

Modern Theory of 2nd-Order Methods (Dec 2019)

Lecture 2: Accelerated 2nd-order methods

Yurii Nesterov (CORE/INMA, UCLouvain)

Minicourse: January 20-23, 2020 (Munich)

Contents

Composite convex optimization problem

Composite Cubic Newton Method

Complexity bounds

Non-degeneracy for 2nd-order methods

Estimating sequences

Accelerated Cubic Newton

Problem formulation

Consider the following optimization problem in the composite form:

$$\min_{x \in \text{dom } \psi} \left\{ F(x) \stackrel{\text{def}}{=} f(x) + \psi(x) \right\},$$

where

- ▶ $f(\cdot)$ is a convex function with Lipschitz-continuous Hessian:

$$(A_2) \quad \|f''(x) - f''(y)\| \leq L_2 \|x - y\|, \quad x, y \in \text{dom } \psi,$$

- ▶ $\psi(\cdot)$ is a simple closed convex function.

Example. $\psi(\cdot)$ is an indicator function of a closed convex set.

Main inequalities: for all $x, y \in \text{dom } \psi$ we have

$$(B_2) \quad \|f'(y) - f'(x) - f''(x)(y - x)\| \leq \frac{L_2}{2} \|y - x\|^2,$$

$$(C_2) \quad f(y) \leq f(x) + \langle f'(x), y - x \rangle$$

$$+ \frac{1}{2} \langle f''(x)(y - x), y - x \rangle + \frac{L_2}{6} \|y - x\|^3.$$

Main operation

Define the composite Cubic Newton Step

$$T_M(x) = \arg \min_{y \in \text{dom } \psi} \left\{ \langle f'(x), y - x \rangle + \frac{1}{2} \langle f''(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3 + \psi(y) \right\}.$$

Assumption: $T_M(x)$ is computable.

Optimality condition: at point $T = T_M(x)$, we have

$$(D_2) \quad \begin{cases} \langle f'(x) + f''(x)(T - x) + \frac{M}{2} r_M(x)(T - x), y - T \rangle \\ + \psi(y) \geq \psi(T), \quad y \in \text{dom } \psi, \end{cases}$$

where $r_M(x) = \|T - x\|$.

NB: Denote $g_\psi(T) = - \left(f'(x) + f''(x)(T - x) + \frac{M}{2} r_M(x)(T - x) \right)$.

Then $g_\psi(T) \in \partial\psi(T)$. Therefore,

$$\boxed{F'(T) \stackrel{\text{def}}{=} f'(T) + g_\psi(T)}$$

is indeed a subgradient of function $F(\cdot)$ at T .

Main properties

If $f(T) \leq f(x) + \langle f'(x), T - x \rangle + \frac{1}{2} \langle f''(x)(T - x), T - x \rangle + \frac{M}{6} r^3$, then

$$(E_2) \quad F(x) - F(T_M(x)) \geq \frac{M}{3} r_M^3(x)$$

Proof. Substituting in (D_2) $y = x$, we get

$$\begin{aligned} \psi(x) &\geq \psi(T) + \langle f'(x), T - x \rangle + \langle f''(x)(T - x), T - x \rangle + \frac{M}{2} r^3 \\ &\geq \psi(T) + f(T) - f(x) + \frac{1}{2} \langle f''(x)(T - x), T - x \rangle + \frac{3M - M}{6} r^3 \\ &\geq F(T) - F(x) + \psi(x) + \frac{M}{3} r^3. \end{aligned} \quad \square$$

$$(F_2) \quad \|F'(T_M(x))\| \leq \frac{L_2 + M}{2} r_M^2(x)$$

Proof. Indeed,

$$\begin{aligned} \|F'(T)\| &= \|f'(T) - f'(x) - f''(x)(T - x) - \frac{M}{2} r(T - x)\| \\ &\leq \|f'(T) - f'(x) - f''(x)(T - x)\| + \frac{M}{2} r^2 \\ &\stackrel{(B_2)}{\leq} \frac{L_2 + M}{2} r^2. \end{aligned} \quad \square$$

Last property

$$\langle F'(T), x - T \rangle \geq \frac{M-L_2}{2} r_M^3(x)$$

Proof. Indeed,

$$\begin{aligned} & \langle F'(T), x - T \rangle \\ &= \langle f'(T) - f'(x) - f''(x)(T - x) - \frac{M}{2}r(T - x), x - T \rangle \\ &\stackrel{(B_2)}{\geq} \frac{M-L_2}{2} r^3. \end{aligned}$$

□

Corollary. If $M \geq L_2$, then

$$(G_2) \quad \langle F'(T), x - T \rangle \geq \frac{M-L_2}{2} \left[\frac{2}{L_2+M} \|F'(T)\| \right]^{3/2}$$

Proof: Use (F_2) .

