Acceleration methods in Optimization

Aude Rondepierre

Institut de Mathématiques de Toulouse, INSA de Toulouse & LAAS-CNRS

CIMI-ANITI School on Optimisation - September 2th, 2021

The framework

$$\min_{x \in \mathbb{R}^N} F(x)$$

where $F : \mathbb{R}^N \to \mathbb{R}$ is a differentiable convex function admitting at least one minimizer x^*.

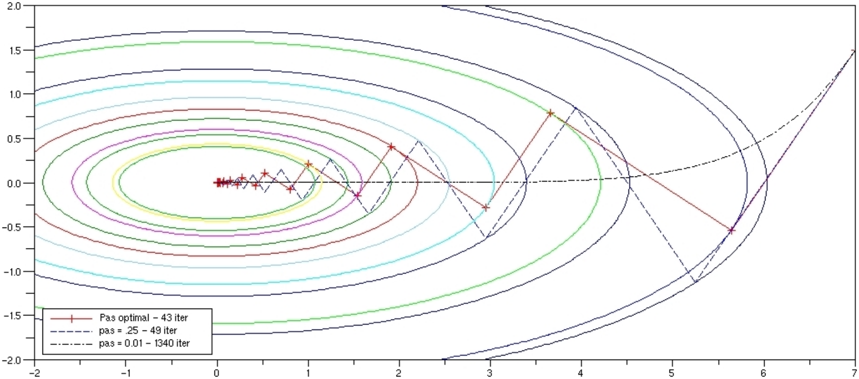
The gradient descent (GD) method

$$\left| \begin{array}{rcl} x_0 & \in & \mathbb{R}^N \\ x_{k+1} & = & x_k - s\nabla F(x_k), \;\; s > 0. \end{array} \right.$$

- A very simple algorithm, does not require second order derivative.
- Each iteration is of the order of N operations.

A quadratic example

$$\min_{(x,y)\in\mathbb{R}^2} f(x,y) = \frac{1}{2}x^2 + \frac{7}{2}y^2, \qquad d_k = -\nabla f(X_k) = \begin{pmatrix} -x_k \\ -7y_k \end{pmatrix}.$$



Gradient method is often slow ; the convergence is very dependent on scaling.

Methods with improved convergence rate

Methods with improved convergence

- Conjugate gradient method
- Accelerated gradient method
- Quasi-Newton methods

Methods with improved convergence rate

Methods with improved convergence

- Conjugate gradient method
- Accelerated gradient method
- Quasi-Newton methods

Methods with improved convergence rate

Methods with improved convergence

- Conjugate gradient method
- Accelerated gradient method
- Quasi-Newton methods

Natural extensions to composite optimization $F(x) = f(x) + g(x)$ where

- f is a convex differentiable function with a L-Lipschitz gradient
- g is a convex lsc (possibly nonsmooth but quite simple) function.

$\hookrightarrow$ Motivation: application to least square problems, LASSO:

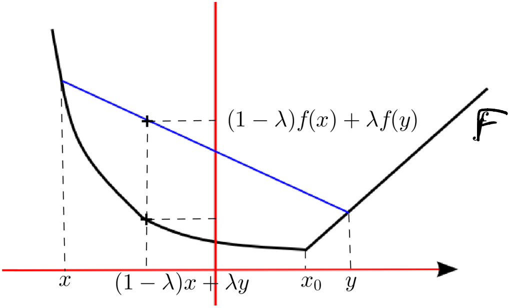
$$\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - b\|^2 + \|x\|_1.$$

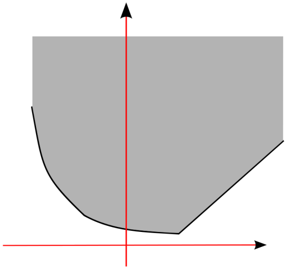
Applications in Image and Signal processing, machine learning,...

1. **Preliminaries: local geometry of convex functions**
2. **Gradient descent methods**
3. **Accelerated gradient methods**
   - The Heavy ball method
   - The Nesterov's accelerated gradient method
   - Natural extension to composite optimization
   - Some numerical experiments
4. **Improving the state of the art results with the help of the geometry**
5. **Conclusion**

A function $F : \mathbb{R}^n \to \mathbb{R}$ is said convex if:

$$\forall (x, y) \in \mathrm{dom}(F)^2,\ \forall \lambda \in [0, 1],\ F((1-\lambda)x + \lambda y) \leq (1-\lambda)F(x) + \lambda F(y).$$

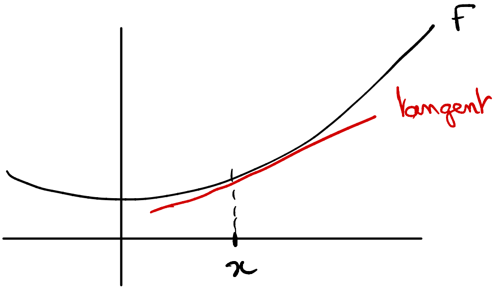




F convex iff its epigraph is convex

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. F is convex if:

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \; f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$



$\Rightarrow$ Monotonicity of the gradient:

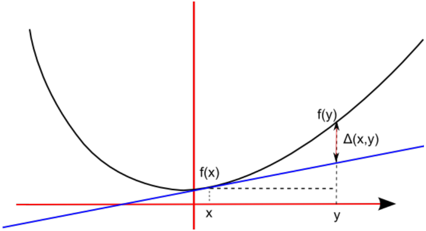
$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \; \langle \nabla F(y) - \nabla F(x), y - x \rangle \geqslant 0.$$

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and $L > 0$. The function F has a L-Lipschitz gradient iff:

$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \ \|\nabla F(x) - \nabla F(y)\| \leqslant L\|x - y\|.$$

1. Quadratic upper bound: For all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$, we have:

$$F(y) \leqslant \underbrace{F(x) + \langle \nabla F(x), y - x \rangle}_{\text{linear approximation}}) + \underbrace{\frac{L}{2}\|y - x\|^2}_{=\Delta(x,y)}$$



Geometry of convex functions

Proof of the quadratic upper bound (2/3)

$$\varphi(r) = F(x + r(y-x)) \qquad \nabla \varphi(r) = \langle \nabla F(x + r(y-x)), y - x \rangle$$

$$F(y) - F(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(r)\, dr$$

$$= \int_0^1 \langle \nabla F(x + r(y-x)), y - x \rangle \, dr$$

$$F(y) - F(x) - \langle \nabla F(x), y - x \rangle$$

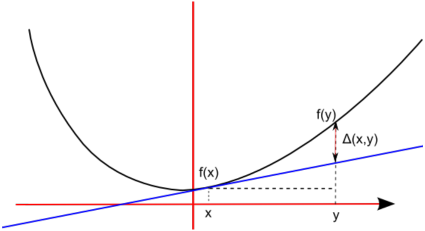
$$= \int_0^1 \langle \nabla F(x + r(y-x)) - \nabla F(x), y - x \rangle \, dr$$

$$\leq \int_0^1 \overbrace{\| \nabla F(x + r(y-x)) - \nabla F(x) \|}^{\leq L \| x + r(y-x) - x \|} \, \| y - x \| \, dr$$

$$\leq L \int_0^1 r \| y - x \|^2 \, dr = \frac{L}{2} \| y - x \|^2$$

$F : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex i.e. that there exists $\mu > 0$ such that:

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n,\ F(y) \geqslant F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2.$$



Geometry of convex functions
Strong convexity (2/2)

