\mathsf{T}(x - x^*) = 0\,.$$

  Since x^* is optimal for $f(x)$, we have $\nabla f(x^*)^\mathsf{T}(x - x^*) = 0$. It yields that

  $$\nabla f(x^*) \in \operatorname{span}\{\nabla g_1(x^*), \nabla g_2(x^*), \dots, \nabla g_m(x^*)\}\,.$$

- For general (not affine) $g(x)$, to obtain the optimality conditions, we require that the neighborhood of x^* has some "good" properties. These properties can be guaranteed by the *implicit function theorem*, which needs *regularity*.

Definition (*Regularity*)
If $\nabla g_1(x), \dots, \nabla g_m(x)$ are linearly independent at x (or equivalently, $Dg(x) \in \mathbb{R}^{m \times n}$ is a full-rank matrix where $m \le n$), then x is said to be a *regular point*. Otherwise, x is called a *critical point*.

Now we can present the Lagrange multipliers conditions.

Definition (*Lagrangian*)
Let $f(x), g_1(x), \dots, g_m(x) : \mathbb{R}^n \to \mathbb{R}$ be real-valued functions, and $g(x) = (g_1(x), \dots, g_m(x))^\mathsf{T}$. Then the *Lagrangian function* (or simply, *Lagrangian*) is given by

$$\mathcal{L}(x, \lambda) = f(x) + \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x) = f(x) + \lambda^\mathsf{T} g(x)\,,$$

where $\lambda = (\lambda_1, \dots, \lambda_m)^\mathsf{T}$ are called *Lagrange multipliers*.

Theorem (*The method of Lagrange multipliers*, first-order condition)
Suppose x^* is a *local optimal point* of $f(x)$, subject to $g(x) = 0$. Namely, there exists a neighborhood of x^* such that $f(x^*) \le f(x)$ for all x in the neighborhood and $g(x) = 0$. If x^* is feasible, and is a regular point, then there exists Lagrange multipliers $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)^\mathsf{T}$ such that

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \lambda_1^* \nabla g_1(x^*) + \cdots + \lambda_m^* \nabla g_m(x^*) = \mathbf{0}\,.$$

Note that $\nabla_\lambda \mathcal{L}(x, \lambda) = g(x)$. So x^* is feasible iff $\nabla_\lambda \mathcal{L}(x^*, \lambda^*) = \mathbf{0}$. Overall, we have

$$\nabla \mathcal{L}(x^*, \lambda^*) = \mathbf{0}\,.$$

In particular, for equality constrained convex optimization problems, regularity is not neccessary and the first-order condition is further sufficient for the optimal solutions.

Theorem (*The method of Lagrange multipliers*, for convex problems)
Suppose $f(x)$ is a convex function and $g(x)$ is an affine function. Then x^* is the *minimum point* of $f(x)$, subject to $g(x) = 0$, **if and only if** there exists Lagrange multipliers $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)^\mathsf{T}$ such that $\nabla \mathcal{L}(\mathbf{x}, \lambda^*) = \mathbf{0}$, namely,

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \lambda_1^* \nabla g_1(x^*) + \cdots + \lambda_m^* \nabla g_m(x^*) = \mathbf{0}\,,$$
$$\nabla_\lambda \mathcal{L}(x^*, \lambda^*) = \begin{pmatrix} g_1(x^*) \\ \vdots \\ g_m(x^*) \end{pmatrix} = \mathbf{0}\,.$$

5.2 Algorithms: Newton's method

Equality constrained quadratic problems

Recall that in Newton's method, we use the second-order Taylor series to approximate the objective function, and use the global minimum point of the Taylor series as the next iteration. Now we apply this idea again.

