Un tour d'horizon des résultats récents sur les algorithmes inertiels dans un cadre déterministe

Aude Rondepierre

Joint work with Jean-François Aujol, Charles Dossal,

Julien Hermant, Hippolyte Labarrière

Institut de Mathématiques de Toulouse, INSA de Toulouse

Congrès SMAI 2025 - MS Optimisation, un hommage à Hedy Attouch

The setting: Large scale optimization

Let:

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

where $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, convex or not, is assumed to have at least one minimizer x^*.

Includes the composite case: $F = f + h$ where f is a convex differentiable function and h is a convex lower semicontinuous (lsc) simple function.

Finding critical points of F / minimizers of F

- First order optimization methods i.e. methods that can only use the values of the function F and/or the values of its gradient (or subgradient).
- Convergence rates in term of decrease of $F(x_k) - F(x^*)$? Convergence rates on $\|x_k - x^*\|$?

Gradient Descent

Cauchy (1857) - Polyak (1964)

Assume that F is a convex differentiable function having a $L-$ *Lipschitz* gradient and at least one minimizer x^*. The gradient descent (GD) is defined by

$$x_{n+1} = x_n - s\nabla F(x_n) \text{ with } s \leqslant \frac{1}{L}.$$

1. The sequence $(F(x_n))_{n\in\mathbb{N}}$ is non increasing.
2. If F is convex, then $(x_n)_{n\in\mathbb{N}}$ weakly converges to a $x^* \in \arg\min(F)$ and:

$$\forall n \in \mathbb{N},\ F(x_n) - F(x^*) \leqslant \frac{\|x_0 - x^*\|^2}{2sn}$$

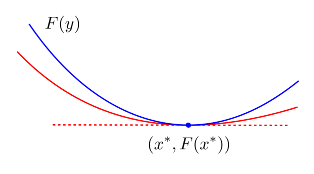
The number of iterations to reach $F(x_n) - F(x^*) \leqslant \varepsilon$ is in $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$.

3. The rate $\frac{1}{n}$ can not be improved assuming only convexity.

Stronger assumptions, better rates

Quadractic growth condition, a relaxation of strong convexity

If F is μ-strongly convex i.e. that $G(x) := F(x) - \frac{\mu}{2}\|x\|^2$ is convex,
or
If F satisfies some quadratic growth condition around its minimizers:



There exists $\mu > 0$ such that:

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

Then the iterates generated by GD with $s = \frac{1}{L}$, satisfy:

$$F(x_n) - F(x^*) = \mathcal{O}\left((1-\kappa)^n\right), \qquad \kappa = \frac{\mu}{L}.$$

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

Very slow if $\frac{\mu}{L} \ll 1$, but not pessimistic: rate achieved for $F(x_1, x_2) = \frac{\mu}{2}x_1^2 + \frac{L}{2}x_2^2$.

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

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

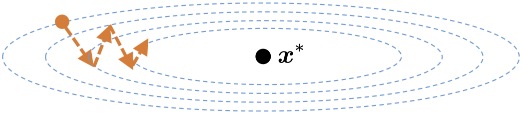
- $\hookrightarrow$ The rate in $\mathcal{O}\left(\frac{1}{k}\right)$ for GD is not optimal.
- $\hookrightarrow$ Can we do better with first order methods if F is convex ? if F is strongly convex ?

Yes, using inertial schemes.

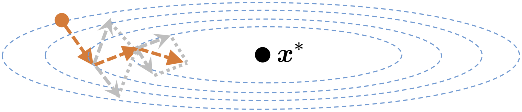
The Heavy ball method

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

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



gradient descent



heavy-ball method

The Heavy Ball method

The dynamical system intuition

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.

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:

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

with: $s = h^2$ and $a = 1 - \alpha h$ (a: inertia parameter - α: friction parameter).

