Zero-Order and First-Order Optimization Algorithms I

Yinyu Ye
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, U.S.A.

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

(Chapters 7 and 8)

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) that converge to a "solution" – or at least they are designed to be so.

Recall that scalars $\{x^k\}$ converges to 0 if and only if for all real numbers $\varepsilon > 0$ there exists a positive integer K such that

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

Then $\{\mathbf{x}^k\}$ converges to solution $\mathbf{x}^*$ if and only if $\{\|\mathbf{x}^k - \mathbf{x}^*\|\}$ converges to 0.

We study algorithms that produce iterates according to

- well determined rules–Deterministic Algorithm
- 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.

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 first-order and/or second-order 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 an iterate (trial solution) belongs to the set of solutions. For example, the residuals of the KKT conditions converge to zero.

Generic Algorithms for Minimization and Global Convergence Theorem

A Generic Algorithm: A point to set mapping in a subspace of R^n.

Theorem 1 *(Page 222, L&Y) 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 (merit) 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 ($\mathbf{x}^k \rightarrow \bar{\mathbf{x}} \in X$ and $A(\mathbf{x}^k) = \mathbf{y}^k \rightarrow \bar{\mathbf{y}}$ imply $\bar{\mathbf{y}} \in A(\bar{\mathbf{x}})$).

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

Descent Direction Methods

In this case, merit function $z(\mathbf{x}) = f(\mathbf{x})$, that 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 within the feasible region.

(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).

Algorithm Complexity and Speeds I

The intrinsic computational cost/time of an algorithm depends on

- number of decision variables n: cost of the inner product of two vectors, cost of solving system of linear equations
- number of constraints m: cost of the product of a matrix and a vector, cost of the product of two matrices
- number of nonzero data entries NNZ: sparse matrix/data representation
- the desired accuracy $0\epsilon < 1$: the cost could be propotional to $\frac{1}{\epsilon^2},\ \frac{1}{\epsilon},\ \log(\frac{1}{\epsilon}),\ \log[\log(\frac{1}{\epsilon})], ...$
- problem difficulty or complexity measures such as the Lipschiz constant β, the condition number of a matrix, etc

Algorithm Complexity and Speeds II

- Finite versus infinite convergence. For some classes of optimization problems there are algorithms that obtain an exact solution—or detect the unboundedness–in a finite number of iterations.

- Polynomial-time versus exponential-time. The solution time grows, in the worst-case, as a function of problem sizes (number of variables, constraints, accuracy, etc.).

- Convergence order and rate. If there is a positive number γ such that

$$\|\mathbf{x}^k - \mathbf{x}^*\| \leq \frac{O(1)}{k^\gamma}\|\mathbf{x}^0 - \mathbf{x}^*\|,$$

  then $\{\mathbf{x}^k\}$ converges arithmetically to $\mathbf{x}^*$ with power γ. If there exists a number $\gamma \in [0,1)$ such that

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq \gamma\|\mathbf{x}^k - \mathbf{x}^*\| \quad (\Rightarrow \|\mathbf{x}^k - \mathbf{x}^*\| \leq \gamma^k\|\mathbf{x}^0 - \mathbf{x}^*\|),$$

  then $\{\mathbf{x}^k\}$ converges geometrically or linearly to $\mathbf{x}^*$ with rate γ. If there exists a number $\gamma \in [0,1)$

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq \gamma\|\mathbf{x}^k - \mathbf{x}^*\|^2 \text{ after } \gamma\|\mathbf{x}^k - \mathbf{x}^*\| < 1$$

  then $\{\mathbf{x}^k\}$ converges quadratically to $\mathbf{x}^*$ (such as $\left\{(\frac{1}{2})^{2^k}\right\}$).

Algorithm Classes

Depending on information of the problem being used to create a new iterate, we have

(a) Zero-order algorithms. Popular when the gradient and Hessian information are difficult to obtain, e.g., no explicit function forms are given, functions are not differentiable, etc.

(b) First-order algorithms. Most popular now-days, suitable for large scale data optimization with low accuracy requirement, e.g., Machine Learning, Statistical Predictions...

(c) Second-order algorithms. Popular for optimization problems with high accuracy need, e.g., some scientific computing, etc.

One-Variable Optimization: Golden Section (Zero Order) Method

Assume that the one variable function $f(x)$ is Unimodel in interval $[a\ b]$, that is, for any point $x \in [a_r\ b_l]$ such that $a \le a_r < b_l \le b$, we have that $f(x) \le \max\{f(a_r),\ f(b_l)\}$. How do we find x^* within an error tolerance ϵ?

0) Initialization: let $x_l = a,\ x_r = b$, and choose a constant $0 < r < 0.5$;

1) Let two other points $\hat{x}_l = x_l + r(x_r - x_l)$ and $\hat{x}_r = x_l + (1-r)(x_r - x_l)$, and evaluate their function values.

2) Update the triple points $x_r = \hat{x}_r, \hat{x}_r = \hat{x}_l, x_l = x_l$ if $f(\hat{x}_l) < f(\hat{x}_r)$; otherwise update the triple points $x_l = \hat{x}_l, \hat{x}_l = \hat{x}_r, x_r = x_r$; and return to Step 1.

In either cases, the length of the new interval after one golden section step is $(1-r)$. If we set $(1-2r)/(1-r) = r$, then only one point is new in each step and needs to be evaluated. This give $r = 0.382$ and the linear convergence rate is 0.618.

