More First-Order Optimization Algorithms

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

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

Chapters 3, 8, 13
The SDM for Unconstrained Convex Lipschitz Optimization

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

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

(1)

**Theorem 1** For convex Lipschitz optimization the Steepest Descent Method generates a sequence of solutions such that

$$f(x^{k+1}) - f(x^*) \leq \frac{\beta}{k+2} \|x^0 - x^*\|^2 \quad \text{and} \quad \min_{l=0,\ldots,k} \|\nabla f(x^l)\|^2 \leq \frac{4\beta^2}{(k+1)(k+2)} \|x^0 - x^*\|^2,$$

where $x^*$ is a minimizer of the problem.

**Proof:** For simplicity, we let $\delta^k = f(x^k) - f(x^*) (\geq 0)$, $g^k = \nabla f(x^k)$, and $\Delta^k = x^k - x^*$ in the rest of proof. As we have proved for general Lipschitz optimization

$$\delta^{k+1} - \delta^k = f(x^{k+1}) - f(x^k) \leq -\frac{1}{2\beta} \|g^k\|^2,$$

that is

$$\delta^k - \delta^{k+1} \geq \frac{1}{2\beta} \|g^k\|^2.$$
Furthermore, from the convexity,

$$-\delta^k = f(x^*) - f(x^k) \geq (g^k)^T(x^* - x^k) = -(g^k)^T\Delta^k,$$

that is $\delta^k \leq (g^k)^T\Delta^k$. \hspace{1cm} (3)

Thus, from (2) and (3)

$$\delta^{k+1} = \delta^{k+1} - \delta^k + \delta^k$$

$$\leq -\frac{1}{2\beta} \|g^k\|^2 + (g^k)^T\Delta^k$$

$$= -\frac{\beta}{2} \|x^{k+1} - x^k\|^2 - \beta(x^{k+1} - x^k)^T\Delta^k, \quad \text{(using (1))}$$

$$= -\frac{\beta}{2} (\|x^{k+1} - x^k\|^2 + 2(x^{k+1} - x^k)^T\Delta^k)$$

$$= -\frac{\beta}{2} (\|\Delta^{k+1} - \Delta^k\|^2 + 2(\Delta^{k+1} - \Delta^k)^T\Delta^k)$$

$$= \frac{\beta}{2} (\|\Delta^k\|^2 - \|\Delta^{k+1}\|^2).$$ \hspace{1cm} (4)

Sum up (4) from 1 to $k + 1$, we have

$$\sum_{l=1}^{k+1} \delta^l \leq \frac{\beta}{2} (\|\Delta^0\|^2 - \|\Delta^{k+1}\|^2) \leq \frac{\beta}{2} \|\Delta^0\|^2.$$
From the proof of the Corollary 1 of last lecture, we have $\delta^0 \leq \frac{\beta}{2} \| \Delta^0 \|^2$. Thus,

$$\sum_{l=0}^{k+1} \delta^l \leq \beta \| \Delta^0 \|^2,$$

and

$$\sum_{l=0}^{k+1} \delta^l = \sum_{l=0}^{k+1} (l + 1 - l)\delta^l = \sum_{l=0}^{k+1} (l + 1)\delta^l - \sum_{l=0}^{k+1} l\delta^l = \sum_{l=1}^{k+2} l\delta^{l-1} - \sum_{l=1}^{k+1} l\delta^l = (k + 2)\delta^{k+1} + \sum_{l=1}^{k+1} l\delta^{l-1} - \sum_{l=1}^{k+1} l\delta^l = (k + 2)\delta^{k+1} + \sum_{l=1}^{k+1} l(\delta^{l-1} - \delta^l) \geq (k + 2)\delta^{k+1} + \sum_{l=1}^{k+1} l \frac{1}{2\beta} \| g^{l-1} \|^2,$$

where the first inequality comes from (2). Let $\| g' \| = \min_{l=0,\ldots,k} \| g^l \|$. Then we finally have

$$(k + 2)\delta^{k+1} + \frac{(k + 1)(k + 2)/2}{2\beta} \| g' \|^2 \leq \beta \| \Delta^0 \|^2,$$

which completes the proof.
The Accelerated Steepest Descent Method (ASDM)

