iPiano: Inertial Proximal Algorithm for Non-Convex Optimization

Peter Ochs

University of Freiburg
Germany

17.01.2017

joint work with: Thomas Brox and Thomas Pock

Overview

What can you expect from this talk:

- First order optimization algorithms.
- Motivation from computer vision, but results are abstract/not application specific.
- Main focus is on certain non-smooth non-convex optimization problems.
- Non-smooth analysis is required for the details.
  For intuition, smooth analysis is sufficient.

Overview:

- Motivation for inertial methods.
- Algorithm for a class of non-smooth non-convex optimization problems: iPiano.
- Application examples.
- Convergence analysis.

Gradient descent dynamical system

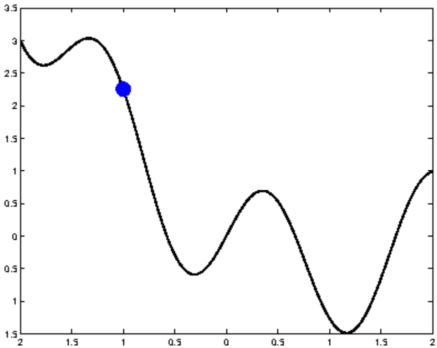
- Smooth optimization problem:

$$\min_{x\in\mathbb{R}^N} f(x)\,.$$

- Consider the (time-continuous) **gradient descent dynamical system**

$$\dot{X}(t) = -\nabla f(X(t))\,.$$

- Solution is a curve $X\colon\ [0,+\infty) \to \mathbb{R}^N$ with time-derivative $\dot{X}(t)$.
- Objective values are non-increasing.



Gradient descent dynamical system

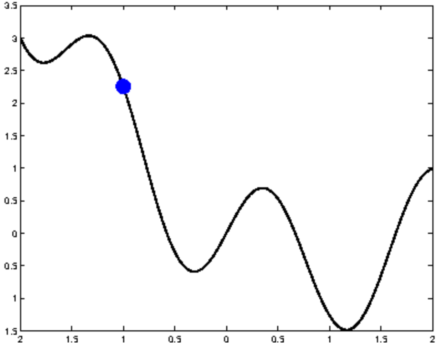
- Smooth optimization problem:

$$\min_{x\in\mathbb{R}^N} f(x)\,.$$

- Consider the (time-continuous) **gradient descent dynamical system**

$$\dot{X}(t) = -\nabla f(X(t))\,.$$

- Solution is a curve $X\colon\ [0,+\infty) \to \mathbb{R}^N$ with time-derivative $\dot{X}(t)$.
- Objective values are non-increasing.

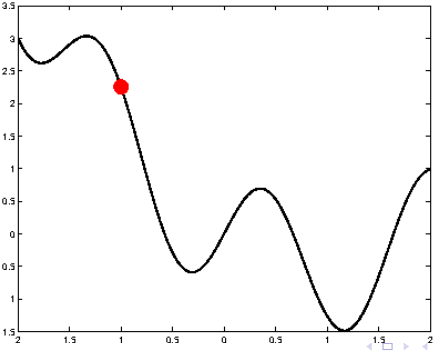


Heavy-ball dynamical system

- **Heavy-ball dynamical system**:

$$\ddot{X}(t) = -\gamma\dot{X}(t) - \nabla f(X(t))$$

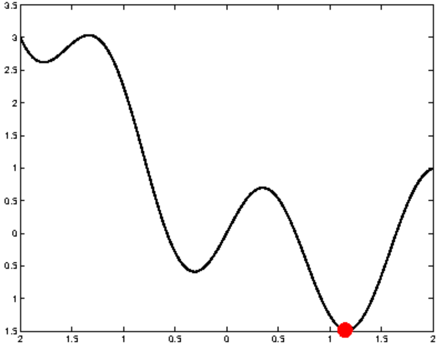
- The system describes the motion of a ball on the graph of the objective function f.
- $\ddot{X}(t)$ is the second derivative ($\sim$ acceleration). $\rightsquigarrow$ models **inertia** / **momentum**.
- $-\gamma\dot{X}$ is a viscous friction force ($\gamma > 0$).



- **Heavy-ball dynamical system**:

$$\ddot{X}(t) = -\gamma \dot{X}(t) - \nabla f(X(t))$$

- The system describes the motion of a ball on the graph of the objective function f.
- $\ddot{X}(t)$ is the second derivative ($\sim$ acceleration). $\rightsquigarrow$ models **inertia** / **momentum**.
- $-\gamma \dot{X}$ is a viscous friction force ($\gamma > 0$).



Inertial methods can speed up convergence

- Polyak investigates multi-step methods in the paper:

  [*Some methods for speeding up the convergence of iteration methods*. Polyak, 1964].