Consider a quadratic problem:

$$\begin{aligned} \min \quad & f(x) = \frac{1}{2} x^\mathsf{T} Q x + w^\mathsf{T} x \\ \text{subject to} \quad & g(x) = Ax - b = \mathbf{0}\,. \end{aligned}$$

The Lagrangian is given by

$$\mathcal{L}(x, \lambda) = \frac{1}{2} x^\mathsf{T} Q x + w^\mathsf{T} x + \lambda^\mathsf{T} (Ax - b)\,.$$

So the condition of Lagrange multipliers can be written as follows: x^* is the optimal point iff there exists λ^* such that

$$\begin{pmatrix} Q & A^\mathsf{T} \\ A & \mathbf{0} \end{pmatrix} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} -w \\ b \end{pmatrix}\,.$$

This equation is called the *KKT system*, and the coefficient matrix $\begin{pmatrix} Q & A^\mathsf{T} \\ A & \mathbf{0} \end{pmatrix}$ is called the *KKT matrix*.
Remark: KKT matrix is *nonsingular* if

- $\ker(Q) \cap \ker(A) = \{\mathbf{0}\}$, or equivalently,
- $Ax = \mathbf{0}$ and $x \neq \mathbf{0}$ implies that $x^\mathsf{T} Q x > 0$.

Algorithm

We now present the Newton's method with equality constraints. Again, we use the second-order Taylor series to approximate $f(x)$:

$$f(x) \approx \tilde{f}(x) = f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2}(x - x_k)^\mathsf{T}\nabla^2 f(x_k)(x - x_k)\,.$$

Then we find the optimal solution to $\tilde{f}$ subject to $g(x) = Ax - b = \mathbf{0}$. Let $d = x^* - x_k$. Then d satisfies the following KKT system:

$$\begin{pmatrix} \nabla^2 f(x_k) & A^\mathsf{T} \\ A & \mathbf{0} \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ \mathbf{0} \end{pmatrix}.$$

Thus the method is described as

> given a starting point x_0
> repeat
> compute the descent direction d by solving the KKT system
> choose a proper step size η_k
> $x_{k+1} \leftarrow x_k + \eta_k d$
> $k \leftarrow k + 1$
> until $d^\mathsf{T}\nabla^2 f(x_k) d \le \delta$ for some sufficiently small δ

Convergence analysis

Assume that $g(x) = Ax - b$ is an affine function, where $A \in \mathbb{R}^{m\times n}$ with $m < n$ is a full-rank matrix. Then there exists $F \in \mathbb{R}^{n\times(n-m)}$ such that $\mathrm{im}(F) = \ker(A)$, namely,

$$\{Fz : z \in \mathbb{R}^{n-m}\} = \{x \in \mathbb{R}^n : Ax = \mathbf{0}\}\,.$$

Then the equality constrained optimization

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g(x) = \mathbf{0}\,. \end{aligned}$$

is equivalent to the following unconstrained optimization

$$\min_z \quad f(\tilde{x} + Fz)\,.$$

The Newton's method with equality constraint is also equivalent to the Newton's method to solve $\min_z f(\tilde{x} + Fz)$. Therefore, the convergence analysis results are the same as the results in the unconstrained case.

6. Inequality constrained optimization

6.1 Optimality condition: KKT condition

We now consider the general optimization problem with inequality constraints:

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) = 0 \quad 1 \le i \le m\,, \\ & h_j(x) \le 0 \quad 1 \le j \le \ell\,. \end{aligned}$$

Again, the first problem is to determine whether x^* is the optimal solution for a given x^*.
We first define *active* and *inactive* constraints.

Definition (*Active/Inactive constraints*)
Suppose $h(x) = (h_1(x), \dots, h_\ell(x))^\mathsf{T}$. Then the j-th constraint h_j is said to be *active* at some feasible point x, if $h_j(x) = 0$. Otherwise (i.e., $h_j(x) < 0$) it is said to be *inactive*.

We also need the general Lagrangian functions.

