**MATH: Operations Research** **2014-15 First Term**

# Handout 11: Gradient method

*Instructor: Junfeng Yang* *December 2, 2014*

## Contents

## 11.1 General framework

Consider $\min_{x\in R^n} f(x)$ with $f \in C^1(R^n)$. Assume that the current point is $x_k$.

**Motivation 1** The direction $d_k = -\nabla f(x_k)$ decreases $f$ fastest at $x_k$. Step forward in this direction with ceratin step length $h_k > 0$:

$$x_{k+1} = x_k - h_k \nabla f(x_k).$$

$f_{k+1} < f_k$ if $\nabla f(x_k) \neq 0$ and $h_k$ sufficiently small.

**Motivation 2** To obtain $x_{k+1}$, we minimize a simple approximate function of $f$ at $x_k$:

$$f(x) \approx f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2h_k}\|x - x_k\|^2,$$

where $h_k > 0$. Again, we obtain $x_{k+1} = x_k - h_k \nabla f(x_k)$.

**Algorithm 1 (Gradient method)** *Initialization: choose $x_0 \in R^n$, set $k = 0$.*

1. *Compute $\nabla f(x_k)$;*
2. *Compute a step size $h_k > 0$ satisfying certain conditions;*
3. *Compute $x_{k+1} = x_k - h_k \nabla f(x_k)$;*
4. *If stopping criterion is satisfied, stop; Otherwise, $k$++, go to step 2.*

**Remark 11.1.1** *At each iteration, the search direction of gradient method is $d_k = -\nabla f(x_k)$, which is locally optimal in the sense that $f$ decreases the fastest at $x_k$ along this direction. $h_k > 0$ is called the step size (or step length) at the $k$th iteration. Usually stopping criterion is $\|\nabla f(x_k)\| \leq \epsilon$ for some $\epsilon > 0$. Other criteria are also useful. Advantages of gradient method: simple and inexpensive (2nd order derivative not required).*

**Theorem 11.1 (Global convergence with exact line search)** *Let $f \in C^1$ and $\{x_k\}$ be the sequence of points generated by gradient method with exact line search rule. Then, any limit point $\bar{x}$ of $\{x_k\}$ is a stationary point of $f$, i.e.,*

$$\nabla f(\bar{x}) = 0.$$

(Proof: page 109, Theorem 3.1.2a in Yuan-Sun book.)

**Remark 11.1.2** *For $f \in C^1$, in general one can only guarantee that any limit point of the sequence generated by gradient method is a stationary point. Without stronger assumption on $f$, this is the best convergence result one can get. The convergence of gradient method with inexact line search cannot be stronger.*

**Example 11.1.1** *Consider minimizing the following 2-dimensional function over $R^n$:*

$$f(x) = f(x^{(1)}, x^{(2)}) = \frac{1}{2}(x^{(1)})^2 + \frac{1}{4}(x^{(2)})^4 - \frac{1}{2}(x^{(2)})^2.$$

*Clearly $f$ is smooth (infinite times continuously differentiable) and*

$$\nabla f(x) = \begin{pmatrix} x^{(1)} \\ (x^{(2)})^3 - x^{(2)} \end{pmatrix} \quad \textit{and} \quad \nabla^2 f(x) = \begin{pmatrix} 1 & 0 \\ 0 & 3(x^{(2)})^2 - 1 \end{pmatrix}.$$

*Three stationary points:* $x_1^* = (0,0)$, $x_2^* = (0,-1)$, $x_3^* = (0,1)$. $f(x_1^*) = 0$ *and* $f(x_1^*+\epsilon e_2) = \frac{\epsilon^4}{4} - \frac{\epsilon^2}{2}$ *for* $0 < \epsilon \ll 1$. *Thus* $x_1^*$ *is not local minimizer.* $x_2^*$ *and* $x_3^*$ *are local minimizers.*

*Suppose we minimize* $f$ *by gradient method and start with* $x_0 = (1,0)$. *Since the second coordinate of* $x_0$ *is* $0$*, that of* $\nabla f(x_0)$ *is also* $0$*, thus that of* $x_1$ *is zero, ... The second coordinate of the whole sequence* $\{x_k\}$ *generated by gradient method is* $0$*. Thus,* $\{x_k\}$ *converges to* $x_1^*$*. Note that this is true no matter what step size rule is used.*

