Optimization Algorithms

Yinyu Ye

Department of Management Science and Engineering

Stanford University

Stanford, CA 94305, U.S.A.

http://www.stanford.edu/~yyye

Introduction

Optimization algorithms tend to be iterative procedures.

Starting from a given point $\mathbf{x}^0$, they generate a sequence $\{\mathbf{x}^k\}$ of iterates (or trial solutions).

We study algorithms that produce iterates according to well determined rules–Deterministic Algorithm rather than some random selection process–Randomized Algorithm.

The rules to be followed and the procedures that can be applied depend to a large extent on the characteristics of the problem to be solved.

Classes algorithms

Depending on information of the problem used to create a new iterate:

(a) Zero-order algorithms;

(b) First-order algorithms;

(c) Second-order algorithms.

Finite versus convergent iterative methods. For some classes of optimization problems (e.g., linear and quadratic programming) there are algorithms that obtain a solution—or detect that the objective function is unbounded—in a finite number of iterations. For this reason, we call them finite algorithms.

Most algorithms encountered in nonlinear programming are not finite, but instead are convergent—or at least they are designed to be so. Their object is to generate a sequence of trial or approximate solutions that converge to a “solution.”

The meaning of “solution”

What is meant by a solution may differ from one algorithm to another.

In some cases, one seeks a local minimum; in some cases, one seeks a global minimum; in others, one seeks a stationary or KKT point of some sort as in the method of steepest descent discussed below.

In fact, there are several possibilities for defining what a solution is.

Once the definition is chosen, there must be a way of testing whether or not a point (trial solution) belongs to the set of solutions.

The general idea

One selects a starting point and generates a possibly infinite sequence of trial solutions each of which is specified by the algorithm.

The idea is to do this in such a way that the sequence of iterates generated by the algorithm converges to a point of the set of solutions of the problem.

Convergence to some other sort of point is undesirable—as is failure of the sequence to converge at all.

Convergent sequences of real vectors

Let $\{x_k\}$ be a sequence of real numbers. Then $\{x_k\}$ converges to 0 if and only if for all real numbers $\varepsilon > 0$ there exists a positive integer K such that

$$|x_k| < \varepsilon \quad \text{for all } k \geq K.$$

Let $\{\mathbf{x}^k\}$ be a sequence of real vectors. Then $\{\mathbf{x}^k\}$ converges to $\mathbf{x}^*$ if and only if $\{\|\mathbf{x}^k - \mathbf{x}^*\|\}$ converges to 0.

Types of convergence

If there exists a number $c \in [0, 1)$ and an integer K such that

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq c\|\mathbf{x}^k - \mathbf{x}^*\| \quad \text{for all } k \geq K,$$

then $\{\mathbf{x}^k\}$ converges q-linearly to $\mathbf{x}^*$. If there exists a number $c \in [0, 1)$ and an integer K such that $c\|\mathbf{x}^k - \mathbf{x}^*\| < 1$ and

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq c\|\mathbf{x}^k - \mathbf{x}^*\|^2 \quad \text{for all } k \geq K,$$

then $\{\mathbf{x}^k\}$ converges q-quadraticly to $\mathbf{x}^*$. If $\{c_k\}$ is a sequence of nonnegative reals converging to 0 and

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq c_k\|\mathbf{x}^k - \mathbf{x}^*\| \quad \text{for all } k \geq K,$$

then $\{\mathbf{x}^k\}$ converges q-superlinearly to $\mathbf{x}^*$.

Examples of convergence

The arithmetic convergence: $\left\{\frac{1}{k}\right\}$.

The q-linear convergence: $\left\{(\frac{1}{2})^k\right\}$.

The q-quadratic convergence: $\left\{(\frac{1}{2})^{2^k}\right\}$.

The q-superlinear convergence: $\left\{(\frac{1}{log(k+1)})^k\right\}$.

Unconstrained minimization of smooth functions

The meaning of the term smooth function differs from one author to another, but it is generally agreed that it means at least continuously differentiable.

