Recent advances on the acceleration of first-order methods in convex optimization

Juan PEYPOUQUET
Universidad Técnica Federico Santa María

Second Workshop on Algorithms and Dynamics for Games and Optimization

Santiago, January 25, 2016

Content

Content

- Basic first-order descent methods
- Nesterov's acceleration
- Dynamic interpretation
  - Damped Inertial Gradient System (DIGS)
- Properties of DIGS trajectories and accelerated algorithms
- A first-order variant bearing second-order information in time and space

Content

- Basic first-order descent methods
- Nesterov's acceleration
- Dynamic interpretation
  - Damped Inertial Gradient System (DIGS)
- Properties of DIGS trajectories and accelerated algorithms
- A first-order variant bearing second-order information in time and space

Content

- Basic first-order descent methods
- Nesterov's acceleration
- Dynamic interpretation
  - Damped Inertial Gradient System (DIGS)
- Properties of DIGS trajectories and accelerated algorithms
- A first-order variant bearing second-order information in time and space

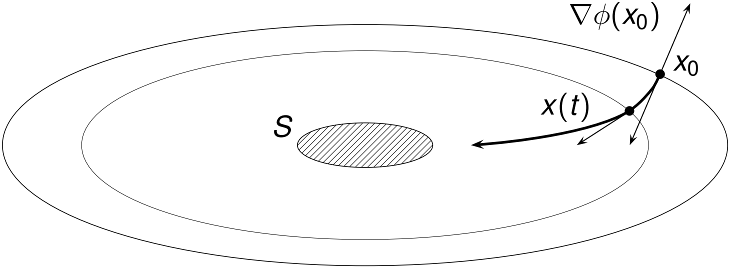
Content

- Basic first-order descent methods
- Nesterov's acceleration
- Dynamic interpretation
  - Damped Inertial Gradient System (DIGS)
- Properties of DIGS trajectories and accelerated algorithms
- A first-order variant bearing second-order information in time and space

BASIC DESCENT METHODS

Basic (first-order) descent methods

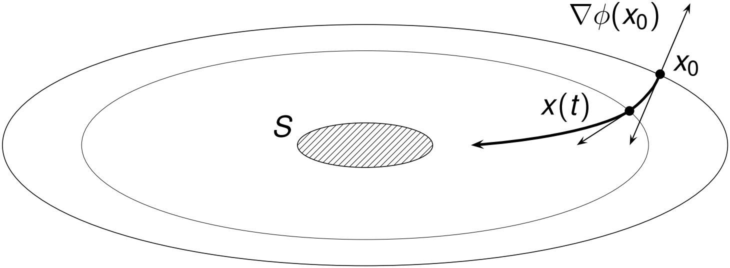
Steepest descent dynamics: $\dot{x}(t) = -\nabla\phi(x(t))$, $x(0) = x_0$



$$\frac{d}{dt}\phi(x(t)) = \langle \nabla\phi(x(t)), \dot{x}(t)\rangle = -\|\nabla\phi(x(t))\|^2 = -\|\dot{x}(t)\|^2$$

Basic (first-order) descent methods

Steepest descent dynamics: $\dot{x}(t) = -\nabla\phi(x(t))$, $x(0) = x_0$



$$\frac{d}{dt}\phi(x(t)) = \langle \nabla\phi(x(t)), \dot{x}(t)\rangle = -\|\nabla\phi(x(t))\|^2 = -\|\dot{x}(t)\|^2$$

Basic (first-order) descent methods

Explicit discretization → gradient method (Cauchy 1847):

$$\frac{x_{k+1} - x_k}{\lambda} = -\nabla\phi(x_k) \quad \iff \quad x_{k+1} = x_k - \lambda\nabla\phi(x_k).$$

Implicit discretization → proximal method (Martinet 1970):

$$\frac{z_{k+1} - z_k}{\lambda} = -\nabla\phi(z_{k+1}) \quad \iff \quad z_{k+1} + \lambda\nabla\phi(z_{k+1}) = z_k.$$

Basic (first-order) descent methods

Explicit discretization → gradient method (Cauchy 1847):

$$\frac{x_{k+1} - x_k}{\lambda} = -\nabla\phi(x_k) \quad \Longleftrightarrow \quad x_{k+1} = x_k - \lambda\nabla\phi(x_k).$$

Implicit discretization → proximal method (Martinet 1970):

$$\frac{z_{k+1} - z_k}{\lambda} = -\nabla\phi(z_{k+1}) \quad \Longleftrightarrow \quad z_{k+1} + \lambda\nabla\phi(z_{k+1}) = z_k.$$

