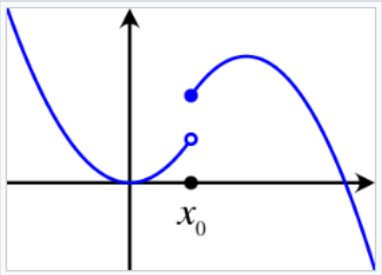
# Nonlinear Optimization: Self-Study

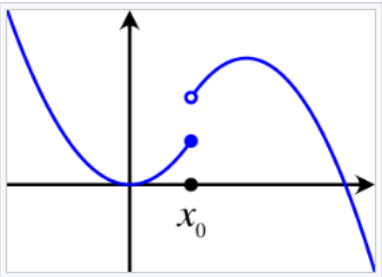
Book: Nonlinear Programming 3rd Ed. - Bertsekas

**Patrick Emami**

## Contents




An upper semi-continuous function.
The solid blue dot indicates $f(x_0)$.




A lower semi-continuous function.
The solid blue dot indicates $f(x_0)$.

Figure 1: Upper and lower semi-continuity (source: `https://en.wikipedia.org/wiki/Semi-continuity`

# 1 Appendix A - Mathematical Background

Definitions, propositions, and theorems useful for understanding the material presented in these notes. Presented here are specifically the concepts from **Appendix A** I was unfamiliar with.

**Definition A.2.** *We say that a vector $x \in \mathbb{R}^n$ is a* limit point of a subsequence $\{x^k\}$ *in $\mathbb{R}^n$ if there exists a subsequence of $\{x^k\}$ that converges to $x$.*

**Definition A.4.** *Let $X$ be a subset of $\mathbb{R}^n$.*

*(a) A real-valued function $f : X \to \mathbb{R}$ is called* upper semicontinuous *(respectively,* lower semicontinuous*) at a vector $x \in X$ if $f(x) \geq \limsup_{k\to\infty} f(x_k)$ [respectively, $f(x) \leq \liminf_{k\to\infty} f(x_k)$] for every sequence $\{x_k\} \subset X$ that converges to $x$.* (See Figure 1).

*(b) A function $f : X \to \mathbb{R}$ is called* coercive *if for every sequence $\{x_k\} \subset X$ such that $\|x_k\| \to \infty$, we have $\lim_{k\to\infty} f(x_k) = \infty$.*

**Proposition A.23** (**Second Order Expansion**). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open sphere $S$ centered at a vector $x$.*

*(a) For all $y$ such that $x + y \in S$,*

$$f(x+y) = f(x) + y'\nabla f(x) + \frac{1}{2}y'\left(\int_0^1 \left(\int_0^t \nabla^2 f(x+\tau y)d\tau\right)dt\right)y.$$

*(b) For all $y$ such that $x + y \in S$, there exists an $\alpha \in [0, 1]$ such that*

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

*(c) For all $y$ such that $x + y \in S$ there holds,*

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

**Proposition A.24 (Descent Lemma).** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, and let $x$ and $y$ be two vectors in $\mathbb{R}^n$. Suppose that*

$$\|\nabla f(x + ty) - \nabla f(x)\| \leq Lt\|y\|, \qquad \forall t \in [0, 1],$$

*where $L$ is some scalar. Then*

$$f(x + y) \leq f(x) + y'\nabla f(x) + \frac{L}{2}\|y\|^2.$$

# 2 Appendix B - Convex Analysis

These results are presented without proof; most are actually omitted from the book anyways since the author refers the reader to his book on Convex Optimization.

## 2.1 Appendix B.1 - Convex Sets and Functions, 1/16/17

### Some definitions and properties of convex sets and functions

A subset $C$ of $\mathbb{R}^n$ is called *convex* if

$$\alpha x + (1 - \alpha)y \in C, \qquad \forall x, y \in C, \quad \forall \alpha \in [0, 1].$$

Some important properties of convex sets, presented without proof:

(a) For any collection $\{C_i \mid i \in I\}$ of convex sets, the set intersection $\cap_{i \in I} C_i$ is convex.

(b) The vector sum of two convex sets is convex.

(c) The image of a convex set under a linear transformation is convex.

(d) If $C$ is a convex set and $f : C \to \mathbb{R}$ is a convex function, the level sets $\{x \in C \mid f(x) \leq \alpha\}$ and $\{x \in C \mid f(x) < \alpha\}$ are convex for all scalars $\alpha$.

A function $f : C \to \mathbb{R}$ is called *convex* if

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

The function $f$ is *concave* if $-f$ is convex. The function $f$ is called *strictly convex* if the above inequality is strict for all $x, y \in C$ with $x \neq y$, and all $\alpha \in (0, 1)$

A special case of Jensen's Inequality gives us the following

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \leq \sum_{i=1}^{m} \alpha_i f(x_i)$$

for $x_1, ..., x_m \in C$, $\alpha_1, ..., \alpha_m \geq 0$, and $\sum_{i=1}^{m} \alpha_i = 1$.

The following provides means for recognizing convex functions

(a) A linear function is convex.

(b) Any vector norm is convex.

(c) The weighted sum of convex functions, with positive wieghts, is convex.

**Characterizations of Differentiable Convex Functions**

(a) $f$ is convex over $C$ if and only if

$$f(z) \geq f(x) + (z - x)'\nabla f(x) \qquad \forall x, z \in C$$

Note that one can easily picture this for the simple case of the quadratic function.

