Generalized Nesterov's accelerated proximal gradient algorithms with convergence rate of order $o(1/k^2)$

Huynh Van Ngai

University of Quy Nhon

Variational Analysis and Optimisation Webinar
March 24, 2021

Outline

Outline

1. An overview- gradient descent and proximal point methods
2. Nesterov's accelerated gradient descent and some extensions
3. Generalized accelerated proximal gradient algorithm
4. Generalized accelerated forward-backward scheme

Gradient descent method

Consider the unconstrained optimization problem:

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function.

Gradient descent (GD) algorithm: Starting an initial point $x_0 \in \mathbb{R}^n$, GD iterates the following update:

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

Convergence

- If f is $L-$smooth (i.e., ∇f is $L-$Lipschitz for $L > 0$), and if step size is small enough ($\alpha_k \leq 2/L$), then the sequence $\{x_k\}$ converges to a stationary point (if it exists) of f. As f is convex, it converges to the global minimizer x^* of f.
- Convergence rate: $O(1/k)$, i.e., for some $c > 0$,

$$f(x_k) - f(x^*) \leq c/k.$$

- If f is not differentiable, $\nabla f(x_k)$ is replaced by a sub-gradient $x_k^* \in \partial f(x_k)$.

A little history

Gradient descent is generally attributed to Baron Augustin-Louis Cauchy (1789-1857), who first suggested it in 1847, but its convergence properties for non-linear optimization problems were first studied by Haskell Curry in 1944.

References.

- A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *C.R. Acad. Sci. Paris*, 25: 536-538, 1847.
- H.B. Curry. The method of steepest descent for non-linear minimization problems. *Quart. Appl. Math.* 2(3): 258-261, 1944.
- C. Lemaréchal. Cauchy and the Gradient Method. *Doc. Math. Extra.*, 251-254, 2012.

Gradient projection (GP) method

Consider the constrained optimization problem:

$$\begin{aligned} &\min f(x) \\ &x \in C \subseteq \mathbb{R}^n \end{aligned} \tag{2}$$

where $C \subseteq \mathbb{R}^n$ is a closed convex subset, and f is continuously differentiable.

In the constrained optimality theorem, if $\bar{x} \in C$ is a local minimum of (2), then

$$\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0, \quad \forall x \in C. \tag{3}$$

The (GP) method consists of the iteration:

$$x_{k+1} = P_C[x_k - \alpha_k \nabla f(x_k)]. \tag{4}$$

$P_C(z)$: the (unique) projection of $z \in \mathbb{R}^n$ on C.

The (GP) algorithm has been proposed by Goldstein in 1964. The same method was independently proposed by Levitin and Polyak one year later. This method is nowadays referred as Goldstein-Levitin-Polyak gradient projection method.

Refs.

- A. Goldstein. Convex programming in Hilbert space. *Bull. Amer. Math. Soc.*, 70(5), 709-710, 1964.
- E. S. Levitin and B. T. Polyak. Constrained minimization problems. *USSR Comput. Math. Math. Phys.* 6, 1-50,1966 (English transl. of paper in Zh. Vychisl. Mat. i Mat. Fiz., 6, 787-823, 1965).

Proximal point algorithm (PPA)

Consider the optimization problem (1), where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper convex function.
The *proximal point algorithm* is the following iteration:

$$x_{k+1} = \mathrm{Argmin}\left\{ f(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 : \quad x \in \mathbb{R}^n \right\} := \mathrm{prox}_{\alpha_k f}(x_k).$$

(PPA) is related closely to the celebrated Tikhonov Regularization.

Historical References

- B. Martinet. Régularisation d'inéquations variationelles par approximations successives. *Revue Francais d'Informatque et Recherche Opérationelle,* 1970.
- T. R. Rockafellar. Monotone operators and the proximal point algorithm, *SIAM Journal on Control and Optimization,* 14(5), 877-898, 1976.

