The differential inclusion modeling FISTA algorithm and optimality in the case $b < 3$

Vassilis Apidopoulos

Joint work with Aujol J.-F. and Dossal Ch.

Institut de Mathématiques de Bordeaux
Université de Bordeaux

Journées annuelles 2017 des GdR MOA et MIA
20 october 2017

The minimization problem

$$\min\{F(x) : x \in \mathcal{H}\} \qquad \text{(P)}$$

- $\mathcal{H} = \mathbb{R}^d$, $d \geq 1$.
- $F = f + g : \mathcal{H} \longrightarrow \overline{\mathbb{R}}$, where :
  - $f : \mathcal{H} \longrightarrow \mathbb{R}$ proper, convex, differentiable with ∇f L-Lipschitz
  - $g : \mathcal{H} \longrightarrow \overline{\mathbb{R}}$ proper, convex, lower semi-continuous.
- $x^* \in \arg\min\{F\} \neq \emptyset$

Solutions of (P) : $\qquad \{\nabla f(x)\} + \partial g(x) = \partial F(x) \ni 0$

Dynamical system

$$\ddot{x}(t) + \frac{b}{t}\dot{x}(t) + \partial F(x(t)) \ni 0 \qquad \text{(DI)}$$

where $b > 0$.

Discretization of (DI)-FISTA (Su et al '15, Attouch et al '16)

Time step : $h > 0$

Explicit discretization with respect to ∇f

Implicit discretization with respect to ∂g

Algorithm 1 FISTA (Beck '09 et al, Nesterov '83/'04)

Let $x_0, x_1 \in \mathbb{R}^d$ and $\gamma = h^2$. For all $n \geq 1$, define $\{x_n\}_{n\in\mathbb{N}}$, $\{y_n\}_{n\in\mathbb{N}}$ as follows :

$$\begin{aligned} y_n &= x_n + \frac{n}{n+b}(x_n - x_{n-1}) \\ x_{n+1} &= Prox_{\gamma g}(y_n - \gamma \nabla f(y_n)) \\ &= \left(Id + \partial \gamma g\right)^{-1}(y_n - \gamma \nabla f(y_n)) \end{aligned} \tag{1}$$

$$Prox_{\gamma g}(x) = \left(Id + \partial \gamma g\right)^{-1}(x) = \arg\min\{g(y) + \frac{\|x-y\|^2}{2\gamma}\}$$

The differential setting

If g differentiable and $\mathcal{H}$ possibly infinite dimensional :

$$\text{(DI)} \quad \Leftrightarrow \quad \ddot{x}(t) + \frac{b}{t}\dot{x}(t) + \nabla F(x(t)) = 0 \qquad \text{(DE)}$$

Theorem (Asymptotic properties for (DE) : Su et al.'15, Attouch et al '15/'16/'17, May '16, Aujol et al'17)

If $b < 3$:

$$F(x(t)) - F(x^*) \sim O\left(t^{-\frac{2b}{3}}\right) \ , \quad \|\dot{x}(t)\| \sim O\left(t^{-p}\right), \forall p < \frac{2b}{3} \qquad (2)$$

If $b \geq 3$: $\quad F(x(t)) - F(x^*) \sim O\left(t^{-2}\right) \ , \quad \|\dot{x}(t)\| \sim O\left(t^{-1}\right)$

$$(3)$$

If $b > 3$:

$$F(x(t)) - F(x^*) \sim o\left(t^{-2}\right) \ , \quad \|\dot{x}(t)\| \sim o\left(t^{-1}\right) \ , \quad x(t) \rightharpoonup x^* \qquad (4)$$

The differential inclusion

For $t_0 > 0$, $b > 0$ and $t \geq t_0$:

$$\begin{cases} \ddot{x}(t) + \frac{b}{t}\dot{x}(t) + \partial F(x(t)) \ni 0 \\ x(t_0) = x_0 \quad \dot{x}(t_0) = v_0 \end{cases} \tag{DI}$$

General framework for (DI) (Paoli 2000)

$$\begin{cases} \ddot{x}(t) + \partial F(x(t)) \ni h(t, x(t), \dot{x})(t) \\ x(t_0) = x_0 \quad \dot{x}(t_0) = v_0 \end{cases} \tag{GDI}$$

where h is Lipschitz continuous in its last two arguments uniformly with respect to the first one.