A Generic Algorithm: Let f be a smooth function on R^n. We seek $\mathbf{x}^* \in R^n$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in R^n$.

It may be necessary to adopt a more modest goal than finding a global minimum.

We might just look for a local minimum or a stationary point of the problem, f.

Assume we can decide whether any given point is a “solution.”

Global Convergence Theorem and Descent Direction Methods

Theorem 1 *Let A be an "algorithmic mapping" defined over set X, and let sequence $\{\mathbf{x}^k\}$, starting from a given point $\mathbf{x}^0$, be generated from*

$$\mathbf{x}^{k+1} \in A(\mathbf{x}^k).$$

Let a solution set $S \subset X$ be given, and suppose

i) all points $\{\mathbf{x}^k\}$ are in a compact set;

ii) there is a continuous function $z(\mathbf{x})$ such that if $\mathbf{x} \notin S$, then $z(\mathbf{y}) < z(\mathbf{x})$ for all $\mathbf{y} \in A(\mathbf{x})$; otherwise, $z(\mathbf{y}) \leq z(\mathbf{x})$ for all $\mathbf{y} \in A(\mathbf{x})$;

iii) the mapping A is closed at points outside S.

Then, the limit of any convergent subsequences of $\{\mathbf{x}^k\}$ is a solution in S.

Descent Direction Methods: $z(\mathbf{x})$ is just the objective itself.

(A1) Test for convergence If the termination conditions are satisfied at $\mathbf{x}^k$, then it is taken (accepted) as a "solution." In practice, this may mean satisfying the desired conditions to within some tolerance. If so, stop. Otherwise, go to step (A2).

(A2) Compute a search direction, say $\mathbf{d}^k \neq \mathbf{0}$. This might be a direction in which the function value is known to decrease.

(A3) Compute a step length, say α_k such that

$$f(\mathbf{x}^k + \alpha_k \mathbf{d}^k) < f(\mathbf{x}^k).$$

This may necessitate a one-dimensional (or line) search.

(A4) Define the new iterate by setting

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k$$

and return to step (A1).

Zero-Order Algorithms: the "Simplex" Method

(1) : Start with a Simplex with $d+1$ corner points and their objective function values.

(2) Reflection: Compute other $d+1$ corner points each of them is an additional corner point of a reflection simplex. If a point is better than its counter point, then the reflection simplex is an improved simplex, and select the most improved simplex and go to Step1; otherwise go to Step 3.

(3) Contraction: Compute the $d+1$ edge-middle points and subdivide the simplex into smaller $d+1$ simplexes, and select the best smaller simplex and go to Step 1.

