Superfast second-order methods for Unconstrained Convex Optimization

Yurii Nesterov, CORE/INMA

UCLouvain, Belgium

One World Optimization Seminar (OWOS)

June 29, 2020

Recent developments: Tensor Methods

Problem: $\min_{x \in \mathbb{E}} f(x)$ where $f(\cdot)$ is a differentiable function on $\mathbb{E}$.

Taylor approximation:

$$\Omega_{x,p}(y) = f(x) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(x)[y-x]^k, \quad y \in \mathbb{E},$$

where $D^k f(x)[h]^k$ is the kth-order derivative of $f(\cdot)$ at $x \in \mathbb{E}$ along direction $h \in \mathbb{E}$.

Lipschitz continuity $\|D^p f(x) - D^p f(y)\| \le L_p \|x - y\|$ $\quad x, y \in \mathbb{E}$,

where the norm $\|\cdot\|$ is Euclidean and $p \ge 1$.

Augmented Taylor approximation:

$$\hat{\Omega}_{x,p,H}(y) = \Omega_{x,p}(y) + \frac{H}{(p+1)!}\|y - x\|^{p+1}, \; y \in \mathbb{E}.$$

Main property: $f(y) \le \hat{\Omega}_{x,p,L_p}(y)$ $\quad$ for all $y \in \mathbb{E}$.

Implementability ($p \geq 1$)

Th. (N.2019) If $f(\cdot)$ is convex and $H \geq pL_p$, then $\hat{\Omega}_{x,p,H}(\cdot)$ is *convex*.

NB: For $p = 3$, function $\tau^3 + H\tau^4$, $\tau \in \mathbb{R}$, is *never* convex.

Corollary. The point $\boxed{T_{p,H}(x) = \arg\min\limits_{y \in \mathbb{E}} \hat{\Omega}_{x,p,H}(y)}$ is computable.

Basic Tensor Method: $\boxed{x_{k+1} = T_{p,H}(x_k)}$ Convergence: $O(k^{-p})$.

Accelerated Tensor Methods. Convergence: $O(k^{-(p+1)})$.

(Baes 2009, N.2019. Tool: Estimating sequences.)

Extensions (Monteiro, Svaiter (2014) for $p = 2$) $O(k^{-(3p+1)/2})$.

NB: Very expensive line search (Bubeck, Jiang, Lee, Li, Sidford (2019), Gasnikov, Gorbunov, Kovalev, Mohhamed, Chernousova (2019)).

Maximal rate (Agarwal, Hazan (2017), Arjevani, Shamir, Shiff (2017))

$$O(k^{-(3p+1)/2}) : \quad p = 2 \Rightarrow O(k^{-7/2}), \quad p = 3 \Rightarrow O(k^{-5}).$$

Main difficulty: Implementation of Tensor Step.

Accelerated 3rd-order method (N.2019)

Assumption: $\|D^3 f(x) - D^3 f(y)\| \leq L_3 \|x - y\|$, $x, y \in \mathbb{E}$.

Augmented Taylor Polynomial:

$$\hat{\Omega}_{x,p,H}(h) = f(x) + \langle f'(x), h \rangle + \tfrac{1}{2}\langle f''(x)h, h\rangle + \tfrac{1}{6} D^3 f(x)[h]^3 + \tfrac{H}{24}\|h\|^4.$$

Main Theorem: $\boxed{D^3 f(x)[h] \preceq f''(x) + \frac{L_3}{2}\|h\|^2 I}$ for all $x, h \in \mathbb{E}$,

where I is the identity matrix.

Proof: $\forall x, h \in \mathbb{E} \Rightarrow 0 \preceq f''(x - h) \preceq f''(x) - D^3 f(x)[h] + \frac{L_3}{2}\|h\|^2 I$. $\square$

Corollary: for function $\rho_x(h) = \frac{1}{2}\langle f''(x)h, h\rangle + \frac{L_3}{4}\|h\|^4$, we have

$$\left(1 - \tfrac{1}{\sqrt{2}}\right) \rho_x''(h) \preceq \hat{\Omega}''_{x,p,6L_3}(h) \preceq \left(1 + \tfrac{1}{\sqrt{2}}\right) \rho_x''(h).$$

Thus, we can use *relative non-degeneracy condition*!

(Bauschke, Bolte, Teboulle (2016), Lu, Freund, N. (2018))

Relative non-degeneracy

Convex problem: $f^* = \min_{x \in \mathbb{E}} f(x)$.

Scaling function: $\rho(\cdot)$ is strictly convex.

Relative non-degeneracy: $\mu\rho''(x) \preceq f''(x) \preceq L\rho''(x) \quad \forall x \in \mathbb{E}$.

Bregman distance: $\beta_\rho(x, y) = \rho(y) - \rho(x) - \langle \rho'(x), y - x \rangle$.

Main property: $\mu\beta_\rho(x, y) \leq \beta_f(x, y) \leq L\beta_\rho(x, y) \quad \forall x, y \in \mathbb{E}$.

Bregman-Distance Gradient Method (BDGM):

$$x_{k+1} = \arg\min_{x \in \mathbb{E}} [f(x_k) + \langle f'(x_k), x - x_k \rangle + L\beta_\rho(x_k, x)], \ k \geq 0.$$

(Nonsmooth: Beck, Teboulle, *ORLetters*(2003). Smooth: N. *MP*(2005).)