There is an *accelerated* steepest descent method (Nesterov 83) that works as follows:

\[
\lambda^0 = 0, \quad \lambda^{k+1} = \frac{1 + \sqrt{1 + 4(\lambda^k)^2}}{2}, \quad \alpha^k = \frac{1 - \lambda^k}{\lambda^{k+1}}, \tag{7}
\]

\[
\tilde{x}^{k+1} = x^k - \frac{1}{\beta} \nabla f(x^k), \quad x^{k+1} = (1 - \alpha^k)\tilde{x}^{k+1} + \alpha^k\tilde{x}^k. \tag{8}
\]

Note that \((\lambda^k)^2 = \lambda^{k+1}(\lambda^{k+1} - 1), \lambda^k > k/2\) and \(\alpha^k \leq 0\).

One can prove:

**Theorem 2**

\[
f(\tilde{x}^{k+1}) - f(x^*) \leq \frac{2\beta}{k^2} \|x^0 - x^*\|^2, \quad \forall k \geq 1.
\]
Convergence Analysis of ASDM

Again for simplification, we let $\Delta^k = \lambda^k x^k - (\lambda^k - 1) \tilde{x}^k - x^*$, $g^k = \nabla f(x^k)$ and $\delta^k = f(\tilde{x}^k) - f(x^*) (\geq 0)$ in the following.

Applying Lemma 1 for $x = \tilde{x}^{k+1}$ and $y = \tilde{x}^k$, convexity of $f$ and (8) we have

$$\delta^{k+1} - \delta^k = f(\tilde{x}^{k+1}) - f(x^k) + f(x^k) - f(\tilde{x}^k)$$
$$\leq -\frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 + f(x^k) - f(\tilde{x}^k)$$
$$\leq -\frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 + (g^k)^T (x^k - \tilde{x}^k)$$
$$= -\frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 - \beta (\tilde{x}^{k+1} - x^k)^T (x^k - \tilde{x}^k).$$  \hspace{1cm} (9)

Applying Lemma 1 for $x = \tilde{x}^{k+1}$ and $y = x^*$, convexity of $f$ and (8) we have

$$\delta^{k+1} = f(\tilde{x}^{k+1}) - f(x^k) + f(x^k) - f(x^*)$$
$$\leq -\frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 + f(x^k) - f(x^*)$$
$$\leq -\frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 + (g^k)^T (x^k - x^*)$$
$$= -\frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 - \beta (\tilde{x}^{k+1} - x^k)^T (x^k - x^*).$$  \hspace{1cm} (10)
Multiplying (9) by $\lambda^k(\lambda^k - 1)$ and (10) by $\lambda^k$ respectively, and summing the two, we have

\[
(\lambda^k)^2 \delta^{k+1} - (\lambda^{k-1})^2 \delta^k \leq - (\lambda^k)^2 \frac{\beta}{2} \|\tilde{x}^{k+1} - x^k\|^2 - \lambda^k \beta (\tilde{x}^{k+1} - x^k)^T \Delta^k
\]

\[
= - \frac{\beta}{2} ((\lambda^k)^2 \|\tilde{x}^{k+1} - x^k\|^2 + 2\lambda^k (\tilde{x}^{k+1} - x^k)^T \Delta^k)
\]

\[
= - \frac{\beta}{2} \left( \|\lambda^k \tilde{x}^{k+1} - (\lambda^k - 1)\tilde{x}^k - x^*\|^2 - \|\Delta^k\|^2 \right)
\]

\[
= \frac{\beta}{2} \left( \|\Delta^k\|^2 - \|\lambda^k \tilde{x}^{k+1} - (\lambda^k - 1)\tilde{x}^k - x^*\|^2 \right).
\]

Using (7) and (8) we can derive

\[
\lambda^k \tilde{x}^{k+1} - (\lambda^k - 1)\tilde{x}^k = \lambda^{k+1} x^{k+1} - (\lambda^{k+1} - 1)\tilde{x}^{k+1}.
\]

Thus,

\[
(\lambda^k)^2 \delta^{k+1} - (\lambda^{k-1})^2 \delta^k \leq \frac{\beta}{2} (\|\Delta^k\|^2 - \|\Delta^{k+1}\|^2).
\]

(11)