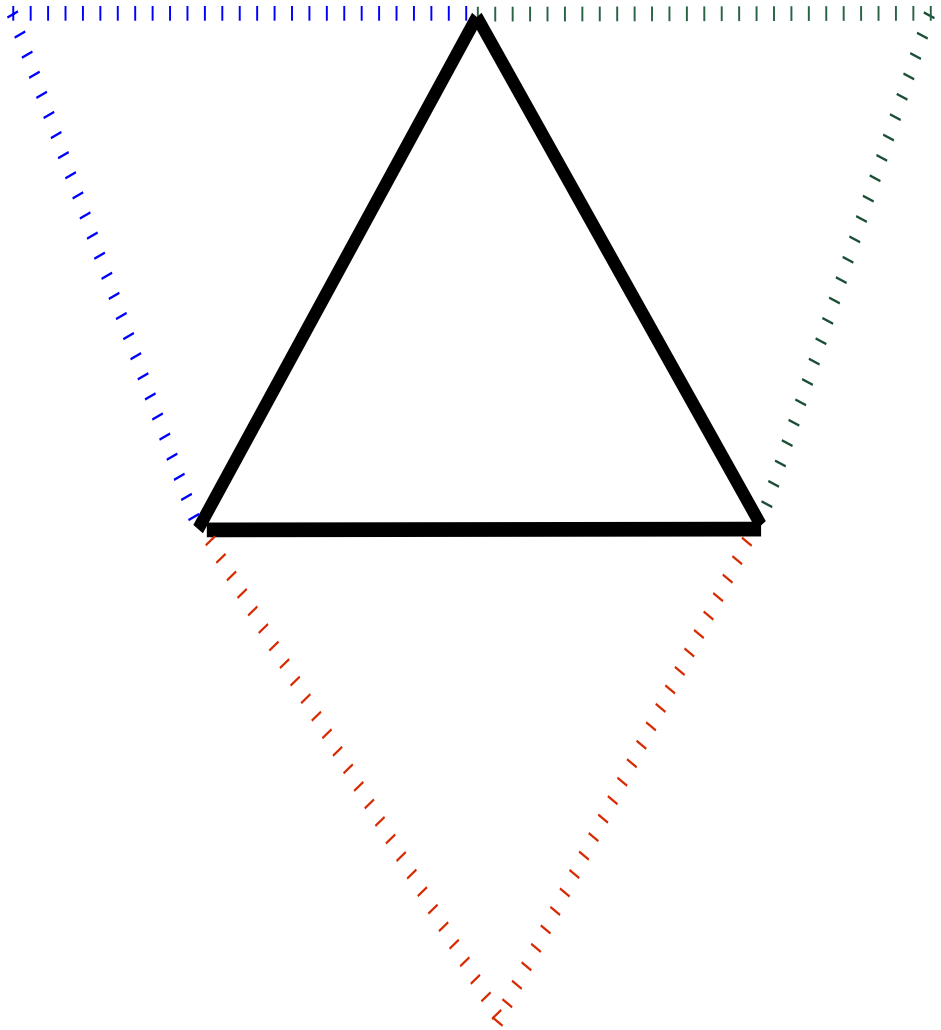

Figure 1: Reflection Simplexes

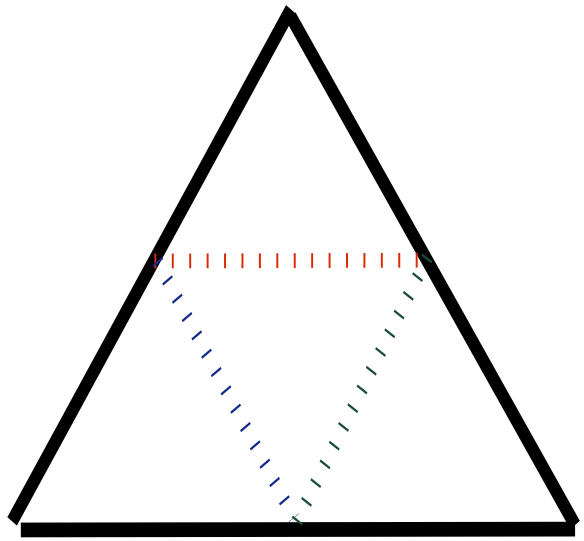

Figure 2: Contraction Simplexes

The Simplex Algorithm for Linear Programming

1. Initialization Start at a corner point of the feasible polyhedron .

2. Test for Optimality. Compute the $d + 1$ neighboring corner points in the feasible region and their objective values. If none of them make an improvement, stop and claim optimality for the current corner point; otherwise, select the best new corner point and go to Step 2.

