A Dynamical System Perspective of Optimization in Data Science

Jalal Fadili
Normandie Université-ENSICAEN, GREYC CNRS UMR 6072

CUHK MATH-IMS colloquium
11 March 2022

Join work with H. Attouch, S. Kungurtsev, Z. Chbani, H. Riahi

Normandie Université

cnrs

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

Some notations

- $\mathcal{H}$: a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$.
- $\Gamma(\mathcal{H})$ the class of functions $f : \mathcal{H} \to]-\infty, +\infty]$ that are :
  - proper : $\mathrm{dom}(f) \neq \emptyset$,
  - lower semicontinuous (closed epigraph).
- $\Gamma_0(\mathcal{H}) = \Gamma(\mathcal{H}) \cap \{\text{convex}\}$.
- $\nabla f : \mathcal{H} \to \mathcal{H}$ the (Fréchet) gradient of a differentiable function.
- $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$: the (Fenchel) subdifferential of $f \in \Gamma_0(\mathcal{H})$, i.e.

$$\partial f(x) \stackrel{\mathrm{def}}{=} \{u \in \mathcal{H} : \ f(y) \geq f(x) + \langle u, y - x \rangle, \ \forall y \in \mathcal{H}\} .$$

- $C^1_L(\mathcal{H})$ is the class of $C^1(\mathcal{H})$ functions with L-Lipschitz gradient.

Motivations

- Design efficient and provably fast algorithms to solve

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

  - f and $g \in \Gamma(\mathcal{H})$;
  - f smooth enough;
  - $\operatorname*{Argmin}_{\mathcal{H}} (f + g) \neq \emptyset$.
- In data science : $f(x) = \mathbb{E}_w \left[\ell(x, w)\right]$, g a regularizer.
- Problem can be convex or not, smooth or non-smooth.
- First order criticality condition

$$x^\star \in \mathrm{Crit}(f + g) \iff 0 \in \partial_G (f + g)(x^\star).$$

- ∂_G is some generalized (set)-valued subdifferential operator (e.g., Clarke, limiting), or even another set-valued field enjoying favorable calculus rules.

Motivations

$$\min_{x\in\mathcal{H}} f(x)+g(x)$$

$$x^\star \in \mathrm{Crit}(f+g) \iff 0\in \partial_G(f+g)(x^\star)$$

- **Key idea** : exploit relation between optimization and equilibria of *dissipative dynamical systems* of the form

$$\begin{cases} H\left(\left\{\frac{d^i}{dt^i}x(t)\right\}_{i\in[n]}\right)+\partial_G(f+g)(x(t))\ni 0, \quad t>0 \\ x(0)=x_0. \end{cases}$$

- Typically in this talk : $n\in\{1,2\}$.
- Algorithms developed as appropriate *discretization* of a dynamical system.
- The dynamical system perspective offers :
  - a powerful way to understand the geometry underlying the dynamics ;
  - a versatile framework to obtain fast, scalable and new algorithms ;
  - equip these algorithms with provable convergence guarantees (including fast rates) though proper Lyapunov analysis.

For simplicity: $g{=}0$ in the sequel

Gradient descent

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

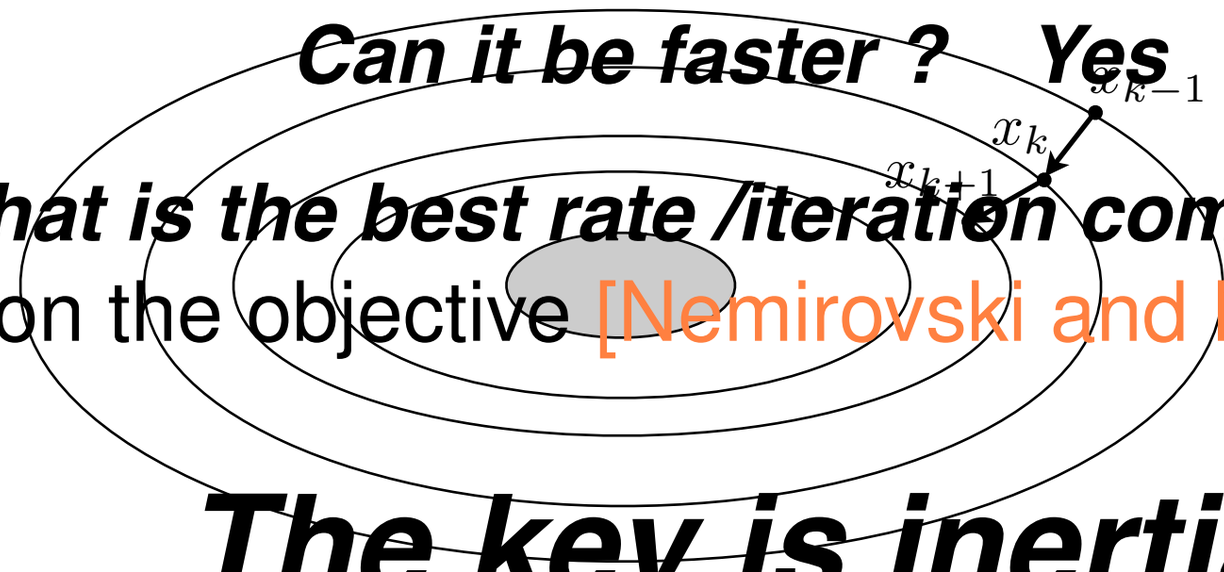
- Gradient descent dynamic ([Cauchy 1847]) : $t \in [0, +\infty[$

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

- Temporal discretization $\longrightarrow$ $x_{k+1} = x_k - \mu_k \nabla f(x_k)$, $0 < \inf_k \mu_k \leq \sup_k \mu_k < 2/L$.

- $(x_k)_k$ weakly converges to a minimizer and $f(x_k)$ converges to $\min_{\mathcal{H}} f$ at the rate $O(1/k)$ (in fact even $o(1/k)$).

Can it be faster ? Yes



What is the best rate /iteration complexity ?

