Superfast second-order methods for Unconstrained Convex Optimization

Yurii Nesterov, CORE/INMA

(UCLouvain, Belgium)

Optimization Without Borders
(Sirius, Rissia)

Online presentation, July 12, 2021

Very old story: Proximal-Point Method

Proximal approximation for $f(\cdot)$: Moreau-Yosida regularization

$$\varphi_\lambda(x) = \min_y \left\{ f(y) + \tfrac{1}{2\lambda}\|y - x\|^2 \right\}, \quad \lambda > 0,$$

where the norm is Euclidean: $\|x\| = \langle Bx, x\rangle^{1/2}$, $B \succ 0$.

(Moreau (1965), Yosida (1980)).

Transformed into a method by Martinet (1978):

$$x_{k+1} = \arg\min_{y \in \mathbb{E}} \left\{ f(y) + \tfrac{1}{2\lambda}\|y - x_k\|^2 \right\}, \; k \geq 0.$$

Convergence rate: $O(k^{-1})$.

Remarks

- Not better than the usual Gradient Method.
- Much more difficult iteration.
- 2^{nd} birth: link to Augmented Lagrangian (Rockafellar 1976).
- Accelerated by Guller (1992) using the Fast Gradient Technique. (Sensitive to inaccuracy?)
- Extensions onto the entropy-like distances (Teboulle, Iusem, Svaiter (1992-1994))

High-order Proximal Points

Problem: $\boxed{f^* = \min\limits_{x \in \mathbb{E}} \; f(x)}$ where $f(\cdot)$ is a differentiable closed convex function. Denote by x^* its optimal solution.

For $p \geq 1$, denote $d_{p+1}(x) = \frac{1}{p+1} \| x \|^{p+1}$.

Proximal-point operator of order $p \geq 1$: Choose $H > 0$ and compute

$$\mathrm{prox}^p_{f/H}(\bar{x}) = \arg\min_{x \in \mathbb{E}} \left\{ f^p_{\bar{x},H}(x) \equiv f(x) + H d_{p+1}(x - \bar{x}) \right\}.$$

Hint. Let $T = \mathrm{prox}^p_{f/H}(\bar{x})$. Then

$$f'(T) + H \| T - \bar{x} \|^{p-1} B(T - \bar{x}) = 0,$$

and $f(\bar{x}) - f(T) \geq \langle f'(T), \bar{x} - T \rangle = H \| T - \bar{x} \|^{p+1} = H^{-\frac{1}{p}} \| f'(T) \|_*^{\frac{p+1}{p}}$.

Since $f(T) - f^* \leq R \| f'(T) \|_*$, we get $f(\bar{x}) - f(T) \geq \mathrm{c}(f(T) - f^*)^{\frac{p+1}{p}}$.

NB: • This is the rate $O(k^{-p})$, with $p \geq 1$.

• The classical proximal-point algorithm is of the 1st order.

• No assumptions on $f(\cdot)$ except convexity!

Implementable versions

Inexact iteration:

$$(*) \quad T \in \mathcal{A}^p_H(\bar{x}, \beta) = \left\{ x \in \mathbb{E} : \; \|\nabla f^p_{\bar{x},H}(x)\|_* \leq \beta \|\nabla f(x)\|_* \right\},$$

where $\beta \in [0, 1)$ is a tolerance parameter.

Basic method: $x_{k+1} \in \mathcal{A}^p_H(x_k, \beta), k \geq 0.$ **Convergence** $\boxed{O(k^{-p})}$

Accelerated method. Choose $x_0 \in \mathbb{E}$, $\beta \in [0, \frac{1}{p}]$, $H > 0$, and $\psi_0(x) = d_{p+1}(x - x_0)$. Define $A_k = \frac{2(1-\beta)}{H} \left(\frac{k}{2p+2} \right)^{p+1}$.

Iteration $k \geq 0$**.**

1. Compute $v_k = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$ and choose $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k$.

2. Compute $x_{k+1} \in \mathcal{A}^p_H(y_k, \beta)$, and update

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

Convergence $\boxed{O\left(k^{-(p+1)}\right)}$ Again, no assumptions on $f(\cdot)$ yet.

Main question: How we can ensure (*)?

Bi-Level Unconstrained Minimization (BLUM)

Upper level: Choose the order of the method $p \geq 1$ and the proximal-point scheme.

NB: Its rate of convergence does not depend on the properties of the objective.

