Fast algorithmic methods for optimization and learning. Recent trends.

Hedy ATTOUCH

Université Montpellier
Institut Montpelliérain Alexander Grothendieck, UMR CNRS 5149

GDR MOA 2022 workshop
Mini-Course
Nice, Université Côte d'Azur

October 12th, 2022

Introduction

Convex differentiable optimization

- $\mathcal{H}$ real Hilbert space, $\langle x, x\rangle = \|x\|^2$.
- $f : \mathcal{H} \to \mathbb{R}$ convex differentiable, $S = \operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$.

$$(\mathcal{P}) \quad \min\{f(x) : x \in \mathcal{H}\}.$$

Objective: develop fast first-order algorithms to solve $(\mathcal{P})$

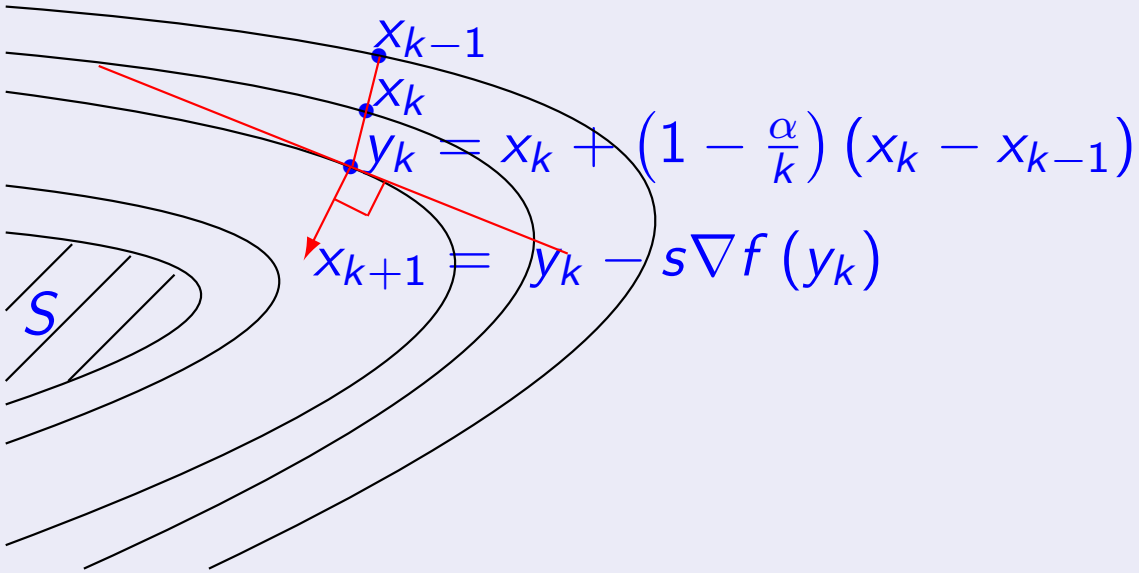
- $x_{k+1} \in x_0 + \operatorname{span}\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)\}$.
- Improve Nesterov accelerated gradient method.

Method: link between algorithms and dynamical systems

- Damped inertial dynamics. Lyapunov analysis.
- High resolution ODE's.
- Time scaling and averaging method.

Nesterov method (1983), ∇f L-Lipschitz.

$$\begin{cases} y_k &= x_k + \left(1 - \frac{3}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \end{cases}$$



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

Optimal order for first-order methods: Nemirovsky-Yudin (1983).

Su-Boyd-Candès dynamic version of Nesterov method

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

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

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

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

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

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

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

Introducing geometric damping driven by the Hessian

$$\ddot{x}(t) + \underbrace{\frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t)}_{\text{damping force}} + \underbrace{b(t)\nabla f(x(t))}_{\text{driving force}} = 0.$$

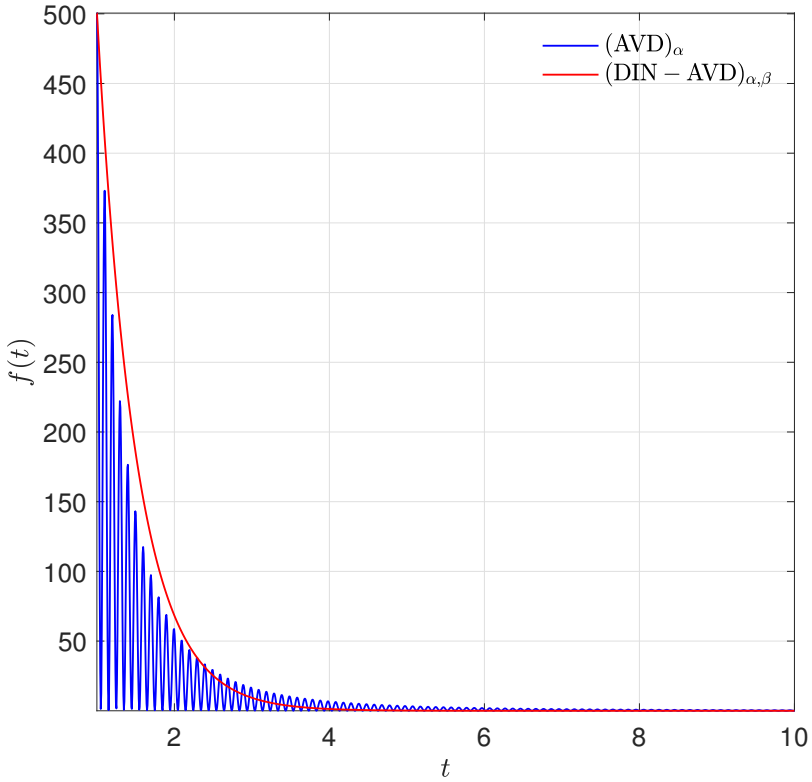
Damping: $\begin{cases} \frac{\alpha}{t}\dot{x}(t) : & \text{accelerated gradient method of Nesterov;} \\ \beta\nabla^2 f(x(t))\dot{x}(t) : & \text{neutralization of oscillations.} \end{cases}$

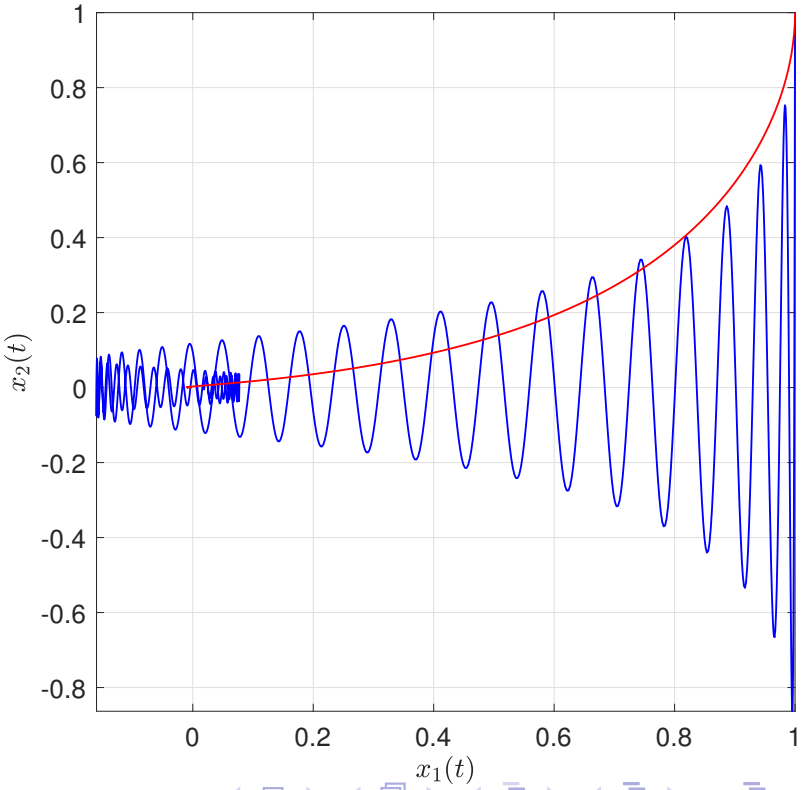
- Provide a better dynamic interpretation of Nesterov method than Su-Boyd-Candès (high resolution ODE versus low resolution ODE).
- Allows to develop new first-order optimization algorithms numerically improving Nesterov.

Hessian driven damping neutralizes oscillations

- $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 1000x_2^2)$: ill-conditioned.
- $\alpha = 3.1$, $\beta = 1$.
- Initial conditions: $(x_1(1), x_2(1)) = (1, 1)$, $(\dot{x}_1(1), \dot{x}_2(1)) = (0, 0)$.

Blue: without Hessian damping, Red: with Hessian damping





Hessian driven damping gives first-order algorithms

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + b(t)\nabla f(x(t)) = 0.$$

$$\nabla^2 f(x(t))\dot{x}(t) = \frac{d}{dt}\nabla f(x(t)) \longrightarrow \text{ first-order algorithms.}$$

(IGAHD) algorithm (ACFR, Math. Prog. 2020)

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

Convergence rates. ∇f L-Lipschitz, $\alpha \geq 3$, $0 < \beta < 2\sqrt{s}$, $sL \leq 1$.

i) $f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right)$ *as* $k \to +\infty$;

ii) $\sum_k k^2 \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_k k^2 \|\nabla f(y_k)\|^2 < +\infty$.