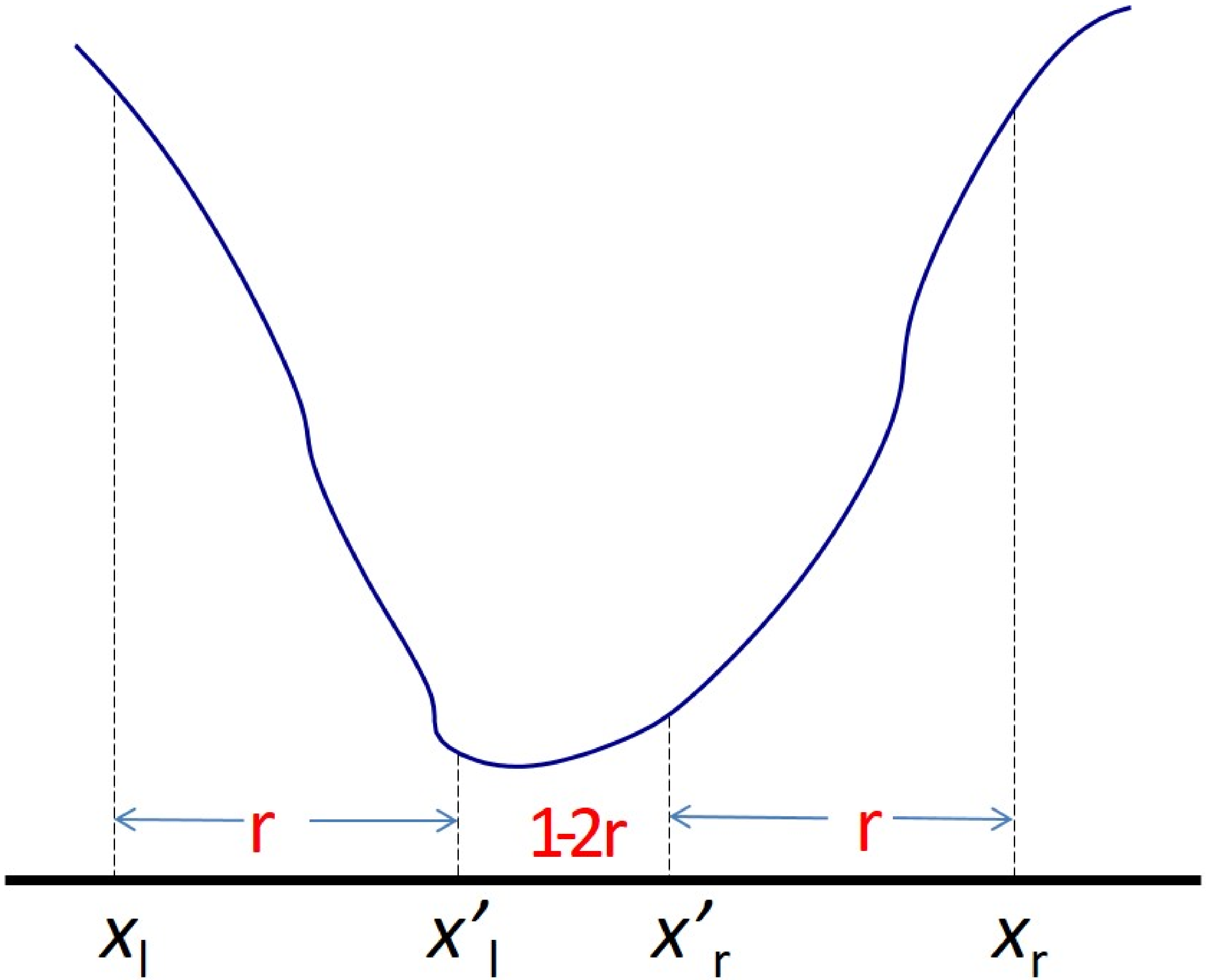

Figure 1: Illustration of Golden Section

One-Variable Optimization: Bisection (First Order) Method

For a one variable problem, an KKT point is the root of $g(x) := f'(x) = 0$.

Assume we know an interval $[a\ b]$ such that $a < b$, and $g(a)g(b) < 0$. Then we know there exists an x^*, $a < x^* < b$, such that $g(x^*) = 0$; that is, interval $[a\ b]$ contains a root of g. How do we find x within an error tolerance ϵ, that is, $|x - x^*| \leq \epsilon$?

0) Initialization: let $x_l = a,\ x_r = b$.

1) Let $x_m = (x_l + x_r)/2$, and evaluate $g(x_m)$.

2) If $g(x_m) = 0$ or $x_r - x_l < \epsilon$ stop and output $x^* = x_m$. Otherwise, if $g(x_l)g(x_m) > 0$ set $x_l = x_m$; else set $x_r = x_m$; and return to Step 1.

The length of the new interval containing a root after one bisection step is $1/2$ which gives the linear convergence rate is $1/2$, and this establishes a linear convergence rate 0.5.