Composite convex optimization Models

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \Phi(x) \tag{5}$$

- f : convex and smooth
- Φ : convex (may be not differentiable)

Examples

- l_1 −regularization optimization

$$\min_{x\in\mathbb{R}^n} f(x) + \|x\|_1$$

- Nuclear norm regularization optimization

$$\min_{X} f(X) + \|X\|_*$$

Proximal Gradient Algorithm (PGA)

$$x_{k+1} = \mathrm{prox}_{\alpha_k \Phi}(x_k - \alpha_k \nabla f(x_k)), \quad k = 0, 1, ...,$$

where $\mathrm{prox}_{\alpha_k \Phi}(x)$ is the proximal operator:

$$\mathrm{prox}_{\alpha_k \Phi}(x) = \mathrm{Argmin}\left\{ \Phi(y) + \frac{1}{2\alpha_k}\|y - x\|^2 : \quad y \in \mathbb{R}^n \right\}.$$

- alternating between gradient updates on f and proximal minimization on Φ
- useful when prox_{Φ} is inexpensive

Refs.

- P. Lions and B. Mercier, Splitting algorithms for the sum of two nonlinear operators, *SIAM J. on Numerical Analysis,* 16, 964-979, 1979.
- S. Boyd and L. Vandenberghe, Convex Optimization, *Cambridge University Press,* 2004.

Nesterov's accelerated gradient algorithm (NAGA)

Consider the unconstrained convex optimization problem: $\min_{x\in\mathbb{R}^n} f(x)$ (f is convex and $L-$smooth).
(NAGA) which was proposed by Nesterov in 1983, iterates the following update scheme:

- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
- $y_{k+1} = (1-\gamma_{k+1})x_{k+1} + \gamma_k x_k$
- $\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}$
- $\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$

(NAGA) has convergence rate $O(1/k^2)$

Theorem

If f is convex and $L-$smooth, and it exists a minimizer x^* of f, then the sequence $\{f(x_k)\}$ produced by (NAGA) converges to the minimum value $f(x^*)$ with rate $O(1/k^2)$, namely,

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

Fast iterative shrinkage-thresholding algorithm (FISTA)-Beck & Teboulle(2009)

Consider again the composite convex optimization (5):

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \Phi(x),$$

where f is $L-$smooth on $\mathbb{R}^n$, and Φ is l.s.c convex, possibly non-differentiable.

(FISTA) Algorithm

- $x_{k+1} = \text{prox}_{L^{-1}\Phi}(y_k - L^{-1}\nabla f(y_k))$.
- $y_{k+1} = (1 - \gamma_{k+1})x_{k+1} + \gamma_k x_k$,
- $\gamma_k = \dfrac{1 - \lambda_k}{\lambda_{k+1}}$,
- $\lambda_k = \dfrac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2}$.

Convergence result:

$$F(x_k) - F(x^*) = O\left(\frac{1}{k^2}\right).$$

Forward-Backward Algorithm- Attouch & Peypouquet (2016)

$$\begin{cases} y_k = x_k + \dfrac{k-1}{k+\alpha-1}(x_k - x_{k-1}), \\ x_{k+1} = \text{prox}_{\kappa\Phi}(y_k - \kappa\nabla f(y_k)), \end{cases} \tag{6}$$

where $\alpha > 3$ and $\kappa \leq L^{-1}$.

Convergence: (Attouch& Peypouquet-(SIOPT 2016)) *If f is $L-$smooth on the whole space then*

- $F(x_k) - F(x^*) = o(1/k^2)$;
- *the whole sequence (x_k) converges (weakly, when the space under consideration is a Hilbert space of infinite dimension) to a minimizer x^*.*

References

- Nesterov Y., A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, *Doklady AN SSSR* (translated as Soviet Math.Docl.) **269**, 543-547, (1983).
- Nesterov Y., Smooth minimization of non-smooth functions, *Math. Program. Ser A.,* **103**, 127-152 (2005).
- Nesterov Y., Gradient methods for minimization composite objective functions, *Math. Prog. Ser. A,* **140**, 125-161, (2013).
- Nesterov Y., Universal Gradient methods for convex optimization problems, *Math. Prog. Ser. A,* **152**, 381-404, (2015).
- Beck A., Teboulle M., A fast iterative shinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imag. Sci.,* **2**(1), 183-202, (2009).
- Attouch H., Peypouquet J., The rate of convergence of Nesterov's accelerated forward-backward method is actually faster $O(1/k^2)$, *SIAM J. Optim.,* **26**(3), 1824-1834, (2016).

