

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 $\{x^k\}$ **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 **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.

Generic Algorithms for Minimization and Global Convergence Theorem

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

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 (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 .

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

- **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 $\left\{ \left(\frac{1}{2}\right)^{2^k} \right\}$).

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.

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, and select the “best” smaller simplex and go to Step 1.

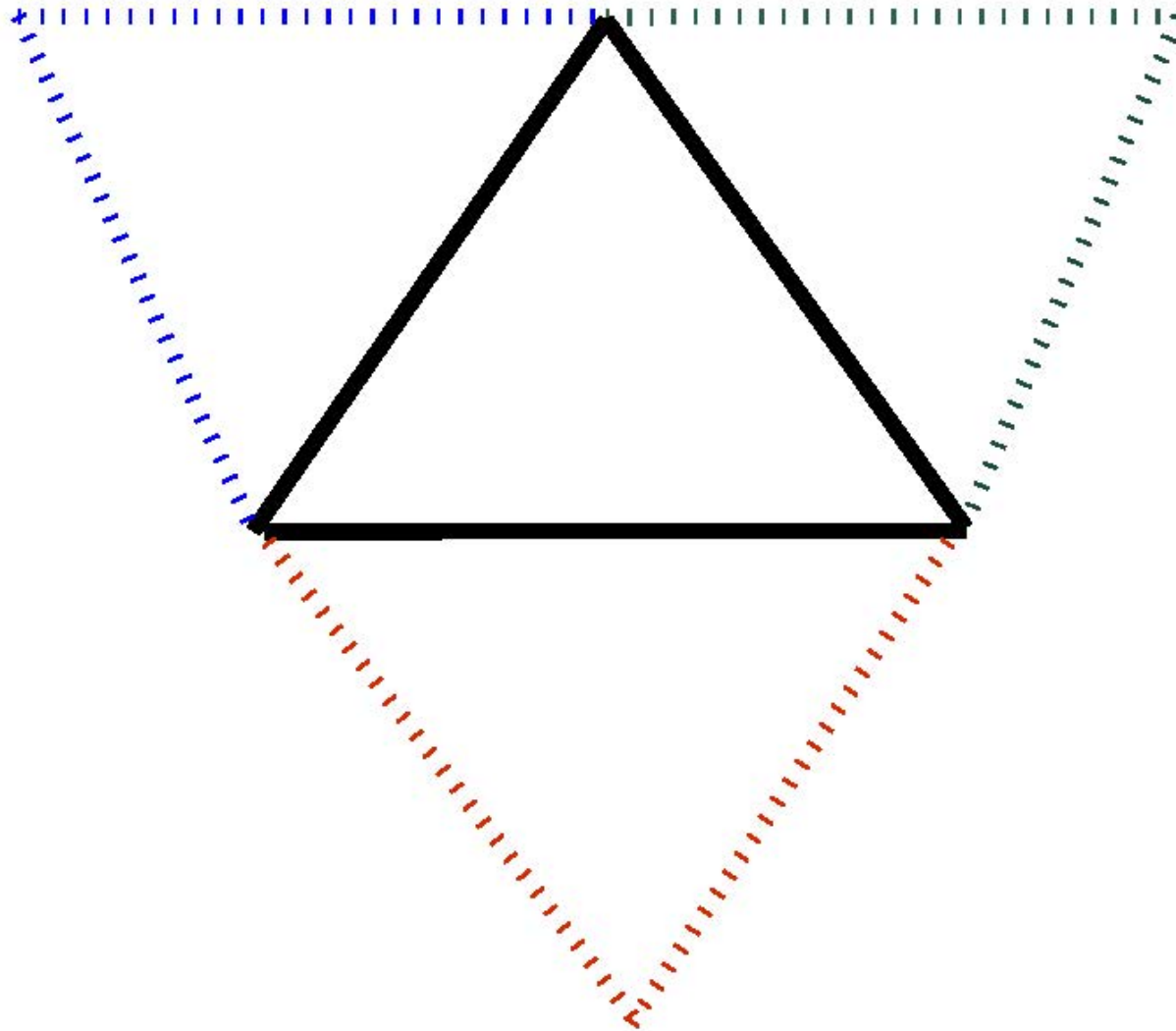


Figure 1: Reflection Simplexes

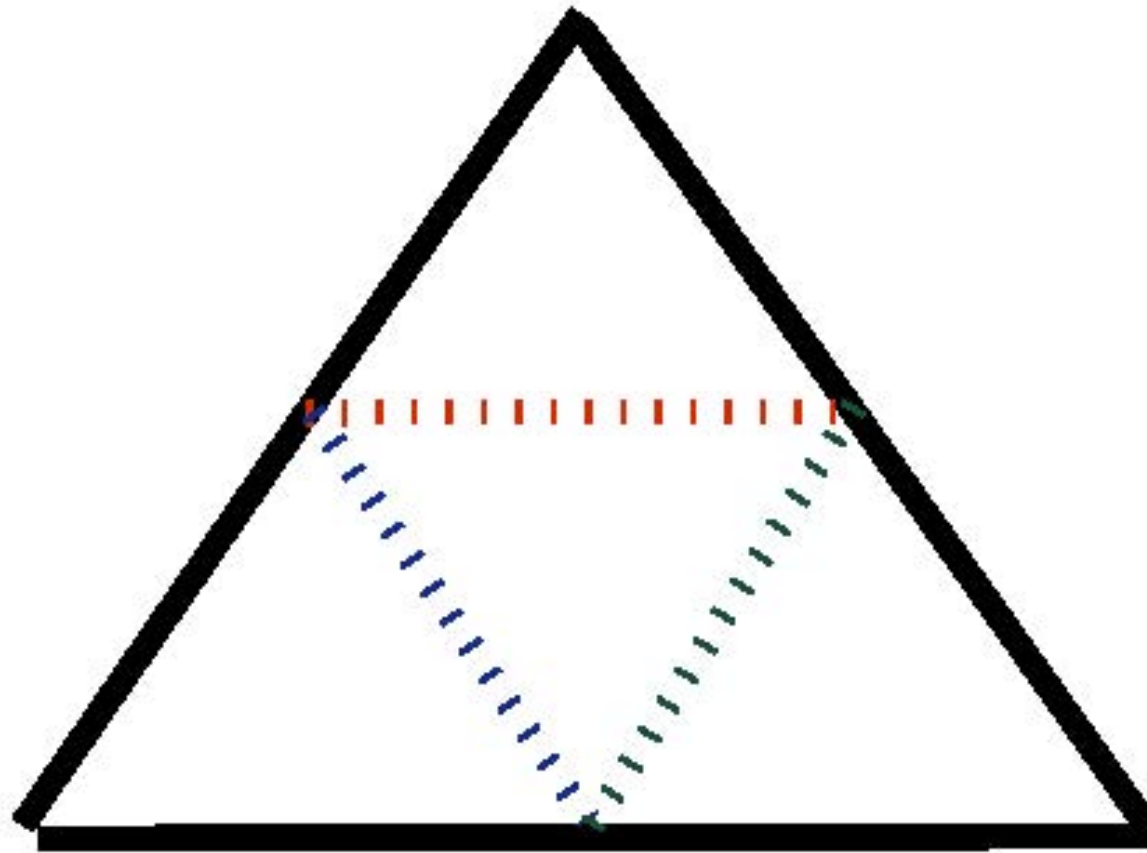


Figure 2: Contraction Simplexes

First-Order Algorithms: the Steepest Descent Method

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

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

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 direction vector $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$ as the search direction at \mathbf{x}^k . The number $\alpha^k \geq 0$, called step-size, is chosen “appropriately,” namely to satisfy

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

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

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**.

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

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 $f(x)$ exists and is given by the solution of the equation

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

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

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

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

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

Thus

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

The recursion for the method of steepest descent now becomes

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

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 2 Let $f : R^n \rightarrow R$ 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

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

where \mathbf{y}^k lies between $\mathbf{x}^k - \bar{\alpha} \nabla f(\mathbf{x}^k)^T$ and $\bar{\mathbf{y}}$. Then $\{\mathbf{y}^k\} \rightarrow \bar{\mathbf{y}}$ and $\{\nabla f(\mathbf{y}^k)\} \rightarrow \nabla f(\bar{\mathbf{y}})$ as $\{\mathbf{x}^k\} \rightarrow \bar{\mathbf{x}}$. 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 , which property is called **global convergence**.

