Pushing the Boundaries of First-Order Optimization: From Gradient Descent to Accelerated Algorithms

Samir Adly

Université de Limoges, Laboratoire XLIM.

Email: samir.adly@unilim.fr

GERAD
Campus de l'Université de Montréal.

Montréal, October 10th, 2023

1. Gradient Descent Method

Gradient methods: some historical aspects

- $\Big(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|\Big)$real Hilbert space, $f : \mathcal{H} \to \mathbb{R}$ continuously differentiable.

- $\min_{x\in\mathcal{H}} f(x) = f^*$: Find $\bar{x} \in \operatorname{argmin}(f)$ such that:

$$f(\bar{x}) = f^*.$$

- Optimality condition: solve the nonlinear equation

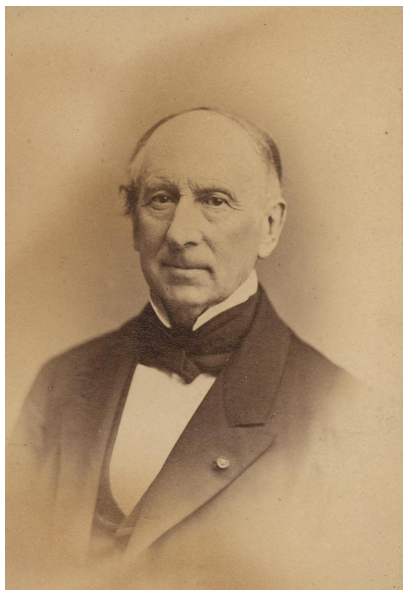
$$\nabla f(x) = 0.$$

Gradient Descent Method (Steepest Descent Method)

$$\begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = x_k - s_k \nabla f(x_k) \end{cases}$$

with $s_k > 0$: the step length or the learning rate.

- Attributed to Cauchy, who first used it in 1847.
- Hadamard proposed a similar method in 1907.
- First proof of convergence is due to Haskell Curry (in 1944).



Cauchy (1789-1857) Hadamard (1865-1963)

Cauchy and the Gradient Method

Claude Lemaréchal

2010 Mathematics Subject Classification: 65K05, 90C30
Keywords and Phrases: Unconstrained optimization, descent method, least-square method

Any textbook on nonlinear optimization mentions that the gradient method is due to Louis Augustin Cauchy, in his *Compte Rendu à l'Académie des Sciences* of October 18, 1847[1] (needless to say, this reference takes a tiny place amongst his fundamental works on analysis, complex functions, mechanics, etc. Just have a look at `http://mathdoc.emath.fr/cgi-bin/oetoc?id=OE_CAUCHY_1_10`: a paper every week).

Cauchy is motivated by astronomic calculations which, as everybody knows, are normally very voluminous. To compute the orbit of a heavenly body, he wants to solve *not the differential equations, but the* [*algebraic*] *equations representing the motion of this body, taking as unknowns the elements of the orbit themselves. Then there are six such unknowns*.[2]. Indeed, a motivation related with operations research would have been extraordinary. Yet, it is interesting to note that equation-solving has always formed the vast majority of optimization problems, until not too long ago.

To solve a system of equations in those days, *one ordinarily starts by reducing them to a single one by successive eliminations, to eventually solve for good the resulting equation, if possible. But it is important to observe that 1° in many cases, the elimination cannot be performed in any way; 2° the resulting equation is usually very complicated, even though the given equations are rather simple*.[3] Something else is wanted.

Thus consider a function

$$u = f(x, y, z, \ldots)$$

THE METHOD OF STEEPEST DESCENT FOR NON-LINEAR MINIMIZATION PROBLEMS*

By HASKELL B. CURRY (*Frankford Arsenal*)

1. Introduction. The problem considered here is that of minimizing a function of n real variables, $G(x_1, \cdots, x_n)$. The object is to find a practical method for evaluating, approximately at least, a stationary point for G.

This problem includes as a special case that of solving a set of simultaneous equations

$$f_i(x_1, \cdots, x_n) = 0 \qquad (i = 1, 2, \cdots, m), \tag{1}$$

because the function

$$G(x_1, \cdots, x_n) = \sum_{k=1}^{m} f_k^2 \tag{2}$$

has a minimum at a solution of (1). It also includes that of determining the parameters $x_1, \cdots, x_n$ of a function $f(u; x_1, \cdots, x_n)$ so as to get the best approximation, in a least square sense, to a function $F(u)$ for certain values of u; the G in this case is of the form given by

$$G(x_1, \cdots, x_n) = \sum_{k=1}^{p} [F(u_k) - f(u_k; x_1, \cdots, x_n)]^2. \tag{3}$$

Certain engineering applications of the latter sort of problem arose in the work of the Engineering Research Section, Fire Control Design Division, at Frankford Arsenal. In these applications, the function $f(u; x_1, \cdots, x_n)$ was sufficiently complicated so that the standard method for dealing with non linear least square problems[1] failed to converge. Two techniques for dealing with this situation were developed by the section under the direction of J. G. Tappert. One of these was an original suggestion of my associate K. Levenberg.[2] The second method is the subject of this note.

This method is not new. Levenberg found it set forth in a paper by Cauchy dated 1847.[3] That it has become a standard procedure in analysis is clear from a recent paper by Courant.[4] Nevertheless it does not appear to be well known to authorities on nu-

* Received Jan. 22, 1944.

[1] See, for example, W. E. Deming, *Some notes on least squares*, U. S. Dept. of Agriculture Graduate School, 1938, p. 31 ff., or E. T. Whittaker and G. Robinson, *The calculus of observations*, Blackie and Son, London, 1940, p. 214. Deming's treatment is also given in his book, *Statistical adjustment of data*, John Wiley & Sons, New York, 1943, p. 52 ff.

[2] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quarterly of Applied Mathematics, **2**, 164 (1944).

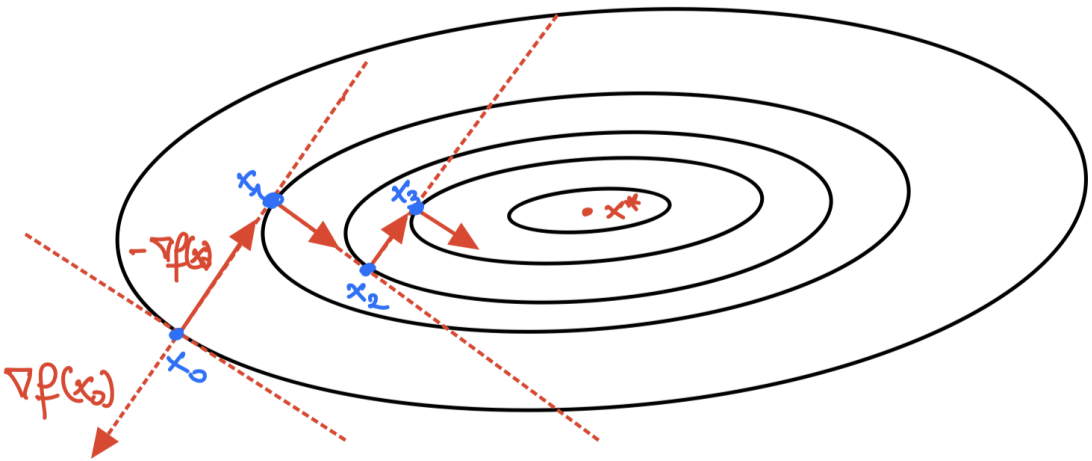
[3] A. L. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes rendus, Ac. Sci. Paris, **25**, 536–538 (1847).

[4] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc. **49**, 1–23 (1943). See especially pp. 17–20. Courant calls the method the "method of

Haskell B. Curry (1900-1982)

Gradient Descent Method: How to choose the step length?

- $$\begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = x_k - \alpha_k \nabla f(x_k) \end{cases}$$



How to choose the step length $s_k > 0$?

Gradient flow as a continuous model

$$\text{(GDM)} \begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = x_k - s\nabla f(x_k) \end{cases}$$

Question. Could we associate a continuous model to the GDM?

Let us introduce the Ansatz $x_k \simeq X(k\,s),\ k \in \mathbb{N}$ for a smooth function $X : [0, +\infty[\to \mathcal{H}$.
For $t = k\,s$, we have

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

Hence,

$$\frac{1}{s}\Big[X(t+s) - X(t)\Big] = -sf(X(t)).$$

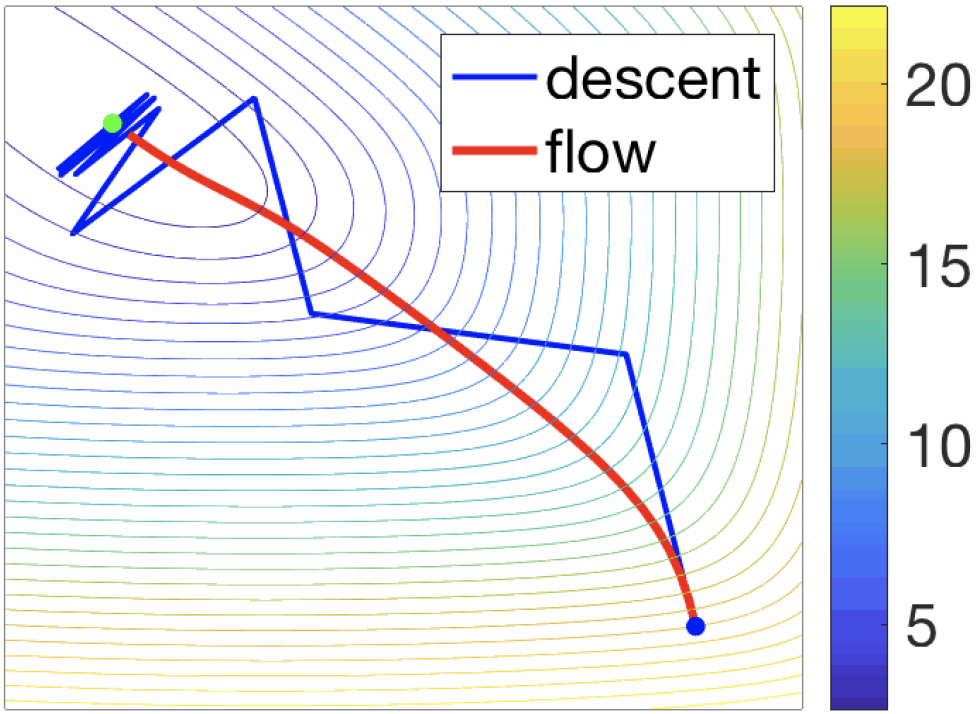
By letting the step length $\alpha \to 0$, we have

$$\text{(GF)} \begin{cases} \dot{X}(t) = -\nabla f(X(t)),\ t \geq 0 \\ X(0) = x_0 \in \mathcal{H}, \end{cases}$$

Gradient Flow

$$\text{(GF)} \begin{cases} \dot{x}(t) = -\nabla f(x(t)), \ t \geq 0 \\ x(0) = x_0 \in \mathcal{H} \end{cases} \xleftarrow[\text{Explicit Euler}]{s \to 0} \xrightarrow{\qquad} \text{(GDM)} \begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = x_k - s\nabla f(x_k) \end{cases}$$

$$\text{discrete iteration} = \frac{\text{continuous time}}{\text{step length}}, \quad k = \frac{t}{s}.$$



Convergence result of the Gradient Flow

$$(\mathrm{GF}) \begin{cases} \dot{x}(t) = -\nabla f(x(t)), \ t \geq 0 \\ x(0) = x_0 \in \mathcal{H} \end{cases}$$

Theorem

Let $f : \mathcal{H} \to \mathbb{R}$ be convex, continuously differentiable and bounded from below.
Assume that $S = \operatorname{argmin}(f) \neq \emptyset$. $f^* = \inf_{\mathcal{H}} f$. Then

(i) $f(x(t)) - f^* \leq \frac{d(x_0,S)^2}{2t}, \ t > 0.$

(ii) $x(t) \rightharpoonup x_\infty \in S$ *weakly as* $t \to +\infty$, $\nabla f(x_\infty) = 0$. *(Bruck (1975), Opial's lemma).*

Convergence result of the Gradient Flow: the convex case

Proof. Consider the Lyapounov function $t \mapsto E(t) = f(x(t)) - f^*$.
We have,

$$E'(t) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\|\dot{x}(t)\|^2 \leq 0.$$

Hence, the function $t \mapsto f(x(t))$ is nonincreasing, i.e. for every $s \leq t$, we have

$$f(x(t)) \leq f(x(s)). \tag{1}$$

We have,

$$f(x(t)) - f(x_0) = -\int_0^t \|\dot{x}(s)\|^2 ds = -\int_0^t \|\nabla f(x(s))\|^2 ds.$$

Since f is bounded below, we get

$$\int_0^{+\infty} \|\dot{x}(s)\|^2 ds = \int_0^{+\infty} \|\nabla f(x(s))\|^2 ds < +\infty.$$

On the other hand, we have for every $y \in \mathcal{H}$

Since f is convex, we obtain

$$\frac{1}{2}\frac{d}{dt}\|x(t) - y\|^2 \leq f(y) - f(x(t)).$$

Using (**??**), we have for every $0 \leq s \leq t$

$$f(x(t)) - f(y) \leq f(x(s)) - f(y).$$

Integrating this inequality, we get

$$\begin{aligned} t\Big(f(x(t)) - f(y)\Big) &\leq \int_0^t \Big(f(x(s)) - f(y)\Big) ds \\ &\leq \frac{1}{2}\|x_0 - y\|^2 - \frac{1}{2}\|x(t) - y\|^2. \\ &\leq \frac{1}{2}\|x_0 - y\|^2 \end{aligned}$$

Hence,

$$f(x(t)) - f(y) \leq \frac{1}{2t}\|x_0 - y\|^2, \ \forall y \in \mathcal{H}.$$

Convergence result of the Gradient Flow: the strongly convex case

Definition

A function $f : \mathcal{H} \to \mathbb{R}$ is μ-strongly convex iff $f - \frac{\mu}{2}\|\cdot\|^2$ is convex, i.e. for every $\lambda \in [0,1]$ and, $x,\ y \in \mathcal{H}$, we have

$$f\Big(\lambda x + (1-\lambda)y\Big) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\lambda(1-\lambda)\mu}{2}\|x-y\|^2.$$

- For differentiable functions, this is equivalent to the μ-strong monotonicity of the gradient∇f, i.e.