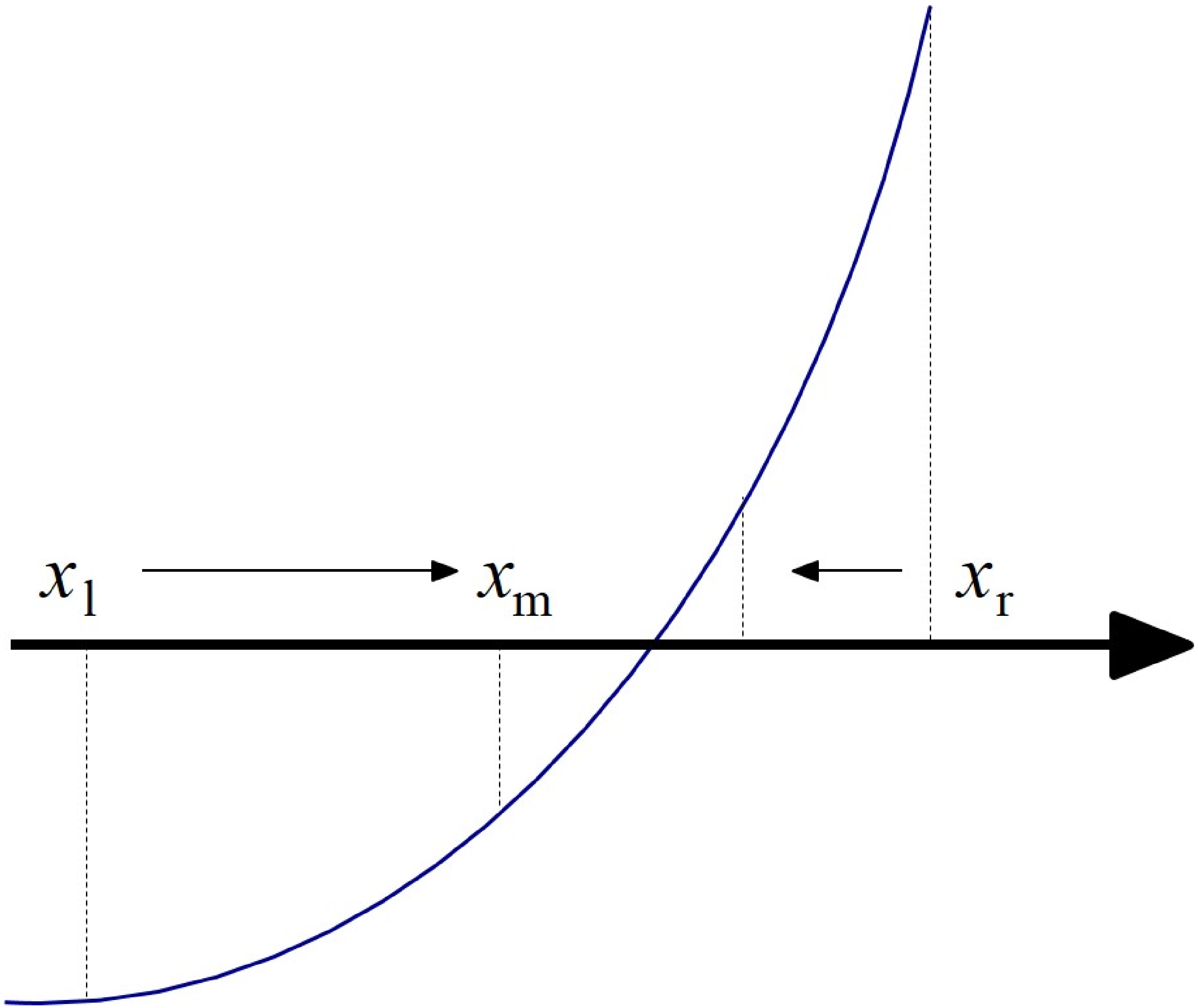

Figure 2: Illustration of Bisection

One-Variable Optimization: Newton's (Second Order) Method

For functions of a single real variable x, the KKT condition is $g(x) := f'(x) = 0$. When f is twice continuously differentiable then g is once continuously differentiable, Newton's method can be a very effective way to solve such equations and hence to locate a root of g. Given a starting point x^0, Newton's method for solving the equation $g(x) = 0$ is to generate the sequence of iterates

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

The iteration is well defined provided that $g'(x^k) \neq 0$ at each step.

For strictly convex function, Newton's method has a linear convergence rate and, when the point is "close" to the root, the convergence becomes quadratic, which leads to the iterations bound of $\log[\log(\frac{1}{\epsilon})]$.