(b) $f$ is strictly convex over $C$ if and only if the above inequality is strict whenever $x \neq z$

(c) if $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$, then $f$ is convex over $C$.

(d) if $\nabla^2 f(x)$ is positive definite for all $x \in C$, then $f$ is **strictly convex** over $C$.

(e) $f$ is **strongly convex** if for some $\sigma > 0$, we have

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

(f) If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and strongly convex in that there is some $\sigma$ satisfying the inequality from above, then $f$ is strictly convex. If in addition, $\nabla f$ satisfies the Lipschitz condition

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

for some $L > 0$, then we have for all $x, y \in \mathbb{R}^n$

$$(\nabla f(x) - \nabla f(y))'(x - y) \geq \frac{\sigma L}{\sigma + L}\|x - y\|^2 + \frac{1}{\sigma + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

(g) If $f$ is twice continuously differentiable over $\mathbb{R}^n$, then $f$ satisfies (e) if and only if the matrix $\nabla^2 f(x) - \sigma I$, where $I$ is the identity, is positive semidefinite for every $x \in \mathbb{R}^n$

## 2.2 Appendix B.1 - Convex Sets and Functions, 1/18/17

### Convex and Affine Hulls

Let $X$ be a subset of $\mathbb{R}^n$. A *convex combination* of elements of $X$ is a vector of the form $\sum_{i=1}^m \alpha_i x_i$, where $x_1, ..., x_m$ belong to $X$ and $\alpha_1, ..., \alpha_m$ are scalars such that

$$\alpha_i \geq 0, \quad i = 1, ..., m, \quad \sum_{i=1}^m \alpha_i = 1.$$

The *convex hull* of $X$ is the set of all convex combinations of elements of $X$. In particular, if $X$ consists of a finite number of vectors $x_1, ..., x_m$, its convex hull is

$$conv(X) = \left\{ \sum_{i=1}^m \alpha_i x_i \,|\, \alpha_i \geq 0, \;\; i = 1, ..., m, \;\; \sum_{i=1}^m \alpha_i = 1 \right\}$$

The *affine hull* of a subspace $S$ is the set of all affine combinations of elements of S,

$$\text{aff}(S) = \left\{ \sum_{i=1}^m \alpha_i x_i \,|\, x_i \in S, \;\; \sum_{i=1}^m \alpha_i = 1 \right\}.$$

A set in a vector space is affine if it contains all of the lines generated by its points. The affine hull is also the intersection of all linear manifolds containing $S$, where linear manifolds are translations of a vector subspace. Note that aff(S) is itself a linear manifold and it contains conv(S).

### Topological Properties of Convex Sets

Let $C$ be a convex subset of $\mathbb{R}^n$. We say that $x$ is a *relative interior point* of $C$ if $x \in C$ and there exists a neighborhood $N$ of $x$ such that $N \cap$ aff(C) $\subset C$, i.e., if $x$ is an interior point of $C$ relative to aff(C). The *relative interior of* $C$ is the set of all relative interior points of $C$.

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then it is continuous. More generally, if $C \subset \mathbb{R}^n$ is convex and $f : C \to \mathbb{R}$ is convex, then $f$ is continuous in the relative interior of $C$. Note that every function that is finite and convex on an open interval is continuous on that interval. The proof for this uses the fact that the left-hand and right-hand derivatives can be shown to exist at every point in the open interval. Alternatively, one can use the fact that $\nabla f$ satisfies the Lipschitz condition.

The set of minimizing points of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ over a closed convex set $X$ is nonempty and compact if and only if all its level sets,

$$L_a = \{x \in X \,|\, f(x) \leq a\}, \quad a \in \mathbb{R},$$

are compact.

## 2.3 Appendix B.2 - Hyperplanes, 1/20/17

A *hyperplane* in $\mathbb{R}^n$ is a set $H = \{x \,|\, a'x = b\}$, where $a$ is a nonzero vector in $\mathbb{R}^n$ and $b$ is a scalar. Note that hyperplanes are convex sets. The hyperplane can also be described as an affine set that is parallel to the subspace $\{x \,|\, a'x = 0\}$. This is because one can describe a hyperplane also as

$$H = \bar{x} + \{x \mid a'x = 0\}$$

for $\bar{x} \in H$, or

$$H = \{x \mid a'x = a'\bar{x}\}.$$

**Theorem 1** (Supporting Hyperplane)**.** *If $C \subset \mathbb{R}^n$ is a convex set and $\bar{x}$ is a point that does not belong to the interior of $C$, there exists a vector $a \neq 0$ such that*

$$a'x \geq a'\bar{x}, \quad \forall x \in C.$$

In fact, one can think of $\bar{x}$ as being on the boundary of $C$. In general, a supporting hyperplane of a convex set $C$ is one that entirely contains $C$ in one of the two closed half-spaces bounded by the hyperplane. Also, $C$ has at least one boundary-point on the hyperplane, and perhaps multiple supporting hyperplanes at a single boundary point.

**Theorem 2** (Separating Hyperplane)**.** *If $C_1$ and $C_2$ are two nonempty and disjoint convex subsets of $\mathbb{R}^n$, there exists a hyperplane that separates them, i.e., a vector $a \neq 0$ such that*

$$a'x_1 \leq a'x_2, \quad x_1 \in C_1,\, x_2 \in C_2.$$

**Theorem 3** (Strict Separation Theorem)**.** *If $C_1$ and $C_2$ are two nonempty and disjoint convex sets such that $C_1$ is closed and $C_2$ is compact, there exists a hyperplane that strictly separates them, i.e., a vector $a \neq 0$ and a scalar $b$ such that*

$$a'x_1 < b < a'x_2, \quad x_1 \in C_1,\, x_2 \in C_2.$$

We can thus characterize a convex set as the intersection of the halfspaces that contain it.

