ISEN 629: Engineering Optimization

Lecture 8

Sergiy Butenko

Industrial and Systems Engineering
Texas A&M University

Fall 2007

Strongly convex functions

- We are looking for a restriction of the functional class $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$, for which we can guarantee a reasonable rate of convergence to a unique solution of the problem

$$\min_{x \in \mathbb{R}^n} f(x), \ f \in \mathcal{F}^1(\mathbb{R}^n).$$

- Let us introduce the global non-degeneracy assumption: assume that there exists a constant $\mu > 0$ such that for any $\bar{x}$ with $f'(\bar{x}) = 0$ and for any $x \in \mathbb{R}^n$ we have

$$f(x) \geq f(\bar{x}) + \frac{1}{2}\mu\|x - \bar{x}\|^2.$$

- We can use the same reasonings as in the "optimization definition" of smooth convex functions to define the class of strongly convex functions.

Strongly convex functions

Definition

A continuously differentiable function $f(x)$ is called strongly convex on $\mathbb{R}^n$ (denoted by $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$) is there exists a constant $\mu > 0$ such that for any $x, y \in \mathbb{R}^n$ we have

$$f(y) \geq f(x) + f'(x)^T(y - x) + \frac{1}{2}\mu\|y - x\|^2.$$

Constant μ is called the convexity parameter of function f.

We will also consider the classes $\mathcal{S}_{\mu,L}^{k,l}(Q)$ with the same meaning of the indices k, l, L as for the class $\mathcal{C}_L^{k,l}(Q)$.

Strongly convex functions

Theorem (2.1.8)

If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$ and $f'(x^*) = 0$ then

$$f(x) \geq f(x^*) + \frac{1}{2}\mu\|x - x^*\|^2$$

for all $x \in \mathbb{R}^n$.

Proof: Since $f'(x^*) = 0$ and f is strongly convex, we have

$$\begin{array}{rcl} f(x) & \geq & f(x^*) + f'(x^*)^T(x - x^*) + \frac{1}{2}\mu\|x - x^*\|^2 \\ & = & f(x^*) + \frac{1}{2}\mu\|x - x^*\|^2. \qquad \square \end{array}$$

Strongly convex functions

Lemma (2.1.4)
If $f_1 \in \mathcal{S}^1_{\mu_1}(\mathbb{R}^n)$, $f_2 \in \mathcal{S}^1_{\mu_2}(\mathbb{R}^n)$ and $\alpha, \beta \geq 0$, then

$$f = \alpha f_1 + \beta f_2 \in \mathcal{S}^1_{\alpha\mu_1+\beta\mu_2}(\mathbb{R}^n).$$

Note that $\mathcal{S}^1_0(\mathbb{R}^n) \equiv \mathcal{F}^1(\mathbb{R}^n)$, thus adding a convex function to a strongly convex function yields a strongly convex function with the same convexity parameter.

Strongly convex functions

Theorem (2.1.9)
Let f be continuously differentiable. Then the inclusion $f \in \mathcal{S}^1_\mu(\mathbb{R}^n)$ is equivalent to each of the following two conditions holding for any $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$:

1. $(f'(x) - f'(y))^T(x-y) \geq \mu\|x-y\|^2$;
2. $\alpha f(x) + (1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y) + \alpha(1-\alpha)\frac{\mu}{2}\|x-y\|^2$.

Theorem (2.1.10)
If $f \in \mathcal{S}^1_\mu(\mathbb{R}^n)$, then for any x and y from $\mathbb{R}^n$ we have

$$f(y) \leq f(x) + f'(x)^T(y-x) + \frac{1}{2\mu}\|f'(x) - f'(y)\|^2;$$

$$(f'(x) - f'(y))^T(x-y) \leq \frac{1}{\mu}\|f'(x) - f'(y)\|^2.$$

Strongly convex functions

Theorem (2.1.11)
Two times continuously differentiable function f belongs to the class $\mathcal{S}^2_\mu(\mathbb{R}^n)$ if and only if for any $x \in \mathbb{R}^n$:

$$f''(x) \succeq \mu I_n.$$

Example:

1. $f(x) = \frac{1}{2}\|x\|^2 \in \mathcal{S}^2_1(\mathbb{R}^n)$ since $f''(x) = I_n$.
2. For a symmetric matrix A satisfying the condition $\mu I_n \preceq A \preceq L I_n$ we have

$$f(x) = \frac{1}{2}x^T A x + b^T x + c \in \mathcal{S}^{\infty,1}_{\mu,L}(\mathbb{R}^n) \subset \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$$

since $f''(x) = A$.