$O(1/k^2)$ on the objective [Nemirovski and Nesterov 1986]

The key is inertia

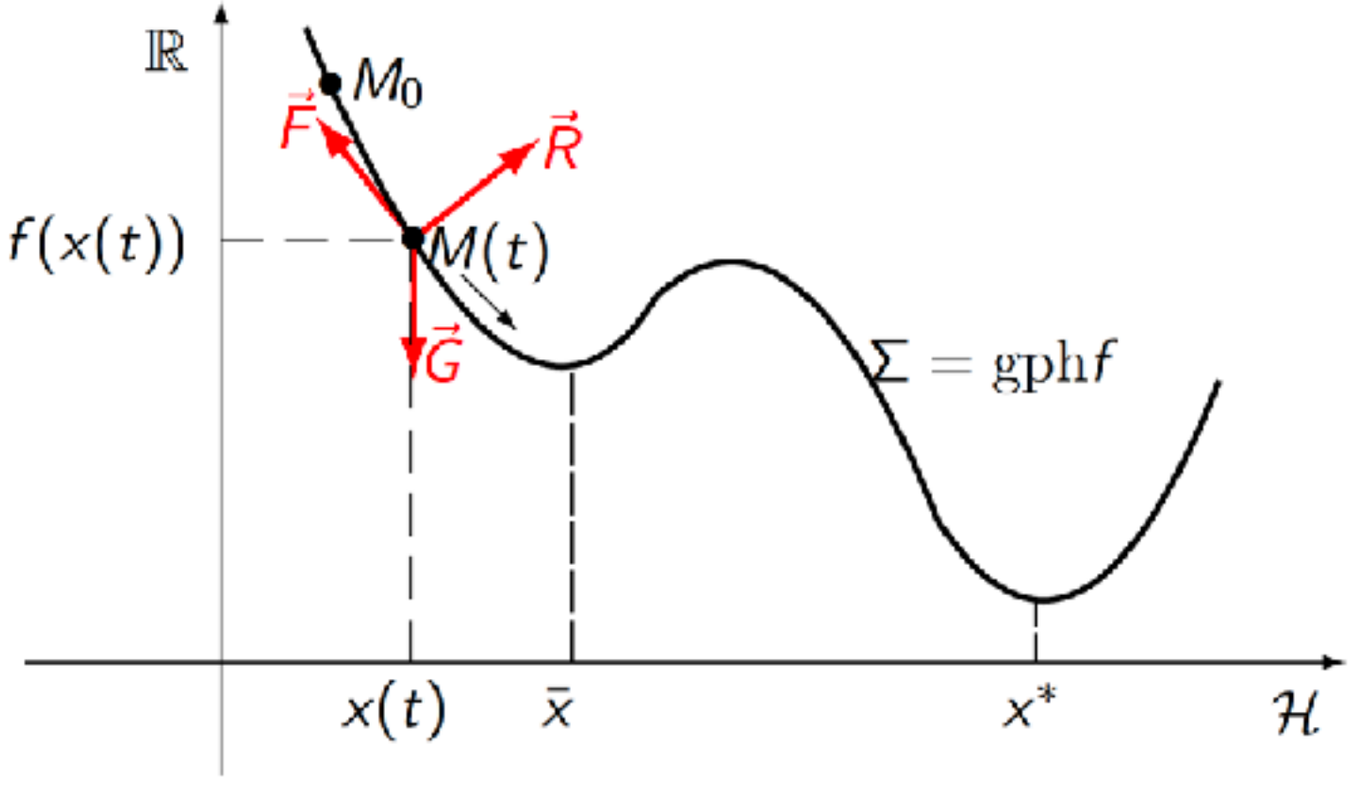
Inertial dynamic: fixed damping

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C_L^1(\mathcal{H})$$

- Fixed viscous damping coefficient $\gamma > 0$, [Polyak, 1964, 1987] $t \in [t_0, +\infty[$, $t_0 > 0$:

$$(\mathrm{HBF}) \qquad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0,$$

- Heavy ball method with friction.



Mechanical interpretation : $\overrightarrow{F}$: friction. $\overrightarrow{R}$: reaction. $\overrightarrow{G}$: gravity.

Inertial dynamic: fixed damping

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

- Fixed viscous damping coefficient $\gamma > 0$, [Polyak, 1964, 1987] $t \in [t_0, +\infty[$, $t_0 > 0$:

$$(\mathrm{HBF}) \qquad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0,$$

- f convex [Alvarez et al. 2000] :
  - $f(x(t)) - \min_{\mathcal{H}} f = O(1/t)$.
  - $x(t) \rightharpoonup x^\star \in \operatorname*{Argmin}_{\mathcal{H}} (f)$.

- f μ-strongly convex, $\gamma = 2\sqrt{\mu}$ [Polyak 1987] :
  - $\frac{\mu}{2} \left\|x(t) - x^\star\right\|^2 \leq f(x(t)) - \min_{\mathcal{H}} f = O(e^{-\sqrt{\mu} t})$.
  - Geometry of $f \leftrightarrow$ Damping coefficient $\leftrightarrow$ Linear rate.

Inertial dynamic: (AVD)

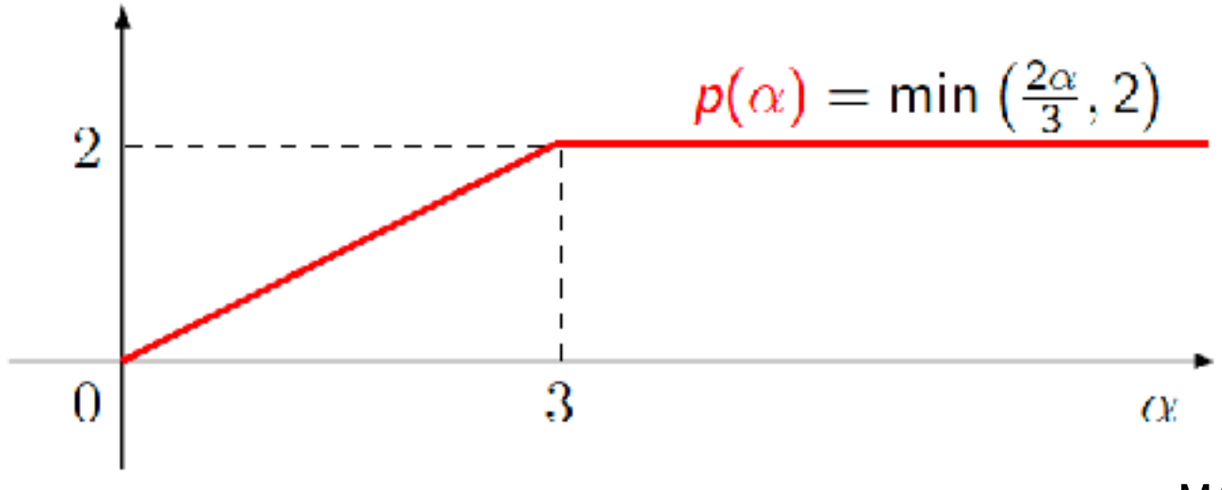
$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

- Inertial dynamic with asymptotically vanishing viscous damping, **(AVD)** $t \in [t_0, +\infty[, t_0 > 0$:

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

[Su et al. 2014, Attouch et al. 2018, Chambolle and Dossal 2015].

- $\alpha > 3 : f(x(t)) - \min_{\mathcal{H}} f = o(1/t^2)$, $x(t) \rightharpoonup x^\star \in \underset{\mathcal{H}}{\mathrm{Argmin}}\,(f)$.
- For $\alpha = 3$, $f(x(t)) - \min_{\mathcal{H}} f = O(1/t^2)$ but convergence of $x(t)$ remains an open problem except in 1D ([Attouch et al. 2019].
- Choosing $\alpha < 3$ leads to a sub-optimal rate $O(1/t^{2\alpha/3})$ [Apidopoulos et al. 2018, Attouch 2017].



