OPTIMIZATION 1

CHAPTER 3

NONLINEAR PROGRAMMING: UNCONSTRAINED MINIMIZATION

Gradient Vector and Hessian Matrix

A real-valued function f defined on a subset of $\mathbb{R}^n$ is said to be **continuous** at $\mathbf{x}$ if $\mathbf{x}_k \to \mathbf{x}$ implies $f(\mathbf{x}_k) \to f(\mathbf{x})$.

Equivalently, f is continuous at $\mathbf{x}$ if given $\epsilon > 0$ there is $\delta > 0$ such that $|\mathbf{y} - \mathbf{x}| < \delta$ implies $|f(\mathbf{y}) - f(\mathbf{x})| < \epsilon$.

If f is continuous on some open set of $\mathbb{R}^n$, we write $f \in C$. If f has continuous partial derivatives of order k on that *open* set, we write $f \in C^k$.

A set of real-valued functions $f_1, f_2, \ldots, f_m$ on a subset $S \subset \mathbb{R}^n$ can be regarded as a single vector-valued function $\mathbf{f} = (f_1, f_2, \ldots, f_m)$. This function assigns a vector

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$$

in $\mathbb{R}^m$ to each vector $\mathbf{x} \in S$.

Such a vector-valued function is said to be **continuous** if each of its component functions is continuous.

If each component of $\mathbf{f} = (f_1, f_2, \ldots, f_m)$ is continuous on some *open* set of $\mathbb{R}^n$, then we write $\mathbf{f} \in C$.

If in addition, each component function has first partial derivatives which are continuous on this set, we write $\mathbf{f} \in C^1$.

In general, if the component functions have continuous partial derivatives of order k, we write $\mathbf{f} \in C^k$.

If $f \in C^1$ is a real-valued function on an open set of $\mathbb{R}^n$ we define the **gradient** of f to be the *row* vector

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right].$$

We sometimes use the alternative notation

$$\boldsymbol{\nabla}_{\mathbf{x}} \boldsymbol{f}(\mathbf{x})$$

for $\nabla f(\mathbf{x})$.

Example 1.1 Let

$$f(\mathbf{x}) = \mathbf{a}^T(\mathbf{x} - \mathbf{x}_0) \quad \text{and} \quad g(\mathbf{x}) = |\mathbf{x} - \mathbf{x}_0|^2.$$

Then

$$\nabla f(\mathbf{x}) = \mathbf{a}^T \quad \text{and} \quad \nabla g(\mathbf{x}) = 2(\mathbf{x}-\mathbf{x}_0)^T \quad \forall\, \mathbf{x} \in \mathbb{R}^n.$$

Significance of the Gradient Vector

The directional derivative of f at $\mathbf{x}$ in the direction of a unit vector $\mathbf{u}$ is equal to

$$D_{\mathbf{u}} f(\mathbf{x}) = \nabla f(\mathbf{x})\mathbf{u} = |\nabla f(\mathbf{x})| \cos\theta,$$

where θ is the angle between $\nabla f(\mathbf{x})^T$ and $\mathbf{u}$.

The directional derivative $D_{\mathbf{u}}f(\mathbf{x})$ is the rate of change of f per unit change in the direction of $\mathbf{u}$ at $\mathbf{x}$. Thus,

- $\nabla f(\mathbf{x})^T$ points in the direction of maximum increase of f at $\mathbf{x}$.
- $-\nabla f(\mathbf{x})^T$ points in the direction of maximum decrease of f at $\mathbf{x}$.
- $\nabla f(\mathbf{x})^T$ is normal to the level curve (or surface) of f at $\mathbf{x}$.

If $f \in C^2$ then we define the **Hessian** of f at $\mathbf{x}$ to be the $n \times n$ matrix denoted $\nabla^2 f(\mathbf{x})$ or $\mathbf{F}(\mathbf{x})$ as

$$\mathbf{F}(\mathbf{x}) := \left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\right] = \begin{bmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n} \end{bmatrix}.$$

Since $f_{x_ix_j} = f_{x_jx_i}$, the Hessian is symmetric.

Example 1.2 If $f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{A}(\mathbf{x} - \mathbf{x}_0)$, where $\mathbf{A} = [a_{ij}]$ is a symmetric matrix, then

$$\nabla f(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_0)^T \mathbf{A} \quad \text{and} \quad \mathbf{F}(\mathbf{x}) = \mathbf{A} \quad \text{for all } \mathbf{x}.$$

Example 1.3 Let f be defined on an open set $G \subset \mathbb{R}^n$ and $\mathbf{x} \in G$. For $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, consider

$$g(t) = f(\mathbf{x} + t\mathbf{d}).$$

The domain of g is an open set I in $\mathbb{R}$ that contains 0.

(a) If f is differentiable, then for all $t \in I$,

$$\boxed{g'(t) = \nabla f(\mathbf{x} + t\mathbf{d})\mathbf{d}}$$

(b) If f is twice differentiable, then for all $t \in I$,

$$\boxed{g''(t) = \mathbf{d}^T \nabla^2 f(\mathbf{x} + t\mathbf{d})\mathbf{d}}$$

For a vector-valued function $\mathbf{f} = (f_1, f_2, \ldots, f_m)$ the situation is similar. If $\mathbf{f} \in C^1$, the **first derivative** is defined as the $m \times n$ matrix

$$\nabla \mathbf{f}(\mathbf{x}) = \left[\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right] = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \frac{\partial f_m(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

For instance, if $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is given by $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ for some $m \times n$ matrix $\mathbf{A}$ and a vector $\mathbf{b} \in \mathbb{R}^m$, then $\nabla \mathbf{f}(\mathbf{x}) = \mathbf{A}$.

If $\mathbf{f} \in C^2$ it is possible to define the m Hessians $\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \ldots, \mathbf{F}_m(\mathbf{x})$ corresponding to the m component functions $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})$.

Given any vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m) \in \mathbb{R}^m$ we note that the real-valued function

$$\boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}) = \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x})$$

has gradient equal to $\boldsymbol{\lambda}^T \nabla \mathbf{f}(\mathbf{x})$ and Hessian, denoted $\boldsymbol{\lambda}^T \mathbf{F}(\mathbf{x})$, equal to

$$\boldsymbol{\lambda}^T \mathbf{F}(\mathbf{x}) := \sum_{i=1}^{m} \lambda_i \mathbf{F}_i(\mathbf{x}).$$

Quadratic Forms and Positive Definite Matrices

Definition 1.1

If $\mathbf{A}$ is an $n \times n$ symmetric matrix, then the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}$$

is called a **quadratic form** in the n variables $x_1, x_2, \ldots, x_n$. The matrix $\mathbf{A}$ is called the **matrix of the quadratic form** f.

Definition 1.2

A symmetric $n \times n$ matrix $\mathbf{A}$ and the quadratic form $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}$ are called

(i) **positive definite** if $f(\mathbf{x}) > 0$ for all $\mathbf{x}$ in $\mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$;

(ii) **negative definite** if $f(\mathbf{x}) < 0$ for all $\mathbf{x}$ in $\mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$;

(iii) **positive semidefinite** if $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$;

(iv) **negative semidefinite** if $f(\mathbf{x}) \leq 0$ for all $\mathbf{x}$;

(v) **indefinite** if f has both positive and negative values.