Contents

1. THE HESSIAN-DRIVEN DAMPING: DIFFERENT PERSPECTIVES

Based on:

- A.-Chbani-Fadili-Riahi (Math. Prog. 2022), arXiv:1907.10536v1
- A.-Fadili (SIOPT 2022), arXiv:2201.11643v2

Hessian driven damping from different perspectives

- High resolution ODE of the Nesterov accelerated gradient method.
- Geometric damping adapted to the function to be minimized.
- Levenberg-Marquard regularization of the Newton method.
- Nonsmooth aspects. Damped shocks in mechanics.
- Strong damping in PDE's, control theory.

1.1 High resolution ODE of Nesterov gradient method

- High resolution method: extensively used in fluid mechanics (physical phenomena occur at multiple scales).
- Idea: not let $h \to 0$. Take into account the terms of order $h = \sqrt{s}$ in the Taylor expansions. Discard the higher order terms.
- The Hessian-driven damping appears in the associated continuous inertial ODE.

Theorem (A.-Fadili, SIOPT, 2022), (Shi-Du-Jordan-Su, Math. Prog., 2021)

Assume that f is $\mathcal{C}^2$. The high resolution ODE *with temporal step size $\sqrt{s}$ of Nesterov accelerated gradient method gives the inertial dynamic*

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

High resolution ODE of Nesterov gradient method

Write (NAG) equivalently as

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

With $s = h^2$, this is equivalent to

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

- Ansatz: $x_k = X(t_k)$ for some smooth curve $t \mapsto X(t)$, $t_k = h(k + c)$.
- Taylor expansion / $h \simeq 0$:

$$x_{k+1} = X(t_{k+1}) = X(t_k) + h\dot{X}(t_k) + \frac{1}{2}h^2\ddot{X}(t_k) + +\frac{1}{6}h^3\dddot{X}(t_k) + \mathcal{O}(h^4) \tag{2}$$

$$x_{k-1} = X(t_{k-1}) = X(t_k) - h\dot{X}(t_k) + \frac{1}{2}h^2\ddot{X}(t_k) - \frac{1}{6}h^3\dddot{X}(t_k) + \mathcal{O}(h^4). \tag{3}$$

By adding (2) and (3), we obtain

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} = \ddot{X}(t_k) + \mathcal{O}(h^2).$$

High resolution ODE of Nesterov gradient method

Moreover, (3) gives

$$\frac{x_k - x_{k-1}}{h} = \dot{X}(t_k) - \frac{1}{2}h\ddot{X}(t_k) + \mathcal{O}(h^2).$$

We also have

$$\begin{aligned}
\nabla f(y_k) &= \nabla f\left(x_k + h\left(1 - \frac{\alpha}{k}\right)\frac{x_k - x_{k-1}}{h}\right) \\
&= \nabla f\left(X(t_k) + h\left(1 - \frac{\alpha}{k}\right)\left(\dot{X}(t_k) - \frac{1}{2}h\ddot{X}(t_k) + \mathcal{O}(h^2)\right)\right) \\
&= \nabla f\left(X(t_k) + h\left(1 - \frac{\alpha}{k}\right)\dot{X}(t_k) + \mathcal{O}(h^2)\right) \\
&= \nabla f(X(t_k)) + h\left(1 - \frac{\alpha}{k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2).
\end{aligned}$$

Putting this with (2) and (3) into (1), we obtain

$$\ddot{X}(t_k) + \frac{\alpha}{kh}\left(\dot{X}(t_k) - \frac{1}{2}h\ddot{X}(t_k)\right) + \nabla f(X(t_k)) + h\left(1 - \frac{\alpha}{k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2) = 0.$$

High resolution ODE of Nesterov gradient method

Equivalently,

$$\left(1-\frac{\alpha}{2k}\right)\ddot{X}(t_k)+\frac{\alpha}{kh}\dot{X}(t_k)+\nabla f(X(t_k))+h\left(1-\frac{\alpha}{k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k)+\mathcal{O}(h^2)=0.$$

Dividing by $\left(1-\frac{\alpha}{2k}\right)$ gives

$$\ddot{X}(t_k)+\frac{\alpha}{h(k-\frac{\alpha}{2})}\dot{X}(t_k)+\left(1+\frac{\alpha h}{2h(k-\frac{\alpha}{2})}\right)\nabla f(X(t_k))$$
$$+h\left(1-\frac{\frac{\alpha}{2}}{k-\frac{\alpha}{2}}\right)\nabla^2 f(X(t_k))\dot{X}(t_k)+\mathcal{O}(h^2)=0.$$

Set $c=-\frac{\alpha}{2}$ and thus $t_k=h(k-\frac{\alpha}{2})$. We obtain

$$\ddot{X}(t_k)+\frac{\alpha}{t_k}\dot{X}(t_k)+\left(1+\frac{\alpha h}{2t_k}\right)\nabla f(X(t_k))+h\left(1-\frac{\frac{\alpha h}{2}}{t_k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k)+\mathcal{O}(h^2)=0.$$

Then neglect the term of order $s=h^2$, and keep terms of order $h=\sqrt{s}$.

Implicit Hessian driven damping

Start from (1): $\frac{x_{k+1}-2x_k+x_{k-1}}{h^2} + \frac{\alpha}{kh}\frac{x_k-x_{k-1}}{h} + \nabla f(y_k) = 0.$

Similar Taylor expansions, but without developping the gradient, give

$$\ddot{X}(t_k) + \frac{\alpha}{h(k-\frac{\alpha}{2})}\dot{X}(t_k) + \frac{kh}{h(k-\frac{\alpha}{2})}\nabla f\left(X(t_k) + h\left(1-\frac{\alpha}{k}\right)\dot{X}(t_k)\right) + \mathcal{O}(h^2) = 0.$$

Set $t_k = h(k-\frac{\alpha}{2})$. We obtain successively

$$\ddot{X}(t_k) + \frac{\alpha}{t_k}\dot{X}(t_k) + \frac{t_k+\frac{\alpha h}{2}}{t_k}\nabla f\left(X(t_k) + h\left(1-\frac{\alpha h}{t_k+\frac{\alpha h}{2}}\right)\dot{X}(t_k)\right) + \mathcal{O}(h^2) = 0.$$

$$\ddot{X}(t_k) + \frac{\alpha}{t_k}\dot{X}(t_k) + \left(1+\frac{\alpha h}{2t_k}\right)\nabla f\Big(X(t_k) + h\dot{X}(t_k)\Big) + \mathcal{O}(h^2) = 0.$$

Then neglect the term of order $s = h^2$, and keep terms of order $h = \sqrt{s}$.

Implicit Hessian (Alesca-Laszlo-Pinta, AMO 2020)

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

Continuous versus discrete dynamic

Accelerated gradient algorithms: Nesterov, Ravine, IGAHD...

(high resolution ODE) $\downarrow$ $\uparrow$ (temporal discretization)

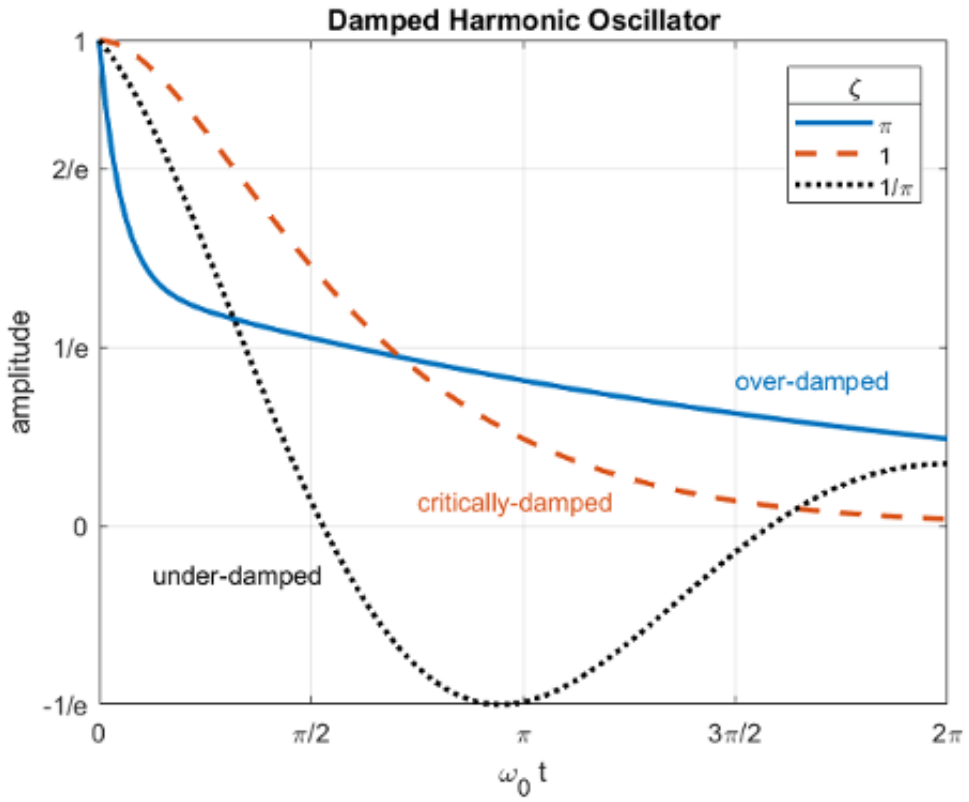
High resolution ODE's, $\theta \in [0,1]$ parameter

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

1.2 Hessian damping is adapted to the geometry of f

Mass-spring system: $\ddot{x}(t) + \beta\dot{x}(t) + \mu x(t) = 0$.