## 11.2 Gradient method for Lipschitz continuous functions

**Gradient method for $C_L^{1,1}(R^n)$ – Global convergence.** Consider $\min_{x\in R^n} f(x)$, where $f \in C_L^{1,1}(R^n)$ and is bounded below and the minimum value of $f$ is attained at $x^*$, i.e.,

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

If $y = x - h\nabla f(x)$ (denote $\nabla f(x)$ by $g(x)$ in the following), then

$$\begin{aligned} f(y) &\le f(x) + g(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 \\ &= f(x) - h\|g(x)\|^2 + \frac{L}{2}h^2\|g(x)\|^2 \\ &= f(x) - h(1 - Lh/2)\|g(x)\|^2. \end{aligned}$$

If $h = 1/L$ (which minimizes the upper bound on the right side), then

$$f(y) = f\left(x - \frac{1}{L}\nabla f(x)\right) \le f(x) - \frac{1}{2L}\|g(x)\|^2.$$

**Corollary 11.2** *Suppose* $f \in C_L^{1,1}(R^n)$ *and* $f(x^*) = \min_{x\in R^n} f(x)$*. It holds that*

$$\frac{1}{2L}\|\nabla f(x)\|^2 \le f(x) - f(x^*) \le \frac{L}{2}\|x - x^*\|^2, \quad \forall x \in R^n.$$

In the following, we let $x_{k+1} = x_k - h_k g_k$, where $h_k > 0$ satisfies certain line search rule.

- Constant step size $h_k \equiv h = \frac{2\eta}{L}$ with $\eta \in (0,1)$:

  $$f_k - f_{k+1} \ge h\left(1 - \frac{Lh}{2}\right)\|g_k\|^2 = \frac{2}{L}\eta(1-\eta)\|g_k\|^2.$$

  The “optimal" choice is $h_k \equiv h = \frac{1}{L}$, i.e., $\eta = 1/2$.
- Exact line search: surely $f_k - f_{k+1} \ge \frac{1}{2L}\|g_k\|^2$.
- Goldstein-Armijo rule: The first condition implies

  $$\beta h_k\|g_k\|^2 \ge f_k - f_{k+1} \ge h_k\left(1 - \frac{Lh_k}{2}\right)\|g_k\|^2.$$

  Thus, $h_k \ge 2(1-\beta)/L$ (therefore, the step size not too small). Further considering the first condition, we get

  $$f_k - f_{k+1} \ge \alpha h_k\|g_k\|^2 \ge \frac{2}{L}\alpha(1-\beta)\|g_k\|^2.$$

- Backtracking line search: suppose the first trail step size is $\hat{h}$, then $h_k \geq \min(\hat{h}, \delta/L)$ because

  $$f(y) \leq f(x) - \frac{h}{2}\|g(x)\|^2, \quad \forall h \in (0, 1/L),$$

  which implies that $h_k \leq 1/L$ is sufficient to satisfy the backtracking condition. As a result,

  $$f_k - f_{k+1} \geq \min(\hat{h}, \delta/L)\|g_k\|^2.$$

In all the above discussed cases, there exists $\omega > 0$ such that

$$f_k - f_{k+1} \geq \frac{\omega}{L}\|g_k\|^2.$$

Sum up for $k = 0, 1, 2, \ldots, N$. Obtain

$$\frac{\omega}{L}\sum_{k=0}^{N}\|g_k\|^2 \leq f_0 - f_{N+1} \leq f_0 - f^*,$$

where $f^* = f(x^*) = \min_{x \in R^n} f(x)$. As a consequence

$$\lim_{k \to \infty} \|g_k\| = 0.$$

Moreover, it holds that

$$\min_{0 \leq k \leq N} \|g_k\| \leq \frac{1}{\sqrt{N+1}}\sqrt{\frac{L}{\omega}(f_0 - f^*)}.$$

The right hand side of this inequality describes the rate of convergence of the sequence

$$\left\{\min_{0 \leq k \leq N} \|g_k\| : N = 0, 1, \ldots\right\}.$$

To obtain a point satisfying $\|g_k\| \leq \epsilon$, an upper bound of required number of iterations is

$$\frac{L}{\omega\epsilon^2}(f_0 - f^*).$$

Without stronger assumption, nothing can be said about the rate of convergence of the sequences $\{f(x_k)\}$ and $\{x_k\}$.