Problem

Consider again the *composition convex* optimization problem of the form

$$\min\{F(x) := f(x) + \Phi(x) : \quad x \in \mathbb{R}^n\}. \tag{7}$$

In what follows we make use of the following assumptions:

(A1) The optimal solution set of problem (7) is nonempty.

(A2) The function $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper lower semicontinuous convex; the function $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function such that its gradient ∇f is $L-$Lipschitz (for some $L > 0$) on $\mathrm{dom}\,\Phi$.

Lower support functions

Definition

For a convex function $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and a point $z \in \mathbb{R}^n$. A convex function $\Psi_z := \Psi_{z,\Phi} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called a lower support function to Φ at z if $\Psi_z \leq \Phi$ and $\Psi_z(z) = \Phi(z)$.

Obviously, the usual two lower support functions of a convex function Φ, at a point z : the first is itself Φ, and the second is the linear function

$$\Psi_z(x) := \Phi(z) + \langle z^*, x - z \rangle, \; x \in \mathbb{R}^n,$$

where $z^* \in \partial\Phi(z)$, when Φ is subdifferentiable at z.

uniformly convex functions

A function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called $p-$uniformly convex with parameter μ, for some $\mu \geq 0$, $p \geq 2$, or called $(\mu, p)-$uniformly convex if for all $x, y \in \mathbb{R}^n$, $\lambda \in [0, 1]$ one has

$$\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y) - \frac{\mu}{p}\lambda(1-\lambda)\|x-y\|^p.$$

When $p = 2$, the function φ is called strongly convex (with parameter μ.) Note that if φ is $(\mu, p)-$uniformly convex, then for all $x, y \in \mathbb{R}^n$, all $x^* \in \partial\varphi(x)$, one has

$$\langle x^*, y - x\rangle \leq \varphi(y) - \varphi(x) - \frac{\mu}{p}\|y - x\|^p. \tag{8}$$

Generalized accelerated proximal gradient algorithm (GAPGA)

Parameters.

- Given a $\rho-$strongly convex function $h : \mathbb{R}^n \to \mathbb{R}$ ($\rho > 0$) which attains minimum at $y_0 \in \mathbb{R}^n$:

$$h(y) \geq h(y_0) + \frac{\rho}{2}\|y - y_0\|^2, \ \ \forall y \in \mathbb{R}^n, \tag{9}$$

- parameters $C, \mu > 0$, $0 < \kappa \leq 1/L$,
- a sequence of positive reals $\{\alpha_k\}$; sequences of nonnegative reals $\{\beta_k\}$, and $\{\gamma_k\}$ as in Section 2. Set

$$A_k = \sum_{i=0}^{k} \alpha_k, \ B_k = \sum_{i=0}^{k} \beta_k,$$

and also assume that $A_k \geq B_k$ for all $k \in \mathbb{N}$, and denote $A_{-1} = B_{-1} = 0$.

Algorithm 1(GAPGA1)

`Initialization:` Initial data: y^0 as in (9). Set $k=0$.

`Repeat:` For $k=0,1,...,$

1. Find

$$
\begin{aligned}
x_k &= \operatorname{argmin}\left\{\Phi(y)+\langle\nabla f(y_k), y-y_k\rangle+\frac{1}{2\kappa}\|y-y_k\|^2 : \quad y\in\mathbb{R}^n\right\}\\
&= \operatorname{prox}_{\kappa\Phi}\left(y_k-\kappa\nabla f(y_k)\right).
\end{aligned}
\tag{10}
$$

2. Find

$$
\begin{aligned}
z_k &= \operatorname{argmin}_{x\in\mathbb{R}^n}\{Ch(x)+\sum_{i=0}^{k-1}\alpha_i[f(y_i)+\langle\nabla f(y_i), x-y_i\rangle+\Psi_{z_i}(x)\\
&+\frac{1}{2}\mu\gamma_i\|x-y_i\|^2]+\alpha_k[f(y_k)+\langle\nabla f(y_k), x-y_k\rangle\\
&+\Phi(x)+\frac{1}{2}\mu\gamma_k\|x-y_k\|^2]\}
\end{aligned}
\tag{11}
$$

