Acceleration of first-order algorithms via inertial dynamics with Hessian damping

Hedy ATTOUCH

Université Montpellier
Institut Montpelliérain Alexander Grothendieck, UMR CNRS 5149

Based on joint work with Z. Chbani, J. Fadili, H. Riahi.

One World Optimization Seminar

January 18, 2021

Introduction

Convex optimization

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

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

Damped inertial dynamic

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

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

Temporal discretization, step size $\sqrt{s}$

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

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

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

Convergence rates

$f : \mathcal{H} \to \mathbb{R}$ convex, ∇f L-Lipschitz, $\alpha \geq 3$, $\quad 0 < \beta < 2\sqrt{s}$, $\quad sL \leq 1$.

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

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

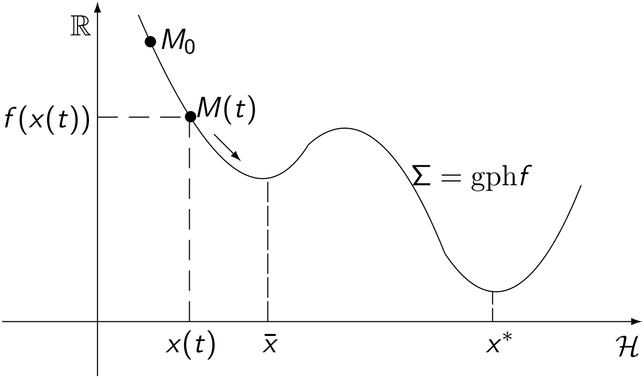
Contents

1. INERTIAL DYNAMICS/ALGORITHMS IN OPTIMIZATION

1. The heavy ball with friction method of Polyak (64,87)

$\gamma > 0$: fixed viscous damping coefficient

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



Exploration of local minima via (HBF)

The heavy ball with friction method of Polyak

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

General convex case

- $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to +\infty$.
- $x(t) \rightharpoonup x_\infty \in S$ weakly (Alvarez, 2000).

Strongly convex case: $f - \frac{\mu}{2}\|\cdot\|^2$ 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 $t \to +\infty$.
- Link between the geometry of f and the damping coefficient.

Recent trends in dissipative autonomous systems:

Haraux-Jendoubi (Springer Briefs, 2015), A.-Bot-Csetnek (JEMS, 2021).

2. Asymptotic vanishing damping: $\gamma(t) \to 0$ as $t \to +\infty$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \quad t \geq t_0.$$

Theorem (Cabot-Engler-Gaddat (TAMS 2009))

The optimization property is satisfied if $\int_{t_0}^{+\infty} \gamma(t)dt = +\infty.$

Define: $p(t) := \exp\left(\int_{t_0}^{t} \gamma(\tau)d\tau\right)$, $\Gamma_\gamma(t) := p(t)\int_t^{+\infty} \frac{ds}{p(s)}$.

Theorem (A.-Cabot (JDE 2017))

$$\gamma(t)\Gamma_\gamma(t) \leq m \quad \text{for some } m < \tfrac{3}{2}.$$

$$\Downarrow$$

$$f(x(t)) - \min_{\mathcal{H}} f = \left(\frac{1}{\int_{t_0}^{t} \Gamma_\gamma(s)\, ds}\right), \quad \|\dot{x}(t)\| = o\left(\frac{1}{\int_{t_0}^{t} \Gamma_\gamma(s)\, ds}\right)^{1/2}.$$

Case: $\gamma(t) = \frac{\alpha}{t}$, $\Gamma_\gamma(t) = \frac{t}{\alpha - 1}$, which gives $\alpha > 3$.

Su-Boyd-Candès model for Nesterov accelerated 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)

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

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

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

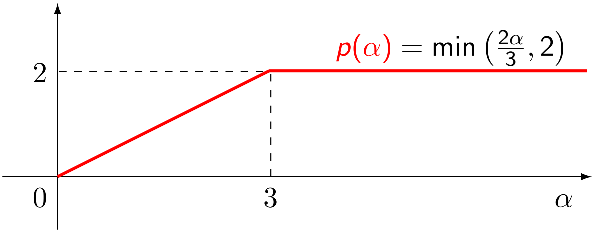
- $\alpha \leq 3$: Apidopoulos-Aujol-Dossal (SIOPT 2018),
  A.-Chbani-Riahi (ESAIM COCV 2019)

