9

CONVEX ANALYSIS AND CONVEX OPTIMIZATION

9 CONVEX ANALYSIS AND CONVEX OPTIMIZATION 428
9.1 Affine sets 430
9.1.1 Basic concepts 430
9.1.2 Affine independence and dimensions 432
9.2 Convex sets and properties 436
9.2.1 Concepts of convex sets 436
9.2.2 Projection theorems 438
9.2.3 Separation theorems 439
9.2.3.1 Separating hyperplane theorem 439
9.2.3.2 Farka's lemma 441
9.3 Convex functions 445
9.3.1 Basic concepts 445
9.3.2 Connection to convex set 447
9.3.3 Strongly convex functions 447
9.3.4 Operations preserve convexity 448
9.3.5 Convexity and derivatives 449
9.3.6 Subgradient 451
9.4 Duality theory 453
9.5 Convex optimization and optimality conditions 457
9.5.1 Local optimality vs. global optimality 457
9.5.2 Unconstrained optimization optimality conditions 458
9.5.3 Constrained optimization optimality conditions 458

9.6 Subgradient methods 463
9.6.1 A generic algorithm for unconstrained problem 463
9.6.2 Convergence under Lipschitz smoothness 465
9.6.3 Projected gradient methods 468
9.6.3.1 Foundations 468
9.6.3.2 Algorithms 470
9.6.4 Proximal gradient methods 471
9.6.4.1 Foundations 471
9.6.4.2 Algorithms 472
9.6.4.3 Case study: sparsity regularization problem 473
9.7 Notes on bibliography 475

9.1 Affine sets

9.1.1 Basic concepts

Definition 9.1.1 (affine sets). *A set $\subset \mathbb{R}^d$ is affine set if for all $x, y \in X$ and $\lambda, \gamma \in \mathbb{R}, \lambda + \gamma = 1$, then $\lambda x + \gamma y \in X$*

Lemma 9.1.1. *Arbitrary intersection of affine sets is an affine set.*

Proof. Let $x, y \in X \cap Y$, then $\lambda x + \gamma y \in X \cap Y$. □

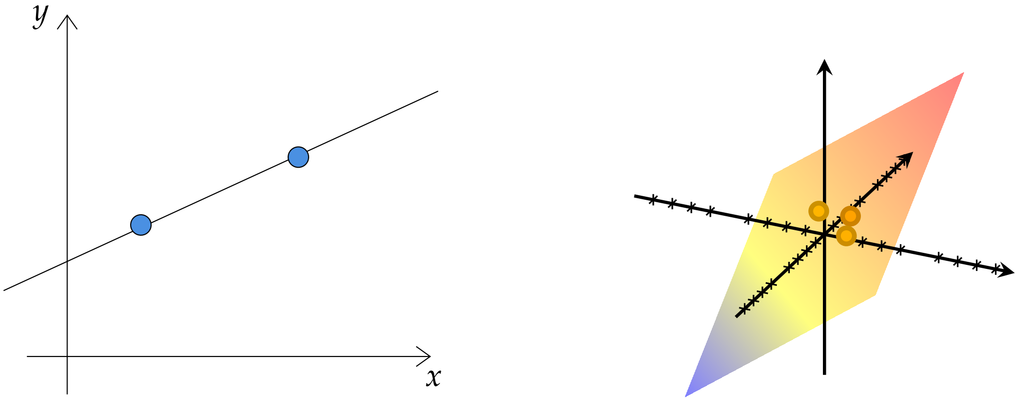
Definition 9.1.2 (affine combination). *Given $x, y \in \mathbb{R}^d$, the affine combination z of x, y refers to*

$$z = ax + by, a + b = 1$$

Definition 9.1.3 (affine hull). *The affine hull of X is defined as*

$$aff(X) = \{\lambda_1 x_1 + ... + \lambda_t x_t : x_1, ..., x_t \in X, \sum_i \lambda_i = 1\}$$

That is, the affine hull the the set of all affine combinations of points in X.



(a) The affine hull of two points in a plane is a line passing through them.

(b) The affine hull of three co-plane points is a plane containing them them.

Figure 9.1.1: Example 2D affine hull and 3D affine hull.

Lemma 9.1.2 (characterization via linear equations). *[1, p. 5]*

- *The solution to the linear equation $Ax = 0$ form a linear subspace X. Moreover, every linear subspace set can be represented as*

$$\{x : Ax = 0\}$$

- *The solution to the linear equation $Ax = b$ form an affine subspace X. Moreover, every affine set can be represented as*

$$\{x : Ax = b\}$$

Proof. (1) (a) nullspace is subspace; (b) consider the orthogonal complement $X^\perp$, and let basis of $X^\perp$ be the rows of A. (2) (a) Let x_1, x_2 be such that $Ax_1 = b, Ax_2 = b$, then

$$A(ax_1 + (1-a)x_2) = ab + (1-a)b = b$$

therefore $ax_1 + ax_2$ is also the solution. Thus the solution form an affine subspace. (b) Let L be the associated linear subspace, then consider the orthogonal complement $L^\perp$, and let basis of $L^\perp$ be the rows of A. Since $X = L + a$, for some $a \in \mathbb{R}^d$, then $X = \{x : A(x-a) = 0\} = \{x : Ax = Aa\}$. □

Definition 9.1.4 (parallel relation in affine sets). *An affine set M is said to be* ***parallel*** *to an affine set L if there exist some a such that*

$$M = L + a$$

Lemma 9.1.3 (translational invariance in affine subspace). *If $X \subset \mathbb{R}^d$ is an affine subspace, then for every $x_0 \in \mathbb{R}^d$, $X - x_0$ is still an affine subspace. That is, affine subspace is translational invariant, and shifted affine subspaces are parallel to each other.*

Proof. Let $x, y \in X$, then $x - x_0, y - x_0 \in X - x_0$, and then we can verify

$$a(x - x_0) + b(x - x_0), a + b = 1$$

also belongs to $X - x_0$. □

Lemma 9.1.4 (affine subspace to linear subspace). *[1, p. 4] If an affine subspace X contains 0, then X is also an linear space.*

Proof. (1) closedness under scalar multiplication: Let $x \in X$, then

$$\lambda x + (1-\lambda)0 = \lambda x \in X, \forall \lambda \in \mathbb{F}$$

(2) closedness under addition: Let $x, y \in X$, then $x + y = 2(\frac{1}{2}(x+y)) \in X$, since $\frac{1}{2}(x+y) \in X$ and X is closed under scalar multiplication. Other properties can be verified easily. □

Lemma 9.1.5 (conversion from affine subspace to linear subspace). *If $X \subset \mathbb{R}^d$ is an affine subspace, then for every $x_0 \in X$, $X - x_0$ is a linear space. Moreover,*

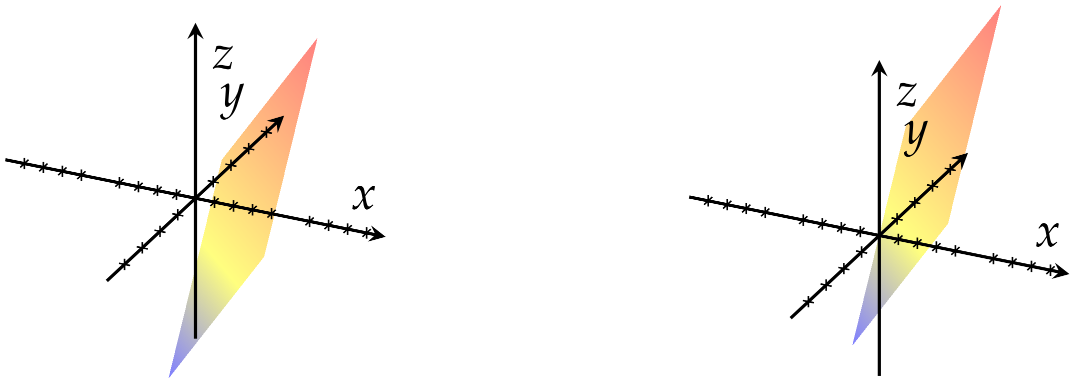
$$X - x_1 = X - x_2, \forall x_1 \neq x_2, x_1, x_2 \in X$$

Proof. (1) $X - x_0$ is an affine subspace containing 0. (2) Let $L_1 = X - x_1$ and $L_2 = X - x_2$, then $L_1 = L_2 + a, a = x_1 - x_2$. Then since L_1 contains 0 then $-a \in L_2$, further implying $a \in L_2$, because $L_2 = a + L_2, a \in L_2$ due to closedness of vector space, therefore

$$L_1 = L_2$$

□

Corollary 9.1.0.1 (associated unique linear subspace). *Every non-empty affine set M is parallel to a unique linear subspace L, which is given as $L = M - x, x \in M$.*



(a) An affine 2D surface that does not contain the origin.

(b) An affine 2D surface that contains the origin; thus it is also a subspace.

Figure 9.1.2: Affine subspace and linear subspace.

9.1.2 Affine independence and dimensions

Definition 9.1.5 (dimension of an affine set). *[1, p. 4] The dimension of an affine set is defined as the dimension of its parallel linear subspace.*

Definition 9.1.6 (affine independence). *A set X is affinely independent if there is no $x \in X$ such that $x \in aff(X - \{x\})$. That is, for every element in X, it can not be written as affine combination of the others.*

Definition 9.1.7 (affine independence,alternative). *[1, p. 7] A set X of $m+1$ points $b_0, b_1, ..., b_m$ are affinely independent if $aff(X)$ has dimension of m; that is the linear subspace*

$$X - b_i, i \in \{0, 1, ..., m\}$$

has dimension of m; or equivalently, the set

$$b_1 - b_0, b_2 - b_0, ..., b_m - b_0$$

are linearly independent.

Lemma 9.1.6 (algebraic criterion for affine independence). *[2, p. 23] The elements $x_0, ..., x_m$ in $\mathbb{R}^n, m \geq 1$, are affinely independent if and only if*

$$\sum_{i=1}^{m} \lambda_i v_i = 0, \sum_{i=0}^{m} \lambda_i = 0 \Rightarrow \lambda_i = 0, \forall i = 0, ..., m$$

Proof. (1) Suppose

$$\sum_{i=0}^{m} \lambda_i v_i = 0, \sum_{i=0}^{m} \lambda_i = 0 \Rightarrow \lambda_i = 0, \forall i = 0, ..., m$$

We can rewrite as

$$\sum_{i=1}^{m} \lambda_i (v_i - v_0) = 0 \Rightarrow \lambda_i = 0, \forall i = 1, ..., m$$

which implies $v_i - v_0, i = 1, ..., m$ are linearly independent. Therefore $v_0, ..., v_m$ are affinely independent. (2) Suppose affinely independent. WLOG, suppose $\lambda_0 \neq 0$, we have

$$\lambda_0 v_0 = -\sum_{i=1}^{m} \lambda_i v_i$$

add both sides with $\sum_{i=1}^{m} \lambda_i v_0$, we have

$$0 = \sum_{i=1}^{m} \lambda_i (v_i - v_0)$$

implies $\lambda_i = 0, i = 1, ..., m \Rightarrow \lambda_0 = 0$, contradicts $\lambda_0 \neq 0$. □

Remark 9.1.1 (linear space analog). A set X is linearly independent if every $x \in X$ cannot be written as the linear combination of the others. Or there is no $x \in X$ such that $x \in span(X - \{x\})$

Theorem 9.1.1 (Barycentric coordinate system). *Let $b_0, b_1, ..., b_m$ be affinely independent, let $M = aff(b_0, ..., b_m)$. Then for any $x \in M$, we have a **unique affine combination** representation for x, given as*

$$x = \sum_{i=0}^{m} a_i b_i, \sum_{i=0}^{m} a_i = 1, a_i \in \mathbb{R}$$

We can view this as the Barycentric coordinate system.