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x-y\|^2, \quad \forall x,\ y \in \mathcal{H}.$$

- Another characterization is

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2.$$

- The parameter $\mu > 0$ measures the curvature of f.

Convergence result of the Gradient Flow: the strongly convex case

Coming back to the proof of the gradient flow convergence, we have for μ-strongly convex functions and every $y \in \mathcal{H}$

$$\frac{1}{2}\frac{d}{dt}\|x(t)-y\|^2 = \langle \nabla f(x(t)), y - x(t)\rangle \leq f(y) - f(x(t)) - \frac{\mu}{2}\|x(t)-y\|^2.$$

In this case, the set of solutions $S = \{x^*\}$. So for $y = x^*$, we have

$$\frac{d}{dt}\|x(t)-x^*\|^2 + \mu\|x(t)-x^*\|^2 \leq 0.$$

Consequently,

$$\|x(t)-x^*\|^2 \leq e^{-\mu t}\|x_0 - x^*\|^2.$$

We deduce the strong convergence of the trajectory $x(t) \to x^*$ as $t \to +\infty$.

Convergence result of the Gradient Descent Method

Let $f : \mathcal{H} \to \mathbb{R}$ be convex and continuously differentiable.
Assume that $S = \operatorname{argmin}(f) \neq \emptyset$.
∇f Lipschitz continuous with modulus $L > 0$, $0 < sL < 2$.

Discrete dynamic: $x_{k+1} = x_k - s\nabla f(x_k), \quad x_0 \in \mathcal{H}$

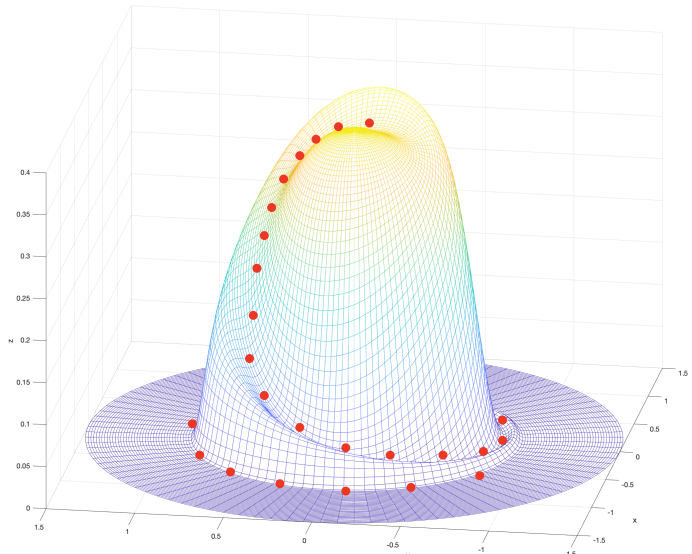
- $f(x_k) - f^* \leq \dfrac{L\operatorname{dist}(x_0, S)^2}{2k} = \mathcal{O}(\dfrac{1}{k})$ as $k \to +\infty$.
- $f(x_{k+1}) - f(x_k) + \dfrac{2 - sL}{2s}\|x_{k+1} - x_k\|^2 \leq 0$ (gradient descent lemma).
- $x_k \rightharpoonup x_\infty \in S$ weakly as $k \to +\infty$.

Gradient flow: the nonconvex case

$$\text{(SD)} \quad \dot{x}(t) + \nabla f(x(t)) = 0.$$

- $f : \mathbb{R}^N \to \mathbb{R}$ real analytic: Lojasiewicz (IHES, 1965).
  Any bounded trajectory converges to a critical point of f.
- Counterexample: J. Palis and W. De Melo(1982), mexican hat (a function in $\mathbb{R}^2$ of class C^∞).
  Without geometric hypothesis on f, $x(\cdot)$ may not converge.

Geometry of f: tame optimization,
KL, complexity.



Łojasiewicz inequality and (SD)

$$\text{(SD)} \quad \dot{x}(t) + \nabla f(x(t)) = 0.$$

Theorem (Łojasiewicz inequality, 1963)

Let $f : U \subset \mathbb{R}^N \to \mathbb{R}$ be real analytic, U be open, $\bar{x} \in U$ be a critical point of f.
Then, there exists $\theta \in [\frac{1}{2}, 1[$, $C > 0$, and a neighbourhood W of $\bar{x}$ s.t.

$$\forall x \in W \quad |f(x) - f(\bar{x})|^\theta \leq C\|\nabla f(x)\|.$$

Theorem (Łojasiewicz, 1984)

$f : U \subset \mathbb{R}^N \to \mathbb{R}$ real analytic. Any bounded trajectory of (SD) has a finite length and hence converges to a critical point of f, as $t \to +\infty$.

2. Acceleration of Gradient-based optimization algorithms

How to accelerate the Gradient Descent Method?

- Polyak's momentum
- Nesterov Accelerated Gradient Method (NAG).

Polyak's momentum

The first improvement of the Gradient Descent Method is due to Polyak in

B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*. Computational mathematics and mathematical physics (1964).

The algorithm is given by

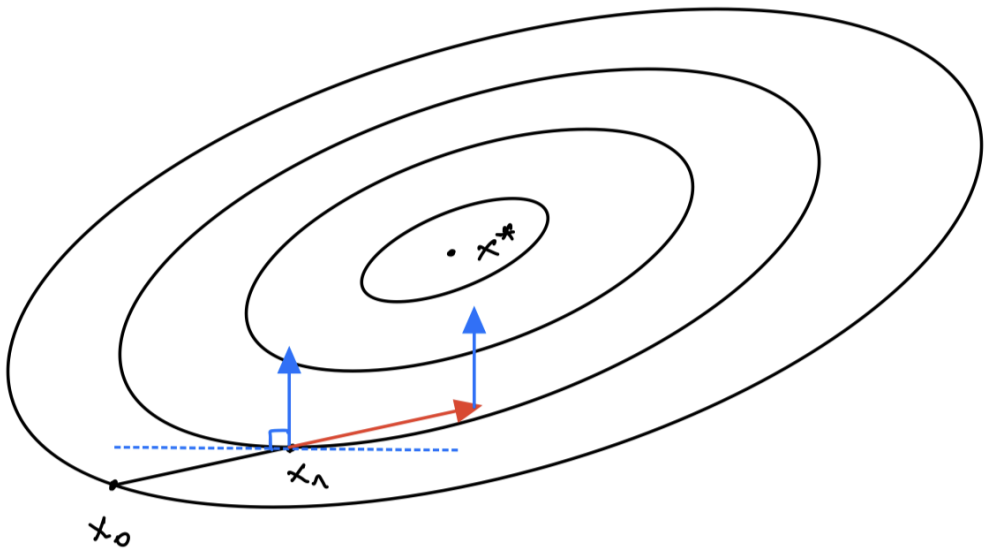
$$\text{(PM)} \begin{cases} x_0,\ x_1 \in \mathcal{H} \\ x_{k+1} = \underbrace{x_k - s\nabla f(x_k)}_{GDM} + \underbrace{\beta(x_k - x_{k-1})}_{momentum}, \end{cases}$$

where $s > 0$ is the step length of the GDM and $\beta > 0$ is the momentum coefficient.

The algorithm is accelerated by giving a momentum from the previous two steps.

Polyak's momentum: interpretation

$$\text{(PM)} \begin{cases} y_k = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_k - s\nabla f(x_k) \end{cases}$$



Remark

Polyak's momentum is not a descent method, i.e. $f(x_{k+1}) \leq f(x_k)$ could be not satisfied.

What is the continuous surrogate of Polyak's momentum?

$$\text{(PM)} \begin{cases} x_0,\ x_1 \in \mathcal{H} \\ x_{k+1} = x_k - s\nabla f(x_k) + \beta(x_k - x_{k-1}) \quad (\star), \end{cases}$$

Set $h = \sqrt{s}$ and $\beta = 1 - \gamma h$ with $\gamma > 0$.
We have

$$(\star) \iff (x_{k+1} - x_k) - (x_k - x_{k-1}) + (1-\beta)(x_k - x_{k-1}) + h^2 \nabla f(x_k) = 0$$
$$\iff \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \gamma \frac{x_k - x_{k-1}}{h} + \nabla f(x_k) = 0.$$

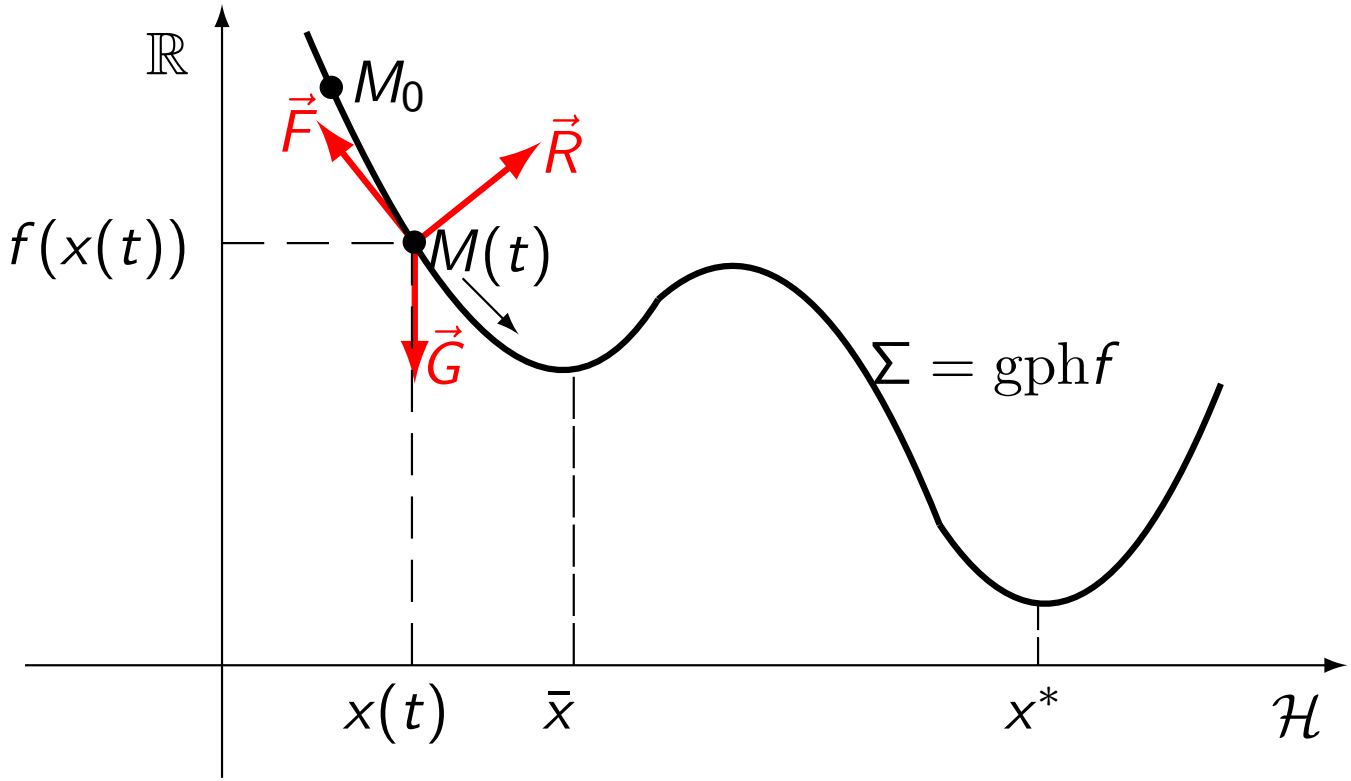
Let us introduce the Ansatz $x_k \simeq X(kh)$ with $k = \frac{t}{h}$. As the step size goes to 0, we get

$$\ddot{X}(t) + \gamma \dot{X}(t) + \nabla f(X(t)) = 0, \quad t \geq 0.$$
$$X(0) = x_0,\ \dot{X}(0) = x_1.$$

The heavy ball with friction method

Fixed viscous damping coefficient $\gamma > 0$, Polyak (1964, 1987)

$$\text{(HBF)} \quad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0,\ x(0) = x_0,\ \dot{x}(0) = x_1.$$



Mechanical interpretation $\vec{G}$ =gravity, $\vec{F}$= friction, $\vec{R}$= reaction

(HBF) in the μ-strongly convex case

Strongly convex functions

$f : \mathcal{H} \to \mathbb{R}$ μ-strongly convex $\iff$ $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

$f : \mathcal{H} \to \mathbb{R}$ μ-strongly convex

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \nabla f(x(t)) = 0.$$

- $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(e^{-\sqrt{\mu}t}\right)$ as $t \to +\infty$.
- Geometry of $f \longleftrightarrow$ Damping coefficient $\longleftrightarrow$ Convergence rate.