This class of functions satisfies a global quadratic growth condition: for any minimizer x^* we have:

$$\forall x \in \mathbb{R}^n,\ F(x) - F(x^*) \geqslant \frac{\mu}{2}\|x - x^*\|^2.$$

and:

$$\forall x \in \mathbb{R}^n,\ \|\nabla F(x)\|^2 \geqslant 2\mu(F(x) - F(x^*)).$$

Łojasiewicz property with $\theta = \frac{1}{2}$

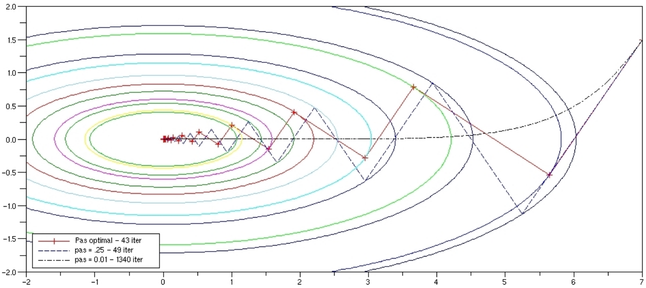
Outline

1. **Preliminaries: local geometry of convex functions**
2. **Gradient descent methods**
3. **Accelerated gradient methods**
   - The Heavy ball method
   - The Nesterov's accelerated gradient method
   - Natural extension to composite optimization
   - Some numerical experiments
4. **Improving the state of the art results with the help of the geometry**
5. **Conclusion**

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a continuously differentiable function having a L-Lipschitz gradient and admitting at least one minimizer.

$$\begin{array}{rcl} x_0 & \in & \mathbb{R}^N \\ x_{k+1} & = & x_k - s_k \nabla F(x_k),\ \ s_k > 0. \end{array}$$

with a fixed step $s_k := s > 0$ or backtracking linesearch.



Step	0.325	0.25	0.125	0.05	0.01
Nb of it.	DV	49	101	263	1340

Properties:

1. GD is a descent method when $s_k < \frac{2}{L}$ for all $k \in \mathbb{N}$, :

$$\forall k \in \mathbb{N},\ F(x_{k+1}) - F(x_k) \leqslant s_k \left(\frac{L}{2} s_k - 1\right) \|\nabla F(x_k)\|^2 \leqslant 0.$$

2. Assume that F is additionally convex and $s_k \leqslant \frac{1}{L}$. The distance to the optimal set decrease. Let $x^* \in \arg\min(F)$.

$$\forall k \in \mathbb{N},\ \|x_{k+1} - x^*\| \leqslant \|x_k - x^*\|.$$

$$x_{k+1} = x_k - s_k \nabla F(x_k)$$

① $$F(x_{k+1}) \leq F(x_k) + \langle \nabla F(x_k), \underbrace{x_{k+1} - x_k}_{-s_k \nabla F(x_k)} \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2$$

$$\Rightarrow F(x_{k+1}) \leq F(x_k) - s_k \|\nabla F(x_k)\|^2 + \frac{L}{2} s_k^2 \|\nabla F(x_k)\|^2$$

$$\leq F(x_k) - s_k \underbrace{\left(1 - \frac{L}{2} s_k\right)}_{\geq 0} \|\nabla F(x_k)\|^2$$

$$\leq F(x_k) \quad \text{if } s_k < \frac{2}{L}$$

② $$\forall k, \; F(x_{k+1}) - F(x_k) \leq \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right)$$

Using a series argument

$$= \sum_{k=0}^{N-1} \left(F(x_{k+1}) - F(x_k) \right) \leq \frac{L}{2} \left(\|x_0 - x^*\|^2 - \|x_N - x^*\|^2 \right)$$

$$\leq \frac{L}{2} \|x_0 - x^*\|^2$$

The gradient descent method

Proof (2/2)

$$\Rightarrow \forall n,\ n\left(F(x_n) - F^*\right) \leqslant \frac{L}{2} \|x_n - x^*\|^2$$

$$\Rightarrow \quad \forall n, \quad F(x_n) - F^* \leq \frac{L \|x_n - x^*\|^2}{2n}$$

General methodology to analyze optimization algorithms

- Interpreting the optimization algorithm as a discretization of a given ODE:

Gradient descent iteration: $$\frac{x_{n+1} - x_n}{h} + \nabla F(x_n) = 0$$

Associated ODE: $$\dot{x}(t) + \nabla F(x(t)) = 0.$$

- Analysis of ODEs using a Lyapunov approach:

$$\mathcal{E}(t) = F(x(t)) - F^*.$$

$$\mathcal{E}(t) = t(F(x(t)) - F^*) + \frac{1}{2}\|x(t) - x^*\|^2.$$

- Building a sequence of discrete Lyapunov energies adapted to the optimization scheme to get the same decay rates

Gradient descent for strongly convex functions

A Lyapunov analysis of the ODE $\dot{x}(t) + \nabla F(x(t)) = 0$ (1/3)

Let:

$$\mathcal{E}(t) = F(x(t)) - F^*.$$

1. $\mathcal{E}$ is a Lyapunov energy (i.e. non increasing along the trajectories $x(t)$):

$$\mathcal{E}'(t) = \langle \nabla F(x(t)), \dot{x}(t) \rangle = -\|\nabla F(x(t))\|^2$$

$$\leq 0$$

$$\Rightarrow \forall t \geq t_0, \quad \mathcal{E}(t) \leq \mathcal{E}(t_0)$$

$$\Rightarrow \forall t \geq t_0, \quad F(x(t)) - F^* \leq F(x_0) - F^*$$

Gradient descent for strongly convex functions
A Lyapunov analysis of the ODE $\dot{x}(t) + \nabla F(x(t)) = 0$ (2/3)

2. Assume now that F is additionally μ-strongly convex. Remember that:

$$\forall y \in \mathbb{R}^N,\ \|\nabla F(y)\|^2 \geqslant 2\mu(F(y) - F^*),$$

$$\mathcal{E}'(t) = -\|\nabla F(x(t))\|^2 \leq -2\mu\left(F(x(t)) - F^*\right)$$

$$\leq -2\mu\, \mathcal{E}(t)$$

$$\Rightarrow \forall t \geq t_0, \quad \mathcal{E}(t) \leq e^{-2\mu t}\, \mathcal{E}(t_0)$$

$$\Rightarrow \forall t \geq t_0, \quad F(x(t)) - F^* \leq e^{-2\mu t}\left(F(x_0) - F^*\right)$$

Gradient descent for strongly convex functions

A Lyapunov analysis of the ODE $\dot{x}(t) + \nabla F(x(t)) = 0$, $x(0) = x_0$

3. Assume that F is only convex. Let $x^* \in \arg\min(F)$.

$$\mathcal{E}(t) = t(F(x(t)) - F^*) + \frac{1}{2}\|x(t) - x^*\|^2.$$

$\mathcal{E}$ is also Lyapunov energy:

$$\mathcal{E}'(t) = t\langle \nabla F(x(t)), \overbrace{\dot{x}(t)}^{= -\nabla F(x(t))} \rangle + F(x(t)) - F^* + \langle x(t) - x^*, \overbrace{\dot{x}(t)} \rangle$$

$$= -t\|\nabla F(x(t))\|^2 + \underbrace{F(x(t)) - F^* - \langle x(t) - x^*, \nabla F(x(t)) \rangle}_{\leq 0 \text{ by convexity.}}$$

$$\leq -t\|\nabla F(x(t))\|^2 \leq 0$$

$$\Rightarrow \forall t \geq t_0, \quad t(F(x(t)) - F^*) \leq \mathcal{E}(t) < \mathcal{E}(t_0)$$

