Fast Algorithmic Methods for Optimization and Learning

1. The convex unconstrained case

Samir Adly

Université de Limoges, Laboratoire XLIM
Email: samir.adly@unilim.fr

Master 2 ACSYON

Université de Limoges (2024-2025).

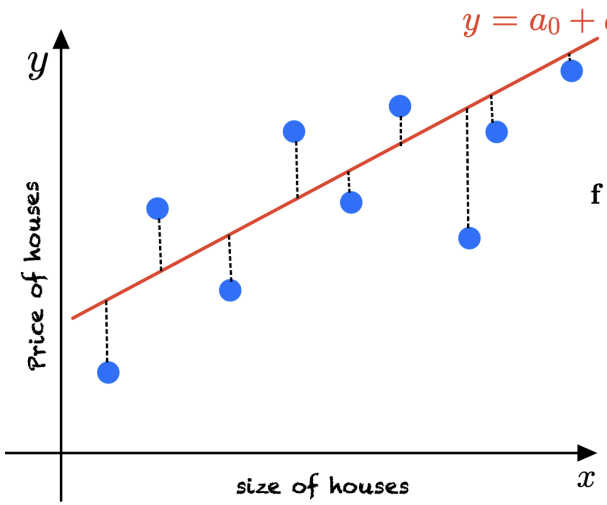
1. Some Motivation from Machine Learning

Examples of learning tasks

- Optical Character Recognition (OCR).
- Text or document classification, spam detection.
- Medical diagnosis: algorithms learn to distinguish healthy from abnormal.
- Fraud detection: credit cards, smartphones, network intrusion.
- Recommendation systems, search engines.
- Unassisted control of vehicles (robot, navigation).

Linear Regression Example

Size of house(sqm.)	Price of house
64	$ 200,000
80	$ 250,000
63	$ 230,000
100	$ 320,000
128	$ 300,000
144	$ 450,000
...	....
...	....
...	....
81	??



Cost function :

$$f(a) = \frac{1}{2m} \sum_{i=1}^{m} \left(a_0 + a_1 x^{(i)} - y_i \right)^2$$

- Least square method

$$\min_{x \in \mathbb{R}^n} f(x) := \|Ax - b\|_2^2,$$

with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

- Solve the normal equations

$$A^T A x = A^T b.$$

2. 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.

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

$$f(x^*) = f^*.$$

- Optimality condition: solve the nonlinear equation

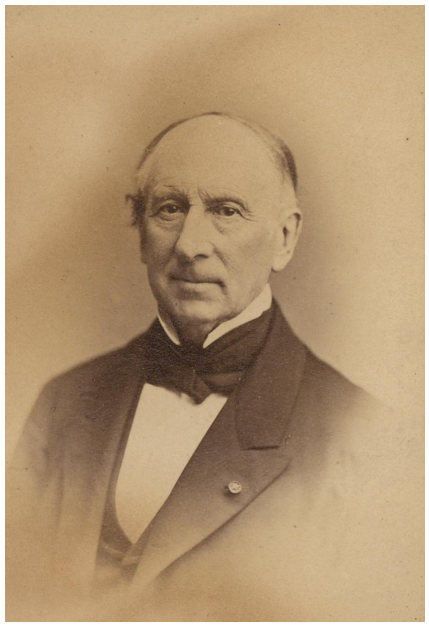
$$\nabla f(x) = 0.$$

Gradient Descent Method (Steepest Descent Method)

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

with $\alpha_k > 0$: 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).

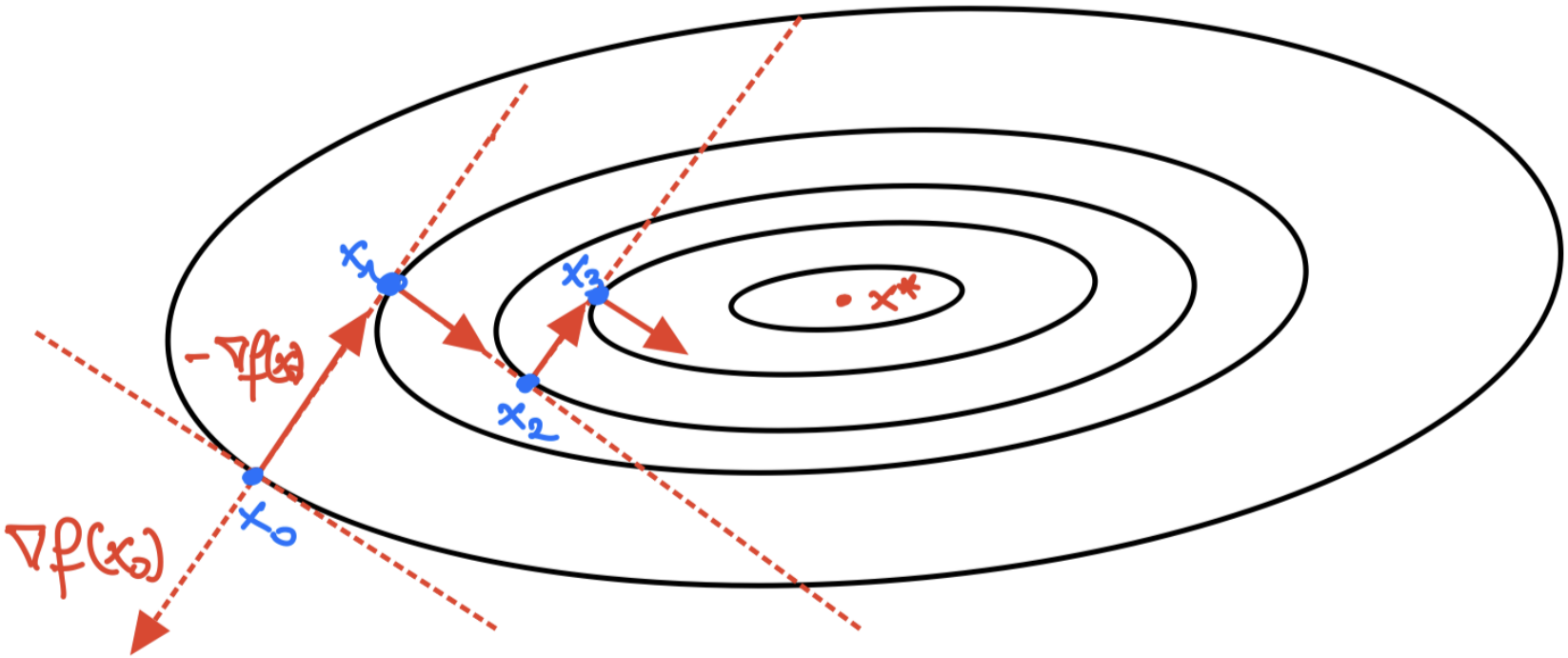
[3] A. L. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes rendus, Ac. Sci. Paris, **25**, 536–538 (1847).

[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)

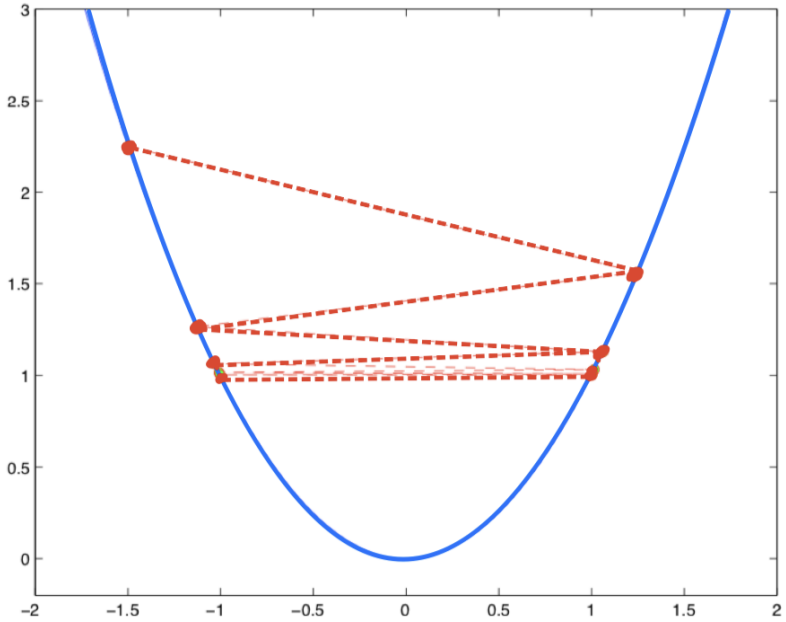
Gradient Descent Method: How to choose the step length?

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

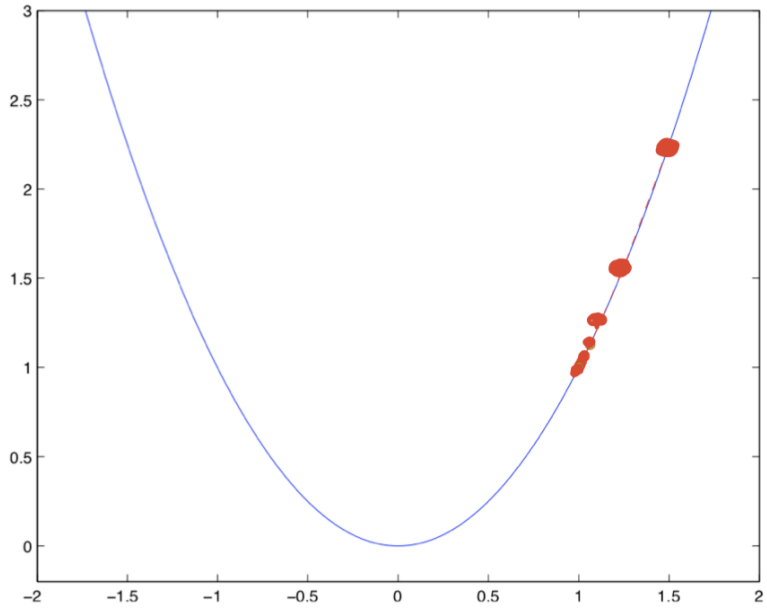


Gradient Descent Method: How to choose the step length?

Big step-size



small step-size



$$\alpha_k = 2 + \frac{3}{2^{k+1}} \qquad p_k = (-1)^{k+1} \qquad \alpha_k = \frac{1}{2^{k+1}}$$

How to choose the step length?

$$\begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = x_k + \alpha_k p_k \end{cases}$$

$p_k \in \mathbb{R}^n$ is a descent direction, i.e. $\langle \nabla f(x_k), p_k \rangle < 0$.

$p_k = -\nabla f(x_k)$ is the steepest descent direction.

- **Two strategies**:
  - **Search Direction Methods**:
    - Exact line search: compute the exact α_k.
    - Inexact line search: find an acceptable α_k (Wolfe conditions or Armijo).
  - **Trust Region Methods**: use information on f to construct a model function m_k whose behavior near the current point x_k is similar to f. The search for a minimiser of m_k is constrained to some region (the trust region).

Steepest descent method for a quadratic problem with exact line search

- Let the quadratic objective $f(x)$ be given by:

$$f(x) = \frac{1}{2} x^T Q x - b^T x,$$

  where Q is symmetric and positive definite.
- The gradient $g(x) = \nabla f(x) = Qx - b$ and the minimiser x^* is the unique solution of $Qx = b$. $f_k = f(x_k)$ and $g_k = g(x_k)$.
- The exact step length that $\min_{\alpha > 0} \phi(\alpha) := f(x_k - \alpha g_k)$ must satisfy:

$$\phi'(\alpha_k) = \frac{d}{d\alpha} f(x_k - \alpha_k g_k) = 0 \Longleftrightarrow \alpha_k = \frac{g_k^T g_k}{g_k^T Q g_k}.$$

- If we use the exact minimiser α_k, the steepest descent iteration is

$$x_{k+1} = x_k - \left(\frac{g_k^T g_k}{g_k^T Q g_k} \right) g_k.$$

Steepest descent method for a quadratic problem with exact line search

- Note that $g_k^T g_{k+1} = 0$ so successive search directions (gradients) are perpendicular.
- This causes the characteristic "zig-zag" pattern of the steepest descent method.

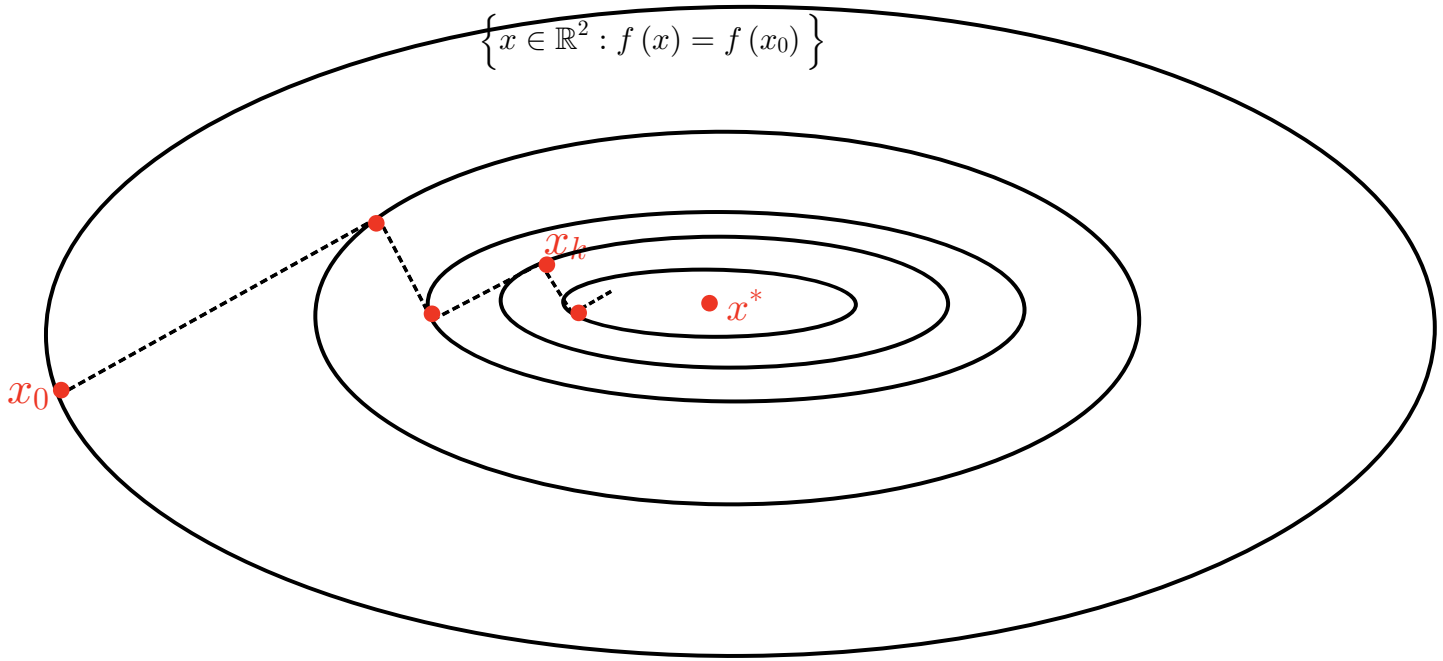


Figure: Typical Steepest Descent steps

- We introduce the weighted norm

$$\|x\|_Q = \sqrt{x^T Q x}, \quad Q = Q^T > 0.$$

- When the steepest descent method with exact line search is applied to a (convex) quadratic function, we have

$$\|x_{k+1} - x^*\|_Q^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \|x_k - x^*\|_Q^2 .$$

  with $\kappa := \kappa(Q) = \lambda_{\max}/\lambda_{\min}$ the condition number.
- We have for the values

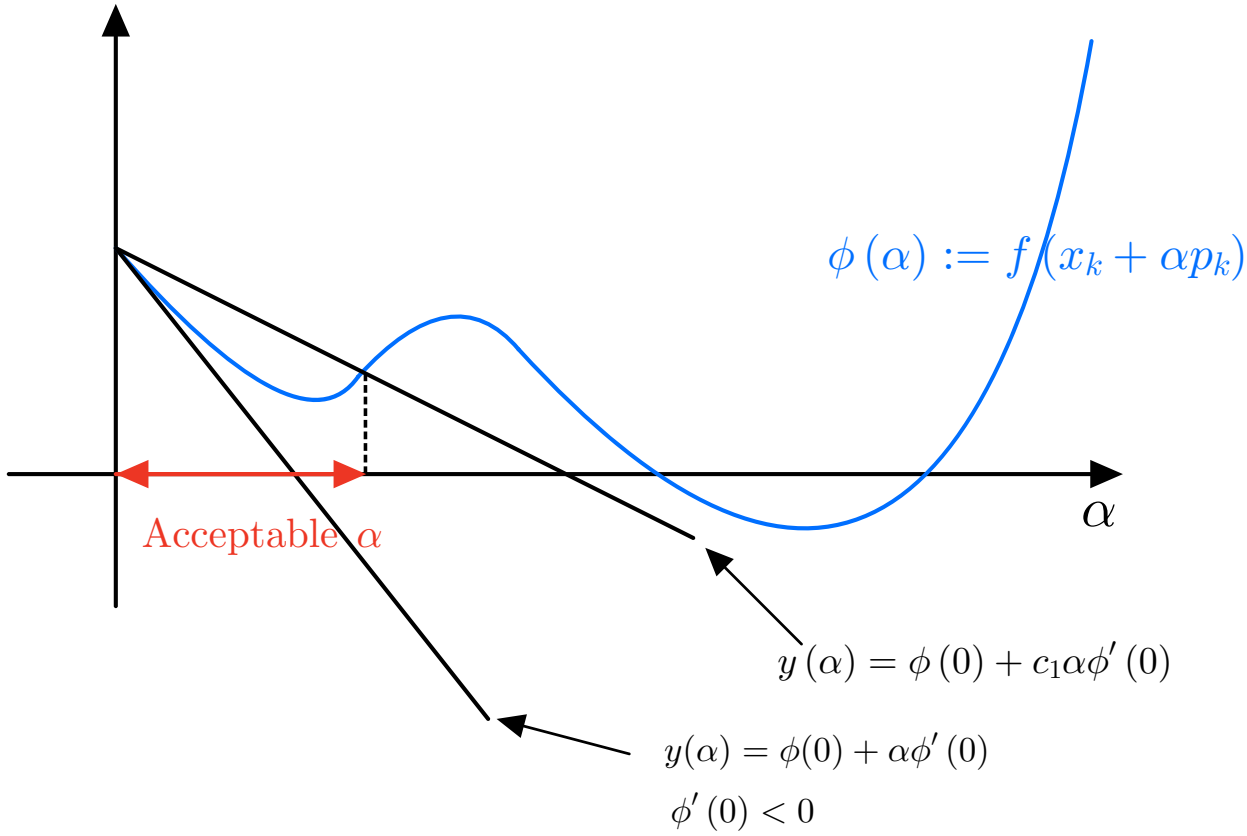
$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right) \Big(f(x_k) - f(x^*)\Big).$$

Inexact line search: the Wolfe Conditions

One widely used inexact line search condition is given by:

$$\phi(\alpha) \leq \phi(0) + c_1 \alpha \phi'(0),$$

for some constant $c_1 \in (0, 1)$.