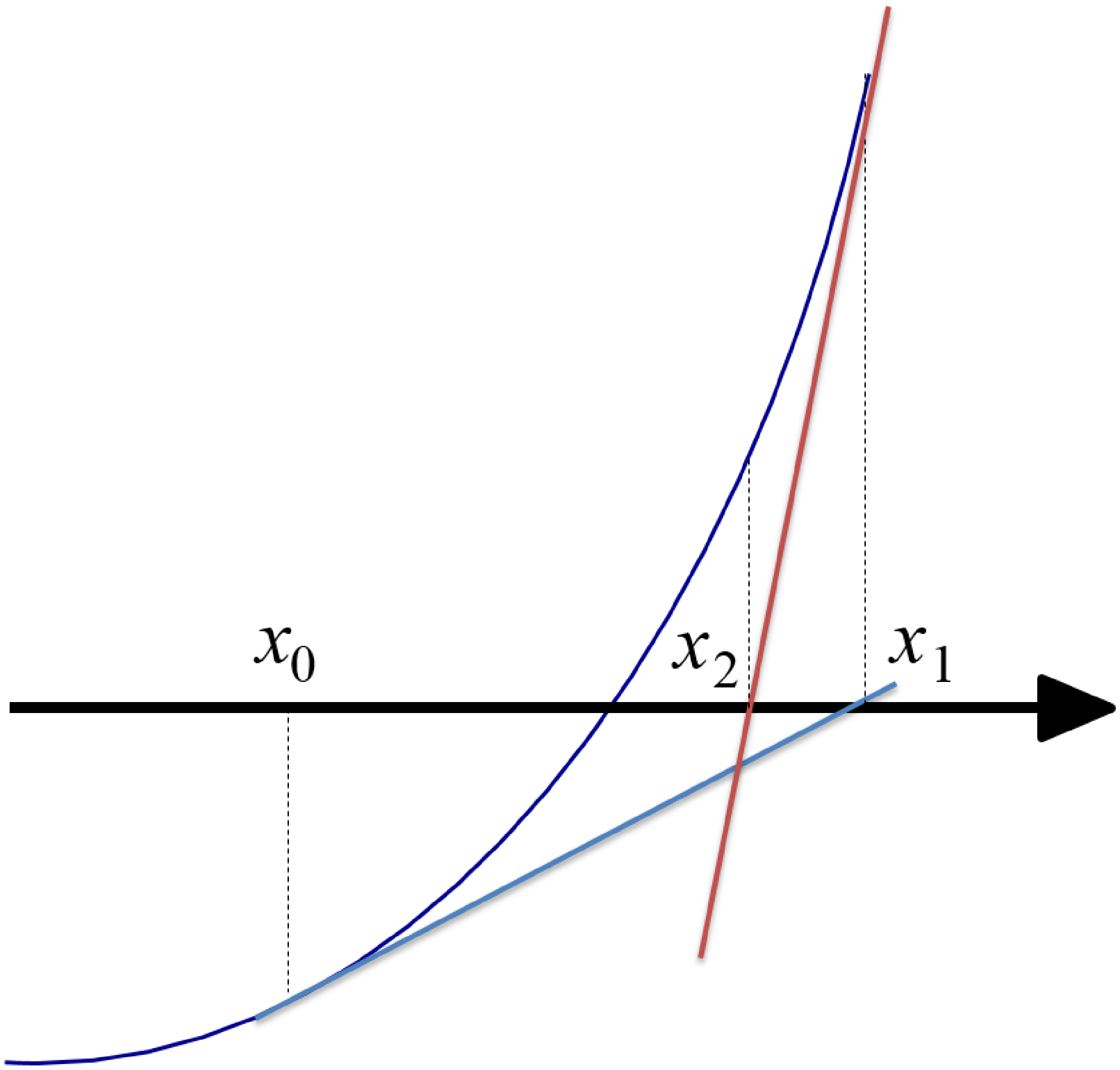

Figure 3: Illustration of Newton's Method

How Close is Close: One-variable Criterion

Theorem 2 *(Smale 86). Let $g(x)$ be an analytic function. Then, if x in the domain of g satisfies*

$$\sup_{k>1} \left| \frac{g^{(k)}(x)}{k! g'(x)} \right|^{1/(k-1)} \leq (1/8) \left| \frac{g'(x)}{g(x)} \right|.$$

Then, x is an approximate root of g.

In the following, for simplicity, let the root be in interval $[0 \;\; R]$.

Corollary 1 *(Y. 92). Let $g(x)$ be an analytic function in R^{++} and let g be convex and monotonically decreasing. Furthermore, for $x \in R^{++}$ and $k > 1$ let*

$$\left| \frac{g^{(k)}(x)}{k! g'(x)} \right|^{1/(k-1)} \leq \frac{\alpha}{8} \cdot \frac{1}{x}$$

for some constant $\alpha > 0$. Then, if the root $\bar{x} \in [\hat{x}, (1 + 1/\alpha)\hat{x}] \subset R^{++}$, $\hat{x}$ is an approximate root of g.

Hybrid of Bisection and Newton I

Note that the interval becomes wider and wider at geometric rate when $\hat{x}$ is increased.

Thus, we may symbolically construct a sequence of points:

$$\hat{x}_0 = \epsilon,\ \hat{x}_1 = (1+1/\alpha)\hat{x}_0, ..., \text{ and } \hat{x}_j = (1+1/\alpha)\hat{x}_{j-1}, ...$$

until $\hat{x}_j = \hat{x}_J \geq R$. Obviously the total number of points, J, of these points is bounded by $O(\log(R/\epsilon))$. Moreover, define a sequence of intervals

$$I_j = [\hat{x}_{j-1}, \hat{x}_j] = [\hat{x}_{j-1}, (1+1/\alpha)\hat{x}_{j-1}].$$

Then, if the root $\bar{x}$ of g is in any one of these intervals, say in I_j, then the front point $\hat{x}_{j-1}$ of the interval is an approximate root of g so that starting from it Newton's method generates an x with $|x - \bar{x}| \leq \epsilon$ in $O(\log\log(1/\epsilon))$ iterations.

Hybrid of Bisection and Newton II

Now the question is how to identify the interval that contains $\bar{x}$?

This time, we bisect the number of intervals, that is, evaluate function value at point $\hat{x}_{j_m}$ where $j_m = [J/2]$. Thus, each bisection reduces the total number of the intervals by a half. Since the total number of intervals is $O(\log(R/\epsilon))$, in at most $O(\log\log(R/\epsilon))$ bisection steps we shall locate the interval that contains $\bar{x}$.

Then the total number iterations, including both bisection and Newton methods, is $O(\log\log(R/\epsilon))$ iterations.

Here we take advantage of the global convergence property of Bisection and local quadratic convergence property of Newton, and we would see more of these features later...

Multi-Variable Optimization 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 $d+1$ middle-face points and subdivide the simplex into smaller $d+1$ simplexes, keep the simplex with the lowest sum of the $d+1$ function values, and go to Step 1.

This method can be also implemented with exhausted enumeration in parallel. The method is suitable for solving problems whose derivatives are difficult to compute.

How to generate the initial $d+1$ points?

