First-Order Optimization Methods

Yinyu Ye

MS&E and ICME, Stanford University

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

(Chapters 7 and 8)

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

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

Line Search

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 1 *Let $f: R^n \to R$ be given. For some given point $\mathbf{x}^0 \in R^n$, let the level set*

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

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,

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

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

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

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.

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.

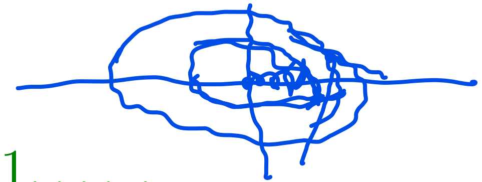
Convergence Speed of the SDM for Strongly Convex QP

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.

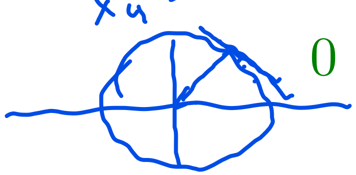
$f(x^*)=0$ $x=0$

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)

< 1



$x_4=\lambda_1$



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

Intuitively, the slow 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 $\mathbf{y}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2 \tag{1}$$

$\frac{1}{2}x^T Q x$

$\nabla f(x) = Qx$

$\|Q(x-y)\|_2 \leq \beta \lambda_n \|x-y\|$

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

x^0 $\quad \ge \varepsilon$

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

$f(x^0) - z^* \ge \frac{K}{2\beta}\varepsilon^2$ $\qquad K = \frac{2\beta\left(f(x^0) - z^*\right)}{\varepsilon^2}$

Theorem 2 *(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 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 $\nabla f(\mathbf{y}) = \nabla f(\mathbf{x}^*) = \mathbf{0}$:

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

$\|\nabla f(x)\| \leq \varepsilon$

$O(\frac{1}{\varepsilon^2})$

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

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

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

$< \varepsilon$

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

$k \approx O(\frac{1}{\varepsilon})$

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

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

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.

$\boxed{\beta}$

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

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

$\frac{1}{\beta}$

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

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

Prove that the selected step-size

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

The Barzilai and Borwein Method

x^k, x^{k-1}, $\nabla f(x^{k-1})$

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

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

$$\frac{\Delta^T Q \Delta}{\Delta^T \Delta}$$

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

$$\alpha^k = \frac{(\Delta_x^k)^T \Delta_g^k}{(\Delta_g^k)^T \Delta_g^k} \quad \text{or} \quad \alpha^k = \frac{(\Delta_x^k)^T \Delta_x^k}{(\Delta_x^k)^T \Delta_g^k}, = \frac{\|\Delta_x^k\|_2^2}{\Delta_x^k Q \Delta_x^k}$$

Q $10^6 \times 10^6$

Then

λ

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

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

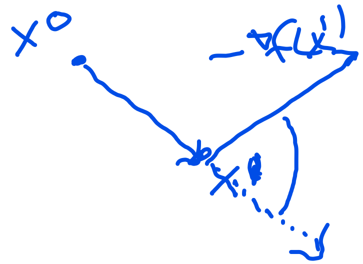
d^k $-\nabla f(x^k)$

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

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.



Double-Directions: The QP Heavy-Ball Method (Polyak 64)

α^g α^m

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{4}{(\sqrt{\lambda_n} + \sqrt{\lambda_1})^2}\nabla f(\mathbf{x}^k) + \left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}\right)(\mathbf{x}^k - \mathbf{x}^{k-1}).$$

where the convergence rate can be improved to

$$\left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}\right)^2. \quad \cancel{<} \quad \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2$$

This is also called the Parallel-Tangent or Conjugate Direction method, where the second direction-term in the formula is nowadays called “acceleration” or“momentum” direction.