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

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

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

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



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

Temporal discretization, h step size, $s = h^2$

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

Different choices for y_k

- Explicit: $y_k = x_k$, Heavy Ball with Friction, Polyak

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

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

$$(\text{IP})_\alpha \qquad x_{k+1} = \text{prox}_{sf}\left(x_k + \left(1 - \tfrac{\alpha}{k}\right)(x_k - x_{k-1})\right).$$

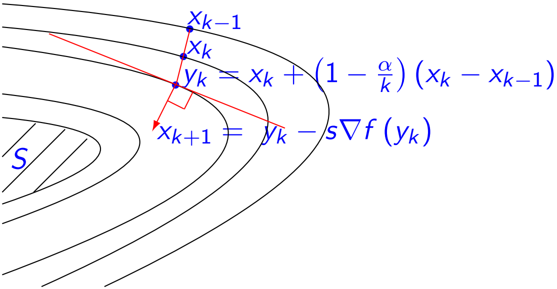
- Nesterov: $\quad y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})$ Inertial Gradient algorithm

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

$$\min \{f(x) : \ x \in \mathcal{H}\}, \ f : \mathcal{H} \to \mathbb{R} \text{ convex differentiable}, \ S = \operatorname{argmin} f \neq \emptyset.$$

Inertial Gradient algorithm, Nesterov (1983, 2004)

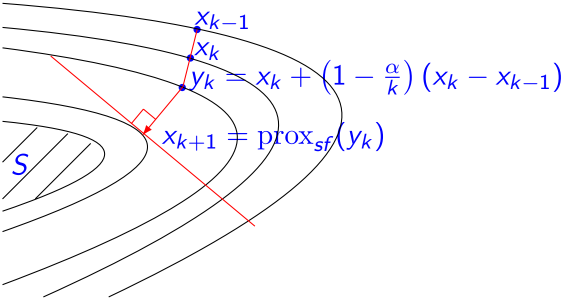
$$(\mathrm{IG})_\alpha \quad \begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla f(y_k) \end{cases}$$



$\min\{f(x):\ x \in \mathcal{H}\},\ f:\mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ closed, convex, $S = \mathrm{argmin} f \neq \emptyset$.

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

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



- $f : \mathcal{H} \to \mathbb{R}$ 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.

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

- $\alpha = 3$: $(f+g)(x_k) - \min_{\mathcal{H}}(f+g) = \mathcal{O}(\frac{1}{k^2})$,

  Nesterov (1983), Beck-Teboulle: FISTA (SIAM J. Imaging 2009).

- $\alpha > 3$: $(f+g)(x_k) - \min_{\mathcal{H}}(f+g) = o(\frac{1}{k^2})$, $x_k \rightharpoonup x_\infty \in S$,

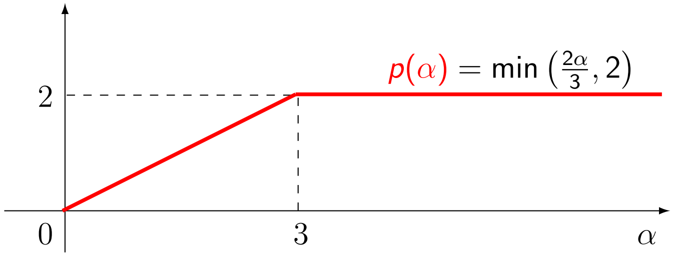
  Chambolle-Dossal (JOTA 2015), A.-Peypouquet (SIOPT 2016).

- $\alpha \leq 3$: $(f+g)(x_k) - \min_{\mathcal{H}}(f+g) = \mathcal{O}\left(\frac{1}{k^{\frac{2\alpha}{3}}}\right)$.