Note

$$\mathbf{A} \text{ is negative definite} \iff -\mathbf{A} \text{ is positive definite}$$
$$\mathbf{A} \text{ is negative semidefinite} \iff -\mathbf{A} \text{ is positive semidefinite}$$

Remark If the quadratic form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ is positive definite, then there is a positive number α such that

$$f(\mathbf{x}) \geq \alpha|\mathbf{x}|^2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Example 1.4 A symmetric matrix with some negative entries may be positive definite. For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

corresponds to the quadratic form

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 2x_1^2 - 2x_1x_2 + x_2^2 \\ &= x_1^2 + (x_1 - x_2)^2. \end{aligned}$$

Since $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} = (x_1, x_2) \neq \mathbf{0}$, $\mathbf{A}$ is positive definite.

Theorem 1.1

Let $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix. Then*

(a) $\mathbf{A}$ *is positive definite if and only if all its eigenvalues are positive;*

(b) $\mathbf{A}$ *is positive semidefinite if and only if all its eigenvalues are nonnegative;*

(c) $\mathbf{A}$ *is negative definite if and only if all its eigenvalues are negative;*

(d) $\mathbf{A}$ *is negative semidefinite if and only if all its eigenvalues are nonpositive;*

(e) $\mathbf{A}$ *is indefinite if and only if it has (at least) one positive eigenvalue and one negative eigenvalue.*

Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. Let $\mathbf{A}_k$ denote the submatrix formed by deleting the last $n - k$ rows and columns of $\mathbf{A}$,

$$\mathbf{A}_k = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}.$$

$\det(\mathbf{A}_k)$ is called the ***k*th leading principal minor** of $\mathbf{A}$.

Theorem 1.2

If $\mathbf{A}$ *is an* $n \times n$*-symmetric matrix, then:*

(a) $\mathbf{A}$ *is positive definite if and only if all the leading principal minors of* $\mathbf{A}$ *are positive (that is,* $\det(\mathbf{A}_k) > 0$ *for* $k = 1, 2, \ldots, n$*);*

(b) $\mathbf{A}$ *is negative definite if and only if the leading principal minors alternate in sign with* $\det(\mathbf{A}_1) = a_{11} < 0$ *(that is,* $(-1)^k \det(\mathbf{A}_k) > 0$ *for* $k = 1, 2, \ldots, n$*).*

Note

(a) A symmetric matrix whose entries are all positive need not be positive definite.

(b) It is not true that if $\mathbf{A}$ is an $n \times n$-symmetric matrix, then $\mathbf{A}$ is positive semidefinite if and only if the principal minors $\det(\mathbf{A}_1), \ldots, \det(\mathbf{A}_n)$ are all nonnegative.

Example 1.5 If

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & \frac{1}{2} \end{bmatrix}$$

then all principal minors of $\mathbf{A}$ are nonnegative but $\mathbf{A}$ is not positive semidefinite since, for example $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$ for $\mathbf{x} = (1, 1, -2)$.

Taylor's Theorem; Linear and Quadratic Approximations

Theorem 1.3

Suppose the function f has a derivative of order n on (a, b) and $x, x_0 \in (a, b)$.

(a) There exists a point ξ between x and x_0 such that

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \cdots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x - x_0)^{n-1} + \frac{f^{(n)}(\xi)}{n!}(x - x_0)^n.$$

Theorem 1.3 (cont'd)

(b) The following representation holds:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + o\big((x - x_0)^n\big),$$

where

$$\lim_{x \to x_0} \frac{o\big((x - x_0)^n\big)}{(x - x_0)^n} = 0.$$

Corollary 1.4

Let $g : I \subset \mathbb{R} \to \mathbb{R}$.

(i) *If* g *is differentiable on an open interval containing* 0, *then*

$$g(t) = g(0) + g'(0)t + o(t).$$

(ii) *If* g *is twice differentiable on an open interval containing* 0, *then*

$$g(t) = g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + o(t^2).$$

Theorem 1.5 (The Mean value theorem)

If $f \in C^1$ in an open set of $\mathbb{R}^n$ that contains the line segment $[\mathbf{x}, \mathbf{y}]$, then there is a $\theta \in (0, 1)$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f\big((1-\theta)\mathbf{x} + \theta\mathbf{y}\big)(\mathbf{y} - \mathbf{x}).$$

Theorem 1.6

If f has continuous second-order partial derivatives on an open set of $\mathbb{R}^n$ that contains the line segment $[\mathbf{x}, \mathbf{y}]$, then there is a $\theta \in (0, 1)$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \mathbf{F}\big(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})\big)(\mathbf{y} - \mathbf{x}).$$

Theorem 1.7

Let f have continuous second-order partial derivatives on an open subset U of $\mathbb{R}^n$ and $U \ni \mathbf{x}$. Then

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \mathbf{F}(\mathbf{x})(\mathbf{y}-\mathbf{x}) + o(|\mathbf{y}-\mathbf{x}|^2),$$

where

$$\lim_{\mathbf{y}\to\mathbf{x}} \frac{o(|\mathbf{y}-\mathbf{x}|^2)}{|\mathbf{y}-\mathbf{x}|^2} = 0.$$

In this chapter we consider optimization problems of the form

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \Omega \end{aligned} \tag{1}$$

where f is a real-valued function and Ω is a subset of $\mathbb{R}^n$.

- f is called the **objective** or **cost** function, Ω is called the **feasible set**.
- In **unconstrained problems**, the constraints are not present and thus, the feasible region Ω is the entire space $\mathbb{R}^n$.

Definition 2.1

Suppose that $f(\mathbf{x})$ is a real-valued function defined on a subset Ω of $\mathbb{R}^n$. A point $\mathbf{x}^* \in \Omega$ is said to be

(i) a **global minimum point** (or **global minimizer**) of f over Ω if

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \Omega;$$

(ii) a **strict global minimum point** (or **strict global minimizer**) of f over Ω if

$$f(\mathbf{x}^*) < f(\mathbf{x}) \text{ for all } \mathbf{x} \in \Omega,\ \mathbf{x} \neq \mathbf{x}^*.$$

$$\mathbf{x}^* \in \Omega \text{ is a global minimum point of } f$$
$$\Longleftrightarrow \ (\forall \mathbf{x} \in \Omega)\big(f(\mathbf{x}^*) \leq f(\mathbf{x})\big).$$

Definition 2.2

A point $\mathbf{x}^* \in \Omega$ is said to be

(iii) a **relative minimum point** (or **local minimizer**) of f over Ω if there is a positive number δ such that
$f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ and $|\mathbf{x} - \mathbf{x}^*| < \delta$;

(iv) a **strict relative minimum point** (or **strict local minimizer**) of f over Ω if there is a positive number δ such that
$f(\mathbf{x}^*) < f(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ and $0 < |\mathbf{x} - \mathbf{x}^*| < \delta$.

$\mathbf{x}^* \in \Omega$ is a relative minimum point of f

$$\iff (\exists \delta > 0)\big(\forall \mathbf{x} \in \Omega,\ |\mathbf{x} - \mathbf{x}^*| < \delta\big)\big(f(\mathbf{x}^*) \leq f(\mathbf{x})\big).$$

The first question that arises in the study of the minimization problem (1) is whether a solution exists.