Lower level: Choose the lower level method for computing inexact proximal-point iteration.

- The order of the lower-level scheme is not necessarily equal to p.
- The complexity of the auxiliary problem depends on the properties of the objective function.

Recent developments: Tensor Methods

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

Taylor approximation:

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

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

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

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

Augmented Taylor approximation:

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

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

NB: The minimum of $\hat{\Omega}_{x,p,H}(\cdot)$ belongs to $\mathcal{A}_H^p(x, \beta)$ for H big enough.

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.

Implementable 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:

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

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

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

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

Bauschke-Bolte-Teboulle(RHS, 2016), Lu-Freund,-Nesterov(LHS, 2018)

Relative non-degeneracy

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

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

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

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

$\boxed{\textbf{It is two!}}$ What about the "lower bound" $O(k^{-7/2})$?

What is the next? Line search

Augmented prox-iteration: for $p \geq 1$ define

$$\text{prox}^p_{f/H}(\bar{x}, \bar{u}) = \arg \min_{x \in \mathbb{E}, \tau \in \mathbb{R}} \left\{ f(x) + H d_{p+1}(x - \bar{x} - \tau \bar{u}) \right\} \in \mathbb{E} \times \mathbb{R}, \qquad (**)$$

where $\bar{x}, \bar{u} \in \mathbb{E}$, and $H > 0$. This is a convex problem!

Main idea: ensure $\langle f'(T), u \rangle = 0$. (Very difficult for Tensor Methods!)

Proximal-Point pth-order Method with Segment Search ($\tau \in [0,1]$)

Initialization. Choose $x_0 \in \mathbb{E}$, $H > 0$, and $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2$.

Iteration $k \geq 0$. **1.** Compute $v_k = \arg \min\limits_{x \in \mathbb{E}} \psi_k(x)$.

2. Compute $(x_{k+1}, \tau_k) = \text{prox}^p_{f/H}(x_k, v_k - x_k)$ with $\tau_k \in (0,1)$.

3. Define $y_k = x_k + \tau_k(v_k - x_k)$ and $g_k = \|f'(x_{k+1})\|_*$.

4. Define a_{k+1} by equation $\frac{a^2_{k+1}}{A_k + a_{k+1}} = \frac{g_k^{(1-p)/p}}{H^{1/p}}$. Set $A_{k+1} = A_k + a_{k+1}$.

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

Rate: $\boxed{O(k^{-(3p+1)/2})}$ **Challenge:** Implementation of (**) for $p = 3$.

Positive Answer

1. For $p = 3$, we can compute and approximate solution to (**) in polynomial time: Nesterov (January 2020) by relative smooothness.

Main features

a) We use bisection for computing an appropriate τ_k.

b) At each internal step, we compute approximately the 3$^{\text{rd}}$-order proximal-point operator, using BDGM with the prox-function defined by the Hessian at starting point.

c) The update of estimating functions is done by combination of two gradients.

2. This 2$^{\text{nd}}$-order scheme converges as k^{-5}. (Compare with $k^{-3.5}$)

3. Our results are valid for functions with bounded 4$^{\text{th}}$ derivative.

Gasnikov, Kamzolov for Moneiro-Swaiter (March 2020).

Conclusion

- Within the framework BLUM, we have two parameters:
  - the order of the upper-level scheme;
  - problem class containing the auxiliary problem.

  They are independent. We need to fill a table of complexity results.
- For two-level schemes, the expected practical efficiency is very high.

  (Small chances to meet worst-worst functions.)
- High sublinear rate: $\epsilon = 10^{-6} \approx 2^{-20}$, $\epsilon^{-1/5} = 16$, $\log_2 \frac{1}{\epsilon} = 20$.
- Many open questions:
  - Lower complexity bounds for high-order proximal-point methods.
  - Constrained/composite minimization (inexact versions).
  - Finer problem classes (strong convexity, uniform convexity).
  - Universal methods.
  - etc.

INTERESTING PROGRAM FOR THE FUTURE RESEARCH!

References

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

Implementation of 3$^{\text{rd}}$-order method with 2$^{\text{nd}}$-order oracle. Convergence $O(k^{-4})$.

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

General framework with high-order proximal point methods.

3. Yurii Nesterov. Inexact accelerated high-order proximal-point methods with auxiliary search procedure. *CORE Discussion Paper* 2020/10 (accepted by SIOPT).

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

THANK YOU FOR YOUR ATTENTION!