  Apidopoulos-Aujol-Dossal (Math Prog '20), A.-Chbani-Riahi (COCV '18)

Convergence rate of the Inertial Proximal Gradient algorithm

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



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

2. INERTIAL DYNAMICS WITH HESSIAN DRIVEN DAMPING

Historical aspects

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

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

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

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

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

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

Related works

Shi-Du-Jordan-Su (arXiv:1810.08907, '18), Lin-Jordan (arXiv:1912.07168, '19)

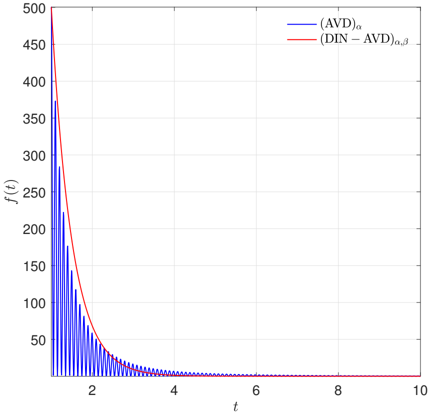
Castera-Bolte-Févotte-Pauwels (hal-02140748, 2019),

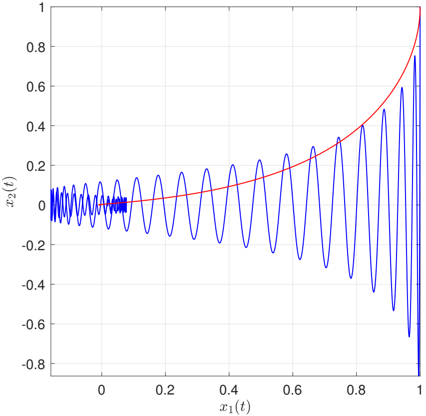
Bot-Csetnek-László (Math. Program., 2019),

Alecsa-László-Pinta (AMO '20), Adly-A. (SIOPT 2020).

Compare $(\mathrm{AVD})_\alpha$ with $(\mathrm{DIN}-\mathrm{AVD})_{\alpha,\beta}$

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





Link with the Newton method

- f convex, $\mathcal{C}^2$, solve $\nabla f(x) = 0$ by the Newton method.
- $\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0$.
- Continuous version: $\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0$: Ill-posed.
- Levenberg-Marquardt regularization, $\gamma(t) > 0$

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

Well-posed: A.-Svaiter (SICON, 2011). Valid with a general maximally monotone operator, closed loop form.

- Dynamical Inertial Newton method

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

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

Lyapunov analysis: A.-Peypouquet-Redont (JDE 2016) (APR for short)

Theorem (APR)

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

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

In addition, when $\beta>0$: $\displaystyle\int_{t_0}^{+\infty} t^2\|\nabla f(x(t))\|^2 dt<+\infty.$

Lyapunov function: $x^*\in\operatorname{argmin}_{\mathcal{H}} f$, $m=\min_{\mathcal{H}} f=f(x^*)$

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

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

Lyapunov analysis

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

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

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

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

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

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

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

Theorem (APR)

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

i) $x(t)\rightharpoonup x_\infty\in\operatorname{argmin}_{\mathcal{H}} f$

ii) $f(x(t))-\min_{\mathcal{H}} f=o\left(\frac{1}{t^2}\right), \quad \|\dot{x}(t)+\beta\nabla f(x(t))\|=o\left(\frac{1}{t}\right).$

iii) $\int_{t_0}^{+\infty} t\left(f(x(t))-\min_{\mathcal{H}} f\right)dt<+\infty, \quad \int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt<+\infty.$

Proof: Lyapunov analysis via the anchoring functions

$h_z(t)=\frac{1}{2}\|x(t)-z\|^2$ $(z\in\operatorname{argmin}_{\mathcal{H}} f)$, and Opial's lemma.

Perturbed dynamic, $e : [t_0, +\infty[\to \mathcal{H}$

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

Theorem (APR)

Let $x : [t_0, +\infty[\to \mathcal{H}$ *be a solution trajectory of* $(\mathrm{DIN-AVD})_{\mathrm{pert}}$.

Suppose that $\displaystyle\int_{t_0}^{+\infty} t\,\|e(t)\|\,dt < +\infty$. *Then,*

- $\alpha \geq 3$, $\beta \geq 0$: $\quad f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\dfrac{1}{t^2}\right)$ *as* $t \to +\infty$.
- $\alpha > 3$, $\beta \geq 0$: $\quad x(t) \rightharpoonup x_\infty \in \mathrm{argmin}_{\mathcal{H}} f$ *as* $t \to +\infty$.
- $\alpha \geq 3$, $\beta > 0$: $\quad \displaystyle\int_{t_0}^{+\infty} t^2\|\nabla f(x(t))\|^2 dt < +\infty$.