Algorithm 1-continued

3. Set Ψ_{z_k} is a support function to Φ at z_k such that

$$\min_{x\in\mathbb{R}^n}\{Ch(x)+\sum_{i=0}^{k-1}\alpha_i[f(y_i)+\langle\nabla f(y_i),x-y_i\rangle$$
$$+\Psi_{z_i}(x)+\frac{1}{2}\mu\gamma_i\|x-y_i\|^2]$$
$$+\alpha_k[f(y_k)+\langle\nabla f(y_k),x-y_k\rangle+\Phi(x)+\frac{1}{2}\mu\gamma_k\|x-y_k\|^2]\}$$
$$=\min_{x\in\mathbb{R}^n}\{Ch(x)+\sum_{i=0}^{k}\alpha_i[f(y_i)+\langle\nabla f(y_i),x-y_i\rangle$$
$$+\Psi_{z_i}(x)+\frac{1}{2}\mu\gamma_i\|x-y_i\|^2]\}. \tag{12}$$

4. Set

$$\tau_k:=\frac{\alpha_{k+1}}{A_{k+1}-B_k},\quad y_{k+1}=\tau_k z_k+(1-\tau_k)x_k.$$

Remarks

- In Nesterov's original accelerated schemes, $\tau_k := \dfrac{\alpha_{k+1}}{A_{k+1}}$. which is a particular case of Algorithm 1 with $\beta_k := 0$, $k \in \mathbb{N}$.
- In Step 3 of Algorithm 1, we can take $\Psi_{z_k} = \Phi$. If we set $\Psi_{z_k} = \Phi$, for all $k \in \mathbb{N}$, Algorithm 1 gives a generalized variant of Nesterov's accelerated dual averaging algorithm.
- An another way to choose Ψ_{z_k} is as follows. As in Step 2, z_k is a minimizer of the convex function in the right hand of (11), then there is $z_k^* \in \partial\Phi(z_k)$ such that

$$\begin{aligned} 0 \in \quad & C\partial h(z_k) + \sum_{i=0}^{k-1} \alpha_i [\nabla f(y_i) + \partial \Psi_{z_i}(z_k)] \\ & + \alpha_k [\nabla f(y_k) + z_k^*] + \mu \sum_{i=0}^{k} \alpha_i \gamma_i (z_i - y_i). \end{aligned} \tag{13}$$

Then the support function

$$\Psi_{z_k}(x) := \langle z_k^*, x - z_k \rangle + \Phi(z_k), \;\; x \in \mathbb{R}^n, \tag{14}$$

Remarks-continued

- When $h(x) := \frac{1}{2}\|x - y_0\|^2$, and for all $k \in \mathbb{N}$, the support function Ψ_{z_k} is defined by (14) for all $k \in \mathbb{N}$, then

$$\begin{aligned} z_{k+1} &= \operatorname{prox}_{\frac{\alpha_{k+1}}{C+\mu\alpha_{k+1}\gamma_{k+1}}\Phi}\left[\frac{1}{C+\mu\alpha_{k+1}\gamma_{k+1}}W_{k+1}\right]; \\ W_{k+1} &:= (C+\mu\alpha_k\gamma_k)z_k - \mu\alpha_k\gamma_k y_k + \alpha_{k+1}\gamma_{k+1}y_{k+1} - \alpha_{k+1}\nabla f(y_{k+1}). \end{aligned} \tag{15}$$

- In particular, when $\mu = 0$, the sequence $\{z_k\}$ is defined recurrently by

$$z_{k+1} = \operatorname{prox}_{\frac{\alpha_{k+1}}{C}\Phi}\left[z_k - \frac{\alpha_{k+1}}{C}\nabla f(y_{k+1})\right]. \tag{16}$$

This is exactly the accelerated scheme of the proximal gradient methods.