$\hookrightarrow$ cv in $O\left(\frac{1}{t}\right)$

$$\mathcal{E}_n = F(x_n) - F^* \qquad \text{with:} \quad x_{n+1} = x_n - s\nabla F(x_n).$$

1. $\mathcal{E}_n$ is a discretization of the Lyapunov energy $\mathcal{E}(t)$. We have:

$$\begin{aligned} \mathcal{E}_{n+1} - \mathcal{E}_n &= F(x_{n+1}) - F(x_n) \leqslant \langle \nabla F(x_n), x_{n+1} - x_n \rangle + \frac{L}{2}\|x_{n+1} - x_n\|^2 \\ &\leqslant -s\left(1 - \frac{L}{2}s\right)\|\nabla F(x_n)\|^2 \end{aligned}$$

If the step s satisfies:

$$s < \frac{2}{L}$$

then the GD is a descent algorithm ($\forall n, F(x_{n+1}) < F(x_n)$) and the values $F(x_n) - F^*$ remain bounded.

$$\mathcal{E}_n = F(x_n) - F^* \qquad \text{with:} \quad x_{n+1} = x_n - s\nabla F(x_n).$$

1. Assume now that F is additionally μ-strongly convex and $h < \frac{2}{L}$. We have:

$$\forall n,\ \|\nabla F(x_n)\|^2 \geqslant 2\mu(F(x_n) - F^*) = 2\mu\mathcal{E}_n,$$

and

$$\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant -s\left(1 - \frac{L}{2}s\right)\|\nabla F(x_n)\|^2.$$

Hence:

$$\mathcal{E}_{n+1} - \mathcal{E}_n \quad \leqslant \quad -2\mu s\left(1 - \frac{L}{2}s\right)\mathcal{E}_n$$

For example if $s = \frac{1}{L}$ we get:

$$\forall n,\ \mathcal{E}_{n+1} - \mathcal{E}_n \quad \leqslant \quad -\kappa\mathcal{E}_n \quad \Rightarrow \quad \mathcal{E}_n \leqslant (1-\kappa)^n\mathcal{E}_0$$

hence:

$$F(x_n) - F^* \leqslant (F(x_0) - F^*)(1-\kappa)^n.$$

The gradient descent method

Convergence rates for convex and strongly convex functions

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex function having a L-Lipschitz gradient. Choosing $s = \frac{1}{L}$ in (GD), we get:

Theorem

$$\forall k \in \mathbb{N},\ \ F(x_k) - F(x^*) \leqslant \frac{L\|x_0 - x^*\|^2}{2k}.$$

The number of iterations to reach $F(x_k) - F^* \leqslant \varepsilon$ is in $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$.

Additionally assume F μ-strongly convex for some $\mu > 0$. Then:

Theorem

$$\forall n \in \mathbb{N}, F(x_n) - F^* \leqslant (1-\kappa)^n (F(x_0) - F^*) \qquad \textit{where } \kappa = \frac{\mu}{L}.$$

The number of iterations to reach $F(x_n) - F^* \leqslant \varepsilon$ is in $\mathcal{O}\left(\log(\frac{1}{\varepsilon})\right)$.

Rmq: the optimal convergence factor is: $F(x_n) - F^* \leqslant \left(\frac{L-\mu}{L+\mu}\right)^n (F(x_0) - F^*)$ attained for $s = \frac{2}{L+\mu}$.

First order method: any iterative algorithm that selects x_{k+1} in the set

$$x_0 + \mathrm{span}\left\{\nabla F(x_0), \dots, \nabla F(x_k)\right\}.$$

Theorem (Nemirovski Yudin 1983, Nesterov 2003)

Let $k \leqslant \frac{N-1}{2}$ and $L > 0$. There exists a convex function F having a L-Lipschitz gradient over $\mathbb{R}^N$ such that for any first order method

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

- $\hookrightarrow$ Suggests that the rate $\frac{1}{k}$ for GD is not optimal !
- $\hookrightarrow$ We will see that recent accelerated gradient methods have a $\frac{1}{k^2}$ convergence rate.

Outline

1. **Preliminaries: local geometry of convex functions**
2. **Gradient descent methods**
3. **Accelerated gradient methods**
   - The Heavy ball method
   - The Nesterov's accelerated gradient method
   - Natural extension to composite optimization
   - Some numerical experiments
4. **Improving the state of the art results with the help of the geometry**
5. **Conclusion**

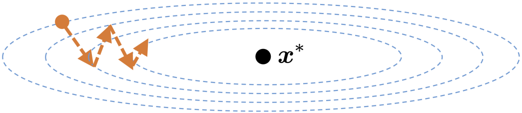
The Heavy Ball method

A first inertial method (Polyak 1964)

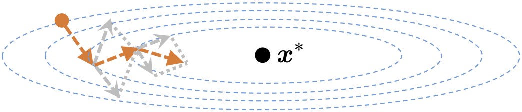
The Heavy ball method

$$\begin{array}{rcl} y_k & = & x_k + a(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - s\nabla F(x_k) \end{array} \quad , \ \alpha \in [0,1], \ s > 0.$$

where $a \in [0,1]$ is an *fixed* inertial coefficient added to mitigate zigzagging.



gradient descent



heavy-ball method

Let us consider:

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla F(x(t)) = 0.$$

- Describe the motion of a body (a heavy ball) in a potential field F subject to a friction proportional to its velocity.
- Natural intuition: the body reaches a minimum of the potential F.

The Heavy Ball method

Link between the continuous ODE and the discrete scheme

The HB algorithm:

$$\begin{array}{rcl} y_k & = & x_k + a(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - s\nabla F(x_k) \end{array}, \ \alpha \in [0,1], \ s > 0.$$

can be seen as a discretization of the second order ODE:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0$$

where:

$$s = h^2, \quad a = 1 - \alpha h.$$

$$x_{k+1} = x_k + a(x_k - x_{k-1}) \to \nabla f(x_k) = 0$$

$$\underbrace{\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1})}_{\ddot{x}(t)} + \frac{(1-a)}{h}\underbrace{\frac{(x_k - x_{k-1})}{h}}_{\dot{x}(t)} + \frac{s}{h^2}\underbrace{\nabla F(x_k)}_{\nabla F(x(t))} = 0$$

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla F(x(t)) = 0$$

Theorem (Global convergence - Polyak 1964)

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex function of class C^2 and having a L-Lipschitz continuous gradient.