**Theorem 11.3 (Local convergence)** *Let $f$ satisfy the following conditions*

1. $f \in C_M^{2,2}(R^n)$.
2. *There exists a local minimum $x^*$ of $f$ such that $\nabla^2 f(x^*) \in S_{++}^n$.*
3. $\ell I_n \preccurlyeq \nabla^2 f(x^*) \preccurlyeq L I_n$ *with $0 < \ell \leq L < \infty$. (essentially assume that $f$ is strongly convex and with Lipschitz gradient around $x^*$.)*
4. *The starting point $x_0$ is sufficiently close to $x^*$:*

   $$r_0 := \|x_0 - x^*\| < \bar{r} := \frac{2\ell}{M}.$$

*Then the gradient method with step size $h_k \equiv \frac{2}{L+\ell}$ converges as follows*

$$\|x_k - x^*\| \leq \frac{\bar{r}r_0}{\bar{r} - r_0}\left(1 - \frac{2}{L/\ell + 3}\right)^k.$$

*This rate of convergence is called linear.*

## 11.3 Gradient method for convex Lipschitz continuous functions

**Gradient method for $\mathcal{F}_L^{1,1}(R^n)$.** Solve $\min_{x\in R^n} f(x)$ by gradient method, where $f \in \mathcal{F}_L^{1,1}(R^n)$. Assume $f^* := f(x^*) = \min_{x\in R^n} f(x) > -\infty$ ($f^*$ is attained at $x^*$).

**Line search condition:** step size $h_k$ satisfies

$$f(x_k - h_k g_k) \leq f(x_k) - \frac{h_k}{2}\|g(x_k)\|^2.$$

Let $y = x - hg(x)$. Since

$$f(y) \leq f(x) - \frac{h}{2}\|g(x)\|^2$$

for $0 < h \leq 1/L$, this line search condition can always be satisfied as long as $h$ is sufficiently small.

Gradient method for $\mathcal{F}_L^{1,1}(R^n)$ Let $x_{k+1} = x_k - h_k g_k$. For constant step size rule with $0 < h_k \equiv h \leq 1/L$, exact line search rule, Goldstein-Armijo step size rule and backtracking line search rule, there exists $\underline{h} > 0$ such that $h_k > \underline{h} > 0$ and

$$f_{k+1} \leq f_k - \frac{\underline{h}}{2}\|g_k\|^2.$$

- For constant step size $h_k \equiv 1/L$ and exact line search: $\underline{h} = 1/L$.
- For Goldstein-Armijo step size rule: $\underline{h} = 4\alpha(1-\beta)/L$.
- For back tracking line search: $\underline{h} = 2\min(\hat{h}, \delta/L)$.

From convexity of $f$, there holds

$$f(x_k) \leq f(x^*) + \nabla f(x_k)^T(x_k - x^*).$$

Thus, it holds that

$$f_{k+1} \leq f_k - \frac{h_k}{2}\|g_k\|^2 \leq f^* + g_k^T(x_k - x^*) - \frac{h_k}{2}\|g_k\|^2.$$

Therefore,

$$\begin{aligned} 0 \leq f_{k+1} - f^* &\leq g_k^T(x_k - x^*) - \frac{h_k}{2}\|g_k\|^2 \\ &= \frac{1}{2h_k}\left(\|x_k - x^*\|^2 - \|x_k - x^* - h_k g_k\|^2\right) \\ &= \frac{1}{2h_k}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right). \end{aligned}$$

Consequently $\|x_{k+1} - x^*\| \leq \|x_k - x^*\|$. For the four types of line search rules, $h_k \geq \underline{h} > 0$. Therefore,

$$\sum_{i=0}^{k-1}(f_{i+1} - f^*) \leq \sum_{i=0}^{k-1}\frac{1}{2h_i}\left(\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2\right) \leq \frac{1}{2\underline{h}}\|x_0 - x^*\|^2.$$