Convergence analysis

Define the *estimate* function:

$$F_k(x) = Ch(x) + \sum_{i=0}^{k-1} \alpha_i [f(y_i) + \langle \nabla f(y_i), x - y_i \rangle + \Psi_{z_i}(x) + \frac{1}{2}\mu\gamma_i \|x - y_i\|^2]$$
$$+\alpha_k [f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \Phi(x) + \frac{1}{2}\mu\gamma_k \|x - y_k\|^2], \quad x \in \mathbb{R}^n. \tag{17}$$

The following theorem gives an estimate for function values $f(x_k) + \Phi(x_k)$, and it is crucial to derive the subsequent convergence rates.

Theorem

Let $\{x_k\}$ and $\{y_k\}$ be sequences generated by Algorithm 1. Suppose that $\kappa \le 1/L$ and the sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\gamma_k\}$ satisfy the condition

$$\left(C\rho + \mu \sum_{i=0}^{k-1} \alpha_i \gamma_i \right) (A_k - B_{k-1}) \ge \alpha_k^2/\kappa, \quad \text{for all } k \in \mathbb{N}. \tag{18}$$

Then one has for all $k \in \mathbb{N}$,

$$\begin{aligned} &\sum_{i=0}^{k} \beta_i [f(x_i) + \Phi(x_i)] + (A_k - B_k)[f(x_k) + \Phi(x_k)] \\ &+ \frac{1}{2}(1/\kappa - L) \sum_{i=0}^{k} (A_i - B_{i-1}) \|x_i - y_i\|^2 \le \min_{x \in \mathbb{R}^n} F_k(x). \end{aligned} \tag{19}$$

Theorem-continued

Moreover, if f is $\mu-$*strong convex, then (19) holds if* $\gamma_k = 1$*,* $k \in \mathbb{N}$*, and the sequences* $\{\alpha_k\}, \{\beta_k\}$ *verifying the condition*

$$\left(C\rho + \mu\sum_{i=0}^{k-1}\alpha_i\right)(A_k - B_{k-1}) \geq \alpha_k^2(\kappa^{-1} - \mu), \quad \text{for all } k \in \mathbb{N}. \tag{20}$$

Convergence 1

Theorem

In Algorithm 1, pick $\alpha_k = k$; $\beta_k = k/2$; $\mu = 0$, *and* $C, \kappa > 0$ *such that* $C\rho \geq \kappa^{-1} \geq L$. *Then condition (18) is satisfied, and therefore for a minimizer* x^* *of problem (7), one has*

$$\lim_{k\to\infty} \min_{i=[k/2],\ldots,k} k^2[f(x_i) + \Phi(x_i) - f(x^*) - \Phi(x^*)] = 0, \tag{21}$$

where $[k/2]$ *stands for the integer part of* $k/2$. *Therefore if* $\{f(x_k) + \Phi(x_k)\}$ *is a decreasing sequence, then*

$$\lim_{k\to\infty} k^2[f(x_k) + \Phi(x_k) - f(x^*) - \Phi(x^*)] = 0. \tag{22}$$

Convergence 2- Uniformly convex case

Theorem

Let f is $(\mu, p)-$uniformly convex with $p > 2$, $\mu > 0$. Let $0 < \kappa \leq L^{-1}$, and $C, \rho, m > 0$ such that

$$m\mu\kappa \geq \begin{cases} 2^{\frac{4}{p-2}} \dfrac{8p}{(p-2)^2} & \text{if } 2 < p < 6, \\ \dfrac{8p}{(p-2)^2} & \text{if } p \geq 6; \end{cases} \tag{23}$$

$$C\rho \geq \begin{cases} \kappa^{-1} & \text{if } 2 < p < 6, \\ \dfrac{p-2}{4} m\mu & \text{if } p \geq 6. \end{cases} \tag{24}$$

Convergence 2-continued

In Algorithm 1, set $\alpha_k = k^{\frac{p+2}{p-2}}$, $\beta_k = 0$, *and* $\gamma_0 = 0$, $\gamma_k = mk^{-2}$ *for* $k \geq 1$. *Then for* x^* *being a minimizer of* $f + \Phi$, *one has for all* $k \in \mathbb{N}$,

$$f(x_k) + \Phi(x_k) - f(x^*) - \Phi(x^*) \leq \frac{2p}{p-2}\left(Ch(x^*) + \frac{1}{2}(p/2)^{\frac{2}{p-2}} m^{\frac{p}{p-2}} (\ln k + 1)\right) k^{-\frac{2p}{p-2}}. \tag{25}$$