**Theorem 4** (Proper Separation)**.** *(a) Let $C_1$ and $C_2$ be two nonempty convex subsets of $\mathbb{R}^n$. There exists a hyperplane that separates $C_1$ and $C_2$, and does not contain both $C_1$ and $C_2$ if and only if*

$$ri(C_1) \cap ri(C_2) = \emptyset.$$

*(b) Let $C$ and $P$ be two nonempty convex subsets of $\mathbb{R}^n$ such that $P$ is the intersection of a finite number of closed halfspaces. There exists a hyperplane that separates $C$ and $P$, and does not contain $C$ if and only if*

$$ri(C) \cap P = \emptyset.$$

See `http://www.unc.edu/~normanp/890part4.pdf` and `http://people.hss.caltech.edu/~kcb/Notes/SeparatingHyperplane.pdf` for proofs of various theorems and properties mentioned above.

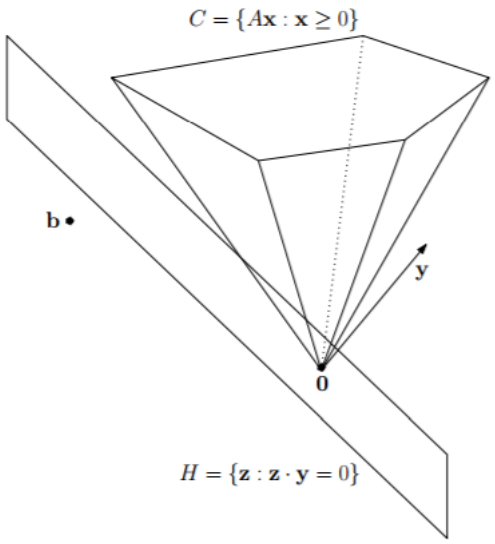



Figure 2: Farkas' Lemma (source: `http://www.sfu.ca/~mdevos/notes/misc/LP.pdf`

## 2.4 Appendix B.3 - Cones and Polyhedral Convexity, 1/23/17

A subset $C$ of a vector space $V$ is a cone (sometimes called linear cone) if for each $x \in C$ and **positive** scalars $\alpha$, the product $\alpha x$ is in $C$. A cone $C$ is a **convex cone** if $\alpha x + \beta y$ belongs to $C$, for positive scalars $\alpha$, $\beta$, and $x, y \in C$. The **polar cone** of $C$ is given by

$$C^{\perp} = \{y \,|\, y'x \leq 0, \quad \forall x \in C\}.$$

The polar cone of a subspace is the orthogonal complement, i.e., $C^{\perp} = -C^{*}$.
A *finitely generated* cone has the form

$$C = \left\{x \,|\, x = \sum_{j=1}^{r} \mu_j a_j, \ \mu_j \geq 0, j = 1, ..., r\right\},$$

where $a_1, ..., a_r$ are some vectors. A cone $C$ is *polyhedral* if it has the form

$$C = \{x \,|\, a_j' x \leq 0, \ j = 1, ..., r\},$$

where $a_1, ..., a_r$ are some vectors. Note that all of these cones are convex.

**Theorem 5** (Polar Cone Theorem)**.** *For any nonempty closed convex cone $C$, we have $(C^{\perp})^{\perp} = C$.*

(*Farkas' Lemma*) Let $x$, $e_1, ..., e_m$, and $a_1, ..., a_r$ be vectors of $\mathbb{R}^n$. We have $x'y \leq 0$ for all vectors $y \in \mathbb{R}^n$ (i.e., $x$ is in a polar cone), such that

$$y'e_i = 0, \ \ \forall i = 1, ..., m \qquad y'a_j \leq 0, \ \ \forall j = 1, ..., r,$$

if and only if $x$ can be expressed as

$$x = \sum_{i=1}^{m} \lambda_i e_i + \sum_{j=1}^{r} \mu_j a_j,$$

where $\lambda_i$ and $\mu_j$ are some scalars with $\mu_j \geq 0$ for all $j$. This is a result stating that a vector is either in a given convex cone or that there exists a hyperplane separating the vector from the cone- there are no other possibilities.

A subset of $\mathbb{R}^n$ is a *polyhedral set* if it is nonempty and it is the intersection of a finite number of closed halfspaces, i.e., if it is of the form

$$P = \{x \mid a_j' x \leq b_j, \;\; j = 1, ..., r\},$$

where $a_j$ are some vectors and $b_j$ are some scalars. A set $P$ is polyhedral if and only if it is the sum of a finitely generated cone and the convex hull of a finite set of points.

## 2.5 Appendix B.4 - Extreme Points and LP, 1/25/17

A vector $x$ is said to be an extreme point of a convex set $C$ if $x$ belongs to $C$ and there do not exist vectors $y, z \in C$, and a scalar $\alpha \in (0, 1)$ such that

$$y \neq x, \quad z \neq x, \quad x = \alpha y + (1 - \alpha) z,$$

Thinking about this, every point on a circle in $\mathbb{R}^2$ is an extreme point of the convex set consisting of these points.