Further considering that $f_{k+1} \leq f_k - \frac{h_k}{2}\|g_k\|^2 \leq f_k$, it holds that

$$k(f_k - f^*) \leq \sum_{i=0}^{k-1}(f_{i+1} - f^*) \leq \frac{1}{2\underline{h}}\|x_0 - x^*\|^2,$$

or, equivalently,

$$0 \le f_k - f^* \le \frac{\|x_0 - x^*\|^2}{2\underline{h}k}, \quad \forall k \ge 1.$$

**Theorem 11.4** (**$O(1/k)$ convergence of gradient method**) *Consider solving $\min_{x\in R^n} f(x)$ by gradient method, where $f \in \mathcal{F}_L^{1,1}(R^n)$ is bounded below and $\min_{x\in R^n} f(x)$ is attained at $x^*$. Using either constant step size $0 < h_k \equiv h \le 1/L$, exact line search, Goldstein-Armijo or backtracking line search, there exists $\underline{h} > 0$ such that*

$$f_{k+1} \le f_k - \frac{\underline{h}}{2}\|g_k\|^2, \quad \forall k \ge 1.$$

*Furthermore, the sequence $\{x_k\}$ satisfies*

$$0 \le f_k - f^* \le \frac{\|x_0 - x^*\|^2}{2\underline{h}k}, \quad \forall k \ge 1,^1$$

*and*

$$\|\nabla f(x_k)\| \le \frac{\sqrt{L/\underline{h}}\|x_0 - x^*\|}{\sqrt{k}}, \quad \forall k \ge 1.^2$$

**Lower complexity bounds for $\mathcal{F}_L^{\infty,1}(R^n)$.** Consider $\min_{x\in R^n} f(x)$, where $f \in \mathcal{F}_L^{1,1}(R^n)$. Suppose we solve this problem by an iterative method $\mathcal{M}$ satisfying the following assumptions:

1. $\mathcal{M}$ only has access to $f(x)$ and $\nabla f(x)$ for any given $x \in R^n$;
2. $\mathcal{M}$ generates a sequence of points $\{x_k\}$ such that

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_{k-1})\}, \quad k \ge 1.$$

Clearly, gradient method is a special case.

**Theorem 11.5** *For any $k$, $1 \le k \le \frac{1}{2}(n-1)$, and any $x_0 \in R^n$, there exists $f \in \mathcal{F}_L^{\infty,1}(R^n)$ such that for any first order method $\mathcal{M}$ described above there hold*

$$f(x_k) - f(x^*) \ge \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2} \quad \textit{and} \quad \|x_k - x^*\|^2 \ge \frac{1}{8}\|x_0 - x^*\|^2,$$

*where $f(x^*) = \min_{x\in R^n} f(x)$.*

Although these bounds hold only for $1 \le k \le (n-1)/2$, they describe the potential performance of first order methods on the initial stage, and they warn us that without stronger assumptions we cannot get better complexity for any first order numerical scheme.

Let $\mathcal{M}$ be an iterative method for solving $\mathcal{P} = \{\min_{x\in R^n} f(x) | f \in \mathcal{F}_L^{1,1}(R^n)\}$. Suppose $\mathcal{M}$ satisfies

1. $\mathcal{M}$ only has access to $f(x)$ and $\nabla f(x)$ for any given $x \in R^n$;
2. $\mathcal{M}$ generates a sequence of points $\{x_k\}$ such that

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_{k-1})\}, \quad k \ge 1.$$

[2]follows from $\frac{1}{2L}\|\nabla f(x)\|^2 \le f(x) - f^*, \forall x \in R^n$. This complexity result is basically the same as for $f \in C_L^{1,1}(R^n)$.

The lower complexity bound

$$f(x_k) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}, \quad 1 \leq k \leq (n-1)/2,$$

for $\mathcal{F}_L^{\infty,1}(R^n)$ implies that the best possible upper bound for $f(x_k) - f(x^*)$ (uniformly for all $k$, irrelevant to the choice of $x_0 \in R^n$ and for all $f \in \mathcal{F}_L^{1,1}(R^n)$) cannot be better than $O(1/k^2)$.