Convergence 3: strongly convex case

Theorem

Let f is $\mu-$strongly convex for some $\mu > 0$, and let q, C such as (??). Then for the sequence $\{x_k\}$ generated by Algorithm 1 with sequences $\alpha_k := q^k, \beta_k = 0$, and $\gamma_k = 1$, $k \in \mathbb{N}$, and a minimizer x^* of problem (7), one has

$$f(x_k) + \Phi(x_k) - f(x^*) - \Phi(x^*) \leq \frac{(q-1)Ch(x^*)}{q^{k+1}-1}, \quad \text{for all } k \in \mathbb{N}. \quad (26)$$

Attouch-Peypouquet's accelerated forward-backward scheme

Assumption: *$f : \mathbb{R}^n \to \mathbb{R}$ is $L-$smooth on the whole space.*

[Attouch & Peypouquet -SIOPT 2016]:

$$\begin{cases} y_k = x_k + \dfrac{k-1}{k+\alpha-1}(x_k - x_{k-1}), \\ x_{k+1} = \text{prox}_{\kappa\Phi}(y_k - \kappa\nabla f(y_k)), \end{cases} \tag{27}$$

- *If $\alpha > 3$ and $\kappa \leq L^{-1}$ then the sequence of functional value $(F(x_k))$ converges to the minimum value with rate $o(1/k^2)$, and moreover the whole sequence (x_k) converges.*

Consider the operator $G_\kappa : \mathbb{R}^n \to \mathbb{R}^n$, defined by

$$G_\kappa(y) = \frac{1}{\kappa}[y - \text{prox}_{\kappa\Phi}(y - \kappa\nabla f(y))], \quad y \in \mathbb{R}^n,$$

and setting

$$z_k = \frac{k+\alpha-1}{\alpha-1} y_k - \frac{k}{\alpha-1} x_k,$$

then we can rewrite the scheme (27) as follows.

$$\begin{cases} z_{k+1} = z_k - \dfrac{\kappa(k+\alpha-1)}{\alpha-1} G_\kappa(y_k), \\ y_k = \dfrac{\alpha-1}{k+\alpha-1} z_k + \dfrac{k}{k+\alpha-1} x_k, \\ x_{k+1} = \text{prox}_{\kappa\Phi}(y_k - \kappa\nabla f(y_k)). \end{cases} \tag{28}$$

- The sequence $\{z_k\}$ in the scheme (28) can be represented equivalently

$$z_{k+1} = \text{argmin}_{x\in\mathbb{R}^n}\left\{\frac{1}{2\kappa}\|x\|^2 + \sum_{i=0}^{k}\alpha_i\langle G_\kappa(y_i), x\rangle\right\},$$

where $\alpha_i = \dfrac{i+\alpha-1}{\alpha-1}$, for $i \in \mathbb{N}$.

Regarding this representation, we will propose the generalized accelerated forward-backward algorithm.

Given a $\rho-$strongly convex function $h : \mathbb{R}^n \to \mathbb{R}$ $(\rho > 0)$; parameters $C, \mu > 0$, $0 < \kappa \leq 1/L$, and a sequence of positive reals $\{\alpha_k\}$; sequences of nonnegative reals $\{\beta_k\}$, and $\{\gamma_k\}$ as in the preceding section, and

$$A_k = \sum_{i=0}^{k}\alpha_k,\ B_k = \sum_{i=0}^{k}\beta_k,$$

with assuming $A_k \geq B_k$ for all $k \in \mathbb{N}$.

Generalized accelerated forward-backward scheme

Algorithm 2 (GAFBA) `Initialization`: Initial data: $x_0 = z_0 = y_0$. Set $k = 0$.