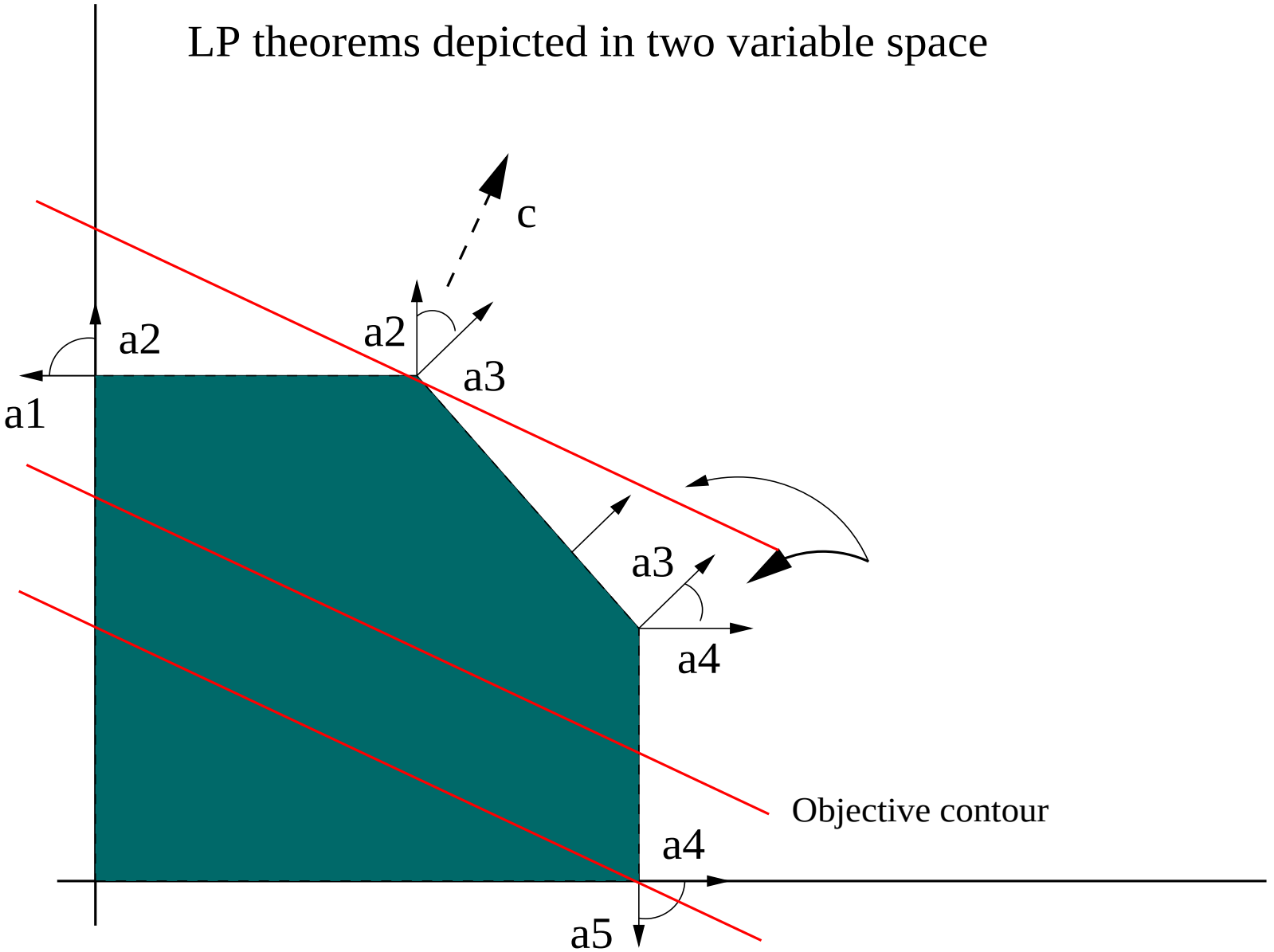

Figure 3: The LP Simplex Method

First-Order Algorithms: the Steepest Descent Method

Let f be a differentiable function and assume we can compute ∇f. We want to solve the unconstrained minimization problem

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

In the absence of further information, we seek a stationary point of f, that is, a point $\mathbf{x}^*$ at which $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Here we choose $\mathbf{p}^k = -\nabla f(\mathbf{x}^k)^T$ as the search direction at $\mathbf{x}^k$. The number $\alpha_k \geq 0$ is chosen "appropriately," namely to satisfy

$$\alpha_k \in \arg\min f(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)^T).$$

Then the new iterate is defined as $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k)^T$.