- A k-step method constructs $x^{(n+1)}$ using the previous k iterations $x^{(n)}, \dots, x^{(n-k+1)}$.
- Gradient descent method is a single-step method.
- Inertial methods are multi-step methods.
- **Heavy-ball method is a 2-step method**.

Evidence in convex optimization:

- Optimal method are usually multi-step methods.
- The Heavy-ball method is optimal for smooth strongly convex functions.

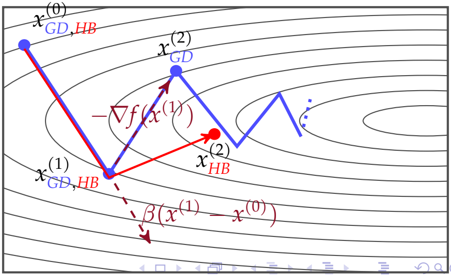
- The (time-discrete) Heavy-ball method has the update rule

$$x^{(k+1)} = x^{(k)} - \alpha\nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\,.$$

- $(x^{(k)})_{k\in\mathbb{N}}$: sequence of iterates.
- $\alpha > 0$: step size parameter.
- $\beta \in [0,1)$: inertial parameter.
- For $\beta = 0$, we recover the gradient descent method.

Some properties:

- It is not a classical descent method.
- It avoids zick-zacking.
- Similarity to conjugate gradient method.



Non-smooth non-convex optimization problems

- Efficiently solving all Lipschitz continuous problems is hopeless [Nesterov, 2004].
- Can take several million years for small problems with only 10 unknowns.

We should exploit the structure of optimization problems

- Develop algorithms for special classes of structured non-convex problems:

$$\min \boxed{\textbf{smooth, non-convex}} + \boxed{\textbf{non-smooth, non-convex, simple}}$$

A generic optimization problem

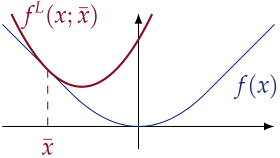
- Non-convex optimization problem with a function $h\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ $\quad (\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\})$

$$\min_{x\in\mathbb{R}^N} h(x)\,; \qquad h(x) := f(x) + g(x)\,.$$

- $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ proper, lower semi-continuous (lsc), **simple**, prox-bounded.

- $f\colon \mathbb{R}^N \to \mathbb{R}$ is smooth with L**-Lipschitz continuous gradient** on $\operatorname{dom} g \subset \mathbb{R}^N$, i.e.

$$|\nabla f(x) - \nabla f(y)| \le L|x-y|\,, \quad \forall x, y \in \operatorname{dom} g\,.$$



- h is **coercive** ($|x| \to +\infty \Rightarrow h(x) \to +\infty$) and bounded from below

Inertial proximal algorithm for nonconvex optimization

Algorithm. (iPiano, [O., Chen, Brox, Pock, 2014], [O., 2015])

- **Optimization problem**:

$$\min_{x\in\mathbb{R}^N} f(x) + g(x)$$

  - f has L-Lipschitz continuous gradient
  - g proper, lsc, prox-bounded

- **Iterations** $(k \geq 0)$: Update $(x^{-1} := x^0 \in \operatorname{dom} g)$

$$x^{(k+1)} \in \operatorname{prox}_{\alpha g}\left(x^{(k)} - \alpha\nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\right)$$

- **Parameter setting**: *See convergence analysis*.

Proximity operator:

- For a proper, lsc, prox-bounded function $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ and $\alpha > 0$, define

$$\mathrm{prox}_{\alpha g}(\bar{x}) := \arg\min_{x\in\mathbb{R}^N} g(x) + \frac{1}{2\alpha}|x - \bar{x}|^2 .$$

- $\mathrm{prox}_{\alpha g}\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ is a set-valued mapping.
- If g is convex, then $\mathrm{prox}_{\alpha g}$ is single-valued.
- If $g = \delta_S$ is the indicator function of a set S, then

$$\mathrm{prox}_{\alpha g}(\bar{x}) = \mathcal{P}_S(\bar{x})$$

  is the **projection onto** S.
- g is **simple**, if $\mathrm{prox}_{\alpha g}$ can be efficiently evaluated for a global minimum.