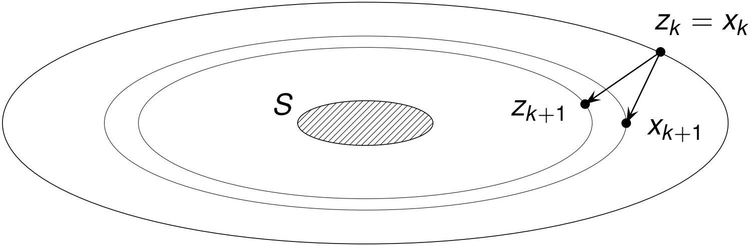
Basic (first-order) descent methods

Gradient

$x_{k+1} = x_k - \lambda \nabla \phi(x_k)$

Proximal

$z_{k+1} + \lambda \nabla \phi(z_{k+1}) = z_k$



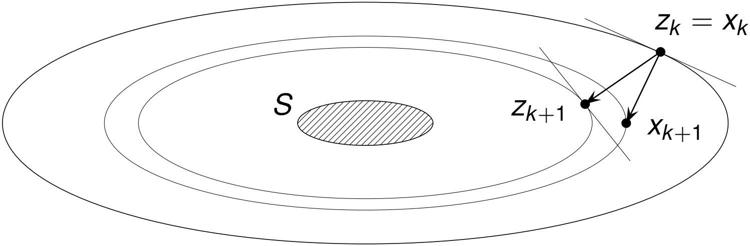
Basic (first-order) descent methods

Gradient

$x_{k+1} = x_k - \lambda \nabla \phi(x_k)$

Proximal

$z_{k+1} + \lambda \nabla \phi(z_{k+1}) = z_k$



Pros and cons

Gradient method

+ Lower computational cost per iteration (explicit formula), easy implementation
– Convergence depends strongly on the regularity of the function (typically $\phi \in \mathcal{C}^{1,1}$) and on the step sizes

Proximal point algorithm

+ More stability, convergence certificate for a larger class of functions ($\nabla\phi \to \partial\phi$), independent of the step size
– Higher computational cost per iteration (implicit formula), often requires inexact computation

Pros and cons

Gradient method

+ Lower computational cost per iteration (explicit formula), easy implementation
– Convergence depends strongly on the regularity of the function (typically $\phi \in \mathcal{C}^{1,1}$) and on the step sizes

Proximal point algorithm

+ More stability, convergence certificate for a larger class of functions ($\nabla\phi \to \partial\phi$), independent of the step size
– Higher computational cost per iteration (implicit formula), often requires inexact computation

Combining smooth and nonsmooth functions

Problem

$$\min\{\Phi(x) := F(x) + G(x) : x \in H\},$$

where F is not smooth but G is.

Forward-Backward Method ($x_k \to x_{k+\frac{1}{2}} \to x_{k+1}$)

$$x_{k+1} + \lambda\partial F(x_{k+1}) \ni x_{k+\frac{1}{2}} = x_k - \lambda\nabla G(x_k)$$

$$x_{k+1} = \mathrm{Prox}_{\lambda F} \circ \mathrm{Grad}_{\lambda G}(x_k)$$

Combining smooth and nonsmooth functions

Problem

$$\min\{\Phi(x) := F(x) + G(x) : x \in H\},$$

where F is not smooth but G is.

Forward-Backward Method ($x_k \to x_{k+\frac{1}{2}} \to x_{k+1}$)

$$x_{k+1} + \lambda \partial F(x_{k+1}) \ni x_{k+\frac{1}{2}} = x_k - \lambda \nabla G(x_k)$$

$$x_{k+1} = \mathrm{Prox}_{\lambda F} \circ \mathrm{Grad}_{\lambda G}(x_k)$$

Combining smooth and nonsmooth functions

Problem

$$\min\{\Phi(x) := F(x) + G(x) : x \in H\},$$

where F is not smooth but G is.

Forward-Backward Method ($x_k \to x_{k+\frac{1}{2}} \to x_{k+1}$)

$$x_{k+1} + \lambda \partial F(x_{k+1}) \ni x_{k+\frac{1}{2}} = x_k - \lambda \nabla G(x_k)$$

$$x_{k+1} = \mathrm{Prox}_{\lambda F} \circ \mathrm{Grad}_{\lambda G}(x_k)$$

Combining smooth and nonsmooth functions

Gradient projection:
Goldstein 1964, Levitin-Polyak 1966, with $F = \delta_C$

General setting:
Lions-Mercier 1979, Passty 1979

Iterative Shrinkage-Thresholding Algorithm (ISTA):
Daubechies-Defrise-DeMol 2004, Combettes-Wajs 2005, for "$\ell^1 + \ell^2$" minimization

$$\Phi(x) = F(x) + G(x) = \mu \|x\|_1 + \frac{1}{2}\|Ax - b\|^2$$

Combining smooth and nonsmooth functions

Gradient projection:
Goldstein 1964, Levitin-Polyak 1966, with $F = \delta_C$

General setting:
Lions-Mercier 1979, Passty 1979

Iterative Shrinkage-Thresholding Algorithm (ISTA):
Daubechies-Defrise-DeMol 2004, Combettes-Wajs 2005, for "$\ell^1 + \ell^2$" minimization

$$\Phi(x) = F(x) + G(x) = \mu\|x\|_1 + \frac{1}{2}\|Ax - b\|^2$$

Combining smooth and nonsmooth functions

Gradient projection:

Goldstein 1964, Levitin-Polyak 1966, with $F = \delta_C$

General setting:

Lions-Mercier 1979, Passty 1979

Iterative Shrinkage-Thresholding Algorithm (ISTA):

Daubechies-Defrise-DeMol 2004, Combettes-Wajs 2005, for "$\ell^1 + \ell^2$" minimization

$$\Phi(x) = F(x) + G(x) = \mu\|x\|_1 + \frac{1}{2}\|Ax - b\|^2$$

Convergence of the forward-backward method

Theorem

Let $\Phi = F + G$, where G is closed and convex, and F is convex with ∇F L-Lipschitz. Assume Φ has minimizers, and let (x_k) be obtained by the FB method with $\lambda \leq 1/L$. Then

- *As $k \to \infty$, (x_k) converges* to a minimizer of Φ; and*
- *$\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-1})$: There is $C > 0$ such that*

$$\Phi(x_k) - \min \Phi \leq \frac{C}{k}.$$

Convergence of the forward-backward method

Theorem

Let $\Phi = F + G$, where G is closed and convex, and F is convex with ∇F L-Lipschitz. Assume Φ has minimizers, and let (x_k) be obtained by the FB method with $\lambda \leq 1/L$. Then

- *As $k \to \infty$, (x_k) converges* to a minimizer of Φ; and*
- $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-1})$*: There is $C > 0$ such that*

$$\Phi(x_k) - \min \Phi \leq \frac{C}{k}.$$

Convergence of the forward-backward method

Theorem

Let $\Phi = F + G$*, where* G *is closed and convex, and* F *is convex with* ∇F L*-Lipschitz. Assume* Φ *has minimizers, and let* (x_k) *be obtained by the FB method with* $\lambda \leq 1/L$*. Then*

- *As* $k \to \infty$*,* (x_k) *converges** *to a minimizer of* Φ*; and*
- $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-1})$*: There is* $C > 0$ *such that*

$$\Phi(x_k) - \min \Phi \leq \frac{C}{k}.$$

Convergence ISTA

Let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be defined by

$$\Phi(x) = \|x\|_1 + \frac{1}{2}\|Ax - b\|^2.$$

Local linear convergence results have been found recently, as well as theoretical convergence rates.

Theorem (Bolte-Nguyen-P.-Suter 2015)

Let (x_k) be obtained by the FB method with step size λ. Then, there is an explicit constant d such that

$$\Phi(x_k) - \min \Phi \leq \frac{\Phi(x_0) - \min \Phi}{(1 + d\lambda)^{2k}}.$$

Convergence ISTA

Let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be defined by

$$\Phi(x) = \|x\|_1 + \frac{1}{2}\|Ax - b\|^2.$$

Local linear convergence results have been found recently, as well as theoretical convergence rates.

Theorem (Bolte-Nguyen-P.-Suter 2015)

Let (x_k) be obtained by the FB method with step size λ. Then, there is an explicit constant d such that

$$\Phi(x_k) - \min\Phi \leq \frac{\Phi(x_0) - \min\Phi}{(1 + d\lambda)^{2k}}.$$

Convergence ISTA

Let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be defined by

$$\Phi(x) = \|x\|_1 + \frac{1}{2}\|Ax - b\|^2.$$

Local linear convergence results have been found recently, as well as theoretical convergence rates.

Theorem (Bolte-Nguyen-P.-Suter 2015)

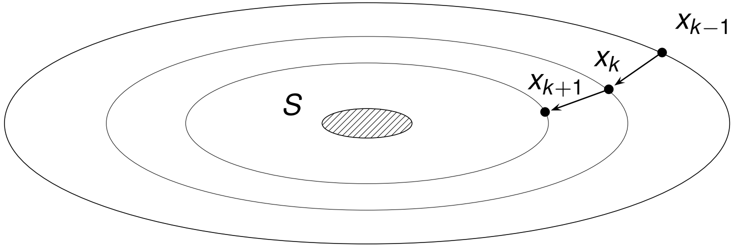
Let (x_k) be obtained by the FB method with step size λ. Then, there is an explicit constant d such that

$$\Phi(x_k) - \min \Phi \leq \frac{\Phi(x_0) - \min \Phi}{(1 + d\lambda)^{2k}}.$$