- underdamping: $\beta < 2\sqrt{\mu}$, overdamping: $\beta > 2\sqrt{\mu}$.
- critical damping: $\beta = 2\sqrt{\mu} \longrightarrow$ optimal rate $|x(t)| \leq Ce^{-\sqrt{\mu}t}$.



Polyak versus Nesterov methods / Combine them

Polyak is good for strongly convex functions

$f : \mathcal{H} \to \mathbb{R}$ μ-strongly convex: $f(x) := g(x) + \frac{\mu}{2}\|x\|^2$ with g convex.

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

- $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(e^{-\sqrt{\mu}t}\right)$ as $t \to +\infty$.
- $\neq$ Nesterov provides convergence rate $\mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$.

Nesterov is good for general convex functions

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

- $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^2}\right)$ as $t \to +\infty$ (when $\alpha \geq 3$).
- $\neq$ Polyak provides convergence rate $\mathcal{O}\left(\frac{1}{t}\right)$.

Historical aspects

Hessian driven damping: clever geometric adaptive damping

- (HBF) $\ddot{x}(t) + \Gamma\dot{x}(t) + \nabla f(x(t)) = 0.$

  $\Gamma : \mathcal{H} \to \mathcal{H}$ anisotropic, Alvarez (SICON 2000).

- $(\mathrm{DIN})_\beta$ $\ddot{x}(t) + \gamma\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0.$

  Alvarez-A.-Bolte-Redont (JMPA 2002), A.-Maingé-Redont (DEA 2012).

- $(\mathrm{DIN}-\mathrm{AVD})_{\alpha,\beta}$ $\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0.$

  A.-Peypouquet-Redont (JDE '16), A.-Chbani-Fadili-Riahi (Math Prog '20)

1.3 Regularization of Newton method

Newton method: f convex, $\mathcal{C}^2$, solve $\nabla f(x) = 0$

Discrete: $\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0.$

Continuous: $\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0.$

Levenberg-Marquardt regularization, (A.-Svaiter, SICON, 2011)

$$\gamma(t)\dot{x}(t) + \nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0.$$

Valid with a general maximally monotone operator, closed-loop form.

Dynamic Inertial Newton method

$$\text{(DIN)} \quad \ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0.$$

1.4 Nonsmooth aspects. Damped shocks in mechanics.

Alvarez-A.-Bolte-Redont (JMPA 2002), A.-Peypouquet-Redont (JDE 2016)

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0.$$

$$\Updownarrow$$

$$\begin{cases} \dot{x}(t) + \beta\nabla f(x(t)) - \left(\frac{1}{\beta} - \frac{\alpha}{t}\right)x(t) + \frac{1}{\beta}y(t) = 0; \\ \dot{y}(t) - \left(\frac{1}{\beta} - \frac{\alpha}{t} + \frac{\alpha\beta}{t^2}\right)x(t) + \frac{1}{\beta}y(t) = 0. \end{cases}$$

- Nonsmooth: $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ nonsmooth, (non) convex, proper.

  Damped shocks in mechanics: A.-Maingé-Redont (DEA 2012).

  The normal component of the velocity is killed during a shock.

- Numerical applications (temporal discretization):

  Castera-Bolte-Févotte-Pauwels (Deep Learning) (JMLR 2021).

1.5 Wave equation with strong damping

Damped wave equation, $u : (x,t) \in \Omega \times [0,+\infty[\to \mathbb{R}$

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + \gamma \dfrac{\partial u}{\partial t} + \beta(-\Delta)^\theta(\dfrac{\partial u}{\partial t}) - \Delta u = 0 \quad x \in \Omega, t > 0; \\ u(0,x) = u_0(x) \quad x \in \Omega \\ \dfrac{\partial u}{\partial t}(0,x) = u_1(x) \quad x \in \Omega \\ u(x,t) = 0 \quad x \in \partial\Omega,\ t > 0. \end{cases}$$

- Dirichlet energy $f(u) = \frac{1}{2}\int_\Omega \|\nabla u(x)\|^2 dx$ on Sobolev space $\mathrm{H}_0^1(\Omega)$.
- Fractional powers of the Laplacian $(-\Delta)^\theta$, $\theta \in [\frac{1}{2}, 1]$.
  $\theta = \frac{1}{2}$ critical case; $\theta = 1$ Hessian driven damping.
- Rich litterature related to linear and nonlinear PDE's.

2. LYAPUNOV ANALYSIS OF DYNAMICS/ALGORITHMS WITH HESSIAN-DRIVEN DAMPING

Based on:

- A.-Chbani-Fadili-Riahi (Math. Prog. 2022), arXiv:1907.10536
- A.-Peypouquet-Redont (JDE. 2016), arXiv:1601.07113

Lyapunov analysis, continuous case

$$(\mathrm{DIN}-\mathrm{AVD})_{\alpha,\beta} \quad \ddot{x}(t)+\frac{\alpha}{t}\dot{x}(t)+\beta\nabla^2 f(x(t))\dot{x}(t)+\nabla f(x(t))=0.$$

Theorem (A.-Peypouquet-Redont, JDE 2016)

Let $x:[t_0,+\infty[\to\mathcal{H}$ *be a solution trajectory of* $(\mathrm{DIN}-\mathrm{AVD})_{\alpha,\beta}$.
Suppose that $\alpha>3,\quad \beta>0$. *Then, as* $t\to+\infty$

- $f(x(t))-\min_{\mathcal{H}} f = o\left(\frac{1}{t^2}\right)$
- $\int_{t_0}^{+\infty} t^2\|\nabla f(x(t))\|^2 dt < +\infty.$
- $x(t) \rightharpoonup x_\infty \in \operatorname{argmin}_{\mathcal{H}} f.$

$$(\mathrm{DIN}-\mathrm{AVD})_{\alpha,\beta} \quad \ddot{x}(t)+\frac{\alpha}{t}\dot{x}(t)+\beta\nabla^2 f(x(t))\dot{x}(t)+\nabla f(x(t))=0.$$

Lyapunov function

$$\mathcal{E}_{\alpha,\beta}(t) := t(t-\beta)\left(f(x(t))-f(x^*)\right)+\tfrac{1}{2}\|(\alpha-1)(x(t)-x^*)+t\left(\dot{x}(t)+\beta\nabla f(x(t))\right)\|^2.$$

Derivation of $\mathcal{E}_{\alpha,\beta}(\cdot)$

$$\dot{\mathcal{E}}_{\alpha,\beta}(t)+\Big((\alpha-3)t-\beta(\alpha-2)\Big)\Big(f(x(t))-f(x^*)\Big)+\beta t(t-\beta)\|\nabla f(x(t))\|^2 \leq 0.$$

- $\alpha > 3$, $\quad t \geq t_1 := \beta\frac{\alpha-2}{\alpha-3} \implies \dot{\mathcal{E}}_{\alpha,\beta}(t) \leq 0$ *i.e.* $\mathcal{E}_{\alpha,\beta}(\cdot)$ decreasing.

$$\implies f(x(t))-\min_{\mathcal{H}} f \leq \frac{\mathcal{E}_{\alpha,\beta}(t_1)}{t(t-\beta)} = \mathcal{O}\left(\frac{1}{t^2}\right).$$

- $\beta > 0$, integration $\implies \displaystyle\int_{t_0}^{\infty} t^2\|\nabla f(x(t))\|^2 dt < +\infty.$

(IGAHD): Inertial gradient algorithm with Hessian damping

A.-Chbani-Fadili-Riahi (Math. Prog. 2020), (ACFR) for short.

$f : \mathcal{H} \to \mathbb{R}$ convex, ∇f L-Lipschitz continuous.

Temporal rescaling of (DIN-AVD)$_{\alpha,\beta}$ (see high resolution ODE)

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

Temporal discretization: $s = h^2$, $\nabla^2 f(x(t))\dot{x}(t) = \frac{d}{dt}\nabla f(x(t))$.

$$\frac{1}{s}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{ks}(x_k - x_{k-1}) + \frac{\beta}{\sqrt{s}}(\nabla f(x_k) - \nabla f(x_{k-1}))$$

$$+\frac{\beta}{k\sqrt{s}}\nabla f(x_{k-1}) + \nabla f(y_k) = 0.$$

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

Choose $y_k \approx$ Nesterov's accelerated gradient method, set $\alpha_k = 1 - \frac{\alpha}{k}$.

(IGAHD): Inertial Gradient Algorithm with Hessian Damping

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) - \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1}) \\ x_{k+1} = y_k - s\nabla f(y_k). \end{cases}$$

Related algorithm: Shi-Du-Jordan-Su (Math. Prog. 2021).

Lyapunov analysis, $x^\star \in \operatorname{argmin}_{\mathcal{H}} f$, $\quad t_k := \frac{k-1}{\alpha-1}$.

$$\mathcal{E}_k := t_k^2(f(x_k) - f(x^\star)) + \tfrac{1}{2s}\|v_k\|^2$$

$$v_k := (x_{k-1} - x^\star) + t_k\Big(x_k - x_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\Big).$$

Theorem (ACFR, 2019)

- $f : \mathcal{H} \to \mathbb{R}$ *convex*, ∇f *L-Lipschitz continuous*, $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$.
- $\alpha \geq 3$, $0 < \beta < 2\sqrt{s}$, $sL \leq 1$.