The Heavy Ball method

Convergence results for strongly convex functions

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

with (Polyak's choice):

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

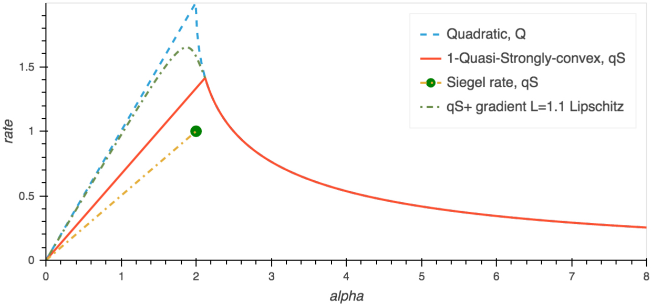
Theorem (Global convergence - [Polyak 1964])

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:

$$
F(x_k) - F^* \leqslant \underbrace{\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k}_{\sim O\left(e^{-2\sqrt{\frac{\mu}{L}}k}\right),\ k\to+\infty} (F(x_0) - F^*).
$$

The Heavy Ball method

How to choose α to optimize the convergence to a minimizer ?



- For strongly convex functions of class C^2 having a L-Lipschitz gradient, the optimal value of α is: $\alpha = 2\sqrt{\mu}$.
- Changing the step and the inertia, [Ghadimi et al. 2015] prove the geometric 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

Nesterov 1983

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

where the sequence $(t_k)_{k\in\mathbb{N}}$ is defined by: $t_1 = 1$ and: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.

A modified version (Chambolle Dossal 2015)

$$\begin{aligned} y_k &= x_k + \frac{k}{k+\alpha}(x_k - x_{k-1}), \quad \alpha \geqslant 3 \\ x_{k+1} &= y_k - s\nabla F(y_k) \end{aligned}$$

For the class of convex functions, the sequence of iterates satisfies:

$$\forall k \in \mathbb{N},\ F(x_k) - F^* \leqslant \frac{(\alpha+1)\|x_0 - x^*\|^2}{2sk^2}$$

and, for the modified version with $\alpha > 3$, weakly converges to a minimizer of F.

Link between the ODE and the optimization scheme

Discretization of an ODE, Su Boyd and Candès (2015)

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

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 $\frac{\alpha}{t}$.

General methodology to analyze optimization algorithms

- Interpreting the optimization algorithm as a discretization of a given ODE.
- Analysis of ODEs using a Lyapunov approach:

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

- 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 2015]
[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]

The Nesterov's accelerated gradient method

For the class of convex functions

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 2015]
[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]

GD vs Nesterov in the strongly convex case

Exponential rate vs Polynomial rate

Assume now that F is additionally μ-strongly convex, or satisfies some quadratic growth condition:

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

Convergence rate for GD

$$\forall n \in \mathbb{N},\ F(x_n) - F^* = \mathcal{O}\left((1-\kappa)^n\right).$$

The number of iterations required to reach an ε-solution is: $n_\varepsilon^{FB} \sim \frac{1}{\kappa} \log\left(\frac{2L}{\varepsilon^2} M_0\right)$.

Convergence rate for Nesterov's accelerated GD [Candès et al 2015], [Attouch Cabot 2017], [ADR 2018].

If F has a unique minimizer,

$$\forall \alpha > 0,\ \forall n \in \mathbb{N},\ F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$$

Nesterov accelerated algorithm for strongly convex functions

Nesterov accelerated algorithm for strongly convex functions

$$y_n = x_n + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(x_n - x_{n-1})$$
$$x_{n+1} = y_n - \tfrac{1}{L}\nabla F(y_n)$$

Theorem (Theorem 2.2.3, Nesterov 2013)

Assume that F is μ-strongly convex for some $\mu > 0$. Let $\varepsilon > 0$. Then for $\kappa = \frac{\mu}{L}$ small enough,