Definition (Shock solution of (DI) Paoli '00, Attouch et al '02)

A function $x : [t_0, +\infty) \longrightarrow \mathbb{R}^d$ is an **energy-conserving shock solution** of (DI) if :

1. $x \in \mathcal{C}^{0,1}([t_0, T]; \mathbb{R}^d)$, for all $T > t_0$
2. $\dot{x} \in BV([t_0, T]; \mathbb{R}^d)$, for all $T > t_0$
3. $x(t) \in \mathcal{D}(F)$, for all $t \geq t_0$
4. For all $\phi \in \mathcal{C}_c^1([t_0, +\infty), \mathbb{R}^+)$ and $v \in \mathcal{C}([t_0, +\infty), \mathcal{D}(F))$, it holds :

$$\int_{t_0}^{T} (F(x(t)) - F(v(t))\phi(t)dt \leq \langle \ddot{x} + \frac{b}{t}\dot{x}, (v - x)\phi\rangle_{\mathcal{M}\times\mathcal{C}} \qquad (5)$$

5. x satisfies the following energy equation for a.e. $t \geq t_0$

$$F(x(t)) - F(x_0) + \frac{1}{2}\|\dot{x}(t)\|^2 - \frac{1}{2}\|v_0\|^2 + \int_{t_0}^{t} \frac{b}{s}\|\dot{x}(s)\|^2 ds = 0 \qquad (6)$$

Approximating ODE

We consider the Moreau-Yosida approximation of F, $\{F_\gamma\}_{\gamma>0}$ and the following approximating ODE :

Approximating ODE

$$\begin{aligned} &\ddot{x}_\gamma(t) + \frac{b}{t}\dot{x}_\gamma(t) + \nabla F_\gamma(x_\gamma(t)) = 0 \\ &x_\gamma(t_0) = x_0 \quad \dot{x}_\gamma(t_0) = v_0 \end{aligned} \tag{ADE}$$

(ADE) falls into the classical theory of differential equations and admits a unique solution $x_\gamma \in \mathcal{C}^2([t_0, +\infty); \mathbb{R}^d)$, for all $\gamma > 0$.

Existence of a shock solution

Theorem

Let $\{F_\gamma\}_{\gamma>0}$ *the Moreau-Yosida approximations of* F. *There exists a subsequence* $\{x_\gamma\}_{\gamma>0}$ *of solutions of* (ADE) *that converges to a shock solution of* (DI), *according to the following scheme :*

$$
\begin{aligned}
&\bullet \quad x_\gamma \underset{\gamma\to 0}{\longrightarrow} x \quad \text{uniformly on } [t_0, T], \forall T > t_0 \\
&\bullet \quad \dot{x}_\gamma \underset{\gamma\to 0}{\longrightarrow} \dot{x} \quad \text{in } L^p([t_0, T]; \mathbb{R}^d) \ , \forall p \in [1, +\infty) \ , \forall T > t_0 \\
&\bullet \quad F_\gamma(x_\gamma) \underset{\gamma\to 0}{\longrightarrow} F(x) \quad \text{in } L^p([t_0, T]; \mathbb{R}^d), \forall p \in [1, +\infty), \forall T > t_0
\end{aligned}
\qquad (\mathcal{AS})
$$

Lemma (A-priori estimates)

$$\sup_{\gamma>0}\{\|x_\gamma\|_\infty, \|\dot{x}_\gamma\|_\infty, \|\nabla F_\gamma(x_\gamma)\|_1, \|\ddot{x}_\gamma\|_1\} < +\infty$$

The case of $\mathcal{D}(F) = \mathbb{R}^d$

Lemma (2^{nd} order a-priori estimate)

If $\mathcal{D}(F) = \mathbb{R}^d$*, we additionally have :* $\sup\limits_{\gamma>0}\{\|\ddot{x}_\gamma\|_\infty\} < +\infty$

Corollary

If $\mathcal{D}(F) = \mathbb{R}^d$*, then the differential inclusion* (DI) *admits a shock solution* x *, such that :*

$$x \in W^{2,\infty}((t_0, T); \mathbb{R}^d) \cap \mathcal{C}^1([t_0, +\infty); \mathbb{R}^d), \text{ for all } T > t_0.$$

Asymptotic analysis

Let x be a shock solution of (DI) obtained as a limit of the approximation scheme $(\mathcal{AS})$ and x^* a minimizer of F.

Notation : $W(t) = F(x(t)) - F(x^*)$

For all λ, $\xi \geq 0$ we consider the Lyapunov function :

$$\mathcal{E}_{\lambda,\xi}(t) = t^2 W(t) + \frac{1}{2}\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 + \frac{\xi}{2}\|x(t) - x^*\|^2$$

Lemma

*For $b \geq 3$, $2 \leq \lambda \leq b-1$ and $\xi = \lambda(b - \lambda - 1)$, $\mathcal{E}_{\lambda,\xi}$ is **essentially non-increasing** in $[t_0, +\infty)$:*

$$\mathcal{E}_{\lambda,\xi}(t) \leq \mathcal{E}_{\lambda,\xi}(s) \ , \quad \text{for a.e. } t_0 \leq s \leq t$$

Convergence rates for $W(t)$ and $\dot{x}(t)$

Theorem (A. et al.'17)