Theorem

If $f : \mathcal{H} \to \mathbb{R}$ is μ-strongly convex and of class $\mathcal{C}^2$, then

$$f(x(t)) - \inf_{\mathcal{H}} f \leq Ce^{-\sqrt{\mu}t}, \ \forall t \geq 0,$$

with $C = f(x_0) - \inf_{\mathcal{H}} f + \mu \mathrm{dist}(x_0, S)^2 + \|x_1\|^2$.

(HBF) in the convex case

$$(\text{HBF}) \begin{cases} \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0, \\ x(0) = x_0,\ \dot{x}(0) = x_1. \end{cases}$$

Theorem (Alvarez (SICON, 2000))

Let $f : \mathcal{H} \to \mathbb{R}$ be convex and of class C^1 such that $S = \operatorname{argmin}(f) \neq \emptyset$.

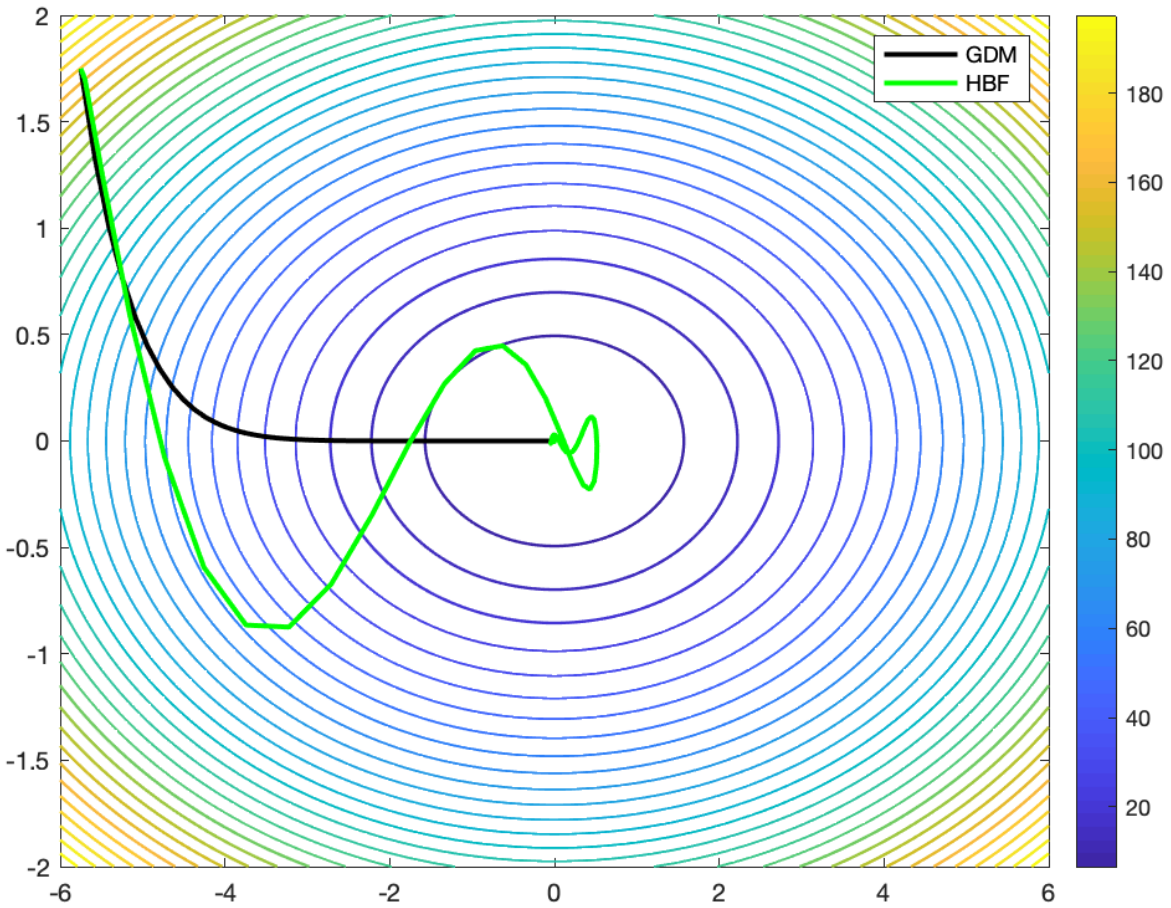
(i) $f(x(t) - \inf_{\mathcal{H}} f \leq \frac{C(x_0,x_1)}{t}$, *with*
$C(x_0, x_1) = \frac{3}{2\gamma}(f(x_0) - \inf_{\mathcal{H}} f) + \gamma \operatorname{dist}(x_0, S)^2 + \frac{5}{4\gamma}\|x_1\|^2$.

(ii) $x(t) \rightharpoonup x_\infty \in S$ *weakly* *as* $t \to +\infty$.

- $E(t) = \frac{1}{2}\|\dot{x}(t)\|^2 + f(x(t))$ the Lyapounov energy function.
  $E'(t) = -\gamma\|\dot{x}(t)\|^2 \leq 0$ (dissipative system).
- $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to +\infty$.

HBF versus GDM

$$\text{(HBF)} \begin{cases} y_k = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_k - s\nabla f(x_k) \end{cases} \qquad \text{(GDM)} \begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = x_k - s\nabla f(x_k) \end{cases}$$



The Heavy Ball with friction: optimal parameters

$$\text{(HBF)} \begin{cases} x_0,\ x_1 \in \mathcal{H} \\ x_{k+1} = x_k - s\nabla f(x_k) + \beta(x_k - x_{k-1}). \end{cases}$$

Let f be of class $\mathcal{C}^2$, μ-strongly convex and L-smooth. The optimal parameters α and β are given by:

$$s = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} \text{ and } \beta = \left[\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right]^2 \text{ with } \kappa = \frac{L}{\mu}.$$

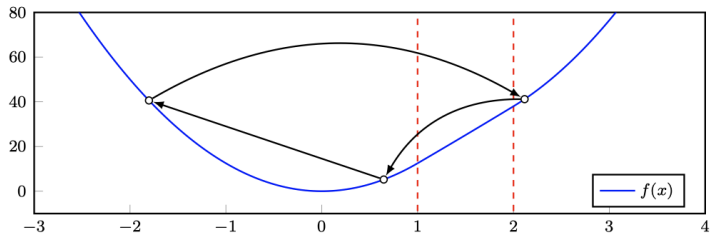
- (HBF) is optimal for $\mathcal{C}^2$, μ-strongly convex, and L-smooth functions.
- Knowledge of both parameters L and μ is crucial for the analysis.
- Tuning parameters s and β for smooth convex functions is unclear.
- For general convex functions, (HBF) converges asymptotically at $\mathcal{O}(1/t)$, not surpassing steepest descent.

The Heavy Ball with friction: some drawbacks

Beside the oscillation problems of the (HBF), it may fail to converge even for strongly convex functions (non $\mathcal{C}^2$).
The following counter-example is given in [LRP] (2015).
Take $f : \mathbb{R} \to \mathbb{R}$ such that

$$f'(x) = \begin{cases} 25x & \text{if} \quad x < 1 \\ x + 24 & \text{if} \quad 1 \leq x < 2 \\ 25x - 24 & \text{if} \quad x \geq 2. \end{cases}$$

The function is L-smooth and μ-strongly convex with $L = 25$ and $\mu = 1$. (HBF) produces a limit cycle with oscillations.



L. Lessard, B. Recht, A. Packard. Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. arXiv:1408.3595 (2015).

Nesterov's Accelerated Gradient Method (NAG)

In 1983, Y. Nesterov introduced an algorithm with momentum

$$\text{(NAG)} \begin{cases} x_{k+1} = y_k - s\nabla f(y_k), \;\; 0 \leq s \leq \frac{1}{L} \\ y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k). \end{cases}$$

with $\beta_k = \frac{k}{k+3}$: the momentum coefficient.

Starting with x_0 and $y_0 = x_0$.

This choice of the extrapolation coefficient is intriguing. It is considered one of the mysterious results in Optimization.

$$\frac{k}{k+3} \simeq 1 - \frac{3}{k} \text{ as } k \to +\infty.$$

Nesterov's Accelerated Gradient Method (NAG)

Докл. Акад. Наук СССР
Том 269 (1983), № 3

Soviet Math. Dokl.
Vol. 27 (1983), No. 2

A METHOD OF SOLVING A CONVEX PROGRAMMING PROBLEM WITH CONVERGENCE RATE $O(1/k^2)$

UDC 51

YU. E. NESTEROV

1. In this note we propose a method of solving a convex programming problem in a Hilbert space E. Unlike the majority of convex programming methods proposed earlier, this method constructs a minimizing sequence of points $\{x_k\}_0^\infty$ that is not relaxational. This property allows us to reduce the amount of computation at each step to a minimum. At the same time, it is possible to obtain an estimate of convergence rate that cannot be improved for the class of problems under consideration (see [1]).

2. Consider first the problem of unconstrained minimization of a convex function $f(x)$. We will assume that $f(x)$ belongs to the class $C^{1,1}(E)$, i.e. that there exists a constant $L > 0$ such that for all $x, y \in E$

$$\|f'(x) - f'(y)\| \leq L\|x - y\|. \tag{1}$$

From $\mathcal{O}\left(\frac{1}{k}\right)$ to $\mathcal{O}\left(\frac{1}{k^2}\right)$