NESTEROV'S ACCELERATION

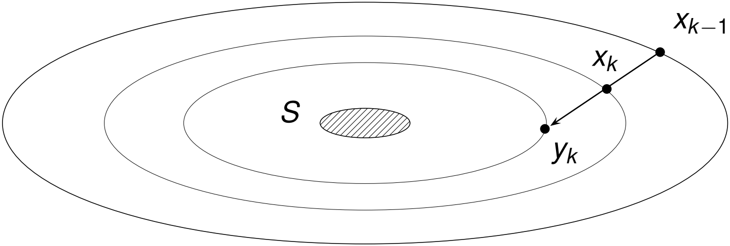
Acceleration

The main idea is the following: Instead of doing



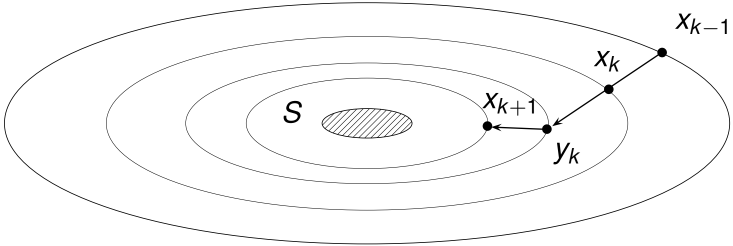
Acceleration

Better try



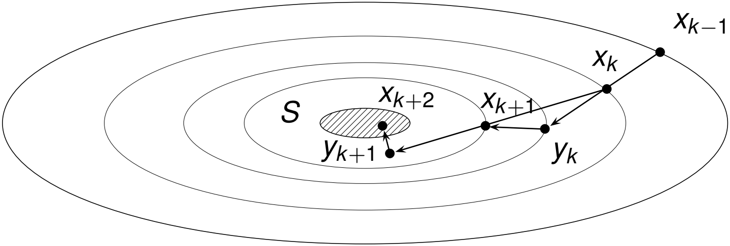
Acceleration

Better try



Acceleration

Better try



Some remarks

- Convergence and its rate are sensitive to the choice of y_k
- This simple procedure (Nesterov 1983) can take the theoretical rate of worst-case convergence for the values from the typical $\mathcal{O}(1/k)$ down to $\mathcal{O}(1/k^2)$
- No convergence proof for the iterates x_k
- Current common practice is

$$y_k = x_k + \left(1 - \frac{3}{k}\right)(x_k - x_{k-1})$$

Keynote example in image processing: FISTA (Beck-Teboulle 2009)

Some remarks

- Convergence and its rate are sensitive to the choice of y_k
- This simple procedure (Nesterov 1983) can take the theoretical rate of worst-case convergence for the values from the typical $\mathcal{O}(1/k)$ down to $\mathcal{O}(1/k^2)$
- No convergence proof for the iterates x_k
- Current common practice is

$$y_k = x_k + \left(1 - \frac{3}{k}\right)(x_k - x_{k-1})$$

Keynote example in image processing: FISTA (Beck-Teboulle 2009)

Some remarks

- Convergence and its rate are sensitive to the choice of y_k
- This simple procedure (Nesterov 1983) can take the theoretical rate of worst-case convergence for the values from the typical $\mathcal{O}(1/k)$ down to $\mathcal{O}(1/k^2)$
- No convergence proof for the iterates x_k
- Current common practice is

$$y_k = x_k + \left(1 - \frac{3}{k}\right)(x_k - x_{k-1})$$

Keynote example in image processing: FISTA (Beck-Teboulle 2009)

Some remarks

- Convergence and its rate are sensitive to the choice of y_k
- This simple procedure (Nesterov 1983) can take the theoretical rate of worst-case convergence for the values from the typical $\mathcal{O}(1/k)$ down to $\mathcal{O}(1/k^2)$
- No convergence proof for the iterates x_k
- Current common practice is

$$y_k = x_k + \left(1 - \frac{3}{k}\right)(x_k - x_{k-1})$$

Keynote example in image processing: FISTA (Beck-Teboulle 2009)

ISTA & FISTA

General case:

- FB: values $\mathcal{O}(k^{-1})$, convergent sequence.
- AFB: values $\mathcal{O}(k^{-2})$.

$\ell^1 + \ell^2$ minimization:

- ISTA: values $\mathcal{O}(Q^k)$, convergent sequence (proved).
- FISTA: values (observed, not proved) $\mathcal{O}(\tilde{Q}^k)$, always strictly faster than ISTA, convergent sequence (observed, not proved).

ISTA & FISTA

General case:

- FB: values $\mathcal{O}(k^{-1})$, convergent sequence.
- AFB: values $\mathcal{O}(k^{-2})$.

$\ell^1 + \ell^2$ minimization:

- ISTA: values $\mathcal{O}(Q^k)$, convergent sequence (proved).
- FISTA: values (observed, not proved) $\mathcal{O}(\tilde{Q}^k)$, always strictly faster than ISTA, convergent sequence (observed, not proved).