Proof. Because $b_0, b_1, ..., b_m$ be affinely independent, then the set

$$b_1 - b_0, ..., b_m - b_0$$

are linearly independent. Because $x \in M$, then x has representation of

$$x = \sum_{i=0}^{m} a_i b_i, \sum_{i=0}^{m} a_i = 1$$

To show uniqueness, we subtract out b_0 on both side, we have

$$x - b_0 = \sum_{i=1}^{m} a_i (b_i - b_0)$$

Suppose we also have

$$x - b_0 = \sum_{i=1}^{m} c_i (b_i - b_0)$$

Then we have

$$\sum_{i=1}^{m} (c_i - a_i)(b_i - b_0) = 0$$

implies $c_i = a_i, i = 1, 2, ..., m$ due to linear independence of $b_1 - b_0, ..., b_m - b_0$. □

Theorem 9.1.2. *Let $X \subset \mathbb{R}^d$. Then the following are equivalent:*

- *X is an affinely independent*
- *For every $x \in X$, the set $\{v - x : v \in X - \{x\}\}$ is linearly independent.*
- *There exists $x \in X$ such that the set $\{v - x : v \in X - \{x\}\}$ is linearly independent.*
- *The set of vectors $\{(x, 1) \in \mathbb{R}^{d+1}, x \in X\}$ is linearly independent.*
- *X is finite set with vectors $x_1, ..., x_m$ such that*

  $$\lambda_1 x_1 + ... + \lambda_m x_m = 0, \lambda_1 + ... + \lambda_m = 0$$

  implies $\lambda_1 = ... = \lambda_m = 0$

Theorem 9.1.3. *Let $X \subset \mathbb{R}^d$. The following are equivalent.*

- *X is an affine subspace*
- *For every $x \in X$, the set $X - x$ is a linear subspace of dimension $0 \leq m \leq d$*
- *There exist affinely independent vectors $x_1, ..., x_{m+1}$ for some $0 \leq m \leq d$ such that every $x \in X$ can be written as $x = \sum_i \lambda_i x_i, \sum_i \lambda_i = 1$*
- *There exists a matrix $A \in \mathbb{R}^{(d-m)\times d}$ with full row rank and a vector $b \in \mathbb{R}^m$ for some $0 \leq m \leq d$ such that*
  $$X = \{x \in \mathbb{R}^d : Ax = b\}$$

9.2 Convex sets and properties

9.2.1 Concepts of convex sets

Definition 9.2.1 (convex set, strictly convex set).

- *A set S is said to be a convex set if and only if for any two points $x, y \in S$, we have*
$$\alpha x + (1-\alpha)y \in S, \forall \alpha \in [0,1].$$
- *A set S is said to be a strictly convex set if and only if for any two points $x, y \in S$, we have*
$$\alpha x + (1-\alpha)y \in intS, \forall \alpha \in (0,1),$$
where $intS$ is the interior of S.

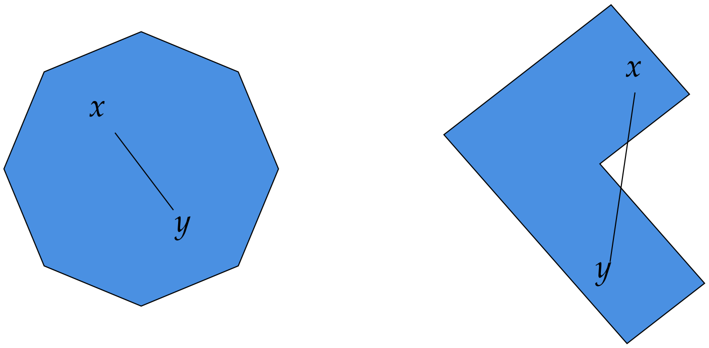


Figure 9.2.1: (left) A convex set. (right) A non-convex set.

Example 9.2.1.

- A convex set can be either open or closed. For example, (a,b) and $[a,b]$ are both convex sets.
- Let $a \in \mathbb{R}^d, \delta \in \mathbb{R}$, then the set $H = \{x \in \mathbb{R}^d : \langle a, x \rangle = \delta\}$ is called hyperplane. The set $H^+ = \{x \in \mathbb{R}^d : \langle a, x \rangle \geq \delta\}$ and $H^- = \{x \in \mathbb{R}^d : \langle a, x \rangle \leq \delta\}$ are halfspaces. Hyperplanes and halfspaces are all convex sets.

Definition 9.2.2 (dimensionality of a convex set). *Let X be a convex set, then $dim(X)$ equals to the maximum number of affinely independent points in X.*

Lemma 9.2.1 (preservation of convexity).

- *The intersection of **arbitrary** collection of convex sets $\cap_{i\in I} C_i$ is a convex set.*
- *The image of any linear function f defined on a convex set X is convex.*
- *If X is convex, then the set $aX = \{ax : x \in X\}, a \in \mathbb{R}$ is also convex. Particularly, $-X$(the symmetric set with respect to origin) is convex.*
- *If X, Y are convex, then $X + Y$ and $X - Y = X + (-Y)$ is also convex.*

Proof. (1) Let $x, y \in \cap_{i\in I} C_i$, then its convex optimization belongs to any convex set C_i, and therefore in $\cap_{i\in I} C_i$.(2)$af(x) + (1-a)f(y) = f(ax + (1-a)y) \in Im(f)$. (3)(4) directly follow from definition,(4) will use (3). □

Remark 9.2.1 (Empty set is convex). Note that intersection might result in empty set. Even so, it does not violate above lemma, since empty set is also convex.

Lemma 9.2.2. *[3, p. 44] A convex combination of a finite number of elements of a convex set X also belongs to that set.*

Proof. Use induction to prove, starting from n=2. Suppose $n = k$ holds, then for $n = k+1$, we have

$$\lambda_0 a_0 + \lambda_1 a_1 + ... + \lambda_k a_k = \lambda_0 a_0 + (1-\lambda_0)(\frac{\lambda_1}{1-\lambda_0} a_1 + ... + \frac{\lambda_k}{1-\lambda_0} a_k) \in X, \sum_i \lambda_i = 1$$

since $(\frac{\lambda_1}{1-\lambda_0} a_1 + ... + \frac{\lambda_k}{1-\lambda_0} a_k) \in X$ by assumption. □

Theorem 9.2.1 (convex hull as the smallest containing affine set). *The convex hull of X, denoted as $conv(X)$,given as*

$$conv(X) = \{\lambda_1 x_1 + ... + \lambda_t x_t : x_1, ..., x_t \in X, \lambda_1, ..., \lambda_t \geq 0, \sum_i \lambda_i = 1\}$$

is the smallest convex hull containing X. Moreover, a set X is convex if $X = conv(X)$.

Proof. It is easy to see $X \subseteq conv(X)$. To prove it is the smallest: let C be any other convex set that containing X, then we have $conv(X) \subseteq C$ due to above lemma. □

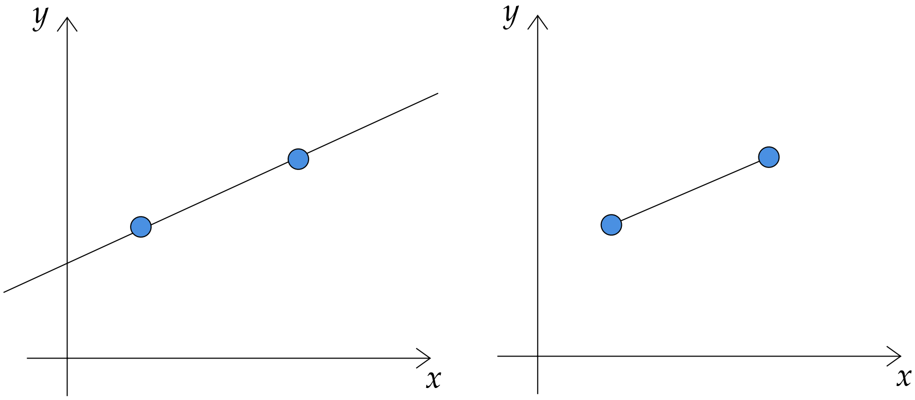


Figure 9.2.2: (left) The affine hull of two points in a plane is a line passing through them. (right) The convex hull of two points in a plane is a line segment containing them them.

9.2.2 Projection theorems

Definition 9.2.3 (projection onto a convex set). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a nonempty closed convex set. The projection of a given vector $y \in \mathbb{R}^n$ onto $\mathcal{X}$ is*

$$Proj_{\mathcal{X}}(y) = \arg\min_{x \in \mathcal{X}} \|x - y\|_2 .$$

Theorem 9.2.2 (projection theorem for convex set). *[4, p. 19] Let $\mathcal{X}$ be a closed convex set in $\mathbb{R}^n$, let $\|.\|$ be the Euclidean norm. Then, we have*

- *For every $x \in \mathbb{R}^n$, the following optimalization problem*

$$\min_{y \in X} \|y - x\|^2$$

  has an unique global minimum solution x^*.
- *(obtuse angle theorem) $x^* = Proj_{\mathcal{X}}(y), y \in \mathbb{R}^n$ is the unique global minimum if and only if*

$$(y - x^*)^T (x - x^*) \leq 0, \forall x \in \mathcal{X}$$

- *The projection mapping $Proj_{\mathcal{X}}(y) : \mathbb{R}^n \to \mathcal{X}$ is continuous and non-expansive mapping, i.e.*

$$\|Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2)\| \leq \|y_1 - y_2\| .$$

Proof. (1) For every $x \in \mathcal{X}$, it is easy to see that the function $f(y) = \|y - x\|^2$ is coercive in the set $\mathcal{X}$. Therefore, from Theorem 3.1.7, we have at least one global minimizer in $\mathcal{X}$.

This minimizer will be unique [Lemma 9.5.2] since $f(y) = \|y-x\|^2$ is strictly convex(The Hessian is positive definite matrix).
(2) Use the first order necessary and sufficient condition for $f(x)$ [section 9.5] gives

$$\nabla f^T(x-x^*) \geq 0, \forall x \in \mathcal{X} \Leftrightarrow (x-x^*)^T(x^*-y) \geq 0, \forall x \in \mathcal{X}$$

(3) From (2), we have

$$\begin{aligned} &(y_1 - Proj_{\mathcal{X}}(y_1))^T(x - Proj_{\mathcal{X}}(y_1)) \leq 0, \forall x \in \mathcal{X} \\ \implies &(y_1 - Proj_{\mathcal{X}}(y_1))^T(Proj_{\mathcal{X}}(y_2) - Proj_{\mathcal{X}}(y_1)) \leq 0, \forall x \in \mathcal{X} \\ &(y_2 - Proj_{\mathcal{X}}(y_2))^T(Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2)) \leq 0, \forall x \in \mathcal{X} \end{aligned}$$

Add together, we have

$$\begin{aligned} -(Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2))^T(y_1 - y_2 - (Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2))) \leq 0 \\ \|Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2)\|^2 \leq (Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2))^T(y_1 - y_2) \\ \leq \|Proj_{\mathcal{X}}(y_1) - Proj_{\mathcal{X}}(y_2)\| \|y_1 - y_2\| \end{aligned}$$

where we use the Cauchy inequality.