Historical NAG $\alpha = 3$

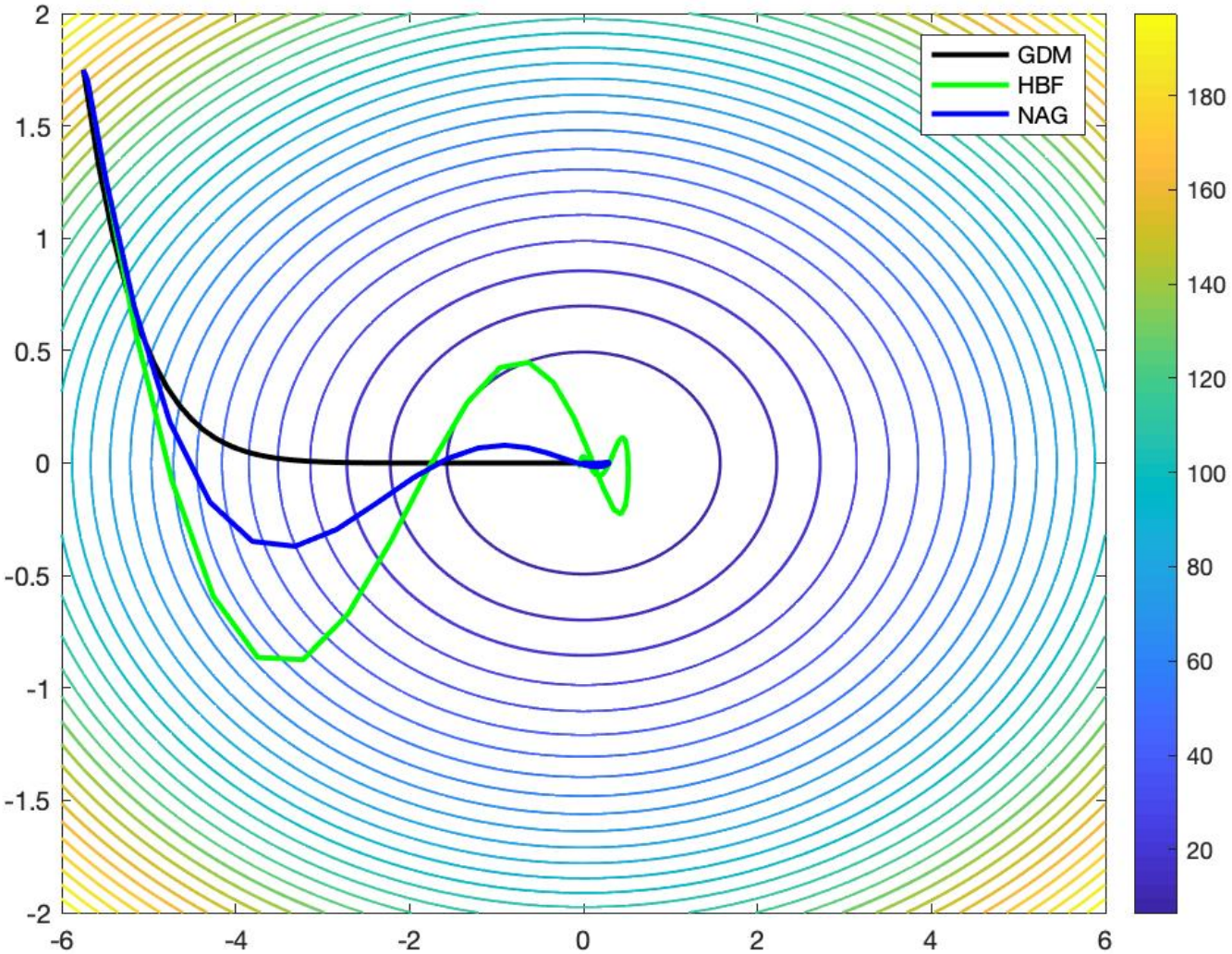
- Suppose that f is convex and L-smooth, then

$$f(x_k) - f^* \leq \frac{2L\,\mathrm{dist}(x_0, S)^2}{(k+1)^2} = \mathcal{O}\left(\frac{1}{k^2}\right).$$

- Convergence of the iterates is an open problem.
- Optimal rate among all first-order gradient based methods.
- Nemirovsky-Yudin (1983), Nesterov (2004), Drori-Teboulle (2012).
- Gradient-based first-order method is a black-box algorithm:

$$\left(x_0, g_0, \ldots, x_k; k_k\right) \mapsto x_{k+1} \in x_0 + \mathrm{Span}(g_0, \ldots, g_k).$$

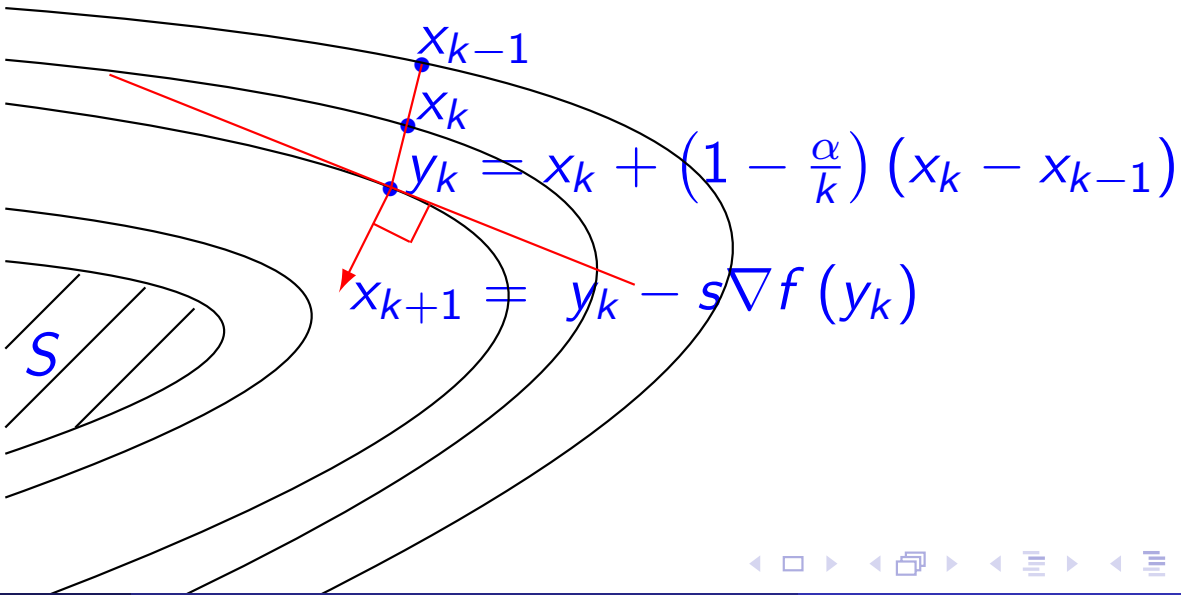
A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interseciences, 1983.


GDM
HBF
NAG
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-6
-4
-2
0
2
4
6
180
160
140
120
100
80
60
40
20

Nesterov's Accelerated Gradient Method (NAG)

$$(\mathrm{NAG})_\alpha \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla f(y_k). \end{cases}$$

- The gradient step is applied to y_k, which is obtained by linear extrapolation from x_k and x_{k-1}.
- Note the subtle tuning of the extrapolation coefficient $0 < \alpha_k := 1 - \frac{\alpha}{k} < 1$ which tends to one from below as $k \to +\infty$.



$$(\mathrm{NAG})_\alpha \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla f(y_k). \end{cases}$$

- $\alpha = 3$: Historical NAG. $f(x_k) - f^* \leq O(1/k^2)$ (Nesterov 1983). The convergence of the sequence (x_k) is an open question.

- $\alpha > 3$: $x_k \rightharpoonup x_\infty \in S$ (Chambolle-Dossal, 2015). $f(x_k) - f^* = o(1/k^2)$ (Attouch-Peypouquet, 2016).

- $0 < \alpha \leq 3$: $f(x_k) - f^* = \mathcal{O}(1/k^{\frac{2\alpha}{3}})$. Apidopoulos-Aujol-Dossal, Attouch-Chbani-Riahi (2016).

NAG for strongly convex function (NAG-SC)

$f : \mathcal{H} \to \mathbb{R}$ μ-strongly convex function.

$$(\mathrm{NAG-SC}) \begin{cases} y_{k+1} &= x_k - s\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(y_{k+1} - y_k). \end{cases}$$

Equivalently,

$$x_{k+1} = x_k - s\nabla f(x_k) + \Big(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\Big)(x_k - x_{k-1}) - s\Big(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\Big)\Big(\nabla f(x_k) - \nabla f(x_{k-1}\Big).$$

Like the heavy ball with the gradient correction term

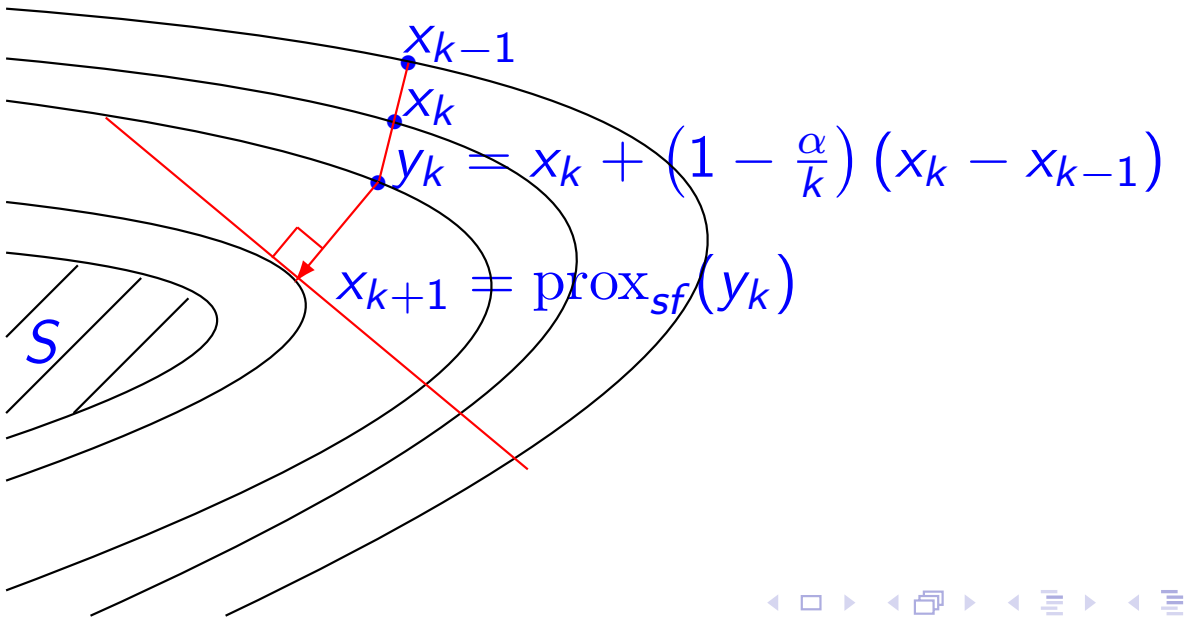
$$s\Big(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\Big)\Big(\nabla f(x_k) - \nabla f(x_{k-1}\Big).$$

Nonsmooth convex case: Inertial Proximal Algorithm

$\min\{f(x) : x \in \mathcal{H}\}, \ f \in \Gamma_0(\mathcal{H}), \ S = \operatorname{argmin} f \neq \emptyset.$

Inertial Proximal algorithm, $\operatorname{prox}_{sf}(y) := \operatorname{argmin}_{\xi\in\mathcal{H}}\{f(\xi) + \frac{1}{2s}\|y-\xi\|^2\}$

$$(\mathrm{IP})_\alpha \quad \begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{sf}(y_k). \end{cases}$$



The composite problem and the LASSO

$$\min_{x \in \mathcal{H}} h(x) := f(x) + g(x),$$

with f convex and L-smooth and $g \in \Gamma_0(\mathcal{H})$.

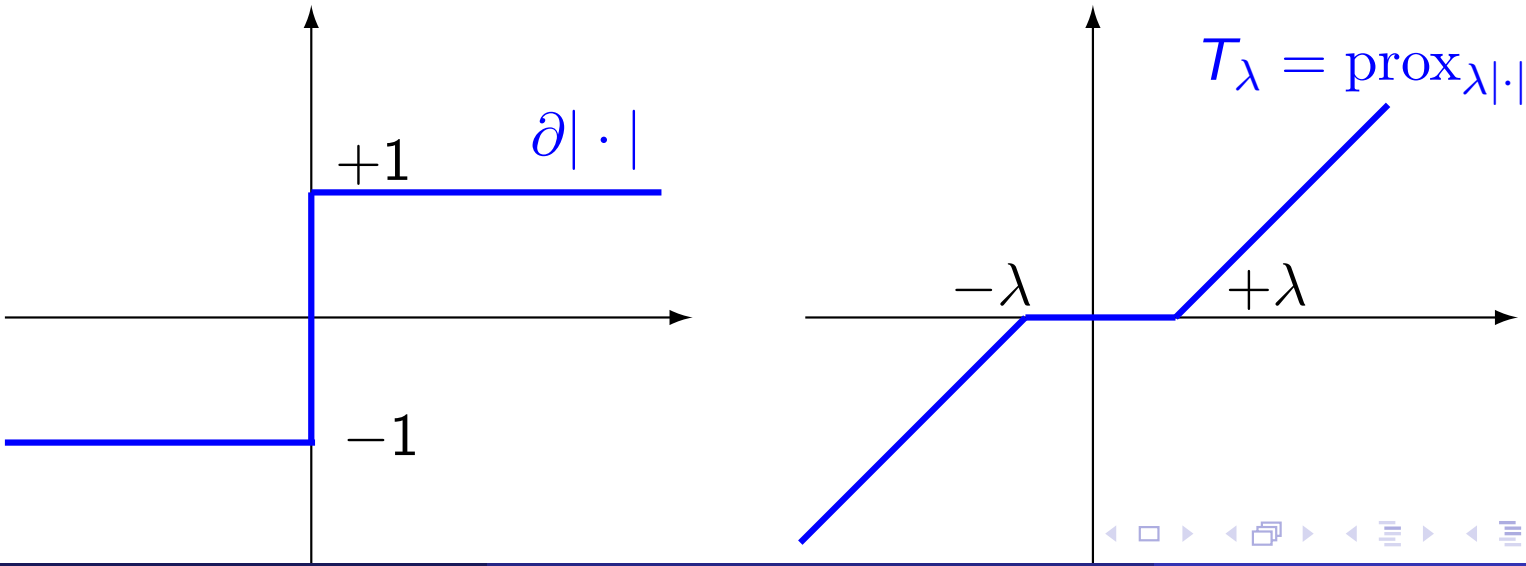
$$\begin{aligned}
0 \in \partial(f+g)(x) &\iff 0 \in \nabla f(x) + \partial g(x) \\
&\iff 0 \in s\nabla f(x) + s\partial g(x),\ s > 0. \\
&\iff x \in x + s\nabla f(x) + s\partial g(x) \\
&\iff x - s\nabla f(x) \in \Big(\mathrm{I} + s\partial g\Big)(x). \\
&\iff x = \mathrm{prox}_{sg}\Big(x - s\nabla f(x)\Big).
\end{aligned}$$

Forward-backward algorithm

$$(\mathrm{FB}) \begin{cases} x_0 \in \mathcal{H},\ 0 < s \leq \frac{1}{L} \\ x_{k+1} = \mathrm{prox}_{sg}\Big(x_k - s\nabla f(x_k)\Big) \end{cases}$$

- LASSO: $\min_{x\in\mathbb{R}^n} \underbrace{\|Ax-b\|_2^2}_{f(x)} + \underbrace{\lambda\|x\|_1}_{g(x)}$, with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

$$\mathrm{prox}_{\lambda\|\cdot\|_1} = \Big(T_\lambda(x_1), \ldots, T_\lambda(x_n)\Big)$$



Iterative Shrinkage-Thresholding Algorithm (ISTA)

$$\text{(FB)} \begin{cases} x_0 \in \mathcal{H},\ 0 < s \leq \frac{1}{L} \\ x_{k+1} = \text{prox}_{s\lambda\|\cdot\|_1}\Big(x_k - s\nabla f(x_k)\Big) \end{cases}$$

$h = f + \lambda\|\cdot\|_1$.

$$h(x_{k+1}) - h^* \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

- Possibility of a backtracking version.