Convergence: for $\gamma = \frac{\mu}{L}$ and $k \geq 0$ we have

$$\beta_\rho(x_{k+1}, x^*) \leq (1 - \gamma)\beta_\rho(x_{k+1}, x^*) - \frac{1}{2L}(f(x_k) - f^*).$$

Our case: $\mu = 1 - \frac{1}{\sqrt{2}}, \ L = 1 + \frac{1}{\sqrt{2}}, \ \gamma = 3 - 2\sqrt{2} > \frac{1}{6}$.

Accelerated 3rd-order method

Let $x_0 \in \mathbb{E}$, $\psi_0(x) = \frac{1}{4}\|x - x_0\|^4$, $A_k = \frac{10}{7L_3}\left(\frac{2}{3}\right)^3\left(\frac{k}{4}\right)^4$, $a_{k+1} = A_{k+1} - A_k$.

Iteration $k \geq 0$: **1.** Define $v_k = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$ and $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_k}{A_{k+1}} v_k$.

2. Set $\varphi_k(h) = \langle f'(y_k), h\rangle + \frac{1}{2}\langle f''(y_k)h, h\rangle + \frac{1}{6} D^3 f(y_k)[h]^3 + \frac{6L_3}{24}\|h\|^4$,

$\rho_k(h) = \frac{1}{2}\langle f''(y_k)h, h\rangle + \frac{L_3}{4}\|h\|^4$. Set $h_{k,0} = 0$ and iterate BDGM:

$$h_{k,i+1} = \arg\min_{h \in \mathbb{E}} \left\{ \langle \varphi'_k(h_{k,i}), h - h_{k,i}\rangle + L\beta_{\rho_k}(h_{k,i}, h) \right\}, \quad i \geq 0.$$

When stop at i_k, define $x_{k+1} = y_k + h_{k,i_k}$.

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

Convergence: $O(k^{-4})$. **Question:** What is the order of this method?

NB: We use $D^3 f(y_k)[h]^2 = \lim\limits_{\tau \to 0} \frac{1}{\tau^2}[f'(y_k + \tau h) + f'(y_k - \tau h) - 2f'(y_k)]$.

WHAT ABOUT THE "LOWER BOUND" $O(k^{-7/2})$?

Implementation details

What do we need:

1. Justification of tensor methods with inexact tensor steps.

2. Justification of BDGM with inexact gradients.

What do we get:

Second-order method with the rate of convergence $O(k^{-4})$.

Complexity of iteration: $O(\ln \frac{1}{\epsilon})$ calls of oracle.

Problem class: functions with bounded fourth derivative.

Conclusion

1. Denote $M_p(f) = \sup\limits_{x \in \mathbb{E}} \| D^p f(x) \|$. Then $M_3(f) \leq \sqrt{2 M_2(f) M_4(f)}$.

Thus, there is no contradiction with the lower bound $O(k^{-7/2})$.

2. Expansion of the lower-order methods onto the field of high-order methods.

The rate of convergence is the same!

3. Old situation: Problem class $\Leftrightarrow$ Order of the method.

This is 1D-picture.

4. New situation: 2D-picture.

Parameters: Order of bounded derivative $+$ Order of the method.

We need to fill the table!

Some hints for future research

1. Functions with bounded 2rd derivative $\boxed{\geq O(k^{-2})}$

Worst function: $f_2(x) = |x^{(1)}|^2 + \sum\limits_{i=1}^{k-1} |x^{(i+1)} - x^{(i)}|^2 - x^{(1)}$.

NB: Derivatives of order $p \geq 3$ are zeros. No help from bounding them.

2. Functions with bounded 3rd derivative $\boxed{\geq O(k^{-7/2})}$

Worst function: $f_3(x) = |x^{(1)}|^3 + \sum\limits_{i=1}^{k-1} |x^{(i+1)} - x^{(i)}|^3 - x^{(1)}$.

NB: 3rd derivative is discontinuous. There is no high-order derivatives.

3. Functions with bounded 4th derivative $\boxed{\geq O(k^{-5})}$

Worst function: $f_4(x) = |x^{(1)}|^4 + \sum\limits_{i=1}^{k-1} |x^{(i+1)} - x^{(i)}|^4 - x^{(1)}$.

NB: Derivatives of order $p \geq 5$ are zeros. No help from bounding them.

Hint: Bounds (1) and (3) are indeed unimprovable.

Some references

1. Yurii Nesterov. Inexact basic tensor methods. *CORE Discussion Paper* 2019/23 (November 2019).

- Acceptable accuracy for auxiliary problem in tensor methods.
- Convergence rate $O(k^{-6})$ for FGM used inside Cubic Reg. Newton.

2. Yurii Nesterov. Superfast second-order methods for unconstrained convex optimization. *CORE Discussion Paper* 2020/07 (January 2020).

Presented at OWOS 29/06/2020.

3. Yurii Nesterov. Inexact accelerated high-order proximal-point methods. *CORE Discussion Paper* 2020/08 (February 2020).

Bi-Level Unconstrained Minimization (BLUM) based on high-order proximal-point operators. Will be presented at SIAM MDS TOMORROW.

4. Yurii Nesterov. Inexact accelerated high-order proximal-point methods with auxiliary search procedure. *CORE Discussion Paper* 2020/10.

2nd-order implementation of 3rd-order scheme with the rate $O(k^{-5})$.

THANK YOU FOR YOUR ATTENTION!