Long-standing questions

- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(\tilde{Q}^k)$ true for FISTA $(\ell^1 + \ell^2)$?
- Is AFB always strictly faster than FB?
- What about FISTA and ISTA?
- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$ optimal for AFB (in general)?
- Are AFB sequences convergent?
- What about FISTA?

Long-standing questions

- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(\tilde{Q}^k)$ true for FISTA $(\ell^1 + \ell^2)$?
- Is AFB always strictly faster than FB?
- What about FISTA and ISTA?
- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$ optimal for AFB (in general)?
- Are AFB sequences convergent?
- What about FISTA?

Long-standing questions

- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(\tilde{Q}^k)$ true for FISTA $(\ell^1 + \ell^2)$?
- Is AFB always strictly faster than FB?
- What about FISTA and ISTA?
- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$ optimal for AFB (in general)?
- Are AFB sequences convergent?
- What about FISTA?

Long-standing questions

- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(\tilde{Q}^k)$ true for FISTA $(\ell^1 + \ell^2)$?
- Is AFB always strictly faster than FB?
- What about FISTA and ISTA?
- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$ optimal for AFB (in general)?
- Are AFB sequences convergent?
- What about FISTA?

Long-standing questions

- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(\tilde{Q}^k)$ true for FISTA ($\ell^1 + \ell^2$)?
- Is AFB always strictly faster than FB?
- What about FISTA and ISTA?
- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$ optimal for AFB (in general)?
- Are AFB sequences convergent?
- What about FISTA?

Long-standing questions

- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(\tilde{Q}^k)$ true for FISTA $(\ell^1 + \ell^2)$?
- Is AFB always strictly faster than FB?
- What about FISTA and ISTA?
- Is $\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$ optimal for AFB (in general)?
- Are AFB sequences convergent?
- What about FISTA?

Dynamic Interpretation

Discretization of DIGS

A finite-difference discretization of

$$(DIGS) \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial F(x(t)) + \nabla G(x(t)) \ni 0.$$

gives

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{kh^2}(x_k - x_{k-1}) + \partial F(x_{k+1}) + \nabla G(y_k) \ni 0,$$

where y_k (specified later) is related to the segment $[x_{k-1}, x_k]$.

Discretization of DIGS

Rewriting

$$\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})+\frac{\alpha}{kh^2}(x_k-x_{k-1})+\partial F(x_{k+1})+\nabla G(y_k)\ni 0,$$

with $\lambda = h^2$, we obtain

$$x_{k+1}+\lambda\partial F(x_{k+1})\ni x_k+\left(1-\frac{\alpha}{k}\right)(x_k-x_{k-1})-\lambda\nabla G(y_k).$$

Thus, if we set $y_k = x_k + \left(1-\frac{\alpha}{k}\right)(x_k - x_{k-1})$, we obtain

$$x_{k+1}+\lambda\partial F(x_{k+1})\ni y_k-\lambda\nabla G(y_k).$$

Discretization of DIGS

Rewriting

$$\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})+\frac{\alpha}{kh^2}(x_k-x_{k-1})+\partial F(x_{k+1})+\nabla G(y_k)\ni 0,$$

with $\lambda = h^2$, we obtain

$$x_{k+1}+\lambda\partial F(x_{k+1})\ni x_k+\left(1-\frac{\alpha}{k}\right)(x_k-x_{k-1})-\lambda\nabla G(y_k).$$

Thus, if we set $y_k = x_k + \left(1-\frac{\alpha}{k}\right)(x_k - x_{k-1})$, we obtain

$$x_{k+1}+\lambda\partial F(x_{k+1})\ni y_k-\lambda\nabla G(y_k).$$

Discretization of DIGS

Therefore, a finite-difference discretization of

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial F(x(t)) + \nabla G(x(t)) \ni 0.$$

naturally yields

$$\begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= \mathrm{Prox}_{\lambda F} \circ \mathrm{Grad}_{\lambda G}(y_k) \end{cases}$$

Construction due to Su-Boyd-Candès 2014.

Discretization of DIGS

Therefore, a finite-difference discretization of

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial F(x(t)) + \nabla G(x(t)) \ni 0.$$

naturally yields

$$\begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= \mathrm{Prox}_{\lambda F} \circ \mathrm{Grad}_{\lambda G}(y_k) \end{cases}$$

Construction due to Su-Boyd-Candès 2014.