Structured minimization: $\min_{\mathcal{H}}(f+g)$

- $f : \mathcal{H} \to \mathbb{R}$ convex, $\mathcal{C}^1$, ∇f L-Lipschitz continuous; $0 < s \leq \frac{1}{L}$.
- $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ convex, lower semicontinuous, proper.

Inertial Proximal Gradient algorithm

$$(\mathrm{IPG})_\alpha \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= \mathrm{prox}_{sg}\left(y_k - s\nabla f(y_k)\right) \end{cases}$$

- $\alpha = 3$: $(f+g)(x_k) - \min_{\mathcal{H}}(f+g) = \mathcal{O}(\frac{1}{k^2})$,
  Beck-Teboulle: FISTA (SIAM J. Imaging 2009).
- $\alpha > 3$: $(f+g)(x_k) - \min_{\mathcal{H}}(f+g) = o(\frac{1}{k^2})$, $x_k \rightharpoonup x_\infty \in S$,
  Chambolle-Dossal (JOTA 2015), Attouch-Peypouquet (SIOPT 2016).
- $\alpha \leq 3$: $(f+g)(x_k) - \min_{\mathcal{H}}(f+g) = \mathcal{O}\left(1/k^{\frac{2\alpha}{3}}\right)$.
  Apidopoulos-Aujol-Dossal (Math Prog '20), Attouch-Chbani-Riahi (COCV '18)

4. Understanding the acceleration phenomenon from the perspective of limiting ODEs

A continuous ODE associated to NAG

$$(\mathrm{NAG})_\alpha \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla f(y_k). \end{cases}$$

Question: Is there any continuous (in time) ODE which is the limit of $(\mathrm{NAG})_\alpha$ by taking the step size $s \to 0$?

Journal of Machine Learning Research 17 (2016) 1-43 Submitted 3/15; Revised 10/15; Published 9/16

A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights

Weijie Su SUW@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104, USA

Stephen Boyd BOYD@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Emmanuel J. Candès CANDES@STANFORD.EDU
Departments of Statistics and Mathematics
Stanford University
Stanford, CA 94305, USA

Su-Boyd-Candès model

Let us set $h = \sqrt{s}$. We have

$$x_{k+1} = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) - s\nabla f(y_k).$$

Hence,

$$\frac{x_{k+1} - x_k}{h} = \left(1 - \frac{\alpha}{k}\right)\frac{x_k - x_{k-1}}{h} - h\nabla f(y_k).$$

We introduce the Ansatz $x_k \simeq X(kh)$ for a smooth curve $X : [0, +\infty[\to \mathcal{H},\ t \mapsto X(t)$ with $t = kh = k\sqrt{s}$.

$$\frac{x_{k+1} - x_k}{h} = \dot{X}(t) + \frac{h}{2}\ddot{X}(t) + o(h).$$

$$\frac{x_k - x_{k-1}}{h} = \dot{X}(t) - \frac{h}{2}\ddot{X}(t) + o(h).$$

$$h\nabla f(y_k) = h\nabla f(X(t)) + o(h).$$

By identification with the coefficients of h, we get

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

Inertial dynamic with an asymptotic vanishing damping.

$$\lim_{s \to 0} \max_{0 \le k \le \frac{T}{h}} \|X(kh) - x_k\| = 0.$$

Su-Boyd-Candès model for NAG

Asymptotic Vanishing Damping

$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0.$$

- $\alpha \geq 3$: Su-Boyd-Candès (NIPS 2014), link with Nesterov

$$f(x(t)) - f^* = \mathcal{O}\left(\frac{1}{t^2}\right) \text{ as } t \to +\infty.$$

- $\alpha > 3$: Attouch-Chbani-Peypouquet-Redont (Math. Prog. 2018)

$$f(x(t)) - f^* = o\left(\frac{1}{t^2}\right),\ x(t) \rightharpoonup x_\infty \in S \text{ as } t \to +\infty.$$

- $\alpha \leq 3$: Apidopoulos-Aujol-Dossal (SIOPT 2018), Attouch-Chbani-Riahi (ESAIM COCV 2019)

$$f(x(t)) - f^* = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right) \text{ as } t \to +\infty.$$

Low versus high resolution ODE of NAG

Remark

- Gradient-based optimization algorithms can be studied from the perspective of limiting ODEs.
- Existing ODEs do not distinguish between two different algorithms: Nesterov's accelerated gradient method for strongly convex functions and Polyak's heavy-ball method.
- SDJS introduced a limiting process that uses high-resolution ODEs: take the step size s small but non-vanishing.
- High resolution ODEs are more accurate than low resolution ODE.

B. Shi, S. S. Du, M. I. Jordan, W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program., 2021.

Low versus high-resolution ODE for NAG-SC and HBF

$$\text{(HBF)}\ x_{k+1} = x_k - s\nabla f(x_k) + \beta(x_k - x_{k-1}).$$

$$\text{(NAG-SC)}\ x_{k+1} = x_k - s\nabla f(x_k) + \Big(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\Big)(x_k - x_{k-1}) - s\Big(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\Big)\Big(\nabla f(x_k) - \nabla f(x_{k-1})\Big).$$

- The low resolution ODE for (HBF) and (NAG-SC) is
$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \nabla f(x(t)) = 0.$$
- The high-resolution ODE for the (HBF) is
$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + (1+\sqrt{\mu s})\nabla f(x(t)) = 0.$$
- The high-resolution ODE for the (NAG-SC) is
$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \sqrt{s}\nabla^2 f(x(t))\dot{x}(t) + (1+\sqrt{\mu s})\nabla f(x(t)) = 0.$$
- The high-resolution ODE for the (NAG-C), convex case, is
$$\ddot{x}(t) + \tfrac{\alpha}{t}\dot{x}(t) + \sqrt{s}\nabla^2 f(x(t))\dot{x}(t) + (1+\tfrac{\alpha\sqrt{s}}{2t})\nabla f(x(t)) = 0.$$

5. The Ravine method: a little known method.

Ravine method. Link with Nesterov method

Gel′fand, I. M. [Gel′fand, Izrail′ Moiseevich]; **Cetlin, M. L. [Tsetlin, M. L.]**
The principle of nonlocal search in automatic optimization systems.
Soviet Physics Dokl. **6** 1961 192–194

The authors suggest a computational procedure for determining the minimum of general functions of n variables. At the kth step, a gradient method is used to pass from the point X_k to the point A_k. Then a line from A_{k-1} (obtained earlier) to A_k is extended an appropriate distance to the new approximation, X_{k+1}. A gradient method then leads to A_{k+1}, etc. Though no numerical results are provided, it is claimed that, for functions involving eight to ten variables, computing times are cut by factors of hundreds over straight search and gradient (not conjugate gradient) methods. *R. Kalaba*

Ravine method. Link with Nesterov method

In Nesterov accelerated gradient, (y_k) follows the Ravine method.

$$(\mathrm{NAG})_\alpha \left\{ \begin{array}{rcl} y_k & = & x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - s\nabla f(y_k) \end{array} \right.$$

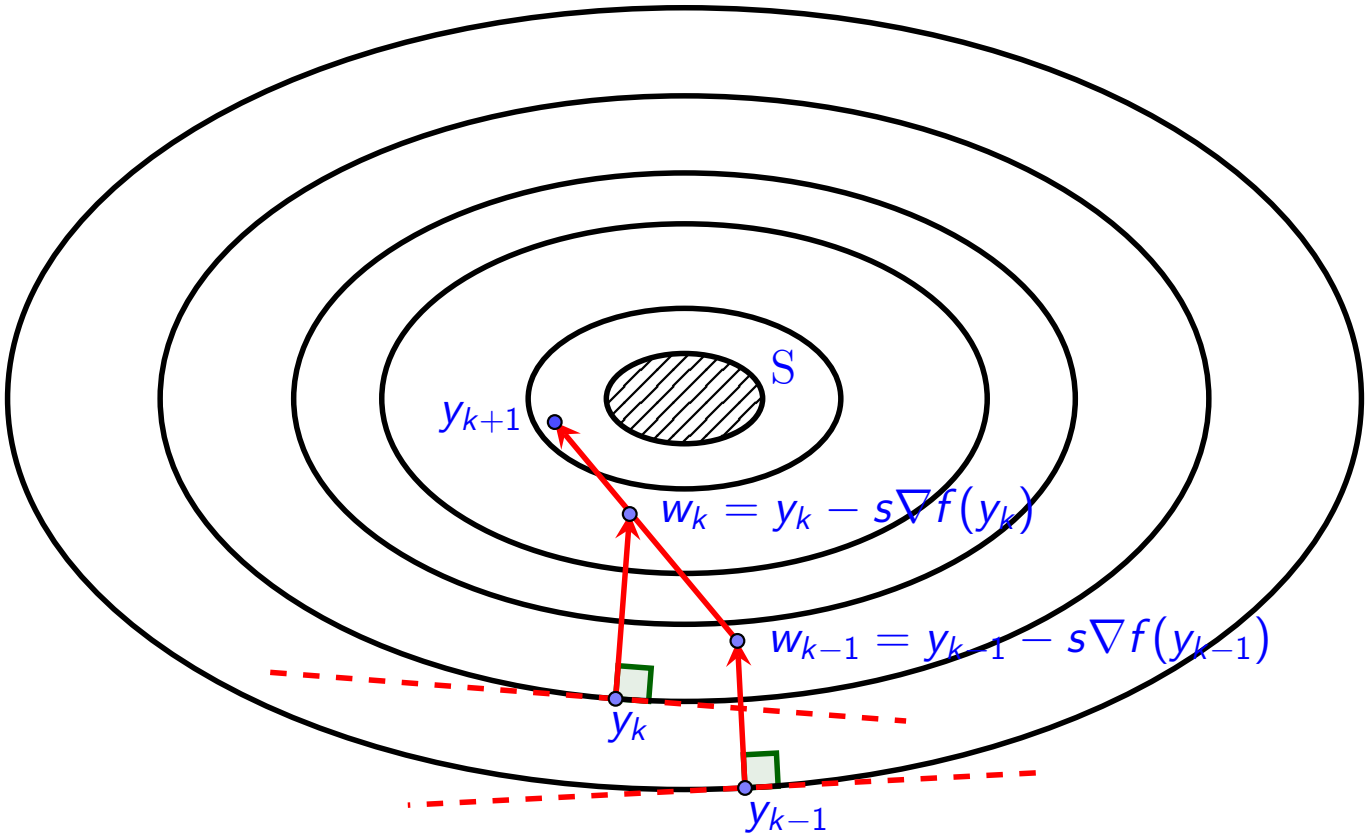
$$\begin{aligned} y_{k+1} &= x_{k+1} + \left(1 - \tfrac{\alpha}{k+1}\right)(x_{k+1} - x_k) \\ &= y_k - s\nabla f(y_k) + \left(1 - \tfrac{\alpha}{k+1}\right)\Big(y_k - s\nabla f(y_k) - (y_{k-1} - s\nabla f(y_{k-1}))\Big). \end{aligned}$$

$$(\mathrm{Ravine})_\alpha \left\{ \begin{array}{rcl} w_k & := & y_k - s\nabla f(y_k) \\ y_{k+1} & = & w_k + \left(1 - \frac{\alpha}{k+1}\right)(w_k - w_{k-1}). \end{array} \right.$$