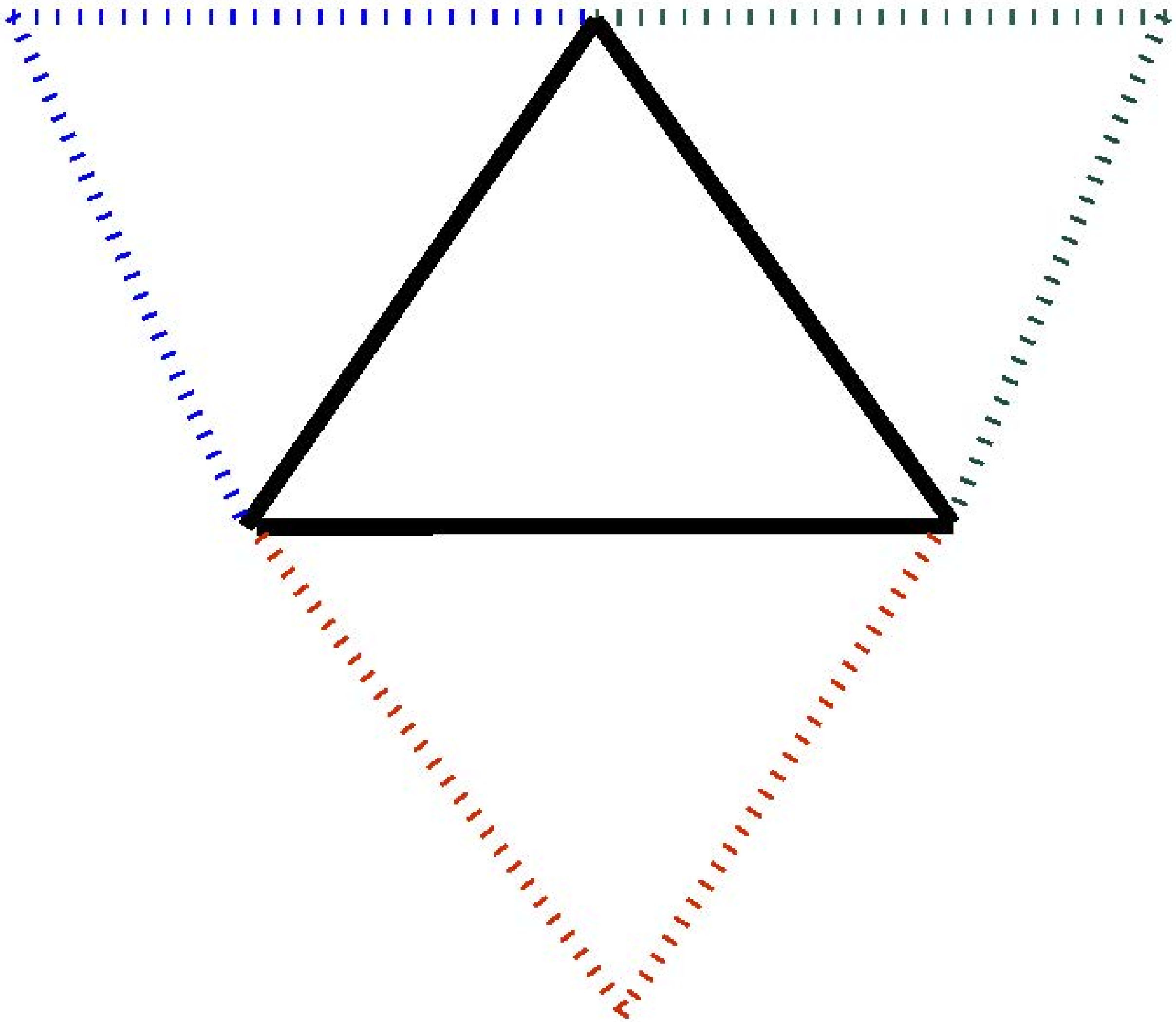

Figure 4: Reflection Simplexes

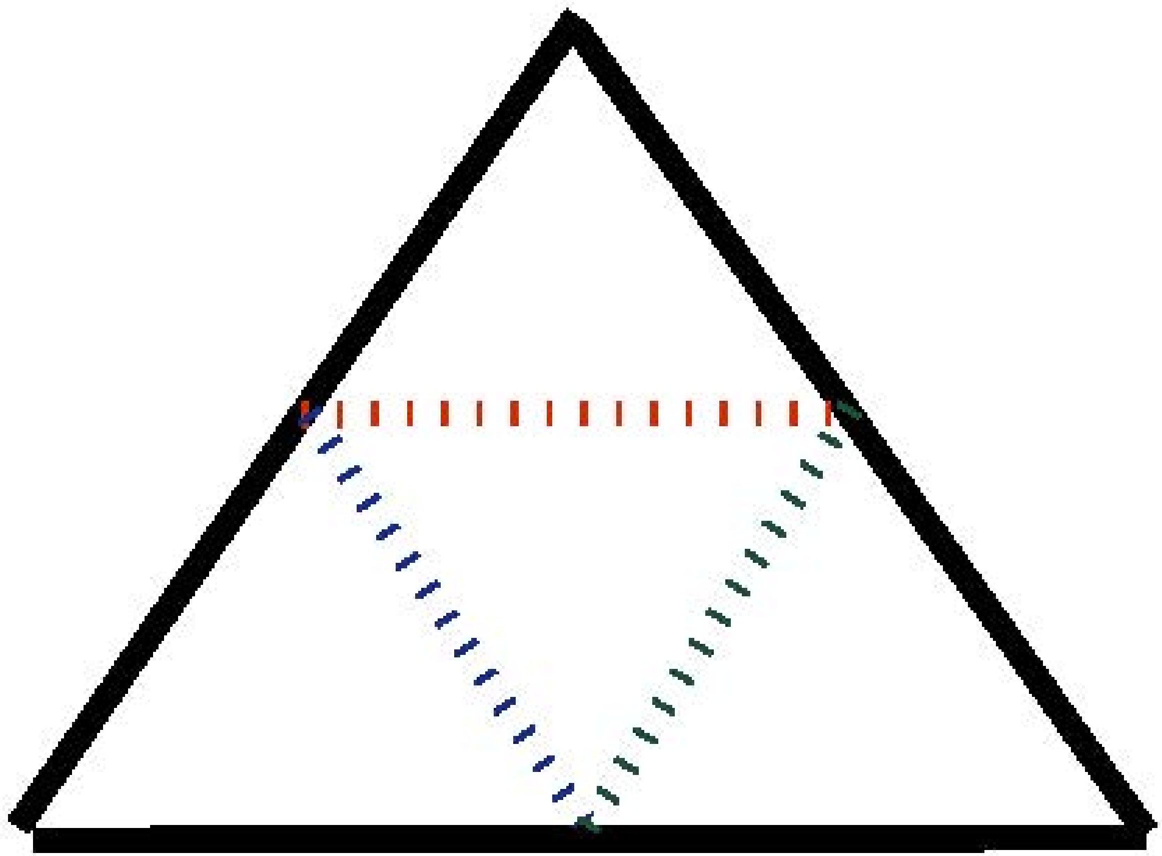

Figure 5: Contraction Simplexes

Multi-Variable Optimization Zero-Order Algorithms: the Finite-Difference Gradient

$$\nabla f(\mathbf{x}^k)_j \sim \frac{1}{\delta}\left(f(\mathbf{x}^k + \delta \mathbf{e}_j) - f(\mathbf{x}^k)\right) \ \forall j$$

for a small $\delta (> 0)$, and they can be estimated in parallel.

Check ZeroorderNLP.m and ZeroordersubNLP.m, which is modified from the derivative-free nonlinear optimization solver “SOLNP”. For more advanced one, see “SOLNP+”!

First-Order Algorithm: the Steepest Descent Method (SDM)

Let f be a differentiable function and assume we can compute gradient (column) vector ∇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 first-order KKT or stationary point of f, that is, a point $\mathbf{x}^*$ at which $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Here we choose direction vector $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$ as the search direction at $\mathbf{x}^k$, which is the direction of steepest descent.

The number $\alpha^k \geq 0$, called step-size, is chosen "appropriately" as

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

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

In some implementations, step-size α^k is fixed through out the process – independent of iteration count k

SDM Example: Unconstrained Quadratic Optimization

Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x}$ 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 $f(\mathbf{x})$ exists and is given by the solution of the system of linear equations

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

or equivalently

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

The iterative scheme becomes, from $\mathbf{d}^k = -(Q\mathbf{x}^k + \mathbf{c})$,

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

To compute the step size, α^k, we consider

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

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

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

Thus

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

The recursion for the method of steepest descent now becomes

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

Therefore, minimize a strictly convex quadratic function is equivalent to solve a system of equation with a positive definite matrix. The former may be ideal if the system only needs to be solved approximately.

Iterate Convergence of the Steepest Descent Method

The following theorem gives some conditions under which the steepest descent method will generate a sequence of iterates that converge .