Properties of DIGS Trajectories

Basic properties

Theorem (Attouch-Chbani-P.-Redont 2015)

If $\alpha > 0$, then

- $\lim_{t \to +\infty} \Phi(x(t)) = \inf(\Phi) \in \mathbb{R} \cup \{-\infty\}$.
- *Every weak limit point of $x(t)$, as $t \to \infty$, minimizes Φ.*
- *Either Φ has minimizers and all trajectories are bounded, or it does not and all trajectories diverge to $+\infty$ in norm.*
- *If Φ is bounded from below, then $\lim_{t \to +\infty} \|\dot{x}(t)\| = 0$.*

Rate of convergence

Theorem (Su-Boyd-Candès 2014)

If $\alpha \geq 3$ and Φ has minimizers, then every solution satisfies

$$\Phi(x(t)) - \min(\Phi) \leq \frac{C}{t^2},$$

where C depends on α and the initial data.

Rate of convergence

The exponent 2 is sharp. More precisely, we have the following:

Theorem (ACPR)

For each $p > 2$, there is Φ such that Φ has minimizers and every solution satisfies

$$\Phi(x(t)) - \min(\Phi) = \frac{C}{t^p}.$$

Rate of convergence

If Φ is strongly convex, convergence is arbitrarily fast, as α grows.

Theorem (ACPR)

Let Φ be strongly convex and let x^* be its unique minimizer. Every solution satisfies

$$\Phi(x(t)) - \min(\Phi) \leq \frac{C}{t^{\frac{2}{3}\alpha}} \qquad \text{and} \qquad \|x(t) - x^*\| \leq \frac{D}{t^{\frac{1}{3}\alpha}},$$

where C and D depend on α, the strong convexity parameter and the initial data.

Convergence of the solutions

Theorem (ACPR, May)

If $\alpha > 3$ and Φ has minimizers, then

- *$x(t)$ converges weakly, as $t \to +\infty$, to a minimizer of Φ.*
- *Convergence is strong if either Φ is uniformly convex, int(Argmin(Φ)) $\neq \emptyset$, or Φ is even.*
- *$\|\dot{x}(t)\| = o(t^{-1})$.*
- *$\Phi(x(t)) - \min(\Phi) = o(t^{-2})$.*

Convergence of the solutions

Theorem (ACPR, May)

If $\alpha > 3$ and Φ has minimizers, then

- *$x(t)$ converges weakly, as $t \to +\infty$, to a minimizer of Φ.*
- *Convergence is strong if either Φ is uniformly convex, int(Argmin(Φ)) $\neq \emptyset$, or Φ is even.*
- $\|\dot{x}(t)\| = o(t^{-1})$.
- $\Phi(x(t)) - \min(\Phi) = o(t^{-2})$.

Convergence of the solutions

Theorem (ACPR, May)

If $\alpha > 3$ and Φ has minimizers, then

- $x(t)$ *converges weakly, as* $t \to +\infty$, *to a minimizer of* Φ.
- *Convergence is strong if either* Φ *is uniformly convex,* $int(Argmin(\Phi)) \neq \emptyset$, *or* Φ *is even.*
- $\|\dot{x}(t)\| = o(t^{-1})$.
- $\Phi(x(t)) - \min(\Phi) = o(t^{-2})$.

Convergence of the solutions

Theorem (ACPR, May)

If $\alpha > 3$ and Φ has minimizers, then

- *$x(t)$ converges weakly, as $t \to +\infty$, to a minimizer of Φ.*
- *Convergence is strong if either Φ is uniformly convex, int(Argmin(Φ)) $\neq \emptyset$, or Φ is even.*
- $\|\dot{x}(t)\| = o(t^{-1})$.
- $\Phi(x(t)) - \min(\Phi) = o(t^{-2})$.

PROPERTIES OF ACCELERATED ALGORITHMS

Back to accelerated algorithms

Recall that

$$\begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= \mathrm{Prox}_{\lambda F} \circ \mathrm{Grad}_{\lambda G}(y_k) \end{cases}$$

Theorem (ACPR)

If $\alpha > 0$, then

- $\lim_{k \to +\infty} \Phi(x_k) = \inf(\Phi)$*; and*
- *every weak limit point of x_k, as $k \to +\infty$, minimizes Φ.*

Back to accelerated algorithms

Recall that

$$\begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \text{Prox}_{\lambda F} \circ \text{Grad}_{\lambda G}(y_k) \end{cases}$$

Theorem (ACPR)

If $\alpha > 0$, then

- $\lim_{k \to +\infty} \Phi(x_k) = \inf(\Phi)$; *and*
- *every weak limit point of x_k, as $k \to +\infty$, minimizes Φ.*

Back to accelerated algorithms

Theorem (ACPR)

If $\alpha \geq 3$ *and* Φ *has minimizers, then*

$$\Phi(x_k) - \min \Phi = \mathcal{O}(k^{-2})$$

and

$$\|x_k - x_{k-1}\| = \mathcal{O}(k^{-1}).$$

Back to accelerated algorithms

Theorem (ACPR,AP)