Strongly convex case

Theorem (ACFR)

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*

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

Suppose that $0 \le \beta \le \frac{1}{2\sqrt{\mu}}$. *Then, for all* $t \ge t_0$

$$i) \quad \frac{\mu}{2}\|x(t) - x^\star\|^2 \le f(x(t)) - \min_{\mathcal{H}} f \le Ce^{-\frac{\sqrt{\mu}}{2}(t-t_0)},$$

where $C := f(x(t_0)) - \min_{\mathcal{H}} f + \mu \mathrm{dist}(x(t_0), S)^2 + \|\dot{x}(t_0) + \beta\nabla f(x(t_0))\|^2$.

$$ii) \quad e^{-\sqrt{\mu}t}\int_{t_0}^{t} e^{\sqrt{\mu}s}\|\nabla f(x(s))\|^2 ds \le C_1 e^{-\frac{\sqrt{\mu}}{2}t}.$$

Moreover, $\int_{t_0}^{\infty} e^{\frac{\sqrt{\mu}}{2}t}\|\dot{x}(t)\|^2 dt < +\infty$.

3. INERTIAL ALGORITHMS
with
HESSIAN DRIVEN DAMPING

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

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

Temporal rescaling of $(\text{DIN-AVD})_{\alpha,\beta}$

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

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 (arXiv 2018).

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

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

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

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*

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

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

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

Reinforced version of the gradient descent lemma.
Since $s \leq \frac{1}{L}$, f convex, and ∇f is L-lipschitz continuous,

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

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

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

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

Strongly convex case

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

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

Explicit time discretization with centered finite differences

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

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

Theorem (ACFR)

$\beta \leq \frac{1}{\sqrt{\mu}},\ L \leq \min\left\{\frac{\sqrt{\mu}}{8\beta}, \frac{\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}}{2\beta\mu+\frac{1}{\sqrt{s}}+\frac{\sqrt{\mu}}{2}}\right\}$. *Set* $q = \frac{1}{1+\frac{1}{2}\sqrt{\mu s}}$, $\theta = \frac{1}{1+\sqrt{\mu s}}$.

(i) $\|x_k - x^\star\| = \mathcal{O}\left(q^{k/2}\right)$ *and* $f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(q^k\right)$ *as* $k \to +\infty$.

(ii) $\theta^k \sum_{p=0}^{k-2} \theta^{-j}\|\nabla f(x_j)\|^2 = \mathcal{O}\left(q^k\right)$ *as* $k \to +\infty$.

Nonsmooth convex case

$f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ μ-strongly convex.
Idea: replace f with its Moreau envelope.
Preserves the infimal value and the solution set.

Proximal calculus (Bauschke-Combettes)

- f μ-strongly convex $\Longrightarrow$ f_λ strongly convex with modulus $\frac{\mu}{1+\lambda\mu}$.
- $\nabla f_\lambda(x) = \frac{1}{\lambda}\left(x - \mathrm{prox}_{\lambda f}(x)\right)$,
- $\mathrm{prox}_{\theta f_\lambda}(x) = \frac{\lambda}{\lambda+\theta} x + \frac{\theta}{\lambda+\theta} \mathrm{prox}_{(\lambda+\theta) f}(x)$.

$$
\begin{cases}
y_k = x_k + (1-a)(x_k - x_{k-1}) + \frac{\beta\sqrt{s}}{\lambda}(1-a)\left(x_k - \mathrm{prox}_{\lambda f}(x_k)\right) \\
x_{k+1} = \frac{\lambda}{\lambda+\theta} y_k + \frac{\theta}{\lambda+\theta} \mathrm{prox}_{(\lambda+\theta) f}(y_k).
\end{cases}
$$

Similar type of convergence rates.

4. 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

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

(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)

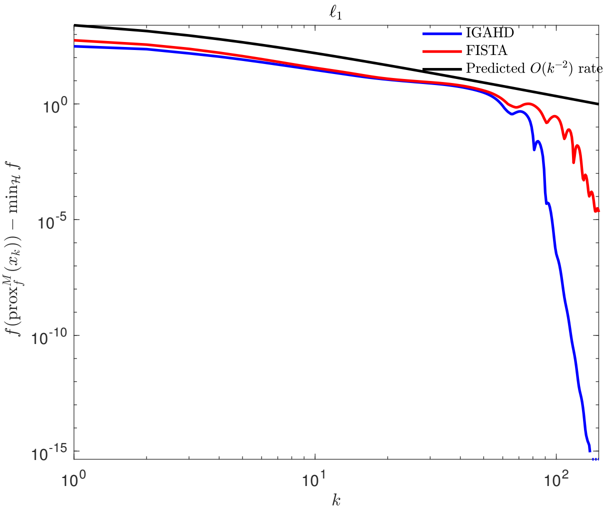
Assumptions: $0 < \lambda \|A\|_2^2 < 1$, $\alpha \geq 3$, $0 \leq \beta < 2\sqrt{s}$, $s \leq 1$.