$(x_k)_{k\in\mathbb{N}}$ *generated by* (IGAHD). *Then* $(\mathcal{E}_k)_{k\in\mathbb{N}}$ *is non-increasing and*

$$i)\ f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right) \text{ as } k \to +\infty;$$

$$ii) \sum_k k^2\|\nabla f(y_k)\|^2 < +\infty \text{ and } \sum_k k^2\|\nabla f(x_k)\|^2 < +\infty.$$

$iii)$ *If* $\alpha > 3$, *then* (x_k) *converges weakly to some* $x^\star \in \operatorname{argmin}_{\mathcal{H}} f$.

Reinforced version of the gradient descent lemma, f convex, $s \leq \frac{1}{L}$

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

Write it successively at $y = y_k$ and $x = x_k$, then at $y = y_k$, $x = x^\star$.

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(y_k), y_k - x_k\rangle - \frac{s}{2}\|\nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(x_k) - \nabla f(y_k)\|^2$$
$$f(x_{k+1}) \leq f(x^\star) + \langle \nabla f(y_k), y_k - x^\star\rangle - \frac{s}{2}\|\nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(y_k)\|^2.$$

Linear combination of the two above equations gives

$$t_{k+1}^2(f(x_{k+1}) - f(x^\star)) \leq (t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - f(x^\star)) + t_k^2(f(x_k) - f(x^\star))$$
$$+ t_{k+1}\langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star\rangle - \frac{s}{2}t_{k+1}^2\|\nabla f(y_k)\|^2$$
$$-\frac{s}{2}(t_{k+1}^2 - t_{k+1})\|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{s}{2}t_{k+1}\|\nabla f(y_k)\|^2.$$

Since $\alpha \geq 3$ we have $t_{k+1}^2 - t_{k+1} - t_k^2 \leq 0$...

Numerical experiments

Regularized Least Square (signal/image, machine learning, statistics)

$$\text{(RLS)} \quad \min_{x\in\mathbb{R}^n} \left\{ f(x) := \frac{1}{2}\|Ax-b\|^2 + g(x) \right\}$$

- A linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, $m \leq n$, $b \in \mathbb{R}^m$.
- $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc. convex: regularizer.

Work with the metric $\|x\|_M^2 = \langle Mx, x\rangle$, where $M = \lambda^{-1} I - A^* A$.

$$0 < \lambda\|A\|^2 < 1 \Longrightarrow M \text{ is symmetric positive definite.}$$

Apply (IGAHD) to f^M: Moreau envelope of f in the metric M

$$f^M(x) := \min_{\xi\in\mathbb{R}^n} \left\{ f(\xi) + \tfrac{1}{2}\|x-\xi\|_M^2 \right\}.$$

f^M is convex $\mathcal{C}^1$; ∇f^M (in the metric M) is 1-Lipschitz, and

$$\nabla f^M(x) = x - \operatorname{prox}_{\lambda g}\left(x - \lambda A^*(Ax-b)\right).$$

(IGAHD) for (RLS)

Initialize: $x_0 \in \mathbb{R}^n$, $x_1 \in \mathbb{R}^n$

$$\begin{cases} z_k = x_k - \mathrm{prox}_{\lambda g}(x_k - \lambda A^*(Ax_k - b)); \\ y_k = x_k + (1 - \frac{\alpha}{k})(x_k - x_{k-1}) - \beta\sqrt{s}(z_k - z_{k-1}) - \frac{\beta\sqrt{s}}{k} z_k; \\ x_{k+1} = y_k - s\left(y_k - \mathrm{prox}_{\lambda g}(y_k - \lambda A^*(Ay_k - b))\right). \end{cases}$$

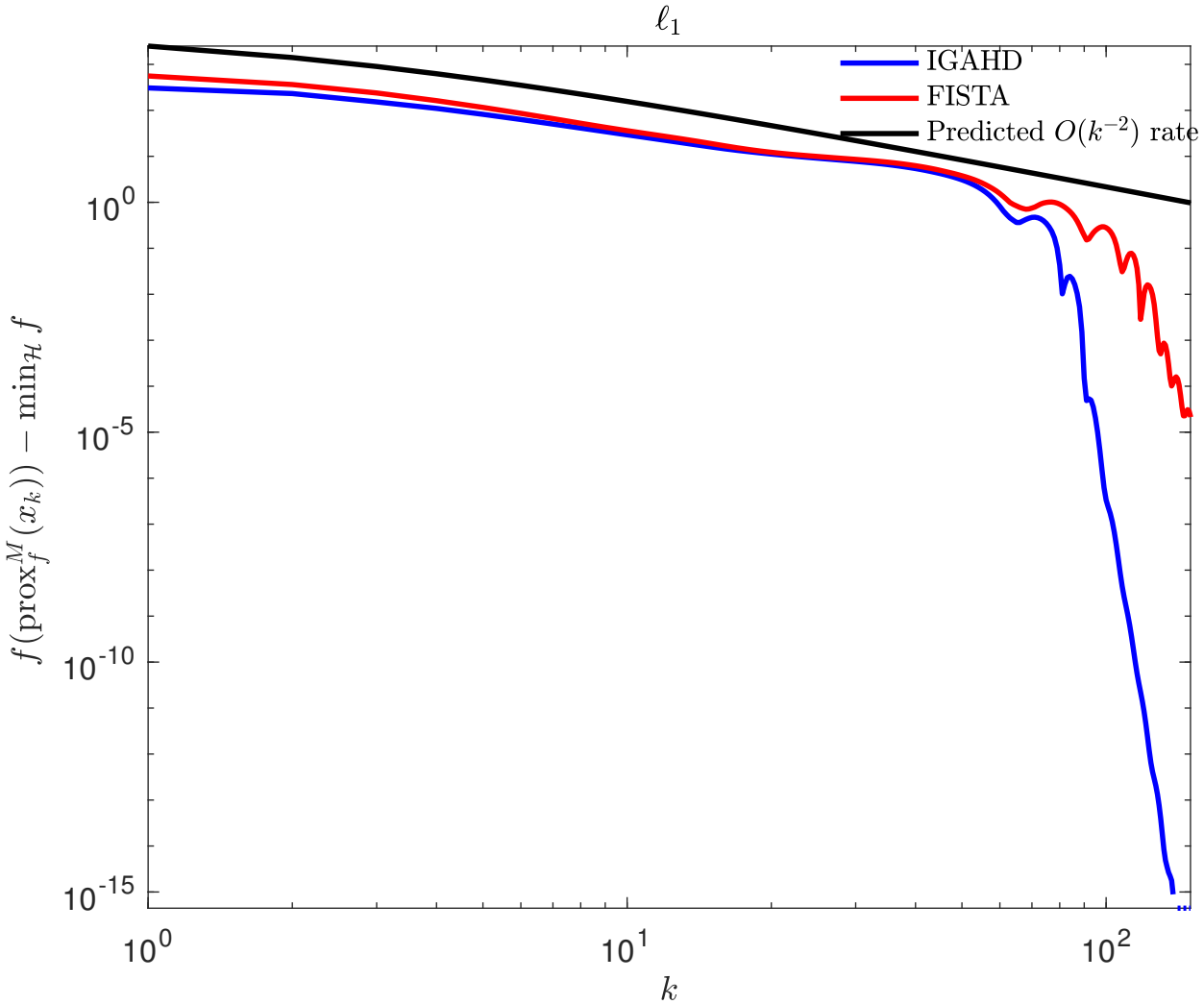
Theorem (ACFR, 2019)

Assumptions: $0 < \lambda \|A\|_2^2 < 1$, $\alpha \geq 3$, $0 \leq \beta < 2\sqrt{s}$, $s \leq 1$.
Let (x_k) *be generated by* (IGAHD) *for* (RLS). *Then,*

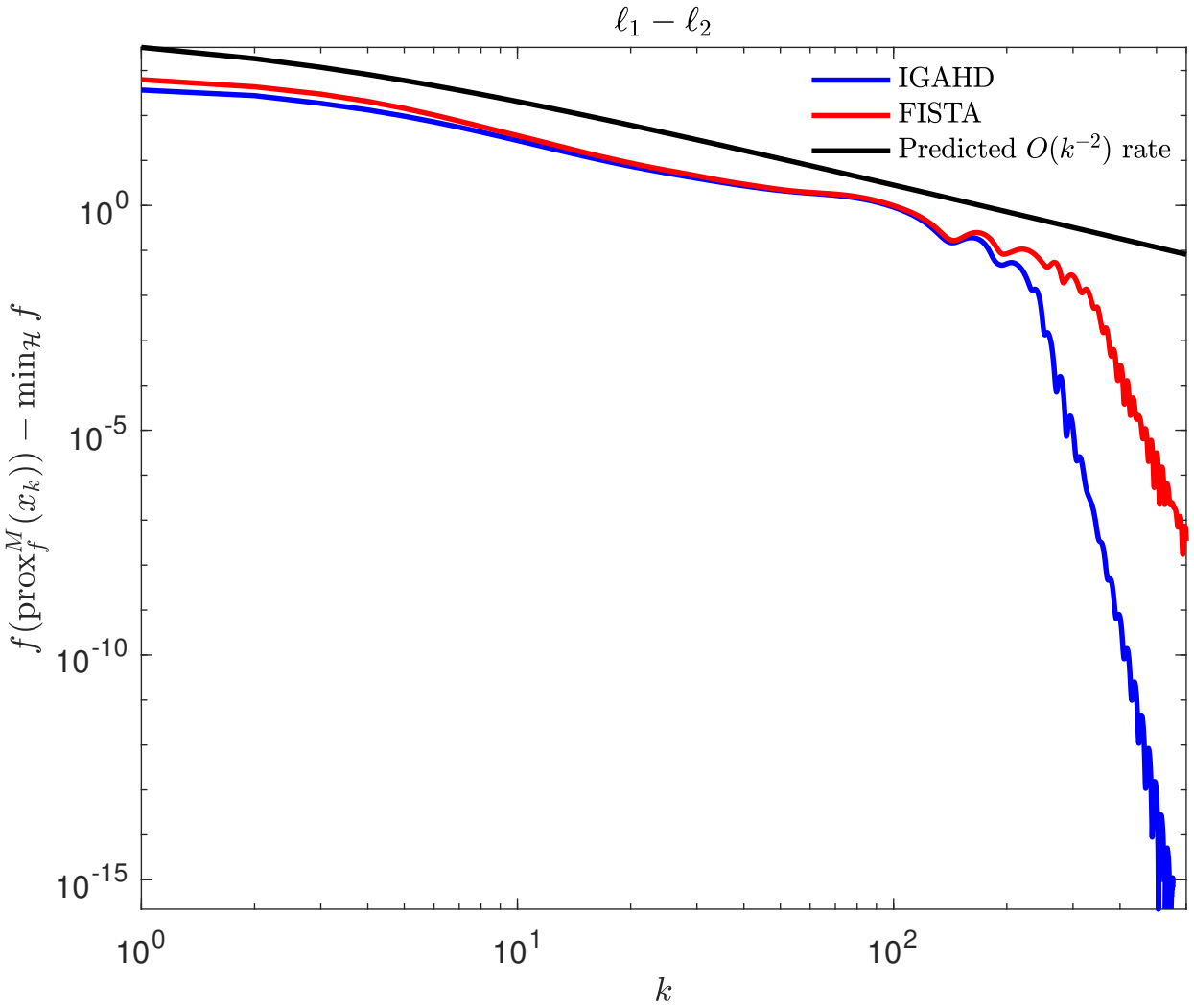
$$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathcal{H}} f = \mathcal{O}(k^{-2}), \quad \sum_k k^2 \|\nabla f(x_k)\|^2 < +\infty,$$