Sum up (11) from 1 to $k$ we have

\[
\delta^{k+1} \leq \frac{\beta}{2(\lambda^k)^2} \|\Delta^1\|^2 \leq \frac{2\beta}{k^2} \|\Delta^0\|^2
\]

since $\lambda^k \geq k/2$ and $\|\Delta^1\| \leq \|\Delta^0\|$. 

\[7\]
First-Order Algorithms for Conic Constrained Optimization

Consider the conic nonlinear optimization problem: \( \min f(x) \quad \text{s.t.} \quad x \in K. \)

- **Nonnegative Linear Regression**: given data \( A \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}^m \)

  \[
  \min f(x) = \frac{1}{2} \|Ax - b\|^2 \quad \text{s.t.} \quad x \geq 0; \quad \text{where} \quad \nabla f(x) = A^T(Ax - b).
  \]

- **Semidefinite Linear Regression**: given data \( A_i \in S^n \) for \( i = 1, \ldots, m \) and \( b \in \mathbb{R}^m \)

  \[
  \min f(X) = \frac{1}{2} \|AX - b\|^2 \quad \text{s.t.} \quad X \succeq 0; \quad \text{where} \quad \nabla f(X) = A^T(AX - b).
  \]

\[
AX = \begin{pmatrix}
A_1 \cdot X \\
\vdots \\
A_m \cdot X
\end{pmatrix} \quad \text{and} \quad A^T y = \sum_{i=1} y_i A_i.
\]
I: SDM Followed by Feasible-Region-Projection

- $\hat{x}^{k+1} = x^k - \frac{1}{\beta} \nabla f(x^k)$
- $x^{k+1} = \text{Proj}_K(\hat{x}^{k+1})$

For examples:

- if $K = \{x : x \geq 0\}$, then
  $$x^{k+1} = \text{Proj}_K(\hat{x}^{k+1}) = \max\{0, \hat{x}^{k+1}\}.$$

- $K = \{X : X \succeq 0\}$, then factorize $\hat{X}^{k+1} = V^T \Lambda V$ and let
  $$X^{k+1} = \text{Proj}_K(\hat{X}^{k+1}) = V^T \max\{0, \Lambda\} V.$$

The drawback is that the eigenvalue-factorization may be costly in each iteration (using low-rank structure and/or $L^T D L$ factorization in Suggested Project I?)
II: The Gradient-Projection Method for Conic Constrained Optimization

Starting from a feasible solution $x^0 \geq 0$ and let the iterative mapping be

$$x^{k+1} = x^k + \alpha^k d^k;$$  \hspace{1cm} (12)

where

- for every $j$: $x^k$:

$$d^k_j = \begin{cases} 
-\nabla f(x^k)_j & \text{if } x^k_j > 0 \text{ or } \nabla f(x^k)_j < 0, \\
0 & \text{otherwise.}
\end{cases}$$

- the stepsize can be chosen from line-search (keeping feasibility) or

$$\alpha^k = \min\left\{ \frac{1}{\beta}, \alpha_{max}^k \right\}$$

where $\alpha_{max}^k$ is the largest stepsize $\alpha$ such that $x^{k+1} = x^k + \alpha d^k \geq 0$.

Does it converge? What is the convergence speed? See more details in HW3. And the extension to SDP cone?
III: The Gradient-Projection Method for Equality Constrained Optimization

Consider the conic nonlinear optimization problem: $\min f(x) \quad \text{s.t.} \quad Ax = b$.

Starting from a feasible solution $x^0$ and let the iterative mapping be

$$d^k = -(I - A^T(ATA)^{-1}A) \nabla f(x^k)$$

$$x^{k+1} = x^k + \alpha^k d^k; \quad (13)$$

where the stepsize can be chosen from line-search or simply

$$\alpha^k = \frac{1}{\beta}$$

and $\beta$ is the global Lipschitz constant.

Does it converge? What is the convergence speed? See more details in HW3. And the extension to SDP cone?
IV: Affine Scaling SDM for Conic Constrained Optimization

At the \( k \)th iterate with \( \mathbf{x}^k > \mathbf{0} \), let \( D^k \) be a diagonal matrix such that

\[
D_{jj}^k = \min\{1, x_j^k\}, \quad \forall j
\]

and

\[
\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k (D^k)^2 \nabla f(\mathbf{x}^k)
\]

where step-size

\[
\alpha^k = \min\left\{ \frac{1}{\beta}, \frac{1}{2\|D^k \nabla f(\mathbf{x}^k)\|_{\infty}} \right\}.
\]

Is \( \mathbf{x}^k > \mathbf{0}, \quad \forall k \)? Does it converge? What is the convergence speed? See more details in HW3.
IV: Affine Scaling for SDP Cone?