□

Composite Cubic Newton Method

Consider the following scheme.

1. Choose $x_0 \in \mathbb{R}^n$ and $M_0 \leq L_2$.

2. **k th iteration ($k \geq 0$)**

(H_2) a) Find the smallest $i_k \geq 0$ such that for $T = T_{2^{i_k} M_k}(x_k)$ we have
 $f(T) \leq f(x_k) + \langle f'(x_k), T - x_k \rangle + \frac{1}{2} \langle f''(x)(T - x), T - x \rangle + \frac{M 2^{i_k}}{6} r^3$,
where $r = \|T - x_k\|$.

b) Define $x_{k+1} = T_{2^{i_k} M_k}(x_k)$, $M_{k+1} = \max\{M_0, \frac{1}{2} 2^{i_k} M_k\}$.

Clearly, $M_0 \leq M_k \leq 2L_2$. Therefore, by (E_2) and (F_2) we have

$$(I_2) \quad F(x_k) - F(x_{k+1}) \geq \frac{M_0}{3} \left[\frac{2}{3L_2} \|F'(x_{k+1})\| \right]^{3/2}$$

NB. $\|F'(x_k)\| \rightarrow 0$, and this is true for potentially nonsmooth function!

Complexity analysis

Assumption: $D_0 = \max_{x \in \text{dom } \psi} \{\|x - x^*\| : F(x) \leq F(x_0)\} < +\infty$.

Then $\|F'(T)\| \geq \frac{1}{D_0} \langle F'(T), T - x^* \rangle \geq \frac{1}{D_0} (F(T) - F(x^*))$.

Consequently, in view of (I_2) we have

$$F(x_k) - F(x_{k+1}) \geq \frac{M_0}{3} \left[\frac{2}{3L_2D_0} \right]^{3/2} (F(x_{k+1}) - F(x^*))^{3/2}.$$

Lemma. Let $\xi_k - \xi_{k+1} \geq \xi_{k+1}^{1+\alpha}$, $k \geq 0$, with $\alpha \in (0, 1]$. Then

$$\xi_k \leq \left[(1 + \xi_0^\alpha) \cdot \frac{1+\alpha}{\alpha k} \right]^{1/\alpha}, \quad k \geq 1$$

Rate of convergence:

$$(J_2) \quad F(x_k) - F(x^*) \leq O\left(\frac{L_2 D_0^3}{k^2}\right)$$

Number of calls of oracle

In method (H_2), for some $k \geq 0$, the number of calls of oracle can be big.

Can we bound the total number of these calls?

Note that $M_{k+1} = \max\{M_0, \frac{1}{2}2^{i_k} M_k\} \geq \frac{1}{2}2^{i_k} M_k$.

Therefore, $i_k \leq 1 + \log_2 \frac{M_{k+1}}{M_k}$.

Consequently, the total number of calls N_T after T steps is bounded:

$$\begin{aligned} N_T &= \sum_{k=0}^T (1 + i_k) \leq 2(T + 1) + \sum_{k=0}^T \log_2 \frac{M_{k+1}}{M_k} \\ &= 2(T + 1) + \log_2 \frac{M_{T+1}}{M_0} \leq 2(T + 1) + \log_2 \frac{2L_1}{M_0}. \end{aligned}$$

Thus, the average number of calls of oracle per iteration is only TWO!

Uniformly convex function

Assumption. Function $f(\cdot)$ is uniformly convex of degree three:

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{\sigma_3}{3} \|y - x\|^3$$

for all $x, y \in \text{dom } \psi$ with $\sigma_3 > 0$.

Main property. Then we have

$$F(y) \geq F(T) + \langle F'(T), y - T \rangle + \frac{\sigma_3}{3} \|y - T\|^3, \quad y \in \text{dom } \psi.$$

Minimizing both sides of this inequality in y , we get

$$(K_2) \quad F(T) - F(x^*) \leq \frac{2}{3\sqrt{\sigma_3}} \|F'(T)\|^{3/2}$$

Corollary. Let us choose $M_0 = L_2$. Then, in view of (I_2) and (K_2) we have

$$F(x_k) - F(x_{k+1}) \geq \frac{1}{3} \sqrt{\frac{2\sigma_3}{3L_2}} \left(F(x_{k+1}) - F(x^*) \right).$$

This is the linear rate of convergence, which is proportional to $\sqrt{\frac{\sigma_3}{L_2}}$.