Nonexpansiveness directly implies continuity: as $y_1 \to y_2$, $Proj_{\mathcal{X}}(y_1) \to Proj_{\mathcal{X}}(y_2)$. □

9.2.3 Separation theorems

9.2.3.1 *Separating hyperplane theorem*

Theorem 9.2.3 (separating hyperplane theorem). *[3, p. 170] Let $C \subset \mathbb{R}^d$ be a closed convex set and let $x \in \mathbb{R}^d, \notin C$. Then there exists a halfspace that contain C and does not contain x. More precisely, there exists $a \in \mathbb{R}^d, \delta \in \mathbb{R}$ such that $\langle a, y \rangle \leq \delta$ for all $y \in C$ and $\langle a, x \rangle > \delta$; that is*

$$\langle a, x \rangle > \langle a, y \rangle, \forall y \in C;$$

or equivalently,

$$\langle a, x \rangle > \sup_{y \in C} \langle a, y \rangle.$$

Proof. (1) If C is empty, then any halfspace that does not contain x will suffice. (2) If C is empty, consider a closed ball $B(x,r)$,where $r = \|x - \bar{x}\|, \bar{x} \in C$, then the set $B(x,r) \cap C$ is a closed and compact set, and therefore there exists x^* as the minimizer of problem $\min_{y\in C\cap B}\|y-x\|$. Let $a = x - x^*$, then from obtuse angle theorem(projection theorem)Theorem 9.2.2 for convex set, we have

$$\langle y - x^*, x - x^* \rangle \leq 0$$

which implies

$$\langle y - x^*, a \rangle \leq 0 \Leftrightarrow \langle a, y \rangle \leq \langle a, x^* \rangle < \langle a, a + x^* \rangle = \langle a, x \rangle$$

□

Corollary 9.2.3.1 (separating hyperplane theorem for convex bodies). *Let C_1 and C_2 be two convex subsets in $\mathbb{R}^n$. If C_1 and C_2 are disjoint, then there exists $a \in \mathbb{R}^n$ such that*

$$\langle a, x_1 \rangle \leq \langle a, x_2 \rangle, \forall x_1 \in C_1, x_2 \in C_2$$

Proof. Consider the convex set $C = C_1 + (-C_2)$ (see Lemma 9.2.1]. Since C does not contain origin, from Theorem 9.2.3, there exists $a \in \mathbb{R}^m, \delta \in \mathbb{R}^m$ such that

$$\langle a, x \rangle \leq \langle a, 0 \rangle = 0 \Rightarrow \langle a, x_1 - x_2 \rangle \leq 0, \forall x_1 \in C_1, x_2 \in C_2$$

□

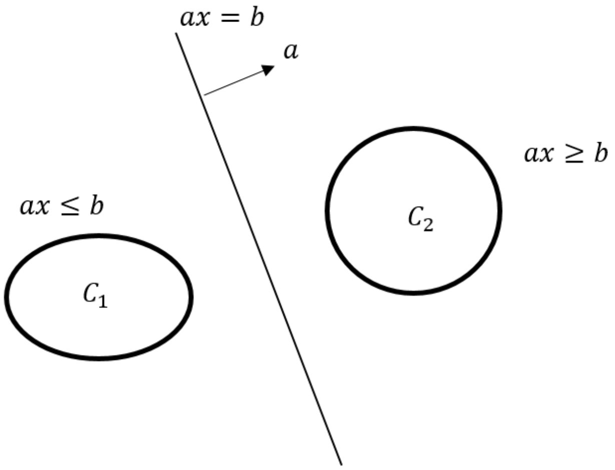


Figure 9.2.3: An illustration of separating hyperplane theorem for two convex bodies.

Corollary 9.2.3.2. *For every closed convex set $X \subseteq \mathbb{R}^d$ there exists a family of tuples $(a^i, \delta^i), a^i \in \mathbb{R}^d, \delta^i \in \mathbb{R}, i \in I$(where I may be an uncountable index set) such that $X =$*

$\cap_{i\in I} H^-(a^i, \delta^i)$. In otherwise, every closed convex set can written as the intersection of some family of halfspaces.

Corollary 9.2.3.3. *Every closed convex set C can e written as the intersection of some family of halfspaces.*

Proof. Consider all the points x_i outside C, then based on separating hyperplane theorem, there exists an halfspaces $H^-(a_i, \delta_i)$, such that $C \subseteq H^-(a_i, \delta_i)$. The intersection of all such halfspaces will be C since every point outside C will not be included in the intersection. □

Definition 9.2.4 (supporting hyperplane). *Given a set $C \subseteq \mathbb{R}^n$. If a vector x_0 belongs to $cl(C)$, a hyperplane with parameter $a \in \mathbb{R}^n, \delta \in R$ such that*

$$\langle a, x \rangle \leq \delta, \langle a, x_0 \rangle = \delta$$

is called a supporting hyperpalne for C.

Theorem 9.2.4 (supporting hyperplane theorem). *[5, p. 67] Let $C \subseteq \mathbb{R}^d$ be a convex set and let $x \in bd(C)$ (the boundary of C). Then there exists $a \in \mathbb{R}^d, \delta \in \mathbb{R}$ such that*

- *$\langle a, y \rangle \leq \delta$ for all $y \in C$; that is, $C \subseteq H^-(a, \delta)$*
- *$\langle a, x \rangle = \delta$*

The hyperplane $\{y \in \mathbb{R}^d : \langle a, y \rangle = \delta\}$ is called a supporting hyperplane for C at x.

Proof. See reference. □

9.2.3.2 *Farka's lemma*

Theorem 9.2.5 (Farkas's lemma, analog in linear equation theory). *Let $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$. Exactly one of the following is true:*

- *$Ax = b$ has a solution*
- *There exists $u \in \mathbb{R}^m$ such that $A^T u = 0$ and $u^T b \neq 0$*

Proof. (1) If $Ax = b$ has a solution, then $b \in \mathcal{R}(A)$. Suppose there exists a u such that $A^T u = 0$, which implies $u \in \mathcal{N}(A^T)$. Since $b \in \mathcal{R}(A), u \in \mathcal{N}(A^T), \mathcal{R}(A) \perp \mathcal{N}(A^T)$, it is impossible to have $u^T b \neq 0$;
(2) If $A^T u = 0$ has a solution u and $u^T b \neq 0$. Then b has nonzero component in $\mathcal{N}(A^T)$, and therefore b cannot lie in the $\mathcal{R}(A)$. □

Theorem 9.2.6 (Farkas' lemma). *Let $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m, b \neq 0$. Exactly one of the following is true:*

- *$Ax = b, x \geq 0$ has a solution.*
- *There exists $u \in \mathbb{R}^m$ such that $A^T u \leq 0$ and $u^T b > 0$.*

Or equivalently, exactly one of the following sets must be empty

- *$\{x | Ax = b, x \geq 0\}$.*
- *$\{u | A^T y \leq 0, b^T y > 0\}$.*

Proof. (1) Suppose $Ax = b, x \geq 0$ holds, then we suppose that there exists $u \in \mathbb{R}^m$ such that $A^T u \leq 0$. Multiply x^T, we have $(Ax)^T u \leq 0 \Rightarrow b^T u \leq 0$, which contradicts $u^T b > 0$. Therefore, when the first case holds, the second case cannot hold at the same time. (2) We can view each column a_i of A is a point in $\mathbb{R}^m$, then the set $C = \{y = Ax, x \geq 0\}$ form a cone. $Ax = b, x \geq 0$ has a solution can be interpreted as b is lying in the cone. Suppose b lying outside the cone, then use Theorem 9.2.3 we know that there exists $u \in \mathbb{R}^m, \delta \in \mathbb{R}$ such that

$$\langle y, u \rangle \leq \delta, \forall y \in C, \langle u, b \rangle > \delta$$

Also, 0 is in the cone, we have $\delta \geq 0$, therefore $\langle u, b \rangle > 0$. To show $\langle a_i, u \rangle \leq 0, \forall i$, suppose there exist some $\langle a_i, u \rangle > 0$, then there exists some scalar $\lambda > 0$ (we can always choose λ large enough) such that $\lambda \langle a_i, u \rangle > \delta$, then $\langle \lambda a_i, u \rangle > \delta$, contradicting the fact that every point in C satisfying $\langle \lambda a_i, u \rangle \leq \delta$. Therefore, $\langle a_i, u \rangle \leq 0, \forall i$ □

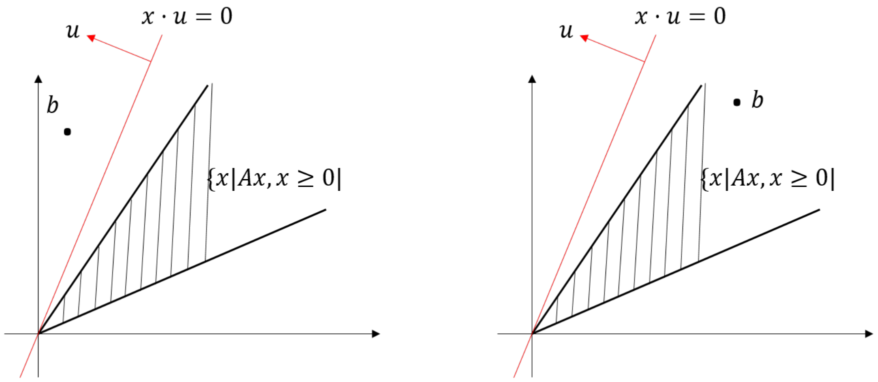


Figure 9.2.4: An illustration of Farkas' lemma. (left) When b lies outside the cone (that is, $Ax = b, x \geq 0$ has no solution), there exists a hyperplane, characterized by normal vector y, seperating b and the cone. (right) When b lies inside the cone (that is, $Ax = b, x \geq 0$ has a solution), there does not exist a hyperplane, characterized by normal vector y, seperating b and the cone.

Corollary 9.2.6.1 (Farkas' lemma variant). *Let $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m, b \neq 0$. Exactly one of the following is true:*

- *$Ax = b, x > 0$ has a solution.*
- *There exists $u \in \mathbb{R}^m$ such that $A^T u \leq 0$ and $u^T b > 0$ (or $A^T u < 0$ and $u^T b \geq 0$).*

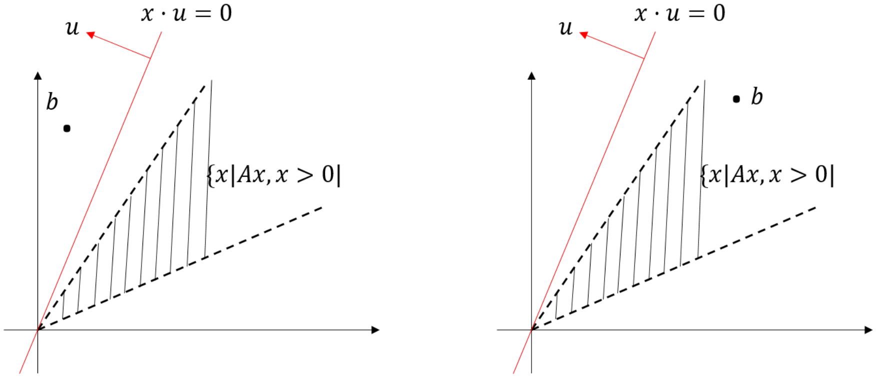


Figure 9.2.5: An illustration of Farkas' lemma variant where the cone is open set. (left) When b lies outside the cone (that is, $Ax = b, x > 0$ has no solution), there exists a hyperplane, characterized by normal vector y, seperating b and the cone. (right) When b lies inside the cone (that is, $Ax = b, x > 0$ has a solution), there does not exist a hyperplane, characterized by normal vector y, seperating b and the cone.

Remark 9.2.2 (financial application). Farkas' lemma in no-arbitrage pricing can be found in [3, p. 168].

9.3 Convex functions

9.3.1 Basic concepts

Definition 9.3.1 (convex function, concave function).