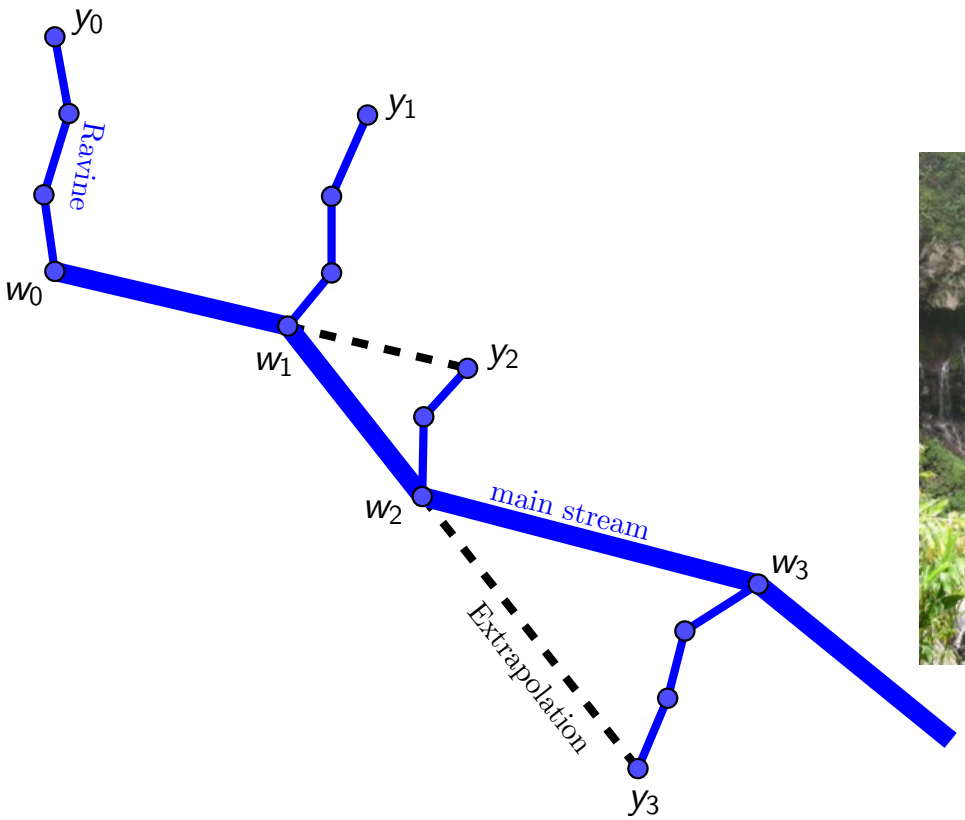
- $(\mathrm{NAG})_\alpha$ extrapolation step + gradient step.
- $(\mathrm{Ravine})_\alpha$ gradient step + extrapolation step.

Geometric view of the Ravine method

Gelfand, Tsetlin (1961), Nesterov (1983), Polyak (2018).



Interpretation of the Ravine method



Link with the Nesterov method

Conversely, if (y_k) follows the Ravine method, *i.e.*

$$(\text{Ravine})_\alpha \quad \left\{ \begin{array}{rcl} w_k & := & y_k - s\nabla f(y_k) \\ y_{k+1} & = & w_k + \left(1 - \frac{\alpha}{k+1}\right)(w_k - w_{k-1}). \end{array} \right.$$

then, (x_k) defined by $x_{k+1} = y_k - s\nabla f(y_k)$ follows $(\text{NAG})_\alpha$:

$$\begin{aligned} y_{k+1} &= y_k - s\nabla f(y_k) + \left(1 - \frac{\alpha}{k+1}\right)\left(y_k - s\nabla f(y_k) - (y_{k-1} - s\nabla f(y_{k-1}))\right) \\ &= x_{k+1} + \left(1 - \frac{\alpha}{k+1}\right)(x_{k+1} - x_k). \end{aligned}$$

$$(\text{NAG})_\alpha \quad \left\{ \begin{array}{l} y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = y_k - s\nabla f(y_k). \end{array} \right.$$

Low resolution ODE of the Ravine method

Equivalent forms of Ravine

$$y_{k+1} = y_k - s\nabla f(y_k) + \left(1 - \tfrac{\alpha}{k+1}\right)\left(y_k - s\nabla f(y_k) - (y_{k-1} - s\nabla f(y_{k-1}))\right)$$

$$y_{k+1} = y_k + \left(1 - \tfrac{\alpha}{k+1}\right)(y_k - y_{k-1}) - s\nabla f(y_k) - s\left(1 - \tfrac{\alpha}{k+1}\right)\left(\nabla f(y_k) - \nabla f(y_{k-1})\right)$$

$$\frac{(y_{k+1} - y_k) - (y_k - y_{k-1})}{h^2} + \frac{\alpha}{kh + h}\frac{y_k - y_{k-1}}{h} + \nabla f(y_k) + (1 - \tfrac{\alpha}{k+1})(\nabla f(y_k) - \nabla f(y_{k-1})) = 0.$$

Ansatz $y_k \approx Y(kh)$

Set $k = t/h$. As $h \to 0$, $Y(t) \approx y_{t/h} = y_k$, $\quad Y(t+h) \approx y_{(t+h)/h} = y_{k+1}$. Taylor expansion of $Y(t)$ at t gives

$$\ddot{Y}(t) + \tfrac{\alpha}{t}\dot{Y}(t) + \nabla f(Y(t)) + o(1) = 0.$$

Letting $h \to 0$ gives that $Y(\cdot)$ is a solution trajectory of $(\mathrm{AVD})_\alpha$

$$\ddot{Y}(t) + \tfrac{\alpha}{t}\dot{Y}(t) + \nabla f(Y(t)) = 0.$$

Theorem

The super-resolution ODE with temporal step-size $\sqrt{s}$ *of* (RAG) *gives the inertial dynamic with Hessian driven damping*

$$\ddot{Y}(t)+\frac{\alpha}{t}\dot{Y}(t)+\sqrt{s}\nabla^2 f(Y(t))\dot{Y}(t)+\left(1+\frac{\alpha\sqrt{s}}{2t}\right)\nabla f(Y(t))$$
$$+\frac{s}{2}\left(\frac{1}{6}Y^{(4)}(t)+\frac{\alpha}{3t}\dddot{Y}(t)-\frac{\alpha}{t}\nabla^2 f(Y(t))\dot{Y}(t)-\nabla^2 f(Y(t))\ddot{Y}(t)-\nabla^3 f(Y(t))\big(\dot{Y}(t),\dot{Y}(t)\big)\right)=0.$$

When neglecting the terms of order higher or equal to 2*, we recover the high-resolution ODE of* (RAG) *of order* 1

$$\ddot{Y}(t)+\frac{\alpha}{t}\dot{Y}(t)+\sqrt{s}\nabla^2 f(Y(t))\dot{Y}(t)+\left(1+\frac{\alpha\sqrt{s}}{2t}\right)\nabla f(Y(t))=0.$$

S. Adly, H. Attouch and J. M. Fadili. Comparative Analysis of Accelerated Gradient Algorithms for Convex Optimization: High and Super Resolution ODE Approach. HAL CNRS (2023).

Theorem

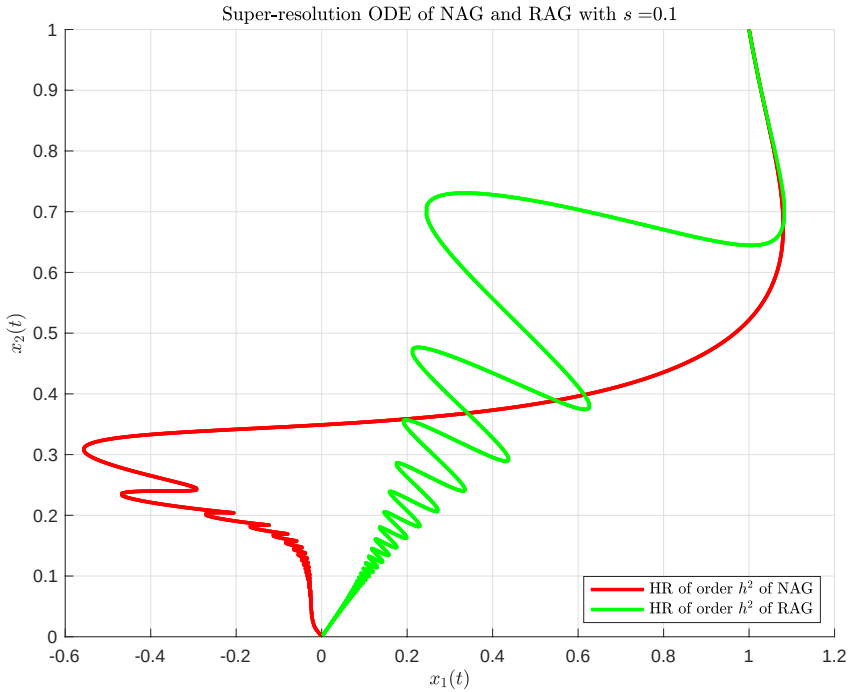
The super-resolution ODE with temporal step-size $\sqrt{s}$ of (NAG) *gives the inertial dynamic with Hessian driven damping*

$$\ddot{X}(t)+\frac{\alpha}{t}\dot{X}(t)+\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)+\left(1+\frac{\alpha\sqrt{s}}{2t}\right)\nabla f(X(t))$$
$$+\frac{s}{2}\left(\frac{1}{6}X^{(4)}(t)+\frac{\alpha}{3t}\dddot{X}(t)-\frac{\alpha}{t}\nabla^2 f(X(t))\dot{X}(t)-\nabla^2 f(X(t))\ddot{X}(t)+\nabla^3 f(X(t))\big(\dot{X}(t),\dot{X}(t)\big)\right)=0.$$

When neglecting the terms of order higher or equal to 2*, we recover the high-resolution ODE of* (NAG) *of order* 1

$$\ddot{X}(t)+\frac{\alpha}{t}\dot{X}(t)+\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)+\left(1+\frac{\alpha\sqrt{s}}{2t}\right)\nabla f(X(t))=0.$$

S. Adly, H. Attouch and J. M. Fadili. Comparative Analysis of Accelerated Gradient Algorithms for Convex Optimization: High and Super Resolution ODE Approach. HAL CNRS (2023).