- *If $\alpha \leqslant 2\sqrt{\mu}$ then:*
$$F(x(t)) - F^* = \mathcal{O}\left(e^{-\alpha t}\right).$$
- *If $\alpha > 2\sqrt{\mu}$ then:*
$$F(x(t)) - F^* = \mathcal{O}\left(e^{-(\alpha - \sqrt{\alpha^2 - 4\mu})t}\right).$$

- $F(x) = \frac{\mu}{2} \|x\|^2 \rightsquigarrow \ddot{x}(t) + \alpha \dot{x}(t) + \mu x(t) = 0$

The Heavy Ball method

Convergence results - In the discrete case

$$\begin{aligned} y_k &= x_k + a(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla F(x_k) \end{aligned}$$

with (Polyak's choice):

$$a = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2, \quad s = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2.$$

Theorem (Global convergence - [Polyak 1964])

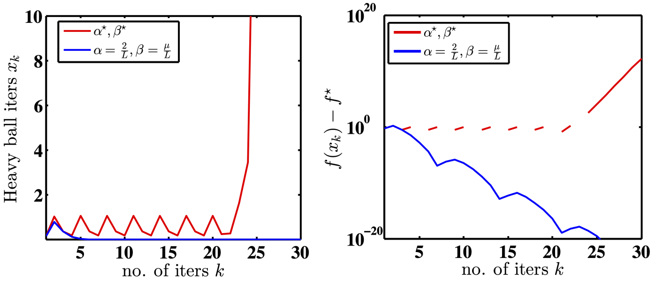
Let $F : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex function of class C^2 and having a L-Lipschitz continuous gradient. If $s < \frac{2}{L}$ then:

$$F(x_k) - F^* \leqslant \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^k (F(x_0) - F^*).$$

Counter example [Ghadimi et al. 2015] Let F be a C^1 μ-strongly convex and L-smooth function (with $\mu = 5$ and $L = 50$) such that:

$$\nabla F(x) = \begin{cases} 50x + 45 & \text{if } x < -1 \\ 5x & \text{if } -1 \leqslant x < 0 \\ 50x & \text{if } x \geqslant 0 \end{cases}$$

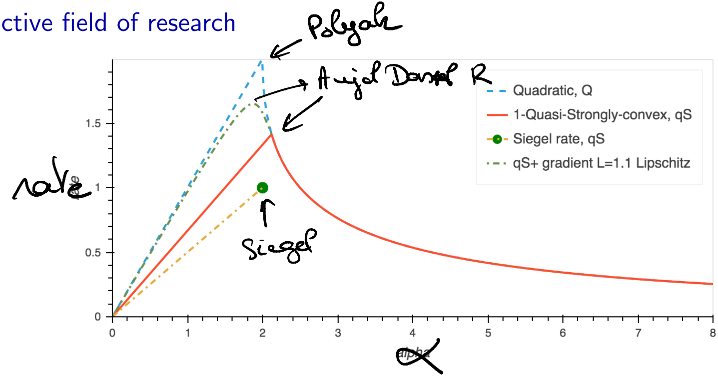
anf F is not of class C^2.



The Heavy Ball method

Convergence results - Without the C^2 assumption

An active field of research



- Nesterov's variant (2013): $F(x_n) - F^* = \mathcal{O}((1-\sqrt{\kappa})^n)$.
- Changing the step and the inertia, [Ghadimi et al. 2015] prove the linear cv for C^1 strongly convex functions having a Lipschitz continuous gradient.
- For strongly convex functions of class C^1 having a L-Lipschitz gradient [Siegel 2019]: when $\alpha = 2\sqrt{\mu}$, $F(x(t)) - F^* = \mathcal{O}\left(e^{-\sqrt{\mu}t}\right)$.

The Nesterov's accelerated gradient method

The historical scheme

Historically Nesterov proposes in 1983 the following inertial method:

$$\begin{aligned} y_k &= x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla F(y_k) \end{aligned}$$

where the sequence $(t_k)_{k\in\mathbb{N}}$ is defined by: $t_1 = 1$ and:

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}.$$

With such an algorithm Nesterov proves that:

$$\forall k \in \mathbb{N},\ F(x_k) - F^* \leqslant \frac{2\|x_0 - x^*\|^2}{sk^2}$$

but he did not prove the convergence of the iterates $(x_k)_{k\in\mathbb{N}}$.

Nesterov inertial scheme

$$\left| \begin{array}{rcl} y_n & = & x_n + \dfrac{n}{n+\alpha}(x_n - x_{n-1}) \\ x_{n+1} & = & y_n - h\nabla F\,(y_n)\,. \end{array} \right.$$

- Initially, Nesterov (1984) proposes $\alpha = 3$.
- Adapted by Beck and Teboulle to composite nonmooth functions (FISTA)

Link between the ODE and the optimization scheme (1/2)

Discretization of an ODE, Su Boyd and Candès (15)

The scheme defined by

$$x_{n+1} = y_n - \cancel{h}s \nabla F(y_n) \text{ with } y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1})$$

can be seen as a semi-implicit discretization of a solution of

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0 \qquad \text{(ODE)}$$

With $\dot{x}(t_0) = 0$. Move of a solid in a potential field with a vanishing viscosity $\frac{\alpha}{t}$.

(Discretization step: $h = \sqrt{s}$ and $x_n \simeq x(n\sqrt{s})$)

General methodology to analyze optimization algorithms

- Interpreting the optimization algorithm as a discretization of a given ODE:

Nesterov iteration: $x_{n+1} = y_n - h\nabla F(y_n)$ with $y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1})$

Associated ODE: $\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0.$

- Analysis of ODEs using a Lyapunov approach:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\dot{x}(t)\|^2.$$

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \frac{1}{2}\left\|(\alpha - 1)(x(t) - x^*) + t\dot{x}(t)\right\|^2..$$

- Building a sequence of discrete Lyapunov energies adapted to the optimization scheme to get the same decay rates

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function and $x^* \in \arg\min(F) \neq \emptyset$.

- If $\alpha \geqslant 3$,

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^2}\right)$$

[Attouch, Chbani, Peypouquet, Redont 2016]

- If $\alpha > 3$, then $x(t)$ cv to a minimizer of F and:

$$F(x(t)) - F(x^*) = o\left(\frac{1}{t^2}\right)$$

[Su, Boyd, Candes 2016]
[Chambolle, Dossal 2017]
[May 2017]

- If $\alpha < 3$ then no proof of cv of $x(t)$ but:

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$$

[Attouch, Chbani, Riahi 2019]
[Aujol, Dossal 2017]

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function with $X^* := \arg\min(F) \neq \emptyset$.

$$\left|\begin{array}{rcl} y_n & = & x_n + \dfrac{n}{n+\alpha}(x_n - x_{n-1}) \\ x_{n+1} & = & y_n - h\nabla F(y_n) \end{array}\right. , \quad \alpha > 0,\ h < \frac{1}{L}$$

- If $\alpha \geqslant 3$
$$F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^2}\right)$$
[Nesterov 1984, Su, Boyd, Candes 2016, Chambolle Dossal 2015, Attouch et al. 2018]
- If $\alpha > 3$, then $(x_n)_{n\geqslant 1}$ cv and:
$$F(x_n) - F(x^*) = o\left(\frac{1}{n^2}\right)$$
[Chambolle, Dossal 2014]
[Attouch, Peypouquet 2015]
- If $\alpha \leqslant 3$
$$F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^{\frac{2\alpha}{3}}}\right).$$
[Attouch, Chbani, Riahi 2018]
[Apidopoulos, Aujol, Dossal 2018]

A first Lyapunov energy

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0$$

$$\mathcal{E}(t) = F(x(t)) - F(x^*) + \frac{1}{2}\|\dot{x}(t)\|^2$$

be the mechanical energy associated to the ODE.

$$\mathcal{E}'(t) = \langle \nabla F(x(t)), \dot{x}(t) \rangle + \langle \ddot{x}(t), \dot{x}(t) \rangle$$

$$= -\frac{\alpha}{t}\|\dot{x}(t)\|^2 \leq 0$$

$$\hookrightarrow \quad F(x(t)) - F^* \leq M$$

A second Lyapunov energy to get the rate $\mathcal{O}\left(\frac{1}{t^2}\right)$ Can we prove that the energy:

$$E(t) = t^2 \left(F(x(t)) - F(x^*)\right) + \frac{t^2}{2}\|\dot{x}(t)\|^2$$

is bounded ? The answer is : NO

The Nesterov's accelerated gradient method
Proof of the convergence rate $\mathcal{O}\left(\frac{1}{t^2}\right)$ under convexity (Su Boyd Candes 2014)