The main result that can be used to address this issue is the theorem of Weierstrass, which states that *if f is continuous and Ω is compact, a solution exists*.

Theorem 2.1

Assume that f is a continuous function and one of the following situations holds:

(a) *The set Ω is closed and bounded;*

(b) *The set Ω is closed and not bounded, but*

$$\lim_{\mathbf{x}\in\Omega,\ |\mathbf{x}|\to\infty} f(\mathbf{x}) = \infty.$$

Then the minimization problem (1) admits at least one global solution.

In many cases the above theorem is enough to ensure the existence of an optimal solution.

- Practical reality, however, both from the theoretical and computational viewpoint, dictates that we must in many circumstances be content with a relative minimum point.
- Thus, in formulating and attacking problem (1) we shall usually consider, implicitly, that we are asking for a relative minimum point. If appropriate conditions hold, this will also be a global minimum point.

Feasible Directions

To derive necessary conditions satisfied by a relative minimum point $\mathbf{x}$, the basic idea is to consider movement away from the point in some given direction.

Along any given direction the objective function can be regarded as a function of a single variable and hence the ordinary calculus of a single variable is applicable.

Definition 2.3

Given a set $\Omega \subset \mathbb{R}^n$ and a point $\mathbf{x} \in \Omega$. A vector $\mathbf{d} \in \mathbb{R}^n$ is called a **feasible direction** at $\mathbf{x}$ if there is a $\delta > 0$ such that

$$\mathbf{x} + t\mathbf{d} \in \Omega \quad \text{for all} \quad t \in [0, \delta].$$

If $\mathbf{d}$ is a feasible direction and $\alpha > 0$ the direction $\alpha\mathbf{d}$ is also feasible. Thus the set of all feasible directions is a cone, called the **cone of feasible directions**.

Example 2.1

(a) At an interior point of Ω each direction is feasible. So the cone of feasible directions in this case is $\mathbb{R}^n$.

(b) If Ω is a convex set and $\mathbf{x} \in \Omega$, then the cone of feasible directions at $\mathbf{x}$ is

$$D = \{t(\mathbf{y} - \mathbf{x}) : t \geq 0, \ \mathbf{y} \in \Omega\}.$$

Theorem 2.2 (First-order necessary conditions)

Let Ω be a subset of $\mathbb{R}^n$ and let $f \in C^1$ be a function on Ω. If $\mathbf{x}^*$ is a relative minimum point of f over Ω, then for any $\mathbf{d} \in \mathbb{R}^n$ that is a feasible direction at $\mathbf{x}^*$, we have

$$\nabla f(\mathbf{x}^*)\mathbf{d} \geq 0.$$

Let f be a function defined on $\Omega \subset \mathbb{R}^n$ and $\mathbf{x} \in \Omega$. We say that the vector $\mathbf{d} \in \mathbb{R}^n$ is a **feasible descent direction** at $\mathbf{x} \in \Omega$ if there exists a $\delta > 0$ such that

$$\mathbf{x} + \alpha \mathbf{d} \in \Omega \quad \text{and} \quad f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x}) \quad \forall \alpha \in (0, \delta].$$

Thus if a descent direction exists at a point $\mathbf{x}$, then it is possible to move a short distance along this direction to a feasible point with a better objective value.

If

$$\nabla f(\mathbf{x})\mathbf{d} < 0,$$

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

Theorem 2.2 says that if $\mathbf{x}^*$ is a relative minimum point of f over Ω, then every feasible direction at $\mathbf{x}^*$ is not descent.

Corollary 2.3 (Unconstrained case)

Let Ω be a subset of $\mathbb{R}^n$, and let $f \in C^1$ be a function on Ω. If $\mathbf{x}^*$ is a relative extremum point of f over Ω and if $\mathbf{x}^*$ is an interior point of Ω, then

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

An interior point $\mathbf{x}^* \in \Omega$ is called

(a) a **critical point** (or **stationary point**) of f if $\nabla f(\mathbf{x}^*) = \mathbf{0}$;

(b) a **saddle point** of f if it is a critical point and for any $r > 0$ there exist points $\mathbf{y}, \mathbf{z}$ in $B(\mathbf{x}^*, r)$ such that $f(\mathbf{y}) < f(\mathbf{x}^*) < f(\mathbf{z})$.

Note that the equation $\nabla f(\mathbf{x}^*) = \mathbf{0}$ is equivalent to n equations

$$\frac{\partial f}{\partial x_1} = 0, \quad \frac{\partial f}{\partial x_2} = 0, \ldots, \frac{\partial f}{\partial x_n} = 0$$

in n unknowns (the components of $\mathbf{x}^*$).

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \iff \frac{\partial f}{\partial x_i} = 0, \quad i = 1, 2, \ldots, n.$$

Example 2.2 Consider the problem

$$\begin{aligned} \text{minimize} \quad & f(x_1, x_2) = (x_1 + 1)^2 + (x_2 - 1)^2 \\ \text{subject to} \quad & x_1 \geq 0, \ \ x_2 \geq 0. \end{aligned}$$

Show that f satisfies the first-order necessary conditions (Theorem 2.2) at the point $(0, 1)$.

Necessary Conditions for a Relative Minimum

Theorem 3.1 (Second-order necessary conditions)

Let Ω be a subset of $\mathbb{R}^n$ and let $f \in C^2$ be a function on Ω. If $\mathbf{x}^*$ is a relative minimum point of f over Ω, then for any $\mathbf{d} \in \mathbb{R}^n$ that is a feasible direction at $\mathbf{x}^*$ we have

(i) $\nabla f(\mathbf{x}^*)\mathbf{d} \geq 0$;

(ii) *if* $\nabla f(\mathbf{x}^*)\mathbf{d} = 0$, *then* $\mathbf{d}^T \mathbf{F}(\mathbf{x}^*)\mathbf{d} \geq 0$.

Example 3.1 Consider the problem

$$\begin{aligned} \text{minimize} \quad & f(x_1, x_2) = x_1^2 - x_1 + x_2 + x_1 x_2 \\ \text{subject to} \quad & x_1 \geq 0, \ x_2 \geq 0. \end{aligned}$$

Show that conditions of Theorem 3.1 are satisfied at the point $(\frac{1}{2}, 0)$.

Corollary 3.2 (Second-order necessary conditions, unconstrained case)

Let $\mathbf{x}^*$ *be an interior point of the set* Ω*, and suppose* $\mathbf{x}^*$ *is a relative minimum point over* Ω *of the function* $f \in C^2$*. Then*

(i) $\nabla f(\mathbf{x}^*) = \mathbf{0}$

(ii) $\mathbf{d}^T \mathbf{F}(\mathbf{x}^*)\mathbf{d} \geq 0$ *for all* $\mathbf{d}$.

Condition (ii) is equivalent to stating that

The Hessian matrix $\mathbf{F}(\mathbf{x}^*)$ is positive semidefinite.