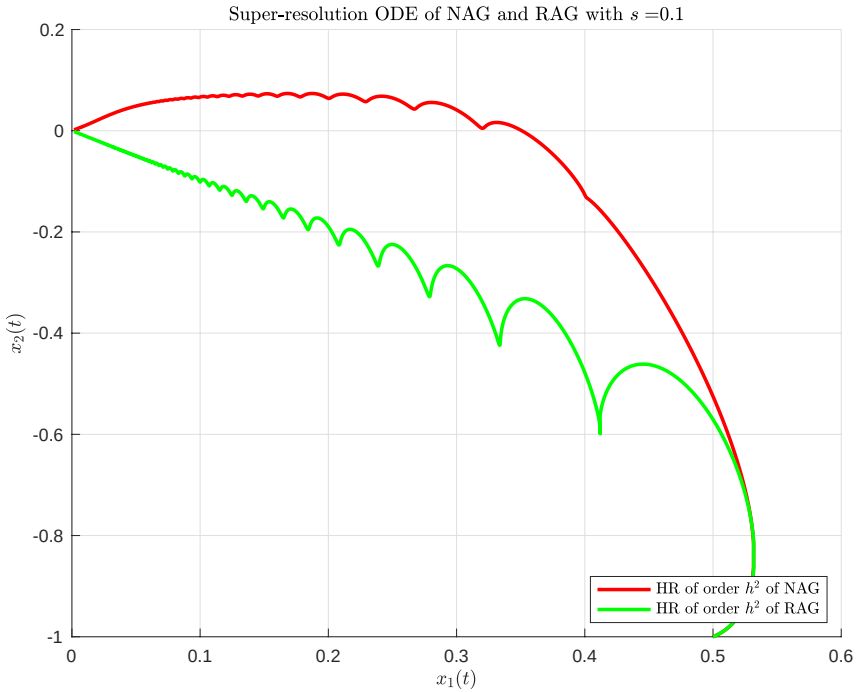
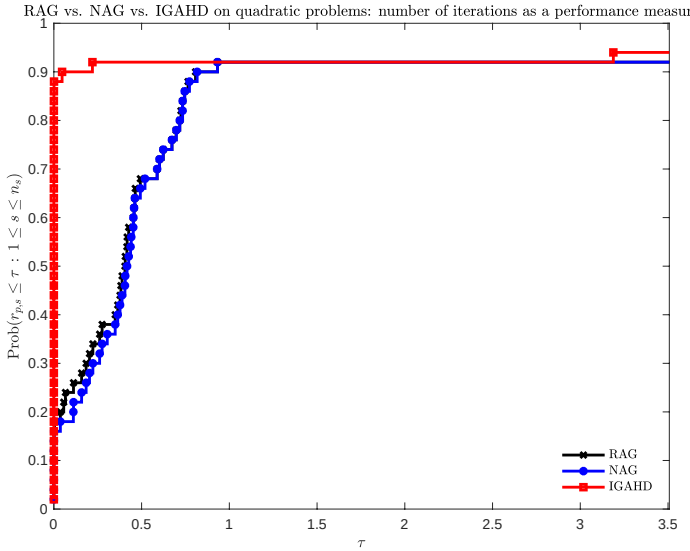
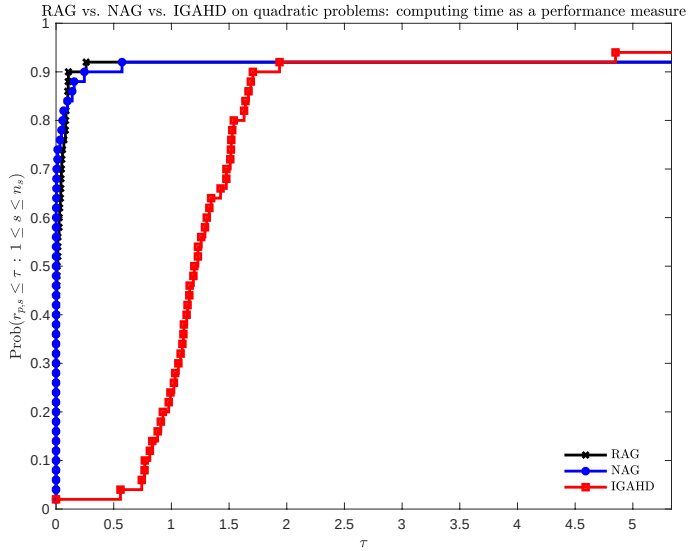
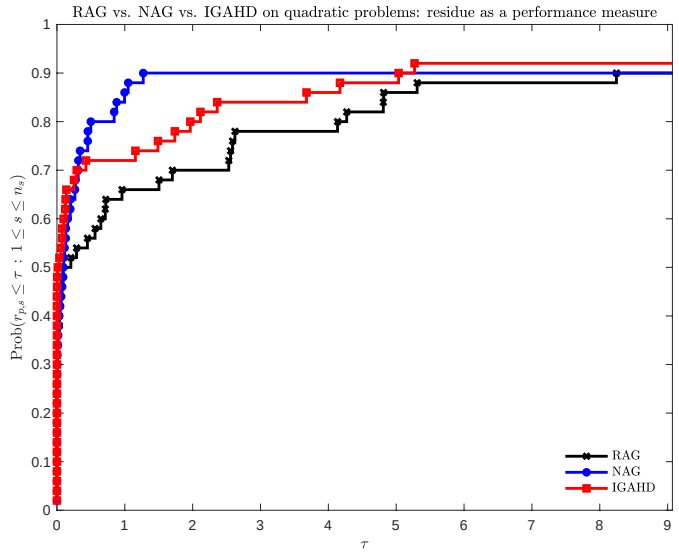


Figure: Trajectories of the super-resolution of order h^2 of NAG and RAG with $s = 0.1$ for different initial conditions.


RAG vs. NAG vs. IGAHD on quadratic problems: number of iterations as a performance measure
$\text{Prob}(r_{p,s} \leq \tau : 1 \leq s \leq n_s)$
τ
RAG
NAG
IGAHD


RAG vs. NAG vs. IGAHD on quadratic problems: computing time as a performance measure
$\text{Prob}(r_{p,s} \leq \tau : 1 \leq s \leq n_s)$
τ
RAG
NAG
IGAHD


RAG vs. NAG vs. IGAHD on quadratic problems: residue as a performance measure
$\text{Prob}(r_{p,s} \leq \tau : 1 \leq s \leq n_s)$
τ
RAG
NAG
IGAHD

5. On the limit form of the Su-Boyd-Candès dynamic version of Nesterov's accelerated gradient method when the viscous parameter becomes large

Based on a paper by S.A. and H. Attouch, (2023).

Analysis of Su-Boyd-Candès Dynamics for Large α

Asymptotic Vanishing Damping (AVD)

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

- A natural approach is to consider the limit as $\alpha \to +\infty$ in the dynamic (AVD).
- However, this approach, as shown below, provides limited insights into the asymptotic behavior of trajectories when α is large.
- Instead, we need to employ a more sophisticated analysis, as indicated by the following result obtained from an elementary energy analysis.

Analysis of the Su-Boyd-Candès dynamic for large α

Proposition

Take $x_0 \in \mathcal{H}$ and $x_1 \in \mathcal{H}$. For each $\alpha \geq 3$, let $x_\alpha : [t_0, +\infty[\to \mathcal{H}$ be the solution trajectory of the Cauchy problem

$$\begin{cases} \ddot{x}_\alpha(t) + \frac{\alpha}{t}\dot{x}_\alpha(t) + \nabla f(x_\alpha(t)) = 0 \\ x_\alpha(t_0) = x_0,\ \dot{x}_\alpha(t_0) = x_1 \end{cases}$$

Then,

- For each $t \geq t_0$, $x_\alpha(t) \to x_0$ strongly in $\mathcal{H}$ as $\alpha \to +\infty$.
- For each T finite, $T > t_0$, we have
$$\sup_{t\in[t_0,T]} \|x_\alpha(t) - x_0\| \leq \frac{M_T}{\alpha - 1},$$

where $M_T = t_0\|x_1\| + T^2\left(\|\nabla f(x_0)\| + L_r T\left(2(f(x_0) - f^*) + \|x_1\|^2\right)^{\frac{1}{2}}\right)$, and L_r is equal to the Lipschitz constant of ∇f on the ball centered at the origin and of radius $r = \|x_0\| + T\left(2(f(x_0) - f^*) + \|x_1\|^2\right)^{\frac{1}{2}}$.

The time rescaling approach

$$\begin{cases} \ddot{x}_\alpha(t) + \frac{\alpha}{t}\dot{x}_\alpha(t) + \nabla f(x_\alpha(t)) = 0 \\ x_\alpha(t_0) = x_0,\ \dot{x}_\alpha(t_0) = x_1 \end{cases}$$

We set $y_\alpha(s) = x_\alpha\left(\sqrt{2(\alpha+1)s}\right)$, which satisfies the differential equation

$$\begin{cases} \frac{2s}{\alpha+1}\ddot{y}_\alpha(s) + \dot{y}_\alpha(s) + \nabla f(y_\alpha(s)) = 0 \\ y_\alpha\left(\frac{t_0^2}{2(\alpha+1)}\right) = x_0,\ \dot{y}_\alpha\left(\frac{t_0^2}{2(\alpha+1)}\right) = \frac{\alpha+1}{t_0}x_1 \end{cases}$$

The time rescaling approach

Theorem

Take $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$. For each $\alpha > 0$, let $x_\alpha : [t_0, +\infty[\to \mathcal{H}$ be the solution trajectory of

$$(\mathrm{AVD})_\alpha \quad \ddot{x}_\alpha(t) + \frac{\alpha}{t}\dot{x}_\alpha(t) + \nabla f(x_\alpha(t)) = 0,$$

which satisfies the Cauchy data $x_\alpha(t_0) = x_0$ and $\dot{x}_\alpha(t_0) = x_1$. Consider the sequence of rescaled trajectories (y_α), $y_\alpha : [\frac{t_0^2}{2(\alpha+1)}, +\infty[\to \mathcal{H}$ defined by

$$y_\alpha(s) = x_\alpha\left(\sqrt{2(\alpha+1)s}\right).$$

Then, the following results are satisfied.

(i) For each $\alpha > 0$, y_α satisfies the differential equation

$$\frac{2s}{\alpha+1}\ddot{y}_\alpha(s) + \dot{y}_\alpha(s) + \nabla f(y_\alpha(s)) = 0, \tag{2}$$

with the Cauchy data $y_\alpha\left(\frac{t_0^2}{2(\alpha+1)}\right) = x_0$ and $\dot{y}_\alpha\left(\frac{t_0^2}{2(\alpha+1)}\right) = \frac{\alpha+1}{t_0}x_1$.

The time rescaling approach

(ii) Suppose now that $\mathcal{H}$ is a finite dimensional Hilbert space. Let us extend the function y_α to $[0, +\infty[$ by setting

$$\tilde{y}_\alpha = y_\alpha \text{ on } [\frac{t_0^2}{2(\alpha+1)}, +\infty[, \quad \tilde{y}_\alpha \equiv x_0 \text{ on } [0, \frac{t_0^2}{2(\alpha+1)}].$$

When α tends to $+\infty$, the sequence $(\tilde{y}_\alpha)$ converges uniformly on the bounded sets of $[0, +\infty[$ to the solution of the following continuous steepest descent

$$\dot{y}(s) + \nabla f(y(s)) = 0, \tag{3}$$

that satisfies $y(0) = x_0$.

- The convexity of f is not required.
- The Cauchy data on the velocity $\dot{y}_\alpha\left(\frac{t_0^2}{2(\alpha+1)}\right) = \frac{\alpha+1}{t_0}x_1$ explodes as $\alpha \to +\infty$. This induces singular perturbation phenomenon.

Time rescaling approach: the convex case

Theorem

Suppose that $\mathcal{H}$ is a general real Hilbert space, and that $f : \mathcal{H} \to \mathbb{R}$ is a **convex** differentiable function.
Then, as $\alpha \to +\infty$, the sequence of rescaled functions $(\tilde{y}_\alpha)$ converges uniformly to y on the bounded intervals of $[0, +\infty[$, where y is the solution of the continuous steepest descent

$$\dot{y}(s) + \nabla f(y(s)) = 0,$$

that satisfies $y(0) = x_0$. Precisely, for each $T > 0$, there exists a constant C_T such that

$$\sup_{s \in [0,T]} \|\tilde{y}_\alpha(s) - y(s)\| \leq \frac{C_T}{\sqrt{\alpha + 1}}.$$

Open question. In the convex case, is the uniform convergence property valid on $[0, +\infty[$?

Time rescaling approach: the convex case

Consider $f(x_1, x_2) = \lambda_1 x_1^2 + \lambda_2 x_2^2$ with $\lambda_1 = 0.02$ and $\lambda_2 = 0.005$ with the initial condition $x_0 = (2, 2)$ and $x_1 = (1, 1)$. Note that f is of the form $f(x) = \langle x, Ax \rangle$ with $A = \mathrm{diag}([\lambda_1, \lambda_2])$.