Relationsship to other methods

$$x^{(k+1)} \in \text{prox}_{\alpha g}\left(x^{(k)} - \alpha \nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\right)$$

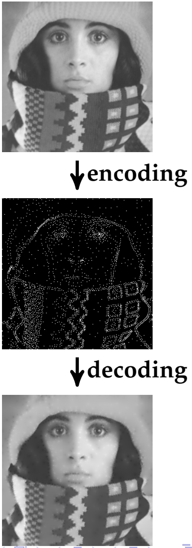
- $g = 0$ and $\beta = 0$: Gradient descent
- $g = \delta_C$ and $\beta = 0$: Projected gradient descent [Goldstein '64], [Levitin, Polyak '66], ...
- $\beta = 0$: Forward–backward splitting [Lions, Mercier '79], [Tseng '91], [Daubechie et al. '04], [Combettes, Wajs '05], [Raguet, Fadili, Peyré '13], [Chouzenoux, Pesquet, Repetti '14], [Fukushima, Mine '81], ...
- $g = 0$: Gradient descent with momentum or Heavy-ball method [Polyak '64], [Zavriev, Kostyuk '93], [Alvarez '04], [Alvarez, Attouch '01], ...
- $f = 0$ and $\beta = 0$: Instance of the proximal point algorithm [Rockafellar '76], ...
- Note the **difference** to Nesterov's method [Nesterov '83]

$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}))) + \beta_k(x^{(k)} - x^{(k-1)})$$

- Generalization to forward–backward splitting [Beck, Teboulle '09], [Nesterov '12], ...

Diffusion based image compression

Diffusion based image compression:



Encoding:

- store image I^0 only in some small number of pixel:
  $c_i = 1$ if pixel i is stored and 0 otherwise

Decoding:

- use $u_i = I_i^0$ for all i with $c_i = 1$
- use linear diffusion in unknown region ($c_i = 0$)
  (solve Laplace equation $Lu = 0$)

$\Rightarrow$ solve for u in

$$C(u - I^0) - (\mathrm{Id} - C)Lu = 0$$

where $C = \mathrm{diag}(c)$, and Id the identity matrix

Diffusion based image compression:

Goal:

- Find a sparse vector c that yields the best reconstruction.

Non-convex optimization problem:

- Math. program with equilibrium constraint

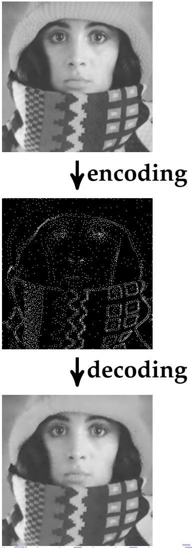
$$\min_{c\in\mathbb{R}^N, u\in\mathbb{R}^N} \frac{1}{2}\|u(c) - I^0\|^2 + \lambda\|c\|_1$$
$$s.t.\ C(u - I^0) - (\mathrm{Id} - C)Lu = 0$$

  where $C = \mathrm{diag}(c)$.

- Can be formulated as

$$\min_{c\in\mathbb{R}^N} \frac{1}{2}\|A^{-1}CI^0 - I^0\|^2 + \lambda\|c\|_1$$

  where $A = C + (C - \mathrm{Id})L$.

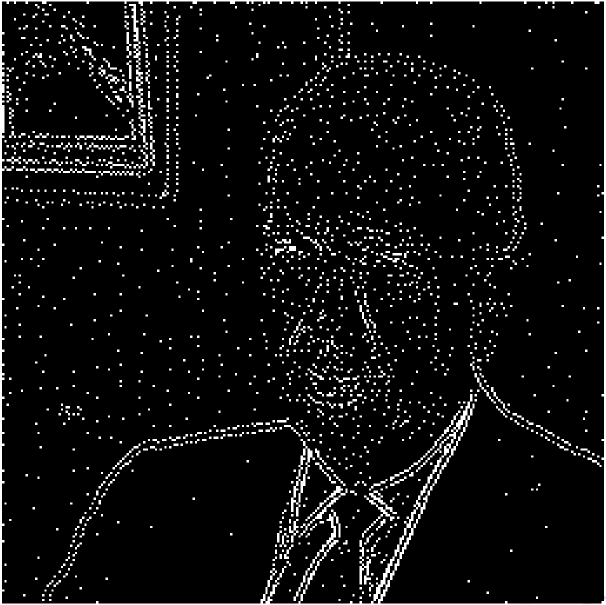


Results for Trui

Results for Trui

Results for Walter

Results for Walter



Results for Walter

Compressive sensing application

Sparse and Low-rank Matrix Decomposition:

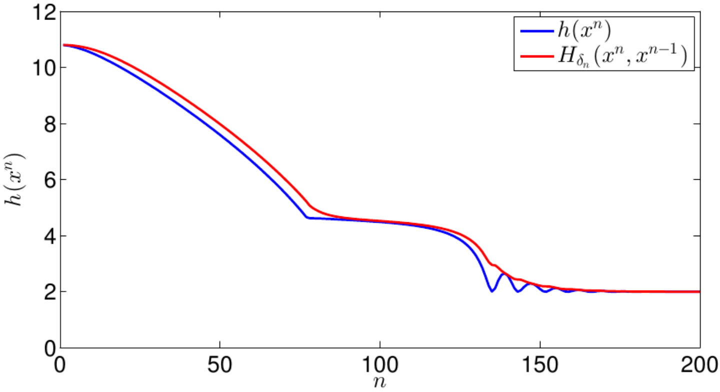
- Let A, X, Y be $M \times N$ matrices.
- Find a decomposition
$$A \approx X + Y\,.$$
- X should have low rank.
- Y should have few non-zero entries.
- Optimization problem:
$$\min_{X,Y\in\mathbb{R}^{M\times N}} \frac{1}{2}\|A - X - Y\|_F^2 + \mathrm{rk}(X) + \|Y\|_0\,.$$

Basic stability result for iPiano

Define $H_\delta(x,y) := h(x) + \delta|x-y|^2$, where $h(x) = f(x) + g(x)$ and $\delta > 0$.