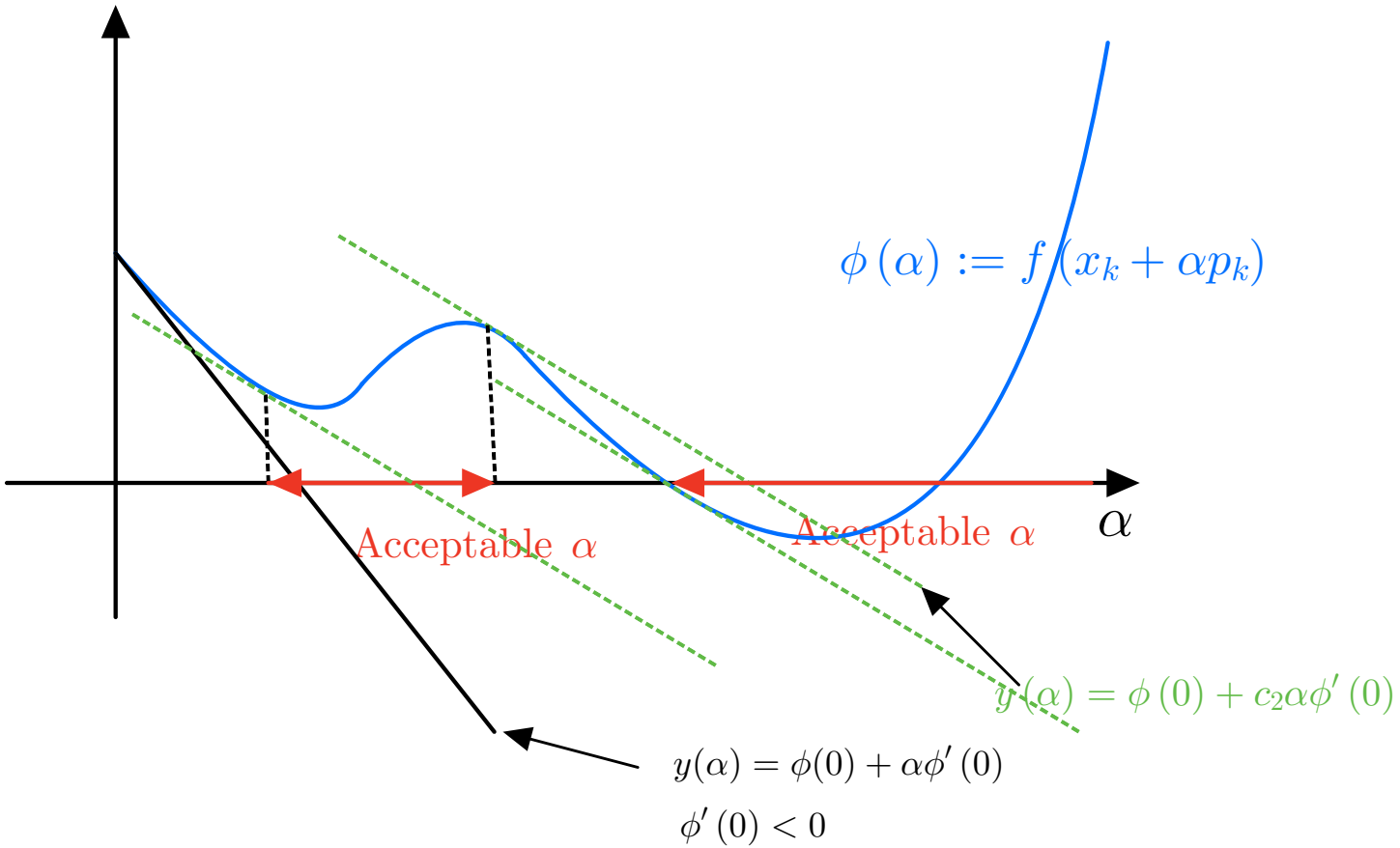


The curvature condition

The curvature condition is given by:

$$\phi'(\alpha) \geq c_2\phi'(0) \iff p_k^T \nabla f(x_k + \alpha p_k) \geq c_2 p_k^T g_k,$$

for some constants $0 < c_1 < c_2 < 1$.



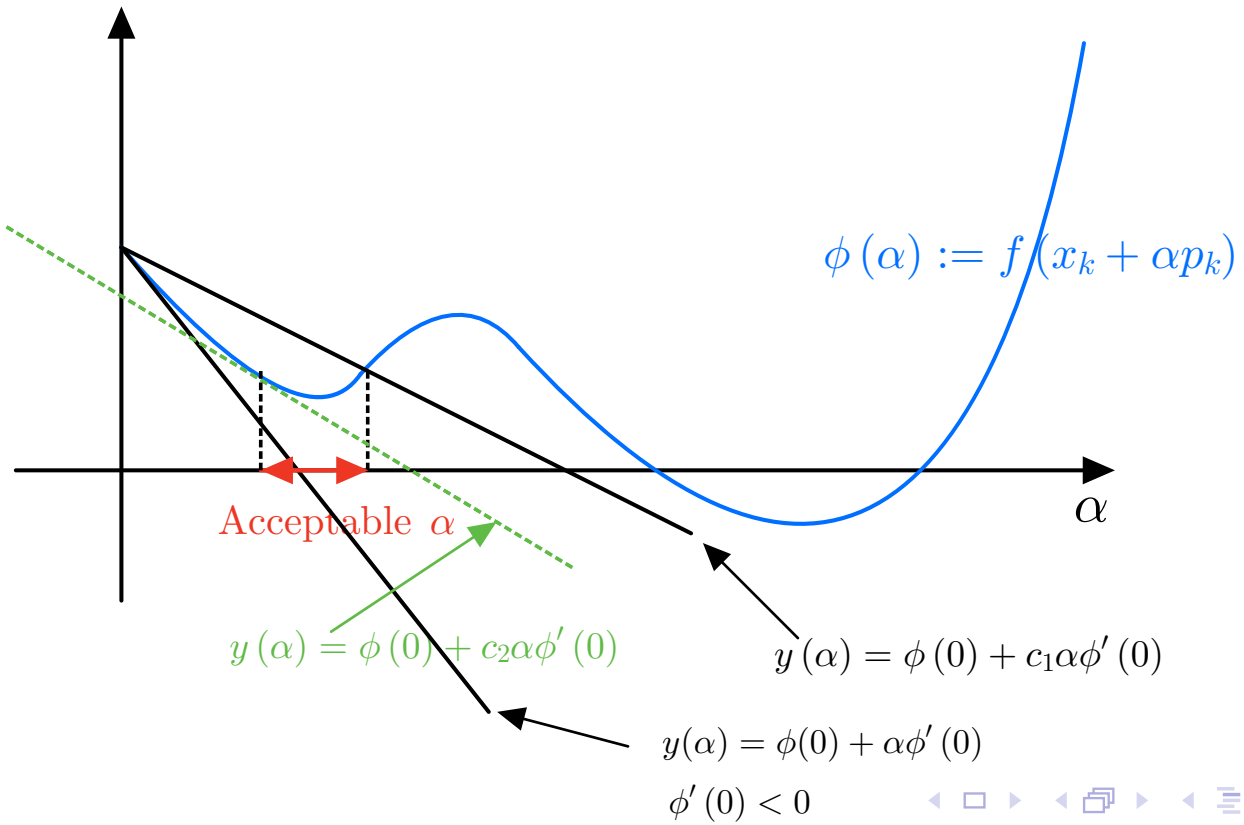
The Wolfe conditions

These two conditions are known collectively as the **Wolfe conditions**.

$$\phi(\alpha) \leq \quad \phi(0) + c_1 \alpha \phi'(0), \tag{W$_1$}$$

$$\phi'(\alpha) \geq \quad c_2 \phi'(0), \tag{W$_2$}$$

with $0 < c_1 < c_2 < 1$.



The Wolfe conditions

$$\phi(\alpha) \leq \quad \phi(0) + c_1 \alpha \phi'(0), \tag{W$_1$}$$
$$\phi'(\alpha) \geq \quad c_2 \phi'(0). \tag{W$_2$}$$

with $0 < c_1 < c_2 < 1$.

Lemma

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is C^1. Let p_k be a descent direction at x_k and assume that f is bounded below along the ray $\{x_k + \alpha p_k \mid \alpha > 0\}$. Then there exists at least one interval of step lengths satisfying the Wolfe conditions (W_1)-(W_2).

Proof. Based on the mean value Theorem. See [NW], Lemma 3.1 (page 35).

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer (2006).

Backtracking Line Search

Algorithm

(Backtracking Line Search)

(1) Choose $\bar{\alpha} > 0$ and $\rho, c \in (0, 1)$

(2) $\alpha := \bar{\alpha}$

(3) while $\phi(\alpha) \geq \phi(0) + c\alpha\phi'(0)$

(4) $\quad \alpha := \rho\alpha$

(5) end

(6) $\alpha_k := \alpha$

Global convergence: Zoutendijk's Theorem

Theorem (**Zoutendijk**)

Consider any iteration of the form

$$x_0 \in \mathbb{R}^n, \quad x_{k+1} = x_k + \alpha_k p_k,$$

- p_k *is a descent direction.*
- α_k *satisfies the Wolfe conditions* (W_1)-(W_2).
- f *is bounded below in* $\mathbb{R}^n$.
- *the gradient* $g \equiv \nabla f$ *is* ***Lipschitz continuous*** *in an open set* $\mathcal{N}$ *containing the level set* $\mathcal{L} = \{x : f(x) \leq f(x_0)\}$, *where* x_0 *is the starting point.*

Then

$$\sum_{k \geq 0} \cos^2(\theta_k) \|g_k\|^2 < \infty \text{ (The series is convergent)}$$

with $\cos\theta_k = \frac{-p_k^T g_k}{\|g_k\| \|p_k\|}$.

Guus Zoutendijk

Guus Zoutendijk (1929–2005)

Guus Zoutendijk was a Dutch mathematician and politician. He worked as a professor of numerical mathematics at Leiden University. In mathematics he was mainly interested in optimization methods, specifically linear and nonlinear programming.

Comments

- The **Zoutendijk condition**, implies that

$$\sum_{k \geq 0} \cos^2(\theta_k) \|g_k\|^2 < \infty \Longrightarrow \cos^2(\theta_k) \|g_k\|^2 \to 0.$$

- If there is a $\delta > 0$ such that $\cos^2 \theta_k \geq \delta > 0$ for all k, then

$$\lim_{k \to \infty} \|g_k\| = 0.$$

- So we know that the gradient norms converge to zero, provided that the search directions are never too close to being perpendicular to the gradient.
- In particular, the **steepest descent method** is guaranteed to produce a sequence of gradients which converges to zero.

Proof.

- $(g_{k+1} - g_k)^T p_k \geq (c_2 - 1) g_k^T p_k$.
- $(g_{k+1} - g_k)^T p_k \leq \alpha_k L \|p_k\|^2$.
- $\alpha_k \geq \frac{c_2 - 1}{L} \frac{g_k^T p_k}{\|p_k\|^2}$.
- $f_{k+1} \leq f_k - c_1 \frac{1 - c_2}{L} \frac{(g_k^T p_k)^2}{\|p_k\|^2}$.
- $f_{k+1} \leq f_k - c \cos^2 \theta_k \|g_k\|^2$, with $c = c_1(1 - c_2)/L$.
- $f_{k+1} \leq f_0 - c \sum_{j=0}^{k} \cos^2 \theta_j \|g_j\|^2$.
- $\sum_{j=0}^{k} \cos^2 \theta_j \|g_j\|^2 < \infty$.

3. Gradient Descent Method with constant setp size

Gradient Descent Method with constant setp size

$\min_{x \in \mathbb{R}^n} f(x)$ $\min_{x \in H} f(x)$ H Hilbert Space

Gradient Descent (GD) Update with Constant Step Size:

$$x_{k+1} = x_k - s\nabla f(x_k), \quad \forall k \geq 0$$

Assumptions:

- **(A1)** f is L-smooth (i.e., continuously differentiable with Lipschitz gradient), where $L < \infty$.
- **(A2)** The optimization is unconstrained: $\min_{x \in \mathbb{R}^d} f(x)$.

Note: Convexity of f is not assumed unless explicitly stated.

Properties of L-Smooth Functions:

$$f(y) \leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2$$

$f(x) \leq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2$

$f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is L-smooth if f is of class $\mathcal{C}^1$ and its gradient ∇f is Lipschitz continuous with module $L > 0$ i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x, y\|$$

$$\forall x, y \in \mathbb{R}^n$$

Example: (i) $f(x) = \frac{1}{2} x^T Q x - b^T x$, $Q \in \mathcal{S}_n^{++}(\mathbb{R})$ is L-smooth. $\nabla f(x) = Qx - b$

$$\nabla^2 f(x) = Q$$

$$\|\nabla f(x) - \nabla f(y)\|_2 = \|Q(x-y)\|_2 \leq \boxed{\|Q\|_2} \|x-y\|_2$$

$$L = \|Q\|_2 = \sqrt{\rho(Q^T Q)} = \sqrt{\rho(Q^2)} = \rho(Q)$$

$$= \lambda_{max}(Q)$$

(ii) $f(x) = \frac{1}{2}\|Ax-b\|_2^2 = \frac{1}{2} x^T Q x - b^T x$

with $Q = A^T A$ $\quad b \leftarrow A^T b$

$A^T A \in \mathcal{S}_n^{+}(\mathbb{R}) \Rightarrow f$ is cvx

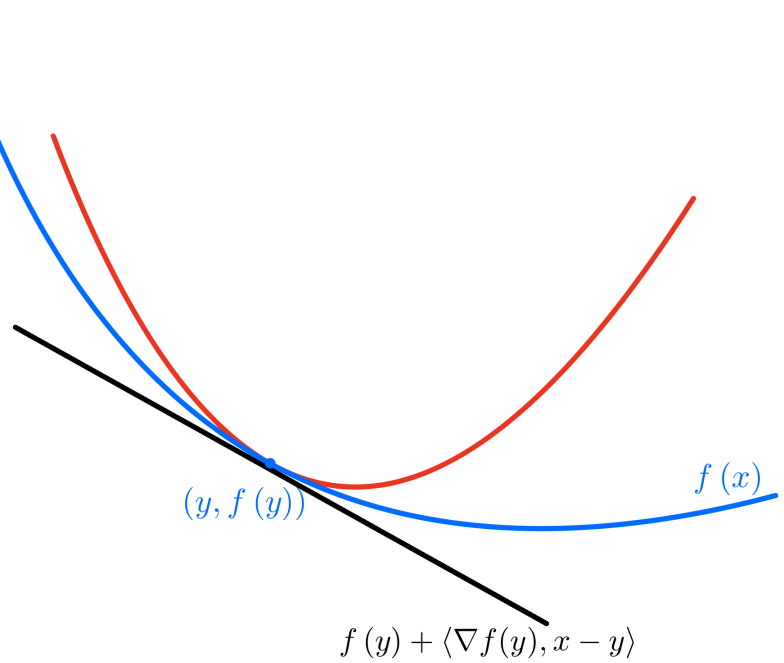
$$\boxed{L = \rho(A^T A)}$$

A function f is L-smooth if its gradient ∇f is Lipschitz continuous with modulus $L > 0$.

Lemma (Descent Lemma)

If f is L-smooth , then for any $x,\ y \in \mathcal{H}$,

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



Descent Lemma

Proof. From the L-smoothness assumption: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.
We apply the fundamental theorem of calculus:

$$f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt.$$

Using the L-smoothness bound and Cauchy-Schwarz ineq.

$$\langle \nabla f(y + t(x - y)) - \nabla f(y), x - y \rangle \leq Lt\|x - y\|^2.$$

Combining terms:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \int_0^1 Lt\|x - y\|^2 dt.$$

Evaluating the integral: $\int_0^1 Ltdt = \frac{L}{2}$, gives the final inequality:

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

Properties of L-Smooth Functions:

$$f(y) \leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2$$

Optimal Update Step: $g(y)$: $\nabla g(y) = \nabla f(x_k) + L(y - x_k) = 0 \Leftrightarrow y = x_k - \frac{1}{L}\nabla f(x_k)$

$$x_{k+1} = \arg\min_{y \in \mathbb{R}^d} \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2 \right\} = x_k - \frac{1}{L}\nabla f(x_k)$$

Function Descent: $f(x_{k+1}) \leq f(x_k) - \frac{1}{2}\|\nabla f(x_k)\|^2 + \frac{L}{2}\left\|\frac{1}{L}\nabla f(x_k)\right\|^2$

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

"$\frac{1}{2L}\|\nabla f(x_k)\|^2$" $\frac{1}{2L} - \frac{1}{L} = -\frac{1}{2L}$

Descent Lemma: If $x_{k+1} = x_k - s\nabla f(x_k)$ with $s \in \left(0, \frac{1}{L}\right]$, then

$$f(x_{k+1}) \leq f(x_k) - \frac{s}{2}\|\nabla f(x_k)\|^2$$

$0 < s < \frac{1}{L} \Leftrightarrow -\frac{1}{2L} \leq s/2$

$\Leftrightarrow -\frac{1}{L}\|\nabla f(x_k)\|^2 \leq -\frac{s}{2}\|\nabla f(x_k)\|^2$

The General Case: L-smooth, Non-convex f

Assumptions: f is L-smooth (potentially non-convex).
From Descent Lemma:

$$f(x_{k+1}) \leq f(x_k) - \frac{s}{2}\|\nabla f(x_k)\|^2 \leq \cdots \leq f(x_0) - \frac{s}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|^2.$$

$f(x_{k+1}) \leq f(x_k) - \frac{s}{2}\|\nabla f(x_k)\|^2$ $\quad \min_x f(x) = f(x^*) = f^*$

Rearranging terms:

$$\frac{s}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|^2 \leq f(x_0) - f(x_{k+1}) \leq f(x_0) - f^*.$$

Lower Bound:

$$\frac{s}{2}(k+1)\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \leq f(x_0) - f^*.$$

Conclusion:

$$\min_{0\leq i\leq k}\|\nabla f(x_i)\| \leq \sqrt{\frac{2(f(x_0)-f^*)}{s(k+1)}}.$$

$\leq \varepsilon$ $\quad \|\nabla f(x_{k+1})\| \leq \varepsilon$

For any $\epsilon > 0$, GDM finds an ϵ-near stationary point if

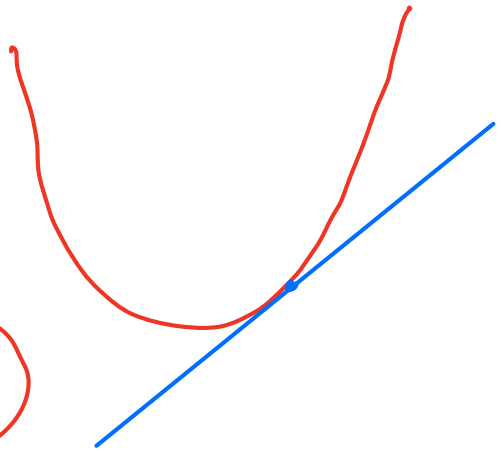
$$k+1 \geq \frac{2(f(x_0)-f^*)}{s\epsilon^2}.$$

The Convex Case: Bounding the Optimality Gap

$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ take $y = x^*$.