Sufficient Conditions for a Relative Minimum

Theorem 3.3 (Second-order sufficient conditions, unconstrained case)

Let $f \in C^2$ be a function defined on a region in which the point $\mathbf{x}^*$ is an interior point. Suppose in addition that

(i) $\nabla f(\mathbf{x}^*) = \mathbf{0}$

(ii) $\mathbf{F}(\mathbf{x}^*)$ *is positive definite.*

Then $\mathbf{x}^*$ is a strict relative minimum point of f.

Note Sufficient conditions for a strict relative maximum point are similar.

Example 3.2 Consider the problem

$$f(x_1, x_2) = \frac{1}{3}x_1^3 + \frac{1}{2}x_1^2 + 2x_1x_2 + \frac{1}{2}x_2^2 - x_2 + 9.$$

ANS. $(2, -3)$: local minimizer.
$(1, -1)$: neither minimizer nor maximizer.

Remark

Let f be twice differentiable on an open set $\Omega \subset \mathbb{R}^n$. If $\mathbf{x}^* \in \Omega$ is a critical point and $\mathbf{F}(x^*)$ is indefinite, then $\mathbf{x}^*$ is a saddle point of f on Ω.

Example 3.3 Consider the problem

$$\begin{aligned} \text{minimize} \quad & f(x,y) = x^2 + y^2 + \beta xy + x + 2y, \\ \text{subject to} \quad & (x,y) \in \mathbb{R}^2. \end{aligned}$$

Remark. Note that if $\mathbf{x}^*$ is a critical point of $f(\mathbf{x})$ and $\mathbf{F}(\mathbf{x}^*)$ is merely positive semidefinite, then nothing can be concluded in general.

In order to develop a theory directed toward characterizing global, rather than local, minimum points, it is necessary to introduce some sort of convexity assumptions.

- Convex functions occur frequently and naturally in many optimization problems that arise in practice.
- Convexity considerations often make it unnecessary to test the Hessians of functions for positive definiteness, a test which can be difficult in practice.

Definitions and Combinations of Convex Functions

Definition 4.1

A function f defined on a convex set Ω is said to be **convex** if, for every $\mathbf{x}_0, \mathbf{x}_1 \in \Omega$ and every $\lambda \in [0, 1]$, there holds

$$f\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1\big) \leq (1-\lambda)f(\mathbf{x}_0) + \lambda f(\mathbf{x}_1).$$

If, for every $\lambda, 0 < \lambda < 1$ and $\mathbf{x}_0 \neq \mathbf{x}_1$, there holds

$$f\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1\big) < (1-\lambda)f(\mathbf{x}_0) + \lambda f(\mathbf{x}_1)$$

then f is said to be **strictly convex**.

Definition 4.2

A function g defined on a convex set Ω is said to be **concave** if the function $f = -g$ is convex. The function g is **strictly concave** if $-g$ is strictly convex.

Let f be defined on Ω. The set

$$E := \left\{(r, \mathbf{x}) \in \mathbb{R}^{n+1} : f(\mathbf{x}) \leq r\right\}$$

is called the **epigraph** of f. It is easy to verify that the set E is convex if and only if f is a convex function.

Theorem 4.1

Let f_1 and f_2 be convex functions on the convex set Ω. Then the function $f_1 + f_2$ is convex on Ω.

Theorem 4.2

Let f be a convex function over the convex set Ω. Then the function cf is convex for any $c \geq 0$.

Theorem 4.3

Let f be a convex function on a convex set Ω. The set $\Gamma_\alpha = \{\mathbf{x} : \mathbf{x} \in \Omega,\ f(\mathbf{x}) \leq \alpha\}$ is convex for every real number α.

We note that, since the intersection of convex sets is also convex, the set of points simultaneously satisfying

$$f_1(\mathbf{x}) \leq \alpha_1$$
$$f_2(\mathbf{x}) \leq \alpha_2$$
$$\vdots$$
$$f_k(\mathbf{x}) \leq \alpha_k$$

where each f_i is a convex function, defines a convex set.

Question: Under what conditions is a function convex?

Properties of Differentiable Convex Functions

Theorem 4.4

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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \Omega.$$

Theorem 4.5

Let f be twice differentiable on an open convex set $\Omega \subset \mathbb{R}^n$. Then

(a) *f is convex on Ω if and only if the Hessian matrix $\mathbf{F}(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in \Omega$.*

(b) *If $\mathbf{F}(\mathbf{x})$ is positive definite for all $\mathbf{x} \in \Omega$, then f is strictly convex on Ω.*

Minimization and Maximization of Convex Functions

Theorem 4.6

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

(i) *the set Γ where f achieves its minimum is convex, and any relative minimum of f is a global minimum;*

(ii) *if, in addition, f is strictly convex over Ω, then there exists at most one minimum point.*

Theorem 4.7

Let $f \in C^1$ be convex on the convex set Ω. If there is a point $\mathbf{x}^* \in \Omega$ such that

$$\nabla f(\mathbf{x}^*)(\mathbf{y} - \mathbf{x}^*) \geq 0 \quad \textit{for all} \quad \mathbf{y} \in \Omega,$$

then $\mathbf{x}^*$ is a global minimum point of f over Ω.
In particular, any stationary point $\mathbf{x}^* \in \Omega$ is a global minimizer of f.

Example 4.1 Determine a global minimum point of the function

$$f(x_1, x_2, x_3) = 3x_1^2 + 2x_1x_2 + x_2^2 + x_2x_3 + 2x_3^2 - 8x_1 - 6x_2 - x_3 + 12$$

on $\mathbb{R}^3$.

ANS. Global minimum point

$$\mathbf{x}^* = (5/13, 37/13, -6/13).$$

Example 4.2 Let

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x} + \alpha,$$

where $\mathbf{Q}$ is an $n \times n$ symmetric matrix, $\mathbf{b} \in \mathbb{R}^n$, and $\alpha \in \mathbb{R}$.

(a) If $\mathbf{Q}$ is positive semidefinite, then f is convex and if $\mathbf{Q}$ is positive definite, then f is strictly convex.

(b) If $\mathbf{Q}$ is positive semidefinite, then optimal solutions of

$$\text{minimize} \quad f(\mathbf{x})$$

are solutions of

$$\mathbf{Q}\mathbf{x} = \mathbf{b}.$$

Example 4.3 There are many applications in which some linear system $\mathbf{Ax} = \mathbf{b}$ of m equations in n unknowns should be consistent but fails to be so because of measurement errors in the entries of $\mathbf{A}$ or $\mathbf{b}$.

In such cases one looks for vectors that come as close as possible to being solutions in the sense that they minimize $|\mathbf{Ax} - \mathbf{b}|$.

We thus have to solve the problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}) = |\mathbf{Ax} - \mathbf{b}|^2. \tag{2}$$

This is called a **least-squares problem**.

The quadratic objective function is given by

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{A}\mathbf{x} + |\mathbf{b}|^2.$$

f is convex because its Hessian

$$\mathbf{F}(\mathbf{x}) = 2\mathbf{A}^T\mathbf{A} \text{ is positive semidefinite for all } \mathbf{x}.$$

Suppose that $m \geq n$ and $\mathbf{A}$ has a **full column rank**; that is, $\text{rank}(\mathbf{A}) = n$. Then $\mathbf{A}^T\mathbf{A}$ is positive definite. In this case the unique global minimizer $\mathbf{x}^*$ of Problem (2) satisfies the equation

$$(\mathbf{A}^T\mathbf{A})\mathbf{x}^* = \mathbf{A}^T\mathbf{b}.$$

(See Example 4.2).