From (AVD) to Nesterov algorithm

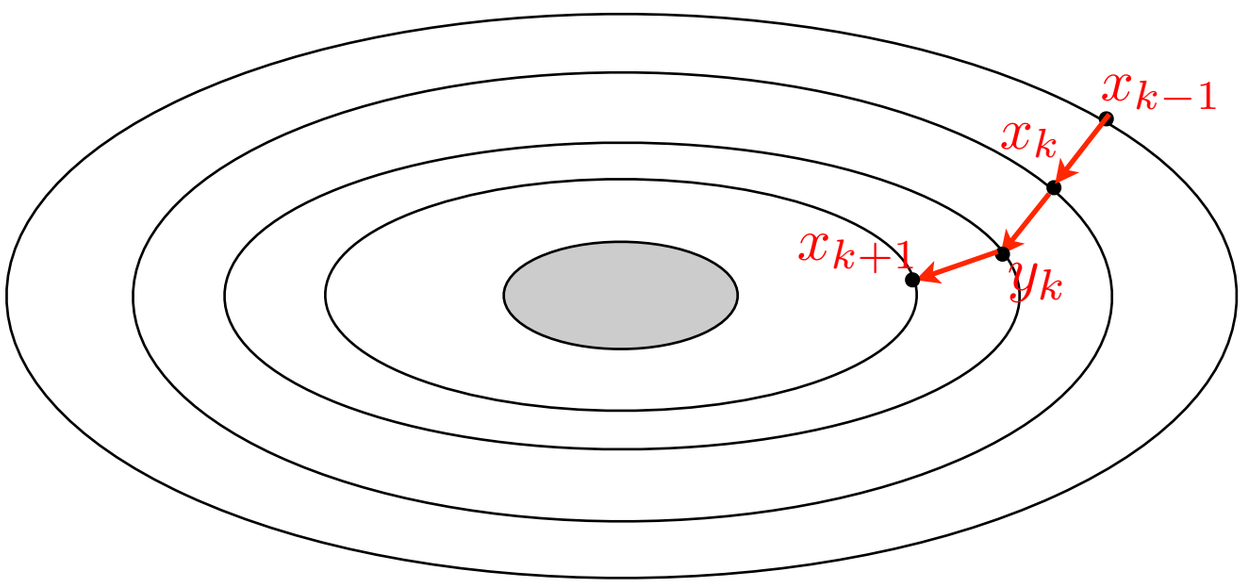
$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

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

- Appropriate temporal discretization $\longrightarrow$ celebrated Nesterov-type accelerated gradient algorithm [Nesterov 1983, Nesterov 2015]

$$(\mathrm{NAG})_\alpha \begin{cases} y_k = x_k + (1-\frac{\alpha}{k})(x_k - x_{k-1}) \\ x_{k+1} = y_k - s\nabla f(y_k), \qquad\qquad s \in]0, 1/L]. \end{cases}$$

- $\alpha > 3 : f(x_k) - \min_{\mathcal{H}} f = o(1/k^2)$, $x_k \rightharpoonup x^\star \in \underset{\mathcal{H}}{\mathrm{Argmin}}\,(f)$.
- For $\alpha = 3$, $f(x_k) - \min_{\mathcal{H}} f = O(1/k^2)$ but convergence of x_k remains open.



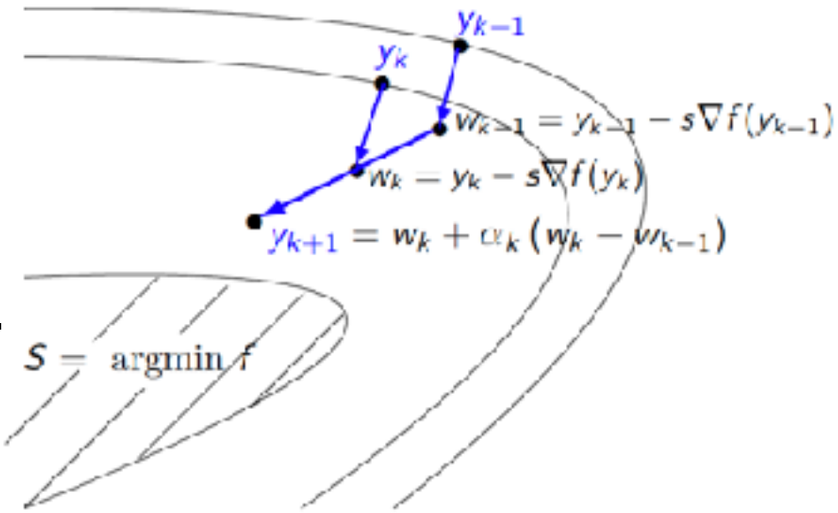
From (AVD) to Ravine algorithm

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

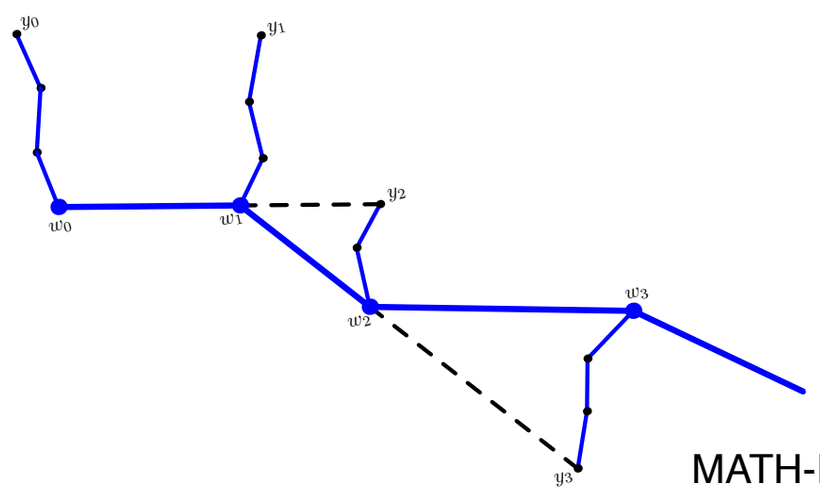
$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad \alpha > 0,$$

- Appropriate temporal discretization $\longrightarrow$ Ravine accelerated gradient algorithm
  [Gelfand and Tsetlin 1961]

$$(\mathrm{RAG})_\alpha \begin{cases} w_k = y_k - s\nabla f(y_k), \\ y_{k+1} = w_k + (1 - \frac{\alpha}{k+1})(w_k - w_{k-1}) \end{cases} \qquad s \in]0, 1/L],$$



- Has long been ignored.
- At the forefront of current research [Attouch and Fadili 2022].
- It mimics the flow of water in the mountains :
  - It first flows rapidly downhill through small, steep ravines.
  - Then it flows along the main river in the valley.