Global non-degeneracy

Standard setting: for convex $f \in C^2(\mathbb{R}^n)$, define positive constants σ_1 and L_1 such that

$$\sigma_1 \|h\|^2 \leq \langle f''(x)h, h \rangle \leq L_1 \|h\|^2$$

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

The value $\gamma_1(f) = \frac{\sigma_1}{L_1}$ is called the *condition number* of f .

(Compatible with definitions in Linear Algebra.)

Geometric interpretation: $\frac{\langle f'(x), x-x^* \rangle}{\|f'(x)\| \cdot \|x-x^*\|} \geq \frac{2\sqrt{\gamma_1(f)}}{1+\gamma_1(f)}, x \in \mathbb{R}^n$.

Complexity: (1st-order methods)

PGM: $O\left(\frac{1}{\gamma_1(f)} \cdot \ln \frac{1}{\epsilon}\right)$, **FGM:** $O\left(\frac{1}{\sqrt{\gamma_1(f)}} \cdot \ln \frac{1}{\epsilon}\right)$.

This does not work for the 2nd-order schemes:

$$f(x_k) - f^* \leq O\left(\frac{L_2 D_0^3}{k^2}\right).$$

Global 2nd-order non-degeneracy

Assumption: for any $x, y \in \mathbb{R}^n$, function $f \in C^2(\mathbb{R}^n)$ satisfies inequalities

$$\|f''(x) - f''(y)\| \leq L_2 \|x - y\|,$$

$$f(y) - f(x) - \langle f'(x), y - x \rangle \geq \frac{1}{3} \sigma_3 \|x - y\|^3,$$

where $\sigma_3 > 0$.

Value $\gamma_2(f) = \frac{\sigma_3}{L_2} \in (0, \frac{1}{2})$ is called the 2nd-order condition number of f .

(Invariant w.r.t. addition of convex quadratic functions.)

Example: for $d(x) = \frac{1}{3} \|x\|^3$, we can prove that $\gamma_2(d) = \frac{1}{4}$.

Complexity bound: (Regularized Cubic Newton)

We have seen that

$$F(x_{k+1}) - F(x^*) \leq \frac{1}{1 + \frac{1}{3} \sqrt{\frac{2\sigma_3}{3L_2}}} (F(x_k) - F(x^*)).$$

Hence, for computing ϵ -solution, we need $O\left(\frac{1}{\sqrt{\gamma_2(f)}} \ln \frac{1}{\epsilon}\right)$ iterations.

Accelerated Newton: Cubic prox-function

Problem: $\min_{x \in \mathbb{R}^n} f(x)$,

where $f(\cdot)$ is convex and $L_2(f) < +\infty$.

Denote $d(x) = \frac{1}{3}\|x\|^3$.

Lemma. Cubic prox-function is *uniformly convex*: for all $x, y \in \mathbb{R}^n$,

$$\langle d'(x) - d'(y), x - y \rangle \geq \frac{1}{2}\|x - y\|^3,$$

$$d(x) - d(y) - \langle d'(y), x - y \rangle \geq \frac{1}{6}\|x - y\|^3.$$

Moreover, its Hessian is Lipschitz continuous:

$$\|d''(x) - d''(y)\| \leq 2\|x - y\|, \quad x, y \in \mathbb{R}^n.$$

Remark. In our constructions, we are going to use $d(\cdot)$ instead of the standard *strongly convex* prox-functions.

Linear Estimating Functions

We recursively update the following sequences.

- ▶ Sequence of estimating functions

$$\psi_k(x) = \ell_k(x) + d(x - x_0), \quad 0 \geq 1,$$

where $\ell_k(\cdot)$ are linear functions ($\ell_0(\cdot) \equiv 0$).

- ▶ A minimizing sequence $\{x_k\}_{k=1}^{\infty}$.
- ▶ A sequence of scaling parameters $\{A_k\}_{k=1}^{\infty}$:

$$A_0 = 0, \quad A_{k+1} = A_k + a_{k+1}, \quad k \geq 0.$$

These objects for all $k \geq 0$ satisfy the following relations:

$$(L_2) \quad A_k f(x_k) \leq \psi_k^* \equiv \min_x \psi_k(x),$$

$$(M_2) \quad \psi_k(x) \leq A_k f(x) + d(x - x_0), \quad \forall x \in \mathbb{R}^n.$$

From these relations, we have $A_k(f(x_k) - f(x^*)) \leq d(x^* - x_0)$.

For $k = 0$, they are satisfied.