If $\mathbf{x}^*$ is a relative minimizer of f and $\mathbf{x}^*$ is an interior point of Ω and $f \in C^1$, then

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

So in principle the optimal solution of the problem can be obtained by finding among all the stationary points of f the one with the minimal function value.

In the majority of problems, however, such an approach is not implementable because it might be a very difficult task to solve the set of (usually nonlinear) equations $\nabla f(\mathbf{x}) = \mathbf{0}$.

Thus instead of trying to find an analytic solution to the stationarity condition, we will consider an iterative algorithm for finding stationary points.

There are many algorithms for solving optimization problems because

- optimization problems can come in so many forms,
- even for particular problems there are also many different algorithms that one could use.

Despite this diversity of both algorithms and problems, all of the algorithms that we will discuss in this course will have the same general form.

Beginning at $\mathbf{x}_0$, optimization algorithms generate a sequence of iterates $\{\mathbf{x}_k\}_{k=0}^{\infty}$ that terminate when either no more progress can be made or when it seems that a solution point has been approximated with sufficient accuracy.

In deciding how to move from one iterate $\mathbf{x}_k$ to the next, the algorithms use information about the function f at $\mathbf{x}_k$ and earlier iterates $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ to find a new iterate $\mathbf{x}_{k+1}$ with a lower function value than $\mathbf{x}_k$.

General Optimization Algorithm

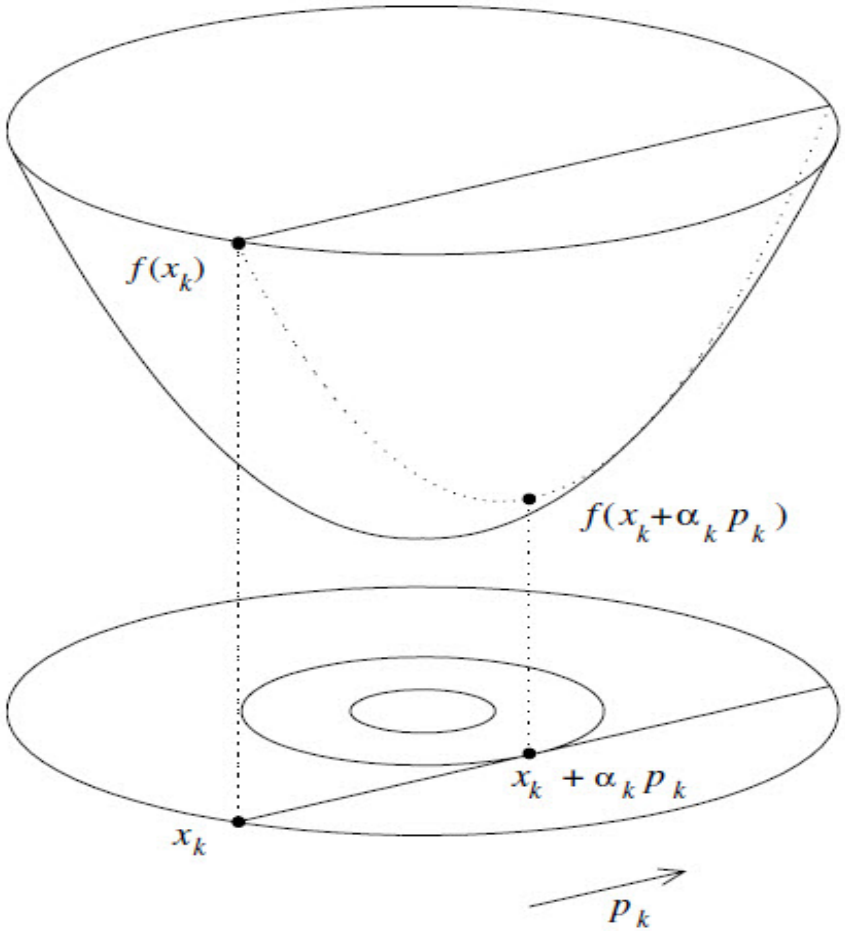
1. Specify some initial guess of the starting point $\mathbf{x}_0$.
2. For $k = 0, 1, \ldots$
   (i) If $\mathbf{x}_k$ is optimal, stop.
   (ii) Determine a search direction $\mathbf{d}_k$.
   (iii) Determine a step length α_k that leads to an improved estimate of the solution:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k.$$

In this algorithm,

- $\mathbf{d}_k$ is a **search direction** that we hope points in the general direction of the solution, or that "improves" our solution in some sense.
- The scalar α_k is a **step length** (or **size**) that determines the point $\mathbf{x}_{k+1}$; once the search direction $\mathbf{d}_k$ has been computed, the step length α_k is found by solving some auxiliary one-dimensional problem.

The process of finding the step size α_k is called a **line search** because it corresponds to a search along the line $\mathbf{x}_k + \alpha \mathbf{d}_k$ defined by α.



Line search

The General Optimization Algorithm with its three major steps (the optimality test, computation of $\mathbf{d}_k$, and computation of α_k) has been the basis for a great many of the most successful optimization algorithms ever developed.

Many details are missing in the above description of the General Optimization Algorithm:

- What is the starting point?
- How to choose the descent direction?
- What step length should be taken?
- What is the stopping criteria?

- The main difference between different methods is the choice of the descent direction.
- An example of a popular stopping criteria is

$$|\nabla f(\mathbf{x}_{k+1})| < \epsilon.$$

There are many choices for step length selection rules. Here are two popular choices:

- **constant step size** $\alpha_k = \bar{\alpha}$ for any k.
- **exact line search** α_k is a minimizer of f along the ray $\mathbf{x}_k + \alpha \mathbf{d}_k$:

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = \min_{\alpha \geq 0} f(\mathbf{x}_k + \alpha \mathbf{d}_k). \qquad (3)$$

However an exact minimization may be expensive and is usually unnecessary.

- Instead, the line search algorithm generates a limited number of trial step lengths until it finds one that loosely approximates the minimum of (3).
- At the new point, a new search direction and step length are computed, and the process is repeated.

The Method of Steepest Descent

Recall that a vector $\mathbf{d} \in \mathbb{R}^n$ is a *descent direction* of f at $\mathbf{x} \in \Omega$ if there exists a $\delta > 0$ such that

$$\mathbf{x} + \alpha\mathbf{d} \in \Omega \quad \text{and} \quad f(\mathbf{x} + \alpha\mathbf{d}) < f(\mathbf{x}) \quad \forall \alpha \in (0, \delta].$$

- If $\nabla f(\mathbf{x})\mathbf{d} < 0$, then $\mathbf{d}$ is a descent direction of f at x.
- If $\mathbf{F}(\mathbf{x})$ is positive definite and $\nabla f(\mathbf{x})^T \neq \mathbf{0}$, then

$$\mathbf{d} = -\mathbf{F}(\mathbf{x})\nabla f(\mathbf{x})^T$$

is a descent direction of f at x.

Steepest descent method is extremely important from a theoretical viewpoint, since it is one of the simplest for which a satisfactory analysis exists.