Let x be a shock solution of (DI) *obtained as a limit of the approximation scheme* ($\mathcal{AS}$) *and x^* a minimizer of F. There exist $C_1, C_2 > 0$, s.t. :*

- *If $b \geq 3$:*

$$\sup_{t \geq t_0} \{\|x(t) - x^*\|\} < +\infty \tag{7}$$

$$W(t) \leq \frac{C_1}{t^2} \quad \textit{and} \quad \|\dot{x}(t)\| \leq \frac{C_2}{t} \quad \textbf{for a.e. } t \geq t_0 \tag{8}$$

- *If $b > 3$: $\int_{t_0}^{+\infty} tW(t)dt < +\infty$ and $\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty$*
  - *The trajectory $\{x(t)\}_{t \geq t_0}$ converges asymptotically to x^*.*
  - $$\textbf{ess} \lim_{t \to \infty} t^2 W(t) = 0 \quad \textit{and} \quad \textbf{ess} \lim_{t \to \infty} t\|\dot{x}(t)\| = 0 \tag{9}$$

The case of low friction $b < 3$

For $\lambda = \frac{2b}{3}$, $\xi = \frac{2b(3-b)}{9} > 0$ and $c = 2 - \frac{2b}{3} > 0$, for all $t \geq t_0$, we consider the energy-function :

$$\begin{aligned}\mathcal{H}(t) &= t^{-c}\mathcal{E}_{\lambda,\xi}(t)\\ &= t^{-c}\left(W(t) + \frac{1}{2}\|\frac{2b}{3}(x(t) - x^*) + t\dot{x}(t)\|^2 + \frac{2b(3-b)}{3}\|x(t) - x^*\|^2\right)\end{aligned}$$

Lemma

For $b \leq 3$, $\mathcal{H}$ is essentially non-increasing in $[t_0, +\infty)$

Corollary (A. et al.'17)

Let $b < 3$, x a shock solution of (DI) obtained as a limit of the approximation scheme ($\mathcal{AS}$) and x^* a minimizer of F.

$$W(t) \leq Ct^{-\frac{2b}{3}} \quad \text{for a.e. } t \geq t_0 \tag{10}$$

Bound estimates when $\mathcal{D}(F) = \mathbb{R}^d$

Let $x \in W^{2,\infty}((t_0, T); \mathbb{R}^d) \cap \mathcal{C}^1([t_0, +\infty); \mathbb{R}^d)$, for all $T > t_0$ a solution of (DI) with $\mathcal{D}(F) = \mathbb{R}^d$.

Key-property

$\dot{x}$, W, $\mathcal{E}_{\lambda,\xi}$, $\mathcal{H}$ are ***locally absolutely continuous functions***

Lemma (Generalized derivation rule (Brezis '73))

$$\frac{d}{dt}(F(x(t))) = \langle z, \dot{x}(t) \rangle \quad \forall z \in \partial F(x(t)) \quad a.e. \ in \ (t_0, T), \forall T > t_0$$

The energies $\mathcal{E}_{\lambda,\xi}$ and $\mathcal{H}$ are non-increasing in $[t_0, +\infty)$ (for suitable values of λ and ξ).

Bound estimates when $\mathcal{D}(F) = \mathbb{R}^d$

Corollary (A. et al.'17)

Let $x \in W^{2,\infty}((t_0, T); \mathbb{R}^d) \cap \mathcal{C}^1([t_0, +\infty); \mathbb{R}^d)$, for all $T > t_0$ be a solution of (DI) with $\mathcal{D}(F) = \mathbb{R}^d$. There exists $C > 0$, s.t. **for all $t \geq t_0$** :

- $b < 3$: $\quad W(t) \leq Ct^{-\frac{2b}{3}}$ and $\|\dot{x}(t)\| \leq (C + \sup_{t \geq t_0}\{\|x(t)\|\})t^{-\frac{b}{3}}$
- $b \geq 3$: $\quad \sup_{t \geq t_0}\{\|x(t)\|\} < +\infty$, $\quad W(t) \leq Ct^{-2}$, $\quad \|\dot{x}(t)\| \leq Ct^{-1}$
- $b > 3$: $\quad W(t) \sim o\left(t^{-2}\right)$, $\quad \|\dot{x}(t)\| \sim o\left(t^{-1}\right)$, $\quad x(t) \to x^*$

Proofs : Well adapted to the differential setting thanks to the absolute continuity property.

Optimal convergence rate for $W(t)$ when $b < 3$

We consider $F(x) = |x|$, for all $x \in \mathbb{R}$ and study (DI) for $b < 3$.