If $\alpha > 3$ and Φ has minimizers, then:

- *x_k converges weakly, as $k \to +\infty$, to a minimizer of Φ.*
- *Strong convergence holds if Φ is even, uniformly convex, or if Argmin(Φ) has nonempty interior.*
- *$\|x_k - x_{k-1}\| = o(k^{-1})$.*
- *$\Phi(x_k) - \min \Phi = o(k^{-2})$.*

Back to accelerated algorithms

Theorem (ACPR,AP)

If $\alpha > 3$ and Φ has minimizers, then:

- x_k *converges weakly, as* $k \to +\infty$, *to a minimizer of* Φ.
- *Strong convergence holds if* Φ *is even, uniformly convex, or if* $Argmin(\Phi)$ *has nonempty interior.*
- $\|x_k - x_{k-1}\| = o(k^{-1})$.
- $\Phi(x_k) - \min \Phi = o(k^{-2})$.

Back to accelerated algorithms

Theorem (ACPR,AP)

If $\alpha > 3$ and Φ has minimizers, then:

- x_k *converges weakly, as* $k \to +\infty$, *to a minimizer of* Φ.
- *Strong convergence holds if* Φ *is even, uniformly convex, or if Argmin*(Φ) *has nonempty interior.*
- $\|x_k - x_{k-1}\| = o(k^{-1})$.
- $\Phi(x_k) - \min \Phi = o(k^{-2})$.

Back to accelerated algorithms

Theorem (ACPR,AP)

If $\alpha > 3$ and Φ has minimizers, then:

- x_k *converges weakly, as* $k \to +\infty$, *to a minimizer of* Φ.
- *Strong convergence holds if* Φ *is even, uniformly convex, or if* $Argmin(\Phi)$ *has nonempty interior.*
- $\|x_k - x_{k-1}\| = o(k^{-1})$.
- $\Phi(x_k) - \min \Phi = o(k^{-2})$.

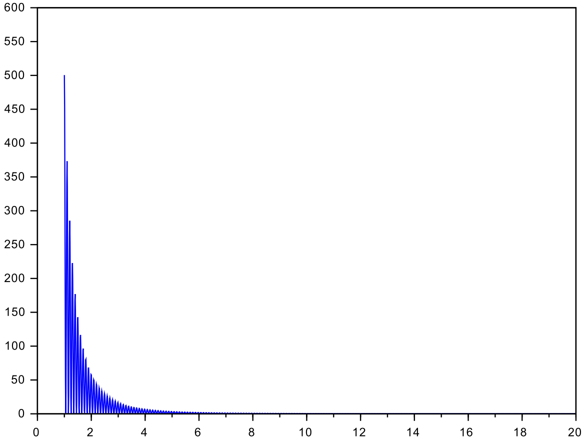
A simple example

We consider the function $\Phi(x_1, x_2) = \frac{1}{2}(x_1^2 + 1000x_2^2)$. We show the behavior of a solution to

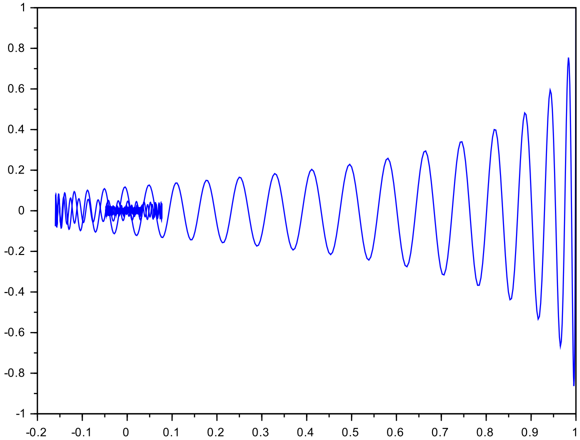
$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\Phi(x(t)) = 0$$

on the interval $[1, 20]$ with $\alpha = 3.1$.

Function values



Trajectory



Can We Do Better?

Idea: Newton / Levenberg-Marquardt

Pros:

- Is fast.
- Compensates the effect of ill-conditioning.

Cons:

- Requires higher regularity (to compute and invert the Hessian).
- Is costly to implement.

Idea: Newton / Levenberg-Marquardt

Pros:

- Is fast.
- Compensates the effect of ill-conditioning.

Cons:

- Requires higher regularity (to compute and invert the Hessian).
- Is costly to implement.

NDIGS

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

Seems much more complicated, but

Proposition (APR 2015)

System (NDIGS) is equivalent to

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

NDIGS

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

Seems much more complicated, but

Proposition (APR 2015)

System (NDIGS) is equivalent to

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

NDIGS

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

Seems much more complicated, but

Proposition (APR 2015)

System (NDIGS) is equivalent to

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

Nonsmooth functions

Using variable $Z = (x, y)$, this is

$$\dot{Z}(t) + \nabla \mathcal{G}(Z(t)) + D(t, Z(t)) \ni 0,$$

where $\mathcal{G}(Z) = \beta \Phi(x)$ and D is a regular linear perturbation.