Theorem 3 *Let $f: R^n \rightarrow R$ be given. For some given point $\mathbf{x}^0 \in R^n$, 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 $f(\bar{\mathbf{x}} - \bar{\alpha}\nabla f(\bar{\mathbf{x}})) + \delta = f(\bar{\mathbf{x}})$. This means that $\bar{\mathbf{y}} := \bar{\mathbf{x}} - \bar{\alpha}\nabla f(\bar{\mathbf{x}})$ 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

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

where $\mathbf{y}^k$ lies between $\mathbf{x}^k - \bar{\alpha}\nabla f(\mathbf{x}^k)$ 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,

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

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

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

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, which property is called global convergence.

This proof can be viewed as a special form of Theorem 1: the SDM algorithm mapping is closed and the objective function is strictly decreasing if not optimal yet.

Convergence Speed of the SDM for Strongly Convex QP

The convergence rate of the steepest descent method applied to convex quadratic functions is known to be linear. Suppose Q is a symmetric positive definite matrix of order n and let its eigenvalues be $0 < \lambda_1 \le \cdots \le \lambda_n$. 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 (page 235, L&Y)

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

where the ratio $\frac{\lambda_n}{\lambda_1}$ is called the condition number of the Hessian matrix.

Intuitively, the rate of linear 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 $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$. To find the next iterate $\mathbf{x}^{k+1}$ we minimize $f(\mathbf{x}^k - \alpha\nabla f(\mathbf{x}^k))$ with respect to $\alpha \ge 0$. At the minimum α^k, the derivative of this function will equal zero. Thus, we obtain $\nabla f(\mathbf{x}^{k+1})^T \nabla f(\mathbf{x}^k) = 0$.

Convergence Speed of the SDM for Minimizing Lipschitz Functions

Let $f(\mathbf{x})$ be differentiable every where and satisfy the (first-order) β-Lipschitz condition, that is, for any two points $\mathbf{x}$ and $\mathbf{y}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\| \tag{1}$$

for a positive real constant β. Then, we have

Lemma 1 *Let f be a β-Lipschitz function. Then for any two points $\mathbf{x}$ and $\mathbf{y}$*

$$f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{2}$$

At the kth step of SDM, we have

$$f(\mathbf{x}) - f(\mathbf{x}^k) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2.$$

The left hand strict convex quadratic function of $\mathbf{x}$ establishes a upper bound on the objective reduction.