We define :

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \overbrace{\frac{1}{2}\left\|(\alpha-1)(x(t)-x^*) + t\dot{x}(t)\right\|^2}^{\geqslant 0}.$$

Using (ODE), a straightforward computation shows that:

$$\begin{aligned}\mathcal{E}'(t) &= -(\alpha-1)t \underbrace{\langle \nabla F(x(t)), x(t)-x^*\rangle}_{\geqslant F(x(t))-F(x^*) \text{ by convexity}} + 2t(F(x(t)) - F(x^*)) \\ &\leqslant \underbrace{(3-\alpha)}_{\leqslant 0} t(F(x(t)) - F(x^*)).\end{aligned}$$

If $\alpha \geqslant 3$ then $\mathcal{E}'(t) \leqslant 0 \Rightarrow \forall t \geqslant t_0,\ \mathcal{E}(t) \leqslant \mathcal{E}(t_0)$

$\Rightarrow \forall t \geqslant t_0,\ t^2(F(x(t)) - F^*) \leqslant \mathcal{E}(t) \leqslant \mathcal{E}(t_0)$

$\Rightarrow \forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant \dfrac{\mathcal{E}(t_0)}{t^2}$

If $\alpha > 3$ $\quad \mathcal{E}'(t) \leqslant \dfrac{3-\alpha}{t}\mathcal{E}(t) \Rightarrow \mathcal{E}(t) \leqslant \dfrac{\mathcal{E}(t_0)}{t^{\alpha-3}},\ t \geqslant t_0$

$$\forall t \geq t_0, \quad \mathcal{E}(t) \leq \frac{\mathcal{E}(t_0)}{t^{\alpha-3}} \Rightarrow \forall t \geq t_0, \ t^2\left(F(x(t)) - F^*\right) \leq \frac{\mathcal{E}(t_0)}{t^{\alpha-3}}$$

$$\Rightarrow \quad t\left(F(x(t)) - F^*\right) \leq \frac{\mathcal{E}(t_0)}{t^{\alpha-2}}$$

$\alpha > 3$

$\alpha - 2 > 1$

$$\Rightarrow \quad \int_{t_0}^{+\infty} t\left(F(x(t)) - F^*\right) < +\infty$$

The continuous, a guideline to analyse the Nesterov scheme

Cv for the class of differentiable convex functions

- Continuous setting:

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \frac{1}{2}\left\|(\alpha - 1)(x(t) - x^*) + t\dot{x}(t)\right\|^2.$$

- Discrete setting:

$$\mathcal{E}_n = n^2(F(x_n) - F(x^*)) + \frac{1}{2h}\left\|(\alpha - 1)(x_n - x^*) + n(x_n - x_{n-1})\right\|^2$$

Using the definition of $(x_n)_{n\geqslant 1}$ and the following convex inequality

$$F(x_n) - F(x^*) \leqslant \langle x_n - x^*, \nabla F(x_n)\rangle$$

we get

$$\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant (3 - \alpha)n(F(x_n) - F(x^*)) \tag{1}$$

1. If $\alpha \geqslant 3$, $\forall n \geqslant 1$, $n^2(F(x_n) - F(x^*)) \leqslant \mathcal{E}_1$
2. If $\alpha > 3$, $\sum_{n\geqslant 1}(\alpha - 3)n(F(x_n) - F(x^*)) \leqslant \mathcal{E}_1$

Natural extension to composite optimization

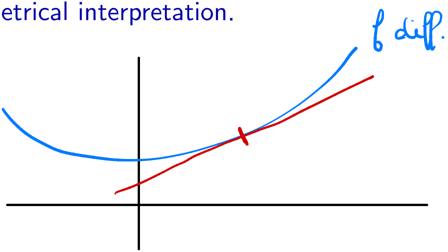
$$\min_x \quad F(x) = \underbrace{f(x)}_{\substack{\text{convex} \\ \text{differentiable} \\ L\text{-Lipschitz gradient}}} + \underbrace{g(x)}_{\substack{\text{convex l.s.c. "simple"} \\ \text{possibly non smooth.}}}$$

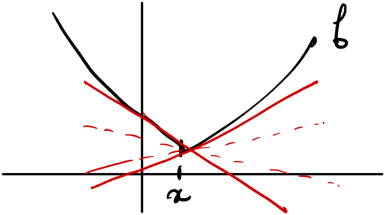
1. Notion of convex subdifferential
2. Proximal operator
3. The Forward Backward algorithm and FISTA

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x_0 \in \mathrm{dom}(f)$. The subdifferential of f at x_0, denoted by $\partial f(x_0)$, is defined by:

$$\partial f(x_0) = \{g \in \mathbb{R}^n \mid \forall x \in \mathrm{dom}(f),\ f(x) \geq f(x_0) + \langle g, x - x_0 \rangle\}.$$

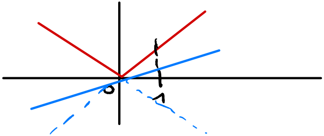
Geometrical interpretation.





__Ex__ $f(x) = |x|,\ x \in \mathbb{R}$

$\partial f(1) = \{+1\}$ $\quad \partial f(0) = [-1, 1]$



Subdifferential of a convex function

Properties

$$\partial f(x_0) = \{g \in \mathbb{R}^n \mid \forall x \in \mathrm{dom}(f),\ f(x) \geq f(x_0) + \langle g, x - x_0 \rangle\}$$

1. If $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex then $\partial f(x)$ is also convex and closed for any $x \in dom(f)$.

2. If f is convex l.s.c then $\partial f(x)$ is additionally non empty and bounded for any $x \in dom(f)$.

Let f be a convex l.s.c. function. Then:

1. If f is differentiable on $\mathrm{dom} f$ then:

$$\forall x \in int(\mathrm{dom}(f)),\ \partial f(x) = \{\nabla f(x)\}.$$

Let f be a convex l.s.c. function. Then:

1. If f is differentiable on $\mathrm{dom} f$ then:

$$\forall x \in int(\mathrm{dom}(f)),\ \partial f(x) = \{\nabla f(x)\}.$$

2. Let $\mathrm{dom}(f) \subseteq \mathbb{R}^n$, $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator and $b \in \mathbb{R}^n$ then $x \mapsto \phi(x) = f(Ax + b)$ is convex l.s.c. and:

$$\forall x \in int(\mathrm{dom}(\phi)),\ \partial\phi(x) = A^\top \partial f(Ax + b).$$

Let f be a convex l.s.c. function. Then:

1. If f is differentiable on $\mathrm{dom} f$ then:
$$\forall x \in int(\mathrm{dom}(f)),\ \partial f(x) = \{\nabla f(x)\}.$$
2. Let $\mathrm{dom}(f) \subseteq \mathbb{R}^n$, $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator and $b \in \mathbb{R}^n$ then $x \mapsto \phi(x) = f(Ax + b)$ is convex l.s.c. and:
$$\forall x \in int(\mathrm{dom}(\phi)),\ \partial \phi(x) = A^\top \partial f(Ax + b).$$
3. Si $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ avec f_1 et f_2 convex l.s.c. on $\mathbb{R}^n$ and $(\alpha_1, \alpha_2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Then:
$$\begin{aligned} \partial f(x) &= \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x) \\ &= \{g \in \mathbb{R}^n \mid \exists (x_1, x_2) \in \partial f_1(x_1) \times \partial f_2(x_2), g = \alpha_1 x_1 + \alpha_2 x_2\}. \end{aligned}$$