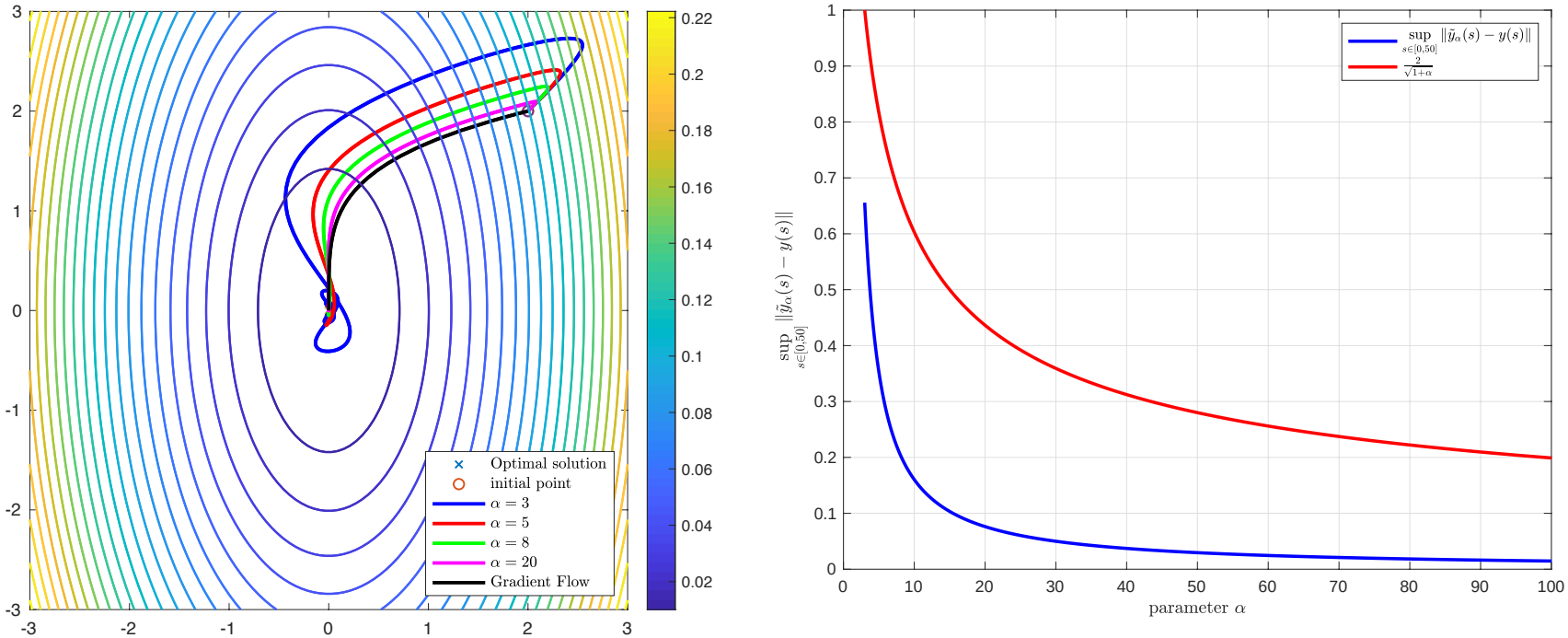
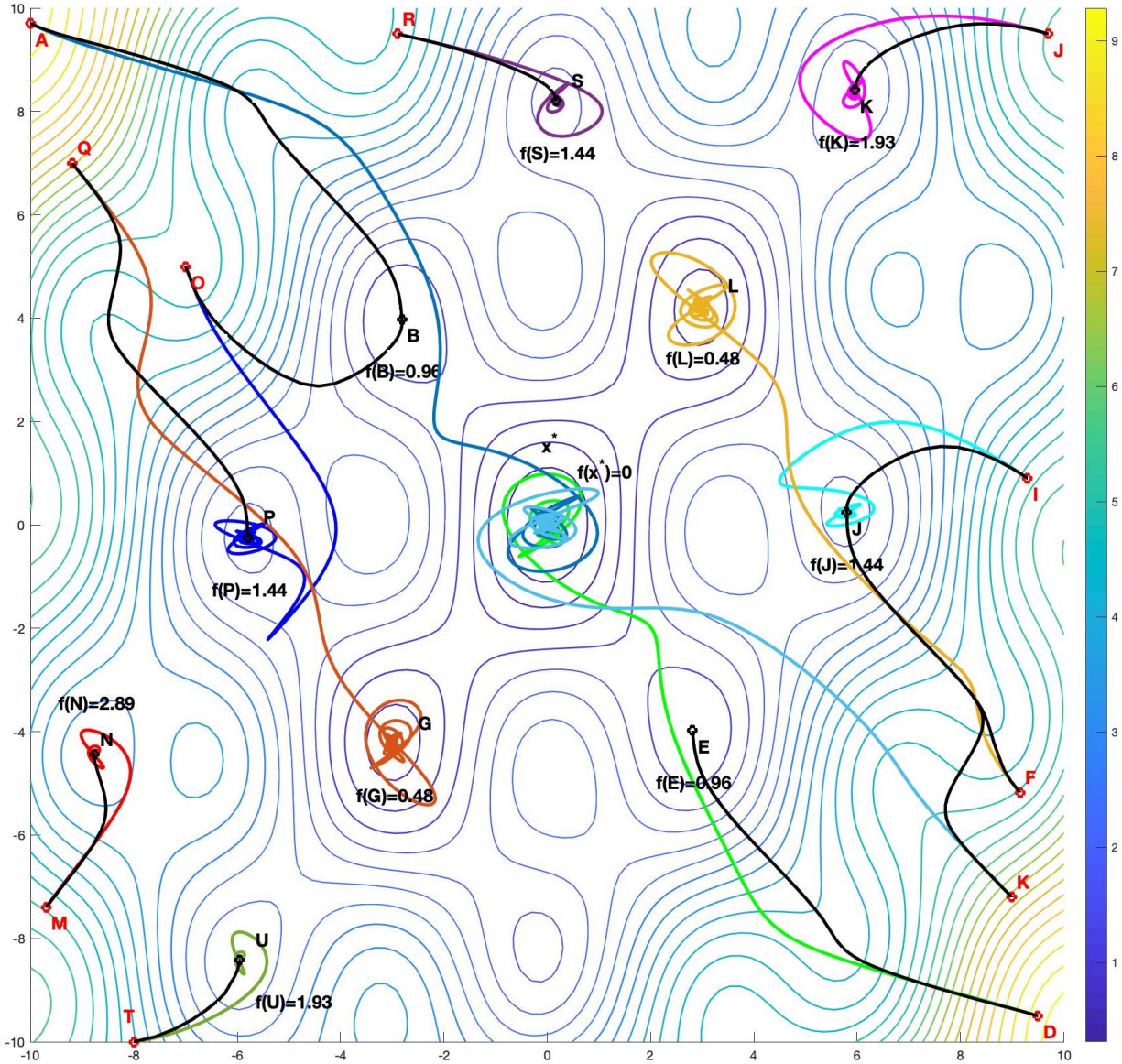


Figure: Illustration on a quadratic convex function.

Illustration on a nonconvex function: exploration of local minima of f.



6. Perspective, open questions

Some open questions

- Comparaison de IGAHD, RAV et NAG (papier Adly-Attouch-Fadili)
- High-resolution ODE
- Large α Adly-Attouch.
- Doubly nonlinear.

Some open questions

$$\left\{\begin{array}{rcl} y_k & = & x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - s\nabla f(y_k) \end{array}\right.$$

- Convergence of NAG's iterates in the critical case $\alpha = 3$ (except in 1D).
- How to tune efficiently the vanishing damping coefficient $\alpha > 3$?
- Extension to nonconvex case: KL theory only works in a finite dimensional framework and for autonomous systems. This is why it cannot be applied directly to $(\mathrm{AVD})_\alpha$ which is a non-autonomous system.
- We have already mentioned that when f is strongly convex, the convergence rate of values is $\mathcal{O}\left(1/t^{\frac{2\alpha}{3}}\right)$, and becomes therefore arbitrarily fast (in the scale of powers of $1/t$) with α large. To exploit this result in the case of a general convex differentiable function $f : \mathcal{H} \to \mathbb{R}$, a natural idea is to use Tikhonov's regularization method.

Thank you very much for your attention

References

- S. Adly, H. Attouch, *On the limit form of the Su-Boyd-Candès dynamic version of Nesterov's accelerated gradient method when the viscous parameter becomes large*, (2021).
- S. Adly, H. Attouch, *Fast optimization via time-scale analysis of inertial dynamics with Hessian-driven damping*, (2022).
- S. Adly, H. Attouch, J. Fadili, *Comparative Analysis of Accelerated Gradient Algorithms for Convex Optimization: High and Super Resolution ODE Approach* (2023).
- S. Adly, H. Attouch, *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping*, SIAM J. Optim., 30(3) (2020), pp. 2134–2162.
- S. Adly, H. Attouch, *Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping*, J. Conv. Analysis, 28 (2) (2021), hal-02557928.
- S. Adly, H. Attouch, *Finite convergence of proximal-gradient inertial algorithms with dry friction damping*, Math. Program., (2020), hal-02388038.

References


C.D. Alecsa, S. László, T. Pinta, *An extension of the second order dynamical system that models Nesterov's convex gradient method*, Applied Mathematics and Optimization, (2020), arXiv:1908.02574v1.


F. Alvarez, H. Attouch, J. Bolte, P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics*, J. Math. Pures Appl., **81**(8) (2002), pp. 747–779.


V. Apidopoulos, J.-F. Aujol, Ch. Dossal, *The differential inclusion modeling the FISTA algorithm and optimality of convergence rate in the case* $b \leq 3$, SIAM J. Optim., 28(1) (2018), pp. 551—574.


V. Apidopoulos, J.-F. Aujol, Ch. Dossal, *Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule*, Math. Program., 180 (2020), pp. 137–156.


H. Attouch, R.I. Boţ, E.R. Csetnek, *Fast optimization via inertial dynamics with closed-loop damping*, Journal of the European Mathematical Society (JEMS), 2021, hal-02910307.

References


H. Attouch, A. Cabot, *Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity*, J. Differential Equations, 263 (9), (2017), pp. 5412–5458.


H. Attouch, A. Cabot, *Convergence of a relaxed inertial proximal algorithm for maximally monotone operators*, Mathematical Programming, 184 (2020), pp. 243–287.


H. Attouch, A. Cabot, *Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions*, Applied Mathematics and Optimization, special issue on Games, Dynamics and Optimization, 80 (3) (2019), pp. 547-598.


H. Attouch, A. Cabot, Z. Chbani, H. Riahi, *Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization*, JOTA, 179 (2018), No.1, pp. 1-36 .


H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *First order optimization algorithms via inertial systems with Hessian driven damping*, Math. Program. (2020).

References

- H. Attouch, Z. Chbani, J. Peypouquet, P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program. Ser. B, 168 (2018), pp. 123–175.
- H. Attouch, Z. Chbani, H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$*. ESAIM COCV, 25 (2019), DOI:10.1051/cocv/2017083.
- H. Attouch, Z. Chbani, H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM J. Optim., 29 (3) (2019), pp. 2227–2256.
- H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Mathematics, Springer, (2011).
- A. Beck, M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), No. 1, pp. 183–202.

References


P. Bégout, J. Bolte, M. A. Jendoubi, *On damped second-order gradient systems*, Journal of Differential Equations, vol. 259, nÂ° 7-8, 2015, pp. 3115–3143.


R. I. Boţ, E. R. Csetnek, *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM J. Control Optim., 54 (2016), pp. 1423-1443.


R. I. Boţ, E. R. Csetnek, S.C. László, *Approaching nonsmooth nonconvex minimization through second order proximal-gradient dynamical systems*, J. Evol. Equ., 18(3) (2018), pp. 1291–1318.


R. I. Boţ, E. R. Csetnek, S.C. László, *Tikhonov regularization of a second order dynamical system with Hessian damping*, Math. Program., DOI:10.1007/s10107-020-01528-8.


R. I. Boţ, E. R. Csetnek, S.C. László, *An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions*, EURO J. Comp. Optim., 4(1) (2016), 3–25.

References


H. Brézis, *Opérateurs maximaux monotones dans les espaces de Hilbert et équations d'évolution*, Lecture Notes 5, North Holland, (1972).


A. Cabot, H. Engler, S. Gadat, *On the long time behavior of second order differential equations with asymptotically small dissipation*, Trans. Amer. Math. Soc., 361 (2009), pp. 5983–6017.


C. Castera, J. Bolte, C. Févotte, E. Pauwels, *An Inertial Newton Algorithm for Deep Learning*. 2019. HAL-02140748.


A. Chambolle, Ch. Dossal, *On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm*, J. Opt. Theory Appl., 166 (2015), pp. 968–982.


D. Davis, W. Yin, *Convergence rate analysis of several splitting schemes*, In: Splitting methods in communication, imaging, science, and engineering, Sci. Comput., pp. 115–163. Springer, (2016).

References


D. Davis, W. Yin, *Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions*, Math. Oper. Res. 42(3), pp. 783–805 (2017).


A. Haraux, M. A. Jendoubi, *Convergence of solutions of second-order gradient-like systems with analytic nonlinearities*, J. Differential Equations, **144** (2), (1999), pp 313–320.


A. Haraux, M. A. Jendoubi, *The Convergence Problem for Dissipative Autonomous Systems*, Classical Methods and Recent Advances, Springer, 2015.


T. Lin, M. I. Jordan, *A Control-Theoretic Perspective on Optimal High-Order Optimization*, arXiv:1912.07168v1 [math.OC] Dec 2019.


S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in: *Les Équations aux Dérivées Partielles*, pp. 87–89, Éditions du centre National de la Recherche Scientifique, Paris 1963.

References

- S. Łojasiewicz, *Sur la géométrie semi- et sous-analytique*, Ann. Inst. Fourier **43**, (1993), 1575-1595.
- M. Muehlebach, M. I. Jordan, *A Dynamical Systems Perspective on Nesterov Acceleration*, (2019), arXiv:1905.07436
- A.S. Nemirovsky, D.B. Yudin, *Problem complexity and method efficiency in optimization*, John Wiley and Sons, 1983.
- Y. Nesterov, *A method of solving a convex programming problem with convergence rate O(1/k2)*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.
- Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer, 2004.

References

B. T. Polyak, *Introduction to Optimization*, New York, Optimization Software, 1987.

B. Shi, S. S. Du, M. I. Jordan, W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program., 2021 https://doi.org/10.1007/s10107-021-01681-8.

W. Su, S. Boyd, E. J. Candès, *A Differential Equation for Modeling Nesterov's Accelerated Gradient Method*, Advances in Neural Information Processing Systems **27** (NIPS 2014).