Now, if $\nabla f(\mathbf{x}^{k+1}) \neq \mathbf{0}$, then $-\nabla f(\mathbf{x}^{k+1})$ is a direction of descent at $\mathbf{x}^{k+1}$; in fact, it is the direction of steepest descent.

Example

Let $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T Q\mathbf{x}$ where $Q \in R^{n\times n}$ is symmetric and positive definite. This implies that the eigenvalues of Q are all positive. The unique minimum $\mathbf{x}^*$ of $f(x)$ exists and is given by the solution of the equation

$$\nabla f(\mathbf{x})^T = \mathbf{c} + Q\mathbf{x} = \mathbf{0},$$

or equivalently

$$Q\mathbf{x} = -\mathbf{c}.$$

The iterative scheme

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$$

becomes

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k(\mathbf{c} + Q\mathbf{x}^k)$$

by virtue of the definition, $\mathbf{p}^k = -(\mathbf{c} + Q\mathbf{x}^k)$.

To compute the step size, α_k, we consider

$$\begin{aligned} & f(\mathbf{x}^k + \alpha \mathbf{p}^k) \\ = \quad & \mathbf{c}^T(\mathbf{x}^k + \alpha \mathbf{p}^k) + \tfrac{1}{2}(\mathbf{x}^k + \alpha \mathbf{p}^k)^T Q(\mathbf{x}^k + \alpha \mathbf{p}^k) \\ = \quad & \mathbf{c}^T \mathbf{x}^k + \alpha \mathbf{c}^T \mathbf{p}^k + \tfrac{1}{2}(\mathbf{x}^k)^{\mathrm{T}} Q\mathbf{x}^k + \alpha(\mathbf{x}^k)^{\mathrm{T}} Q\mathbf{p}^k + \tfrac{1}{2}\alpha^2(\mathbf{p}^k)^{\mathrm{T}} Q\mathbf{p}^k \end{aligned}$$

which is a strictly convex quadratic function of α. Its minimizer α_k is the unique value of α where the derivative $f'(\mathbf{x}^k + \alpha \mathbf{p}^k)$ vanishes, i.e., where

$$\mathbf{c}^T \mathbf{p}^k + (\mathbf{x}^k)^{\mathrm{T}} Q\mathbf{p}^k + \alpha(\mathbf{p}^k)^{\mathrm{T}} Q\mathbf{p}^k = 0.$$

Thus

$$\alpha_k = -\frac{\mathbf{c}^T\mathbf{p}^k + (\mathbf{x}^k)^{\mathrm{T}}Q\mathbf{p}^k}{(\mathbf{p}^k)^{\mathrm{T}}Q\mathbf{p}^k} = \frac{\|\mathbf{p}^k\|^2}{(\mathbf{p}^k)^{\mathrm{T}}Q\mathbf{p}^k}.$$

The recursion for the method of steepest descent now becomes

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \left(\frac{\|\mathbf{p}^k\|^2}{(\mathbf{p}^k)^{\mathrm{T}}Q\mathbf{p}^k}\right)\mathbf{p}^k$$

where $\mathbf{p}^k = -(\mathbf{c} + Q\mathbf{x}^k)$.

Convergence of the steepest descent method

The following theorem gives some conditions under which the steepest descent method will converge.

Theorem. Let $f : R^n \to R$ be given. For some given point $x^0 \in R^n$, let the level set

$$X^0 = \{\mathbf{x} \in R^n : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$$

be bounded. Assume further that f is continuously differentiable on the convex hull of X^0. Let $\{\mathbf{x}^k\}$ be the sequence of points generated by the steepest descent method initiated at $\mathbf{x}^0$. Then every accumulation point of $\{\mathbf{x}^k\}$ is a stationary point of f.

Proof.

Note that the assumptions imply the compactness of X^0. Since the iterates will all belong to X^0, the existence of at least one accumulation point of $\{\mathbf{x}^k\}$ is guaranteed by the Bolzano-Weierstrass Theorem. Let $\bar{\mathbf{x}}$ be such an accumulation point, and without losing generality, $\{\mathbf{x}^k\}$ converge to $\bar{\mathbf{x}}$.