Gradient method (e.g., with $h_k \equiv h = 1/L$) is not optimal for $f \in \mathcal{F}_L^{1,1}(R^n)$ because

$$f_k - f^* \leq \frac{L\|x_0 - x^*\|^2}{2k}, \quad \forall k \geq 1,$$

where the upper bound is only $O(1/k)$, which has a gap with $O(1/k^2)$.

**Algorithm 2 (optimal gradient method for $\mathcal{F}_L^{1,1}(R^n)$)** *1. Initialization: choose $x_0 \in R^n$, set $y_0 = x_0$.*

2. *Repeat for $k = 1, 2, \ldots$*

$$\begin{aligned} x_k &= y_{k-1} - h_k \nabla f(y_{k-1}) \\ y_k &= x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \end{aligned}$$

   *where $0 < h_k \equiv h \leq 1/L$ or determined by, e.g., backtracking.*

**Remark 11.3.1** *1. not a descent method: $f(x_k) > f(x_{k-1})$ can happen;*

2. *convergence:*

$$f(x_k) - f(x^*) \leq \frac{2\|x_0 - x^*\|^2}{\underline{h}(k+1)^2},$$

   *where $\underline{h} = h \in (0,, 1/L]$ for constant step and $\underline{h} = \min\{\hat{h}, \delta/L\}$ for backtracking line search;*

3. *convergence rate optimal: $O(1/\sqrt{\epsilon})$ iterations to reach $f_k - f^* \leq \epsilon$;*

4. *published in 1983, many variants studied until very recently.*

## 11.4 Gradient method for strongly convex Lipschitz continuous functions

**Gradient method for $\mathcal{S}_{\mu,L}^{1,1}(R^n)$.** Line search condition: step size $h_k$ satisfies

$$f(x_k - h_k g(x_k)) \leq f(x_k) - \frac{h_k}{2}\|g(x_k)\|^2.$$

Assume there exists $\underline{h} > 0$ such that

$$f_{k+1} \leq f_k - \frac{\underline{h}}{2}\|g_k\|^2.$$

Therefore,

$$f_{k+1} - f^* \leq f_k - f^* - \frac{\underline{h}}{2}\|g_k\|^2.$$

Strong convexity of $f$ implies that $f(x) - f^* \leq \frac{1}{2\mu}\|g(x)\|^2$ for all $x$. Thus,

$$f_{k+1} - f^* \leq (1 - \mu\underline{h})(f_k - f^*).$$

Note that $\mu\underline{h} < 1$ for all four types of line search rules (assume $\mu < L$). As a result, $f_k$ converges to $f^*$ as

$$f_k - f^* \le (1 - \mu\underline{h})^k (f_0 - f^*).$$

Conclusion: For $\epsilon > 0$, the number of iterations to reach $f_k - f^* \le \epsilon$ is

$$\frac{\log\left((f_0 - f^*)/\epsilon\right)}{\log\left(1 - \mu\underline{h}\right)^{-1}} \approx \frac{1}{\mu h_{\min}} \times \log\left((f_0 - f^*)/\epsilon\right).$$

($\lim_{x\to 0+} \frac{\log(1-x)^{-1}}{x} = 1$.) For $h_k \equiv 1/L = \underline{h}$, it holds that

$$f_k - f^* \le \left(1 - \frac{1}{L/\mu}\right)^k (f_0 - f^*),$$

and the number of iterations to reach $f_k - f^* \le \epsilon$ is approximately

$$\frac{L}{\mu} \times \log\left((f_0 - f^*)/\epsilon\right).$$

This is why $Q_f = L/\mu$ is referred to as the condition number of $f$.

The bound can be slightly improved:

**Theorem 11.6** *If $f \in \mathcal{S}^{1,1}_{\mu,L}(R^n)$ and $0 < h_k \equiv h \le \frac{2}{\mu+L}$, then the gradient method generates a sequence $\{x_k\}$ satisfying*

$$\|x_k - x^*\|^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|x_0 - x^*\|^2.$$

*If $h = \frac{1}{\mu+L}$, then*

$$\|x_k - x^*\| \le \left(\frac{Q_f - 1}{Q_f + 1}\right)^k \|x_0 - x^*\|,$$

$$f_k - f^* \le \frac{L}{2}\left(\frac{Q_f - 1}{Q_f + 1}\right)^{2k} \|x_0 - x^*\|^2,$$

*where $Q_f := L/\mu$. The above rate of convergence is called linear convergence.*

(Assume $\mu < L$, then $\frac{2}{\mu+L} > 1/L$ and and $\left(\frac{Q_f-1}{Q_f+1}\right)^2 < 1 - \frac{1}{Q_f} < 1$.)