Definition (*Lagrangian*)
Let λ and μ denote the multipliers for equality constraints and inequality constraints, respectively. Then the Lagrangian function is given by

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{\ell} \mu_j h_j(x) = f(x) + \lambda^\mathsf{T} g(x) + \mu^\mathsf{T} h(x)\,.$$

Then we have the following Karush-Kuhn-Tucker (KKT) conditions.

Theorem (*KKT conditions*)
Suppose x^* is a feasible solution (namely, $g(x) = 0$ and $h(x) \le 0$) and is *regular* for all equality constraints and all *active* inequality constraints. If x^* is a *local optimal point* then there exists Lagrange/KKT multipliers λ^*, μ^* such that

- $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \mathbf{0}$, that is, $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^\ell \mu_j^* \nabla h_j(x^*) = \mathbf{0}$.
- $\mu_j^* \ge 0$ for all $j = 1, 2, \dots, \ell$.
- $\mu_j^* h_j(x^*) = 0$ for all $j = 1, 2, \dots, \ell$.
- $g(x^*) = 0$ and $h(x^*) \le 0$.

6.2 Duality of general optimization

Let D denote the domain of the problem, and Ω denote the feasible set. We usually assume $D = \mathbb{R}^n$ unless otherwise specified.

Lagrange dual functions and problems

For any $x \in D$, if x is a feasible solution, then $g(x) = 0$ and $h(x) \leq 0$. So for any λ and $\mu \geq \mathbf{0}$, we have

$$f(x) = \mathcal{L}(x, \lambda, \mathbf{0}) \geq \mathcal{L}(x, \lambda, \mu)\,.$$

It yields that $f(x) = \max_{\mu \geq \mathbf{0}, \lambda} \mathcal{L}(x, \lambda, \mu)$. If x is infeasible, then $\mathcal{L}(x, \lambda, \mu)$ can be unbounded above, and thus $\max_{\mu \geq \mathbf{0}, \lambda} \mathcal{L}(x, \lambda, \mu) = \infty$. So, the original problem is equivalent to

$$\inf_{x \in D} \sup_{\mu \geq \mathbf{0}} \mathcal{L}(x, \lambda, \mu)\,.$$

We now consider another optimization direction. For any fixed λ and μ, let

$$\phi(\lambda, \mu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \mu)\,.$$

Then $\phi(\lambda, \mu)$ is called the *Lagrange dual function*. Let Ω be the feasible set. It is easy to see that

$$\phi(\lambda, \mu) \leq \inf_{x \in \Omega} \mathcal{L}(x, \lambda, \mu) \leq \inf_{x \in \Omega} f(x)$$

if $\mu \geq \mathbf{0}$. So $\phi(\lambda, \mu)$ gives a lower bound to the original optimization problem. We would like to maximize the dual function, which is equivalent to solve the following optimization:

$$\begin{aligned} \max \quad & \phi(\lambda, \mu) \\ \text{subject to} \quad & \mu \geq \mathbf{0}\,. \end{aligned}$$

This optimization problem is called the *Lagrange dual problem*.

An interesting fact is that dual functions are always *concave*, thus dual problems are always convex, regardless of whether the primal problem is convex or not.

Remark: This fact reveals that the dual problem of the dual is not necessary the primal, since the primal problem may not be convex.

Theorem
For any (not necessarily convex) optimization problem, its Lagrange dual function is concave.

▼ Proof
For any fixed x, $\mathcal{L}(x, \lambda, \mu)$ is an affine function of λ and μ, so

$$\phi(\lambda, \mu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \mu)$$

is a pointwise minimum of a family of affine functions, which implies that $-\phi(\lambda, \mu)$ is a convex function.

Weak and strong duality

Let f^* and ϕ^* be the optimal value of the primal and the dual, respectively. Then we have the following result.

Theorem (*Weak duality*)

$$f^* \geq \phi^* .$$

The optimal value of the dual is a lower bound of the optimal value of the primal. In the linear programming case, we have showed that the optimal value of the primal is identical to the optimal value of the dual. However it may not be true in general.