Let f be a convex l.s.c. function. Then:

1. If f is differentiable on $\mathrm{dom} f$ then:

$$\forall x \in int(\mathrm{dom}(f)),\ \partial f(x) = \{\nabla f(x)\}.$$

2. Let $\mathrm{dom}(f) \subseteq \mathbb{R}^n$, $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator and $b \in \mathbb{R}^n$ then $x \mapsto \phi(x) = f(Ax + b)$ is convex l.s.c. and:

$$\forall x \in int(\mathrm{dom}(\phi)),\ \partial \phi(x) = A^\top \partial f(Ax + b).$$

3. Si $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ avec f_1 et f_2 convex l.s.c. on $\mathbb{R}^n$ and $(\alpha_1, \alpha_2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Then:

$$\begin{aligned} \partial f(x) &= \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x) \\ &= \{g \in \mathbb{R}^n \mid \exists (x_1, x_2) \in \partial f_1(x_1) \times \partial f_2(x_2), g = \alpha_1 x_1 + \alpha_2 x_2\}. \end{aligned}$$

4. Let $g : \mathbb{R} \cup \{+\infty\} \to \mathbb{R} \cup \{+\infty\}$ a non-decreasing convex function. Let: $h = g \circ f$. Then:

$$\forall x \in int(\mathrm{dom}(f)),\ \partial h(x) = \{\eta_1 \eta_2 \mid \eta_1 \in \partial g(f(x)), \eta_2 \in \partial f(x)\}. \quad (2)$$

1. $f(x) = |x|$

$$\partial f(x) = [-1, 1]$$

2. $f(x, y) = x + 2|y|$

$$\partial f(x,0) = \left\{ \begin{pmatrix} 1 \\ 2g \end{pmatrix} ; g \in \partial|\cdot|(0) \right\}$$

$$= \left\{ \begin{pmatrix} 1 \\ 2g \end{pmatrix} \middle/ g \in [-1, 1] \right\}$$

3. $f(x, y) = y^2 + |x^2 + 3y|$

4. $f(x) = \|x\|_2^2$, $x \in \mathbb{R}^N$.

$$\partial f(x) = \{2x\}$$

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The point $x^* \in \mathbb{R}^n$ is a global minimum point of f si et seulement si:

$$0 \in \partial f(x^*).$$

Elementary proof: If f differentiable, $\partial f(x^*) = \{\nabla f(x^*)\}$

$$0 \in \partial f(x^*) \iff \nabla f(x^*) = 0$$

x^* global minimum pt of f

$$\iff \forall x \in \mathbb{R}^n, \quad f(x) \geq f(x^*)$$

$$\iff \forall x \in \mathbb{R}^n, \quad f(x) \geq \underline{f(x^*) + \langle 0, x - x^* \rangle}$$

$$\iff 0 \in \partial f(x^*)$$

$$\operatorname{prox}_f(x) = \arg\min_{u \in \mathbb{R}^n} f(u) + \frac{1}{2}\|u - x\|_2^2.$$

Examples

- $f(x) = \chi_X(x) = \begin{cases} 0 & \text{if } x \in X \\ +\infty & \text{otherwise,} \end{cases}$ X convexe

$$\min_{u \in \mathbb{R}^n} f(u) + \frac{1}{2}\|u - x\|^2 = \min_{u \in X} \frac{1}{2}\|u - x\|^2 \rightsquigarrow P_X(x)$$

- $f(x) = 0$ $\quad \min_{u} 0 + \frac{1}{2}\|u - x\|_2^2 \rightsquigarrow u^* = x$

$$\operatorname{prox}_0(x) = \operatorname{id}(x)$$

- $f(x) = \alpha\|x\|_1 \rightsquigarrow$ FISTA.

$$\operatorname{prox}_{\gamma f}(x) = \max(0, |x| - \gamma) * \operatorname{sign}(x)$$

1. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex l.s.c. function. Then $\text{prox}_f(x)$ exists and is unique for any $x \in \mathbb{R}^n$. Moreover

$$p = \text{prox}_f(x) \Leftrightarrow \; x - p \in \partial f(p)$$

2. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $t > 0$. Then:

$$\bar{x} \in \arg\min_{x \in \mathbb{R}^n} f(x) \Longleftrightarrow \bar{x} = \text{prox}_{tf}(\bar{x}).$$

Proximal point algorithm

$$\begin{aligned} x_0 &\in \mathbb{R}^n \\ x_{k+1} &= prox_{tf}(x_k), \quad t > 0. \end{aligned}$$

Let:

$$\min_{x \in \mathbb{R}^n} f(x) = g(x) + h(x)$$

where:

- $g : \mathbb{R}^n \to \mathbb{R}$ convex differentiable function having a L-Lipschitz gradient:

$$\|\nabla g(x_1) - \nabla g(x_2)\| \leq L\|x_1 - x_2\|, \ \forall (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n. \quad (3)$$

- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex l.s.c. function: simple = we can compute its prox.

Optimality condition

$$0 \in \partial f(x) \Leftrightarrow x = \text{prox}_{sh}(x - s\nabla g(x)).$$

$$T = \text{prox}_{sh} \circ (I - s\nabla g)$$

Forward-Backward and FISTA algorithms

Forward-Backward algorithm

$$
\begin{aligned}
x_0 &\in \mathbb{R}^n \\
x_{k+1} &= \mathrm{prox}_{sh}(x_k - s\nabla g(x_k)),\ s > 0.
\end{aligned}
$$

If $s < \frac{2}{L}$ then the sequence $(x_k)_{k\in\mathbb{N}}$ converge to a global minimum point of f and

$$\forall k \in \mathbb{N},\ F(x_k) - F(x^*) \leqslant \frac{2\|x_0 - x^*\|^2}{sk}$$

FISTA algorithm

$$
\begin{aligned}
y_k &= x_k + \alpha_k(x_k - x_{k-1}) \\
x_{k+1} &= \mathrm{prox}_{sg}(y_k - s\nabla f(y_k))
\end{aligned}
$$

If $s < \frac{1}{L}$ then the sequence $(k)_{k\in\mathbb{N}}$ converge to a global minimum point of f and

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

Some numerical experiments

Outline

Improving the state of the art results with the help of the geometry
Joint work with J.-F. Aujol, Ch.Dossal

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function and $x^* \in \arg\min(F) \neq \emptyset$.