For minimizing general convex functions, we can let

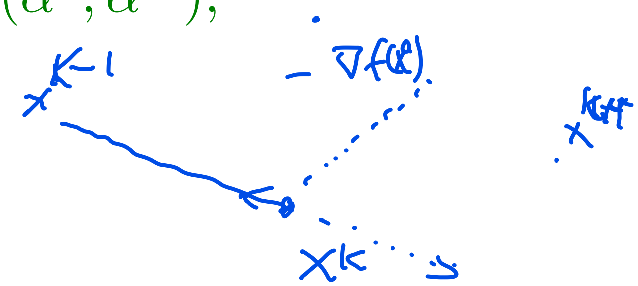
$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^g \nabla f(\mathbf{x}^k) + \alpha^m(\mathbf{x}^k - \mathbf{x}^{k-1}) = \mathbf{x}^k + \mathbf{d}(\alpha^g, \alpha^m),$$

where the pair of step-sizes (α^g, α^m) can be chosen to

$$\min_{(\alpha^g, \alpha^d)} \nabla f(\mathbf{x}^k)\mathbf{d}(\alpha^g, \alpha^m) + \frac{1}{2}\mathbf{d}(\alpha^g, \alpha^m)\nabla^2 f(\mathbf{x}^k)\mathbf{d}(\alpha^g, \alpha^m),$$

$\left\| \begin{pmatrix} \alpha^g \\ \alpha^m \end{pmatrix} \right\| \le 1$

where $\mathbf{x}^1$ can be computed from the SDM step.



DRSOM: The Close-Form Step-Size from Quadratic Approximation

Let $\mathbf{d}^k = \mathbf{x}^k - \mathbf{x}^{k-1}$, $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ and $H^k = \nabla^2 f(\mathbf{x}^k)$, then the step-sizes can be chosen from

$$\begin{pmatrix} (\mathbf{g}^k)^T H^k \mathbf{g}^k & -(\mathbf{d}^k)^T H^k \mathbf{g}^k \\ -(\mathbf{d}^k)^T H^k \mathbf{g}^k & (\mathbf{d}^k)^T H^k \mathbf{d}^k \end{pmatrix} \begin{pmatrix} \alpha^g \\ \alpha^m \end{pmatrix} = \begin{pmatrix} \|\mathbf{g}^k\|^2 \\ -(\mathbf{g}^k)^T \mathbf{d}^k \end{pmatrix}.$$

$f(x)$

Hd

Hg

If the Hessian $\nabla^2 f(\mathbf{x}^k)$ is not available, one can approximate

$$H^k \mathbf{g}^k \sim \nabla(\mathbf{x}^k + \mathbf{g}^k) - \mathbf{g}^k \quad \text{and} \quad H^k \mathbf{d}^k \sim \nabla(\mathbf{x}^k + \mathbf{d}^k) - \mathbf{g}^k \sim -(\mathbf{g}^{k-1} - \mathbf{g}^k);$$

$\lim_{\epsilon \to 0} \frac{\nabla f(x+\epsilon d) - \nabla f(x)}{\epsilon} = H(x) \cdot \nabla f(x)$

SNL

Logic

or for some small $\epsilon > 0$:

Hd

$d = x^k - x^{k-1}$

$\nabla f(x^k) - \nabla f(x^{k-1})$

$$H^k \mathbf{g}^k \sim \frac{1}{\epsilon}(\nabla(\mathbf{x}^k + \epsilon \mathbf{g}^k) - \mathbf{g}^k) \quad \text{and} \quad H^k \mathbf{d}^k \sim \frac{1}{\epsilon}(\nabla(\mathbf{x}^k + \epsilon \mathbf{d}^k) - \mathbf{g}^k).$$

$\frac{1}{2}\|A^1 x - b^1\|^2 + \frac{1}{2}\|A^2 x - b^2\|^2 + \frac{1}{2}\|A^3 x - b^3\|^2$

x

DC#1 DC#2 DC#3

Application in **Federated-Learning**.