Convexity: Let $x^* \in \arg\min f(x)$. Convexity gives the lower bound:

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle.$$



Goal: Bound $f(x_{k+1}) - f(x^*)$: $x_{k+1} = x_k - s\nabla f(x_k)$

$$f(x^*) \geq f(x_k) + \frac{1}{s}\langle x_k - x_{k+1}, x^* - x_k \rangle.$$

Using algebraic manipulation and the descent lemma, we have:

$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k, v_{k+1} \rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2$

$$f(x^*) \geq f(x_{k+1}) + \frac{1}{2s}(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2).$$

Distance to Minimizer: Set $h(x) = \frac{1}{2}\|x - x^*\|^2$

$$\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \leq 2s(f(x^*) - f(x_{k+1})) \leq 0.$$

$0 \leq h(x_{k+1}) \leq h(x_k)$ $h(x_k)$ is non-decreasing.

This shows GDM always moves closer to the minimizers.

D.L. $\Rightarrow$ $f(x_k) \geq f(x_{k+1}) + \frac{1}{2s}\|x_k - x_{k+1}\|^2$

$f(x^*) \geq f(x_{k+1}) + \frac{1}{2s}\|x_k - x_{k+1}\|^2 + \frac{1}{s}\langle x_k - x_{k+1}, x^* - x_k\rangle$

$f(x^*) \geq f(x_{k+1}) + \frac{1}{s}\left[\frac{1}{2}\|x_k - x_{k+1}\|^2 + \langle x_{k+1} - x_k, x_k - x^*\rangle\right]$

We have,

$\frac{1}{2}\|x_{k+1} - x^*\|^2 - \frac{1}{2}\|x_k - x^*\|^2 = \frac{1}{2}\|x_{k+1} - x_k\|^2 + \langle x_{k+1} - x_k, x_k - x^*\rangle$

$v_k = x_k - x^* \qquad v_{k+1} = x_{k+1} - x^*, \quad v_{k+1} - v_k = x_{k+1} - x_k$

$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \frac{1}{2}\|v_{k+1} - v_k\|^2 + \langle v_{k+1} - v_k, v_k\rangle$

The Convex Case: Bounding the Optimality Gap

Bound on Optimality Gap:

$$f(x_{m+1}) - f(x^*) \leq \frac{1}{2s}(\|x_m - x^*\|^2 - \|x_{m+1} - x^*\|^2).$$

Summing over m:

$$\sum_{m=0}^{k}[f(x_{m+1}) - f(x^*)] \leq \frac{1}{2s}\|x_0 - x^*\|^2.$$

Convergence Rate: Since $f(x_1) \geq f(x_2) \geq \ldots$:

$$(k+1)[f(x_{k+1}) - f(x^*)] \leq \frac{1}{2s}\|x_0 - x^*\|^2,$$

$$f(x_{k+1}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2s(k+1)}. \Rightarrow f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2sk}$$

Iterations to Reach ϵ-Optimality:

$$k \geq \left\lceil \frac{\|x_0 - x^*\|^2}{2s\epsilon} \right\rceil.$$

$f(x_k) - f(x^*) \leq O(1/k)$

Definition of strongly convex function

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.

if f is L-smooth and μ-strongly cvx then

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

$f(x)=\frac{1}{2}x^T Q x - b^T x \qquad Q \in \mathcal{S}_n^{++}(\mathbb{R})$

f is μ-strongly convex with $\mu=\lambda_{min}(Q)$

$\langle Qx, x\rangle \geq \lambda_1 \|x\|^2$

We have:

$$0 \leq \boxed{\mu \leq L}$$

The best solut° for SMCO:

- ~~Monday 14/10 $8^H00 \rightarrow 11^H00$~~
- Tuesday 15/10 $14^H00 \rightarrow 17^H00$

Strongly Convex Case: Distance and Convergence

Assumption: f is μ-strongly convex. For all k:

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

Following similar steps:

$$f(x^*) \geq f(x_{k+1}) + \frac{1}{2s}\|x_{k+1} - x^*\|^2 - \left(\frac{1}{2s} - \frac{\mu}{2}\right)\|x_k - x^*\|^2.$$

Rearranging:

$$\|x_{k+1} - x^*\|^2 \leq (1 - \mu s)\|x_k - x^*\|^2.$$

When $s \leq \frac{1}{L}$, we have $\mu s \in (0, 1]$, so:

$$\|x_{k+1} - x^*\|^2 \leq (1 - \mu s)^{k+1}\|x_0 - x^*\|^2.$$

Convergence: To achieve $\|x_{k+1} - x^*\| \leq \epsilon$, we need:

$$k = O\left(\frac{1}{\mu s}\log\left(\frac{\|x_0 - x^*\|}{\epsilon}\right)\right).$$

Exercise Solution

Exercise: Prove:

$$f(x_{k+1}) - f(x^*) \leq (1 - \mu s)^{k+1} (f(x_0) - f(x^*)) .$$

What about $\|\nabla f(x_{k+1})\|$?

Solution: From the Descent Lemma:

$$f(x_{k+1}) \leq f(x_k) - \frac{s}{2}\|\nabla f(x_k)\|^2.$$

Using strong convexity: $f(x_k) - f(x^*) \geq \frac{\mu}{2}\|x_k - x^*\|^2$. $\leftarrow \nabla f(x^*) = 0$
Also, $\|\nabla f(x_k)\| \geq \mu\|x_k - x^*\|$, so: $\|x_k - x^*\| \leq \frac{1}{\mu}\|\nabla f(x_k)\|$.
Substitute back into the strong convexity inequality:

$$f(x_k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{1}{\mu}\|\nabla f(x_k)\|\right)^2 = \frac{1}{2\mu}\|\nabla f(x_k)\|^2.$$

Rewriting: $\|\nabla f(x_k)\|^2 \leq 2\mu(f(x_k) - f(x^*))$.
Substitute into the Descent Lemma: $f(x_{k+1}) \leq f(x_k) - s\mu(f(x_k) - f(x^*))$.
Simplify: $f(x_{k+1}) - f(x^*) \leq (1 - \mu s)(f(x_k) - f(x^*))$.
By induction: $f(x_{k+1}) - f(x^*) \leq (1 - \mu s)^{k+1}(f(x_0) - f(x^*))$.

f μ-strong cvx $\Leftrightarrow$ ∇f is μ strong monotone

$\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle \geq \mu \|x_k - x^*\|^2$

$\mu \|x_k - x^*\|^2 \leq \langle \nabla f(x_k), x_k - x^* \rangle \leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\|$

$$\|x_k - x^*\| \leq \frac{1}{\mu} \|\nabla f(x_k)\|$$

$f(x_k) - f(x^*) \leq \frac{\mu}{2} \|x_k - x^*\|^2 \leq \frac{1}{2\mu} \|\nabla f(x_k)\|^2$

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu (f(x) - f^*)$$

Known as Polyak-Łojasiewicz Inequality (PL)

---

f μ-strong cvx $\Leftrightarrow$ $\langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle \geq \mu \|x - x^*\|^2$

$\|\nabla f(x)\| \cdot \|x - x^*\| \geq \langle \nabla f(x), x - x^* \rangle \geq \mu \|x - x^*\|^2$

$$\|\nabla f(x)\| \geq \mu \|x - x^*\|$$

f μ-strong cvx $\Leftrightarrow$ $f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle + \frac{\mu}{2} \|x - x^*\|^2$ ($\nabla f(x^*) = 0$)

$\Leftrightarrow$ $f(x) - f^* \geq \frac{\mu}{2} \|x - x^*\|^2$

Proof of PL inequality: f μ-st. cvx $\Rightarrow \frac{1}{2} \|\nabla f(x)\|^2 \geq \mu (f(x) - f^*)$

f st$_g$ cvx $\Leftrightarrow$ $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$ $\forall y, \forall x$

$\Rightarrow \min_y f(y) \geq \min_y \underbrace{\left(f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \right)}_{g(y)}$

f^*

$\nabla g(y) = \nabla f(x) + \mu (y - x) = 0 \Leftrightarrow y - x = -\frac{1}{\mu} \nabla f(x)$

$y^* = x - \frac{1}{\mu} \nabla f(x)$

$g(y^*) = f(x) + \langle \nabla f(x), -\frac{1}{\mu} \nabla f(x) \rangle + \frac{\mu}{2} \cdot \frac{1}{\mu^2} \|\nabla f(x)\|^2$

$f^* \geq f(x) - \frac{1}{\mu} \|\nabla f(x)\|^2 + \frac{1}{2\mu} \|\nabla f(x)\|^2 = f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2$

Hence, $\frac{1}{2\mu}\|\nabla f(x)\|^2 \geq f(x) - f^* \quad \forall x$

$$\boxed{\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left(f(x) - f^*\right)}$$

PL inequality

Proof of : $f(x_{k+1}) - f^* \leq (1-\mu s)^{k+1}\left(f(x_0) - f^*\right)$

PL : $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2$

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

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

$f(x_{k+1}) \leq f(x_k) - s\|\nabla f(x_k)\|^2 + \frac{s^2 L}{2}\|\nabla f(x_k)\|^2$

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

$f(x_{k+1}) \leq f(x_k) + \underbrace{\frac{s(sL-2)}{2}}_{-s\left(2 - \frac{sL}{2}\right)}\|\nabla f(x_k)\|^2 \leftarrow$ choose $s = \frac{1}{L}$ $\quad -\frac{2s + s^2 L}{2} = \frac{s(sL-2)}{2}$

$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \leq f(x_k) - \boxed{\frac{\mu}{L}}\left(f(x_k) - f^*\right)$

$f(x_{k+1}) - f^* \leq \underbrace{f(x_k) - f^* - \frac{\mu}{L}\left(f(x_k) - f^*\right)}_{(1-s\mu)\left(f(x_k) - f^*\right)}$

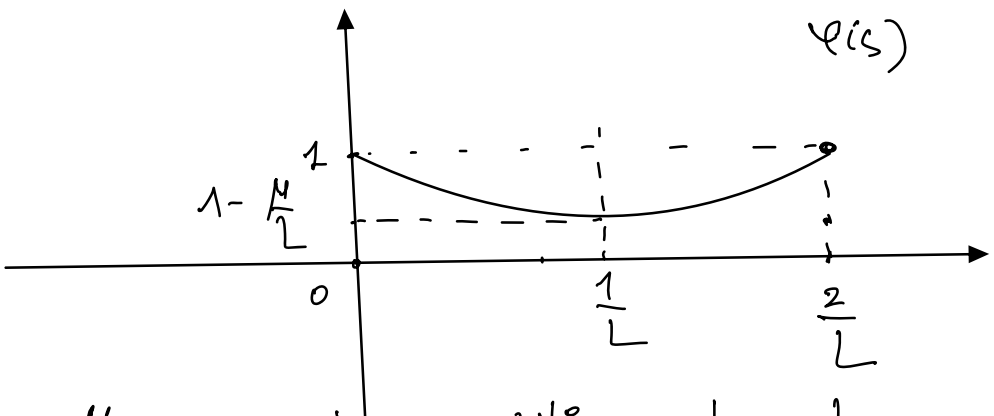
$$\boxed{f(x_{k+1}) - f^* \leq (1-\mu s)^{k+1}\left(f(x_0) - f^*\right)}$$

If we keep s : $f(x_{k+1}) - f^* \leq \underbrace{\left(1 - \mu s(2 - sL)\right)}_{\varphi(s)}\left(f(x_k) - f^*\right)$

we set $\min_{s \in [0, \frac{2}{L}]} \varphi(s)$

$\varphi(s) = 1 - \left(2\mu s - \mu s^2 L\right) = \mu L s^2 - 2\mu s + 1$

$\varphi'(s) = 2\mu L s - 2\mu = 0 \iff \boxed{s = \frac{1}{L}}$ $\quad \varphi\left(\frac{1}{L}\right) = \mu L \cdot \frac{1}{L^2} - \frac{2\mu}{2} + 1 = 1 - \frac{\mu}{2}$



$\frac{\mu}{L} = \chi$ = the condit° number !

Gradient Norm Bound:
Using L-smoothness:

$$\|\nabla f(x_{k+1})\| \leq L\|x_{k+1} - x^*\|.$$

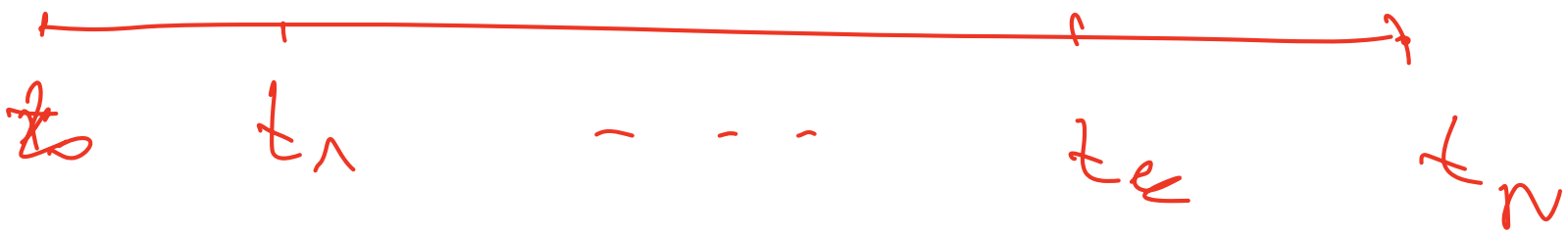
From the distance contraction:

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

Therefore:

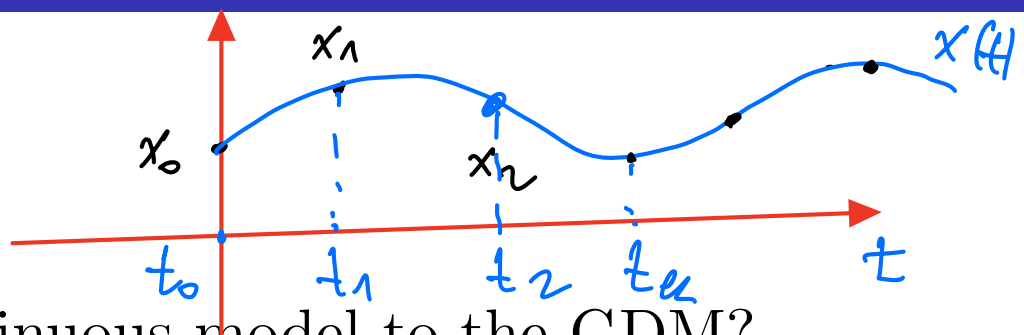
$$\|\nabla f(x_{k+1})\| \leq L(1 - \mu s)^{\frac{k+1}{2}} \|x_0 - x^*\|.$$

4. Gradient Descent Method: The continuous cas (The gradient flow)



Gradient flow as a continuous model

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



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

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

For $t = k\alpha$, we have $t + \alpha = k\alpha + \alpha = (k+1)\alpha$

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

Hence,

$$\frac{1}{\alpha}\Big[X(t+\alpha) - X(t)\Big] = -\nabla f(X(t)).$$

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

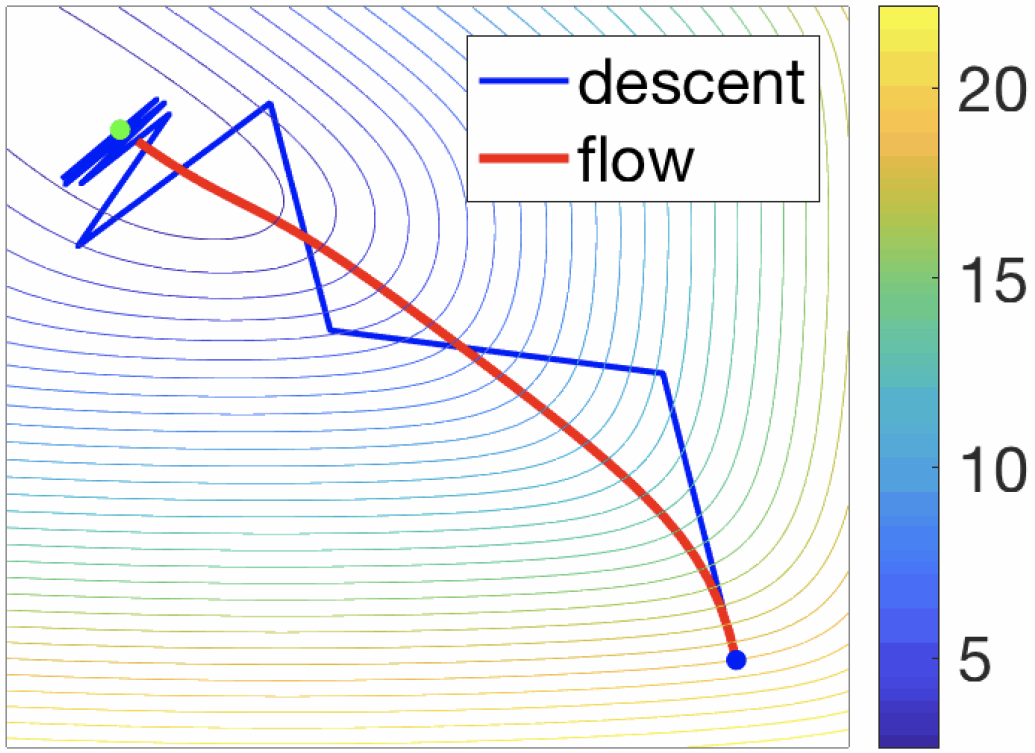
$x(t) \xrightarrow[t \to \infty]{} x_\infty$

$\nabla f(x_\infty) = 0$

Gradient Flow

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

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



Introduction to Euler Methods

- Euler Methods are fundamental numerical techniques for solving initial value problems (IVPs) of Ordinary Differential Equations (ODEs).
- Applicable to first-order ODEs of the form:

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0$$

$t \in [0, +\infty[$
$t \in [0, T]$

- Serve as the foundation for more advanced numerical methods.
- Two primary variants:
  1. **Explicit Euler Method**
  2. **Implicit Euler Method**

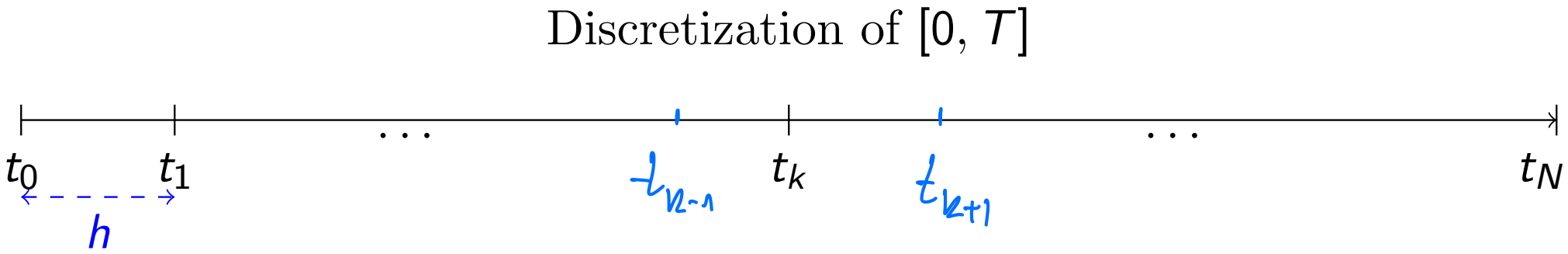


Figure: Discretization of the interval $[0, T]$ using the sequence t_k, $k \in \mathbb{N}$.