From (AVD) to Ravine algorithm

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

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

- Appropriate temporal discretization $\longrightarrow$ Ravine accelerated gradient algorithm [Gelfand and Tsetlin 1961]

$$(\mathrm{RAG})_\alpha \begin{cases} w_k = y_k - s\nabla f(y_k), & s \in]0, 1/L], \\ y_{k+1} = w_k + (1 - \frac{\alpha}{k+1})(w_k - w_{k-1}) \end{cases}$$

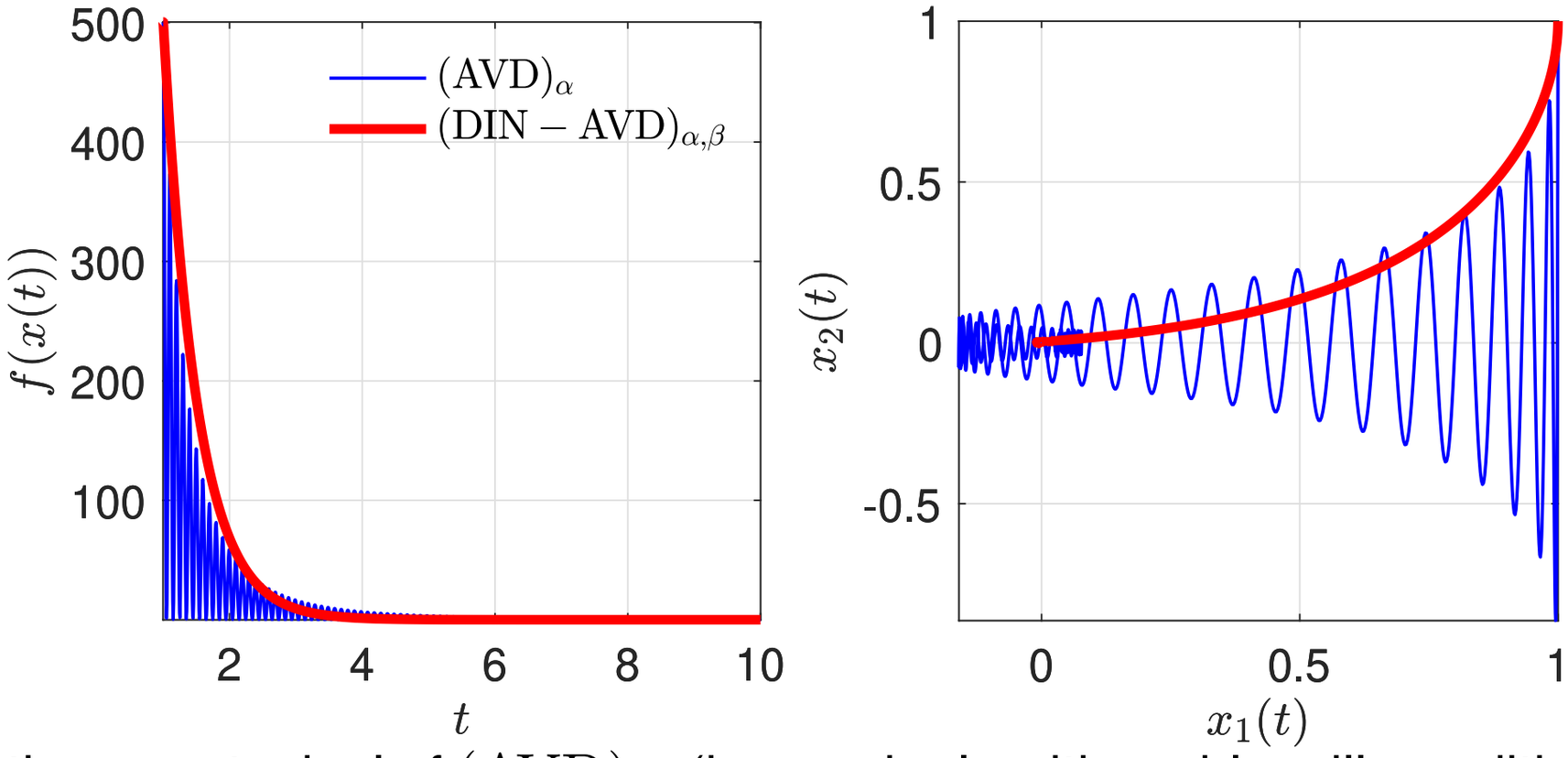
- $(\mathrm{NAG})_\alpha$ vs $(\mathrm{RAG})_\alpha$ [Attouch and Fadili 2022] :
  - Can be deduced from each other by reversing the order of extrapolation and gradient descent.
  - They have similar dynamic interpretation (both low and high resolution).
  - They enjoy the same convergence properties : for $\alpha > 3$
    - $f(w_k) - \min_{\mathcal{H}} f \leq f(y_k) - \min_{\mathcal{H}} f = o(1/k^2)$,
    - $w - \lim y_k = w - \lim x_k = w - \lim w_k \in \underset{\mathcal{H}}{\mathrm{Argmin}}\,(f)$.

Oscillations and geometric damping

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

- $\mathcal{H} = \mathbb{R}^2$, $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 1000x_2^2)$
- $\alpha = 3.1$.
- $(x_1(1), x_2(1)) = (1, 1)$, $(\dot{x}_1(1), \dot{x}_2(1)) = (0, 0)$.

A proposal: neutralize these oscillations by geometric damping



- Wild oscillations are typical of $(\mathrm{AVD})_\alpha$ (increasingly with problem ill-conditioning) and are due to viscous damping [Liang, Fadili and Peyré 2016] : workarounds in practice use warm restart.

Inertial dynamic with combined damping

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathcal{H})$$

- Inexact inertial dynamic with viscous and geometric damping, $t \in [t_0, +\infty[$, $t_0 > 0$:

(ISEHD-PERT) $$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\frac{d}{dt}\Big(\nabla f(x(t)) + e(t)\Big) + \nabla f(x(t)) + e(t) = 0$$

Viscous damping

Geometric Hessian-driven damping

Exogenous error

- If f is also C^2 :

$$\frac{d}{dt}(\nabla f(x(t))) = \nabla^2 f(x(t))\dot{x}(t),$$

hence the name Inertial System with *Explicit* Hessian Damping (ISEHD).

- Encompasses many related works as special cases [Alvarez et al. 2002 ; Attouch et al. 2012, 2016 ; Shi et al. 2018 ; Lin et al. 2019 ; Bot et al. 2019].

First-order formulation

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\frac{d}{dt}\Big(\nabla f(x(t)) + e(t)\Big) + \nabla f(x(t)) + e(t) = 0$$

$$\Updownarrow$$

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

- Wise formulation to show well-posedness.
- Natural extension to non-smooth $f \in \Gamma_0(\mathcal{H}) : \nabla f \to \partial f$.
- Natural extension to finding zeros of maximal monotone set-valued operators.
- Natural extension to the non-convex setting.