Definition (*Strong duality*)
We say the *strong duality* holds if $f^* = \phi^*$.

The duality has a geometric interpretation. Let C be the region

$$C \triangleq \{(p, q, t) : \exists\, x \in D \text{ s.t. } h(x) \leq p, g(x) = q, f(x) \leq t\} .$$

Then $(\mathbf{0}, \mathbf{0}, f^*)$ is the lowest intersection point of C and the f-axis, and $(\mathbf{0}, \mathbf{0}, \phi^*)$ is the highest one over all intersection point of supporting hyperplanes to C and the f-axis.

Duality condition for convex optimization

Using the geometric interpretation of the duality, we have the following propositions and conditions to justify strong duality.

Proposition If the optimization problem is convex, then C is a convex set.

Theorem (*Slater's condition*)
Suppose the optimization problem is convex, and there exists a *strictly* feasible solution $x_0 \in \operatorname{rel\,int} D$ such that

$$h_j(x_0) < 0$$

for all inequality constraints. Then the strong duality holds.

Remark: The Slater's condition can be refined. It is easy to see that if $\exists\, x_0 \in \operatorname{rel\,int} D$ such that $h_j(x_0) < 0$ for all **nonlinear** inequality constraints h_j, then strong duality holds.

Finally, we restate the KKT condition (using duality terms) and reveal the connection between KKT conditions and the strong duality.

Definition (*KKT conditions*, restated)
Consider an optimization problem. Let x^* be a variable and λ^*, μ^* be the multipliers of equality and inequality constraints respectively. Then we say KKT conditions hold at x^* with Lagrange (or KKT) multipliers λ^*, μ^* if all of the followings hold.

1. (*Stantionarity*) $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \mathbf{0}$.
2. (*Primal feasibility*) $g(x^*) = \mathbf{0}$ and $h(x^*) \leq \mathbf{0}$.
3. (*Dual feasibility*) $\mu^* \geq \mathbf{0}$.
4. (*Complementary slackness*) $\mu_j^* h_j(x^*) = 0$ for all $j = 1, 2, \ldots, \ell$.

Theorem
For any convex optimization problem, if x^* has multipliers λ^* and μ^* satisfying KKT conditions, then strong duality holds. In particular, x^* is an optimal solution to the primal problem, and (λ^*, μ^*) is an optimal solution to the dual problem.

Remark: KKT conditions are not necessary for optimality, but are necessary if with regularity. Moreover, for convex problems, KKT conditions are sufficient for optimality.

Theorem
For any optimization problem, if the primal has a (finite) optimal solution x^*, the dual has a (finite) optimal solution (λ^*, μ^*), and the strong duality holds, then (λ^*, μ^*) are multipliers of x^*, satisfying KKT conditions.

6.3 Algorithms: projected gradient descent

As the last section, we present an algorithm for general optimization problems. Recall the gradient method for unconstrained optimization use the update rule $x_{k+1} = x_k - \eta_k \nabla f(x_k)$ for the k-th iteration. When there are constraints, x_{k+1} may not be feasible, even if x_k is. One approach to deal with it is *projecting* x_{k+1} onto the feasible set Ω. Let $y_k = x_k - \eta_k \nabla f(x_k)$. Then x_{k+1} is the projection of y_k onto Ω, namely,

$$x_{k+1} = \mathcal{P}_\Omega(y_k) \triangleq \arg\min_{x \in \Omega} \|x - y_k\|^2 .$$

This method is called the *projected gradient descent*. The projected gradient descent is a special case of the proximal gradient descent. In fact, consider the indicator function

$$I_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ +\infty & \text{if } x \notin \Omega \end{cases} .$$

Then $I_\Omega(\cdot)$ is a convex function if Ω is convex, and thus the projection operator $\mathcal{P}_\Omega(\cdot)$ is exactly the same as the proximal operator $\mathrm{prox}_{I_\Omega}(\cdot)$.