Explicit Euler Method

- Definition:

$$x_{k+1} = x_k + h \cdot f(t_k, x_k)$$

- Procedure:
  1. Start with initial condition x_0 at t_0.
  2. Compute subsequent values using the slope at the current point.
- Advantages:
  - Simple to implement.
  - Computationally inexpensive per step.
- Disadvantages:
  - Stability issues, especially for stiff equations.
  - Lower accuracy (first-order).

Implicit Euler Method

- Definition:

$$x_{k+1} = x_k + h \cdot f(t_{k+1}, x_{k+1})$$

- Procedure:
  1. Start with initial condition x_0 at t_0.
  2. Solve the nonlinear equation for x_{k+1} at each step:

$$g(X) = X - x_k - h \cdot f(t_{k+1}, X) = 0.$$

- Advantages:
  - Unconditionally stable; suitable for stiff ODEs.
  - Better handling of certain types of problems.
- Disadvantages:
  - More computationally intensive due to the need to solve equations at each step.
  - Implementation complexity.

Explicit Euler Method

$x(t) = x_0 e^{-at}$ $\lim_{t \to +\infty} x(t) = 0$

ODE:

$$x'(t) = -ax(t), \quad x(0) = x_0, \quad a > 0$$

Update formula:

$$x_{k+1} = x_k + h(-ax_k) = x_k(1 - ah)$$

Expression in terms of x_0:

$$x_k = x_0(1 - ah)^k$$

We should have $\lim_{k \to +\infty} x_k = 0$

$|1 - ah| < 1$

Condition for stability:

$$0 < h \leq \frac{2}{a}$$

$-1 < 1 - ah < 1$

$-2 < -ah < 0$

$0 < ah < 2$

Growth factor:

$$(1 - ah)^k$$

Implicit Euler Method

ODE:

$$x'(t) = -ax(t), \quad x(0) = x_0, \quad a > 0$$

Update formula:

$$x_{k+1} = x_k + h(-ax_{k+1}) \implies x_{k+1} = \frac{x_k}{1 + ah}$$

Expression in terms of x_0:

$$x_k = \frac{x_0}{(1 + ah)^k} = x_0 \left[(1+ah)^{-1}\right]^k$$

Condition for stability: Unconditionally stable

Growth factor:

$$(1 + ah)^{-k}$$

h small

$(1+ah)^{-1} \approx 1 - ah$

$\frac{1}{1+x} \approx 1 - x + o(x)$

Stability and Convergence

- Stability:
  - **Explicit Euler**:
    - Conditionally stable.
    - Stability region limited; restrictive step size for stiff problems.
  - **Implicit Euler**:
    - Unconditionally stable.
    - Suitable for larger step sizes in stiff equations.
- Convergence:
  - Both methods are first-order accurate:

$$\text{Local truncation error} = \mathcal{O}(h^2)$$

$$\text{Global error} = \mathcal{O}(h)$$

- Trade-offs:
  - Explicit: Simpler but requires smaller step sizes for stability.
  - Implicit: More complex but allows for larger step sizes.

Comparison and Applications

Feature	**Explicit Euler**	**Implicit Euler**
Stability	Conditionally stable	Unconditionally stable
Complexity	Simple implementation	Requires solving equations
Computational Cost	Lower per step	Higher per step
Suitability	Non-stiff ODEs	Stiff ODEs
Accuracy	First-order	First-order

Applications:

- **Explicit Euler**:
  - Smooth functions.
- **Implicit Euler**:
  - Nonsmooth functions.

Convergence result of the Gradient Flow

$\forall x_0 \in H$, $\exists!\, x(\cdot; x_0)$ trajectory Solution of (GF) $t \in [0, +\infty[$

Say few words about Existence and uniqueness

Cauchy-Lipschitz thm for ODE's.

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

Theorem

Let $f : \mathcal{H} \to \mathbb{R}$ be convex, continuously differentiable and bounded from below.

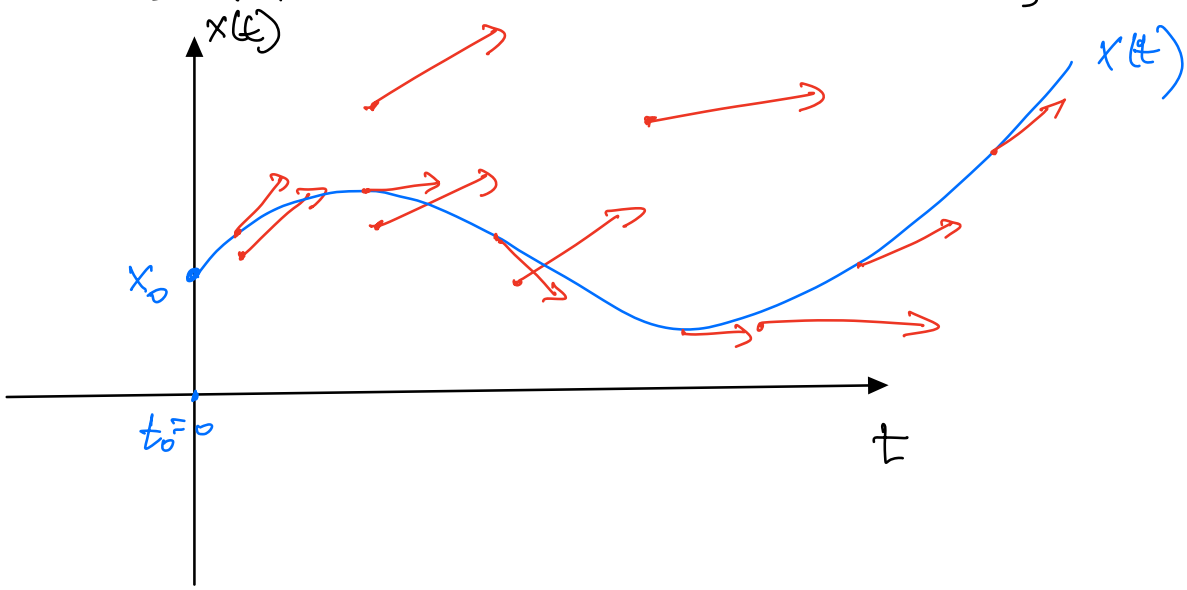
Assume that $S = \operatorname{argmin}(f) \neq \emptyset$. $f^* = \inf_{\mathcal{H}} f$. Then

(i) $f(x(t)) - f^* \leq \frac{d(x_0, S)^2}{2t}$, $t > 0$. $f(x(t)) - f^* \leq O(1/t)$ $t \to \infty$

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

Geometrical interpretation:

$f : \mathbb{R}^2 \to \mathbb{R}$ C^1 scalar $\qquad \nabla f : \mathbb{R}^2 \to \mathbb{R}^2$ vector field



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

$t \to \infty$

$$f(x(t)) - f^* \leq O\left(\frac{1}{t}\right)$$

Convergence result of the Gradient Flow

Proof. (i) Consider the Lyapounov function $t \mapsto E(t) = t(f(x(t)) - f^*) + \frac{1}{2}\|x(t) - x^*\|^2$, with $x^* \in S$. We have

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

Hence,

$$t\big(f(x(t)) - f^*\big) \leq E(t) \leq E(0) = \frac{1}{2}\|x_0 - x^*\|^2.$$

(ii) We have

$$\frac{1}{2}\frac{d}{dt}\|x(t) - x^*\|^2 = \langle \nabla f(x(t), x^* - x(t))\rangle \leq f(x^*) - f(x(t)) \leq 0.$$

Hence, $\lim_{t\to+\infty} \|x(t) - x^*\|$ exists for every $x^* \in S$.
Since every convex function is weakly lsc, we deduce that if $x(t)$ converges weakly to some x_∞, then

$$\liminf_{t\to+\infty} f(x(t)) \geq f(x_\infty).$$

Therefore, $f^* = f(x_\infty)$, which means that $x_\infty \in S$. Opial's Lemma allows to conclude.

Opial's Lemma. Let $S \subset \mathcal{H}$ be a nonempty set and $x : [0, +\infty[\to \mathcal{H}$ such that:

(i) for every $x^* \in S$, $\lim_{t\to+\infty} \|x(t) - x^*\|$ exists.

(ii) every weak sequential limit point of $x(t)$, as $t \to +\infty$, belongs to S.

Then $x(t) \rightharpoonup x_\infty \in S$ as $t \to +\infty$.

Aleksandr Mikhailovich Lyapunov

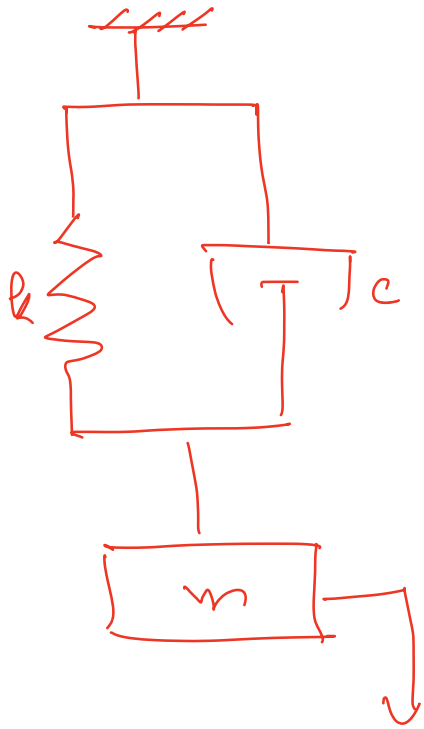
Aleksandr Lyapunov (1857-1918): a Russian mathematician best known for his development of the stability theory of dynamical systems.

$$\dot{x}(t) = \Phi(x(t)), \quad x(0) = x_0, \quad t \geq 0.$$

Lyapunov's direct method

If V is locally positive definite and its orbital derivative $\dot{V}(x) \leq 0$ is locally in x and for all t, then the origin of the system is locally stable (in the sense of Lyapunov).

- $\dot{V}(x) = \langle \nabla V(x(t)), \dot{x}(t) \rangle = \langle \nabla V(x(t)), \Phi(x(t)) \rangle$.



$$m\ddot{x}(t) = \sum F = -kx(t) - c\dot{x}(t)$$

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = 0$$

$c > 0$: coeff. of viscosity $\boxed{m=1}$

$c\dot{x}$ = viscous fricto

$$x(t_0) = x_0, \quad \dot{x}(t_0) = x_1$$

$$E(t) = \underbrace{\frac{1}{2}\dot{x}(t)^2}_{\text{kinetic energy}} + \underbrace{\frac{k}{2}x(t)^2}_{\text{potential energy}} \geq 0$$

$$E'(t) = \dot{x}(t)\ddot{x}(t) + kx(t)\dot{x}(t)$$

$$= \dot{x}(t)\underbrace{\left[\ddot{x}(t) + kx(t)\right]}_{-c\dot{x}(t)}$$

$$\boxed{E'(t) = -c\dot{x}(t)^2 \leq 0}$$

Convergence result of the Gradient Flow: the strongly convex case

if f is very smooth : $\nabla^2 f(x) \geq \mu I_n$

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

$f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \frac{\mu}{2}\|x-x_0\|^2 \leq f(x)$ $\forall x, \forall x_0$

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

- 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 $t \to +\infty$.

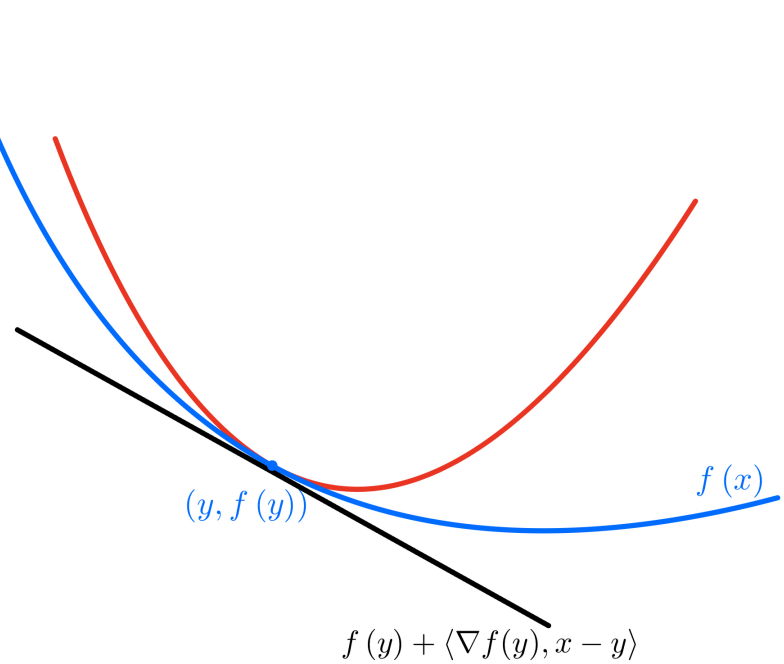
Convergence result of the Gradient Flow

A function f is L-smooth if its gradient ∇f is Lipschitz continuous with modulus $L > 0$.

Lemma (Descent Lemma)

If f is L-smooth , then for any $x,\ y \in \mathcal{H}$,

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



Using this Lemma, we get

$$0 \leq f(x(t)) - f^* \leq \underbrace{\langle \nabla f(x^*), x(t) - x^* \rangle}_{=0} + \frac{L}{2}\|x(t) - x^*\|^2.$$

$$\leq \frac{1}{2} L e^{-\mu t} \|x_0 - x^*\|^2.$$

$\lim_{t \to +\infty} f(x(t)) = f^*$

Convergence result of the Gradient Descent Method

Gradient flow : flot de gradient : $\begin{cases} \dot{x}(t) = -\nabla f(x(t)) \\ x(t_0) = x_0 \end{cases}$.

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 $L > 0$, $0 < sL < 2. \iff 0 < s < \frac{2}{L}$

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 $k \to +\infty$.

Homework : (i) Prove the gradient descent Lemma.
(ii) Prove $f(x_k) - f^* \leq \dfrac{L\operatorname{dist}(x_0, S)^2}{2k}$ as $k \to +\infty$
(iii) $x_k \rightharpoonup x_\infty \in S$ as $k \to +\infty$ (take $H = \mathbb{R}^n$).

Proximal Point Algorithm

Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ convex lsc and proper, i.e. $f \in \Gamma_0(\mathcal{H})$.
Assume that $S = \operatorname{argmin}(f) \neq \emptyset$.
Consider the following nonsmooth gradient flow

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

where the subdifferential ∂f is defined by

$$p \in \partial f(x) \iff \langle p, y - x \rangle \leq f(y) - f(x), \ \forall y \in \mathcal{H}.$$

Bruck (70-80)

- $x(t) \rightharpoonup x_\infty \in S$ weakly as $t \to +\infty$, i.e. $0 \in \partial f(x_\infty)$.
  Extension to a maximally monotone operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$:

$$\dot{x}(t) \in -A(x(t)), \ t \geq 0.$$

Implicit discretization

Using an implicit Euler discretization, we have

$$\frac{1}{s}\Big(x_{k+1} - x_k\Big) \in -\partial f(x_{k+1}) \iff x_{k+1} = (\mathrm{I} + s\partial f)^{-1}(x_k).$$

Jean-Jacques Moreau

The proximal operator is defined by

$$\mathrm{prox}_{sf}(x) = \mathrm{argmin}_{y\in\mathcal{H}}\left\{f(y) + \frac{1}{2s}\|y - x\|^2\right\} = (\mathrm{I} + s\partial f)^{-1}(x).$$

For every $f \in \Gamma_0(\mathcal{H})$, the operator $\mathrm{prox}_f : \mathcal{H} \to \mathcal{H}$ is single-valued and nonexpansive.
For a closed convex subset $C \subset \mathcal{H}$, we have $\mathrm{prox}_{I_C} = \mathrm{Proj}_C$, with I_C the indicator function of C.

The proximal point algorithm

$\min_{x \in \mathbb{R}^n} f(x) =$ $\quad 0 \in \partial f(x) \Leftrightarrow x \in x + s\partial f(x) \quad \forall s > 0$

$\Leftrightarrow x \in (Id + s\partial f)(x)$

$\Leftrightarrow x = \text{prox}_{sf}(x)$

$$(\text{PPA}) \begin{cases} x_0 \in \mathcal{H} \\ x_{k+1} = \text{prox}_{sf}(x_k) \end{cases}$$

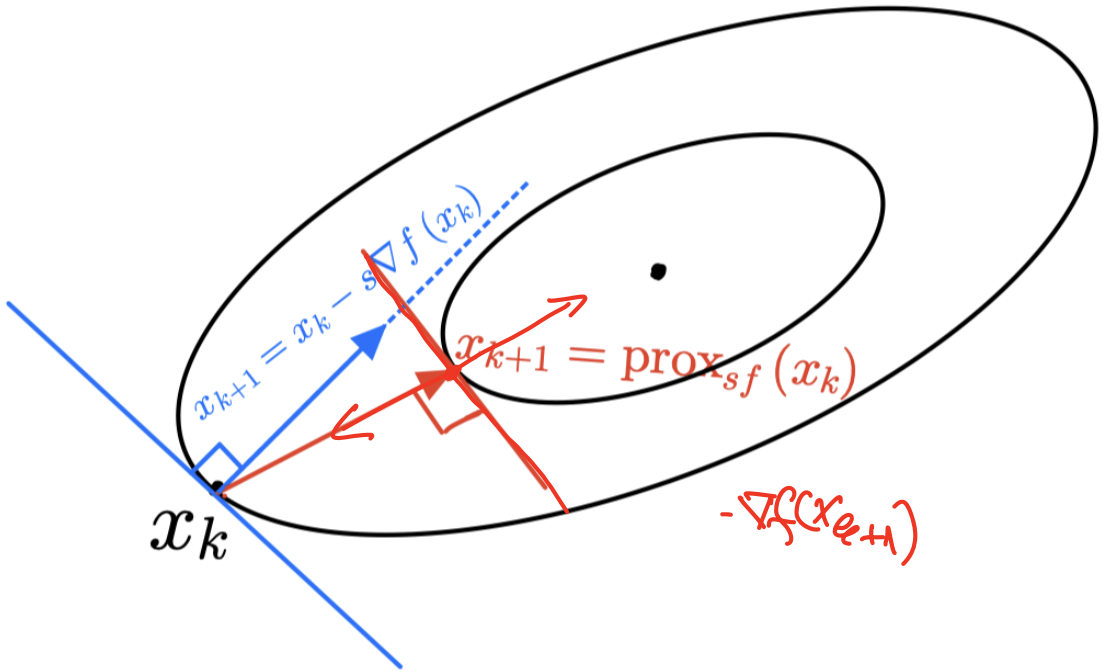
B. Martinet, R.T. Rockafellar (1970) and O. Güler (1991).

- $x_k \rightharpoonup x_\infty \in S$ weakly as $k \to +\infty$.
- $f(x_k) - \min(f) = O(\frac{1}{k})$.

PPA versus GDM

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

$x_{k+1} = (I + s\nabla f)^{-1}(x_k) \Leftrightarrow x_{k+1} + s\nabla f(x_{k+1}) = x_k \Leftrightarrow x_{k+1} = x_k - s\nabla f(x_{k+1})$



Gradient flow: the nonconvex case

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

- $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ convex: Bruck (1970-80), maximally monotone operators, semi-groups of contractions:

  $x(t) \rightharpoonup x_\infty \in \operatorname{argmin} f \neq \emptyset$ weakly as $t \to +\infty$.