Outline

- Continuous dynamics.
- Discrete algorithms.
- Numerical results.
- Conclusion.

Outline

- Continuous dynamics.
- Discrete algorithms.
- Numerical results.
- Conclusion.

Continuous dynamics: f smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathcal{H})$$

(ISEHD-PERT) $$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\frac{d}{dt}\Big(\nabla f(x(t)) + e(t)\Big) + \nabla f(x(t)) + e(t) = 0$$

Theorem *Assume that* e *is* C^1 *with* $\int_{t_0}^{+\infty} t\|e(t)\|dt < +\infty$ *and* $\int_{t_0}^{+\infty} t\|\dot{e}(t)\|dt < +\infty$.

Let x *be a solution trajectory to* (ISEHD-PERT) *for* $\alpha > 3$ *and* $\beta > 0$ *such that* $t_0 > \frac{\beta(\alpha-2)}{\alpha-3}$. *Then the following holds :*

If not verified, convergence to a noise dominated region

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

Fast sublinear convergence rate

(ii) $\int_{t_0}^{+\infty} t^2\|\nabla f(x(t))\|^2 dt < +\infty.$

(iii) $\int_{t_0}^{+\infty} t\left(f(x(t)) - \inf_{\mathcal{H}} f\right) dt < +\infty.$

Integrability
$\beta \neq 0$ important

(iv) If, moreover, $t_0 \geq \frac{\beta(\alpha-2-\varepsilon)}{\alpha-3-\varepsilon}$ *for some* $\varepsilon \in]0, \alpha-3[$, *then*

(a) $x(t)$ *converges weakly to a minimizer of* f.

Convergence of the trajectory

(b) $f(x(t)) - \inf_{\mathcal{H}} f = o\left(t^{-2}\right)$ *and* $\|\dot{x}(t)\| = o\left(t^{-1}\right)$ *as* $t \to +\infty$.

Even faster asymptotic rates

Continuous dynamics: f non-smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H})$$

$$\begin{cases} \dot{x}(t) + \beta\Big(\partial f(x(t)) + e(t)\Big) - \Big(\frac{1}{\beta} - \frac{\alpha}{t}\Big)\, x(t) + \frac{1}{\beta} y(t) \ni 0; \\ \dot{y}(t) - \Big(\frac{1}{\beta} - \frac{\alpha}{t} + \frac{\alpha\beta}{t^2}\Big)\, x(t) + \frac{1}{\beta} y(t) = 0. \end{cases}$$

Theorem *Let $f \in \Gamma_0(\mathcal{H})$ such that $S \neq \emptyset$. Suppose that $e(\cdot) \in W^{1,1}(t_0, T; \mathcal{H})$ for all $T > t_0$, with $\int_{t_0}^{+\infty} t\|e(t) + \beta\dot{e}(t)\|dt < +\infty$. Ther*

If not verified, convergence to a noise dominated region

(x, y) of the above system with $\alpha \geq 3$ and $\beta > 0$, the following holds :

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

Fast sublinear convergence rate

(ii) $\int_{t_0}^{+\infty} t^2\|\xi(t)\|^2 dt < +\infty, \quad \xi(t) \in \partial f(x(t)).$

(iii) $\int_{t_0}^{+\infty} t\left(f(x(t)) - \inf_{\mathcal{H}} f\right) dt < +\infty.$

Integrability
$\beta \neq 0$ important

(iv) If, moreover, $t_0 \geq \frac{\beta(\alpha-2-\varepsilon)}{\alpha-3-\varepsilon}$ *for some* $\varepsilon \in]0, \alpha - 3[$, *then*

(a) $x(t)$ *converges weakly to a minimizer of* f.

Convergence of the trajectory

(b) $f(x(t)) - \inf_{\mathcal{H}} f = o\left(t^{-2}\right)$ *and* $\|\dot{x}(t)\| = o\left(t^{-1}\right)$ *as* $t \to +\infty$.

Even faster asymptotic rates

Continuous dynamics: f strongly convex

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathcal{H})$$

$$\ddot{x}(t) + \underbrace{2\sqrt{\mu}}_{\text{Yields the best rate}}\dot{x}(t) + \beta\frac{d}{dt}\Big(\nabla f(x(t)) + e(t)\Big) + \nabla f(x(t)) + e(t) = 0.$$

Theorem *Suppose that $f : \mathcal{H} \to \mathbb{R}$ is μ-strongly convex for some $\mu > 0$. Let $x(\cdot) : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the above system. Suppose that $0 \leq \beta \leq \dfrac{1}{2\sqrt{\mu}}$ and $\displaystyle\int_{t_0}^{+\infty} \|e(t) + \beta\dot{e}(t)\|dt < +\infty$. Then,*

(i) for all $t \geq t_0$

Fast exponential convergence — Noise dominated term

$$\frac{\mu}{2}\|x(t) - x^\star\|^2 \leq f(x(t)) - \min_{\mathcal{H}} f = O\left(e^{-\frac{\sqrt{\mu}}{2}(t-t_0)} + \|e(t) + \beta\dot{e}(t)\|\right).$$

(ii) Suppose moreover that for some $p > 0$, $\|e(t) + \beta\dot{e}(t)\| = O\left(\dfrac{1}{t^p}\right)$, as $t \to +\infty$. Then as $t \to +\infty$

Noise dominated sublinear convergence rate

$$f(x(t)) - \inf_{\mathcal{H}} f = O\left(\frac{1}{t^p}\right), \quad \|x(t) - x^\star\|^2 = O\left(\frac{1}{t^p}\right), \quad \|\dot{x}(t)\|^2 = O\left(\frac{1}{t^p}\right)$$

and $e^{-\sqrt{\mu}t}\displaystyle\int_{t_0}^{t} e^{\sqrt{\mu}s}\|\nabla f(x(s))\|^2 ds = O\left(\frac{1}{t^p}\right)$. Integrability

Outline

- Continuous inertial dynamics with Hessian damping:
- Discrete algorithms.
- Numerical results.
- Conclusion.