More advanced algorithms are often motivated by an attempt to modify the basic steepest descent technique in such a way that the new algorithm will have superior convergence properties.

The method of steepest descent remains, therefore, not only the technique most often first tried on a new problem but also the standard of reference against which other techniques are measured.

This method can be justified by the following geometrical argument.

If we want to minimize a function $f(\mathbf{x})$ and if our current trial point is $\mathbf{x}_k$ then we can expect to find better points by moving away from $\mathbf{x}_k$ along the direction which *causes f to decrease most rapidly*.

This direction of steepest descent is given by the *negative gradient*: the steepest descent method is a line search method that moves along

$$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)^T$$

at every step.

The Idea of the Method of Steepest Descent

At each stage of the iteration, the method searches for the next point by minimizing the function in the direction of the negative gradient at the current point.

The Method

Let f have continuous first partial derivatives on $\mathbb{R}^n$.

The gradient $\nabla f(\mathbf{x})$ is defined as an n-dimensional row vector.

For convenience we define the n-dimensional column vector $\mathbf{g}(\mathbf{x}) := \nabla f(\mathbf{x})^T$.

We sometimes also write $\mathbf{g}_k$ for

$$\mathbf{g}(\mathbf{x}_k) = \nabla f(\mathbf{x}_k)^T.$$

Note that $\mathbf{d} := -\mathbf{g}(\mathbf{x}) = -\nabla f(\mathbf{x})^T$ is the *steepest descent direction* of f at $\mathbf{x}$ whenever $\nabla f(\mathbf{x})^T \neq \mathbf{0}$.

The method of steepest descent is defined by the iterative algorithm

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k}$$

where α_k is a nonnegative scalar minimizing $f(\mathbf{x}_k - \alpha \mathbf{g}_k)$,

$$f(\mathbf{x}_k - \alpha_k \mathbf{g}_k) = \min_{\alpha \geq 0} f(\mathbf{x}_k - \alpha \mathbf{g}_k).$$

In words, from the point $\mathbf{x}_k$ we search along the direction of the negative gradient $-\mathbf{g}_k$ to a minimum point on this ray; this minimum point is taken to be $\mathbf{x}_{k+1}$.

Algorithm *(The Steepest Descent Method)*

Step 0. Let $0 < \varepsilon \ll 1$ be the termination tolerance. Given an initial point $\mathbf{x}_0 \in \mathbb{R}^n$. Set $k = 0$.

Step 1. If $|\mathbf{g}_k| \leq \varepsilon$, stop ; otherwise go to Step 2.

Step 2. Find the step length factor α_k, such that

$$f(\mathbf{x}_k - \alpha_k \mathbf{g}_k) = \min_{\alpha \geq 0} f(\mathbf{x}_k - \alpha \mathbf{g}_k);$$

Step 3. Compute $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$.

Step 4. Set $k := k + 1$, return to Step 1.

The Quadratic Case

Essentially all of the important local convergence characteristics of the method of steepest descent are revealed by an investigation of the method when applied to quadratic problems.

Consider

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x},$$

where $\mathbf{Q}$ is a positive definite symmetric $n \times n$ matrix.

With $\mathbf{Q}$ is positive definite, it follows that all of its eigenvalues are positive.

We assume that these eigenvalues are ordered

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Since $\mathbf{Q}$ is positive definite, f is *strictly convex*.

The unique minimum point of f can be found directly, by setting the gradient to zero, as the vector $\mathbf{x}^*$ satisfying

$$\boxed{\mathbf{Qx} = \mathbf{b}}$$

Thus the method of steepest descent can be expressed as

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k} \tag{4}$$

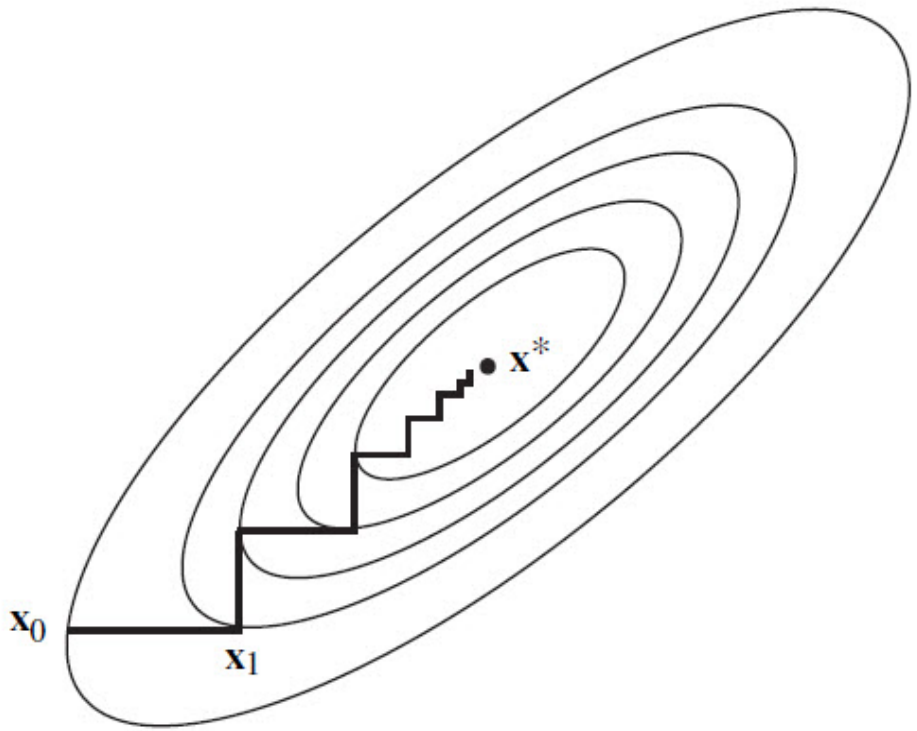
where

$$\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$$

and where α_k minimizes $f(\mathbf{x}_k - \alpha \mathbf{g}_k)$.

Hence the method of steepest descent (4) takes the explicit form

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k} \right) \mathbf{g}_k} \tag{5}$$



Steepest descent

Example 6.1 Compute the first three iterates $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ in the steepest descent sequence $\{\mathbf{x}_k\}$ for

$$f(x_1, x_2) = (x_1 - 1)^2 + 2(x_2 - 2)^2$$

starting with the initial point $\mathbf{x}_0 = (0, 3)$.

ANS.

k	$\mathbf{x}_k$	$f(\mathbf{x}_k)$	$\mathbf{g}_k$	α_k
0	$\begin{bmatrix} 0 \\ 3 \end{bmatrix}$	3	$\begin{bmatrix} -2 \\ 4 \end{bmatrix}$	$\frac{5}{18}$
1	$\frac{1}{9}\begin{bmatrix} 5 \\ 17 \end{bmatrix}$	$\frac{2}{9}$	$\frac{1}{9}\begin{bmatrix} -8 \\ -4 \end{bmatrix}$	$\frac{5}{12}$
2	$\frac{1}{27}\begin{bmatrix} 25 \\ 56 \end{bmatrix}$	$\frac{4}{3^5}$	$\frac{1}{27}\begin{bmatrix} -4 \\ 8 \end{bmatrix}$	$\frac{5}{18}$
3	$\frac{1}{243}\begin{bmatrix} 235 \\ 484 \end{bmatrix}$	$\frac{8}{3^8}$	$\frac{1}{235}\begin{bmatrix} -16 \\ -8 \end{bmatrix}$	$\frac{5}{12}$

Introduce the **error function**

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

Let λ and Λ be, respectively, the smallest and largest eigenvalues of $\mathbf{Q}$.