Some important facts about extreme points:

1. if $H$ is a hyperplane that passes through a boundary point of $C$ and contains $C$ in one of its halfspaces, then every extreme point of $C \cap H$ is also an extreme point of $C$.
2. $C$ has at least one extreme point if and only if it does not contain a line, i.e., a set $L$ of the form $L = \{x + \alpha d \mid \alpha \in \mathbb{R}\}$ with $d \neq 0$.
3. Let $C$ be a closed convex subset of $\mathbb{R}^n$, and let $C^*$ be the set of minima of a concave function $f : C \to \mathbb{R}$ over $C$. Then if $C$ is closed and contains at least one extreme point, and $C^*$ is nonempty, then $C^*$ contains some extreme point of $C$.

**Proposition B.19.** *Let $C$ be a closed convex set and let $f : C \to \mathbb{R}$ be a concave function. Assume that for some invertible $n \times n$ matrix $A$ and some $b \in \mathbb{R}^n$ we have*

$$Ax \geq b, \;\; \forall x \in C.$$

*Then if $f$ attains a minimum over $C$, it attains a minimum at some extreme point of $C$.*

Now, important facts concerning polyhedral sets.

Let $P$ be a polyhedral set in $\mathbb{R}^n$.

1. If $P$ has the form
$$P = \{x \mid a_j' x \leq b_j, \; j = 1, ..., r\},$$
then a vector $v \in P$ is an extreme point of $P$ if and only if the set
$$A_v = \{a_j \mid a_j' v = b_j, \; j \in \{1, ..., r\}\},$$
contains $n$ linearly independent vectors.

2. If $P$ has the form
$$P = \{x \mid Ax = b,\ x \geq 0\},$$
where $A$ is a given $m \times n$ matrix and $b$ is a given vector, then a vector $v \in P$ is an extreme point of $P$ if and only if the columns of $A$ corresponding to the nonzero coordinates of $v$ are linearly independent.

3. (*Fundamental Theorem of Linear Programming*) Assume that $P$ has at least one extreme point. Then if a linear function attains a minimum over $P$, it attains a minimum at some extreme point of $P$.

*Proof.* For (3): Since $P$ is polyhedral, it has a representation
$$P = \{x \mid Ax \geq b\},$$
for some $m \times n$ matrixc $A$ and some $b \in \mathbb{R}^n$. If $A$ had rank less than $n$, then its nullspace would contain some nonzero vector $\bar{x}$, so $P$ would contain a line parallel to $\bar{x}$, contradicting the existence of an extreme point. Thus $A$ has rank $n$ and hence it must contain $n$ linearly independent rows that constitute an $n \times n$ invertible submatrix $\hat{A}$. If $\hat{b}$ is the corresponding subvector of $b$, we see that every $x \in P$ satisfies $\hat{A}x \geq \hat{b}$. The result then follows by B.19. □

### 2.6 Appendix B.5 - Differentiability Issues, 1/27/17

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we say that a vector $d \in \mathbb{R}^n$ is a *subgradient* of $f$ at a point $x \in \mathbb{R}^n$ if
$$f(z) \geq f(x) + (z - x)'d,$$
The set of all subgradients of a convex function is called the *subdifferential* of $f$ at $x$, and is denoted by $\partial f(x)$.

A vector $x^* \in X$ minimizes $f$ over a convex set $X \subset \mathbb{R}^n$ if and only if there exists a subgradient $d \in \partial f(x^*)$ such that
$$d'(z - x^*) \geq 0, \quad \forall z \in X.$$
For the special case where $X = \mathbb{R}^n$, we obtain a basic necessary and sufficient condition for unconstrained optimality of $x^*$, namely $0 \in \partial f(x^*)$.

## 3 Chapter 1 - Unconstrained Optimization: Basic Methods

### 3.1 Chapter 1.1 - Optimality Conditions, 1/30/17

**Proposition 1.1.1** (**Necessary Optimality Conditions**). *Let $x^*$ be an unconstrained local minimum of $f : \mathbb{R}^n \to \mathbb{R}$, and assume that $f$ is continuously differentiable in an open set $S$ containing $x^*$. Then*
$$\nabla f(x^*) = 0.$$
*If in addition $f$ is twice continuously differentiable within $S$, then*
$$\nabla^2 f(x^*) : \textit{positive semidefinite.}$$

*Proof.* Fix some arbitrary $d \in \mathbb{R}^n$. Then, using the chain rule to differentiate the function $g(\alpha) = f(x^* + \alpha d)$ of the scalar $\alpha$, we have

$$0 \leq \lim_{\alpha \downarrow 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \frac{dg(0)}{d\alpha} = d' \nabla f(x^*),$$

where the inequality follows because we assume that $x^*$ is a local minimum. Since $d$ is arbitrary, we can replace it with $-d$ and the inequality still holds. Therefore, $d' \nabla f(x^*) = 0$ for all $d \in \mathbb{R}^n$, which shows that $\nabla f(x^*) = 0$.