Proximal scheme: f convex smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathcal{H})$$

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\frac{d}{dt}\Big(\nabla f(x(t)) + e(t)\Big) + \nabla f(x(t)) + e(t) = 0.$$

Implicit-time discretization

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \frac{\alpha}{kh}\frac{x_{k+1} - x_k}{h} + \frac{\beta}{h}(\nabla f(x_{k+1}) - \nabla f(x_k)) + \nabla f(x_{k+1}) \approx_{\xi_k} 0.$$

Inexact computation

Inexact Inertial Proximal Algorithm with Hessian Damping (IIPAHD)

Input : $x_0, x_1, \alpha \geq 1, s = h^2$;

for $k \geq 1$ **do**

$$\alpha_k = 1 - \frac{\alpha}{k+\alpha};$$

$$\gamma_k = \alpha_k(\beta\sqrt{s} + s);$$

$$y_k = x_k + \alpha_k(x_k - x_{k-1}) + \beta\sqrt{s}\alpha_k\nabla f(x_k) + \xi_k;$$

$$x_{k+1} = \mathrm{prox}_{\gamma_k f}(y_k).$$

Viscous damping

Interacting dampings

One gradient evaluation

Inexact computation

One prox evaluation

$$\mathrm{prox}_{\gamma f}(x) = \operatorname*{Argmin}_{z\in\mathcal{H}} \frac{1}{2\gamma}\|z - x\|^2 + f(z).$$

Proximal scheme: f convex smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1(\mathcal{H})$$

Input : $x_0, x_1, \alpha \geq 1, s = h^2$;

for $k \geq 1$ **do**

$$\alpha_k = 1 - \tfrac{\alpha}{k+\alpha}\,;$$

$$\gamma_k = \alpha_k(\beta\sqrt{s} + s)\,;$$

$$y_k = x_k + \alpha_k(x_k - x_{k-1}) + \beta\sqrt{s}\alpha_k\nabla f(x_k) + \xi_k\,;$$

$$x_{k+1} = \operatorname{prox}_{\gamma_k f}(y_k).$$

Viscous damping

Interacting dampings

Inexact computation

$$\operatorname{prox}_{\gamma f}(x) = \operatorname*{Argmin}_{z\in\mathcal{H}} \frac{1}{2\gamma}\|z - x\|^2 + f(z).$$

Theorem *Suppose that* $\alpha > 3$ *and* $k\,\|\xi_k\| \in \ell^1_+(\mathbb{N})$. *Then, for any sequence* $(x_k)_{k\in\mathbb{N}}$ *of (IIPAHD)*

(i) $f(x_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$; Fast sublinear asymptotic rate

(ii) $\sum_k k^2\|\nabla f(x_k)\|^2 < +\infty$; Summability

(iii) x_k *converges weakly to a minimizer of* f. Convergence of the iterates

Properties of the continuous dynamics are preserved

Proximal scheme: f convex non-smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H})$$

- Rely on the Moreau-Yosida regularization, $\lambda > 0$,

$$f^\lambda(x) = \min_{z\in\mathcal{H}} \frac{1}{2\lambda}\|z-x\|^2 + f(z).$$

- $f^\lambda \in \Gamma_0(\mathcal{H}) \cap C^1_{1/\lambda}(\mathcal{H})$, and $\operatorname*{Argmin}_{\mathcal{H}} f = \operatorname*{Argmin}_{\mathcal{H}} f^\lambda$.
- Apply (IPAHD) to f^λ using Moreau proximal calculus :

$$\nabla f_\lambda(x) = \frac{1}{\lambda}\left(x - \operatorname{prox}_{\lambda f}(x)\right),$$

$$\operatorname{prox}_{\theta f_\lambda}(x) = \frac{\lambda}{\lambda+\theta}x + \frac{\theta}{\lambda+\theta}\operatorname{prox}_{(\lambda+\theta)f}(x).$$

Proximal scheme: f convex non-smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H})$$

Input : x_0, x_1, $\lambda > 0$, $\alpha \geq 1$, $s = h^2$;

$$\operatorname{prox}_{\gamma f}(x) = \operatorname*{Argmin}_{z\in\mathcal{H}} \frac{1}{2\gamma}\|z-x\|^2 + f(z).$$

for $k \geq 1$ **do**

$$\alpha_k = 1 - \frac{\alpha}{k+\alpha};$$

$$\gamma_k = \frac{\lambda}{\lambda+\alpha_k(\beta\sqrt{s}+s)};$$

$$y_k = x_k + \alpha_k(x_k - x_{k-1}) + \frac{\beta\sqrt{s}}{\lambda}\alpha_k\left(x_k - \operatorname{prox}_{\lambda f}(x_k)\right) + \xi_k;$$

$$x_{k+1} = \gamma_k y_k + (1-\gamma_k)\operatorname{prox}_{\lambda\gamma_k^{-1} f}(y_k).$$

Viscous damping

Interacting dampings

Inexact computation

Two prox evaluations

Theorem *Suppose that* $\alpha > 3$ *and* $k\,\|\xi_k\| \in \ell^1_+(\mathbb{N})$. *Then, for any sequence* $(x_k)_{k\in\mathbb{N}}$ *of (IIPAHD)*

(i) $f(\operatorname{prox}_{\lambda f}(x_k)) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$; Fast sublinear asymptotic rate

(ii) $\sum_k k^2\|x_{k+1} - \operatorname{prox}_{\lambda f}(x_{k+1})\|^2 < +\infty$; Summability

(iii) x_k *converges weakly to a minimizer of* f. Convergence of the iterates

Properties of the continuous dynamics are preserved

Gradient scheme: f convex smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\frac{d}{dt}\Big(\nabla f(x(t)) + e(t)\Big) + (1+\beta/t)\Big(\nabla f(x(t)) + e(t)\Big) = 0.$$

Explicit-time discretization

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

Inexact computation

Inexact Inertial Gradient Algorithm with Hessian Damping (IIGAHD)

Input : $x_0, x_1, \alpha \geq 3, s \leq 1/L, \beta \in [0, 2\sqrt{s}[$;

for $k \geq 1$ **do**

$$\alpha_k = 1 - \frac{\alpha}{k+\alpha};$$

$$y_k = x_k + \alpha_k(x_k - x_{k-1}) - \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1}) + \xi_k;$$

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

Viscous damping

Hessian damping

Interacting term

Inexact computation

Three gradient evaluations

IIGAHD: f convex smooth

$$\min_{x\in\mathcal{H}} f(x), \quad f \in \Gamma_0(\mathcal{H}) \cap C^1_L(\mathcal{H})$$

Input : $x_0, x_1, \alpha \geq 3, s \leq 1/L, \beta \in [0, 2\sqrt{s}[$;

for $k \geq 1$ **do**

$$\alpha_k = 1 - \tfrac{\alpha}{k+\alpha};$$

$$y_k = x_k + \alpha_k(x_k - x_{k-1}) - \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) - \tfrac{\beta\sqrt{s}}{k}\nabla f(x_{k-1}) + \xi_k;$$

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

Viscous damping

Hessian damping

Interacting term

Inexact computation

Theorem *Consider the sequence $(x_k)_{k\in\mathbb{N}}$ of (IIGAHD). Assume that $\alpha > 3$, $0 \leq \beta < 2\sqrt{s}$, $s \leq 1/L$ and $k\xi_k \in \ell^1_+(\mathbb{N})$. Then, the following hold :*

(i) $f(x_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$;

(ii) $\sum_k k^2\|\nabla f(x_k)\|^2 < +\infty$;

(iii) x_k *converges weakly to a minimizer of* f.