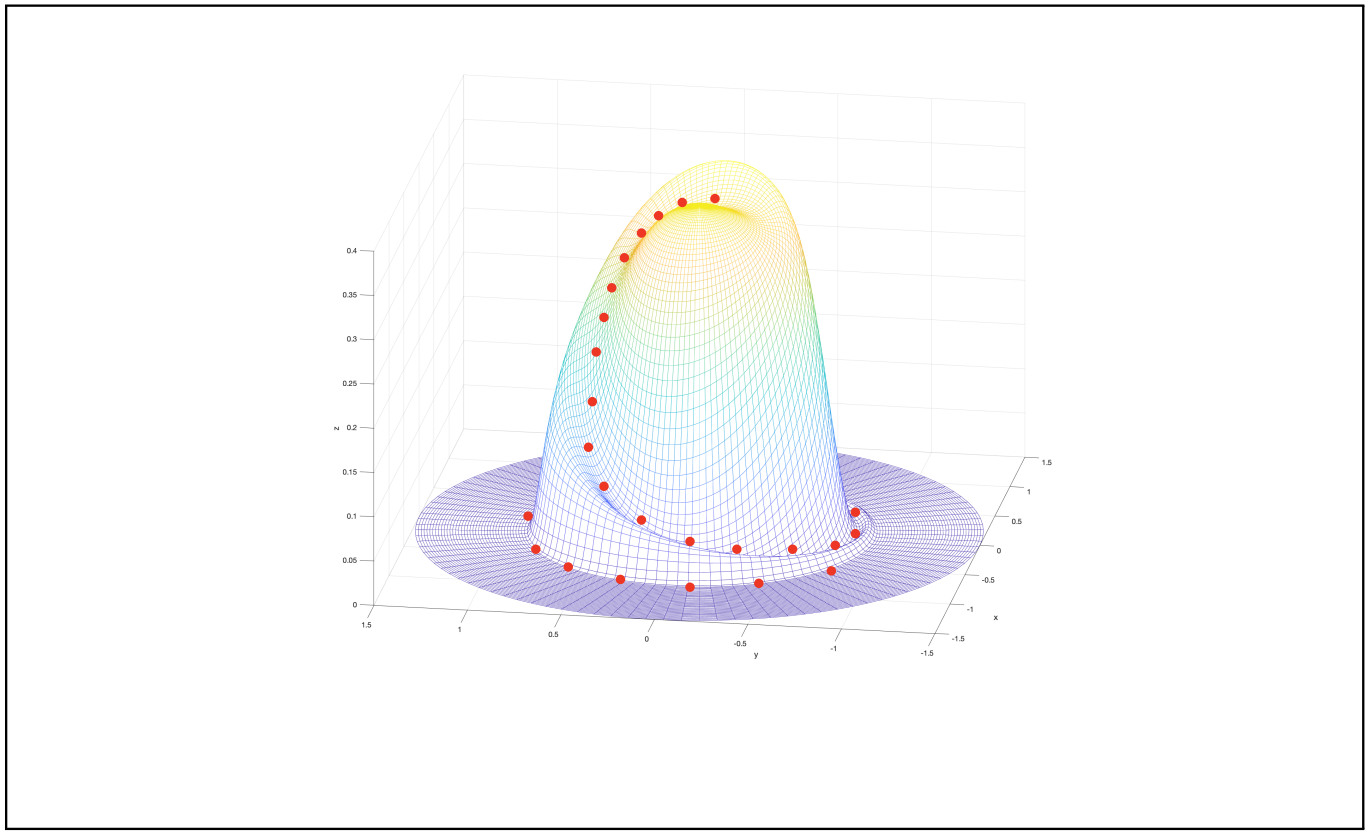
  $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}(\frac{1}{t})$ as $t \to +\infty$.

- $f : \mathbb{R}^N \to \mathbb{R}$ 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.

J. Palis and W. De Melo counterexample (1982): mexican hat



$$f(r,\theta) = e^{-\frac{1}{1-r^2}}\left[1 - \frac{4r^4}{4r^4 + (1-r^2)^4}\sin\left(\theta - \frac{1}{1-r^2}\right)\right]$$

Łojasiewicz inequality: the nonconvex case

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

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 (GF) has a finite length and hence converges to a critical point of f, as $t \to +\infty$.

3. 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

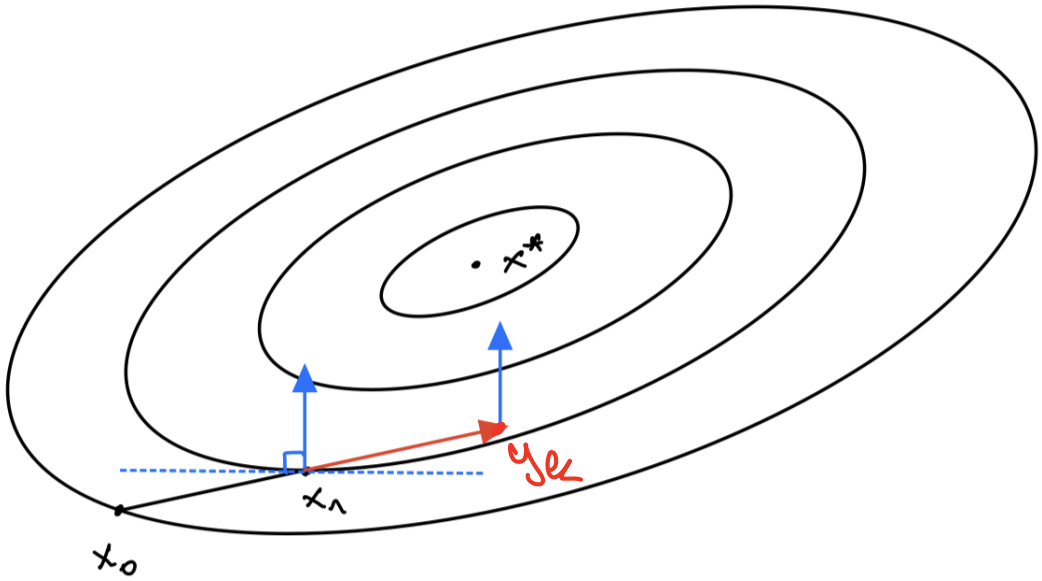
$$(\text{PM}) \begin{cases} x_0,\ x_1 \in \mathcal{H} \\ x_{k+1} = \underbrace{x_k - \alpha \nabla f(x_k)}_{GDM} + \underbrace{\beta(x_k - x_{k-1})}_{momentum}, \end{cases}$$

where $\alpha > 0$ 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

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



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?

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

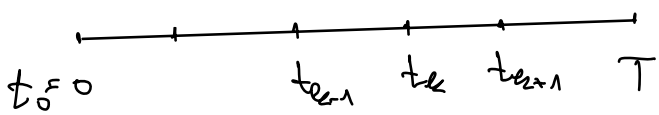
Set $h = \sqrt{\alpha}$ and $\beta = 1 - \gamma h$ with $\gamma > 0$.
We have

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

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

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

$$X(0) = x_0,\ \dot{X}(0) = x_1.$$



$t_{k+1} - t_k = h > 0$

$$x(t+h) = x(t) + h\dot{x}(t) + \frac{h^2}{2}\ddot{x}(t) + h^2\varepsilon(h)$$

$$x(t-h) = x(t) - h\dot{x}(t) + \frac{h^2}{2}\ddot{x}(t) + h^2\varepsilon(-h)$$

$o(h^2)$

$$x(t+h) + x(t-h) = 2x(t) + h^2\ddot{x}(t) + h^2[\varepsilon(h) + \varepsilon(-h)]$$

$\lim_{h\to 0} \varepsilon(h) = 0$

$$\ddot{x}(t) \simeq \frac{x(t+h) - 2x(t) + x(t-h)}{h^2}$$ h small

$$\dot{x}(t) \simeq \frac{x(t+h) - x(t)}{h}$$

$t \leftarrow t_k$ $\quad x(t_k) = x_k$

$x(t_k + h) = x(t_{k+1}) = x_{k+1}$

$x(t_k - h) = x_{k-1}$

$$\ddot{x}(t_k) \simeq \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2}$$ discrete acceleration

$$\dot{x}(t_k) \simeq \frac{x_{k+1} - x_k}{h}$$ discrete velocity

$x(t) \leftarrow$ posit°

$\dot{x}(t) \leftarrow$ velocity

$\ddot{x}(t) \leftarrow$ acceleration

Centred finite difference

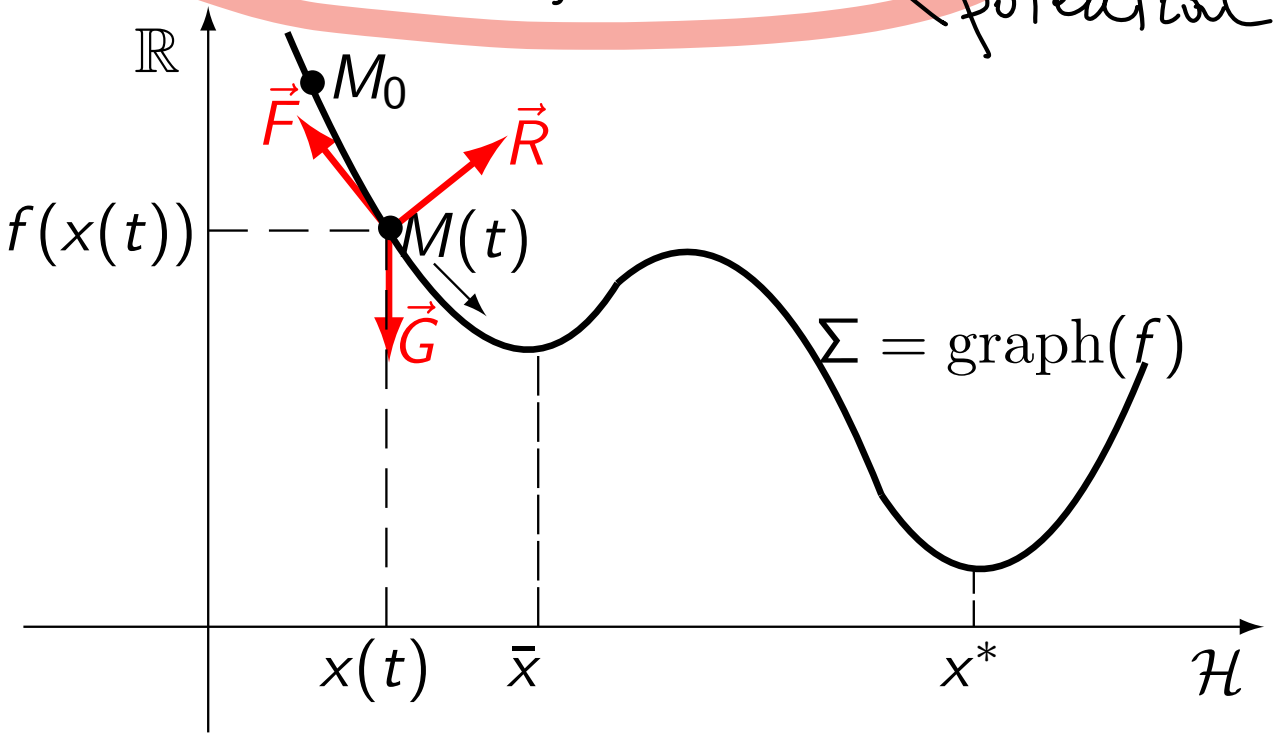
The heavy ball with friction method

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

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

viscous frict

potential

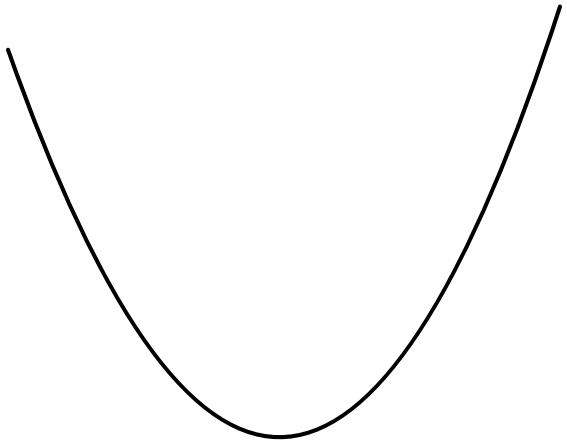


$\vec{E} = -\overrightarrow{\text{grad}}\, V$

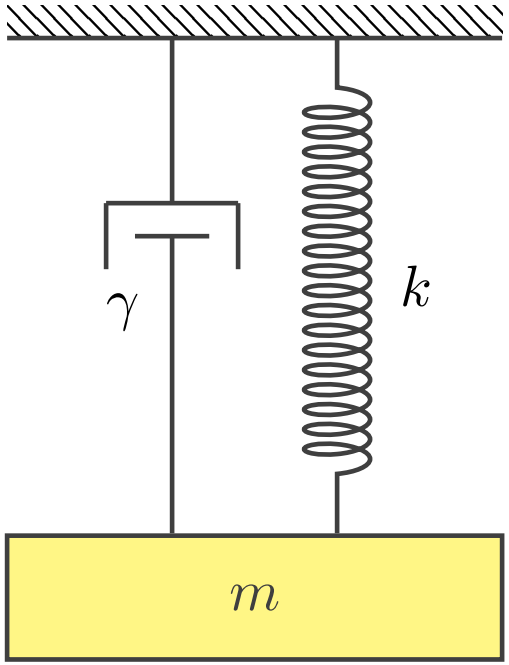
Mechanical interpretation $\vec{G}$ =gravity, $\vec{F}$= friction, $\vec{R}$= reaction

Figure taken from Attouch-Goudou-Redont (2000).

Mass-Spring-Damper



$\frac{1}{2}x^2$



$f(x) = \frac{1}{2}x^2$ is μ-strongly cvx with $0 < \mu \leq 1$

$$m\ddot{x}(t) + \gamma\dot{x}(t) + kx(t) = 0.$$

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

(HBF) in the μ-strongly convex case

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

- $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(e^{-\sqrt{\mu}t}\right) \quad \text{as } t \to +\infty.$
- 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 $C = f(x_0) - \inf_{\mathcal{H}} f + \mu \text{dist}(x_0, S)^2 + \|x_1\|^2.$

To prove this Theorem we the following Lyapunov function

$$\mathcal{E}(t) := f(x(t)) - \min_{\mathcal{H}} f + \frac{1}{2}\|\sqrt{\mu}(x(t) - x^*) + \dot{x}(t)\|^2,$$

where x^* is the unique minimizer of f.

(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* $t \to +\infty$.

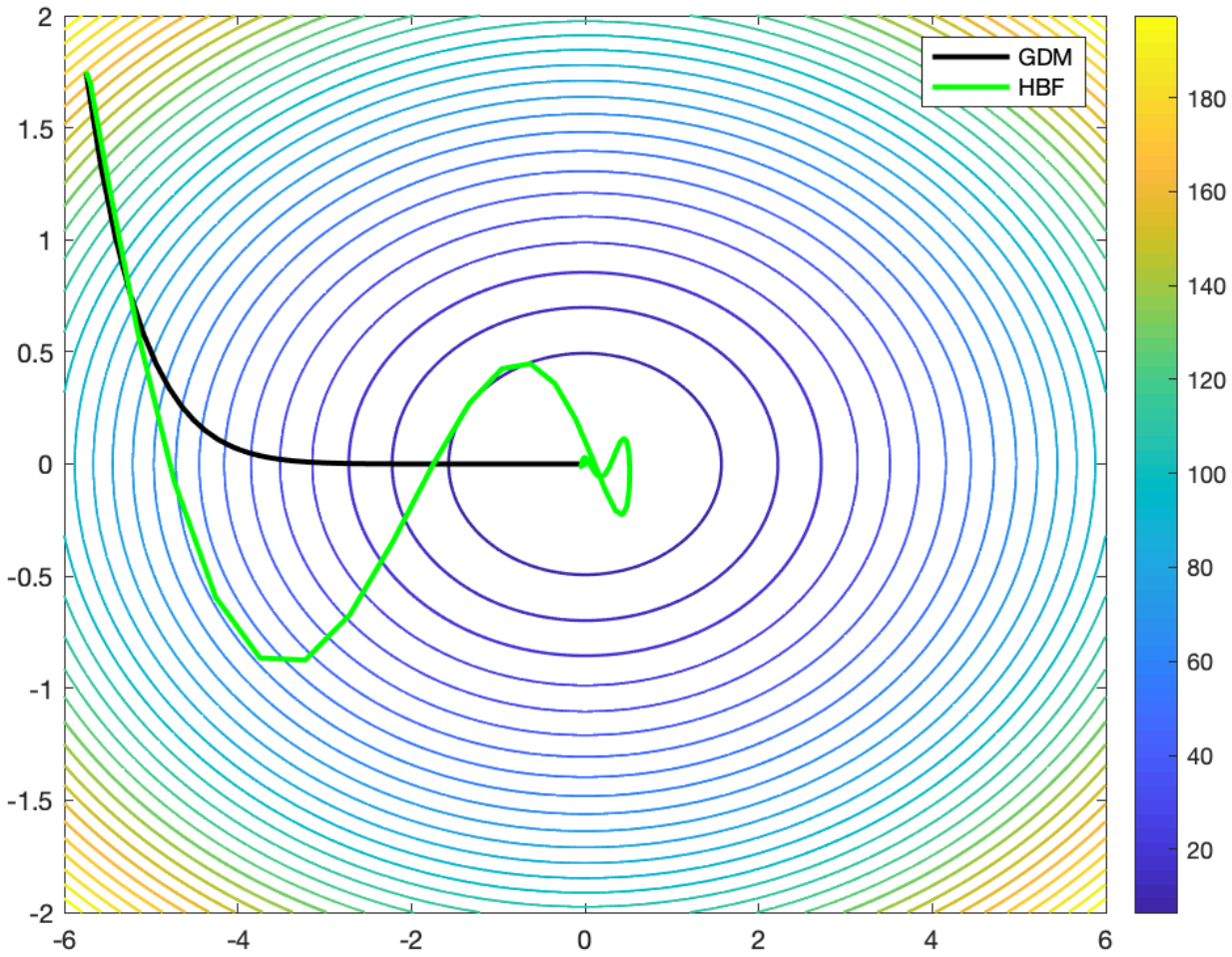
- $E(t) = \frac{1}{2}\|\dot{x}(t)\|^2 + f(x(t))$ the Lyapounov energy function.
  $E'(t) = -\gamma\|\dot{x}(t)\|^2 \leq 0$ (dissipative system).
- $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to +\infty$.

The Heavy Ball with friction

- B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*. Computational mathematics and mathematical physics (1964).
- H. Attouch, X. Goudou, P. Redont. The heavy ball with friction method. I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Commun. Contemp. Math. 2 (2000).
- F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM J. Control Optim. (2000).
- A. Haraux, M. Jendoubi. Decay estimates to equilibrium for some evolution equations with an analytic nonlinearity. Asymptot. Anal. 26 (2001).