Assume that $f$ is twice continuously differentiable, and let $d$ be any vector in $\mathbb{R}^n$. For all $\alpha \in \mathbb{R}$, the second order expansion yields

$$f(x^* + \alpha d) - f(x^*) = \alpha \nabla f(x^*)' d + \frac{\alpha^2}{2} d' \nabla^2 f(x^*) d + o(\alpha^2).$$

Using the condition $\nabla f(x^*) = 0$ and the local optimality of $x^*$, we see that there is a sufficiently small $\epsilon > 0$ such that for all $\alpha$ with $\alpha \in (0, \epsilon)$,

$$0 \leq \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \frac{1}{2} d' \nabla^2 f(x^*) d + \frac{o(\alpha^2)}{\alpha^2}.$$

Taking the limit as $\alpha \to 0$ and using $\lim_{\alpha \to 0} o(\alpha^2)/\alpha^2 = 0$, we obtain $d' \nabla^2 f(x^*) d \geq 0$, showing that $\nabla^2 f(x^*)$ is positive semidefinite. □

For the convex case where both $f$ and the constraint set $X$ are convex;

**Proposition 1.1.2.** *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. If in addition $f$ is strictly convex over $X$, then $f$ has at most one global minimum over $X$. Moreover, if $f$ is strongly convex and $X$ is closed, then $f$ has a unique global minimum over $X$.*

The proof consists of simple applications of the convexity definitions from **Appendix B**.

**Proposition 1.1.3** (**Necessary and Sufficient Conditions for Convex Case**). *Let $X$ be a convex set and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function over $X$.*

*(a) If $f$ is continuously differentiable, then*

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

*is a necessary and sufficient condition for a vector $x^* \in X$ to be a global minimum of $f$ over $X$.*

*(b) If $X$ is open and $f$ is continuously differentiable over $X$, then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for a vector $x^* \in X$ to be a global minimum of $f$ over $X$.*

**Proposition 1.1.5** (**Second Order Sufficient Optimality Conditions**). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open set $S$. Suppose that a vector $x^* \in S$ satisfies the conditions*

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) \ : \text{ positive definite.}$$

*Then, $x^*$ is a strict unconstrained local minimum of $f$. In particular, there exist scalars $\gamma > 0$ and $\epsilon > 0$ such that*

$$f(x) \geq f(x^*) + \frac{\gamma}{2} \|x - x^*\|^2, \qquad \forall x \ \text{ with } \ \|x - x^*\| < \epsilon.$$

## 3.2 Chapter 1.2 - Gradient Methods - Convergence, 2/7/17

### Gradient Methods

Most of the interesting algorithms for unconstrained minimization of a continuously differentiable function are *iterative descent methods*. Given a vector $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$, consider the half line of vectors

$$x_a = x - \alpha \nabla f(x), \qquad \forall \alpha \geq 0.$$

More generally, consider the half line of vectors

$$x_a = x + \alpha d, \qquad \forall \alpha \geq 0,$$

where the direction vector $d \in \mathbb{R}^n$ makes an angle with $\nabla f(x)$ that is greater than 90 degrees, i.e.,

$$\nabla f(x)' d < 0.$$

We have $f(x_a) = f(x) + \alpha \nabla f(x)' d + o(\alpha)$ from the first order expansion about $x$. For $\alpha$ near zero, the term $\alpha \nabla f(x)' d$ dominates $o(\alpha)$ and as a result, for positive but sufficiently small $\alpha$, $f(x + \alpha d)$ is smaller than $f(x)$. This forms the basis for a broad class of iterative algorithms;

$$x^{k+1} = x^k + \alpha^k d^k, \qquad k = 0, 1, ...,$$

with proper choice of direction $d^k$. This algorithm is known as the *gradient method* when $d^k = -\nabla f(x^k)$.

### Selecting the Descent Direction

Many gradient methods are specified in the form

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k),$$

where $D^k$ is a positive definite symmetric matrix. Since $d^k = -D^k \nabla f(x^k)$, the descent condition $\nabla f(x^k)' d^k < 0$ is written as

$$\nabla f(x^k)' D^k \nabla f(x^k) > 0,$$

and holds thanks to the positive definiteness of $D^k$.

For the *steepest descent* algorithm,

$$D^k = I, \quad k = 0, 1, ...$$

The name is derived from the property of the (normalized) negative gradient direction

$$d^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

Among all directions $d \in \mathbb{R}^n$ that are normalized so that $\|d\| = 1$, it is the one that minimizes the slope $\nabla f(x^k)' d$ of the cost $f(x^k + \alpha d)$ along the direction $d$ at $\alpha = 0$. By the Schwartz inequality,

$$\nabla f(x^k)' d \geq -\|\nabla f(x^k)\| \cdot \|d\| = -\|\nabla f(x^k)\|,$$

and equality is obtained with the aforementioned negative gradient direction for $d^k$.

*Newton's Method* involves selecting

$$D^k = (\nabla^2 f(x^k))^{-1}, \quad k = 0, 1, ...,$$

provided $\nabla^2 f(x^k)$ is positive definite. Taking the quadratic approximation of $f$ around the current point and setting the first derivative to 0 produces the desired gradient direction. Therefore, in general, the Newton iteration is

$$x^{k+1} = x^k - \alpha^k (\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$

Note that Newton's method finds the global minimum of a positive definite quadratic function in a single iteration (assuming $\alpha^k = 1$).

*Diagonally Scaled Steepest Descent* sets $D^k$ to be an $n \times n$ diagonal matrix where the diagonal entries are positive scalars to ensure positive definiteness. A popular choice is to choose the diagonal entries so that $D^k$ approximates the inverted second partial derivative of $f$ w.r.t. to $x_i$.

Other choices include computing the Hessian at the first iteration and only recomputing every $p > 1$ iterations (or never), as well as using a finite-difference approximation to the Hessian.