- If $\alpha \geqslant 3$,

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^2}\right)$$

[Attouch, Chbani, Peypouquet, Redont 2016]

- If $\alpha > 3$, then $x(t)$ cv to a minimizer of F and:

$$F(x(t)) - F(x^*) = o\left(\frac{1}{t^2}\right)$$

[Su, Boyd, Candes 2016]
[Chambolle, Dossal 2017]
[May 2017]

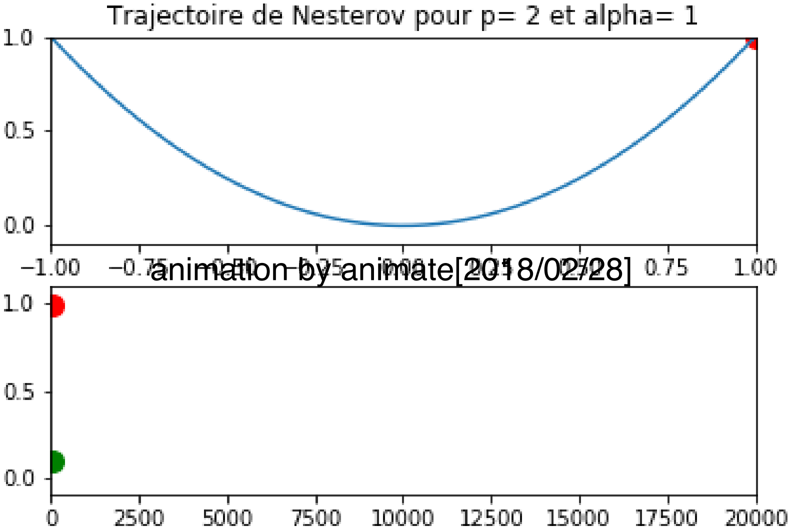
- If $\alpha < 3$ then no proof of cv of $x(t)$ but:

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$$

[Attouch, Chbani, Riahi 2019]
[Aujol, Dossal 2017]

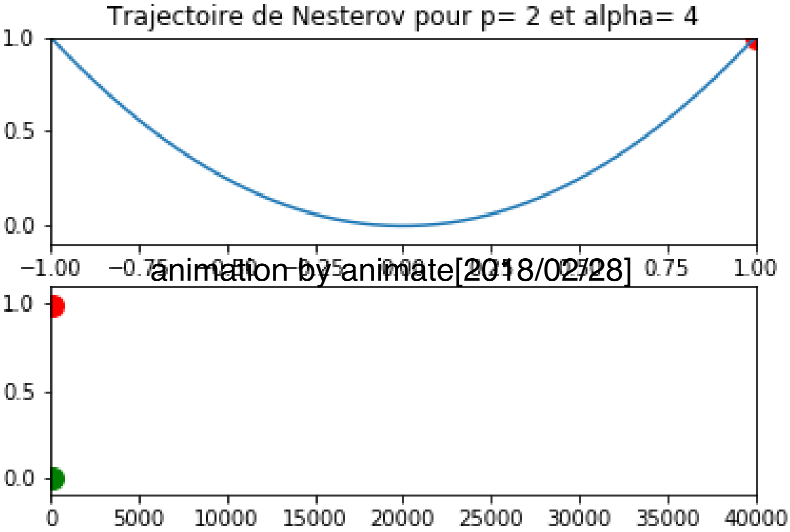
First Example : $F(x) = x^2$ and $\alpha = 1$ - State of the art rate: $\mathcal{O}(\frac{1}{n^{2/3}})$

In blue $F(x_n)$, in orange $n \times (F(x_n) - F^*)$



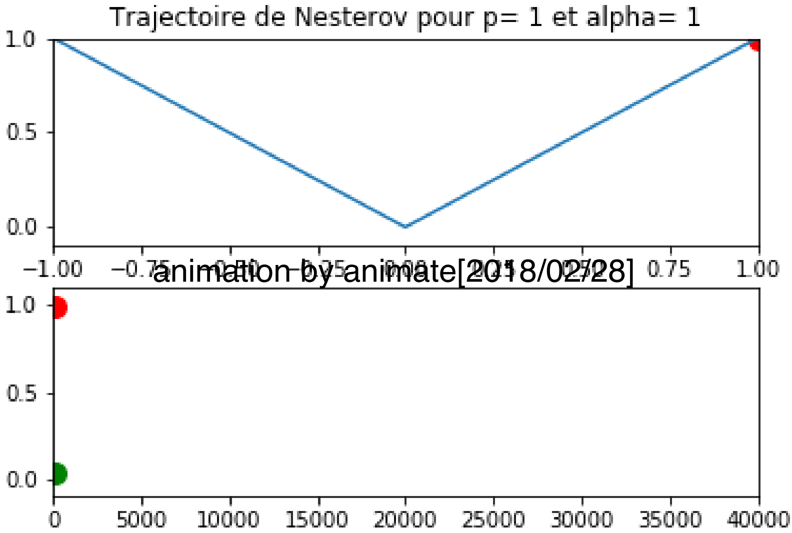
Second Example : $F(x) = x^2$ and $\alpha = 4$ - State of the art rate: $\mathcal{O}(\frac{1}{n^2})$

In blue $F(x_n)$, in orange $n^4 \times (F(x_n) - F^*)$



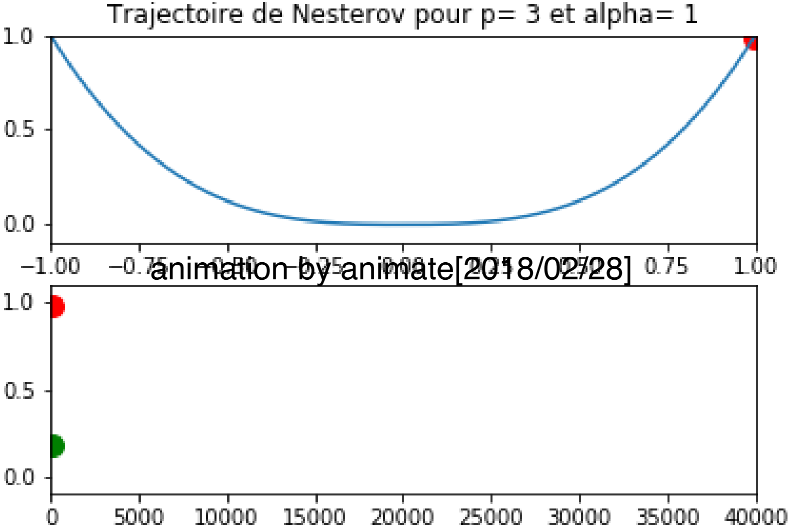
A Third Example FISTA : $F(x) = |x|$ and $\alpha = 1$

In blue $F(x_n)$, in orange $n^{\frac{2}{3}} \times F(x_n)$



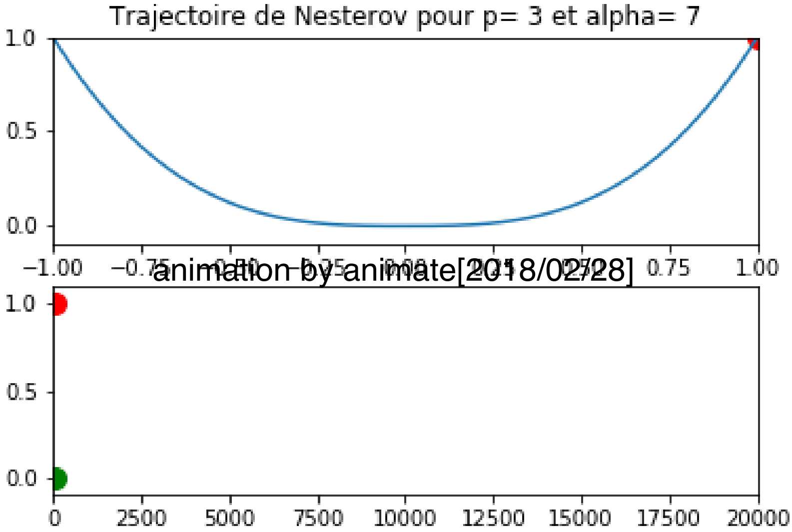
Fourth Example : $F(x) = |x|^3$ and $\alpha = 1$ - State of the art rate: $\mathcal{O}(\frac{1}{n^{2/3}})$

In blue $F(x_n)$, in orange $n^{\frac{6}{5}} \times (F(x_n) - F^*)$



Fifth Example : $F(x) = |x|^3$ and $\alpha = 7$ - State of the art rate: $\mathcal{O}(\frac{1}{n^2})$

In blue $F(x_n)$, in orange $n^6 \times (F(x_n) - F^*)$



Definition

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable function. F has the $\mathcal{H}(\gamma)$ property for some $\gamma \geqslant 1$ if

$$\forall x \in \mathbb{R}^n,\ F(x) - F(x^*) \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^* \rangle.$$

Flatness properties

- If $(F - F^*)^{\frac{1}{\gamma}}$ is convex, then F satisfies $\mathcal{H}(\gamma)$.
- If F satisfies $\mathcal{H}(\gamma)$ then for any $x^* \in X^*$, there exist $C > 0$ and $\eta > 0$ such that

$$\forall x \in B(x^*, \eta),\ F(x) - F(x^*) \leqslant C\|x - x^*\|^\gamma.$$

- Ensures that F is sufficiently flat (*at least as flat as* $\|x\|^\gamma$) around its set of minimizers.
- May prevent from bad oscillations of the solution.