- *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* ***convex function*** *is for $x_1, x_2 \in \mathbb{R}^n, \alpha \in [0,1]$, then*

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

- *A function f is* ***concave*** *if $-f$ is convex.*
- *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* ***strictly convex function*** *is for $x_1, x_2 \in \mathbb{R}^n, x_1 \neq x_2, \alpha \in (0,1)$, then*

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

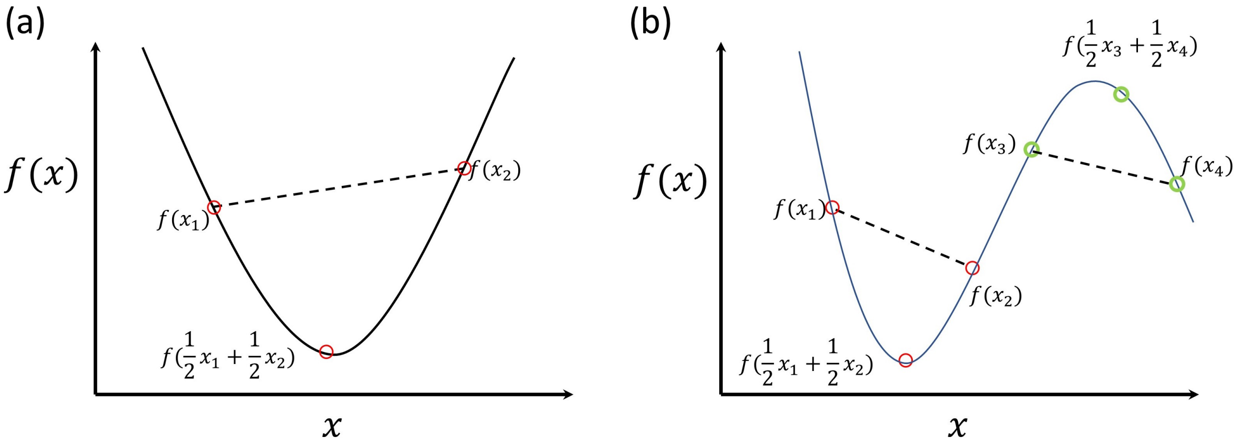


Figure 9.3.1: Demonstration of convex functions. (a) A convex function satisfying $f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$. (b) A non-convex function where the green points does not satisfy the relation.

Note 9.3.1 (the domain needs to convex). **Note that the domain C here also has to be convex.** We emphasize that the domain have to convex to avoid the situation that when x_1 and x_2 are in S but their convex combination is not in S.

Example 9.3.1 (Examples of convex and concave functions).

- Affine functions $Ax + b$ are both convex and concave.
- Power function x^a is convex for $a \geq 1$. concave if $a \in (0,1]$.

- Negative entropy $x\ln(x), x > 0$ is convex
- $\log(x)$ is concave.
- For $x \in \mathbb{R}^n$, 2-norm $\|x\|_2^2 = x^T x$ is both convex and strictly convex.

Definition 9.3.2 (proper and closed convex functions). *[5, p. 8] Let $\mathcal{X} \subseteq \mathbb{R}^n$($\mathcal{X}$ is not necessarily a convex set) and $f : \mathcal{X} \to \bar{\mathbb{R}}$ with*

$$\bar{\mathbb{R}} \triangleq \mathbb{R} \cup \{-\infty, \infty\}.$$

We have the following definitions.

- *The **effective domain** of f is*

$$dom(f) \triangleq \{x \in \mathcal{X} : f(x) < \infty\}.$$

- *The function f is said to be **proper** if*

$$f(x) > -\infty, \forall x \in \mathbb{R}^n, \text{ and } f(x) < \infty \text{ for some } x \in \mathbb{R}^n.$$

- *The function f is said to closed if its epigraph is closed.*

Remark 9.3.1. A affine function with nonzero slope is not a proper function.

Theorem 9.3.1 (Convexity and continuity). *(convexity implies continuity) If f is convex, then it is continuous.*

Proof. (informal) consider the epigraph a discontinuous function, then we can find out near the discontinuous point, the convexity condition will be violated. □

Theorem 9.3.2 (convexity is equivalent to convexity along all lines). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the function $g(\alpha) = f(x + \alpha p), \alpha \in \mathbb{R}$ is convex for all $x, y \in \mathbb{R}$.*

Proof. We can easily show that

$$g(\lambda\alpha) + g((1-\lambda)\beta) \geq g(\lambda\alpha + (1-\lambda)\beta).$$

□

9.3.2 Connection to convex set

Definition 9.3.3 (epigraph). *Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, the* ***epigraph*** *[Figure 9.3.2] of f is defined as the set of points in $\mathbb{R}^n \times \mathbb{R}$:*

$$\{(x,y)|x \in \mathbb{R}^n, y \in \mathbb{R}, y \geq f(x)\}.$$

Lemma 9.3.1 (convex function and epigraph definition). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph is convex set.*

Proof. (forward) If we take $(x_1, t_1), (x_2, t_2) \in epi(f)$, then

$$\theta t_1 + (1-\theta)t_2 \geq \theta_1 f(x_1) + (1-\theta)f(x_2) \geq f(\theta x_1 + (1-\theta)x_2) \implies \theta(x_1,t_1) + (1-\theta)(x_2,t_2) \in epi(f).$$

(backward) Because $(x_1, f(x_1)), (x_2, f(x_2)) \in epi(f)$, by convexity of epigraph $(\theta(x_1, f(x_1)) + (1-\theta)(x_2, f(x_2)) \in epi(f)$), we have

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta_1 f(x_1) + (1-\theta)f(x_2)$$

□

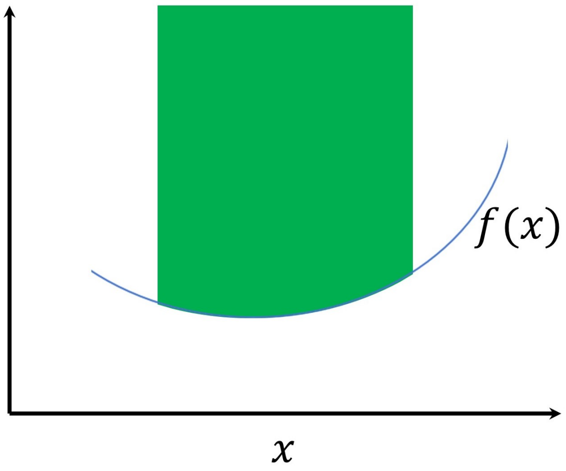


Figure 9.3.2: The epigraph (green area) of a convex function.

9.3.3 Strongly convex functions

Definition 9.3.4 (strongly convex function). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathcal{X} \to \bar{\mathbb{R}}$ be a proper convex function. f is* ***strongly convex*** *if there exists $\alpha > 0$ such that*

$$f(x) - \alpha\|x\|^2$$

is a convex function.

Remark 9.3.2 (interpretation). A strongly convex function is not as flat as a regular convex function. The shape of a strongly convex function is more bowl-like.

Lemma 9.3.2 (strongly convex is strictly convex). *If f is strongly convex, then f is strictly convex and also convex.*

Proof. Based on the definition of strong convexity, we have

$$f(\lambda x + (1-\lambda)y) - \alpha\|\lambda x + (1-\lambda)y\|^2 \leq \lambda f(x) + (1-\lambda)f(y) - \lambda\alpha\|x\|^2 - (1-\lambda)\alpha\|y\|^2$$

Rearrange and we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda\alpha\|x\|^2 - (1-\lambda)\alpha\|y\|^2 + \alpha\|\lambda x + (1-\lambda)y\|^2.$$

Since $\|x\|^2$ is strictly convex, we have

$$\lambda\alpha\|x\|^2 + (1-\lambda)\alpha\|y\|^2 - \alpha\|\lambda x + (1-\lambda)y\|^2 > 0, \forall x, y, x \neq y, \forall \lambda \in (0,1).$$

Therefore,

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

□

9.3.4 Operations preserve convexity

Lemma 9.3.3. *[5, p. 12] The following operations preserve convexity:*

- *(addition) If f_1 and f_2 are convex functions, then $f_1 + f_2$ is convex.*
- *(maximization) If $f_1, f_2, ..., f_k$ are convex functions, then $\max(f_1, f_2, ..., f_k)$ is convex function.*
- *(composition) If $g : \mathbb{R} \to \mathbb{R}$ is non-decreasing and convex, $h : \mathbb{R}^n \to \mathbb{R}$ is convex, then $g \circ h$ is convex.*

9.3.5 Convexity and derivatives

Theorem 9.3.3 (first derivative and linear under-estimator). *Let $\mathcal{D}$ be an open set and convex set in $\mathbb{R}^n$, and let $f : \mathcal{D} \to \mathbb{R}$ be differentiable on $\mathcal{D}$. Then,*

- *f is convex on $\mathcal{D}$ if and only if*

$$y - x \geq \nabla f(y-x), \forall x, y \in \mathcal{D}$$

- *f is strictly convex on $\mathcal{D}$ if and only if above inequality is strict whenever $x \neq z$*

Proof. (1, forward)Since f is convex, we have

$$f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda) f(x), \forall \lambda \in [0,1], x, y \in \mathrm{dom}(f),$$

By rearranging, we have

$$\begin{aligned} f(x + \lambda(y-x)) &\leq f(x) + \lambda(f(y) - f(x)) \\ \Rightarrow f(y) - f(x) &\geq \frac{f(x + \lambda(y-x)) - f(x)}{\lambda}, \forall \lambda \in (0,1] \end{aligned}$$

As $\lambda \to 0$, we use the definition of gradient and get

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

(1, backward) Let $z = \lambda x + (1-\lambda)y$. We have

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

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

If we multiply the first by λ and the second by $(1-\lambda)$ and add together, we have

$$\begin{aligned} \lambda f(x) + (1-\lambda) f(y) &\geq f(z) + \nabla f^T(z)(\lambda x + (1-\lambda)y - z) \\ &= f(z) = f(\lambda x + (1-\lambda)y) \end{aligned}$$

(2) Similar to (1). □

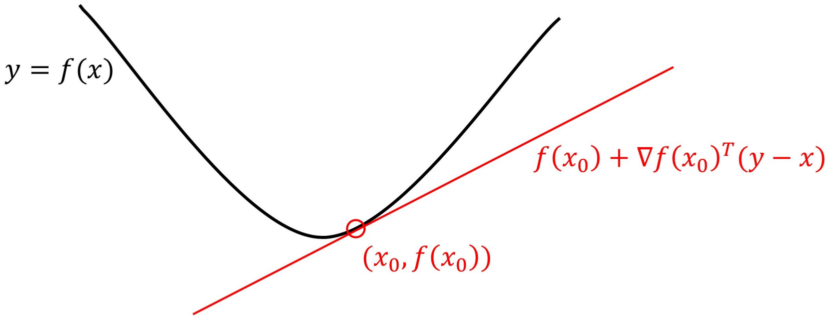


Figure 9.3.3: Illustration of linear underestimator.

Theorem 9.3.4 (convexity and second derivative). *Let $\mathcal{D}$ be an open set and convex set in $\mathbb{R}^n$, and let $f : \mathcal{D} \to \mathbb{R}$ be twice differentiable on $\mathcal{D}$. Then, f is convex on $\mathcal{D}$ if and only if*

$$\nabla^2 f \geq 0, \forall x \in \mathcal{D}.$$

Proof. (forward) We first prove the case that dimensionality is 1. Let $y > x$, then

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

We then have

$$f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x).$$

Diving both sides by $(y - x)^2$ gives

$$\frac{f'(y) - f'(x)}{y - x} \geq 0, \forall x, y, x \neq y.$$

As we let $y \to x$, we get $f''(x) \geq 0$. Now we consider general dimensionality. From the zero-order condition, convexity is equivalent to convexity along all lines in the domain [Theorem 9.3.2]. Let $v \in \mathcal{D}$, then $v^T \nabla(x) v \geq 0$ indicate $\nabla^2 f \geq 0$.