HBF versus GDM

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



The Heavy Ball with friction: optimal parameters

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

Homework. Let's assume that f is a quadratic function:

$$f(x) = \frac{1}{2} x^\top Q x - b^\top x + c,$$

where Q is a symmetric positive definite matrix with eigenvalues in the range $[\mu, L]$. Show that the optimal parameters α and β are given by:

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

The Heavy Ball with friction: optimal parameters

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

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

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

- The Heavy ball with friction (HBF) is optimal for $\mathcal{C}^2$, μ-strongly convex and L-smooth functions.
- The analysis relies on the knowledge of both parameters L and μ.
- For smooth convex functions, the way how to tune the parameters α and β is not clear.
- For general convex functions, the asymptotic convergence rate of (HBF) is $\mathcal{O}(1/t)$ (in the worst case). This is however not better than the steepest descent method.

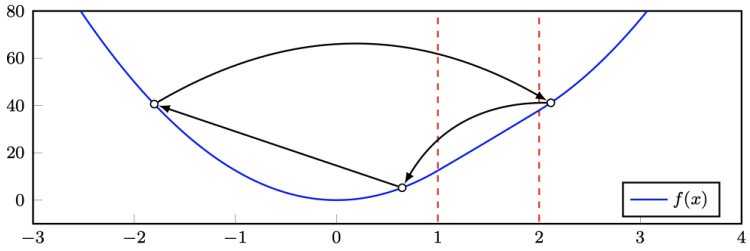
The Heavy Ball with friction: some drawbacks

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

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

Homework: Code this example.

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).

The Heavy Ball with friction

- B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*. Computational mathematics and mathematical physics (1964).
- H. Attouch, X. Goudou, P. Redont. The heavy ball with friction method. I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Commun. Contemp. Math. 2 (2000).
- F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM J. Control Optim. (2000).
- A. Haraux, M. Jendoubi. Decay estimates to equilibrium for some evolution equations with an analytic nonlinearity. Asymptot. Anal. 26 (2001).

Nesterov's Accelerated Gradient (1983)

Yurii Nesterov introduced a momentum-based optimization algorithm that significantly accelerates convergence:

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

where the momentum coefficient β_k is defined as:

$$\beta_k = \frac{\theta_k - 1}{\theta_{k+1}}, \quad \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}.$$

Initialization: Start with x_0, $y_0 = x_0$, and $\theta_1 = 1$.

This advanced setup of β_k using t_k values is critical for optimizing the rate of convergence and is a key aspect of why NAG is highly effective in deep learning optimization.

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

- 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.

Theorem (Nemirovsky-Yudin (1983))

For any $k \leq \frac{N-1}{2}$, and any $x_0 \in \mathbb{R}^N$, there exists a convex function $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^N)$ such that any first-order algorithm satisfies

$$f(x_k) - \min_{\mathcal{H}} f \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2} \text{ and } \|x_k - x^*\|^2 \geq \frac{1}{8}\|x_0 - x^*\|^2.$$

The order of complexity of Nesterov's algorithm is "optimal".

Nesterov's Accelerated Gradient Method (NAG)

NAG can be rewritten as

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

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.

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

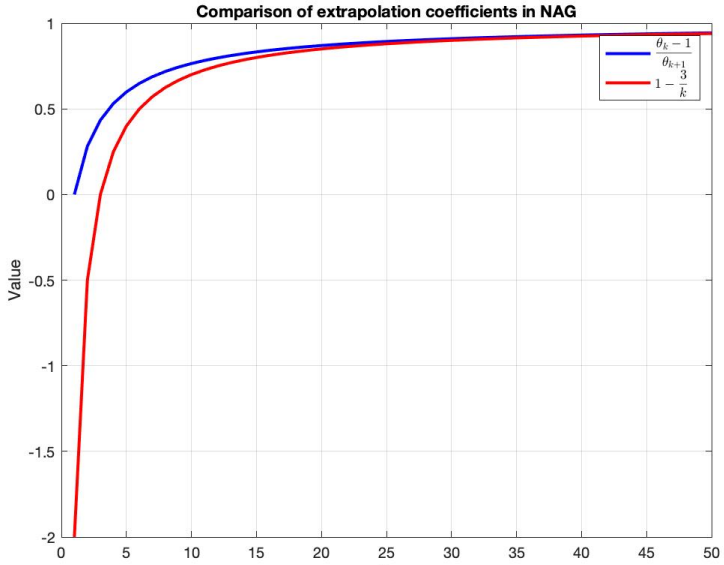
Nesterov's Accelerated Gradient Method (NAG)

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

where the momentum coefficient β_k is defined as:

$$\beta_k = \frac{\theta_k - 1}{\theta_{k+1}}, \quad \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2} \text{ and } \beta_k = 1 - \frac{3}{k}.$$

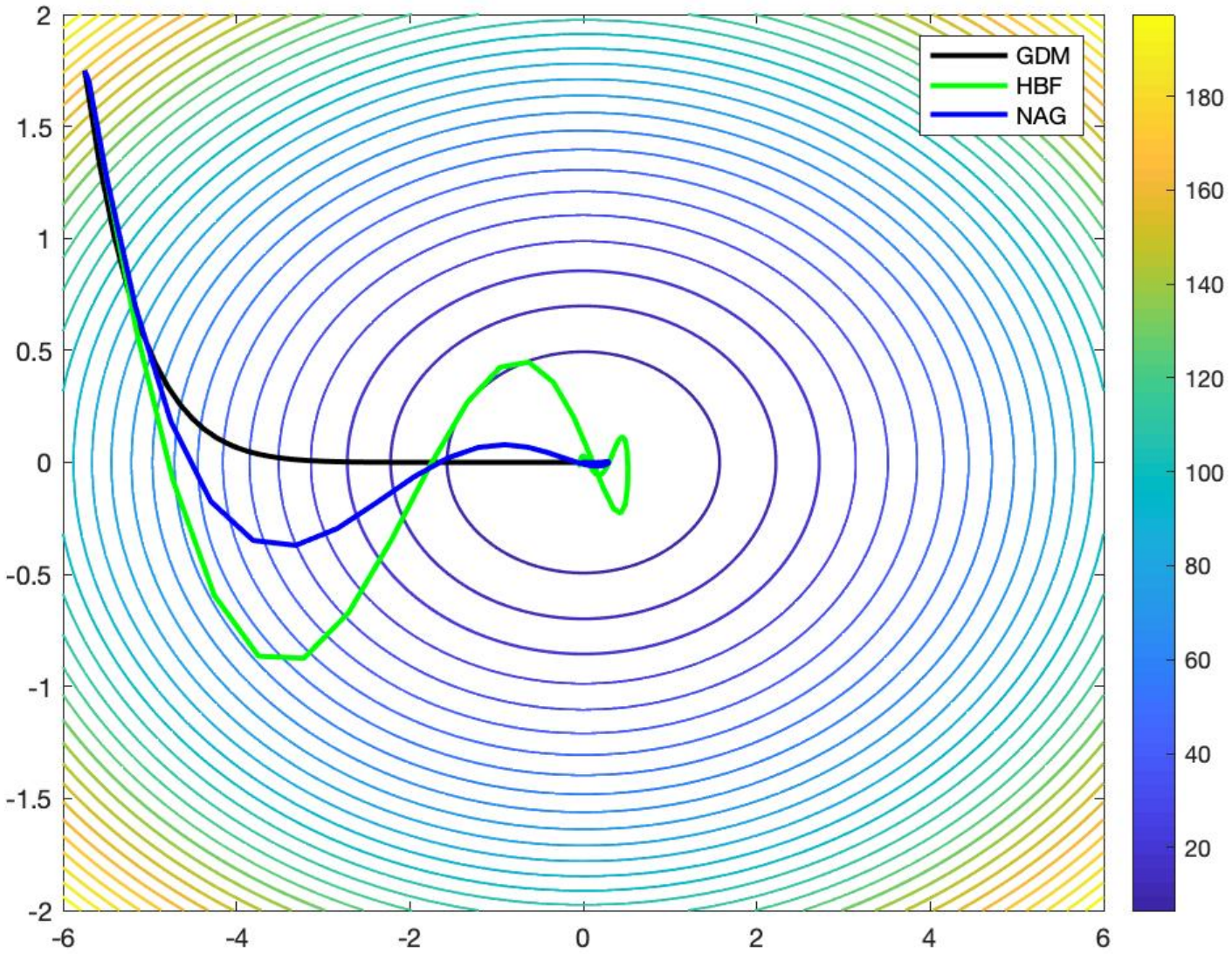
Chambolle-Dossale (JOTA, 2015).



Nesterov's Accelerated Gradient Method (NAG)

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

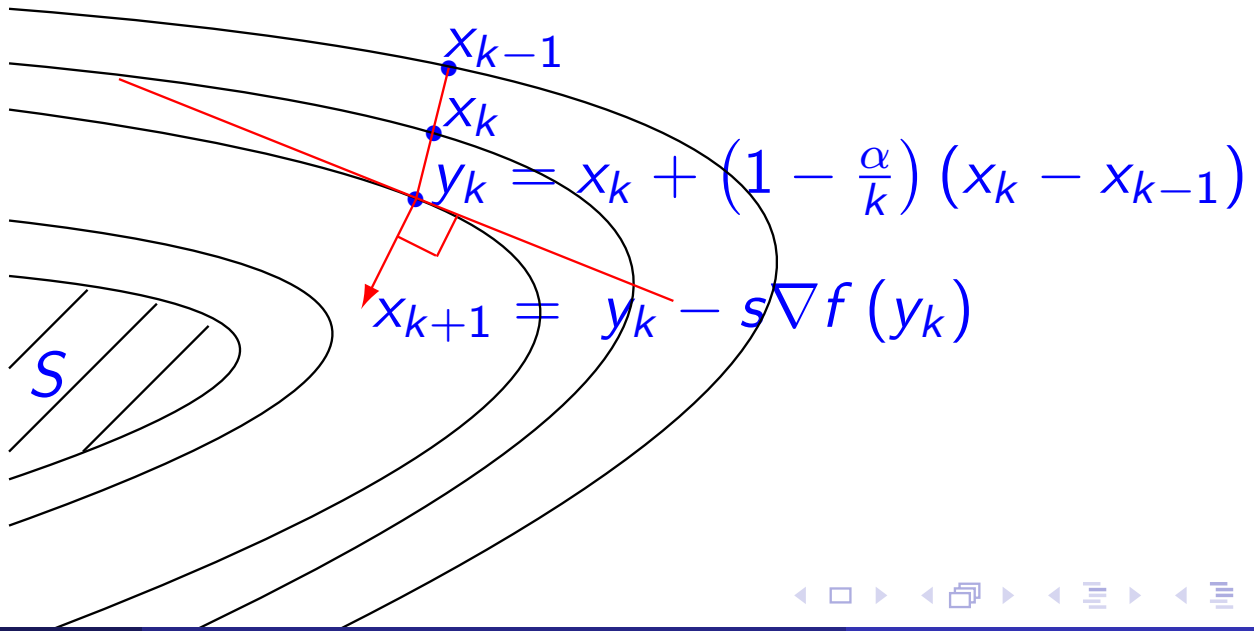
- The gradient step is applied to y_k, which is obtained by linear extrapolation from x_k and x_{k-1}.
- Note the subtle tuning of the extrapolation coefficient $0 < \alpha_k := 1 - \frac{\alpha}{k} < 1$ which tends to one from below as $k \to +\infty$.


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)

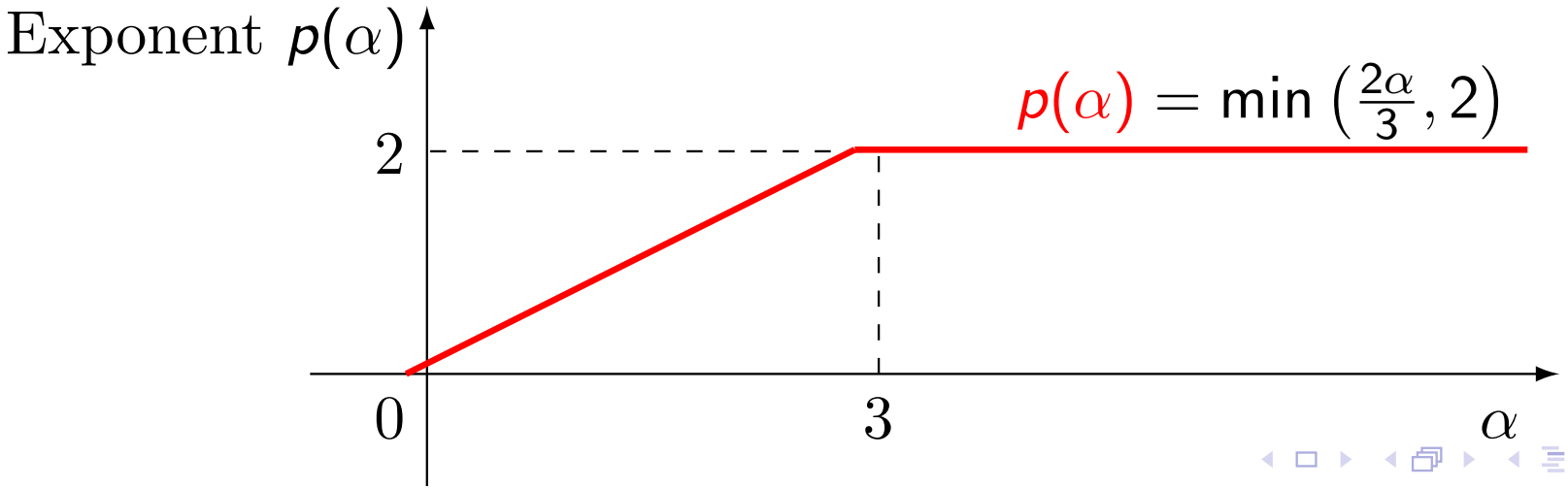
$$(\mathrm{NAG})_\alpha \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla f(y_k). \end{cases}$$

- 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

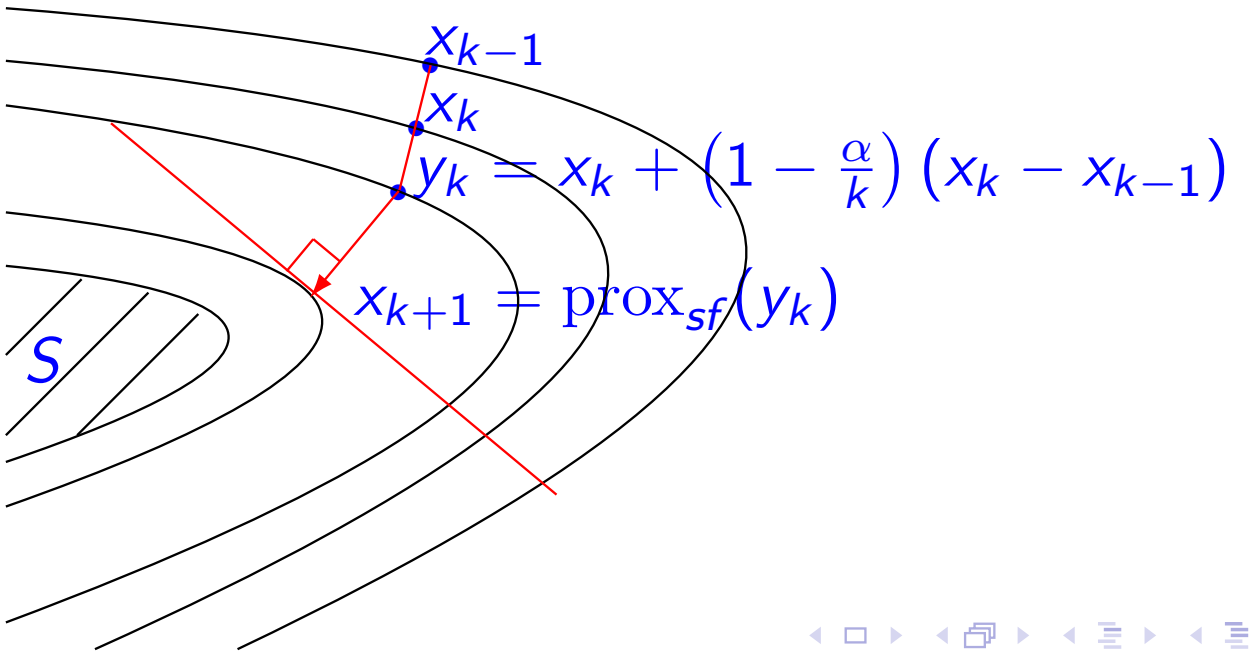
$$s\Big(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\Big)\Big(\nabla f(x_k) - \nabla f(x_{k-1}\Big).$$

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 $g \in \Gamma_0(\mathcal{H})$.

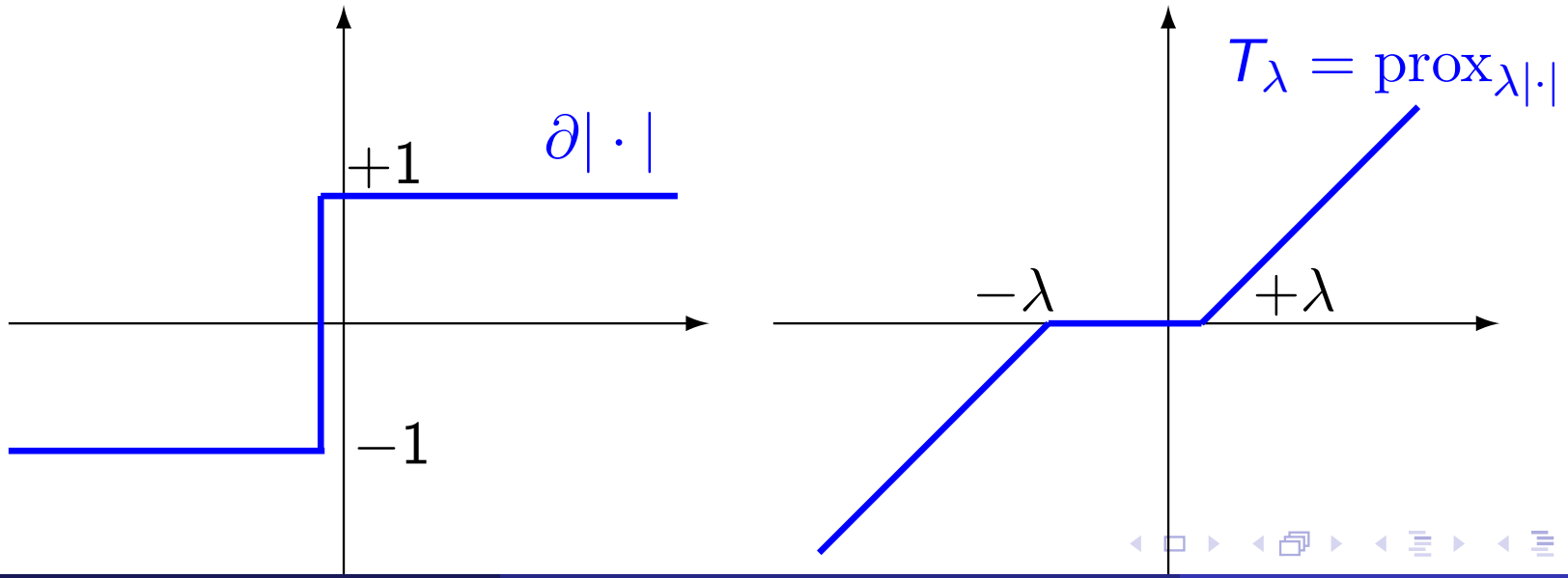
$$\begin{aligned}
0 \in \partial(f+g)(x) &\iff 0 \in \nabla f(x) + \partial g(x) \\
&\iff 0 \in s\nabla f(x) + s\partial g(x), \ s > 0. \\
&\iff x \in x + s\nabla f(x) + s\partial g(x) \\
&\iff x - s\nabla f(x) \in \Big(\mathrm{I} + s\partial g\Big)(x). \\
&\iff x = \mathrm{prox}_{sg}\Big(x - s\nabla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

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)

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

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

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

- Possibility of a backtracking version.