The *Gauss Newton Method* applies to the problem of minimizing the sum of squares of real-valued functions $g_1, g_2, ..., g_m$,

$$\text{minimize } f(x) = \frac{1}{2}\|g(x)\|^2 = \frac{1}{2}\sum_{i=1}^{m}(g_i(x))^2$$

$$\text{subject to } x \in \mathbb{R}^n.$$

We choose

$$D^k = (\nabla g(x^k) \nabla g(x^k)')^{-1}, \qquad k = 0, 1, ...$$

$\nabla g(x^k) \nabla g(x^k)'$ is positive definite and hence invertible if and only if the matrix $\nabla g(x^k)$ has rank $n$. The Gauss-Newton method takes the form

$$x^{k+1} = x^k - \alpha^k (\nabla g(x^k) \nabla g(x^k)')^{-1} \nabla g(x^k) g(x^k).$$

**Stepsize Selection**

The *Minimization Rule* chooses $\alpha^k$ such that the cost function is minimized along the direction $d^k$, i.e., $\alpha^k$ satisfies

$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k).$$

The *Limited Minimization Rule* simply uses a fixed scalar $s$ and chooses $\alpha^k$ that yields the greatest cost reduction over all stepsizes in the interval $[0, s]$, i.e.,

$$f(x^k + \alpha^k d^k) = \min_{\alpha \in [0,s]} f(x^k + \alpha d^k).$$

These can be implemented using one-dimensional line search algorithms. In practice, the line search is stopped once a stepsize $\alpha^k$ satisfies some termination criterion since the minimum cannot always

be computed exactly. These methods trade off more function and/or gradient evaluations for fewer required iterations, because of the greater cost reduction per iteration they achieve.

Since line minimization can sometimes incur considerable additional computation, there are alternatives based on successive stepsize reduction. For example, the simplest rule initially selects a stepsize, and if the corresponding vector $x^k + sd^k$ does not yield an improved value of $f$, the stepsize is reduced by a certain factor until the value of $f$ is improved. This often works in practice but is theoretically unsound, because the cost improvement obtained at each iteration may not be substantial enough to guarantee convergence to a minimum.

The *Armijo rule* modifies the scheme described above by introducing scalars $s$, $\beta$, and $\sigma$ with $0 < \beta < 1$, and $0 < \sigma < 1$. We set $\alpha^k = \beta^{m_k} s$, where $m_k$ is the first nonnegative integer $m$ for which

$$f(x^k) - f(x^k - \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(x^k)' d^k.$$

The stepsizes $\beta^m s$ are tried succesively until the above inequality is satisfed. This ensures that the cost improvement is sufficiently large. Usually, $\sigma$ is chosen close to zero, e.g., $\sigma \in [10^{-5}, 10^{-1}]$. The reduction factor $\beta$ is chosen from 1/2 to 1/10 depending on the confidence we have on the initial stepsize $s$. Many Newton-like methods incorporate some implicit scaling of the direction $d^k$, which makes $s = 1$ a good stepsize choice. See Figure 1.2.7 for a nice depiction of the Armijo rule in action.

Choosing a *constant step size* is usually only successful when an appropriate value is known or can be determined fairly easily. A *diminishing stepsize* such that $\alpha^k \to 0$ does not guarantee descent at each iteration, but descent becomes more likely as the stepsize diminishes. To ensure that progress can be maintained even when far from a stationary point, we require that

$$\sum_{k=0}^{\infty} \alpha^k = \infty.$$

Generally, this has good theoretical convergence properties but the associated convergence rate tends to be slow. Hence, this is useful in situations where slow convergence is inevitable, e.g., in singular problems or when the gradient is calculated with error.

**Convergence Results**

Gradient methods are guided downhill by local information about the $f$. The most we can expect from a gradient method is that it converges to a stationary point. One way to encourage a gradient descent method to not get stuck at nonstationary points is by prevent the descent direction from asymptotically becoming orthogonal to the gradient direction. For e.g., if the eigenvalues of the positive definite symmetric matrix $D^k$ are bounded above and bounded away from zero. A more general condition is as follows:

Consider the sequence $\{x^k, d^k\}$ generated by a given gradient method. We say that the direction sequence $\{d^k\}$ is *gradient related* to $\{x^k\}$ if the following property can be shown:

*For any subsequence $\{x^k\}_{k \in K}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k \in K}$ is bounded and satisfies*

$$\lim_{k \to \infty,\, k \in K} \sup \nabla f(x^k)' d^k < 0.$$

This is a "nonorthongality" type of condition, which is quite general. Roughly, this means that $d^k$ does not become "too small" or "too large" relative to $\nabla f(x^k)$, and that the angle between $d^k$ and $\nabla f(x^k)$ does not get "too close" to 90 degrees. We care about this because we do not want our gradient method to converge onto a nonstationary point; in fact, it follows that if a subsequence $\{\nabla f(x^k)\}$ tends to a nonzero vector (i.e., a nonstationary point), the corresponding subsequence of directions $d^k$ is bounded and does not tend to be orthogonal with the gradient.