(backward) Suppose $f''(x) \geq 0$. By mean value theorem, we have

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x),$$

for some $z \in [x, y]$. Then we have

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

□

Corollary 9.3.4.1 (strong convexity and second derivative). *If f is a strongly convex function, then there exists a $\sigma > 0$ such that*

$$\nabla^2 f \geq \sigma I.$$

In other words, the smallest eigenvalue of the Hessian of f is uniformly lower bounded by σ everywhere.

Lemma 9.3.4 (relation to convex function). *Suppose a differentiable function $f : X \to \mathbb{R}$ is strongly convex such that there exists $\sigma > 0$ and $f(x) - \sigma\|x\|^2$ is convex.*

Then

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\|x-y\|^2, \forall x, y \in X$$

Proof. Because $f(x) - \sigma\|x\|^2$ is convex, we have

$$f(y) - \sigma\|y\|^2 \geq f(x) - \sigma\|x\|^2 + (\nabla f(x) - 2\sigma x)^T(y-x),$$

Rearrange and we have

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\|x-y\|^2.$$

□

Note 9.3.2 (convex and strongly convex). If $f \in C^2$ and the domain is $\mathbb{R}$, then

- f is convex if and only if $f''(x) \geq 0, \forall x$.
- f is strongly convex if and only if $f''(x) \geq m > 0, \forall x$.

Note that for strongly convex function, $f''(x)$ is uniformly bounded away from 0. closely.

Example 9.3.2.

- $f(x) = x^4$ has $f''(x) = 12x^2 \geq 0$, so f is a convex function and strictly convex function. It is not strongly convex because $\{f''(x_n)\} \to 0$ for the sequence $\{1/n\}$.

9.3.6 Subgradient

Definition 9.3.5 (subgradient, subdifferential). *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper convex function and $x \in dom(f)$. Any $g \in \mathbb{R}^n$ satisfying*

$$f(y) \geq f(x) + g^T(y - x), \forall y \in \mathbb{R}^n$$

is called a ***subgradient*** *of f at x.*

The set of all subgradients of f at x is called ***subdifferential*** *of f at x, denoted as $\partial f(x)$.*

Remark 9.3.3.

- If $g \in \partial f(x)$, then the epigraph of f lies above the linear underestimator

$$l(y) \triangleq f(x) + g^T(y - x).$$

- By convention, if $x \notin dom(f)$, then $\partial f(x) = \emptyset$.

Example 9.3.3. Consider $f(z) = |z|$. For $x \neq 0$, $\partial f(x) = \{1\}$; For $x = 0$, $\partial f(0) = [-1, 1]$.

Lemma 9.3.5 (basic properties of subgradient).

- *$g \in \mathbb{R}^n$ is a subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x_0 \in \mathbb{R}^n$ if and only if a hyperplane with normal vector $(g, -1) \in \mathbb{R}^{n+1}$ supports $epi(f)$ at $(x_0, f(x_0))$.*
- *If f is convex and differentiable, then $\nabla f(x)$ is a subgradient of f at x.*

Proof. (1) The hyperplane supporting at $(x_0, f(x_0))$ is given by the set $\{(x,y)|(g,-1)(x,y)^T = (g,-1)(x_0, f(x_0))\}$. (forward) Let (x,y) be a point in the $epi(f) = \{(x,y)|x \in \mathbb{R}^n, y \in \mathbb{R}, y \geq f(x)\}$, then we have $(g,-1)(x,y)^T \leq (g,-1)(x_0, f(x_0))$; this, the every point in $epi(f)$ is belonging to the half-space of $\{(x,y)|(g,-1)(x,y)^T \leq (g,-1)(x_0, f(x_0))\}$. Therefore, we have showed that g is subgradient by definition. since $(x, f(x))$ is in the epigraph. And g is a subgradient by definition. (backward) If g is subgradient, then

$$f(x) \geq f(x_0) + g^T(x - x_0)$$

and

$$y \geq f(x) \geq f(x_0) + g^T(x - x_0)$$

where (x,y) are in epigraph. Then, $epi(f)$ is supported by the hyperplane. (2) From Theorem 9.3.3. □

9.4 Duality theory

Definition 9.4.1 (constrained convex optimization problem). *Let $f : \mathcal{X} \to \mathbb{R}$ and $c_i : \mathcal{X} \to \mathbb{R}, i = 1, ..., m$ be convex functions, and let $\mathcal{X} \subseteq \mathbb{R}^n$ be a nonempty convex set. We consider the following convex optimization problem.*

$$\min_{x \in \mathbb{R}^n} f(x), \text{ subject to } x \in \mathcal{X} = \{x : c(x) \leq 0\}$$

where

$$c(x) = \begin{bmatrix} c_1(x) \\ c_2(x) \\ \cdots \\ c_m(x) \end{bmatrix}.$$

Definition 9.4.2 (primal function and dual function). *For convex optimization problem [Definition 9.4.1], define the Lagrangian $\mathcal{L} : \mathcal{X} \times \mathbb{R}^m \to \mathbb{R}$ as*

$$\mathcal{L}(x, y) = f(x) + c(x)^T y = f(x) + \sum_{i=1}^{m} c_i(x) y_i$$

where $y \in \mathbb{R}^m, y \geq 0$ is called dual variables and $x \in \mathcal{X}$ is called primal variables. The ***primal*** *function $p(x)$ is defined as*

$$p(x) = \begin{cases} \sup_{y \in \mathcal{Y}} \mathcal{L}(x, y), \text{ if } x \in \mathcal{X} \\ \infty, \text{ if } x \notin \mathcal{X} \end{cases},$$

and the ***dual function*** *is defined as*

$$d(y) = \begin{cases} \inf_{x \in \mathcal{C}} \mathcal{L}(x, y), \text{ if } y \in \mathcal{Y} \\ -\infty, \text{ if } y \notin \mathcal{Y} \end{cases},$$

where $\mathcal{Y} = \{y : y_i \geq 0, \forall i = 1, 2, ..., m\}$.

Definition 9.4.3 (primal problem and dual problem). *The **primal problem** is defined as*

$$\min_{x\in\mathbb{R}^n} p(x) = \min_{x\in\mathcal{X}}\{\sup_{y\in\mathcal{Y}} \mathcal{L}(x,y)\}.$$

*The **dual problem** is defined as*

$$\max_{y\in\mathbb{R}^m} d(y) = \max_{y\in\mathcal{Y}}\{\inf_{x\in\mathcal{X}} \mathcal{L}(x,y)\}.$$

Lemma 9.4.1 (primal function "is" original function). *The primal function is equivalent to*

$$p(x) = \begin{cases} f(x), \text{if } x \in \mathcal{X}, c(x) \leq 0 \\ \infty, \text{otherwise} \end{cases}$$

Proof. When $c(x) \leq 0$, y will take 0 to maximize $\mathcal{L}(x,y)$ in y; that is $p(x) = \sup_{y\in\mathcal{Y}} \mathcal{L}(x,y) = \mathcal{L}(x,0) = f(x)$. When there exist one $c_i(x) > 0$, $p(x) = \sup_{y\in\mathcal{Y}} = \infty$. □

Remark 9.4.1 (dual function is explicit only for simple cases). Note that dual function has explicit form only for simple optimization problems, such as linear optimization.

Lemma 9.4.2 (The dual function is a concave function). *The dual function defined as*

$$d(y) = \begin{cases} \inf_{x\in\mathcal{C}} \mathcal{L}(x,y), \text{ if } y \in \mathcal{Y} \\ -\infty, \text{ if } y \notin \mathcal{Y} \end{cases},$$

where $\mathcal{Y} = \{y : y_i \geq 0, \forall i = 1,2,...,m\}$, is a concave function.

Proof.

$$\begin{aligned} d(\lambda y_1 + (1-\lambda)y_2) &= \inf_x \mathcal{L}(x, \lambda y + (1-\lambda)y) \\ &= \inf_x \mathcal{L}(x, \lambda y) + \mathcal{L}(x, (1-\lambda)y) \\ &\geq \inf_x \mathcal{L}(x, \lambda y) + \inf_x \mathcal{L}(x, (1-\lambda)y) \\ &\geq d(\lambda y_1) + d((1-\lambda)y_2)c. \end{aligned}$$

□

Example 9.4.1 (linear programming). [6, lec 2] Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$ and consider the linear program

$$\min_{x\in\mathbb{R}^n} f(x) = g^T x, \ subject\ to\ Ax - b \leq 0$$

and assume that $Ax - b \leq 0$ is feasible. The Lagrangian and dual functions are given by

$$\mathcal{L}(x,y) = g^T x + (Ax-b)^T y$$

and

$$d(y) = \begin{cases} \inf_{x\in\mathbb{R}^n} g^T x + (Ax-b)^T y = \inf_{x\in\mathbb{R}^n} x^T(g + A^T y) - b^T y, if\ y \geq 0 \\ -\infty, otherwise \end{cases}$$

Note that for linear optimization $\inf_{x\in\mathbb{R}^n} x^T(g+A^Ty) - b^Ty$, if $g + A^Ty \neq 0$, then $\inf_{x\in\mathbb{R}^n} x^T(g+A^Ty) - b^Ty = -\infty$. Therefore, we can further simplify the dual function to

$$d(y) = \begin{cases} -b^T y, if\ y \geq 0, g + A^T y = 0 \\ -\infty, otherwise \end{cases}$$

Maximize the dual function can be written as

$$\max_{y\in\mathbb{R}^m} -b^T y, subject\ to\ g + A^T y = 0, y \geq 0.$$

Theorem 9.4.1 (weak duality I). *For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have*

$$d(y) \leq p(x).$$

Moreover, if x is a primal feasible point, then

$$d(y) \leq f(x).$$

Let $f_{opt} = \inf_{x\in\mathcal{X}} f(x), subject\ to\ x \in \mathcal{X}, c(x) \leq 0$. If it is feasible, then

$$\sup_{d\in\mathbb{R}^m} d(y) \leq \inf_{x\in\mathcal{X}} f(x) < \infty.$$

Proof. (1)

$$d(y) = \inf_x L(x,y) \leq L(x,y) \leq \sup_y L(x,y) = p(x)$$

(2) When x is primal feasible, then $p(x) = f(x)$. □

Definition 9.4.4 (strong duality). *If the equality $d(y^*) = p(x^*)$ holds, we say the* ***strong duality*** *holds. However, strong duality does not always hold. For certain type of constraints, only some simple conditions need to be satisfied to ensure strong duality.*

Consider convex optimization of the form minimize $\quad f_0(x)$ subject to $g_i(x) \leq 0, \quad i = 1, \ldots, m \; Ax = b$ where $f, g_1, ...g_m$ are convex functions.

Then if there exists an $x \in relintD$ such that non-affine inequalities are strictly satisfied, i.e., $g_i(x) < 0, i = 1, ..., m, Ax = b$, which is the so-called **Slater's condition**, then strong duality holds.