Fast ISTA (FISTA)

Beck-Teboulle 2009

1. Initialize: $x_0 = y_0 \in \mathbb{R}^n$, $\theta_0 = 1$, $s := \frac{1}{L}$ with L being a Lipschitz constant of ∇f, $k := 0$
2. $x_{k+1} = \text{prox}_{sg}(y_k - s\nabla f(y_k))$
3. $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$, $y_{k+1} = x_{k+1} + \left(\frac{\theta_k - 1}{\theta_{k+1}}\right)(x_{k+1} - x_k)$, $k := k + 1$

When $g = 0$, FISTA recovers Nesterov's accelerated gradient method.

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.

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

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
$$\ddot{x}(t) + \tfrac{\alpha}{t}\dot{x}(t) + \sqrt{s}\nabla^2 f(x(t))\dot{x}(t) + (1+\tfrac{\alpha\sqrt{s}}{2t})\nabla f(x(t)) = 0.$$

See Hedy's Lecture tomorrow.

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)

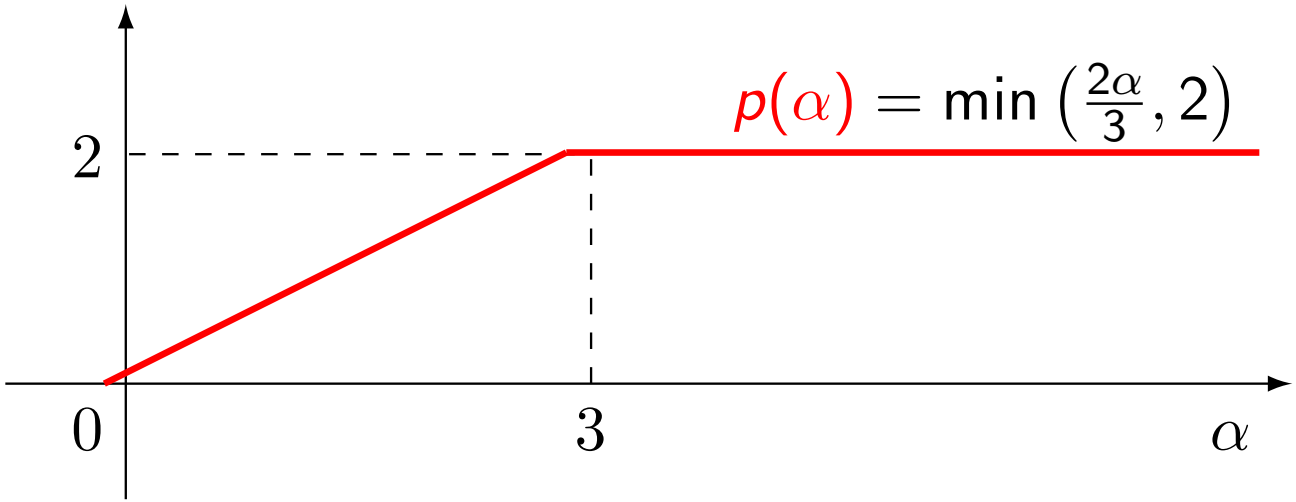
$$f(x(t)) - f^* = o\left(\frac{1}{t^2}\right),\ x(t) \rightharpoonup x_\infty \in S \text{ as } t \to +\infty.$$

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

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

Convergence rates: global view

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



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

Optimal rate: $\begin{cases} \alpha \geq 3 : \ f(x) = \|x\|^r, \ r \to +\infty; \\ \alpha < 3 : \ f(x) = \|x\|. \end{cases}$

Su-Boyd-Candès model: Lyapounov analysis

Asymptotic Vanishing Damping

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

- $E(t) = \frac{2t^2}{\alpha-1}\Big(f(x(t)) - f^*\Big) + (\alpha-1)\|x(t) + \frac{t}{\alpha-1}\dot{x}(t) - x^*\|^2.$
- $E'(t) \leq -\frac{2(\alpha-3)t}{\alpha-1}\Big(f(x(t) - f^*\Big) \leq 0.$
- $\frac{2t^2}{\alpha-1}\Big(f(x(t)) - f^*\Big) \leq E(t) \leq E(0) = (\alpha-1)\|x_0 - x^*\|^2.$

Convergence results.

$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \qquad t \geq t_0 > 0.$$

Theorem (SBC, ACPR, AP, ACR, AAD, AACR)

Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function with $S = \operatorname{argmin} f \neq \emptyset$. Let $x(\cdot)$ be a solution trajectory of $(\mathrm{AVD})_\alpha$. Then, as $t \to +\infty$

- $\alpha \geq 3$: $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^2}\right)$.
- $\alpha > 3$: $f(x(t)) - \min_{\mathcal{H}} f = o\left(\frac{1}{t^2}\right)$, $\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty$, $x(t) \rightharpoonup x_\infty \in S$.
- $0 < \alpha \leq 3$: $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$.

Lyapunov analysis. Case $\alpha \geq 3$

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

Lyapunov function, $z \in \operatorname{argmin} f$, $\lambda > 0$, $\xi \geq 0$ real parameters

$$\mathcal{E}_\lambda(t) := t^2\Big(f(x(t)) - \min_{\mathcal{H}} f\Big) + \tfrac{1}{2}\|\lambda(x(t) - z) + t\dot{x}(t)\|^2.$$

Show that $\lambda > 0$ judiciously chosen $\Longrightarrow \frac{d}{dt}\mathcal{E}_\lambda(t) \leq 0$.

Structure of the Lyapunov function

- Global energy: $W(t) := [f(x(t)) - \min_{\mathcal{H}} f] + \frac{1}{2}\|\dot{x}(t)\|^2$;
- Anchor function: $h(t) := \frac{1}{2}\|x(t) - z\|^2$;

$$\mathcal{E}_\lambda(t) = t^2 W(t) + \lambda^2 h(t) + \lambda t\dot{h}(t).$$

Lyapunov analysis. Case $\alpha \geq 3$

Classical derivation chain rule + $(\mathrm{AVD})_\alpha$ + convexity of f

$$\frac{d}{dt}\mathcal{E}_\lambda(t) \leq -(\lambda-2)\,t\,(f(x(t)) - \min_{\mathcal{H}} f) \\ -\lambda(\alpha-1-\lambda)\langle x(t)-z, \dot{x}(t)\rangle - (\alpha-1-\lambda)\,t\,\|\dot{x}(t)\|^2.$$

Take $\lambda = \alpha - 1$:

$$\frac{d}{dt}\mathcal{E}_{\alpha-1}(t) \leq -(\alpha-3)\,t\,(f(x(t)) - \min_{\mathcal{H}} f). \tag{3}$$

$\mathcal{E}_{\alpha-1}(\cdot)\downarrow$. Hence for $t \geq t_0$, $t^2(f(x(t)) - \min_{\mathcal{H}} f) \leq \mathcal{E}_{\alpha-1}(t) \leq \mathcal{E}_{\alpha-1}(t_0)$

$$\alpha \geq 3: \quad f(x(t)) - \min_{\mathcal{H}} f \leq \frac{\mathcal{E}_{\alpha-1}(t_0)}{t^2}.$$

By integration of (3)

$$\alpha > 3: \quad \int_{t_0}^{+\infty} t\big(f(x(t)) - \min_{\mathcal{H}} f\big)\,dt \leq \frac{\mathcal{E}_{\alpha-1,0}(t_0)}{\alpha-3} < +\infty.$$

3e. $o\left(\frac{1}{t^2}\right)$ rate of convergence for $\alpha > 3$.

Global energy: $W(t) := (f(x(t)) - \inf_{\mathcal{H}} f) + \frac{1}{2}\|\dot{x}(t)\|^2$.

Classical derivation chain rule, and $(\mathrm{AVD})_\alpha$ give

$$\frac{d}{dt}(t^2 W(t)) + (\alpha - 1)t\|\dot{x}(t)\|^2 = 2t(f(x(t)) - \inf_{\mathcal{H}} f).$$

$$\Longrightarrow \int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty \quad \text{and} \quad \lim_{t\to+\infty} t^2 W(t) \text{ exists},$$

$$\int_{t_0}^{+\infty} t\,(f(x(t)) - \min_{\mathcal{H}} f)\,dt < +\infty, \int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty \Longrightarrow \int_{t_0}^{+\infty} tW(t)dt < +\infty,$$

which combined with $\lim_{t\to+\infty} t^2 W(t)$ exists, implies

$$\lim_{t\to+\infty} t^2 W(t) = 0, \text{ that is } W(t) = o\left(1/t^2\right). \tag{4}$$

Lyapunov analysis. General case $0 < \alpha < +\infty$

$$\mathcal{E}(t) = t^{2p}\left[f(x(t)) - \min_{\mathcal{H}} f\right] + \tfrac{1}{2}\|\lambda(t)(x(t)-z) + t^p\dot{x}(t)\|^2 + \tfrac{\xi(t)}{2}\|x(t)-z\|^2$$

Use derivation chain rule, constitutive equation, convex inequality:

$$\begin{aligned}
\tfrac{d}{dt}\mathcal{E}(t) \quad &\leq \quad t^p\left[2pt^{p-1} - \lambda(t))\right](f(x(t)) - \min_{\mathcal{H}} f) \\
&+ \quad \left[\xi(t) + t^p\dot{\lambda}(t) - (\alpha - p)t^{p-1}\lambda(t) + \lambda(t)^2\right]\langle x(t) - z, \dot{x}(t)\rangle \\
&- \quad t^p\left[(\alpha-p)t^{p-1} - \lambda(t)\right]\|\dot{x}(t)\|^2 + \left[\lambda(t)\dot{\lambda}(t) + \tfrac{\dot{\xi}(t)}{2}\right]\|x(t)-z\|^2.
\end{aligned}$$

- $(\mathbf{H}_1)$: $\lambda(t) = 2pt^{p-1} \implies$ first term equal to zero.
- $(\mathbf{H}_2)$: $\xi(t) = 2(\alpha - 4p + 1)pt^{2(p-1)} \implies$ second term equal to zero.
- $(\mathbf{H}_3)$: $\alpha + 1 \geq 4p \implies \xi \geq 0$, and hence $\mathcal{E} \geq 0$.
- $(\mathbf{H}_4)$: $\alpha \geq 3p \implies$ third term less or equal than zero.
- $(\mathbf{H}_5)$: $1 \geq p \implies$ fourth term less or equal than zero.

Take $p = \min(1, \frac{\alpha}{3}, \frac{\alpha+1}{4}) = \min(1, \frac{\alpha}{3})$.

Lyapunov analysis. General case $0 < \alpha < +\infty$

$$\mathcal{E}(t) = t^{2p}\left[f(x(t)) - \min_{\mathcal{H}} f\right] + \frac{1}{2}\|\lambda(t)(x(t)-z)+t^p\dot{x}(t)\|^2 + \frac{\xi(t)}{2}\|x(t)-z\|^2$$

$p = \min(1, \frac{\alpha}{3}) \Longrightarrow \frac{d}{dt}\mathcal{E}(t) \leq 0$, *i.e.* $\mathcal{E}(\cdot)$ is nonincreasing.

Since $\xi(t) \geq 0$, we get for $t \geq t_0$

$$f(x(t)) - \min_{\mathcal{H}} f \leq \frac{\mathcal{E}(t_0)}{t^{2p}}.$$

- Case $\alpha \geq 3$. Then $p = 1$ and $f(x(t)) - \min_{\mathcal{H}} f \leq \frac{\mathcal{E}(t_0)}{t^2}$.
- Case $\alpha \leq 3$. Then $p = \frac{\alpha}{3}$ and $f(x(t)) - \min_{\mathcal{H}} f \leq \frac{\mathcal{E}(t_0)}{t^{\frac{2\alpha}{3}}}$.

Weak convergence of the orbits for $\alpha > 3$

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

Theorem (ACPR, Math. Prog. 2016)

Suppose $\alpha > 3$. *Let* $x : [t_0, +\infty[\to \mathcal{H}$ *be an orbit of* $(\mathrm{AVD})_\alpha$.

Then, $w - \lim_{t\to+\infty} x(t) = x_\infty$ *exists, for some* $x_\infty \in \operatorname{argmin} f$.

Lemma (Opial)

Let $S \subset \mathcal{H}$, $S \neq \emptyset$, *and* $x : [t_0, +\infty[\to \mathcal{H}$ *a map. Assume that*

(i) *for every* $z \in S$, $\lim_{t\to+\infty} \|x(t) - z\|$ *exists;*

(ii) *every weak sequential cluster point of* $x(\cdot)$ *belongs to* S.

Then, $w - \lim_{t\to+\infty} x(t) = x_\infty$ *exists, for some element* $x_\infty \in S$.

Weak convergence of the orbits for $\alpha > 3$

Lemma (differential inequality)

Let $t_0 > 0$, $\alpha > 1$, *and* $w : [t_0, +\infty[\to \mathbb{R}$ *that satisfies*

$$\dot{w}(t) + \tfrac{\alpha}{t} w(t) \leq m(t),$$

for some $m : [t_0, +\infty[\to \mathbb{R}^+$ *such that* $t \mapsto t m(t) \in L^1(t_0, +\infty)$. *Then*

$$w^+ \in L^1(t_0, +\infty).$$

Given $z \in \operatorname{argmin} f$, set $h(t) := \frac{1}{2}\|x(t) - z\|^2$.

$$\dot{h}(t) = \langle x(t) - z, \dot{x}(t) \rangle,$$
$$\ddot{h}(t) = \langle x(t) - z, \ddot{x}(t) \rangle + \|\dot{x}(t)\|^2.$$

Combining these two equations, and using $(\mathrm{AVD})_\alpha$, we obtain

$$\ddot{h}(t) + \frac{\alpha}{t}\dot{h}(t) = \|\dot{x}(t)\|^2 + \langle x(t) - z, \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) \rangle, \tag{5}$$
$$= \|\dot{x}(t)\|^2 + \langle x(t) - z, -\nabla f(x(t)) \rangle. \tag{6}$$

Weak convergence of the orbits for $\alpha > 3$

Equivalently

$$\ddot{h}(t) + \frac{\alpha}{t}\dot{h}(t) + \langle x(t) - z, \nabla f(x(t))\rangle = \|\dot{x}(t)\|^2.$$

By monotonicity of ∇f and $\nabla f(z) = 0$, we infer

$$\ddot{h}(t) + \frac{\alpha}{t}\dot{h}(t) \leq \|\dot{x}(t)\|^2. \tag{7}$$

According to the above Lemma and $\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty$, we obtain $\dot{h}^+ \in L^1(t_0, +\infty)$. Therefore,

- (i) for every $z \in S$, $\lim_{t\to+\infty} \|x(t) - z\|$ exists;
- (ii) of Opial lemma follows from $f(x(t)) \to \min_{\mathcal{H}} f$, and the lower sci. of the convex continuous function f for the weak topology.

5. Sharpness of the convergence rates.

5a. Sharpness of the results: $\alpha \geq 3$

$\mathcal{H} = \mathbb{R}$, $f(x) = c|x|^\gamma$, $c > 0$, $\gamma > 0$, parameters.

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + c\gamma|x(t|^{\gamma-2}x(t) = 0.$$

Nonnegative, completely damped solutions of $(\mathrm{AVD})_\alpha$:

$$x(t) = \frac{1}{t^\theta}, \quad \theta > 0.$$

Replacing $x(\cdot)$ in $(\mathrm{AVD})_\alpha$ gives $\gamma > 2$, $\theta = \frac{2}{\gamma-2}$, $\alpha > \frac{\gamma}{\gamma-2}$, and

$$f(x(t)) = \frac{2}{\gamma(\gamma-2)}(\alpha - \frac{\gamma}{\gamma-2})\frac{1}{t^{\frac{2\gamma}{\gamma-2}}}.$$

As $\gamma \uparrow +\infty$, $\frac{2\gamma}{\gamma-2} \downarrow 2$: f becomes very flat around its minimizer.

5b. Sharpness of the results: $\alpha \leq 3$

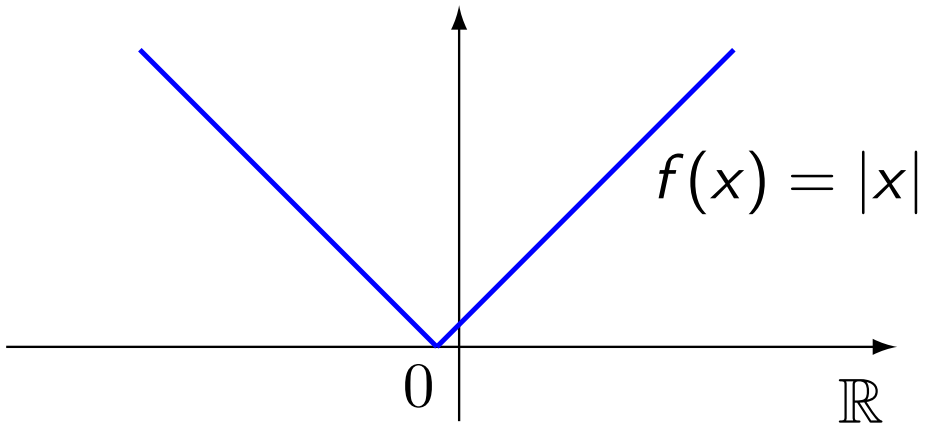
Optimality of the estimate $W(t) = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$.

Aujol-Dossal: take $f(x) = |x|^\delta$, and let δ tend to 1. For $\delta > 1$,

$$W(t) \geq Kt^{-\frac{2\alpha}{1+2/\delta}}.$$

The optimality of the decay rate is obtained by letting δ tend to one.
$\delta = 1 \longrightarrow$ Differential inclusion: Attouch-Cabot, Aujol-Dossal.