Complexity analysis

Denote $v_k = \arg \min_x \psi_k(x)$.

For some $a_{k+1} > 0$ and $M = 2L_2$, define

$$\tau_k = \frac{a_{k+1}}{A_k + a_{k+1}} \in (0, 1],$$

$$y_k = (1 - \tau_k)x_k + \tau_k v_k,$$

$$x_{k+1} = T_M(y_k),$$

$$\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle].$$

The last recursion implies (M_2) for all $k \geq 0$.

It remains to ensure (L_2) .

Justification of the method

Assume that (L_2) is valid for some $k \geq 0$. Then

$$\begin{aligned}\psi_{k+1}^* &= \min_x \left\{ \psi_k(x) + a_{k+1}(f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle) \right\} \\ &\geq \min_x \left\{ \psi_k^* + \frac{1}{6} \|x - v_k\|^3 + a_{k+1}(f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle) \right\} \\ &\geq \min_x \left\{ A_k f(x_k) + \frac{1}{6} \|x - v_k\|^3 + a_{k+1}(f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle) \right\} \\ &\geq \min_x \left\{ A_{k+1} f(x_{k+1}) + \frac{1}{6} \|x - v_k\|^3 + \langle f'(x_{k+1}), a_{k+1}(x - x_{k+1}) + A_k(x_k - x_{k+1}) \rangle \right\} \\ &= A_{k+1} f(x_{k+1}) - \frac{2}{3} \sqrt{2} \left(a_{k+1} \|f'(x_{k+1})\| \right)^{3/2} + A_{k+1} \langle f'(x_{k+1}), y_k - x_{k+1} \rangle.\end{aligned}$$

In view of (G_2) , we have $\langle f'(x_{k+1}), y_k - x_{k+1} \rangle \geq \frac{L_2}{2} \left[\frac{2}{3L_2} \|f'(T)\| \right]^{3/2}$.

This gives us the equation for a_{k+1} :

$$a_{k+1}^{3/2} = \frac{A_k + a_{k+1}}{2\sqrt{3}L_2}$$

Rate of convergence

Let the sequence $\{A_k\}_{k \geq 0}$ be defined by the following recursion

$$A_0 = 0, \quad a_{k+1}^{3/2} = \gamma(A_k + a_{k+1}),$$

$$A_{k+1} = A_k + a_{k+1}, \quad k \geq 0,$$

where $\gamma > 0$. Let us estimate from below its rate of growth.

Since function $\tau^{1/3}$ is concave for $\tau \geq 0$, we have

$$\begin{aligned} A_{k+1}^{1/3} - A_k^{1/3} &\geq \frac{1}{3} A_{k+1}^{-2/3} (A_{k+1} - A_k) \\ &= \frac{1}{3} A_{k+1}^{-2/3} (\gamma A_{k+1})^{2/3} = \frac{1}{3} \gamma^{2/3}. \end{aligned}$$

Thus, we have proved that $A_k \geq \gamma^2 \left(\frac{k}{3}\right)^3$.

in our case, $\gamma = \frac{1}{2\sqrt{3L_2}}$. Hence,

$$A_k \geq \frac{1}{12L_2} \left(\frac{k}{3}\right)^3$$

Accelerated CNM

Initialization: Choose $x_0 \in \mathbb{R}^n$. Define $\psi_0(x) = d(x - x_0)$.

Iteration k , ($k \geq 0$): Compute $v_k = \arg \min_{x \in \mathbb{R}^n} \psi_k(x)$,

and $a_{k+1} > 0$ from the equation $a_{k+1}^{3/2} = \frac{A_k + a_{k+1}}{2\sqrt{3}L_2}$. Set $A_{k+1} = A_k + a_{k+1}$.

Choose $y_k = \frac{A_k}{A_{k+1}}x_k + \frac{a_{k+1}}{A_{k+1}}v_k$, and compute $x_{k+1} = T_{2L_2}(y_k)$.

Update $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle]$.

Rate of convergence:
$$f(x_k) - f^* \leq 4L_2 \left(\frac{3}{k}\right)^3 \|x_0 - x^*\|^3$$

Remark: For updating $\psi_k(x)$, we need to update only one vector:

$$s_0 = 0, \quad s_{k+1} = s_k + a_{k+1}f'(x_{k+1}), \quad k \geq 0.$$

Then v_k can be computed by an explicit expression.

Open questions

1. Problem classes.
2. Non-degenerate problems: geometric interpretation?
3. Complexity of strongly convex functions.
(1st-order schemes?)
4. Consequences for polynomial-time methods.