9.5 Convex optimization and optimality conditions

9.5.1 Local optimality vs. global optimality

Lemma 9.5.1 (local minimum implies global minimum). *If X is a convex subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is convex over X, then a local minimum of f over X is also a global minimum.*

Proof. let x^* be a local minimum, suppose there is a point $x' \neq x^*$ being a global minimum. Then

$$f(x^* + a(x' - x^*)) \leq af(x^*) + (1-a)f(x') < af(x^*) + (1-a)f(x^*) < f(x^*)$$

for $a \in [0,1]$ which contradicts that x^* is local minimum. □

Note: The domain of convex function has to be convex, which is the definition of convex functions on a general set. Otherwise, for $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$, the element $(\lambda x + (1-\lambda)y)$

Lemma 9.5.2 (uniqueness of global minimum under strict convexity). *[4, p. 17] If X is a convex subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** over X, then f over X has at most one global minimum. In other words, **the set of minimizers is either empty or a singleton.***

Proof. Suppose there are two global minimum at x_1, x_2, then the middle $(x_1 + x_2)/2$ will have a smaller value based on the definition of **strict convexity**. □

Remark 9.5.1 (global minimizer might not exist). Consider $f(x) = e^{-x}$. It is strictly convex since $f''(x) = e^{-x}$. The global minimizer is $-\infty \notin \mathbb{R}$. However, for strongly convex function, the global minimizer will exist on closed convex set.

Lemma 9.5.3 (existence and uniqueness of global minimizer under strongly convexity). *[4, p. 17] If X is a **convex and closed** subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is **strongly convex** over X, then f over X has one and only one global minimum.*

Proof. Note that from the properties of strongly convex function [subsection 9.3.3], it can be showed that it is a coercive function. A coercive function on a closed set has a global minimizer [Theorem 3.1.7]. Moreover, strongly convex function is also strictly convex function, therefore, the minimizer is unique. □

9.5.2 Unconstrained optimization optimality conditions

Theorem 9.5.1 (necessary and sufficient optimality condition, differentiable function). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex* ***differentiable*** *function and consider the problem*

$$\min_{x\in\mathbb{R}^n} f(x).$$

Then x^* is a global minimizer if and only if

$$0 = \nabla f(x^*).$$

Proof. Let $g \triangleq \partial f(x^*)$, then from the property of first derivative and convexity [Theorem 9.3.3]

$$f(y) \geq f(x^*) + g^T(y - x^*), \forall y \in \mathcal{X}.$$

Take $g = 0$, then

$$f(y) \geq f(x^*), \forall y \in \mathcal{X}.$$

□

Theorem 9.5.2 (necessary and sufficient optimality condition via subdifferential, nonsmooth function). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and consider the problem*

$$\min_{x\in\mathbb{R}^n} f(x).$$

Then x^* is a global minimizer if and only if

$$0 \in \partial f(x^*).$$

Proof. Let $g \in \partial f(x^*)$, then based on the definition of subgradient we have

$$f(y) \geq f(x^*) + g^T(y - x^*), \forall y \in \mathcal{X}.$$

Take $g = 0$, then

$$f(y) \geq f(x^*), \forall y \in \mathcal{X}.$$

□

9.5.3 Constrained optimization optimality conditions

Definition 9.5.1 (general constrained convex optimization). *The convex constrained optimization problem is given as*

$$\min_{x \in \mathbb{R}^n} f(x), x \in \mathcal{X},$$

for some nonempty set $\mathcal{X} \subseteq \mathbb{R}^n$ and function $f : \mathbb{R}^n \to (-\infty, \infty]$

Theorem 9.5.3 (necessary and sufficient condition). *[4, p. 17]Let X be a convex set and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function over X.*

1. *If f is continuously differentiable, then*

   $$\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in X$$

   is a necessary and sufficient condition for x^* to be a global minimum of f over X.
2. *If X is open and f is continuously differentiable, then*

   $$\nabla f(x^*) = 0, \forall x \in X$$

   is a necessary and sufficient condition for x^* to be a global minimum of f over X

Proof. (1) (a)From $f(x) - f(x^*) \geq f(x^*)(x - x^*)$ and $f(x^*)(x - x^*) \geq 0$, we have

$$f(x) \geq f(x^*), \forall x \in X$$

; (b) If $f(x^*)(x - x^*) < 0$, then

$$(f(x + a(x^* - x)) - f(x)) = \nabla a f(x^*)^T(x - x^*) + o(a(x - x^*)) < 0$$

, as $a > 0, a \to$, therefore, we can decrease the function, which is a contraction. (2)(a) If $\nabla f(x^*) = 0$, from (1) we can prove the forward direction (b) suppose $\nabla f(x^*) > 0$, then we can decrease it by move in $-\nabla f(x^*)$ in sufficiently small steps. □

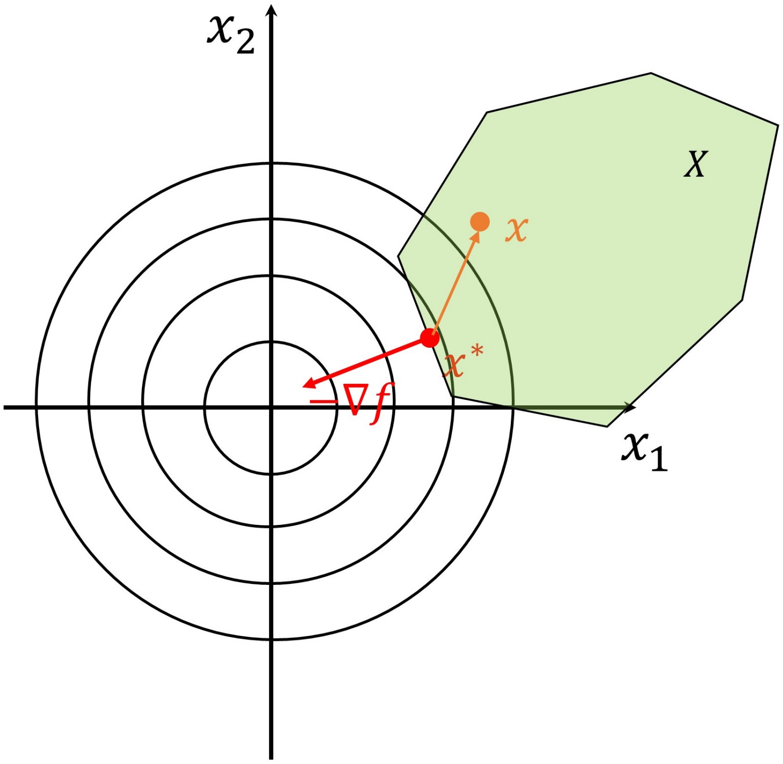


Figure 9.5.1: Demonstration of optimality condition $\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in X$ when x^* lies on the boundary of X.

Remark 9.5.2.

- If X is not a open set, then we cannot guarantee there is a neighborhood around x^*, and therefore we have to use the first item, because $\nabla f(x^*)$ might not exist if x^* is at the boundary of X.
- No matter X is open or not, as long as X is convex, we can always use the first item.
- For unconstrained optimization on $\mathbb{R}^m$, we usually use the the second item because $\mathbb{R}^m$ is considered open.

Remark 9.5.3 (Why we do not need second order condition)**.**

- We do not need second order necessary condition because if f is twice differentiable, then $\nabla f \geq 0$ as implies by convexity of f.
- We do not need second order sufficient condition like because $\nabla f > 0$ if f is twice differentiable because we rely on the condition $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ in the proof, which is a much stronger condition.

- The second order sufficient condition $\nabla f > 0$ in general nonlinear optimization is a 'strong' condition that directly guarantees the existence of 'strict' local minimum, which does not direct counterpart in convex optimization. In convex optimization, if we want 'strict' local minimum, we certainly need 'extra information/conditions'.

Theorem 9.5.4 (KKT condition, inequality constraint). *Consider an convex optimization problem given by*

$$\begin{aligned} &\min \quad f(x) \\ &\textit{subject to } h_i(x) \leq 0, i = 1, \dots m \\ &\ell_j(x) = 0, j = 1, \dots r \end{aligned}$$

where $f(x), h_1, ..., h_m$ are convex function, $l_1, ..., l_r$ are linear equality constraints. The ***Karush-Kuhn-Tucker conditions (KKT conditions)*** *are:*

- *stationarity:* $0 \in \partial f(x) + \sum_{i=1}^m u_i \partial h_i(x) + \sum_{j=1}^r v_j \partial \ell_j(x)$
- *complementary:* $u_i h_i = 0, \forall i$
- *primal feasibility:* $h_i(x) \leq 0, l_j(x) = 0, \forall i, j$
- *dual feasibility:* $u_i \geq 0, \forall i$

Assume strong duality holds, then there exists x^*, u^*, v^* satisfying the KKT condition if and only if x^*, u^*, v^* are primal and dual solutions.

Proof. (forward)

Note that

$$\begin{aligned} f\left(x^*\right) &= g\left(u^*, v^*\right) \\ &= \min_x f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* \ell_j(x) \\ &\leq f\left(x^*\right) + \sum_{i=1}^m u_i^* h_i\left(x^*\right) + \sum_{j=1}^r v_j^* \ell_j\left(x^*\right) \\ &\leq f\left(x^*\right) \end{aligned}$$

Therefore above equality holds. We must have

- x^* minimizes $L\left(x, u^*, v^*\right)$ over x, i.e.,

$$\begin{aligned} &0 \in \partial_x L\left(x^*, u^*, v^*\right) \\ &0 \in \partial f\left(x^*\right) + \sum u_i^* \partial h_i\left(x^*\right) + \sum v_j^* \partial \ell_j\left(x^*\right) \end{aligned}$$

which is the stationary condition.

- $\sum u_i^* h_i(x^*) = 0$, i.e.
  $$u_i^* h_i(x) = 0 \text{ for all } i$$
  which is the complementary slackness condition.
- The primal and dual feasibility of (x^*, u^*, v^*) hold.

(backward) If there exists x^*, u^*, v^* that satisfy the KKT conditions, then

$$\begin{aligned} g(u^*, v^*) &= f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{r} v_j^* \ell_j(x^*) \\ &= f(x^*) \end{aligned}$$

where the first equality holds from stationarity, and the second holds from complementary slackness. Therefore, the duality gap is zero, so x^* and u^*, v^* are primal and dual optimal respectively. □

Remark 9.5.4 (compare with KKT conditions in nonlinear constrained optimization)**.**

- For convex optimization with inequality constraints, we only need first order condition, whereas general nonlinear optimization will require second order condition.

Example 9.5.1. Consider the support vector machine optimization problem given by

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p, \beta_0\in\mathbb{R}, \xi\in\mathbb{R}} \quad & \tfrac{1}{2}\|\beta\|_2^2 + C\textstyle\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, i = 1, \cdots, n \\ & y_i\left(x_i^T\beta + \beta_0\right) \geq 1 - \xi_i, i = 1, \cdots, n \end{aligned}$$

where $x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$. Denote the dual variables by $v, w \geq 0$. The KKT condition gives

$$0 = \sum_{i=1}^{n} w_i y_i, \quad \beta = \sum_{i=1}^{n} w_i y_i x_i, \quad w = C\mathbf{1} - v$$

and complementary slackness condition:

$$v_i\xi_i = 0, w_i\left(1 - \xi_i - y_i\left(x_i^T\beta + \beta_0\right)\right) = 0, \quad i = 1, \cdots, n.$$

9.6 Subgradient methods

9.6.1 A generic algorithm for unconstrained problem

We first consider minimizing an unconstrained function $f(x)$, which is convex and defined on $\mathbb{R}^n$. Subgradient methods employ a similar idea in gradient descent, with gradients replaced by subgradient. Starting from an iterate $x^{(0)}$, we update the iterate via

$$x_k = x_{k-1} - \alpha_k \cdot g_{k-1}, k = 1, 2, 3, \cdots$$

Compared to general gradient methods, the step size α_k can be chosen in a much simpler way and still have good convergence property. Most common two ways are:

- Fixed step size: $\alpha_k = s$ for all $k = 1, 2, 3 \cdots$
- Diminishing step size: choose α_k to satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty \,, \; \sum_{k=1}^{\infty} \alpha_k = \infty$$

A generic subgradient algorithm is in the following algorithm 16.