Let (x_k) *be generated by* (IGAHD) *for* (RLS). *Then,*

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

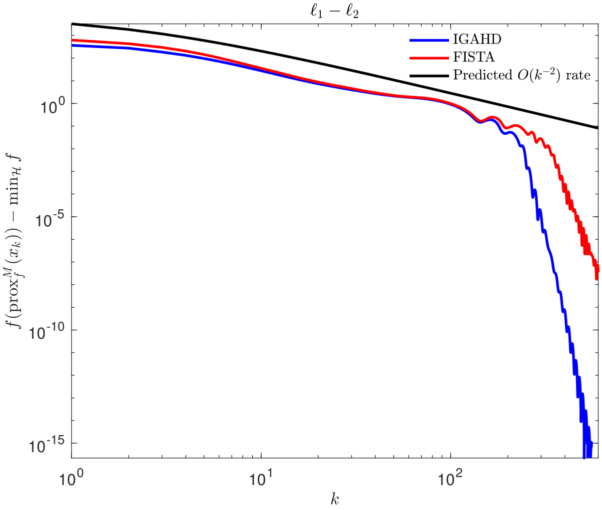
where $\mathrm{prox}_f^M(x_k) := \mathrm{prox}_{\lambda g}(x_k - \lambda A^*(Ax_k - b))$.

Lasso:



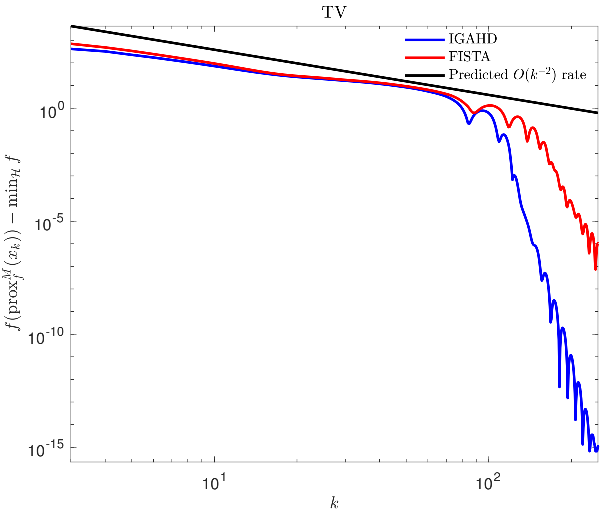
ℓ_1
IGAHD
FISTA
Predicted $O(k^{-2})$ rate
10^0
10^{-5}
10^{-10}
10^{-15}
$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathcal{H}} f$
10^0
10^1
10^2
k

Group Lasso:

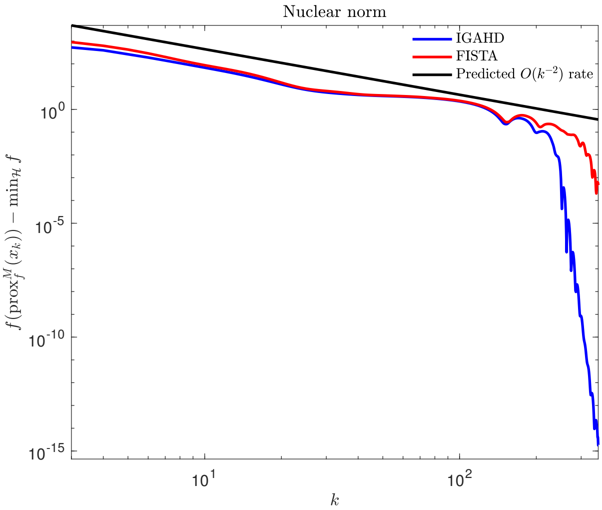


$\ell_1 - \ell_2$
IGAHD
FISTA
Predicted $O(k^{-2})$ rate
10^0
10^{-5}
10^{-10}
10^{-15}
$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathcal{H}} f$
10^0
10^1
10^2
k

TV (total variation):



Nuclear norm:



5. RELATED SYSTEMS

1. First-order in time and space equivalent formulation

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).
- Numerical applications (temporal discretization):

  Castera-Bolte-Févotte-Pauwels (Deep Learning) (HAL 2019).