**Lower complexity bounds for $\mathcal{S}^{\infty,1}_{\mu,L}(R^n)$.** Consider $\min_{x\in R^n} f(x)$, where $f \in \mathcal{S}^{\infty,1}_{\mu,L}(R^n)$, $\mu > 0$ and $Q_f = L/\mu > 1$. Suppose we solve this problem by an iterative method $\mathcal{M}$ satisfying:

1. $\mathcal{M}$ only has access to $f(x)$ and $\nabla f(x)$ for any given $x \in R^n$;
2. $\mathcal{M}$ generates a sequence of points $\{x_k\}$ such that

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_{k-1})\}, \quad k \ge 1.$$

**Theorem 11.7** *For any $x_0 \in R^n$ and any constants $L$ and $\mu$ ($L > \mu > 0$), there exists a function $f \in \mathcal{S}^{\infty,1}_{\mu,L}(R^n)$ such that for any first order method $\mathcal{M}$ described above there hold*

$$\|x_k - x^*\| \ge \left(\frac{\sqrt{Q_f} - 1}{\sqrt{Q_f} + 1}\right)^k \|x_0 - x^*\|,$$

$$f(x_k) - f(x^*) \ge \frac{\mu}{2}\left(\frac{\sqrt{Q_f} - 1}{\sqrt{Q_f} + 1}\right)^{2k} \|x_0 - x^*\|^2,$$

*where* $f(x^*) = \min_{x \in R^n} f(x)$.

**Optimal gradient method for $\mathcal{S}^{1,1}_{\mu,L}(R^n)$.** Gradient method (e.g., with $h_k \equiv h = 1/(\mu + L)$) is not optimal for $f \in \mathcal{S}^{1,1}_{\mu,L}(R^n)$ because

$$f_k - f^* \leq \frac{L}{2} \left( \frac{Q_f - 1}{Q_f + 1} \right)^{2k} \|x_0 - x^*\|^2, \quad \forall k \geq 1,$$

where the upper bound has a gap with $O\left( (\sqrt{Q_f} - 1)/(\sqrt{Q_f} + 1) \right)^{2k}$.

**Algorithm 3 (optimal gradient method)** *1. Initialization: choose* $x_0 \in R^n$ *and* $\alpha_0 \in (0,1)$. *Set* $y_0 = x_0$ *and* $q = \mu/L$.

2. *For* $k \geq 0$, *repeat*

   (a) $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$;

   (b) *compute* $\alpha_{k+1} \in (0,1)$ *from* $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$.

   (c) *set* $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ *and*

$$y_{k+1} = x_{k+1} + \beta_k (x_{k+1} - x_k).$$

**Theorem 11.8** *Let* $\gamma_0 = \frac{\alpha_0(\alpha_0 L - \mu)}{1 - \alpha_0}$ *and* $C = f(x_0) - f^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2$. *If* $\alpha_0 \geq \sqrt{\mu/L}$, *then*

$$f_k - f^* \leq C \times \min \left\{ \left( 1 - \sqrt{\frac{\mu}{L}} \right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\}.$$

**Remark 11.4.1**

- *in step 2.1 of the algorithm, line search can be incorporated;*
- *algorithmic framework can be more general;*
- *convergence rate optimal;* [3]
- *also optimal for* $f \in \mathcal{F}^{1,1}_L(R^n)$ *(set* $\mu = 0$*).*