Fast sublinear asymptotic rate

Summability β>0

Convergence of the iterates

Hessian damping should not be too large to preserve acceleration of AVD

A few words on the non-convex case

$$\min_{x\in\mathcal{H}} f(x) \overset{\text{def}}{=} \mathbb{E}_w\left[\ell(x,w)\right], \quad f\in\Gamma(\mathcal{H})$$

- $Z(t)=(x(t),y(t))\in\mathcal{H}\times\mathcal{H}$.
- First order formulation is equivalent to

$$\dot{Z}(t)+\partial_C G(Z(t))+D(t,Z(t))\ni 0, \quad Z(t_0)=(x_0,y_0),$$

where $G(Z)=\beta f(x)$, and D is the affine operator

$$D(t,Z)=\left(-\left(\frac{1}{\beta}-\frac{\alpha}{t}\right)x+\frac{1}{\beta}y,-\left(\frac{1}{\beta}-\frac{\alpha}{t}+\frac{\alpha\beta}{t^2}\right)x+\frac{1}{\beta}y\right).$$

- Take $D(t,Z)=D(t_0,Z)$.
- Take iid samples w_k.
- Stochastic approximation à la Robbins-Monro :

$$Z_{k+1}-Z_k\in-\gamma_k\left(\partial_C\ell(x_k,w_k)+D(t_0,Z_k)\right),$$

$\lim_{k\to+\infty}\gamma_k=0, \gamma_k\notin\ell^1_+(\mathbb{N})$.

Theorem *Assume that $\mathcal{H}$ is finite-dimensional, that f is a tame function (e.g. semialgebraic) and take $\gamma_k=o\left(\frac{1}{\log k}\right)$. Conditioning on the event that $(Z_k)_{k\in\mathbb{N}}$ is uniformly bounded, then with probability 1, each accumulation point of $(x_k)_{k\in\mathbb{N}}$ is a critical point of f and f is constant there.*

Outline

- Continuous inertial dynamics with Hessian damping:
- Discrete algorithms:
- Numerical results.
- Conclusion.

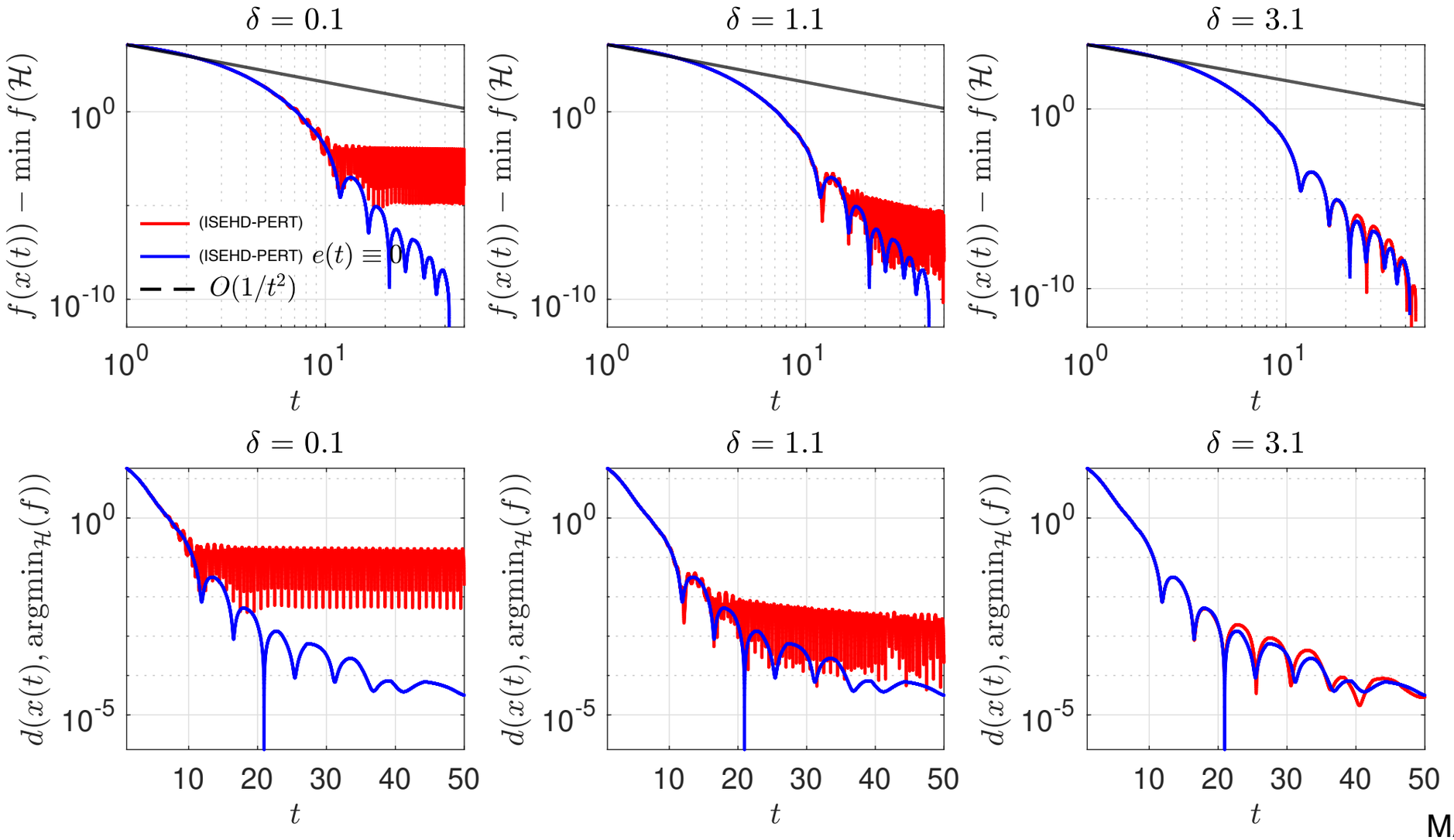
Simple inexact setup

$$f(x_1, x_2) = (x_1 - 1)^4 + (x_2 - 5)^2 + 0.1(|x_1| + |x_2|)$$

$$e(t) = \frac{\cos(2\pi t)}{t^\delta} \text{ with } \delta \in \{0.1, 1.1, 3.1\}$$

$$f(x(t)) - f^\star \leq O\left(\frac{1}{t^2}\right) + \frac{C\left(\int_{t_0}^{t} \tau \|e(\tau)\| d\tau\right)^2}{t^2}.$$

$$\sim t^{-2(\delta-1)}, \delta \in]1, 2]$$



Large-scale exact setup

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

- $A:\mathbb{R}^n\to\mathbb{R}^m$ is linear, $m\leq n$.
- $g\in\Gamma_0(\mathbb{R}^n)$ acts as a regularizer (typically prox-friendly).
- (RLS) is extremely popular in a variety of fields : variational inverse problems in data processing, machine learning, statistics, etc.
- Fits our framework by working with the metric $\|x\|_M^2=\langle Mx,x\rangle$, $M=\lambda^{-1}I-A^*A$.
- $0<\lambda\left\|A\right\|^2<1\Rightarrow M$ is symmetric definite-positive.
- Apply IGAHD algorithm to f^M, the Moreau envelope of f in the metric M,