where $\mathrm{prox}_f^M(x_k) := \mathrm{prox}_{\lambda g}(x_k - \lambda A^*(Ax_k - b))$.

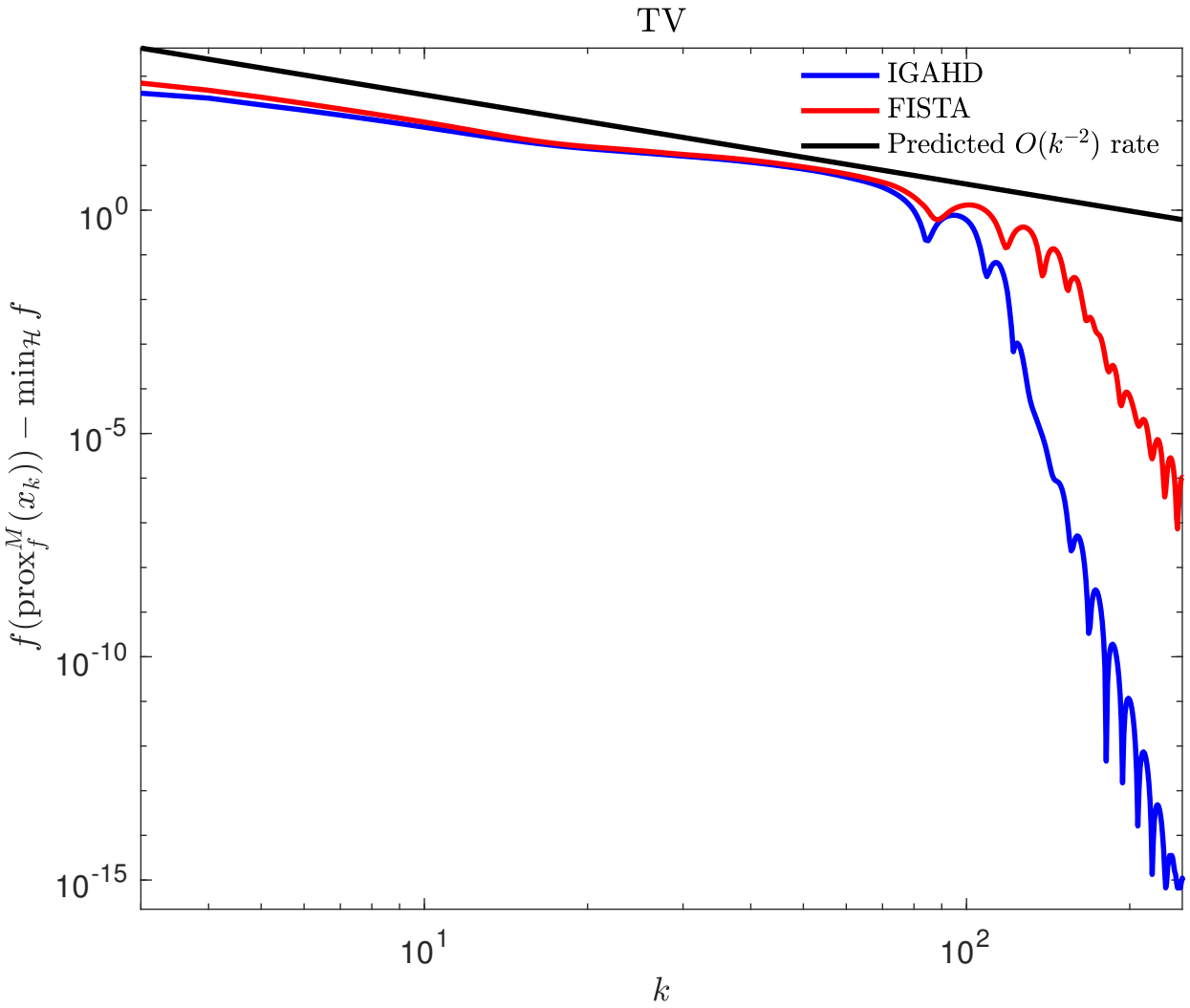
Lasso:


ℓ_1
IGAHD
FISTA
Predicted $O(k^{-2})$ rate
10^0
10^{-5}
10^{-10}
10^{-15}
$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathcal{H}} f$
10^0
10^1
10^2
k

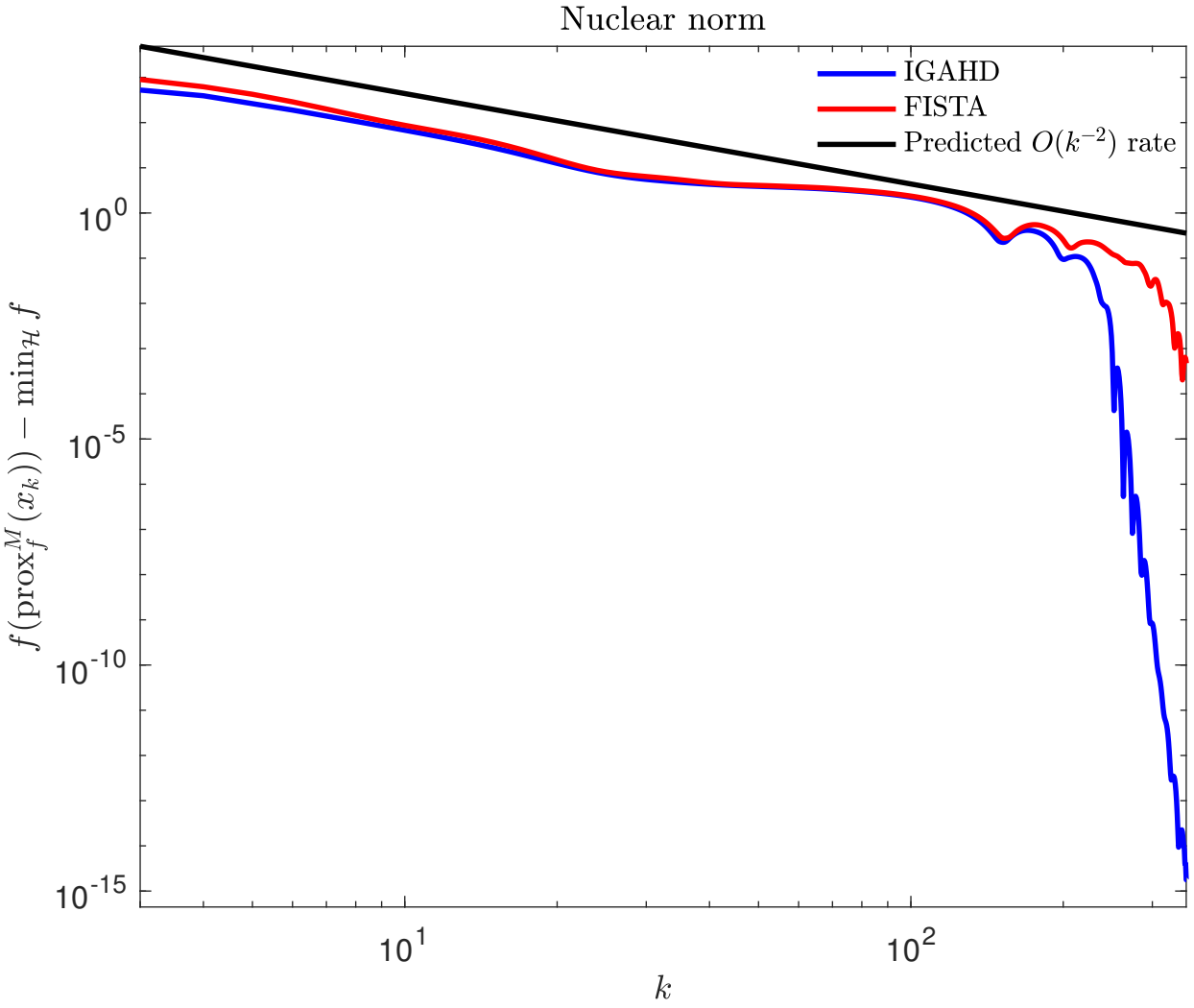
Group Lasso:


$\ell_1 - \ell_2$
IGAHD
FISTA
Predicted $O(k^{-2})$ rate
10^0
10^{-5}
10^{-10}
10^{-15}
$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathcal{H}} f$
10^0
10^1
10^2
k

TV (total variation):



Nuclear norm:



3. RAVINE METHOD.
LINK WITH NESTEROV METHOD.

Based on:

- A.-Fadili (SIOPT 2022), arXiv:2201.11643v2

Ravine method. Link with Nesterov method

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

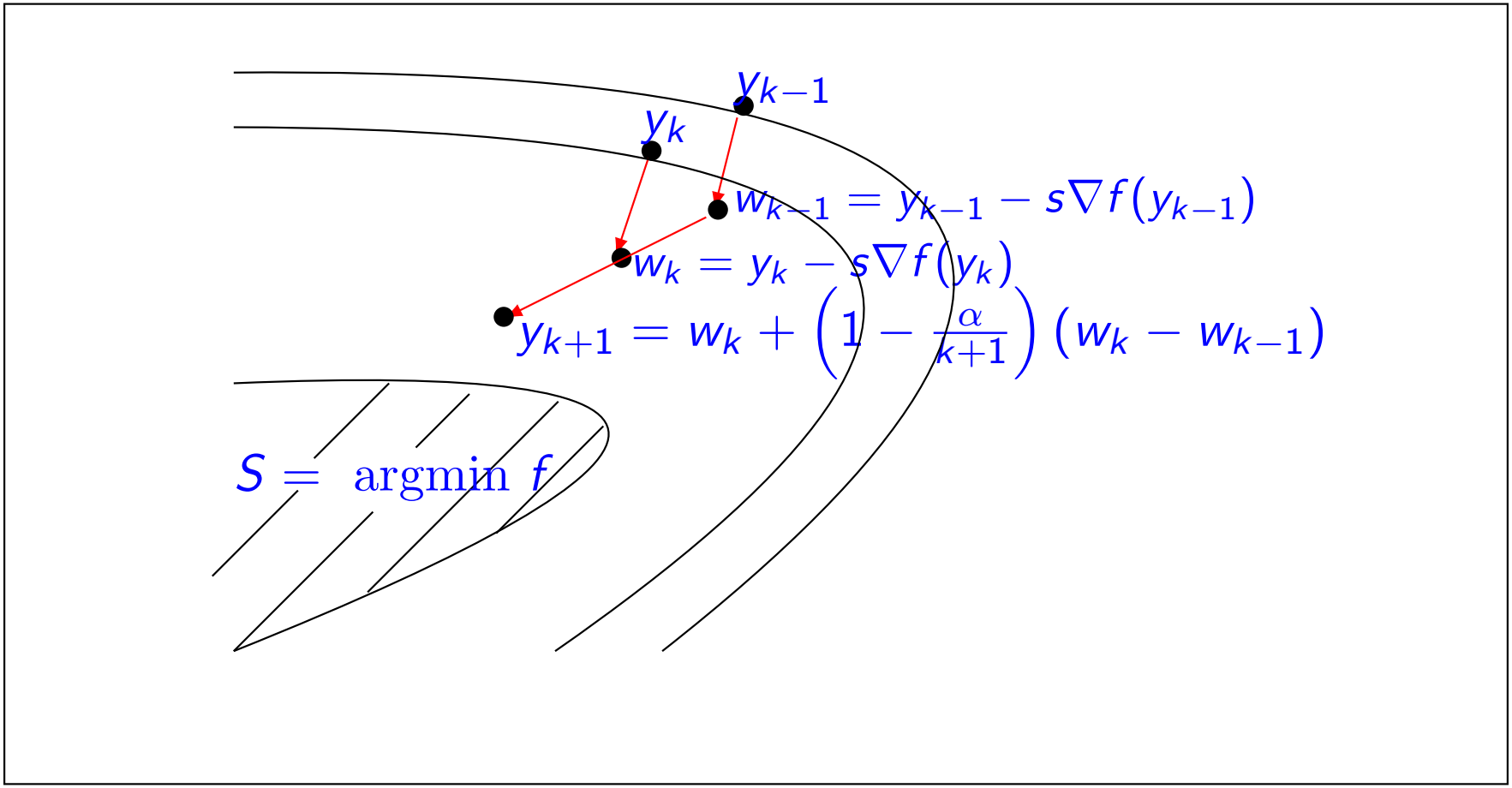
$$(\mathrm{NAG})_\alpha \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla f(y_k) \end{cases}$$

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

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

Geometric view of the Ravine method

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



Link with the Nesterov method

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

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

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

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

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

Low resolution EDO of the Ravine method

Equivalent forms of Ravine

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

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

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

Ansatz $y_k \approx Y(kh)$

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

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

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

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

High resolution EDO of the Ravine method

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

$t_k = kh$. $Y(t_k) \approx y_k$. Don't let $h \to 0$. Taylor expansion gives

Theorem (Shi-Du-Jordan-Su, Math. Prog. 2021)

The high resolution ODE with temporal step size $\sqrt{s}$ of Ravine

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

gives the inertial dynamic with Hessian driven damping

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

Convergence rates of the Ravine method

Theorem (ACFR, SDJS)

Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function such that ∇f is L-Lipschitz continuous, and $S = \operatorname{argmin} f \neq \emptyset$. Let (y_k) be a solution trajectory of the Ravine method with $\alpha > 3$, and $sL \leq 1$. Let (x_k) the associated trajectory generated by $(\mathrm{NAG})_\alpha$. Then, as $k \to +\infty$

- $f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right), \quad \sum_k k^2 \|\nabla f(y_k)\|^2 < +\infty.$
- $w - \lim y_k = w - \lim x_k = z \in S.$

- Convex inequality and $-\frac{1}{s}(x_{k+1} - y_k) = \nabla f(y_k)$ give

$$\begin{aligned} f(y_k) - \min_{\mathcal{H}} f &\leq f(x_k) - \min_{\mathcal{H}} f + \tfrac{1}{s} \langle x_{k+1} - y_k, x_k - y_k \rangle \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \tfrac{1}{s} \left(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\| \right) \|x_k - x_{k-1}\|. \end{aligned}$$

Then conclude thanks to $f(x_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$, $\|x_k - x_{k-1}\| = o\left(\frac{1}{k}\right)$.

- $\|y_k - x_k\| \leq \|x_k - x_{k-1}\| \to 0.$

Nesterov method versus Ravine method

Dual structure

- Nesterov: Extrapolation step, then Gradient step.
- Ravine: Gradient step, then Extrapolation step

Similar low and high resolution ODE

- $\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad \ddot{y}(t) + \frac{\alpha}{t}\dot{y}(t) + \nabla f(y(t)) = 0.$
- High resolution of Nesterov and Ravine $\rightarrow$ Hessian damping.

Similar asymptotic convergence rates

- $f(x_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right), \; f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right).$
- $\sum_k k^2 \|\nabla f(x_k)\|^2 < +\infty, \quad \sum_k k^2 \|\nabla f(y_k)\|^2 < +\infty.$

(IGADH) performs numerically better than Nesterov! ($\neq$ discretizations)

4. ACCELERATED DYNAMICS AND ALGORITHMS VIA TIME SCALE AND AVERAGING METHODS

Based on:

A.-Bot-Nguyen, arXiv:2208.08260

Time scaling of the steepest descent

Change of time variable $t = \tau(s)$ in the dynamic (SD)

$$\text{(SD)} \quad \dot{z}(t) + \nabla f(z(t)) = 0 \tag{4}$$

- $\tau(\cdot) : \mathbb{R}^+$ to $\mathbb{R}^+$, increasing, class $\mathcal{C}^1$, $\lim_{s \to +\infty} \tau(s) = +\infty$.
- $y(s) := z(\tau(s))$.

$$\dot{y}(s) = \dot{\tau}(s)\dot{z}(\tau(s)) \tag{5}$$

$$\dot{z}(\tau(s)) + \nabla f(z(\tau(s))) = 0. \tag{6}$$

According to (5) and (6), we obtain

$$\dot{y}(s) + \dot{\tau}(s)\nabla f(y(s)) = 0. \tag{7}$$

The convergence rate becomes

$$f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{\tau(s)}\right) \text{ as } s \to +\infty. \tag{8}$$

Averaging

Let us attach to $y(\cdot)$ the new function $x : [s_0, +\infty[\to \mathcal{H}$ defined by

$$\dot{x}(s) + \frac{1}{\dot{\tau}(s)}(x(s) - y(s)) = 0, \tag{9}$$

with $x(s_0) = x_0$ given in $\mathcal{H}$. Equivalently

$$y(s) = x(s) + \dot{\tau}(s)\dot{x}(s). \tag{10}$$

By temporal derivation of (10) we get

$$\dot{y}(s) = \dot{\tau}(s)\ddot{x}(s) + (1 + \ddot{\tau}(s))\dot{x}(s). \tag{11}$$

Replacing $\dot{y}(s)$ as given by (11) in (7) we get

$$\ddot{x}(s) + \frac{1 + \ddot{\tau}(s)}{\dot{\tau}(s)}\dot{x}(s) + \nabla f\Big(x(s) + \dot{\tau}(s)\dot{x}(s)\Big) = 0. \tag{12}$$

Link with Nesterov method

According to the Su-Boyd-Candès model for Nesterov method, take

$$\frac{1+\ddot{\tau}(s)}{\dot{\tau}(s)} = \frac{\alpha}{s},$$

which gives $\tau(s) = \frac{s^2}{2(\alpha-1)}$. Equations (12) become

Inertial System with Implicit Hessian Damping

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \frac{s}{\alpha-1}\dot{x}(s)\right) = 0. \tag{13}$$

and we have the convergence rate, as $s \to +\infty$

$$f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right).$$

We will show that this convergence rate is inherited by $x(\cdot)$.