Theorem 6.1 (Steepest descent; quadratic case)

For any $\mathbf{x}_0 \in \mathbb{R}^n$ the method of steepest descent (5) *converges to the unique minimum point $\mathbf{x}^*$ of f. Furthermore, with*

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

there holds at every step k,

$$E(\mathbf{x}_{k+1}) \leq \left(\frac{\Lambda - \lambda}{\Lambda + \lambda}\right)^2 E(\mathbf{x}_k).$$

Example 6.2 Let us take

$$\mathbf{Q} = \begin{bmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{bmatrix}$$

$$\mathbf{b} = (0.76, 0.08, 1.12, 0.68).$$

For this matrix it can be calculated that $\lambda = 0.52$, $\Lambda = 0.94$ and hence $r := \Lambda/\lambda = 1.8$.

Step k	$f(\mathbf{x}_k)$
0	0
1	-2.1563625
2	-2.1744062
3	-2.1746440
4	-2.1746585
5	-2.1746595
6	-2.1746595

Solution point $\mathbf{x}^* = (1.534965, 0.1220097, 1.975156, 1.412954)$

In this section we present Newton's method in its most basic or "classical" form.

Newton's method: one-dimensional case

Newton's Method is a procedure for finding numerical approximations to zeros of functions.

Newton's method corresponds to approximating a function g by its tangent line at the point x_k. The point where the tangent line crosses the x-axis (i.e., a zero of the tangent line) is taken as the new estimate of the solution x^* of the equation $g(x) = 0$.

If g is a differentiable function, then the $(k+1)^{st}$ estimate x_{k+1} for a zero of g is obtained from the kth estimate x_k by the formula

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)},$$

provided that $g'(x_k) \neq 0$.

A minimum point x^* of a differentiable function f satisfies the necessary condition $f'(x^*) = 0$.

Now we determine a minimum point of the function f, applying the above procedure to $g := f'$. We thus have the following formula

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

One can easily verify that if $f''(x_k) > 0$, then x_{k+1} is the minimum point of the Taylor approximation

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

of the function f around the point x_k.

Newton's method: *n*-dimensional case

Newton's method to minimize one-dimensional functions can be generalized to solve the unconstrained problem of functions of multiple variables.

The idea behind Newton's method is that the function f being minimized is approximated locally by a quadratic function, and this approximate function is minimized exactly.

Near $\mathbf{x}_k$ we can approximate f by a Taylor polynomial of order 2:

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T \mathbf{F}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k).$$

The right-hand side is minimized at

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\mathbf{F}(\mathbf{x}_k)\right]^{-1} \nabla f(\mathbf{x}_k)^T}$$

and this equation is the **pure form** of Newton's method.

Algorithm

Step 1. Choose $\mathbf{x}_0 \in \mathbb{R}^n$ and set $k = 0$.

Step 2. Calculate $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^T$.

Step 3. If $\mathbf{g}_k = \mathbf{0}$ (or $|\mathbf{g}_k| \leq \epsilon$), stop. Otherwise, go to Step 4.

Step 4. Set $\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\mathbf{F}(\mathbf{x}_k)\right]^{-1} \nabla f(\mathbf{x}_k)^T$, $k = k + 1$ and go to Step 2.

In practice,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k,$$

where $\mathbf{d}_k$ solves the equation

$$\mathbf{F}(\mathbf{x}_k)\mathbf{d} = -\mathbf{g}_k.$$

Example 7.1 We apply the method to the problem

$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^2} \quad f(x_1, x_2) = x_1^2 + e^{x_2} - x_2$$

with $\mathbf{x}_0 = (2, 1)$.

ANS.

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ e^{x_2} - 1 \end{bmatrix}^T, \quad \mathbf{x}_{k+1} = \begin{bmatrix} 0 \\ x_{k,2} - 1 + e^{-x_{k,2}} \end{bmatrix}$$

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ e^{-1} \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ e^{-1} - 1 + e^{e^{-1}} \end{bmatrix} \approx \begin{bmatrix} 0 \\ 0.0601 \end{bmatrix}.$$

Order of Convergence

Let the sequence $\{r_k\}$ converge to r^*. The **order of convergence** of $\{r_k\}$ is defined as the supremum of the nonnegative numbers p satisfying

$$0 \leq \limsup_{k \to \infty} \frac{|r_{k+1} - r^*|}{|r_k - r^*|^p} < \infty.$$

Example 7.2

(a) The sequence with $r_k = a^k$ where $0 < a < 1$ converges to zero with order unity, since $r_{k+1}/r_k = a$.

(b) The sequence with $r_k = a^{(2^k)}$ for $0 < a < 1$ converges to zero with order two, since $r_{k+1}/r_k^2 = 1$.

In Newton's method, the procedure converges quadratically to the solution $\mathbf{x}^*$ of the system $\nabla f(\mathbf{x}) = \mathbf{0}$, if the starting point $\mathbf{x}_0$ is "close enough" to $\mathbf{x}^*$.

Theorem 7.1

Let $f \in C^3$, and assume that at the local minimum point $\mathbf{x}^*$, the Hessian $\mathbf{F}(\mathbf{x}^*)$ is positive definite. Then if started sufficiently close to $\mathbf{x}^*$, the points generated by Newton's method converge to $\mathbf{x}^*$. The order of convergence is at least two.

Conjugate direction methods can be regarded as being somewhat intermediate between the method of steepest descent and Newton's method.

They are motivated by the desire to accelerate the typically slow convergence associated with steepest descent while avoiding the information requirements associated with the evaluation, storage, and inversion of the Hessian as required by Newton's method.

In this section we consider the quadratic problem

$$\text{minimize} \quad \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{6}$$

where $\mathbf{Q}$ is an $n \times n$ symmetric positive definite matrix.

The techniques works out for this problem can be extended, by approximation, to more general problems because in a neighborhood of a minimum point, general functions can often be well approximated by quadratic functions.

Conjugate direction methods, especially the method of conjugate gradients, have proved to be extremely effective in dealing with general objective functions and are considered among the best general purpose methods.

Conjugate Directions

Definition 8.1

Given a symmetric matrix $\mathbf{Q}$, two vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be **Q-orthogonal**, or **conjugate with respect to** $\mathbf{Q}$, if $\mathbf{u}^T\mathbf{Q}\mathbf{v} = 0$. A finite set of vectors $\mathbf{d}_0, \mathbf{d}_1, ..., \mathbf{d}_k$ is said to be a **Q-orthogonal set** if $\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_j = 0$ for all $i \neq j$.

If $\mathbf{Q} = \mathbf{O}$, any two vectors are conjugate, while if $\mathbf{Q} = \mathbf{I}$, conjugacy is equivalent to the usual notion of orthogonality.