Theorem (A. et al.'17)

Let x be a solution of (DI), with $F(x) = |x|$ and $b < 3$ such that $x(t_0) \neq 0$. There exists a constant $K > 0$, such that for any $T > t_0$, there exists $t > T$ such that :

$$W(t) = |x(t)| \geq \frac{K}{t^{\frac{2b}{3}}} \tag{11}$$

We have that $x^* = 0$ and

$$W(t) = F(x(t)) - F(x^*) = |x(t)| = z(t)x(t) \quad \text{with } z(t) \in \partial F(x(t)) \tag{P1-H}$$

Lemma

Let $b < 3$ *and* x *a solution to* (DI) *with* $F(x) = |x|$, *such that* $x(t_0) = x_0 \neq 0$. *Then* $\exists K_1 > 0$ *s.t.* $\boldsymbol{\mathcal{H}(t) \geq K_1}$.

Proof : For $\lambda = \frac{2b}{3}$, $\xi = \frac{2b(3-b)}{9} > 0$ and $c = 2 - \frac{2b}{3} > 0$, for all $t \geq t_0$

$$\boldsymbol{\mathcal{H}(t) = t^{-c}(t^2|x(t)| + \frac{1}{2}|\lambda x(t) + t\dot{x}(t)|^2 + \frac{\xi}{2}|x(t)|^2)}$$

By differentiating $\mathcal{H}$ and using (P1-H) :

$$\begin{aligned} \dot{\mathcal{H}}(t) &= -\frac{2b(9-b^2)}{27} t^{-c-1}|x(t)|^2 \\ &\geq C_0 t^{c-5}\mathcal{H}(t) \quad \text{a.e. in } (t_0, +\infty) \end{aligned}$$

with $\boldsymbol{c = 2 - \frac{2b}{3} \in (0,2)}$.

It follows that : $\mathcal{H}(t) \geq K_1$, where $K_1 = \mathcal{H}(t_0)e^{C_0' t_0^{c-4}}$

Proof of Optimality : From Lemma :

$$t^2|x(t)| + \underbrace{\frac{1}{2}|\lambda x(t) + t\dot{x}(t)|^2 + \frac{\xi}{2}|x(t)|^2}_{=v(t)} = \mathcal{E}_{\lambda,\xi}(t) = \mathcal{H}(t)t^c \geq K_1 t^c \quad (12)$$

Let $T > t_0$. We can always find $t > T$, $\theta \in (0,1)$, such that :

$$v(t) \leq \theta K_1 t^c \quad (13)$$

Hence there exists $t > T$, $\theta \in (0,1)$, s.t. :

$$t^2 W(t) = t^2|x(t)| \geq \underbrace{(1-\theta)K_1}_{=K} t^c \quad (14)$$

□

Thank you for the Attention

For more details :

- Vassilis Apidopoulos, Jean-François Aujol and Charles Dossal. *The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case* $b \leq 3$*. Preprint, 2017*. https://hal.archives-ouvertes.fr/hal-01517708/
- Vassilis Apidopoulos, Jean-François Aujol and Charles Dossal. *Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule. Preprint, 2017*. https://hal.archives-ouvertes.fr/hal-01551873/

[1] Vassilis Apidopoulos, Jean-François Aujol and Charles Dossal, *The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case* $b \leq 3$, Preprint, 2017

[2] Hedy Attouch, Alexandre Cabot and Patrick Redont, *The dynamics of elastic shocks via epigraphical regularization of a differential inclusion. Barrier and penalty approximations*, Advances in Mathematical Sciences and Applications, 2002.

[3] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Mathematical Programming, 2016.

[4] Hedy Attouch, Zaki Chbani and Hassan Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, Preprint, 2017.

[5] Jean-François Aujol and Charles Dossal, *Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for* $b > 0$, Preprint, 2017.

[6] Amir Beck and Marc Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences, 2009.

[7] Haim Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, Lecture Notes 5, North Holland, 1973.

[8] Antonin Chambolle and Charles Dossal, *On the convergence of the iterates of the Fast Iterative Shrinkage/Thresholding Algorithm*, Journal of Optimization Theory and Applications, 2015.

[9] Ramzi May, *Asymptotic for a second order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Mathematics, 2017.

[10] Yurii Nesterov, *A method for solving the convex programming problem with convergence rate* $O(k^{-2})$, Dokl. Akad. Nauk SSSR, 269(3):543–547, 1983.

[11] Yurii Nesterov, *Introductory lectures on convex optimization: A basic course 2*, Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[12] Laetitia Paoli, *An existence result for vibrations with unilateral constraints: case of a nonsmooth set of constraints*, Mathematical Models and Methods in Applied Sciences, 2000.

[13] Michelle Schatzman, *A class of nonlinear differential equations of second order in time*, Nonlinear Analysis: Theory, Methods & Applications, 1978.

[14] Weijie Su, Stephen Boyd and Emmanuel J. Candes, *A differential equation for modeling Nesterov?s accelerated gradient method: theory and insights*, Journal of Machine Learning Research, 2016.