Optimal convergence rates for Nesterov acceleration

Vasileios Apidopoulos, Jean-François Aujol,
Charles Dossal, Aude Rondepierre

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

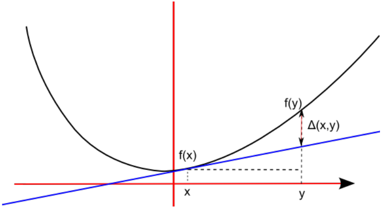
Séminaire MIAT - INRAE Toulouse Vendredi 16 avril 2021

The setting (1/2)

How to build an efficient sequence to estimate

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

where $F : \mathbb{R}^N \to \mathbb{R}$ is a differentiable convex function with a L-Lipschitz continuous gradient and at least one minimizer x^*.



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

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

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

The setting (2/2)

Possible extensions to

- Composite functions:
$$F(x) = f(x) + g(x)$$
where f is a convex differentiable function with a L-Lipschitz gradient and g is a convex lsc (possibly nonsmooth but quite simple) function.

  $\hookrightarrow$ Application to least square problems, LASSO:
$$\min_{x\in\mathbb{R}^N} \frac{1}{2}\|Ax-b\|^2 + \|x\|_1$$

- Constrained optimization:
$$\arg\min_{x\in C} F(x) \Leftrightarrow \arg\min_{x\in\mathbb{R}^N} F(x) + i_C(x).$$

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

Two examples of algorithms

Let $F : \mathbb{R}^N \to \mathbb{R}$ is a differentiable convex function with a L-Lipschitz continuous gradient and at least one minimizer x^*.

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

Explicit Gradient Descent

$$x_{n+1} \quad = \quad x_n - h\nabla F(x_n), \qquad h < \tfrac{2}{L}$$

Inertial Gradient Descent

$$\left| \begin{array}{lcl} y_n & = & x_n + \alpha_n(x_n - x_{n-1}) \\ x_{n+1} & = & y_n - h\nabla F(y_n) \end{array} \right. , \ \alpha_n \in [0,1], \ h < \frac{1}{L}.$$

Outline of the talk

How to exploit the geometry of F to get good or optimal convergence rates ?

A methodology to analyze optimization algorithms

- Link between optimization algorithms and ODEs. A guideline to study the optimization algorithms
- Analysis of ODEs using a Lyapunov approach
- Building a sequence of Lyapunov energies adapted to the optimization scheme to get the same decay rates

Illustration on two algorithms

1. Gradient descent algorithm
2. Nesterov scheme

Assume that F is μ-strongly convex i.e. that there exists $\mu > 0$ such that:

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

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

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

Explicit Gradient Descent

Assume that F is μ-strongly convex. The explicit gradient algorithm $x_{n+1} = x_n - h\nabla F(x_n)$ ensures that for any $h \leqslant \frac{1}{L}$,

$$F(x_n) - F^* \leqslant (1-\kappa)^n (F(x_0) - F^*) \qquad \text{where } \kappa = \frac{\mu}{L}.$$

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

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

Let:

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

1. Proving that $\mathcal{E}$ is non increasing only ensures that $F(x(t)) - F^*$ is bounded.

$$\mathcal{E}'(t) = \langle \nabla F(x(t)), \dot{x}(t) \rangle = -\|\nabla F(x(t))\|^2 \leqslant 0$$

hence:

$$F(x(t)) - F^* \leqslant F(x_0) - F^*.$$

Let:

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

1. Proving that $\mathcal{E}$ is non increasing only ensures that $F(x(t)) - F^*$ is bounded.

$$\mathcal{E}'(t) = \langle \nabla F(x(t)), \dot{x}(t) \rangle = -\|\nabla F(x(t))\|^2 \leqslant 0$$

hence:

$$F(x(t)) - F^* \leqslant F(x_0) - F^*.$$

2. Assume now that F is additionaly μ-strongly convex. Then we can prove:

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

hence:

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

and we deduce:

$$\forall t \geqslant t_0, \ F(x(t)) - F^* \leqslant (F(x_0) - F^*)e^{-2\mu(t-t_0)}.$$

Gradient descent for strongly convex functions
From the continuous to the discrete

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

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

If the step h satisfies:

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

then the GD is a descent algorithm:

$$\forall n, F(x_{n+1}) < F(x_n)$$

and the values $F(x_n) - F^*$ are bounded.