Algorithm 16: A generic subgradient algorithm

```
Input: Initial guess $x_0 \in \mathbb{R}^n$
Set $k = 0$.
repeat
    Obtain a subgradient $g_k \in \partial f(x_k)$
    if $g_k = 0$ then
        return a global minimizer $x_k$.
    end
    Choose $\alpha_k > 0$
    Set $x_{k+1} = x_k - \alpha_k g_k$.
    Record the best iterate so far $x_k^{best}$ that has the minimum $f(x)$.
    set $k = k + 1$.
until termination condition;
```

Theorem 9.6.1. *Consider the subgradient algorithm [algorithm 16]. Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies:*

- *f has its subgradient bounded by $\|g\|_2 < G$.*
- *$||x_0 - x^*|| \leq R$ which means it is bounded.*

Then we have

- *For a fixed step size $\alpha_k = s$, we have*

$$\lim_{k\to\infty} f(x_k^{best}) \leq f(x^*) + \frac{G^2 s}{2}$$

- *For a step size α_k satisfying $\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \sum_{k=1}^{\infty} \alpha_k < \infty$, we have*

$$\lim_{k\to\infty} f(x_k^{best}) = f(x^*)$$

Proof.

$$\begin{aligned} ||x_{k+1} - x^*||^2 =& ||x_k - t_k g_k - x^*||^2 \\ =& ||x_k - x^*||^2 - 2t_k(g_k)^T(x_k - x^*) + \alpha_k^2 ||g_k||^2 \end{aligned}$$

By definition of the subgradient method, we have

$$f(x^*) \geq f(x_k) + g_k(x^* - x_k)$$

Using this inequality, we have

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - 2t_k(f(x_k) - f(x^*)) + \alpha_k ||g_k||^2.$$

Sum up k terms, we can arrive at

$$||x_{k+1} - x^*||^2 \leq ||x_1 - x^*||^2 - 2\sum_{i=1}^{k} t_i(f(x_i) - f(x^*)) + \sum_{i=1}^{k} \alpha_i^2 ||g_i||^2.$$

Then we have

$$\begin{aligned} 0 \leq ||x_{k+1} - x^*||^2 \leq R^2 - 2\sum_{i=1}^{k} \alpha_i(f(x_i) - f(x^*)) + \sum_{i=1}^{k} \alpha_i^2 G^2 \\ 2\sum_{i=1}^{k} \alpha_i(f(x^{(i)}) - f(x^*)) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 G^2 \\ 2(\sum_{i=1}^{k} \alpha_i)(f(x_k^{best}) - f(x^*)) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 G^2 \end{aligned}$$

For a constant step size $\alpha_i = s$:

$$\frac{R^2 + G^2 s^2 k}{2sk} \to \frac{G^2 s}{2}, \text{ as } k \to \infty,$$

and for diminishing step size, we have:

$$\sum_{i=0}^{k} \alpha_i^2 \leq 0, \sum_{i=0}^{k} \alpha_i = \infty$$

therefore,

$$\frac{R^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i} \to 0, \text{ as } k \to \infty,$$

□

Remark 9.6.1 (estimation of convergence speed). If we take $\alpha_i = R/(G\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following bound:

$$\frac{R^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i} = \frac{R^2 + G^2}{\sqrt{k}}.$$

That is, subgradient method has convergence rate of $O(1/\sqrt{k})$, and to get $f(x_{best}^{(k)}) - f(x^*) \leq \epsilon$, needs $O(1/\epsilon^2)$ iterations.

9.6.2 Convergence under Lipschitz smoothness

Definition 9.6.1 (Lipschitz continuous function). *A function f is $L-$smoothness if there exists a constant $L \geq 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \forall x, y \in \mathcal{X}.$$

Example 9.6.1. Consider a quadratic function

$$f(x) = \frac{1}{2}x^T Bx + c^T x, x \in \mathbb{R}^n.$$

We have [Theorem 4.13.1]

$$\nabla f(x) = Bx + c, \|\nabla f(x) - \nabla f(y)\|_2 = \|B(x - y)\| \leq \|B\|_2 \|x - y\|.$$

Therefore, if the max magnitude of eigenvalues, i.e., $\max_{1 \leq i \leq n} |\lambda_i|$ equals L, then f has a Lipschitz continuous gradient constant L.

Lemma 9.6.1 (properties of Lipschitz continuous function). *Suppose f is convex and L -smooth. We have*

- $$\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{L}{2}\|x-y\|_2^2, \forall x, y \in \mathcal{X}$$
- $$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2, \forall x, y \in \mathcal{X}$$
- $$[\nabla f(x) - \nabla f(y)]^{\dot{T}}(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \forall x, y \in \mathcal{X}$$

Proof. (1)

$$\begin{aligned}
& f(y) - f(x) - \nabla f(x)^T(y-x) \\
&= \int_0^1 \nabla f(x + t(y-x))^T(y-x)dt - \nabla f(x)^T(y-x) \\
&= \int_0^1 [\nabla f(x + t(y-x)) - \nabla f(x)]^T(y-x)dt \\
&\leq \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\|_2 \|y-x\|_2 dt \text{ (by Cauchy-Schwarz inequality)} \\
&\leq \int_0^1 L \left[t(y-x)\right\|_2 \| y - x\|_2 dt \text{ (by L-smoothness of } f\text{)} \\
&= L\|(y-x)\|_2^2 \int_0^1 t dt \\
&= \frac{L}{2}\|(y-x)\|_2^2
\end{aligned}$$

(2) Let $z = y + \frac{1}{L}(\nabla f(x) - \nabla f(y))$ We have

$$\begin{aligned}
f(y) - f(x) &= f(y) - f(z) + f(z) - f(x) \\
&\geq -\nabla f(y)^T(z-y) - \frac{L}{2}\|y-z\|_2^2 + \nabla f(x)^T(z-x) \\
&= \nabla f(x)^T(y-x) - \{\nabla f(x) - \nabla f(y)\}^T(y-z) - \frac{L}{2}\|y-z\|_2^2 \\
&= \nabla f(x)^T(y-x) + \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2 \\
&= \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2
\end{aligned}$$

(3) Add the following two together.

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \tfrac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$
$$f(x) \geq f(y) + \nabla f(y)^T(x-y) + \tfrac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

□

Lemma 9.6.2 (reduction per subgradient step). *Let f be convex and L -smooth. The subgradient descent step strictly reduces the objective function value at each iteration,*

$$f\left(x_{k+1}\right) - f\left(x_k\right) \leq -\alpha\left(1 - \frac{L}{2}\alpha\right)\left\|\nabla f\left(x_k\right)\right\|_2^2$$

If we take $\alpha = \frac{1}{L}$, we get the maximum reduction, which is given by

$$f\left(x_{k+1}\right) - f\left(x_k\right) \leq -\frac{1}{2L}\left\|\nabla f\left(x_k\right)\right\|_2^2.$$

Proof.

$$\begin{aligned} f\left(x_{t+1}\right) - f\left(x_t\right) &\leq \nabla f\left(x_t\right)^T\left(-\alpha\nabla f\left(x_t\right)\right) + \frac{L}{2}\left\|-\alpha\nabla f\left(x_t\right)\right\|_2^2 \\ &= -\alpha\left(1 - \frac{L}{2}\alpha\right)\left\|\nabla f\left(x_t\right)\right\|_2^2 \end{aligned}$$

□

Theorem 9.6.2 (convergence of subgradient methods on Lipschitz continuous function). *[7] Let f be convex and L -smooth, x^* be an optimal solution. With $\alpha = \frac{1}{L}$ for all $\alpha > 0$, the iterates from subgradient descent satisfies*

$$f\left(x_k\right) - \min_{x\in\mathbb{R}^n} f(x) \leq \frac{2L\left\|x_0 - x^*\right\|_2^2}{k}$$

Proof. Note that we have

- $f\left(x_{k+1}\right) - f\left(x_k\right) \leq -\frac{1}{2L}\left\|\nabla f\left(x_k\right)\right\|_2^2$.

- $\|x_{k+1} - x_*\|_2 \leq \|x_k - x_*\|_2$ This is because

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \left\|x_k - \frac{1}{L}\nabla f\left(x_t\right) - x^*\right\|_2^2 \\ &= \|x_k - x^*\|_2^2 - \frac{2}{L}\nabla f\left(x^*\right)^T\left(x_t - x^*\right) + \frac{1}{L^2}\left\|\nabla f\left(x_t\right)\right\|_2^2 \\ &\leq \|x_k - x^*\|_2^2 - \frac{1}{L^2}\left\|\nabla f\left(x_k\right)\right\|_2^2 \\ &\leq \|x_k - x^*\|_2^2\end{aligned}$$

- $\left\|\nabla f\left(x_i\right)\right\|_2 \geq \frac{f(x_i) - f(x_i)}{\|x_i - x_i\|}$
  because

$$f\left(x_k\right) - f\left(x^*\right) \leq \nabla f\left(x_k\right)^T\left(x_k - x_k\right) \leq \left\|\nabla f\left(x_k\right)\right\|_2 \|x_k - x^*\|_2$$

By combining them, we arrive at

$$f\left(x_{k+1}\right) - f\left(x_k\right) \leq -\frac{1}{2L}\left[\frac{f\left(x_k\right) - f\left(x_*\right)}{\|x_0 - x_*\|_2}\right]^2$$

and set $\epsilon_k = f\left(x_k\right) - f\left(x_k\right)$ and $\beta = \frac{1}{2L_1 x_0 - x_0 + 1/3}$

$$\begin{aligned}\left[f\left(x_{k+1}\right) - f\left(x^*\right)\right] - \left[f\left(x_k\right) - f\left(x^*\right)\right] = \epsilon_{k+1} - \epsilon_k \leq -\tfrac{1}{2L}\tfrac{\epsilon_k^2}{\|x_0 - x_*\|_2^2} = -\beta\epsilon_k^2 \\ \frac{1}{\epsilon_k} - \frac{1}{\epsilon_{k+1}} \leq -\beta\frac{\epsilon_k}{\epsilon_{k+1}} \leq -\beta \\ \Rightarrow \frac{1}{\epsilon_k} + \beta \leq \frac{1}{\epsilon_{k+1}} \\ \Rightarrow \frac{1}{\epsilon_0} + \beta k \leq \frac{1}{\epsilon_k} \\ \Rightarrow \beta k \leq \frac{1}{\epsilon_t} = \frac{1}{f\left(x_k\right) - f\left(x^*\right)}\end{aligned}$$

which implies $f\left(x_k\right) - f\left(x^*\right) \leq \frac{2L\|x_0 - x.\|_2^2}{k}$ □

9.6.3 Projected gradient methods

9.6.3.1 *Foundations*

Now we address convex optimization problems over convex sets. The target problem is

$$\min_{x \in \mathcal{X}} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is **continuously differentiable** and $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set.