Let us minimize the quadratic function

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \ \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2,$$

and let the minimizer be the next iterate. Then it has a close form:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{\beta}\nabla f(\mathbf{x}^k)$$

which is the SDM with the fixed step-size $\frac{1}{\beta}$. Then

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \le -\frac{1}{2\beta}\|\nabla f(\mathbf{x}^k)\|^2, \quad \text{or} \quad f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ge \frac{1}{2\beta}\|\nabla f(\mathbf{x}^k)\|^2.$$

Then, after $K(\ge 1)$ steps, we must have

$$f(\mathbf{x}^0) - f(\mathbf{x}^K) \ge \frac{1}{2\beta}\sum_{k=0}^{K-1}\|\nabla f(\mathbf{x}^k)\|^2. \tag{3}$$

Theorem 4 *(Error Convergence Estimate Theorem) Let the objective function* $p^* = \inf\ f(\mathbf{x})$ *be finite and let us stop the SDM as soon as* $\|\nabla f(\mathbf{x}^k)\| \le \epsilon$ *for a given tolerance* $\epsilon \in (0\ 1)$*. Then the SDM* *terminates in* $\frac{2\beta(f(\mathbf{x}^0)-p^*)}{\epsilon^2}$ *steps.*

Proof: From (3), after $K = \frac{2\beta(f(\mathbf{x}^0)-p^*)}{\epsilon^2}$ steps

$$f(\mathbf{x}^0) - p^* \geq f(\mathbf{x}^0) - f(\mathbf{x}^K) \geq \frac{1}{2\beta}\sum_{k=0}^{K-1} \|\nabla f(\mathbf{x}^k)\|^2.$$

If $\|\nabla f(\mathbf{x}^k)\| > \epsilon$ for all $k = 0, ..., K-1$, then we have

$$f(\mathbf{x}^0) - p^* > \frac{K}{2\beta}\epsilon^2 \geq f(\mathbf{x}^0) - p^*$$

which is a contradiction.

Corollary 2 *If a minimizer* $\mathbf{x}^*$ *of* f *is attainable, then the SDM terminates in* $\frac{\beta^2\|\mathbf{x}^0-\mathbf{x}^*\|^2}{\epsilon^2}$ *steps.*

The proof is based on Lemma 1 with $\mathbf{x} = \mathbf{x}^0$ and $\mathbf{y} = \mathbf{x}^*$ and noting $\nabla f(\mathbf{y}) = \nabla f(\mathbf{x}^*) = \mathbf{0}$:

$$f(\mathbf{x}^0) - p^* = f(\mathbf{x}^0) - f(\mathbf{x}^*) \leq \frac{\beta}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

The SDM for Unconstrained Convex Lipschitz Optimization

Here we consider $f(\mathbf{x})$ being convex and differentiable everywhere and satisfying the (first-order) β-Lipschitz condition. Given the knowledge β, we again adopt the fixed step-size rule:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{\beta}\nabla f(\mathbf{x}^k). \tag{4}$$

The following lemma is instrumental for establishing the global convergence rate of the Steepest Descent Method in this case.

Lemma 2 *It holds for all x and $y \in R^n$ that*

$$f(\mathbf{x}) - f(\mathbf{y}) - [\nabla f(\mathbf{x})]^T(\mathbf{x}-\mathbf{y}) \;\le\; -\frac{1}{2\beta}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2. \tag{5}$$

Proof: Fix an $\mathbf{x} \in R^n$. Define $F(\mathbf{y}) = f(\mathbf{y}) + [\nabla f(\mathbf{x})]^T(\mathbf{x}-\mathbf{y})$ for $\mathbf{y} \in R^n$. Then (5) is equivalent to $F(\mathbf{x}) - F(\mathbf{y}) \le -\|\nabla F(\mathbf{y})\|^2/(2\beta)$. This inequality holds because ∇F is β-Lipschitz and $F(\mathbf{x})$ is the global minimum of F, as F is convex and $\nabla F(\mathbf{x}) = 0$.

Theorem 5 *For convex Lipschitz optimization the Steepest Descent Method generates a sequence of solutions such that*

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \ \le \ \frac{\beta}{2(k+1)}\|\mathbf{x}^0 - \mathbf{x}^*\|^2, \tag{6}$$

$$\min_{0\le l\le k} \|\nabla f(\mathbf{x}^l)\| \ \le \ \frac{\sqrt{2}\beta}{\sqrt{(k+1)(k+2)}}\|\mathbf{x}^0 - \mathbf{x}^*\|, \tag{7}$$

where we assume that $\mathbf{x}^*$ *is a minimizer of the problem.*

Proof: According to Lemma 2, for the gradient method (4), we have

$$\begin{aligned} f(\mathbf{x}^k) - f(\mathbf{x}^*) &\le [\nabla f(\mathbf{x}^k)]^T(\mathbf{x}^k - \mathbf{x}^*) - \tfrac{1}{2\beta}\|\nabla f(\mathbf{x}^k)\|^2 \\ &= \beta(\mathbf{x}^k - \mathbf{x}^{k+1})^T(\mathbf{x}^k - \mathbf{x}^*) - \tfrac{\beta}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 \\ &= \tfrac{\beta}{2}(\mathbf{x}^k - \mathbf{x}^{k+1})^T(\mathbf{x}^k + \mathbf{x}^{k+1}) - 2\mathbf{x}^*) \\ &= \tfrac{\beta}{2}(\|\mathbf{x}^k - \mathbf{x}^*\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2). \end{aligned} \tag{8}$$