Fast convergence rate

Theorem (ABN, Arxiv 2022)

Let $f : \mathcal{H} \to \mathbb{R}$ be a convex $\mathcal{C}^1$ function, whose gradient is Lipschitz continuous on the bounded sets, and such that $S = \operatorname{argmin} f \neq \emptyset$. Suppose that $\alpha > 3$. Let $x : [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \frac{s}{\alpha - 1}\dot{x}(s)\right) = 0.$$

Then the following convergence properties are satisfied:

1. $f(x(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right)$
2. $\|\nabla f(x(s))\| \leq \frac{C}{s}; \quad \int_{s_0}^{+\infty} s^3 \left\|\nabla f\left(x(s) + \frac{s}{\alpha - 1}\dot{x}(s)\right)\right\|^2 ds < +\infty;$
3. $x(s)$ *converges weakly as* $s \to +\infty$, *and its limit belongs to* S.

From the convergence properties of the steepest descent,

$$f\left(z\left(t\right)\right)-\inf_{\mathcal{H}} f=\frac{\varepsilon_0(t)}{t}$$

for a positive function $\varepsilon_0(\cdot)$ such that $\lim_{t\to+\infty}\varepsilon_0(t)=0$ as $t\to+\infty$. From the definition of $y(s):=z(\tau(s))$ we get

$$f\left(y\left(s\right)\right)-\inf_{\mathcal{H}} f=\frac{\varepsilon(s)}{\tau(s)}, \tag{14}$$

with $\varepsilon(s)=\varepsilon_0(\tau(s))$. According to $\tau(s)=\frac{s^2}{2(\alpha-1)}$, we get

$$f\left(y\left(s\right)\right)-\inf_{\mathcal{H}} f=o\left(\frac{1}{s^2}\right). \tag{15}$$

Let us show that (15) is inherited by $x(\cdot)$. This will result from interpreting going from y to x as an averaging process.

For simplicity, suppose that $x(s_0) = y(s_0)$. By definition (9) of x

$$s\dot{x}(s) + (\alpha - 1)x(s) = (\alpha - 1)y(s). \tag{16}$$

After multiplication of (16) by $s^{\alpha-2}$, and integrating we get

$$x(s) \quad = \quad \frac{s_0^{\alpha-1}}{s^{\alpha-1}} y(s_0) + \frac{\alpha - 1}{s^{\alpha-1}} \int_{s_0}^{s} u^{\alpha-2} y(u) du. \tag{17}$$

Averaging process

$$x(s) = \int_{s_0}^{s} y(u)\, d\mu_s(u), \tag{18}$$

where μ_s is the probability measure on $[s_0, s]$

$$\mu_s := \frac{s_0^{\alpha-1}}{s^{\alpha-1}} \delta_{s_0} + (\alpha - 1) \frac{u^{\alpha-2}}{s^{\alpha-1}} du.$$

Jensen's inequality, f convex

$$\begin{aligned} f(x(s)) - \inf_{\mathcal{H}} f &= (f - \inf_{\mathcal{H}} f)\left(\int_{s_0}^{s} y(u) d\mu_s(u)\right) \\ &\leq \int_{s_0}^{s} \left(f(y(u)) - \inf_{\mathcal{H}} f\right) d\mu_s(u) \\ &\leq \int_{s_0}^{s} \frac{\varepsilon(u)}{\tau(u)} d\mu_s(u). \end{aligned}$$

Let us explicit the last above integral. We get, for all $s \geq s_0$

$$\begin{aligned} f(x(s)) - \inf_{\mathcal{H}} f &\leq 2(\alpha - 1) \int_{s_0}^{s} \frac{\varepsilon(u)}{u^2} d\mu_s(u) \\ &\leq 2(\alpha - 1) s_0^{\alpha-3} \varepsilon(s_0) \frac{1}{s^{\alpha-1}} + \frac{2(\alpha-1)^2}{s^{\alpha-1}} \int_{s_0}^{s} \varepsilon(u) u^{\alpha-4} du. \end{aligned}$$

Equivalently

$$s^2(f(x(s)) - \inf_{\mathcal{H}} f) \leq 2(\alpha-1)s_0^{\alpha-3}\varepsilon(s_0)\frac{1}{s^{\alpha-3}} + \frac{2(\alpha-1)^2}{s^{\alpha-3}}\int_{s_0}^{s}\varepsilon(u)u^{\alpha-4}du.$$

Therefore, for $\alpha > 3$

$$\limsup_{s\to+\infty} s^2(f(x(s)) - \inf_{\mathcal{H}} f) \quad \leq \quad \limsup_{s\to+\infty} \frac{2(\alpha-1)^2}{s^{\alpha-3}}\int_{s_0}^{s}\varepsilon(u)u^{\alpha-4}du.$$

Then conclude with the help of the following lemma.

Lemma (convergence implies ergodic convergence)

Let $a : [s_0, +\infty[\to \mathbb{R}$ be a positive real valued function which verifies $\lim_{u\to+\infty} a(u) = 0$. Take $\alpha > 1$. Then $\lim_{s\to+\infty} A(s) = 0$, where

$$A(s) = \frac{1}{s^{\alpha-1}}\int_{s_0}^{s} a(u)u^{\alpha-2}du.$$

5. FROM GRADIENT FLOWS TO DOUBLY NONLINEAR EVOLUTION SYSTEMS

Based on:

Adly-A., September 2022

A doubly nonlinear evolution system with dry friction

Our approach is built on the doubly nonlinear evolution equation

$$\text{(GSDF)} \qquad \dot{z}(t) + \partial\phi(\dot{z}(t)) + \nabla f(z(t)) \ni 0 \tag{19}$$

Model case: $\phi =$ dry friction potential (simplified version of Coulomb)

$$\phi(\xi) = r\|\xi\|.$$

We have $\quad \partial\phi(\xi) = r\frac{\xi}{\|\xi\|}$ if $\xi \neq 0, \quad \partial\phi(0) = \mathbb{B}(0, r).$

Equivalently

$$\dot{z}(t) - (I + \partial\phi)^{-1}(-\nabla f(z(t)) = 0$$

which gives, with $f =$ general differentiable function

$$\dot{z}(t) + \nabla f(z(t)) - \mathrm{proj}_{\mathbb{B}(0,r)}(\nabla f(z(t)) = 0. \tag{20}$$

If $r = 0$ (no dry friction) $\longrightarrow$ Gradient flow.

A doubly nonlinear evolution system with dry friction

Theorem

Let $z : [t_0, +\infty[\to \mathcal{H}$ *be a global solution trajectory of* (GSDF). *Then*

i) $\displaystyle\int_{t_0}^{+\infty} \|\dot{z}(t)\|^2 dt < +\infty$;

ii) $\displaystyle\int_{t_0}^{+\infty} \|\dot{z}(t)\| dt < +\infty$;

iii) $z(\cdot)$ *converges strongly as* $t \to +\infty$ *to* z_∞ *with* $\|\nabla f(z_\infty)\| \leq r$.

iv) If $\|\nabla f(z_\infty)\| < r$, *then the trajectory stops at* z_∞ *after a finite time.*

Suppose moreover that f *is a convex function. Then*

v) $t \mapsto \|\dot{z}(t)\|$ *and* $t \mapsto \operatorname{dist}(\nabla f(z(t)), \mathbb{B}(0, r))$ *are decreasing.*

vi) $\|\dot{z}(t)\| = \operatorname{dist}(\nabla f(z(t)), \mathbb{B}(0, r)) = o\left(\dfrac{1}{t}\right)$ *as* $t \to +\infty$.

Dual view, Riemannian structure, Bregman distance

Dual variable: $g(t) := \nabla f(z(t))$.

Riemannian structure associated with the Hessian of f^*

According to the Fenchel conjugaison formula, we get

$$\frac{d}{dt}\left(\partial f^*(g(t))\right) + g(t) - \mathrm{proj}_{\mathbb{B}(0,r)}(g(t)) \ni 0. \tag{21}$$

When f is smooth,

$$\nabla^2 f^*(g(t))\dot{g}(t) + g(t) - \mathrm{proj}_{\mathbb{B}(0,r)}(g(t)) = 0.$$

Bregman distance

$$D(g(t), g_\infty) = f^*(g_\infty) - f^*(g(t)) - \langle \nabla f^*(g(t)), g_\infty - g(t) \rangle .$$

$t \mapsto D(g(t), g_\infty)$ is decreasing.