Strongly convex functions

Note that the class $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ is described by

$$(f'(x) - f'(y))^T(x-y) \geq \mu\|x-y\|^2,$$

$$\|f'(x) - f'(y)\| \leq L\|x-y\|.$$

The value $Q_f = L/\mu \geq 1$ is called the condition number of function f. The first inequality can be strengthened:

Theorem (2.1.12)
If $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, then for any $x, y \in \mathbb{R}^n$ we have

$$(f'(x) - f'(y))^T(x-y) \geq \frac{\mu L}{\mu+L}\|x-y\|^2 + \frac{1}{\mu+L}\|f'(x) - f'(y)\|^2.$$

Lower complexity bounds for $\mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^n)$

We consider the following problem class (note that $\mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^n) \subset \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$).

Model:	$\min_{x\in\mathbb{R}^n} f(x),\ f \in \mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^n),\ \mu > 0.$
Oracle:	First-order local black box.
Approximate solution:	$\bar{x}:\ f(\bar{x}) - f^* \leq \epsilon,\ \|\bar{x} - x^*\|^2 \leq \epsilon.$

Assumption: An iterative method $\mathcal{M}$ generates a sequence of test points $\{x_k\}$ such that

$$x_k \in x_0 + \mathrm{Lin}\{f'(x_0), \dots, f'(x_{k-1})\},\ k \geq 1.$$

Lower complexity bounds for $\mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^n)$

- The lower complexity bound will be given in terms of the condition number $Q_f = \frac{L}{\mu}$.
- We will consider infinite-dimensional problems ($n = \infty$).
- Consider $\mathbb{R}^\infty = l_2$, the space of all sequences $x = \{x^{(i)} : i \geq 1\}$ with finite norm

$$\|x\|^2 = \sum_{i=1}^{\infty} \left(x^{(i)}\right)^2 < \infty.$$

Lower complexity bounds for $\mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^n)$

Theorem (2.1.13)

For any $x_0 \in \mathbb{R}^\infty$ and any constants $\mu > 0, Q_f > 1$ there exists a function $f \in \mathcal{S}_{\mu,\mu Q_f}^{\infty,1}(\mathbb{R}^\infty)$ such that for any first-order method $\mathcal{M}$ we have

$$\|x_k - x^*\|^2 \geq \left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \|x_o - x^*\|^2,$$

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

where x^* is the minimum of f and $f^* = f(x^*)$.

Gradient method on $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

- Next we will look at performance of the constant-step gradient method on the class $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$.
- Recall the scheme of the gradient method:
  0. Choose $x_0 \in \mathbb{R}^n$.
  1. k-th iteration ($k \geq 0$):
     a) Compute $f(x_k), f'(x_k)$, and h_k.
     b) $x_{k+1} = x_k - h_k f'(x_k)$.
- We will analyze the simplest variant, where the step size is constant, $h_k = h > 0$.
- Similar results hold for other step-size rules as well.
- For the problem $\min_{x\in\mathbb{R}^n} f(x)$, let x^* denote an optimal solution and $f^* = f(x^*)$.

Gradient method on $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Theorem (2.1.14)

Let $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $0 < h < \frac{2}{L}$. Then the gradient method generates a sequence $\{x_k\}$ such that

$$f(x_k) - f^* \le \frac{2(f(x_0) - f^*)\|x_0 - x^*\|^2}{2\|x_0 - x^*\|^2 + kh(2 - Lh)(f(x_0) - f^*)}.$$

To choose the step size that minimizes the above upper bound, we need to maximize $\phi(h) = h(2 - Lh)$ with respect to h. We have $h^* = \frac{1}{L}$, and the corresponding bound is

$$f(x_k) - f^* \le \frac{2L(f(x_0) - f^*)\|x_0 - x^*\|^2}{2L\|x_0 - x^*\|^2 + k(f(x_0) - f^*)}.$$

Gradient method on $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Recall a property of the class $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$:

$$0 \le f(y) - f(x) - f'(x)^T(y - x) \le \frac{L}{2}\|x - y\|^2.$$

Using this inequality with $x = x^*$ and $y = x_0$, we have

$$\begin{aligned} f(x_0) - f^* &\le f'(x^*)^T(x_0 - x^*) + \tfrac{L}{2}\|x_0 - x^*\|^2 \\ &= \tfrac{L}{2}\|x_0 - x^*\|^2. \end{aligned}$$

Since our upper bound in

$$f(x_k) - f^* \le \frac{2L(f(x_0) - f^*)\|x_0 - x^*\|^2}{2L\|x_0 - x^*\|^2 + k(f(x_0) - f^*)},$$

is increasing with respect to $f(x_0) - f^*$, we have

$$f(x_k) - f^* \le \frac{2L\frac{L}{2}\|x_0 - x^*\|^2\|x_0 - x^*\|^2}{2L\|x_0 - x^*\|^2 + k\frac{L}{2}\|x_0 - x^*\|^2} = \frac{2L\|x_0 - x^*\|^2}{k + 4}.$$