At the $k$th iterate with $X^k \succ 0$. Let $D^k$ be the symmetric matrix $D^k$ such that

$$D^k = V^T \min\{1, \Lambda\} V,$$

where $X^k = V^T \Lambda V$.

the new SDM iterate would be

$$X^{k+1} = X^k - \alpha_k D^k \nabla f(X^k) D^k.$$

Choose step-size is chosen such that the smallest eigenvalue of $X^{k+1}$ is at most halved from the one of $X^k$? Does it converge? What is the convergence speed? See more details in HW3.
V: Reduced Gradient Method – the Simplex Algorithm for LP

\[ \text{LP: } \min c^T x \text{ s.t. } Ax = b, \ x \geq 0, \]

where \( A \in \mathbb{R}^{m \times n} \) has a full row rank \( m \).

**Theorem 3** (The Fundamental Theorem of LP in Algebraic form) Given (LP) and (LD) where \( A \) has full row rank \( m \),

i) if there is a feasible solution, there is a basic feasible solution (Carathéodory’s theorem);

ii) if there is an optimal solution, there is an optimal basic solution.

**High-Level Idea:**

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

2. **Test for Optimality.** Compute the reduced gradient vector at the corner. If no descent and feasible direction can be found, stop and claim optimality at the current corner point; otherwise, select a new corner point and go to Step 2.
LP theorems depicted in two variable space

Figure 1: The LP Simplex Method
When a Basic Feasible Solution is Optimal

Suppose the basis of a basic feasible solution is $A_B$ and the rest is $A_N$. One can transform the equality constraint to

$$A_B^{-1}Ax = A_B^{-1}b,$$

so that $x_B = A_B^{-1}b - A_B^{-1}A_Nx_N$.

That is, we express $x_B$ in terms of $x_N$, the non-basic variables are are active for constraints $x \geq 0$.

Then the objective function equivalently becomes

$$c^T x = c_B^T x_B + c_N^T x_N = c_B^T A_B^{-1}b - c_B^T A_B^{-1}A_Nx_N + c_N^T x_N$$

$$= c_B^T A_B^{-1}b + (c_N^T - c_B^T A_B^{-1}A_N)x_N.$$

Vector $r^T = c^T - c_B^T A_B^{-1}A$ is called the Reduced Gradient/Cost Vector where $r_B = 0$ always.

**Theorem 4** If Reduced Gradient Vector $r^T = c^T - c_B^T A_B^{-1}A \geq 0$, then the BFS is optimal.

**Proof:** Let $y^T = c_B^T A_B^{-1}$ (called Shadow Price Vector), then $y$ is a dual feasible solution ($r = c - A^T y \geq 0$) and $c^T x = c_B^T x_B = c_B^T A_B^{-1}b = y^T b$, that is, the duality gap is zero.
The Simplex Algorithm Procedures

0. Initialize  Start a BFS with basic index set $B$ and let $N$ denote the complementary index set.

1. Test for Optimality: Compute the Reduced Gradient Vector $r$ at the current BFS and let

$$r_e = \min_{j \in N} \{ r_j \}.$$\

If $r_e \geq 0$, stop – the current BFS is optimal.

2. Determine the Replacement: Increase $x_e$ while keep all other non-basic variables at the zero value (inactive) and maintain the equality constraints:

$$x_B = A_B^{-1}b - A_B^{-1}A_e x_e \geq 0.$$\

If $x_e$ can be increased to $\infty$, stop – the problem is unbounded below. Otherwise, let the basic variable $x_o$ be the one first becoming 0.