(What about proof by Theorem 1?)

Convergence Rate of the SDM for 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 $0 < \lambda_1 \leq \dots \leq \lambda_n$. 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

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

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 $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)^T$. 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 $\nabla f(\mathbf{x}^{k+1}) \nabla f(\mathbf{x}^k)^T = 0$.

The Barzilai and Borwein Method

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

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

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

Then

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

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 .

An Explanation why the BB Method Works

For convex quadratic minimization, let the **distinct nonzero eigenvalues** of Hessian Q be $\lambda_1, \lambda_2, \dots, \lambda_K$; and let the step size in the SDM be $\alpha^k = \frac{1}{\lambda_k}$, $k = 1, \dots, K$. 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 .

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

Convergence Speed of the SDM for 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}

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\| \quad (3)$$

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}*

$$f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) \leq \frac{\beta}{2} \|\mathbf{x} - \mathbf{y}\|^2. \quad (4)$$

At the k th 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:

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

so that

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

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

$$f(\mathbf{x}^0) - f(\mathbf{x}^K) \geq \frac{1}{2\beta} \sum_{k=0}^{K-1} \|\nabla f(\mathbf{x}^k)\|^2. \quad (5)$$

Theorem 3 (*Error Convergence Estimate Theorem*) Let the objective function $p^* = \inf f(\mathbf{x})$ be finite and let us stop the SDM as soon as $\|\nabla f(\mathbf{x}^k)\| \leq \epsilon$ 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 (5), after $K = \frac{2\beta(f(\mathbf{x}^0) - p^*)}{\epsilon^2} + 1$ steps

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

If $\|\nabla f(\mathbf{x}^k)\| > \epsilon$ for all $k = 0, \dots, K$, 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 1 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 $\nabla f(\mathbf{y}) = \nabla f(\mathbf{x}^*) = \mathbf{0}$:

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