Time scaling and averaging of the dual of (GSDF)

Theorem

Time scale and averaging of the dual formulation of (GSDF)

$$\frac{d}{dt}\left(\partial f^*(g(t))\right) + g(t) - \mathrm{proj}_{\mathbb{B}(0,r)}(g(t)) \ni 0.$$

Set $t = \tau(s) = \frac{s^2}{2(\alpha-1)}$, $v(s) = g(\tau(s))$, $\dot{w}(s) + \frac{1}{\dot{\tau}(s)}(w(s) - v(s)) = 0$.

The dynamic becomes

$$\nabla^2 f^*\Big(w(s) + \frac{s}{\alpha-1}\dot{w}(s)\Big)\Big(\ddot{w}(s) + \frac{\alpha}{s}\dot{w}(s)\Big)$$
$$+\Big(w(s) + \frac{s}{\alpha-1}\dot{w}(s)\Big) - \mathrm{proj}_{\mathbb{B}(0,r)}\Big(w(s) + \frac{s}{\alpha-1}\dot{w}(s)\Big) = 0.$$

Convergence rate properties of this dynamic: as $s \to +\infty$

$$\mathrm{dist}(\nabla f(w(s)), \mathbb{B}(0,r)) = o\left(\frac{1}{s^2}\right).$$

6. FAST OPTIMIZATION VIA VANISHING TIKHONOV REGULARIZATION

Based on:

A.-Bahlag-Chbani-Riahi, JDE 2022, arXiv:2203.05457

Vanishing Tikhonov regularization

Tikhonov's regularization: $f : \mathcal{H} \to \mathbb{R}$ convex function, $\varepsilon > 0$

Then, $x \mapsto f(x) + \frac{\varepsilon}{2}\|x\|^2$ is ϵ-strongly convex.

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

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

Then, $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(e^{-\sqrt{\mu}t}\right) \quad \text{as } t \to +\infty.$

Tikhonov's dynamical regularization: $f : \mathcal{H} \to \mathbb{R}$ convex function

Replace f by φ_t with $\varphi_t(x) := f(x) + \frac{\varepsilon(t)}{2}\|x\|^2$, $\varepsilon(t) \to 0$ as $t \to +\infty$.
Since φ_t is $\varepsilon(t)$-strongly convex, this gives the nonautonomous dynamic

$$\text{(TRIGS)} \qquad \ddot{x}(t) + \delta\sqrt{\varepsilon(t)}\dot{x}(t) + \nabla f(x(t)) + \varepsilon(t)x(t) = 0.$$

Fast optimization via vanishing Tikhonov regularization

Take $\varepsilon(t) = \frac{1}{t^p}$.

Theorem (A.-Bahlag-Chbani-Riahi, JDE 2022)

Take $0 < p < 2, \quad \delta > 0$. *Let* $x : [t_0, +\infty[\to \mathcal{H}$ *be a solution trajectory of*

$$\ddot{x}(t) + \frac{\delta}{t^{\frac{p}{2}}}\dot{x}(t) + \nabla f\left(x(t)\right) + \frac{1}{t^p}x(t) = 0.$$

Then, we have the following convergence rates: as $t \to +\infty$

$$f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^p}\right)$$

and $x(t)$ *converges strongly to the minimum norm solution.*

$\longrightarrow$ Fast convergence via vanishing damping coefficient.

7. AN SDE PERSPECTIVE ON STOCHASTIC CONVEX OPTIMIZATION

Based on:

A.-Fadili-Maulen, arXiv:2207.02750v1 2022

Stochastic continuous gradient method

$$\text{(SDE)} \begin{cases} dX(t) = -\nabla f(X(t))dt + \sigma(t, X(t))dW(t), & t \geq 0 \\ X(0) = X_0. \end{cases}$$

- $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$: filtered probability space
- $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$: diffusion (volatility) matrix
- W: m-dimensional Brownian motion.

Theorem (A.-Fadili-Maulen, arXiv 2022)

Suppose f convex, $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, $\sigma_\infty(t) := \sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F$, then:

1. $\sup_{t\geq 0} \mathbb{E}[\|X(t)\|^2] < +\infty$.
2. $\forall x^\star \in S$, $\lim_{t\to\infty} \|X(t) - x^\star\|$ exists a.s., $\sup_{t\geq 0} \|X(t)\| < +\infty$ a.s.
3. $\lim_{t\to\infty} \|\nabla f(X(t))\| = 0$ a.s. Hence, $\lim_{t\to\infty} f(X(t)) = \min f$ a.s.
4. There exists an S-valued random variable $x^\star$ such that $\lim_{t\to\infty} X(t) = x^\star$ a.s.

Some open questions concerning Nesterov algorithm

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

- Convergence of the iterates in the critical case $\alpha = 3$?
- Optimal tuning of the parameter $\alpha > 3$.
- Is is possible to obtain $1/k^2$ rate of convergence with autonomous dynamic/algorithms? Damping as a closed loop control (velocity, value...). First results: Lin-Jordan (Arxiv 2019), A.-Bot-Cestnek (JEMS, 2021), Aujol-Dossal-Labarrière-Rondepierre (2021).
- Nesterov versus Tikhonov regularization of Polyak heavy ball.

Some open questions concerning Hessian damping

- Link continuous/discrete dynamics: (IGAHD) and Nesterov have the same high resolution ODE with Hessian driven damping. Still, (IGAHD) performs numerically better than Nesterov and FISTA?
- Proximal-gradient methods associated with (RAVINE), (IGAHD).
- Best tuning of the damping coefficients α, β.
- Compare explicit and implicit Hessian driven damping.
- Accelerated (ADMM) via time scaling and averaging method.
- Restarting method via asymptotic Tikhonov regularization and Hessian driven damping.
- Extension to monotone inclusions. (A.-Laszlo)

THANK YOU FOR YOUR ATTENTION

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, *First-order inertial algorithms involving dry friction damping*, Math. Program, (2021) https://doi.org/10.1007/s10107-020-01613-y
- 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, 84 (2021), pp. 1687–1716.

References

- 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, A. Balhag, Z. Chbani, H. Riahi, *Damped inertial dynamics with vanishing Tikhonov regularization: Strong asymptotic convergence towards the minimum norm solution*, Journal of Differential Equations, 311 (2022), pp. 29–58.

References

- 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.
- H. Attouch, R.I. Boţ, D.-K. Nguyen, *Fast convex optimization via time scale and averaging of the steepest descent* arXiv:2208.08260v1 [math.OC] Aug 2022.
- H. Attouch, H., J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic* features, Math. Program., Ser. B, **116** (2009), pp. 5–16.
- H. Attouch, J. Bolte, P. Redont, A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Lojasiewicz inequality*, Mathematics of Operations Research, **35**(2), (2010), pp. 438–457.

References

- H. Attouch, J. Bolte, B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., **137**(1) (2013), pp. 91–129.
- 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*, Math. Program., 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.

References

- H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *First order optimization algorithms via inertial systems with Hessian driven damping*, Math. Program., 193 (2022), pp. 113–155.
- H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics*, JOTA, 193(1-3) 2022, pp. 704–736.
- H. Attouch, Z. Chbani, J. Peypouquet, P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program. 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.

References

- H. Attouch, J. Fadili, *From the Ravine Method to the Nesterov Method and Vice Versa: A Dynamical System Perspective*, SIAM J. Optim. 32(3) (2022), 10.1137/22M1474357
- H. Attouch, J. Fadili, V. Kungurtsev, *First order optimization via inertial systems with Hessian driven damping subject to perturbation errors*, Evolution Equations and Control Theory, 2022, doi: 10.3934/eect.2022022
- H. Attouch, X. Goudou, P. Redont, *The heavy ball with friction method. The continuous dynamical system...*, Commun. Contemp. Math., 2(1) (2000), pp. 1–34.
- H. Attouch, S. C. László, *Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators*, SIAM J. Optim., 30(4) (2020), 10.1137/20M1333316.
- H. Attouch, S. C. László, *Continuous Newton-like Inertial Dynamics for Monotone Inclusions*, Set Valued and Variational

References

- H. Attouch, J. Peypouquet, *Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators*, Mathematical Programming, 174 (1-2) (2019), pp. 391–432.
- H. Attouch, J. Peypouquet, P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM J. Optim., 24(1) (2014), pp. 232–256.
- H. Attouch, J. Peypouquet, P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261(10), (2016), pp. 5734–5783.
- H. Attouch, B. F. Svaiter, *A continuous dynamical Newton-Like approach to solving monotone inclusions*, SIAM J. Control Optim., 49 (2) (2011), pp. 574–598.

References

- J.-F. Aujol, Ch. Dossal, H. Labarrière, A. Rondepierre, *FISTA restart using an automatic estimation of the growth parameter*, 2021, HAL 03153525 .
- 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.
- P. Bégout, J. Bolte, M. A. Jendoubi, *On damped second-order gradient systems*, Journal of Differential Equations, vol. 259, n° 7-8, 2015, pp. 3115–3143.

References

- 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*, 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*, Journal of Machine Learning Research 22 (2021), pp. 1–31.
- 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.

Bibliography

- 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.
- S. Łojasiewicz, *Sur la géométrie semi- et sous-analytique*, Ann. Inst. Fourier **43**, (1993), 1575-1595.
- P.-E. Maingé, F. Labarre, *Accelerated methods with fastly vanishing subgradients for structured non-smooth minimization*, Numerical Algorithms, 2022.
- R. Maulen, J. Jalal, H. Attouch, *An SDE perspective on on stochastic convex optimization*, arXiv:2207.02750v1 [math.OC] 6 Jul 2022
- M. Muehlebach, M. I. Jordan, *A Dynamical Systems Perspective on Nesterov Acceleration*, (2019), arXiv:1905.07436

References

- 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).