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

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

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

hence:

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

For example si $h \leqslant \frac{1}{L}$ we get:

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

hence:

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

Nesterov inertial scheme

Nesterov inertial scheme/FISTA

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

- Initially, Nesterov (1984) proposes $\alpha = 3$.
- Adapted by Beck and Teboulle to composite nonmooth functions (FISTA)
- For the class of convex functions, if $h < \frac{1}{L}$ and:
  - If $\alpha \geqslant 3$

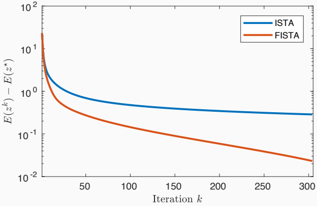
$$F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^2}\right).$$

[Su, Boyd, Candes 2016, Chambolle Dossal 2015, Attouch et al. 2018].

Efficiency of Nesterov-FISTA

$$F(x) = \frac{1}{2}\left\|y - h \star x\right\|_2^2 + \lambda \left\|Wx\right\|_1$$

(a) Input y: motion blur + noise ($\sigma = 2$)



(b) Convergence prof les

(c) Deconvolution ISTA(300)+UDWT

(d) Deconvolution FISTA(300)+UDWT

Some questions

Some questions

- Can we get more accurate rates than $\mathcal{O}\left(\frac{1}{n^2}\right)$ with more information on F?
- Are these bounds tight ?
- What is the role of the inertial parameter α ?
- Is Nesterov scheme really an acceleration of the Gradient descent ?

Some questions

Some questions

- Can we get more accurate rates than $\mathcal{O}\left(\frac{1}{n^2}\right)$ with more information on F?
- Are these bounds tight ?
- What is the role of the inertial parameter α ?
- Is Nesterov scheme really an acceleration of the Gradient descent ?

Answers

- Yes... with strong convexity, Su et al. (15) Attouch et al. (17)
- We give a more accurate answer for more general geometries.

Some questions

Some questions

- Can we get more accurate rates than $\mathcal{O}\left(\frac{1}{n^2}\right)$ with more information on F?
- Are these bounds tight ?
- What is the role of the inertial parameter α ?
- Is Nesterov scheme really an acceleration of the Gradient descent ?

Answers

- Yes... with strong convexity, Su et al. (15) Attouch et al. (17)
- We give a more accurate answer for more general geometries.
- In many numerical problems Nesterov is more efficient, but not always.
- Take-away message: Nesterov may be more efficient than GD... or not.

State of the art

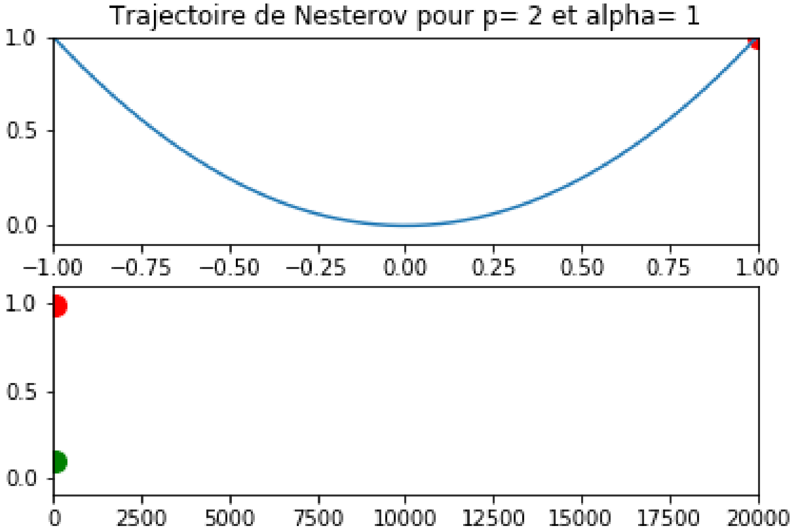
Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function with $X^* := \arg\min(F) \neq \emptyset$.