**Proposition 1.2.1** (**Stationarity of Limit Points for Gradient Methods**). *Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, and assume that $\{d^k\}$ is gradient related and $\alpha^k$ is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of $\{x^k\}$ is a stationary point.*

*Proof.* Consider first the Armijo rule and let $\bar{x}$ be a limit point of $\{x^k\}$. Since $\{f(x^k)\}$ is monotonically nonincreasing, $\{f(x^k)\}$ either converges to a finite value or diverges to $-\infty$. Since $f$ is continuous, $f(\bar{x})$ is a limit point of $\{f(x^k)\}$, so it follows that the entire sequence $\{f(x^k)\}$ converges to $f(\bar{x})$, and

$$f(x^k) - f(x^{k+1}) \to 0. \tag{1.17}$$

Moreover, by the definition of the Armijo rule, we have

$$f(x^k) - f(x^{k+1}) \geq -\sigma \alpha^k \nabla f(x^k)' d^k, \tag{1.18}$$

so the right-hand side in the above relation tends to 0.
Let $\{x^k\}_K$ be a subsequence converging to $\bar{x}$, and assume to arrive at a contradiction that $\bar{x}$ is nonstationary (recall the definition of gradient related!). Since $\{d^k\}$ is gradient related, we have

$$\lim_{k\to\infty,\, k\in K} \sup \nabla f(x^k)' d^k < 0,$$

and therefore from Eqs. (1.17) and (1.18),

$$\{\alpha^k\}_K \to 0.$$

Hence, by the definition of the Armijo rule, we must have for some index $\bar{k} \geq 0$

$$f(x^k) - f(x^k + (\alpha^k/\beta) d^k) < -\sigma(\alpha^k/\beta) \nabla f(x^k)' d^k, \qquad \forall k \in K, k \geq \bar{k}, \tag{1.19}$$

i.e., the initial stepsize $s$ will be reduced at least once $\forall k \in K, k \geq \bar{k}$. Since $\{d^k\}$ is gradient related, $\{d^k\}_K$ is bounded, so there exists a subsequence $\{d^k\}_{\bar{K}}$ of $\{d^k\}_K$ such that

$$\{d^k\}_{\bar{K}} \to \bar{d},$$

where $\bar{d}$ is some vector. From Eq. (1.19), we have

$$\frac{f(x^k) - f(x^k + \bar{\alpha}^k d^k)}{\bar{\alpha}^k} < -\sigma \nabla f(x^k)' d^k, \qquad \forall k \in \bar{K}, k \geq \bar{k}, \tag{1.20}$$

where $\bar{\alpha}^k = (\alpha^k/\beta)$. By using the mean value theorem, this relation is written as

$$-\nabla f(x^k + \tilde{\alpha}^k d^k)' d^k < -\sigma \nabla f(x^k)' d^k, \qquad \forall k \in \bar{K}, k \geq \bar{k},$$

where $\tilde{\alpha}^k$ is a scalar in the interval $[0, \bar{\alpha}^k]$. Taking limits in the above relation we obtain

$$-\nabla f(\bar{x})'\bar{d} \leq -\sigma \nabla f(\bar{x})'\bar{d}$$

or

$$0 \leq (1-\sigma)\nabla f(\bar{x})'\bar{d}.$$

Since $\sigma < 1$, it follows that

$$0 \leq \nabla f(\bar{x})'\bar{d}, \tag{1.21}$$

which contradicts the assumption that $\{d^k\}$ is gradient related. This proves the result for the Armijo rule. For the minimization and limited minimization cases, it can easily be shown that the line of argument just used establishes that any stepsize rule that gives a larger reduction in cost at each iteration than the Armijo rule inherits the convergence properties of the latter. □

**Proposition 1.2.2** (**Convergence of Constant Stepsize**). *Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that the Lipschitz condition holds, and that for all $k$ we have $d^k \neq 0$ and*

$$\epsilon \leq \alpha^k \leq (2-\epsilon)\bar{\alpha}^k, \tag{1.24}$$

*where*

$$\bar{\alpha}^k = \frac{|\nabla f(x^k)'d^k|}{L\|d^k\|^2},$$

*and $\epsilon \in (0,1]$ is a fixed scalar. Then every limit point of $\{x^k\}$ is a stationary point of $f$.*

*Proof.* Using a slightly modified version of A.24, one can show that $f(x) \leq f(y) + \nabla f(x)'(x-y) + \frac{L}{2}\|x-y\|^2$. Plugging $y = x^{k+1}$ and $x = x^k$ into this equation, we obtain

$$\begin{aligned} f(x^k) - f(x^k + \alpha^k d^k) &\geq -\alpha^k \nabla f(x^k)'d^k - \frac{1}{2}(\alpha^k)^2 L\|d^k\|^2 \\ &= \alpha^k(\|\nabla f(x^k)'d^k\| - \frac{1}{2}\alpha^k L\|d^k\|^2). \end{aligned}$$

The rhs of Eq. (1.24) yields

$$|\nabla f(x^k)'d^k| - \frac{1}{2}\alpha^k L\|d^k\|^2 \geq \frac{1}{2}\epsilon|\nabla f(x^k)'d^k|.$$

Using this relation together with the condition $\alpha^k \geq \epsilon$ in the inequality, the cost improvements at iteration $k$ is bounded:

$$f(x^k) - f(x^k + \alpha^k d^k) \geq \frac{1}{2}\epsilon|\nabla f(x^k)'d^k|.$$

If a subsequence $\{x^k\}_K$ converges to a nonstationary point, we must have that $f(x^k) - f(x^{k+1}) \to 0$, and the preceding relation implies that $|\nabla f(x^k)'d^k| \to 0$. This contradicts the assumption that $\{d^k\}$ is gradient related. Hence, every limit point of $\{x^k\}$ is stationary. □

The idea here is that if the curvature of $f$ is bounded by the Lipschitz condition, then one can construct a quadratic function that overestimates $f$. An appropriate constant stepsize can then be obtained within an interval around the scalar that minimizes this quadratic function along the direction $d^k$. In the case of steepest descent, the condition (1.24) becomes

$$\epsilon \leq \alpha^k \leq \frac{2-\epsilon}{L}.$$

Thus a constant stepsize roughly in the middle of the interval $[0, 2/L]$ guarantees convergence. Notice that $f$ is not required to be convex.