Gradient method on $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Recall the lower bounds for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$:

$$f(x_k) - f^* \ge \frac{3L\|x_0 - x^*\|^2}{32(k + 1)^2}.$$

Compare to the upper bound:

$$f(x_k) - f^* \le \frac{2(f(x_0) - f^*)\|x_0 - x^*\|^2}{2\|x_0 - x^*\|^2 + kh(2 - Lh)(f(x_0) - f^*)}$$

or

$$f(x_k) - f^* \le \frac{2L\|x_0 - x^*\|^2}{k + 4}.$$

The gap is quite significant, so, the gradient method is not optimal for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$.

Gradient method on $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$

Theorem (2.1.15)

If $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ and $0 < h < \frac{2}{\mu + L}$, then the gradient method generates a sequence $\{x_k\}$ such that

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

If $h = \frac{2}{\mu + L}$ then

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

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

where $Q_f = L/\mu$.

Gradient method on $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$

Recall the lower complexity bounds for $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$:

$$\|x_k - x^*\|^2 \geq \left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \|x_o - x^*\|^2,$$

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

Compare to the upper bounds:

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

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

Again, the gradient method is not optimal for $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$.

Optimal methods

- The fact that upper bounds differ from the lower bounds by an order of magnitude, in general, does not imply that the method is not optimal (since the lower bounds may be too optimistic).
- However, we will show by constructing a method with the efficiency corresponding to the lower bound, that in our case the lower bounds are exact up to a constant factor.
- The schemes and efficiency bounds for optimal methods are based on the notion of estimate sequence.

Optimal methods

Definition

A pair of sequences $\{\phi_k(x) : k \geq 0\}$ and $\{\lambda_k : k \geq 0\}$, $\lambda_k \geq 0$, is called an estimate sequence of function $f(x)$ if

$$\lambda_k \to 0, k \to \infty,$$

and for any $x \in \mathbb{R}^n$ and all $k \geq 0$ we have

$$\phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x).$$

Optimal methods

Lemma (2.2.1)

If for some sequence $\{x_k\}$ we have

$$f(x_k) \leq \phi_k^* \equiv \min_{x\in\mathbb{R}^n} \phi_k(x),$$

then $f(x_k) - f^* \leq \lambda_k(\phi_0(x^*) - f^*) \to 0, k \to \infty.$

Proof:

$$\begin{aligned} f(x_k) &\leq \phi_k^* = \min_{x\in\mathbb{R}^n} \phi_k(x) \\ &\leq \min_{x\in\mathbb{R}^n}[(1-\lambda_k)f(x) + \lambda_k\phi_0(x)] \\ &\leq (1-\lambda_k)f(x^*) + \lambda_k\phi_0(x^*). \end{aligned}$$

- We can derive the rate of convergence for $\{f(x_k)\}$ directly from the rate of convergence for $\{\lambda_k\}$.
- How to form an estimate sequence that would satisfy the lemma's condition?

Optimal methods

Lemma (2.2.2)

Assume that:

1. $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$,
2. $\phi_0(x)$ *is an arbitrary function on* $\mathbb{R}^n$,
3. $\{y_k : k \geq 0\}$ *is an arbitrary sequence in* $\mathbb{R}^n$,
4. $\{\alpha_k : k \geq 0\} \subset (0,1)$ *is such that* $\sum_{k=0}^{\infty} \alpha_k = \infty$,
5. $\lambda_0 = 1$.

Then the pair of sequences $\{\phi_k(x) : k \geq 0\}, \{\lambda_k : k \geq 0\}$ *recursively defined by*

$$\begin{aligned} \lambda_{k+1} &= (1-\alpha_k)\lambda_k, \\ \phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k[f(y_k) + f'(y_k)^T(x-y_k) + \tfrac{\mu}{2}\|x-y_k\|^2], \end{aligned}$$

is an estimate sequence for $f(x)$.

Optimal methods

Lemma (2.2.3)

Let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2$. *Then the recursive process in the above lemma preserves the canonical form of functions* $\{\phi_k(x)\}$:

$$\phi_k(x) \equiv \phi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2,$$

where the sequences $\{\gamma_k\}$, $\{v_k\}$ *and* $\{\phi_k^*\}$ *are defined as follows:*

$$\begin{aligned} \gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu, \\ v_{k+1} &= \tfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k f'(y_k)], \\ \phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k f(y_k) - \tfrac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2 \\ &\quad + \tfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\tfrac{\mu}{2}\|y_k - v_k\|^2 + f'(y_k)^T(v_k - y_k)\right). \end{aligned}$$