  Maingé-Labarre (Fast convergence results, 2020).

2. Alecsa-László-Pinta dynamic model (AMO 2020).

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

- Implicit (DIN-AVD): Taylor expansion as $t \to +\infty$, $\dot{x}(t) \to 0$,

$$\nabla f\Big(x(t) + (\gamma + \frac{\beta}{t})\dot{x}(t)\Big) \approx \nabla f(x(t)) + (\gamma + \frac{\beta}{t})\nabla^2 f(x(t))\dot{x}(t).$$

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

- Explicit temporal discretization $\longrightarrow$ Nesterov accelerated gradient.
- Fast convergence rates:

$$f\Big(x(t) + (\gamma + \frac{\beta}{t})\dot{x}(t)\Big) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^2}\right).$$

$$\int_{t_0}^{+\infty} t^2 \|\nabla f\Big(x(t) + (\gamma + \frac{\beta}{t})\dot{x}(t)\Big)\|^2 dt < +\infty.$$

- Muehlebach-Jordan (arXiv:1905.07436v1, 2019).

3. Dry friction with Hessian damping, Adly–A. (SIOPT 2020)

$$\ddot{x}(t) + \gamma \dot{x}(t) + \partial\phi(\dot{x}(t)) + \beta \nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0.$$

$\phi : \mathcal{H} \to \mathbb{R}_+$ convex, sharp minimum at the origin, $(\phi(x) = r\|x\|)$;

$f : \mathcal{H} \to \mathbb{R}$ differentiable, not necessarily convex, ∇f L-Lipschitz.

Inertial Proximal-Gradient Algorithm

$$y_k = \frac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \frac{\beta}{1+h\gamma}(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{h}{1+h\gamma}\nabla f(x_k)$$

$$x_{k+1} = x_k + h\,\mathrm{prox}_{\frac{h}{1+h\gamma}\phi}(y_k).$$

Suppose $\gamma \geq L\left(\frac{h}{2} + \beta\right)$. Then,

- Finite length: $\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\| < +\infty, \quad \lim_{k\to\infty} x_k := x_\infty$.
- Approximate critical point: $-\nabla f(x_\infty) \in \partial\phi(0)$.
- Geometric convergence if $-\nabla f(x_\infty) \in \mathrm{int}(\partial\phi(0))$.
- Tolerates errors not converging to zero: $\|e_k\| \leq r' < r$.

6. MONOTONE INCLUSIONS.

Convergence of inertial dynamics for monotone inclusions

$A : \mathcal{H} \to 2^{\mathcal{H}}$ maximally monotone, $J_{\lambda A} = (I + \lambda A)^{-1}$, $A_\lambda = \frac{1}{\lambda}(I - J_{\lambda A})$.

Claim: $x(t) \rightharpoonup x_\infty \in A^{-1}(0)$ as $t \to +\infty$ in the following cases:

$A : \mathcal{H} \to \mathcal{H}$ λ-cocoercive, $\lambda\gamma^2 > 1$, Alvarez-A. (2001), A.-Maingé (2011)

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

$A : \mathcal{H} \to 2^{\mathcal{H}}$ general maximally monotone operator

- A.-Peypouquet (Math. Program. 2019), $\alpha > 2$, $\lambda(t) = (1+\epsilon)\dfrac{t^2}{\alpha^2}$.

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

- A.-László (SIOPT 2020) $\alpha > 1$, $\beta \geq 0$, $\lambda(t) = \lambda t^2$, $\lambda > \frac{1}{(\alpha-1)^2}$.

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta \frac{d}{dt}\left(A_{\lambda(t)}(x(t))\right) + A_{\lambda(t)}(x(t)) = 0.$$

Rate of convergence $\|\dot{x}(t)\| = o\left(1/t\right)$.

Proximal Regularized Inertial Newton Algorithm for Monotone operator

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta \frac{d}{dt}\left(A_{\lambda(t)}(x(t))\right) + A_{\lambda(t)}(x(t)) = 0.$$

Implicit finite-difference scheme: $t_k = kh$, $x_k = x(t_k)$, $\lambda_k = \lambda(t_k)$, $\alpha_k = 1 - \frac{\alpha}{k}$.