The Lipschitz continuity condition also essentially guarantees convergence for a *diminishing stepsize*.

**Proposition 1.2.2** (**Convergence of Diminishing Stepsize**). *Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$. Assume that the Lipschitz condition holds, and that there exist positive scalars $c_1$, $c_2$ such that for all $k$ we have*

$$c_1\|\nabla f(x^k)\|^2 \leq -\nabla f(x^k)'d^k, \qquad \|d^k\|^2 \leq c_2\|\nabla f(x^k)\|^2.$$

*Suppose also that*

$$\alpha^k \to 0, \qquad \sum_{k=0}^{\infty} \alpha^k = \infty.$$

*Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of $f$.*

The section concludes with the presentation and proof of the *Capture Theorem*. This essentially states that local minima which are sufficiently "isolated" tend to attract gradient methods: once the method gets close enough to such a minimum it remains close and converges to it. The conditions $f(x^{k+1}) \leq f(x^k)$ and $\alpha^k \leq s$ under which this theorem holds are satisfied by the Armijo rule and the limited minimization rule. They are also satisfied for a constant and a diminishing stepsize under conditions that guarantee descent at each iteration. The condition $\|d^k\| \leq c\|\nabla f(x^k)\|$ is satisfied if $d^k = -D^k\nabla f(x^k)$ with the eigenvalues of $D^k$ bounded from above.

## 3.3 Chapter 1.3 - Gradient Methods - Rate of Convergence, 7/2/2017

The three main schools of thought for analysis of rates of convergence in nonlinear programming are the computational complexity approach, the informational complexity approach, and local analysis. The latter provides an accurate description of the behavior of a method near the optimal solution by using series approximations. Even though the behavior at the beginning of a method is ignored entirely, it is still the most useful and employed in this book the most.

For the minimization of any twice continuously differentiable function, a quadratic approximation around the optimal solution is quite useful for asymptotic convergence analysis in the general case. We first consider the convergence rate of steepest descent for quadratic functions. For a interactive demonstration of the following, see `http://distill.pub/2017/momentum/`.

Consider a cost function $f$ with positive definite Hessian $Q$. Thus,

$$f(x) = \frac{1}{2}x'Qx, \quad \nabla f(x^k) = Qx, \quad \nabla^2 f(x) = Q.$$

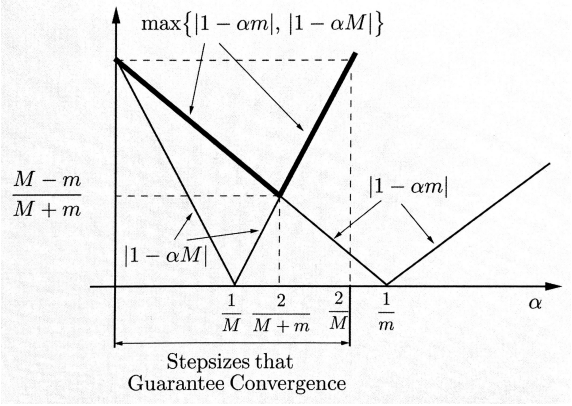



Figure 3: Illlustration of the convergence rate bound. The bound is minimized when $\alpha$ is such that $1 - \alpha m = \alpha M - 1$, i.e., for $\alpha = 2/(M + m)$

The steepest descent method takes the form

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = (I - \alpha^k Q)x^k.$$

Squaring both sides and using the fact that, for min and max eigenvalues $\lambda_1$, $\lambda_n$ of $A$, $\lambda_1 \|x\|^2 \leq x'Ax \leq \lambda_n \|x\|^2$, we have

$$(x^k)'(I - \alpha^k Q)^2 x^k = \|x^{k+1}\|^2 \leq \lambda^* \|x^k\|^2,$$

where $\lambda^*$ is the max eigenvalue of $(I - \alpha^k Q)^2$. Since the eigenvalues of $(I - \alpha^k Q)^2$ are equal to $(1 - \alpha^k \lambda_i)^2$, and $\lambda^* = \max\{(1 - \alpha^k m)^2, (1 - \alpha^k M)^2\}$ where $m$ is the smallest eigenvalue and $M$ is the largest, it follows that for $x^k \neq 0$,

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \max\{|1 - \alpha^k m|, |1 - \alpha^k M|\}.$$

From Figure 3, we can see that

$$\alpha^* = \frac{2}{M + m}.$$

Intuitively, the optimal stepsize causes the max and min eigenvalues to converge at the same rate. The best convergence rate bound for steepest descent with constant stepsize is

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \frac{M - m}{M + m} = \frac{M/m - 1}{M/m + 1}.$$

The ratio $M/m$ is the condition number of $Q$ and determines the convergence rate for a given problem. When the condition number is large, the problem is *ill-conditioned* and converges slowly.