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

- If $\alpha \geqslant 3$
$$F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^2}\right)$$
[Attouch, Peypouquet 2016]
- 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]

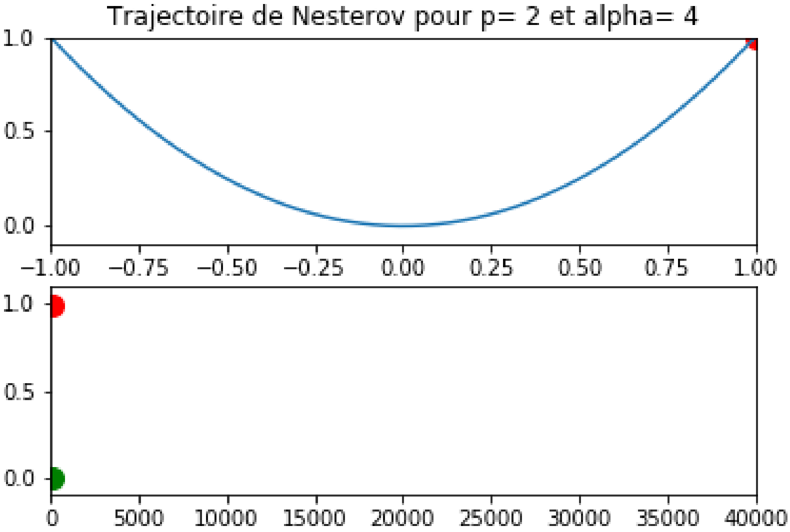
First Example : $F(x) = x^2$ and $\alpha = 1$ - State of the art rate: $\mathcal{O}(\frac{1}{n^{2/3}})$

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



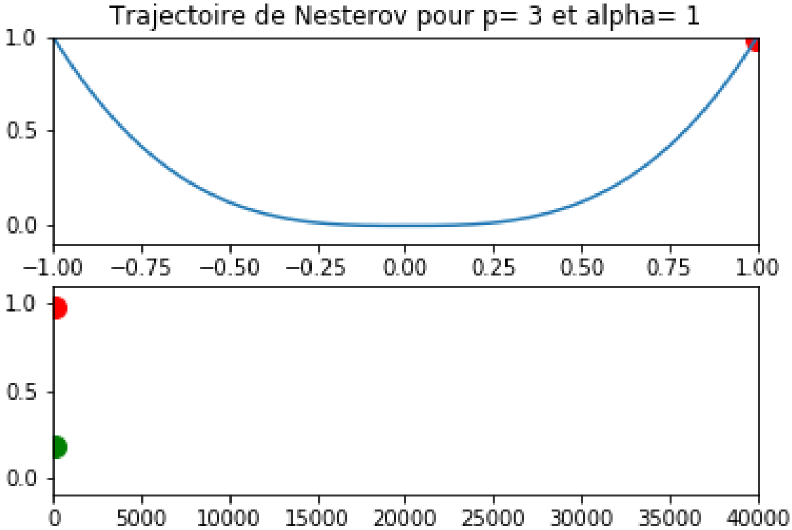
Second Example : $F(x) = x^2$ and $\alpha = 4$ - State of the art rate: $\mathcal{O}(\frac{1}{n^2})$

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



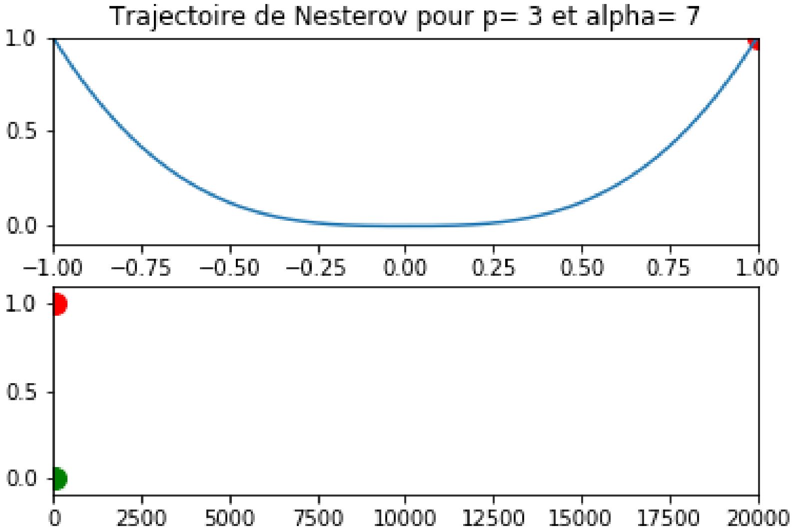
Third Example : $F(x) = |x|^3$ and $\alpha = 1$ - State of the art rate: $\mathcal{O}(\frac{1}{n^{2/3}})$