On the other hand, as we have proved for general Lipschitz optimization case,

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ \ge \ \frac{1}{2\beta}\|\nabla f(\mathbf{x}^k)\|^2. \tag{9}$$

Hence $\{f(\mathbf{x}^k)\}$ is nonincreasing. Consequently,

$$\sum_{l=0}^{k}\left[f(\mathbf{x}^l)-f(\mathbf{x}^*)\right] \;\geq\; (k+1)\left[f(\mathbf{x}^k)-f(\mathbf{x}^*)\right],$$

which renders (6) together with (8). Meanwhile, inequality (7) follows from (8) and

$$\begin{aligned}\sum_{l=0}^{k}[f(\mathbf{x}^l)-f(\mathbf{x}^*)] \quad &\geq \sum_{l=0}^{k}\sum_{i=l}^{k}[f(\mathbf{x}^i)-f(\mathbf{x}^{i+1})]\\ &\geq \frac{1}{4\beta}(k+2)(k+1)\min_{0\leq l\leq k}\|\nabla f(\mathbf{x}^l)\|^2,\end{aligned}$$

where the second inequality uses (9).

Remark When $k=0$, inequalities (6) and (7) reduce to

$$f(\mathbf{x}^0)-f(\mathbf{x}^*) \;\leq\; \frac{\beta}{2}\|\mathbf{x}^0-\mathbf{x}^*\|^2 \quad \text{and} \quad \|\nabla f(\mathbf{x}^0\| \;\leq\; \beta\|\mathbf{x}^0-\mathbf{x}^*\|,$$

which cannot be improved.

Forward and Backward Tracking Step-Size Method

In most real applications, the Lipschitz constant β is unknown. Furthermore, we like to use the smallest localized Lipschitz constant β^k at iteration k such that

$$f(\mathbf{x}^k + \alpha\mathbf{d}^k) - f(\mathbf{x}^k) - \nabla f(\mathbf{x}^k)^T(\alpha\mathbf{d}^k) \leq \frac{\beta^k}{2}\|\alpha\mathbf{d}^k\|^2,$$

where $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$, to decide the step-size $\alpha = \frac{1}{\beta^k}$.

Consider the following step-size strategy: stat at a good step-size guess $\alpha > 0$:

(1): If $\alpha \leq \frac{2(f(\mathbf{x}^k)-f(\mathbf{x}^k+\alpha\mathbf{d}^k))}{\|\mathbf{d}^k\|^2}$ then doubling the step-size: $\alpha \leftarrow 2\alpha$, stop as soon as the inequality is reversed and select the latest α with $\alpha \leq \frac{2(f(\mathbf{x}^k)-f(\mathbf{x}^k+\alpha\mathbf{d}^k))}{\|\mathbf{d}^k\|^2}$;

(2): Otherwise halving the step-size: $\alpha \leftarrow \alpha/2$; stop as soon as $\alpha \leq \frac{2(f(\mathbf{x}^k)-f(\mathbf{x}^k+\alpha\mathbf{d}^k))}{\|\mathbf{d}^k\|^2}$ and return it.

Prove that the selected step-size

$$\frac{1}{2\beta^k} \leq \alpha \leq \frac{1}{\beta^k}.$$

The Barzilai and Borwein Method

There is a steepest descent method (Barzilai and Borwein 88) that chooses the step-size as follows:

$$\Delta_x^k = \mathbf{x}^k - \mathbf{x}^{k-1} \quad \text{and} \quad \Delta_g^k = \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}), \tag{10}$$

$$\alpha^k = \frac{(\Delta_x^k)^T \Delta_g^k}{(\Delta_g^k)^T \Delta_g^k} \quad \text{or} \quad \alpha^k = \frac{(\Delta_x^k)^T \Delta_x^k}{(\Delta_x^k)^T \Delta_g^k},$$

Then

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \nabla f(\mathbf{x}^k). \tag{11}$$

For convex quadratic minimization with Hessian Q, $\Delta_g^k = Q\Delta_x^k$, the two step size formula become

$$\alpha^k = \frac{(\Delta_x^k)^T Q \Delta_x^k}{(\Delta_x^k)^T Q^2 \Delta_x^k} \quad \text{or} \quad \alpha^k = \frac{(\Delta_x^k)^T \Delta_x^k}{(\Delta_x^k)^T Q \Delta_x^k}$$

and it is between the reciprocals of the largest and smallest non-zero eigenvalues of Q (Rayleigh quotient).

An Explanation why the BB Method Works

For convex quadratic minimization, let the distinct nonzero eigenvalues of Hessian Q be λ_1, λ_2, ..., λ_K; and let the step size in the SDM be $\alpha^k = \frac{1}{\lambda_k}$, $k = 1, ..., K$. Then, the SDM terminates in K iterations from any starting point $\mathbf{x}^0$.

In the BB method, α^k minimizes

$$\|\Delta_x^k - \alpha\Delta_g^k\| = \|\Delta_x^k - \alpha Q\Delta_x^k\|.$$

If the error becomes 0 plus $\|\Delta_x^k\| \neq 0$, $\frac{1}{\alpha^k}$ will be a nonzero eigenvalue of Q – this is learning via Rayleigh quotient.

Another interpretation: one-dimensional Newton - (the second choice of) α^k minimizes the quadratic (approximate) objective function along the negative-gradient direction at step $k-1$.

On the other hand, many questions remain open for the BB method.