$\frac{1}{2}\|Ax-b\|_2^2 \leftarrow A^T(Ax-b)$

The Accelerated Steepest Descent Method (ASDM)

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

$$\lambda^0 = 0,\ \lambda^{k+1} = \frac{1+\sqrt{1+4(\lambda^k)^2}}{2},\ \alpha^k = \frac{1-\lambda^k}{\lambda^{k+1}}, \tag{12}$$

$$\tilde{\mathbf{x}}^{k+1} = \mathbf{x}^k - \frac{1}{\beta}\nabla f(\mathbf{x}^k),\ \mathbf{x}^{k+1} = (1-\alpha^k)\tilde{\mathbf{x}}^{k+1} + \alpha^k\tilde{\mathbf{x}}^k. \tag{13}$$

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

One can prove:

Theorem 4

ε

$$\underbrace{f(\tilde{\mathbf{x}}^{k+1}) - f(\mathbf{x}^*) \le \frac{2\beta}{k^2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2,\ \forall k \ge 1.}$$

$\Rightarrow$

$$\boxed{\begin{array}{c} O\left(\frac{1}{\varepsilon^2}\right) \\ O\left(\frac{1}{\varepsilon}\right) \\ O\left(\frac{1}{\sqrt{\varepsilon}}\right) \end{array}}$$

Convergence Analysis of ASDM

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

Applying Lemma 1 for $\mathbf{x} = \tilde{\mathbf{x}}^{k+1}$ and $\mathbf{y} = \tilde{\mathbf{x}}^k$, convexity of f and (13) we have

$$\begin{aligned}
\delta^{k+1} - \delta^k &= f(\tilde{\mathbf{x}}^{k+1}) - f(\mathbf{x}^k) + f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \\
&\leq -\frac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \\
&\leq -\frac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + (\mathbf{g}^k)^T(\mathbf{x}^k - \tilde{\mathbf{x}}^k) \\
&= -\frac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 - \beta(\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k)^T(\mathbf{x}^k - \tilde{\mathbf{x}}^k).
\end{aligned} \tag{14}$$

Applying Lemma 1 for $\mathbf{x} = \tilde{\mathbf{x}}^{k+1}$ and $\mathbf{y} = \mathbf{x}^*$, convexity of f and (13) we have

$$\begin{aligned}
\delta^{k+1} &= f(\tilde{\mathbf{x}}^{k+1}) - f(\mathbf{x}^k) + f(\mathbf{x}^k) - f(\mathbf{x}^*) \\
&\leq -\frac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + f(\mathbf{x}^k) - f(\mathbf{x}^*) \\
&\leq -\frac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + (\mathbf{g}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \\
&= -\frac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 - \beta(\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*).
\end{aligned} \tag{15}$$

Multiplying (14) by $\lambda^k(\lambda^k - 1)$ and (15) by λ^k respectively, and summing the two, we have

$$\begin{aligned}(\lambda^k)^2\delta^{k+1} - (\lambda^{k-1})^2\delta^k &\le -(\lambda^k)^2\tfrac{\beta}{2}\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 - \lambda^k\beta(\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k)^T\Delta^k \\ &= -\tfrac{\beta}{2}((\lambda^k)^2\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + 2\lambda^k(\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k)^T\Delta^k) \\ &= -\tfrac{\beta}{2}(\|\lambda^k\tilde{\mathbf{x}}^{k+1} - (\lambda^k - 1)\tilde{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \|\Delta^k\|^2) \\ &= \tfrac{\beta}{2}(\|\Delta^k\|^2 - \|\lambda^k\tilde{\mathbf{x}}^{k+1} - (\lambda^k - 1)\tilde{\mathbf{x}}^k - \mathbf{x}^*\|^2).\end{aligned}$$

Using (12) and (13) we can derive

$$\lambda^k\tilde{\mathbf{x}}^{k+1} - (\lambda^k - 1)\tilde{\mathbf{x}}^k = \lambda^{k+1}\mathbf{x}^{k+1} - (\lambda^{k+1} - 1)\tilde{\mathbf{x}}^{k+1}.$$

Thus,

$$(\lambda^k)^2\delta^{k+1} - (\lambda^{k-1})^2\delta^k \le \frac{\beta}{2}(\|\Delta^k\|^2 - \|\Delta^{k+1}\|^2.) \tag{16}$$

Sum up (16) from 1 to k we have

$$\delta^{k+1} \le \frac{\beta}{2(\lambda^k)^2}\|\Delta^1\|^2 \le \frac{2\beta}{k^2}\|\Delta^0\|^2$$

since $\lambda^k \ge k/2$ and $\|\Delta^1\| \le \|\Delta^0\|$.