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{kh^2}(x_k - x_{k-1}) + \frac{\beta}{h}(A_{_k}(x_k) - A_{\lambda_{k-1}}(x_{k-1})) + A_{\lambda_{k+1}}(x_{k+1}) = 0.$$

$$(\text{PRINAM}) \begin{cases} y_k = \left(1 - \beta\left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k-1}}\right)\right)x_k + \left(\alpha_k - \frac{\beta}{\lambda_{k-1}}\right)(x_k - x_{k-1}) \\ \qquad\qquad + \beta\left(\frac{1}{\lambda_k}J_{\lambda_k A}(x_k) - \frac{1}{\lambda_{k-1}}J_{\lambda_{k-1}A}(x_{k-1})\right) \\ x_{k+1} = \dfrac{\lambda_{k+1}}{\lambda_{k+1} + s}y_k + \dfrac{s}{\lambda_{k+1} + s}J_{(\lambda_{k+1}+s)A}(y_k). \end{cases}$$

Relaxed proximal inertial algo: A.-Cabot (MP '19), A.-Peypouquet (MP '19).
$A = \partial f$, $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ convex lower semicontinuous proper.

Geometric interpretation of (PRINAM)

$\lim_{\lambda \to +\infty} J_{\lambda A} x = \text{proj}_S(x)$. Hence $J_{(_{k+1}+s)A}(y_k) - y_k \sim \text{proj}_S(y_k) - y_k$.
But only a small step in this direction.

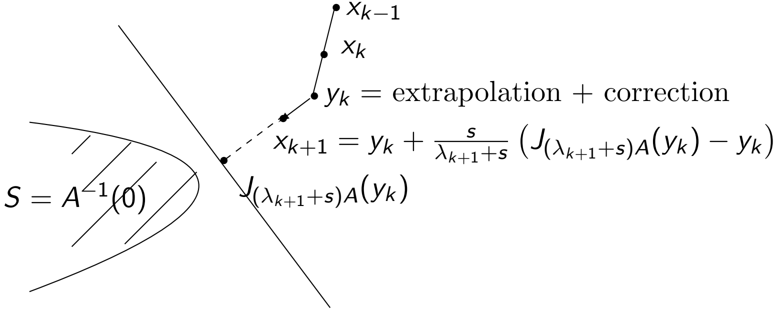


Figure: (PRINAM) algorithm

Theorem (A.-László (SIOPT 2020))

Assumption: $A : \mathcal{H} \to 2^{\mathcal{H}}$ *maximally monotone,* $S = A^{-1}(0) \neq \emptyset$.
$\alpha_k = \frac{t_{k-1}}{t_{k+1}}, t_k = rk + q,\ r > 0, q \in \mathbb{R}$ *and*

$$\lambda_k = \lambda k^2 \ \textit{with} \ \lambda > \frac{(2\beta+s)^2 r^2}{s}.$$

Then, for any sequences (x_k), (y_k) *generated by* (PRINAM)

i) *The speed* $(x_{k+1} - x_k)_{k\geq 1}$ *tends to zero, and*

$$\|x_{k+1} - x_k\| = \mathcal{O}\left(\tfrac{1}{k}\right) \ \textit{as} \ k \to +\infty, \quad \sum_{k\geq 2} k\|x_k - x_{k-1}\|^2 < +\infty$$

$$\|A_{\lambda_k}(x_k)\| = o\left(\frac{1}{k^2}\right) \ \textit{as} \ k \to +\infty, \quad \sum_{k\geq 1} k^3 \|A_{\lambda_k}(x_k)\|^2 < +\infty.$$

ii) *The sequence* (x_k) *converges weakly to some* $\hat{x} \in S$, *as* $k \to +\infty$.

iii) *The sequence* (y_k) *converges weakly to* $\hat{x} \in S$, *as* $k \to +\infty$.
$\|y_k - x_k\| = \mathcal{O}\left(\frac{1}{k}\right)$, *and so* $y_k - x_k$ *converges strongly to zero.*

7. PERSPECTIVE, OPEN QUESTIONS

TAKE AWAY MESSAGE

- A dynamic perspective on accelerated optimization algorithms with viscous and Hessian-driven damping;
- The key is inertia;
- A unified analysis of convergence and integrability;
- New provably accelerated algorithms without explicit Hessian construction.
- Hessian geometric damping neutralizes oscillations; get the best of both world.
- Convergence of trajectories and iterates (Yes);
- Faster asymptotic convergence rates (Yes);
- Inexact/stochastic case (ongoing);
- Operator splitting.