3. Update basis: update $B$ with $x_o$ being replaced by $x_e$, and return to Step 1.
A Toy Example

minimize \(-x_1 - 2x_2\)
subject to \(x_1 + x_3 = 1\)
\(x_2 + x_4 = 1\)
\(x_1 + x_2 + x_5 = 1.5\).

\[
A = \begin{pmatrix}
1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 1 \\
\end{pmatrix}, \quad b = \begin{pmatrix}
1 \\
1 \\
1.5 \\
\end{pmatrix}, \quad c^T = (-1 - 2 0 0 0).
\]

Consider initial BFS with basic variables \(B = \{3, 4, 5\}\) and \(N = \{1, 2\}\).

Iteration 1:

1. \(A_B = I, A_B^{-1} = I, y^T = (0\ 0\ 0)\) and \(r_N = (-1 - 2)\) – it’s NOT optimal. Let \(e = 2\).
2. Increase $x_2$ while

$$x_B = A_B^{-1} b - A_B^{-1} A_x x_2 = \begin{pmatrix} 1 \\ 1 \\ 1.5 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} x_2.$$ 

We see $x_4$ becomes 0 first.

3. The new basic variables are $B = \{3, 2, 5\}$ and $N = \{1, 4\}$.

Iteration 2:

1. 

$$A_B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad A_B^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

$$y^T = (0 \quad -2 \quad 0) \text{ and } r_N = (-1 \quad 2) \text{ – it’s NOT optimal. Let } e = 1.$$
2. Increase $x_1$ while

$$x_B = A_B^{-1}b - A_B^{-1}A_1x_1 = \begin{pmatrix} 1 \\ 1 \\ 0.5 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}x_1.$$ 

We see $x_5$ becomes 0 first.

3. The new basic variables are $B = \{3, 2, 1\}$ and $N = \{4, 5\}$.

**Iteration 3:**

1. 

$$A_B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad A_B^{-1} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

$$y^T = (0 -1 -1) \text{ and } r_N = (1 1) - \text{it's Optimal.}$$

Is the Simplex Method always convergent to a minimizer? Which condition of the Global Convergence Theorem failed?
Value-Iteration for MDP: Fixed-Point Mapping

Let \( y \in \mathbb{R}^m \) represent the cost-to-go values of the \( m \) states, \( i \)th entry for \( i \)th state, of a given policy. The MDP problem entails choosing the optimal value vector \( y^* \) which is a fixed-point of:

\[
y_i^* = \min_{j \in A_i} \{ c_j + \gamma p_j^T y^* \}, \forall i,
\]

The Value-Iteration (VI) Method is, starting from any \( y^0 \), the iterative mapping:

\[
y_i^{k+1} = A(y^k)_j = \min_{j \in A_i} \{ c_j + \gamma p_j^T y^k \}, \forall i.
\]

If the initial \( y^0 \) is strictly feasible for state \( i \), that is, \( y_i^0 < c_j + \gamma p_j^T y^0, \forall j \in A_i \), then \( y_i^k \) would be increasing in the VI iteration for all \( i \) and \( k \).

On the other hand, if any of the inequalities is violated, then we have to decrease \( y_i^1 \) at least to

\[
\min_{j \in A_i} \{ c_j + \gamma p_j^T y^0 \}
\].
### Convergence of Value-Iteration for MDP

**Theorem 5** Let the VI algorithm mapping be $A(v)_i = \min_{j \in A} \{c_j + \gamma p_j^T v, \forall i\}$. Then, for any two value vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^m$ and every state $i$:

$$|A(u)_i - A(v)_i| \leq \gamma \|u - v\|_\infty,$$

which implies $\|A(u)_i - A(v)_i\|_\infty \leq \gamma \|u - v\|_\infty$

Let $j_u$ and $j_v$ be the two arg min actions for value vectors $u$ and $v$, respectively. Assume that $A(u)_i - A(v)_i \geq 0$ where the other case can be proved similarly.

$$0 \leq A(u)_i - A(v)_i = (c_{j_u} + \gamma p_{j_u}^T u) - (c_{j_v} + \gamma p_{j_v}^T v) \leq (c_{j_v} + \gamma p_{j_v}^T u) - (c_{j_v} + \gamma p_{j_v}^T v) = \gamma p_{j_v}^T (u - v) \leq \gamma \|u - v\|_\infty.$$

where the first inequality is from that $j_u$ is the arg min action for value vector $u$, and the last inequality follows from the fact that the elements in $p_{j_v}$ are non-negative and sum-up to 1.

Many research issues in suggested Project III.