Assume $\nabla f(\bar{\mathbf{x}}) \neq 0$. Then there exists a value $\bar{\alpha} > 0$ and a $\delta > 0$ such that $f(\bar{\mathbf{x}} - \bar{\alpha}\nabla f(\bar{\mathbf{x}})^T) + \delta = f(\bar{\mathbf{x}})$. This means that $\bar{\mathbf{y}} := \bar{\mathbf{x}} - \bar{\alpha}\nabla f(\bar{\mathbf{x}})^T$ is an interior point of X^0 and

$$f(\bar{\mathbf{y}}) = f(\bar{\mathbf{x}}) - \delta.$$

For an arbitrary iterate of the sequence, say $\mathbf{x}^k$, the Mean-Value Theorem implies that we can write

$$f(\mathbf{x}^k - \bar{\alpha}\nabla f(\mathbf{x}^k)^T) = f(\bar{\mathbf{y}}) + (\nabla f(\mathbf{y}^k))^T \left(\mathbf{x}^k - \bar{\alpha}\nabla f(\mathbf{x}^k)^T - \bar{\mathbf{y}}\right)$$

where $\mathbf{y}^k$ lies between $\mathbf{x}^k - \bar{\alpha}\nabla f(\mathbf{x}^k)^T$ and $\bar{\mathbf{y}}$. Then $\{\mathbf{y}^k\} \to \bar{\mathbf{y}}$ and $\{\nabla f(\mathbf{y}^k)\} \to \nabla f(\bar{\mathbf{y}})$ as $\{\mathbf{x}^k\} \to \bar{\mathbf{x}}$. Thus, for sufficiently large k,

$$f(\mathbf{x}^k - \bar{\alpha}\nabla f(\mathbf{x}^k)^T) \leq f(\bar{\mathbf{y}}) + \frac{\delta}{2} = f(\bar{\mathbf{x}}) - \frac{\delta}{2}.$$

Since the sequence $\{f(\mathbf{x}^k)\}$ is monotonically decreasing and converges to $f(\bar{\mathbf{x}})$, hence

$$f(\bar{\mathbf{x}}) < f(\mathbf{x}^{k+1}) = f(\mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k)^T) \leq f(\mathbf{x}^k - \bar{\alpha}\nabla f(\mathbf{x}^k)^T) \leq f(\bar{\mathbf{x}}) - \frac{\delta}{2}$$

which is a contradiction. Hence $\nabla f(\bar{\mathbf{x}}) = 0$.

Remark.

According to this theorem, the steepest descent method initiated at any point of the level set X^0 will converge to a stationary point of f. In other words, it is not necessary to start the process in a neighborhood of the (unknown) solution. This property is called global convergence.

The convergence rate of the steepest descent method applied to quadratic functions is known to be linear. Suppose Q is a symmetric positive definite matrix of order n and let its eigenvalues be $0 < \lambda_1 \leq \cdots \leq \lambda_n$. Obviously, the global minimizer of the quadratic form $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x}$ is at the origin.

It can be shown that when the steepest descent method is started from any nonzero point $\mathbf{x}^0 \in R^n$, there will exist constants c_1 and c_2 such that

$$0 < c_1 \leq \frac{f(\mathbf{x}^{k+1})}{f(\mathbf{x}^k)} \leq c_2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 < 1, \; k = 0, 1, \ldots .$$

Intuitively, the slow rate of convergence of the steepest descent method can be attributed the fact that the successive search directions are perpendicular.

Consider an arbitrary iterate $\mathbf{x}^k$. At this point we have the search direction $\mathbf{p}^k = -\nabla f(\mathbf{x}^k)^T$. To find the next iterate $\mathbf{x}^{k+1}$ we minimize $f(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)^T$ with respect to $\alpha \geq 0$. At the minimum α_k, the derivative of this function will equal zero. Thus, we obtain $\nabla f(\mathbf{x}^{k+1}) \nabla f(\mathbf{x}^k)^T = 0.$