So, we can consider

$$(NDIGS') \qquad \dot{Z}(t) + \partial \mathcal{G}(Z(t)) + D(t, Z(t)) \ni 0,$$

for nondifferentiable Φ.

Nonsmooth functions

Using variable $Z = (x, y)$, this is

$$\dot{Z}(t) + \nabla \mathcal{G}(Z(t)) + D(t, Z(t)) \ni 0,$$

where $\mathcal{G}(Z) = \beta\Phi(x)$ and D is a regular linear perturbation.

So, we can consider

$$(NDIGS') \qquad \dot{Z}(t) + \partial \mathcal{G}(Z(t)) + D(t, Z(t)) \ni 0,$$

for nondifferentiable Φ.

Convergence results

Theorem (APR)

Let Φ be closed and convex, and let $\beta > 0$.

- *All the conclusions obtained for the solutions of (DIGS) are also true for the solutions of (NDIGS').*
- *But also $\lim_{t\to\infty} \|\nabla\Phi(x(t))\| = 0$.*
- *If $\nabla\Phi$ is locally Lipschitz-continuous, then $\lim_{t\to\infty} \|\ddot{x}(t)\| = 0$.*

Convergence results

Theorem (APR)

Let Φ be closed and convex, and let $\beta > 0$.

- *All the conclusions obtained for the solutions of (DIGS) are also true for the solutions of (NDIGS').*
- *But also* $\lim_{t\to\infty} \|\nabla\Phi(x(t))\| = 0$.
- *If $\nabla\Phi$ is locally Lipschitz-continuous, then* $\lim_{t\to\infty} \|\ddot{x}(t)\| = 0$.

Convergence results

Theorem (APR)

Let Φ be closed and convex, and let $\beta > 0$.

- *All the conclusions obtained for the solutions of (DIGS) are also true for the solutions of (NDIGS').*
- *But also* $\lim_{t\to\infty} \|\nabla\Phi(x(t))\| = 0$.
- *If $\nabla\Phi$ is locally Lipschitz-continuous, then* $\lim_{t\to\infty} \|\ddot{x}(t)\| = 0$.

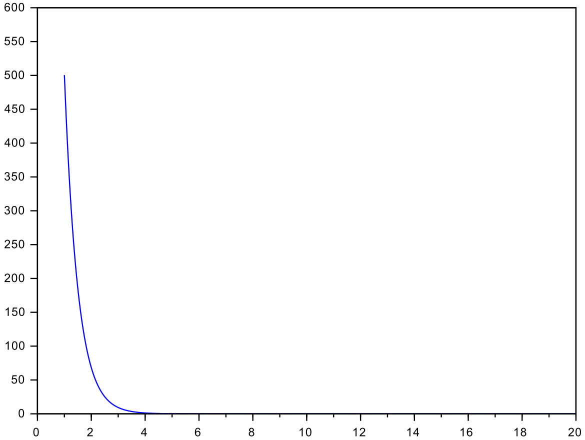
A simple example

We consider the function $\Phi(x_1, x_2) = \frac{1}{2}(x_1^2 + 1000x_2^2)$. We show the behavior of a solution to

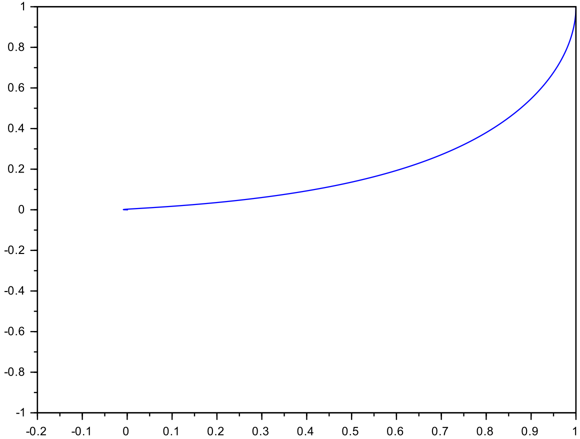
$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0$$

on the interval $[1, 20]$ with $\alpha = 3.1$ and $\beta = 1$.

Function values



Trajectory



Algorithmic implementation

Several discretizations are possible, giving different iterative algorithms.

Conjecture (Work in progress)

An appropriate discretization defines an algorithm with the same convergence properties as the continuous-time system (NDIGS').

Algorithmic implementation

Several discretizations are possible, giving different iterative algorithms.

Conjecture (Work in progress)

An appropriate discretization defines an algorithm with the same convergence properties as the continuous-time system (NDIGS').