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



Fourth 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 $n^6 \times (F(x_n) - F^*)$



Nesterov: from the continuous to the discret

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

The scheme defined by

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

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

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

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

Advantages of the continuous setting

1. A simpler Lyapunov analysis, better insight
2. Optimality of bounds

Nesterov, Proof of the convergence rate $\mathcal{O}\left(\frac{1}{t^2}\right)$ under convexity

A first Lyapunov energy

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

be the mechanical energy associated to the ODE. We have:

$$\mathcal{E}_M'(t) \quad = \quad \langle \nabla F(x(t)), \dot{x}(t)\rangle + \langle \ddot{x}(t), \dot{x}(t)\rangle = -\frac{\alpha}{t}\|\dot{x}(t)\|^2 \leqslant 0.$$

Hence:

$$\begin{aligned} \forall t \geqslant t_0,\ F(x(t)) - F(x^*) \quad &\leqslant \quad \mathcal{E}_M(t) \leqslant \mathcal{E}_M(t_0) \\ &\leqslant \quad F(x_0) - F(x^*) + \frac{1}{2}\|\dot{v}_0\|^2 \end{aligned}$$

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

Nesterov, Proof of the convergence rate $\mathcal{O}\left(\frac{1}{t^2}\right)$ under convexity

We define:

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

Using (ODE), a straightforward computation shows that:

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

1. If $\alpha \geqslant 3$, $\forall t \geqslant t_0$, $t^2(F(x(t)) - F(x^*)) \leqslant \mathcal{E}(t_0)$.
2. If $\alpha > 3$, $\displaystyle\int_{t=t_0}^{+\infty} (\alpha-3)t(F(x(t) - F(x^*))dt \leqslant \mathcal{E}(t_0)$.

If F is convex and if $\alpha \geqslant 3$, the solution of (ODE) satisfies

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

Improving the convergence rate under geometrical assumptions

Assume now that F is μ-strongly convex and satisfies some flatness assumption:

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

for some $\gamma \geqslant 1$.

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

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

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

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

Nesterov, Proof of convergence rate

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

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

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.
3. We deduce that there exists $A \in \mathbb{R}$ such that

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

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.

For the class of convex functions, take: $p = 0$, $\lambda = \alpha - 1, \xi = 0$.
For the class of sharp convex functions, take:

$$p = \frac{2\alpha\gamma}{\gamma+2} - 2,\ \lambda = \frac{2\alpha}{\gamma+2}, \xi = \lambda(\lambda + 1 - \alpha).$$

The continuous, a guideline to analyse the Nesterov scheme

For the class of 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

$$\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant (3-\alpha)n(F(x_n) - F(x^*)) \tag{2}$$

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

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

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

$$\forall \alpha > 0,\ F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^{\frac{2\gamma\alpha}{\gamma+2}}}\right). \tag{3}$$

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 $\mathcal{O}\left(\frac{1}{n^{\alpha}}\right)$.

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
$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{\gamma}{\gamma-2}}}\right) \qquad \text{[Garrigos et al. 2017].}$$

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.

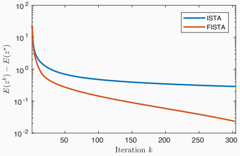
Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. V. Apidopoulos, J.-F. Aujol, Ch. Dossal, A. Rondepierre. December 2018.

Conclusion

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) Input y: motion blur + noise (σ = 2)



(b) Convergence prof les

(c) Deconvolution ISTA(300)+UDWT

(d) Deconvolution FISTA(300)+UDWT

$$F(x) = \frac{1}{2}\|y - h \star x\|_2^2 + \lambda \|Wx\|_1$$

satisfies $\mathcal{L}(2)$.

Conclusion

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.