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 $\alpha > 3$?
- The sequence (y_k) follows the Ravine method. Is is possible to obtain $1/k^2$ rate of convergence with the Ravine method?
- Is is possible to obtain $1/k^2$ rate of convergence with autonomous dynamic/algorithms?

THANK YOU FOR YOUR ATTENTION

ANY QUESTIONS?

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.
- 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.

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, 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, published online June 2019, https://doi.org/10.1007/s10107-019-01412-0.
- 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, Z. Chbani, J. Fadili, H. Riahi, *First order optimization algorithms via inertial systems with Hessian driven damping*, Math. Program. (2020), https://doi.org/10.1007/s10107-020-01591-1.

References

- 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. ATTOUCH, S. C. LÁSZLÓ, *Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators*, (2020), SIAM J. Optim., hal-02549730.

- H. ATTOUCH, S. C. LÁSZLÓ, *Continuous Newton-like Inertial Dynamics for Monotone Inclusions*, Set Valued and Variational Analysis, (2020), https://doi.org/10.1007/s11228-020-00564-y, hal-02577331.

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

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.

R. I. Bot, E. R. Csetnek, S.C. Laszlo, *Tikhonov regularization of a second order dynamical system with Hessian damping*, (2019), Math. Program., DOI:10.1007/s10107-020-01528-8.

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.

References

- 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.
- 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.
- M. Muehlebach, M. I. Jordan, *A Dynamical Systems Perspective on Nesterov Acceleration*, (2019), arXiv:1905.07436
- Y. Nesterov, *A method of solving a convex programming problem with convergence rate O(1/k2)*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

References

- Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer, 2004.
- 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*, arXiv:submit/2440124[cs.LG] 21 Oct 2018.
- 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).

Annex 1. The Ravine method

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

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

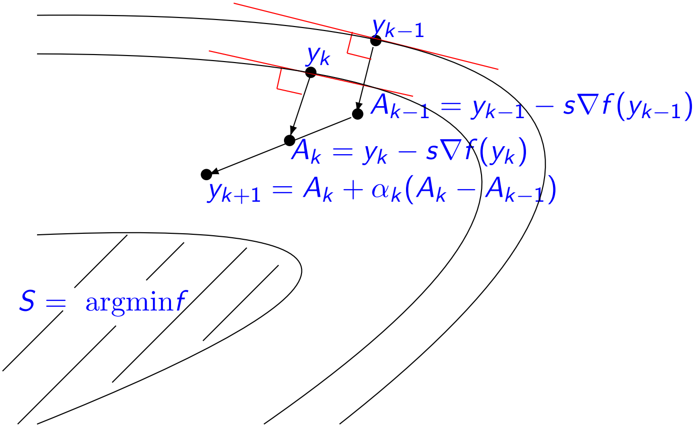
$$y_{k+1} = x_{k+1} + \left(1 - \frac{\alpha}{k+1}\right)(x_{k+1} - x_k)$$

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

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

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

Ravine method: Gelfand, Tsetlin (1961), Nesterov (1983), Polyak ('18).



Shi-Du-Jordan-Su (2018): High-resolution ode, arXiv:1810.08907v3.

Annex 2. General coefficients

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

Theorem (A.-Bahlag-Chbani-Riahi,(EECT)), hal-02940534.

Let $x : [t_0, +\infty[\to \mathcal{H}$ *be a solution trajectory of*

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \left(\frac{\beta}{t} + d(t)t^{\alpha-3}\right)\nabla f(x(t)) = 0$$

where $d(\cdot)$ *is a nonincreasing positive function. Then,*

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

$$b)\ \int_{t_0}^{+\infty} -\dot{d}(t)t^{\alpha-1}(f(x(t)) - \inf_{\mathcal{H}} f)dt < +\infty;$$

$$c)\ \int_{t_0}^{+\infty} t^{\alpha-1}d(t)\|\nabla f(x(t))\|^2 dt < +\infty.$$