`Repeat`: For $k = 0, 1, ...,$

1. Set
$$\tau_k := \frac{\alpha_k}{A_k - B_{k-1}}, \quad y_k = \tau_k z_k + (1 - \tau_k) x_k.$$
2. Find
$$x_{k+1} = \text{prox}_{\kappa\Phi}(y_k - \kappa\nabla f(y_k)). \tag{29}$$
3. Set
$$G_\kappa(y_k) = \frac{1}{\kappa}[y_k - \text{prox}_{\kappa\Phi}(y_k - \kappa\nabla f(y_k))] = \frac{1}{\kappa}(y_k - x_{k+1}).$$
4. Find
$$z_{k+1} = \operatorname{argmin}\{Ch(x) + \sum_{i=0}^{k} \alpha_i[\langle G_\kappa(y_i), x - y_i\rangle + \frac{1}{2}\mu\gamma_i\|x - y_i\|^2] : \ x \in \mathbb{R}^n\} \tag{30}$$

(GAPGA) has convergence rate $o(1/k^2)$

Let $\{x_k\}$ be the sequences defined by Algorithm (GAPGA). Let x^* be a minimizer of problem (7).

Convergence results.

(i). For $\mu = 0$, and any two sequences of positive reals $\{\alpha_k\}$ and $\{\beta_k\}$ with $\alpha_k \geq \beta_k$ for $k \in \mathbb{N}$ and

$$0 < \liminf_{k\to\infty} \frac{\beta_k}{k} \leq \limsup_{k\to\infty} \frac{\alpha_k}{k} < +\infty, \quad \limsup_{k\to\infty} \frac{\beta_k}{\alpha_k} < 1,$$

then we can find $C_0 > 0$ such that for all $C \geq C_0$, one has

$$\lim_{k\to\infty} k^2 \min_{i=[k/2],\ldots,k} [f(x_i) + \Phi(x_i) - f(x^*) - \Phi(x^*)] = 0.$$

Convergence- continued...

(ii) *Suppose that f is $(\mu, p)-$uniformly convex with $\mu > 0, p > 2$. Then one has*

$$f(x_k) + \Phi(x_k) - f(x^*) - \Phi(x^*) = O\left(\frac{\ln k}{k^{2p/(p-2)}}\right).$$

(iii) *If f is $\mu-$strongly convex, then with suitable parameters $q > 1$ $C > 0$, and the sequences $\alpha_k = q^k$, $\beta_k = 0$ and $\gamma_k = 1$, for $k \in \mathbb{N}$, one has*

$$f(x_k) + \Phi(x_k) - f(x^*) - \Phi(x^*) = O\left(q^{-k}\right).$$

Special case

In Algorithm 2, set

- $h(x) := \frac{1}{2}\|x - y_0\|^2, \; x \in \mathbb{R}^n$;
- $\mu = 0$, and sequences $\{\alpha_k\}$ and $\{\beta_k\}$ satisfying the condition

$$A_k - B_{k-1} = \alpha_k^2, \;\; k \in \mathbb{N}.$$

In this special case, Algorithm 2 can be rewritten simply in the following scheme generalizing (27) by Attouch-Peypouquet:

$$\begin{cases} y_k = x_k + \frac{\alpha_{k-1} - 1}{\alpha_k}(x_k - x_{k-1}), \\ x_{k+1} = \text{prox}_{\kappa\Phi}(y_k - \kappa\nabla f(y_k)). \end{cases} \quad (31)$$

Convergence

Theorem

Consider the scheme (31). Let $\{\alpha_k\}$, $\{\beta_k\}$ be sequences of positive reals such that for some $0 < c_1, c_2 < 1$,

$$c_1\alpha_k \leq \beta_k \leq c_2\alpha_k, \quad A_k - B_{k-1} = \alpha_k^2, \quad k \in \mathbb{N}.$$

Then one has

$$\lim_{k\to\infty} k^2[f(x_k) + \Phi(x_k) - f(x^*) - \Phi(x^*)] = 0; \qquad \lim_{k\to\infty} k\|x_{k+1} - x_k\| = 0,$$

where x^* is a minimizer of problem (7). Moreover, the whole sequence $\{x_k\}$ converges to a minimizer of problem (7).

This talk is based on the paper:

- Huynh Van Ngai & Ta Anh Son, Generalized Nesterov's accelerated proximal gradient algorithms with convergence rate of order $o(1/k^2)$, *submitted* (2020).

THANKS!