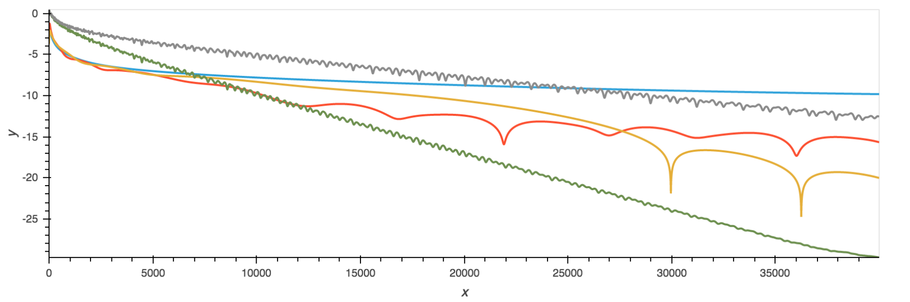
$$\forall n \in \mathbb{N},\ F(x_n) - F(x^*) \leqslant 2(1-\sqrt{\kappa})^n \left(F(x_0) - F(x^*)\right),$$

which means that an ε-solution can be obtained in at most:

$$n_\varepsilon^{NSC} = \frac{1}{|\log(1-\sqrt{\kappa})|} \log\left(\frac{4LM_0}{\varepsilon^2}\right). \tag{1}$$

The iterations require an estimation of $\kappa = \frac{\mu}{L}$!

FISTA in the strongly convex case



$\log(\|g(x_n)\|)$ along the iterations

FB, FISTA with $\alpha = 8$, FISTA with $\alpha = 30$,

NSC with the true value of μ, NSC with $\widetilde{\mu} = \frac{\mu}{10}$.

FISTA is efficient without knowing μ and its convergence rate does not suffer from any underestimation of μ

Convergence rate analysis under some quadratic growth condition

Theorem (Aujol Dossal R. 2023, Aujol Dossal Labarrière R. 2024)

Let $\varepsilon > 0$ *and*

$$\alpha_\varepsilon := 3\log\left(\frac{5\sqrt{L(F(x_0)-F^*)}}{e\varepsilon}\right) \qquad \textit{does not depend on any estimation of } \mu.$$

Let $(x_n)_{n\in\mathbb{R}^N}$ *be a sequence of iterates generated by the Nesterov's accelerated GD with parameter* α_ε. *Then for* $\kappa = \frac{\mu}{L}$ *small enough, an* ε*-solution is reached in at most:*

$$n_\varepsilon^{FISTA} := \frac{8e^2}{3\sqrt{\kappa}}\alpha_\varepsilon = \frac{8e^2}{\sqrt{\kappa}}\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right)$$

iterations.

Theorem (Aujol, Dossal, Labarrière, R. 2024)

If F *satisfies some local quadratic growth condition then, for* α *large enough, the sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by Nesterov GD/FISTA* **strongly** *converges to a minimizer of* F.

Conclusion (1/2)

- Inertial methods can be more efficient than the GD for the class of convex functions having a quadratic growth
  - No need to estimate the growth parameter μ and the convergence rate does not suffer from an underestimation of μ.
  - Strong convergence of the iterates generated by the Nesterov's accelerated GD/FISTA

- Restarting FISTA can improve the convergence rate
  - If F is $\mu-$strongly convex, restarting FISTA each $e\sqrt{\frac{L}{\mu}}$ ensures an exponential decay... but μ may be unknown.
  - Estimation of μ: Alamo et al 2020, Fercoq et al. 2023, Aujol Calatroni Dossal R. Labarrière 2024...

Conclusion (2/2)

- High resolution ODEs enables a more accurate description of the trajectories of the optimization algorithm.
  - Since 2016 Attouch and co-authors combine a Hessian-driven damping term to an asymptotic vanishing damping term resulting in

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

  - The HB scheme

$$\begin{cases} y_n & = x_n + \alpha(x_{n-1} - x_n) \\ x_{n+1} & = y_n - s\nabla F(x_n) \end{cases} \tag{2}$$

is associated to the following High Resolution ODE (Shi et al 2018)

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + (1 + \sqrt{\mu s})\nabla F(x(t)) = 0. \tag{3}$$