Theorems for sharp functions [Aujol, Dossal, R. 2018]

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0.$$

Assume now that F is μ-strongly convex, satisfies the flatness condition $\mathcal{H}(\gamma)$ and admits a unique minimizer x^*. Then:

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha\gamma}{\gamma+2}}}\right) \tag{4}$$

$$x_{n+1} = y_n - h\nabla F(y_n) \quad \text{where:} \quad y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}), \qquad \alpha > 0,\ h < \frac{1}{L}$$

Theorem for sharp functions (Apidopoulos, Aujol, Dossal, R. (2018))

Assume that F is strongly convex and satisfies $\mathcal{H}(\gamma)$ for some $\gamma \in [1,2]$.

$$\forall \alpha > 0,\ F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^{\frac{2\gamma\alpha}{\gamma+2}}}\right). \tag{5}$$

Theorem for sharp functions (Apidopoulos, Aujol, Dossal, R. (2018))

Assume that F is strongly convex and satisfies $\mathcal{H}(\gamma)$ for some $\gamma \in [1,2]$.

$$\forall \alpha > 0,\ F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^{\frac{2\gamma\alpha}{\gamma+2}}}\right). \tag{6}$$

Comments

- For $\gamma = 1$ we recover the decay $\mathcal{O}\left(\frac{1}{n^{\frac{2\alpha}{3}}}\right)$ from [Attouch, Cabot 2018].
- Since ∇F is $L-$Lipschitz and satisfies $\mathcal{L}(2)$, F automatically satisfies $\mathcal{H}(\gamma)$ for some $\gamma > 1$ and thus
$$\frac{2\gamma\alpha}{\gamma+2} > \frac{2\alpha}{3}.$$
- For quadratic functions (i.e. for $\gamma = 2$), we get $\mathcal{O}\left(\frac{1}{n^{\alpha}}\right)$.

Theorems for sharp functions [Aujol, Dossal, R. 2018]

Sketch of proof in the continuous case

1. We define for $(p,\xi,\lambda)\in\mathbb{R}^3$

$$\mathcal{H}(t)=t^p\left(t^2(F(x(t))-F*)+\frac{1}{2}\|(\lambda(x(t)-x^*)+t\dot{x}(t)\|^2+\frac{\xi}{2}\|x(t)-x^*\|^2\right)$$

2. We choose $(p,\xi,\lambda)\in\mathbb{R}^3$ depending on the hypotheses to ensure that $\mathcal{H}$ is bounded. $\mathcal{H}$ may not be non increasing.
   - For the class of convex functions, take: $p=0$, $\lambda=\alpha-1,\xi=0$.
   - For the class of sharp convex functions, take:
     $$p=\tfrac{2\alpha\gamma}{\gamma+2}-2,\ \lambda=\tfrac{2\alpha}{\gamma+2},\xi=\lambda(\lambda+1-\alpha).$$
3. We deduce that there exists $A\in\mathbb{R}$ such that

$$t^{2+p}(F(x(t))-F(x^*))\leqslant A-t^p\frac{\xi}{2}\|x(t)-x^*\|^2$$

4. If $\xi\geqslant 0$ then $F(x(t))-F(x^*)=\mathcal{O}\left(\frac{1}{t^{p+2}}\right)$.
5. If $\xi\leqslant 0$ we must use the strong convexity to conclude.

Convergence rates for flat functions

Theorem for flat functions (Apidopoulos, Aujol, Dossal, R. (2018))

Let $\gamma > 2$. If F has a unique minimizer x^*, if F satisfies the flatness condition $\mathcal{H}(\gamma)$ and the growth condition:

$$\forall x \in \mathbb{R}^n,\ \frac{\mu}{2}\|x - x^*\|^\gamma \leqslant F(x) - F^*$$

Then if $\alpha > \frac{\gamma+2}{\gamma-2}$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2\gamma}{\gamma-2}}}\right).$$

Comments

- Better rate than $o(\frac{1}{n^2})$.
- Better rate than for the Gradient descent: if F satisfies $\mathcal{L}(\gamma)$ with $\gamma > 2$, then

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{\gamma}{\gamma-2}}}\right)$$

[Garrigos et al. 2017].

Application to the linear Least Square problem

Let $A : \mathbb{R}^N \to \mathbb{R}^N$ a positive definite bounded linear operator and $y \in \mathbb{R}^N$. Consider

$$\min_{x\in\mathbb{R}^N} F(x) := \frac{1}{2}\|Ax - y\|^2.$$

- F is convex and has a L-Lipschitz continuous gradient ($L = |||A^*A|||$).
- As a convex quadratic function, we have:

$$F(x) - F(x^*) = \frac{1}{2}\langle \nabla F(x), x - x^*\rangle = \frac{1}{2}\|A(x - x^*)\|^2.$$

  - F satisfies $\mathcal{H}(\gamma)$ for any $\gamma \in [1, 2]$, and $\mathcal{L}(2)$.

- $\forall n,\ x_n \in \{x_0\} + \mathrm{Im}(A^*)$.

Since this problem has a unique solution on the space $\{x_0\} + \mathrm{Im}(A^*)$, our theorem is still applicable and:

$$F(x_n) - F^* = \mathcal{O}\left(\frac{1}{n^\alpha}\right).$$

To sum up

Two ingredients to get better convergence rates on $F(x_n) - F^*$

- A sharpness condition
  - Ensuring that the magnitude of the gradient is not too low in the neighborhood of the minimizers.
- A flatness condition.
  - Ensuring that F is not too sharp in the neighborhood of its minimizers to prevent from bad oscillations of the solution.

Optimal convergence rates for Nesterov acceleration. J.-F. Aujol, Ch. Dossal, A. Rondepierre. May 2018.

Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. V. Apidopoulos, J.-F. Aujol, Ch. Dossal, A. Rondepierre. December 2018.

Outline

Is the Nesterov's method really an acceleration of the GD ?

A first conclusion

- If F is sharp, Gradient Descent is faster than Nesterov.
- If F is flat, Nesterov is faster than Gradient Descent.
- Choose α as large as possible

A second conclusion : it's more complicated

- Constants in big $\mathcal{O}$ or in geometric decays may be important.

  For example in the convex case ($\gamma = 1$), the constant in $\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$ is of the form:

$$\forall t \geqslant \frac{\alpha}{\sqrt{\mu}},\ F(x(t)) - F(x^*) \leqslant CE_m(t_0)\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}}$$

- Nesterov with restart and backtracking may outperform Conjugate Gradient on the least square problem.