Definition 9.6.2 (gradient projection arc). *[6] For any $x_k \in \mathcal{X}$, the* ***gradient projection arc*** *is defined as the set of vectors*

$$\{x_k(\alpha) : \alpha > 0\},$$

where $x_k(\alpha) \triangleq Proj_{\mathcal{X}}(x_k - \alpha\nabla f(x_k))$ and

$$Proj_{\mathcal{X}}(y) \triangleq \arg\min_{x\in\mathcal{X}} \|x - y\|_2 .$$

Lemma 9.6.3 (descent properties of gradient projection arc). *[8, p. 304][6] If $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function and $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed and convex set, then the following properties hold:*

- *If $x_k(\alpha) \neq x_k$ for any $\alpha > 0$, then $d_k(\alpha) \triangleq x_k(\alpha) - x_k$ is a feasible direction satisfying*

$$\nabla f(x_k)^T(x_k(\alpha) - x_k) \leq -\frac{1}{\alpha}\|x_k(\alpha) - x_k\|_2^2 < 0, \forall \alpha > 0.$$

  That is, $\nabla f(x_k)^T d_k(\alpha) < 0$, $d_k(\alpha)$ ***is descent direction for all*** *$\alpha > 0$* ***such that*** *$x_k(\alpha) \neq x_k$*
- *If $x_k(\alpha) = x_k$ for some $\alpha > 0$, then x_k is first-order minimizer for the target problem.*

Proof. (1) From the obtuse angle theorem [Theorem 9.2.2],we have

$$(x_k - \alpha\nabla f(x_k) - x_k(\alpha))^T(x - x_k(\alpha)) \leq 0, \forall x \in \mathcal{X}.$$

Choose $x = x_k$, we obtain

$$(x_k - -\alpha\nabla f(x_k) - x_k(\alpha))^T(x_k - x_k(\alpha)) \leq 0.$$

Rearrange, we have

$$\nabla f(x_k)^T(x_k(\alpha) - x_k) \leq -\frac{1}{\alpha}\|x_k(\alpha) - x_k\|_2^2 < 0, \forall \alpha > 0.$$

(2) From the obtuse angle theorem [Theorem 9.2.2],we have

$$(x_k - \alpha\nabla f(x_k) - x_k(\alpha))^T(x - x_k(\alpha)) \leq 0, \forall x \in \mathcal{X}.$$

If there exists $\alpha > 0$ such that $x_k(\alpha) = x_k$, then

$$\alpha\nabla f(x_k)^T(x - x_k(\alpha)) \geq 0, \forall \alpha > 0.$$

This is exactly the first-order optimality condition [Theorem 9.5.3]. □

Theorem 9.6.3 (convergence with constant step size). *[6] If*

- $f : \mathbb{R}^n \to \mathbb{R}$ *is a continuously differentiable function*
- ∇f *is Lipschitz continuous with constant* L
- $\mathcal{X} \subseteq \mathbb{R}^n$ *be a closed and convex set*
- $\alpha \in (0, 2/L)$.
- *The sequence* $\{x_k\}$ *is generated by*

$$x_{k+1} = Proj_{\mathcal{X}}(x_k - \alpha \nabla f(x_k)).$$

Then,

- ***monotonically decreasing along*** x_k*:*

$$f(x_{k+1}) \leq f(x_k) - (\frac{1}{\alpha} - \frac{L}{2})\|x_{k+1} - x_k\|_2^2.$$

- *Every limit point of* $\{x_k\}$ *is a* ***first order solution****.*

9.6.3.2 *Algorithms*

Algorithm 17: Gradient projection algorithm with constant step size

```
Input: Input any $x_{in} \in \mathbb{R}^n$
Set $x_0 = Proj_{\mathcal{X}}(x_{in})$ such that $x_0$ is feasible and set $k = 0$.
Choose $\epsilon > 0$, $\alpha_0 > 0$, and $\eta_d \in (0,1)$.
repeat
    compute next iterate via:
        $x_{k+1} = Proj_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k)).$
    while $f(x_{k+1}) > l(x_{k+1}; x_k) + \frac{1}{2\alpha_k}\|x_{k+1} - x_k\|^2$ do
        Set $\alpha_k = \eta_d \alpha_k$.
        Compute $x_{k+1} = Proj_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k))$.
    end
    Set $\alpha_{k+1} = \alpha_k$.
    set $k = k + 1$.
until termination condition;
```

In practice, if we do not know L, then how to choose $\alpha > 0$. We have select step size according to [6]:

$$f(x_k(\alpha)) \leq \ell(x_k(\alpha); x_k) + \frac{1}{2\alpha}\|x_k(\alpha) - x_k\|_2^2$$

Consequently, if we decrease α by some factor every time that holds, eventually we will have $\alpha \in (0, 1/L]$ and it will no longer need to be changed.

Algorithm 18: Gradient projection algorithm with adaptive size

```
Input: Input any $x_{in} \in \mathbb{R}^n$
Set $x_0 = Proj_{\mathcal{X}}(x_{in})$ such that $x_0$ is feasible and set $k = 0$.
Choose $\epsilon > 0$, $\alpha_0 > 0$, and $\eta_d \in (0, 1)$.
repeat
    compute next iterate via:
        $$x_{k+1} = Proj_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k)).$$
    while $f(x_{k+1}) > l(x_{k+1}; x_k) + \frac{1}{2\alpha_k}\|x_{k+1} - x_k\|^2$ do
        Set $\alpha_k = \eta_d \alpha_k$.
        Compute $x_{k+1} = Proj_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k))$.
    end
    Set $\alpha_{k+1} = \alpha_k$.
    set $k = k + 1$.
until termination condition;
```

Remark 9.6.2.

- Choose $\alpha_0 \approx 1/L$ if an estimate of the Lipschitz constant L is available.
- Reasonable stopping condition

  $$\|x_{k+1} - x_k\|_\infty \leq \epsilon \max(1, \|x_0\|_\infty)$$

  for some small $\epsilon \approx 10^{-6}$.
- We cannot use $\|\nabla f(x_k)\| \approx 0$ as the termination condition, because we are doing constrained optimization.

9.6.4 Proximal gradient methods

9.6.4.1 *Foundations*

We consider convex optimization whose objective function can be written in two parts, which is given by

$$\min_{x \in \mathbb{R}^n} f(x) + r(x),$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is **differentiable convex function**.

- ∇f is Lipschitz continuous with constant L.
- $r : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a closed proper convex function.

Different choices of $r(x)$ can be

- $r(x) = \delta_{\mathcal{X}}(x)$ will give the constrained optimization on f, and the proximal gradient projection iteration is equivalent to project gradient iteration.
- $r(x) = \lambda\|x\|_1$ gives the sparsity problem.See subsubsection 9.6.4.3.
- $r(x) = \lambda\|x\|_2$ gives the following iteration:

$$x_{k+1} = (\frac{1}{1+\alpha_k})z_k.$$

The method proximal gradient method will carry out iteration given by

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} f(x_k) + \nabla f(x_k)^T(x - x_k) + r(x) + \frac{1}{2\alpha_k}\|x - x_k\|_2^2.$$

The proximal gradient method iteration is equivalent to a two step iteration:

$$\begin{aligned} z_k &= x_k - \alpha\nabla f(x_k) \\ x_{k+1} &= r(x) + \frac{1}{2\alpha_k}\|x - z_k\|_2^2. \end{aligned}$$

9.6.4.2 *Algorithms*

Algorithm 19: Proximal gradient algorithm

```
Input: Input any $x_0 \in \mathbb{R}^n$. Objective function with Liptchize constant $L$
Set $k = 0$, and choose $\epsilon > 0, \alpha_0 < 1/L$, and $\eta_d \in (0,1)$
repeat
    Set $z_k = x_k - \alpha_k\nabla f(x_k)$.
    Compute next iterate via:
        $$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} r(x) + \frac{1}{2\alpha_k}\|x - x_k\|_2^2.$$
    set $k = k + 1$.
until termination condition;
```

Theorem 9.6.4 (convergence of proximal gradient method). *[6] Assume*

- *$f : \mathbb{R}^n \to \mathbb{R}$ is a convex differentiable function*
- *∇f is Lipschitz continuous with constant L*

- *$\mathcal{X} \subseteq \mathbb{R}^n$ be a closed and convex set*
- *$\mathcal{X}^*$ is nonempty.*

The gradient projection algorithm has the following properties:

1. *The inner loop searching for α_k will terminate in finite steps. And $\alpha_k = \bar{\alpha} > 0$ for all sufficiently large k.*
2. *$\{x_k\}$ converges to a **first order solution**.*
3. *The sequence $\{f(x_k)\}$ is monotonically decreasing. The decrease of objective function for all k given by*
$$f(x_{k+1}) \leq f(x_k) - (\frac{1}{\alpha} - \frac{1}{2\alpha_k})\|x_{k+1} - x_k\|_2^2.$$
4. $$f(x_{k+1}) - f_{opt} \leq \frac{\min_{x^* \in \mathcal{X}^*} \|x_0 - x^*\|_2^2}{2(k+1)\bar{\alpha}}$$

9.6.4.3 *Case study: sparsity regularization problem*

Definition 9.6.3 (target problem). *The target problem is*

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \|x\|_1,$$

where

- *$f : \mathbb{R}^n \to \mathbb{R}$ is **continuously differentiable convex function**.*
- *$\lambda > 0, \lambda \in \mathbb{R}$ is the regularization parameter.*

The proximal gradient iteration x_{k+1} is given by(component-wise):

$$[x_{k+1}]_i = \mathcal{S}(x_k - \alpha \nabla f(x_k), \alpha_k \lambda) \triangleq \begin{cases} [x_k - \alpha_k \nabla f(x_k)]_i - \alpha_k \lambda, \ if \ [x_k - \alpha_k \nabla f(x_k)]_i > \alpha_k \lambda \\ [x_k - \alpha_k \nabla f(x_k)]_i + \alpha_k \lambda, \ if \ [x_k - \alpha_k \nabla f(x_k)]_i < -\alpha_k \lambda \\ 0, otherwise \end{cases}.$$

That is, x_{k+1} is the solution to

$$\min_{x \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T (x - x_k) + \lambda \|x\|_1 + \frac{1}{2\alpha_k} \|x - x_k\|_2^2$$

The algorithm is given below.

Algorithm 20: Iterative Shrinkage-Thresholding Algorithm with constant step size for L1 optimization

```
Input: Input initial $x_0 \in \mathbb{R}^n$ and Lipschitz constant $L$ for $\nabla f$
Set $k = 0$, and choose $\alpha \in (0, 1/L]$
repeat
    Evaluate $\nabla f(x_k)$.
    Compute $x_{k+1} = \mathcal{S}(x_k - \alpha \nabla f(x_k); \lambda\alpha)$.
    set $k = k + 1$.
until termination condition;
```

9.7 Notes on bibliography

For convex analysis, see [5]. For convex optimization algorithms, see [8].

For proximal algorithms, see [9].

For finance optimization, see Optimization Methods in Finance

A great online resource is the course page from Professor Stephen Boyd(`http://stanford.edu/class/ee364b/resources.html`). To learn about financial applications of convex optimization, see `http://web.stanford.edu/~boyd/cvxbook/bv_cvxbook_extra_exercises.pdf`).

BIBLIOGRAPHY

1. Rockafellar, R. *Convex Analysis* ISBN: 9780691015866 (Princeton University Press, 1997).
2. Mordukhovich, B. S. & Nam, N. M. An easy path to convex analysis and applications. *Synthesis Lectures on Mathematics and Statistics* **6,** 1–218 (2013).
3. Bertsimas, D. & Tsitsiklis, J. N. *Introduction to linear optimization* (Athena Scientific Belmont, MA, 1997).
4. Bertsekas, D. *Nonlinear programming* ISBN: 9781886529007 (Athena Scientific, 2016).
5. Bertsekas, D. P. *Convex optimization theory* (Athena Scientific Belmont, 2009).
6. Robinson, D. *Convex optimization lecture notes* (Johns Hopkins University, 2015).
7. Nesterov, Y. *Introductory lectures on convex optimization: A basic course* (Springer Science & Business Media, 2013).
8. Bertsekas, D. P. & Scientific, A. *Convex optimization algorithms* (Athena Scientific Belmont, 2015).
9. Parikh, N., Boyd, S. P., *et al.* Proximal Algorithms. *Foundations and Trends in optimization* **1,** 127–239 (2014).