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

- f^M is convex, C_1^1 in the metric M, $\operatorname*{Argmin}_{\mathbb{R}^n} f=\operatorname*{Argmin}_{\mathbb{R}^n} f^M$ and

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

Standard FB fixed point operator

IGAHD for regularized regression

Input : $x_0, x_1, \alpha > 0, \lambda \in]0, 1/\|A\|^2[, \beta \in [0, 2\sqrt{s}[, s \leq 1$;

for $k \geq 1$ **do**

Viscous damping — Hessian damping — Interacting term

$$\alpha_k = 1 - \frac{\alpha}{k} ;$$

$$z_k = x_k - \mathrm{prox}_{\lambda g}(x_k + \lambda A^*(b - Ax_k)) ;$$

$$y_k = x_k + \alpha_k(x_k - x_{k-1}) - \beta\sqrt{s}\,(z_k - z_{k-1}) - \frac{\beta\sqrt{s}}{k} z_{k-1} ;$$

$$x_{k+1} = (1 - s)y_k + s\,\mathrm{prox}_{\lambda g}(y_k + \lambda A^*(b - Ay_k)) .$$

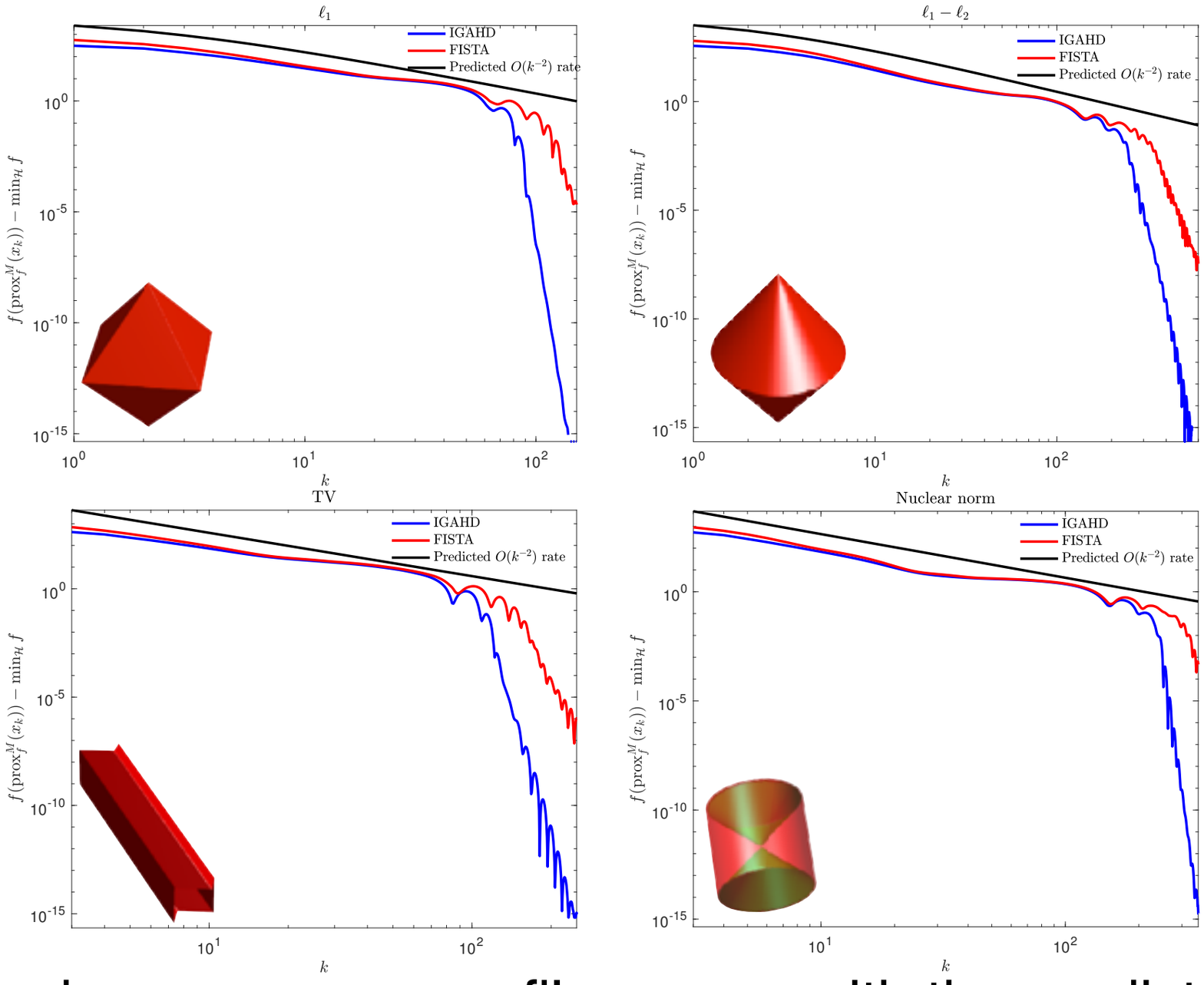
Theorem *Suppose the above assumptions on* $(\lambda, \alpha, \beta, s)$*. The sequence* $(x_k)_{k\in\mathbb{N}}$ *of (IGAHD) for solving (RLS) obeys*

$$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathbb{R}^n} f = O(1/k^2). \quad \mathrm{prox}_{fM}(x) = \mathrm{prox}_{\lambda g}(x + \lambda A^*(y - Ax))$$

If moreover $\beta > 0$*, then*

$$\sum_{k\in\mathbb{N}} k^2 \|\nabla f(x_k)\|^2 < +\infty.$$

Convergence profiles



- Observed convergence profiles agree with the predicted (wort-case) rate.
- (IGAHD) exhibits, as expected, much less oscillations than FISTA, and eventually converges faster.

Take away messages

- A dynamical perspective on accelerated optimization algorithms with errors ad **viscous-Hessian**-driven damping.
- The key is inertia.
- A **unified** analysis of convergence and integrability.
- **Provably accelerated inexact** algorithms without explicit Hessian construction.
- **Hessian** geometric damping **neutralizes** oscillations: get the best of both worlds.
- Convergence of the trajectories and iterates (✓).
- Faster asymptotic rates (✓).
- Inexact deterministic case (✓).
- Stochastic setting (ongoing).
- Operator splitting (ongoing).
- Beyond convexity (ongoing).

Preprints on arxiv and papers on

https://fadili.users.greyc.fr/

Thanks

Any questions ?