$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial(|\cdot|)(x(t)) \ni 0.$$

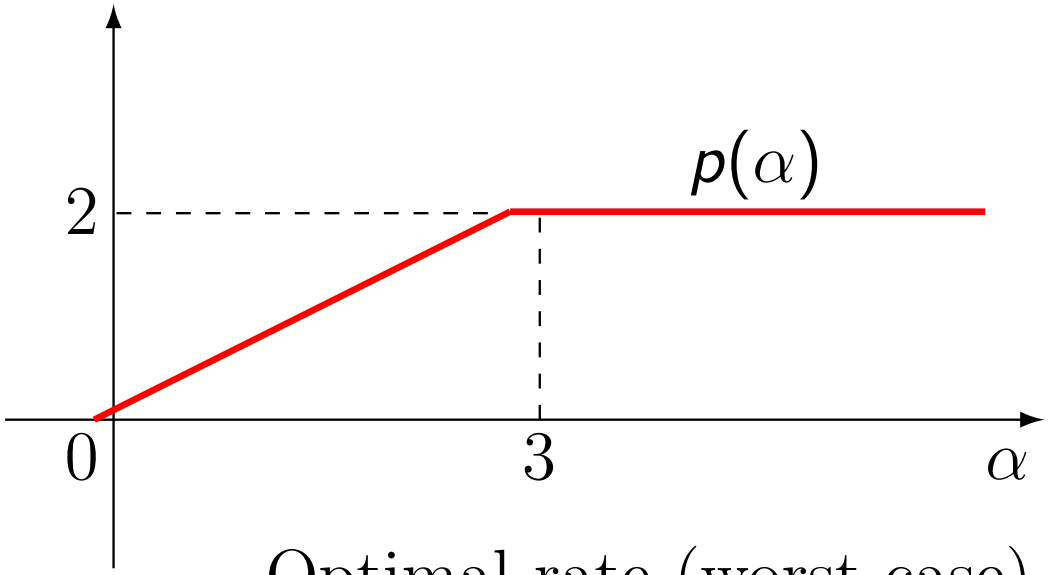


5c. Back to the global picture

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

Rate of convergence of the values as $t \to +\infty$.

$$f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^{p(\alpha)}}\right), \qquad p(\alpha) = \min\left(\frac{2\alpha}{3}, 2\right).$$



Optimal rate (worst case)

6. THE STRONGLY CONVEX CASE

6a. Convergence rate

$f : \mathcal{H} \to \mathbb{R}$ is strongly convex if there exists $\mu > 0$ such that

$$\forall x, y \in \mathcal{H}, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \tag{8}$$

The convergence rates increase indefinitely with larger values of α.

Theorem (ACPR, SBC)

Let $f : \mathcal{H} \to \mathbb{R}$ be strongly convex, and let $x : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of $(\mathrm{AVD})_\alpha$ with $\alpha > 3$. Then $x(t)$ converges strongly, as $t \to +\infty$, to the unique element $x^* \in \operatorname{argmin} f$. Moreover

$$f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(t^{-\frac{2}{3}\alpha}\right), \quad \|x(t) - x^*\|^2 = \mathcal{O}\left(t^{-\frac{2}{3}\alpha}\right)$$

$$\|\dot{x}(t)\|^2 = \mathcal{O}\left(t^{-\frac{2}{3}\alpha}\right).$$

The proof is based on the rescaled Lyapunov function:

$$\mathcal{E}(t) \;=\; t^p \left(t^2 (f(x(t)) - \min_{\mathcal{H}} f) + \frac{1}{2}\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 \right).$$

6b. Case $f(x) = \frac{1}{2}\|x\|^2$.

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

Solution of $(\mathrm{AVD})_\alpha$ with Cauchy data $x(0) = x_0$, $\dot{x}(0) = 0$:

$$x(t) = 2^{\frac{\alpha-1}{2}}\Gamma\left(\frac{\alpha+1}{2}\right)\frac{J_{\frac{\alpha-1}{2}}(t)}{t^{\frac{\alpha-1}{2}}}x_0.$$

$J_{\frac{\alpha-1}{2}}(\cdot)$: first kind Bessel function of order $\frac{\alpha-1}{2}$. For large t,

$$J_\alpha(t) = \sqrt{\frac{2}{\pi t}}\left(\cos\left(t - \frac{\pi\alpha}{2} - \frac{\pi}{4}\right) + \mathcal{O}\left(\frac{1}{t}\right)\right).$$

$$\Rightarrow \quad f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}(t^{-\alpha}).$$

For $\alpha < 2$, we cannot expect the $\mathcal{O}(t^{-2})$ rate of convergence!

6c. Adaptive restart in the strongly convex case

Strategy: maintain high velocity along the orbit $\longrightarrow$ linear convergence

$$(\mathrm{AVD})_\alpha \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad x(0) = x_0, \ \dot{x}(0) = 0.$$

Restarting time: $T(f, x_0) = \sup\{t > 0, \ \forall \tau \in]0, t[, \ \frac{d}{d\tau}\|\dot{x}(\tau)\|^2 > 0\}$.

Before time $T(f, x_0) > 0$, $t \mapsto f(x(t))$ decreases:

$$\frac{d}{dt} f(x(t)) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\frac{\alpha}{t}\|\dot{x}(t)\|^2 - \frac{1}{2}\frac{d}{dt}\|\dot{x}(t)\|^2 \leq 0.$$

At time $T(f, x_0)$, stop and restart, and so on.

Theorem (SBC)

Suppose $f : \mathcal{H} \to \mathbb{R}$ strongly convex, ∇f Lipschitz continuous, $\alpha \geq 3$. Let $x_{sr}(\cdot)$ be an orbit of the speed restarting dynamic. Then

$$f(x_{sr}(t)) - \min_{\mathcal{H}} f \leq c_1 e^{-c_2 t}.$$

Question: adaptive restart for a general convex function f?

Temporal discretization of $(\mathrm{AVD})_\alpha$

$h > 0$ step size, $s = h^2$ and $t = ks$

$$\frac{1}{s}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{ks}(x_k - x_{k-1}) + \nabla f(y_k) = 0.$$

Different choices for y_k

- Nesterov: Inertial Gradient algorithm

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

- Implicit: $y_k = x_{k+1}$, Inertial Proximal algorithm, Güler, Beck-Teboulle

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

Introduction of the Inertial Proximal algorithm (details)

Implicit discretization

$$\frac{1}{s}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{ks}(x_k - x_{k-1}) + \nabla f(x_{k+1}) = 0.$$

After multiplication by s

$$(x_{k+1} - x_k) - (x_k - x_{k-1}) + \frac{\alpha}{k}(x_k - x_{k-1}) + s\nabla f(x_{k+1}) = 0.$$

Equivalently

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

which gives

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

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. Adly and H. Attouch, (2022).

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

Asymptotic Vanishing Damping

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

A first natural approach is to pass directly to the limit as $\alpha \to +\infty$ in the dynamic $(\mathrm{AVD})_\alpha$. Indeed, as shown below, it is a too naive approach which does not give much information on the asymptotic behavior of the trajectories when α is large. The following result follows 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

$$\frac{2s}{\alpha+1}\ddot{y}_\alpha(s) + \dot{y}_\alpha(s) + \nabla f(y_\alpha(s)) = 0, \tag{9}$$

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

$$\dot{y}(s) + \nabla f(y(s)) = 0, \tag{10}$$

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: Example

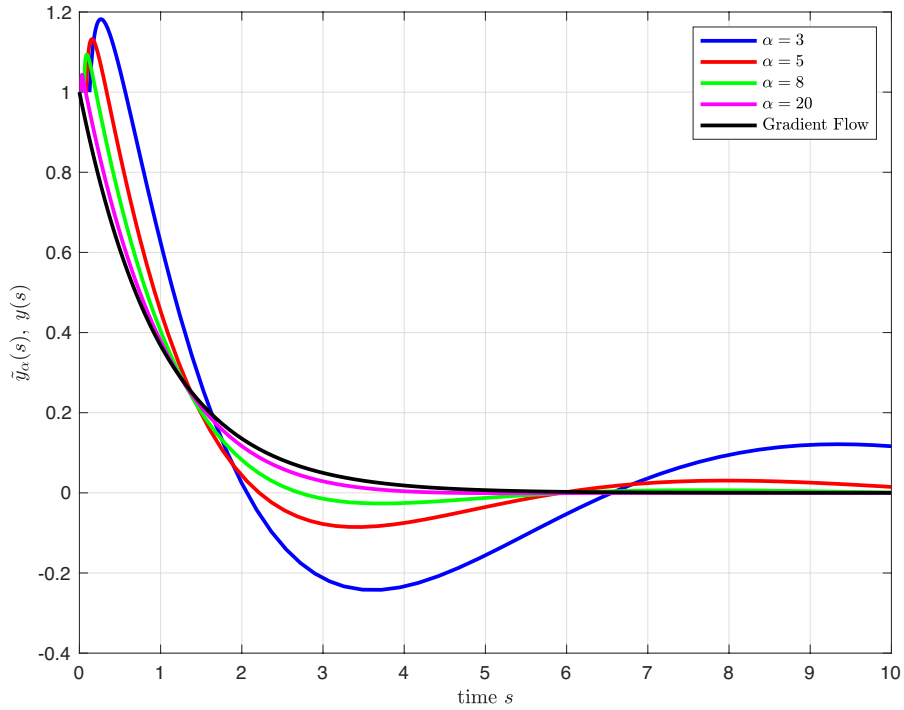
$$(\mathrm{AVD})_\alpha \quad \ddot{x}_\alpha(t) + \frac{\alpha}{t}\dot{x}_\alpha(t) + \lambda\, x_\alpha(t) = 0,$$

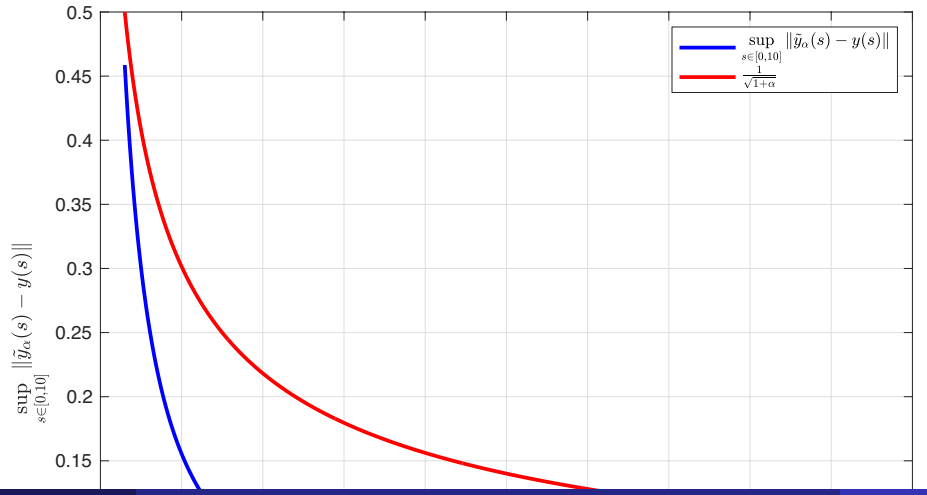
Here $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto f(x) = \frac{1}{2}\lambda x^2$ with $\lambda > 0$, $t_0 = 1$, and $x_0 = 1$ and $x_1 = 0$. Using a symbolic computation software, the solution $t \mapsto x_\alpha(t)$ of $(\mathrm{AVD})_\alpha$ satisfying $x_\alpha(1) = 1$ and $\dot{x}_\alpha(1) = 0$ is given by

$$x_\alpha(t) = \frac{1}{A}\, t^{-\frac{\alpha}{2}+\frac{1}{2}} \left[\mathrm{J}_{\frac{\alpha}{2}+\frac{1}{2}}(\sqrt{\lambda})\mathrm{Y}_{\frac{\alpha}{2}-\frac{1}{2}}(\sqrt{\lambda}t) - \mathrm{Y}_{\frac{\alpha}{2}+\frac{1}{2}}(\sqrt{\lambda})\mathrm{J}_{\frac{\alpha}{2}-\frac{1}{2}}(\sqrt{\lambda}t)\right],$$

with $A = \mathrm{J}_{\frac{\alpha}{2}+\frac{1}{2}}(\sqrt{\lambda})\mathrm{Y}_{\frac{\alpha}{2}-\frac{1}{2}}(\sqrt{\lambda}) - \mathrm{Y}_{\frac{\alpha}{2}+\frac{1}{2}}(\sqrt{\lambda})\mathrm{J}_{\frac{\alpha}{2}-\frac{1}{2}}(\sqrt{\lambda})$.
Here $\mathrm{J}_\nu(\cdot)$ and $\mathrm{Y}_\nu(\cdot)$ are the Bessel functions of the first and second kind respectively of order $\nu \in \mathbb{R}$. Accordingly, we can compute explicitly $f(x_\alpha(t)) - \inf_{\mathcal{H}} f$ for different values of t and α.

Time rescaling approach: the convex case





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 $A = \mathrm{diag}([\lambda_1, \lambda_2])$.

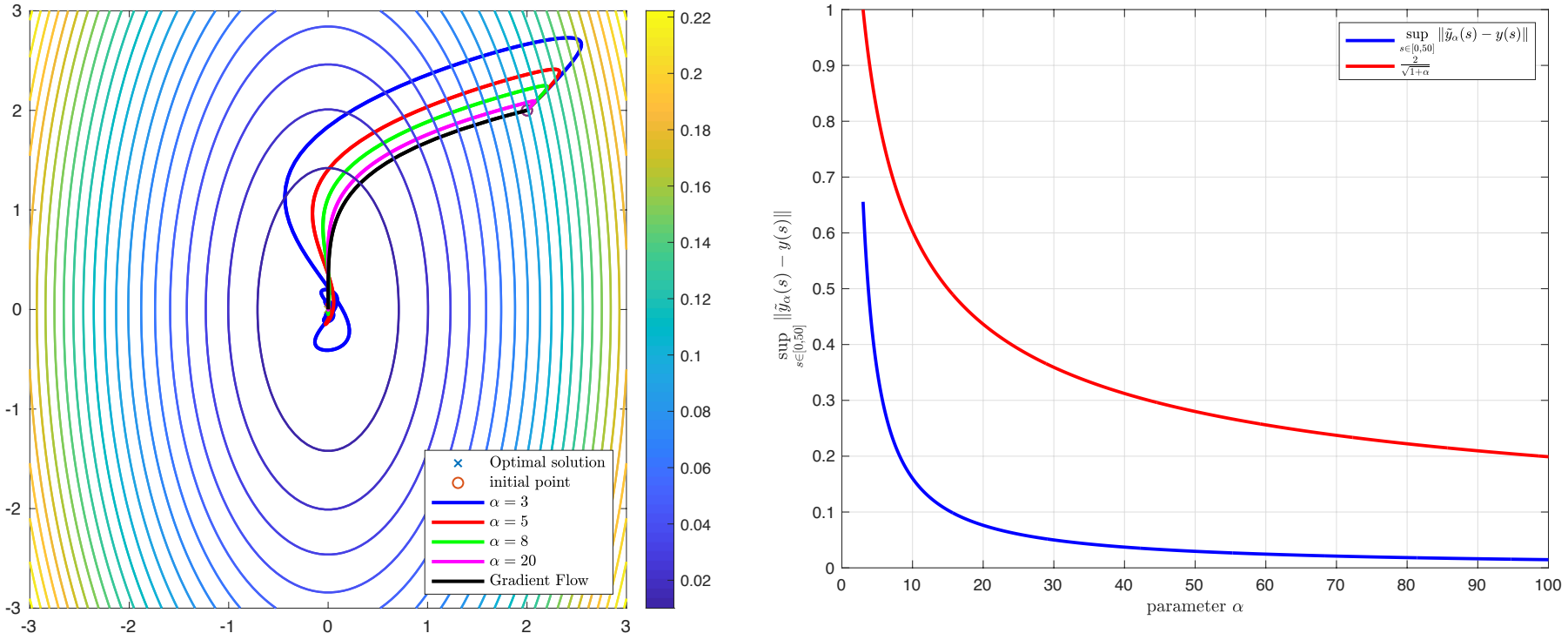
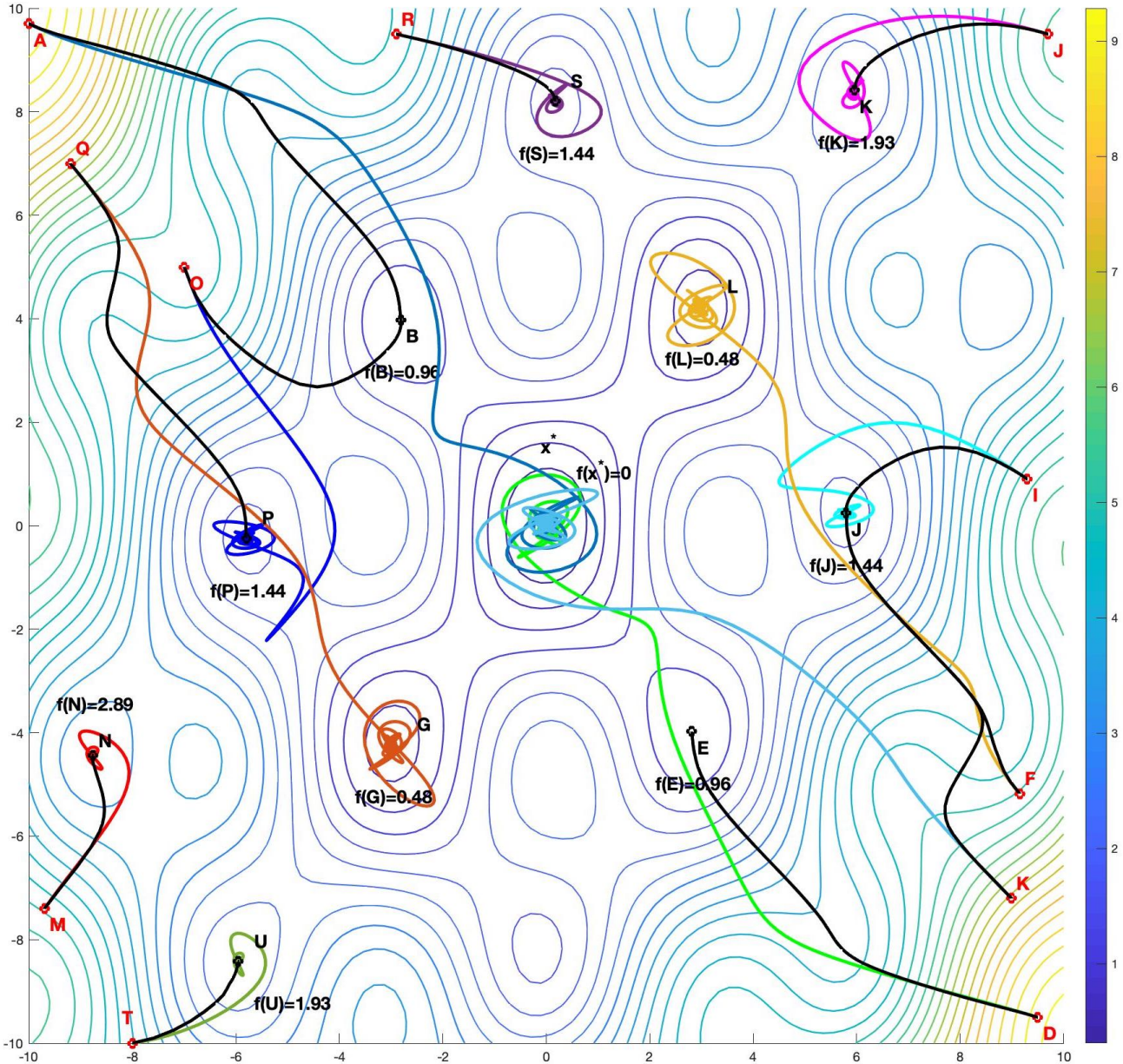


Figure: Illustration on a quadratic convex function.

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



6. Perspective, open questions

Some open questions

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

- 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

The continuation of the course tomorrow with Hedy.

References

- 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.
- 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).

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).