Theorem 8.1

If $\mathbf{Q}$ *is positive definite and the set of nonzero vectors* $\mathbf{d}_0, \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k$ *are* $\mathbf{Q}$*-orthogonal, then these vectors are linearly independent.*

Corresponding to the $n \times n$ positive definite matrix $\mathbf{Q}$ let $\mathbf{d}_0, \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{n-1}$ be n nonzero $\mathbf{Q}$-orthogonal vectors. The unique solution $\mathbf{x}^*$ to the quadratic problem (6) is also the unique solution to the linear equation

$$\mathbf{Qx} = \mathbf{b}.$$

Theorem 8.2 (Conjugate Direction Theorem)

Let $\{\mathbf{d}_i\}_{i=0}^{n-1}$ *be a set of nonzero* $\mathbf{Q}$*-orthogonal vectors. For any* $\mathbf{x}_0 \in \mathbb{R}^n$ *the sequence* $\{\mathbf{x}_k\}$ *generated according to*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k, \quad k \geq 0 \tag{7}$$

with

$$\alpha_k = -\frac{\mathbf{g}_k^T \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k} \quad \text{and} \quad \mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}, \tag{8}$$

converges to the unique solution $\mathbf{x}^*$ *of* $\mathbf{Q}\mathbf{x} = \mathbf{b}$ *after* n *steps, that is,* $\mathbf{x}_n = \mathbf{x}^*$.

A procedure of the type (7)–(8) is called a **conjugate direction method**.

Example 8.1 Applying the procedure (7)–(8) to problem

$$\text{minimize} \quad \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x},$$

where

$$\mathbf{Q} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 11 \\ 13 \end{bmatrix},$$

$\mathbf{d}_0 = (2, 1)$, $\mathbf{d}_1 = (3, -3)$, and $\mathbf{x}_0 = (2, 4)$.

ANS.

$$\mathbf{g}_0 = \begin{bmatrix} -1 & 9 \end{bmatrix}^T, \quad \alpha_0 = -1/3, \quad \mathbf{x}_1 = (4/3, 11/3)$$

$$\mathbf{g}_1 = \begin{bmatrix} -10/3 & 20/3 \end{bmatrix}^T, \ \alpha_1 = 5/9, \ \mathbf{x}_2 = (3, 2) = \mathbf{x}^*.$$

Question:

How to generate a sequence of conjugate directions if a set of such vectors is not available?

Algorithm *(Conjugate gradient method)*

Step 1. Choose $\mathbf{x}_0 \in \mathbb{R}^n$, calculate $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)^T$.

Step 2. If $\mathbf{g}_0 = \mathbf{0}$ (or $|\mathbf{g}_0| \leq \epsilon$), stop; $\mathbf{x}_0$ is the minimum point of f. Otherwise set $\mathbf{d}_0 = -\mathbf{g}_0$, $k = 0$ and go to Step 3.

Step 3. Determine $\alpha_k = -\frac{\mathbf{g}_k^T \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k}$

Step 4. Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$.

Step 5. Calculate $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})^T$.

Step 6. If $\mathbf{g}_{k+1} = \mathbf{0}$ (or $|\mathbf{g}_{k+1}| \leq \epsilon$), stop. Otherwise go to Step 7.

Step 7. If $k < n-1$, go to Step 8; if $k = n-1$, go to Step 10.

Step 8. Calculate $\beta_k = \mathbf{g}_{k+1}^T \mathbf{g}_{k+1} / \mathbf{g}_k^T \mathbf{g}_k$.

Step 9. Set $\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k$, $k = k+1$ and go to Step 3.

Step 10. Set $\mathbf{x}_0 = \mathbf{x}_n$, $\mathbf{d}_0 = -\mathbf{g}_n$, $k = 0$ and go to Step 3.

Example 8.2 We apply the algorithm to the problem

$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^2} \quad (x_1 - 1)^2 + 2(x_2 - 2)^2$$

choosing $\mathbf{x}_0 = (0, 3)$.

ANS. The calculations are summarized as follows:

k	$\mathbf{x}_k$	$\mathbf{g}_k$	$\mathbf{d}_k$	α_k	β_k
0	$(0, 3)$	$(-2, 4)$	$(2, -4)$	$5/18$	$4/81$
1	$\left(\frac{5}{9}, \frac{17}{9}\right)$	$\left(-\frac{8}{9}, -\frac{4}{9}\right)$	$\left(\frac{80}{81}, \frac{20}{81}\right)$	$9/20$	
2	$(1, 2)$				

Quasi-Newton methods are currently among the most widely used Newton-type methods for problems of moderate size, where matrices can be stored.

The method is based on Newton's method but use a different formula to compute the search direction.

The idea underlying many quasi-Newton methods is:

If the Hessian is cumbersome or too expensive or time-consuming to compute, it (or its inverse) is approximated by a suitable (easy to compute) matrix.

Modified Newton Method

A very basic iterative process for solving the problem

$$\text{minimize} \quad f(\mathbf{x})$$

which includes as special cases most of our earlier ones is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{S}_k \nabla f(\mathbf{x}_k)^T \tag{9}$$

where $\mathbf{S}_k$ is a symmetric $n \times n$ matrix and α_k is chosen to minimize

$$\phi(\alpha) := f(\mathbf{x}_k + \alpha \mathbf{d}_k), \quad \alpha > 0, \ \mathbf{d}_k = -\mathbf{S}_k \nabla f(\mathbf{x}_k)^T.$$

If $\mathbf{S}_k$ is the inverse of the Hessian of f,

$$\mathbf{S}_k = [\mathbf{F}(\mathbf{x}_k)]^{-1},$$

we obtain Newton's method, while if

$$\mathbf{S}_k = \mathbf{I}$$

we have steepest descent.

It would seem to be a good idea, in general, to select $\mathbf{S}_k$ as an approximation to the inverse of the Hessian.

We assume that $\mathbf{S}_k$ is positive definite so that the process (9) is guaranteed to be a descent method for small values of α.

We consider the standard quadratic problem

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x},$$

where $\mathbf{Q}$ is symmetric and positive definite.

The algorithm becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{S}_k \mathbf{g}_k,$$

where

$$\begin{aligned} \mathbf{g}_k &= \mathbf{Q}\mathbf{x}_k - \mathbf{b} \\ \alpha_k &= \frac{\mathbf{g}_k^T \mathbf{S}_k \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{S}_k \mathbf{Q} \mathbf{S}_k \mathbf{g}_k}. \end{aligned}$$

Usually, a quasi-Newton method is a procedure of the form

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k,$$

where α_k is determined by exact line search; the search direction is calculated as $\mathbf{d}_k = -\mathbf{S}_k \mathbf{g}_k$, where $\mathbf{S}_k$ is an $n \times n$ symmetric and positive definite matrix which must be updated in each step.

The matrices $\mathbf{S}_k$ are constructed such that $\mathbf{S}_n = \mathbf{Q}^{-1}$ is satisfied when solving a quadratic problem. It is desirable that the construction of these matrices requires the lowest possible computational work and uses only the first derivatives of the objective function.

The fact that $\mathbf{S}_k$ is positive definite, ensures that $\mathbf{d}_k$ is a descent direction. Due to the requirement $\mathbf{S}_n = \mathbf{Q}^{-1}$, n steps of a quasi-Newton method correspond approximately to one step of the Newton method.