**Summary.**

| $f$ | convergence |
|---|---|
| $C^1$ | any limit point $\bar{x}$ satisfies $\nabla f(\bar{x}) = 0$ |
| $C^{1,1}_L$ | global convergence: $\lim_{k\infty} \|g_k\| = 0$ <br> convergence rate: $\min_{0 \leq k \leq N} \|g_k\| \leq C/\sqrt{N}$ |
| | local convergence: $\|x_k - x^*\| \leq C(1 - \frac{1}{Q_f})^k$ |
| $\mathcal{F}^{1,1}_L$ | global sublinear convergence: $f_k - f^* \leq C/k$ |
| | global optimal rate: $f_k - f^* \leq C/k^2$ |
| $\mathcal{S}^{1,1}_{\mu,L}$ | global linear convergence: $f_k - f^* \leq C(1 - \frac{1}{Q_f})^k$ |
| | global optimal rate: $f_k - f^* \leq C(1 - \frac{1}{\sqrt{Q_f}})^k$ |

[3] because $1 - (x-1)^2/(x+1)^2 = O(1/x)$ as $x \to \infty$.

## 11.5 More discussions

**Gradient method for quadratic problems.** Let $A \in S^n$ and $b \in R^n$. Consider solving the following quadratic problem by gradient method

$$\min_{x \in R^n} \left\{ f(x) := \frac{1}{2} x^T A x - b^T x \right\}.$$

1. If $A$ has negative eigenvalues, then $f$ is unbounded below.
2. Suppose $A \in S_+^n$. If $A$ is not positive definite and $b$ lies in the range space of $A$, then there exist infinitely many solutions.
3. The case of interest is when $A$ is positive definite.

Suppose $A$ is positive definite, and $\lambda_1, \lambda_n$ are, resp., the largest and smallest eigenvalues.

- Optimality condition: $\nabla f(x^*) = Ax^* - b = 0$, i.e., the unique optimal solution is $x^* = A^{-1}b$.
- Steepest descent method ($g_k = Ax_k - b$):

$$\begin{aligned} h_k^* &= \frac{g_k^T g_k}{g_k^T A g_k} = \arg\min_{h \geq 0} f(x_k - h g_k) \\ x_{k+1} &= x_k - h_k^* g_k. \end{aligned}$$

- two matrix-vector multiplications per iter, $g_{k+1}^T g_k = 0$ causes zigzagging.
- Convergence: $f_{k+1} - f^* \leq \frac{(Q_f - 1)^2}{(Q_f + 1)^2} (f_k - f^*)$, where $Q_f = \lambda_1 / \lambda_n$.

**Barzilai-Borwein's gradient method.** For solving the following quadratic problem

$$\min_{x \in R^n} \left\{ f(x) := \frac{1}{2} x^T A x - b^T x \right\},$$

where $A$ is positive definite, the gradient method with BB step length iterates as follows (initial point $x_0$)

$$\begin{aligned} h_k^* &= \frac{g_k^T g_k}{g_k^T A g_k}, \\ x_{k+1} &= \begin{cases} x_k - h_k^* g_k, & k = 0; \\ x_k - h_{k-1}^* g_k, & k \geq 1. \end{cases} \end{aligned}$$

**Remark 11.5.1**
- *BB step size is closely related to quasi-Newton method;*
- *per-iteration cost: two matrix-vector multiplications;*
- *the generated sequence $\{f(x_k)\}$ is non-monotone;*
- *Convergence: there exists an integer $m > 0$ such that*

$$\|g_k\| \leq 2(Q_f - 1)^{m-1} 2^{-k/m} \|g_0\|, \quad \forall k \geq 1.$$

- *Many variants of the BB step size, including extension to minimizing non-quadratic problems (assisted by non-monotone line search).*

**Numerical illustrations.** Run the codes `demo_sd, demo_bb, compare_sd_bb` (available online) and check the performance of steepest descent and BB gradient method. Steepest descent method applied to positive definite quadratic problems:

- $f_k$ is monotonically decreasing, but $\|g_k\|$ is non-monotone;
- performance deteriorates as $Q_f$ increases;
- fast convergence at first few iterations, zigzags severely when close to solution for ill-condition problems;
- convergence can be sensitive to initial point $x_0$.

BB gradient method applied to positive definite quadratic problems:

- Neither $f_k$ nor $\|g_k\|$ is monotone;
- performance deteriorates as well as $Q_f$ increases;
- convergence is less sensitive to initial point $x_0$;
- much faster convergence than steepest descent method.

# References

[Nesterov] Yurii Nesterov, Introductory Lectures on Convex Optimization, A Basic Course.