- $(H_\delta(x^{(k)}, x^{(k-1)}))_{k\in\mathbb{N}}$ is monotonically decreasing and thus converging:

$$H_\delta(x^{(k+1)}, x^{(k)}) \le H_\delta(x^{(k)}, x^{(k-1)}) - \gamma|x^{(k)} - x^{(k-1)}|^2 \qquad \text{for some } \gamma > 0\,.$$



Discussion about step size parameters

$$H_\delta(x^{(k+1)}, x^{(k)}) \le H_\delta(x^{(k)}, x^{(k-1)}) - \gamma |x^{(k)} - x^{(k-1)}|^2$$

- Step size restrictions come from $\gamma > 0$.
- Actually, α and β can vary along the iterations.
- Lipschitz constant of ∇f can be estimated "locally" using backtracking.
- Later, γ and δ and the norm can vary along the iterations [O., 2016].

- **General case**:

$$0 < \alpha < \frac{(1-2\beta)}{L} \quad \text{and} \quad \beta \in [0, \frac{1}{2}) .$$

- g **semi-convex** with modulus $m \in \mathbb{R}$ (m maximal such that $g(x) - \frac{m}{2}|x|^2$ is convex):

$$0 < \alpha < \frac{2(1-\beta)}{L-m} \quad \text{and} \quad \beta \in [0, 1) .$$

- g **convex**:

$$0 < \alpha < \frac{2(1-\beta)}{L} \quad \text{and} \quad \beta \in [0, 1) .$$

Definition:

A point $x^* \in \operatorname{dom} h$ is a **critical point** of $h\colon \mathbb{R}^N \to \overline{\mathbb{R}}$, if

$$0 \in \partial h(x^*) \qquad \text{(zero of the limiting subdifferential)}\,.$$

In our case, it is equivalent to

$$-\nabla f(x^*) \in \partial g(x^*)\,.$$

Theorem:

- The sequence $(h(x^{(k)}))_{k\in\mathbb{N}}$ converges.
- There exists a converging subsequence $(x^{k_j})_{j\in\mathbb{N}}$.
- Any limit point $x^* := \lim_{j\to\infty} x^{k_j}$ is a critical point of h and $h(x^{k_j}) \to h(x^*)$ as $j \to \infty$.

Full convergence for iPiano

Theorem:

If $H_\delta(x, y)$ has the **Kurdyka-Łojasiewicz property** at a cluster point (x^*, x^*), then

- $(x^{(k)})_{k\in\mathbb{N}}$ has **finite length**, i.e., $\sum_{k=1}^{\infty} |x^{(k)} - x^{(k-1)}| < \infty$,
- $x^{(k)} \to x^*$ as $k \to \infty$,
- (x^*, x^*) is a critical point of H_δ, and x^* is a critical point of h.

Kurdyka-Łojasiewicz property:

- Weak assumption about the structure of the objective functions.
- **Very hard** to find a function that **does not** have this property.
- Examples on next slide.

Examples of KL functions

- Real analytic functions [Łojasiewicz '63]
- Differentiable functions that are definable in an o-minimal structure [Kurdyka '98]
- Non-smooth lsc functions that are definable in an o-minimal structure [Bolte, Daniilidis, Lewis, Shiota 2007], [Attouch, Bolte, Redont, Soubeyran 2010]
- semi-algebraic functions
  (polynomials, piecewise polynomials, absolute value function, Euclidean distance function, p-norm for $p \in \mathbb{Q}$ (also $p = 0$), ...)
- An o-minimal structure is closed under finite sums and products, composition, and several other important operations

- Bad news: not all functions are KL functions, [Bolte, Daniilidis, Ley, Mazet 2010] construct a $\mathcal{C}^2$ function in $\mathbb{R}^2$ that does not satisfy the KL inequality
- Good news: Such functions are very unlikely to occur in practical applications

Abstract descent algorithms [Attouch et al. 2013]

$$\min_{x\in\mathbb{R}^N} f(x)$$

- $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ proper, lsc
- $(x^{(k)})_{k\in\mathbb{N}}$ sequence of iterates generated by some algorithm
- $a, b > 0$ fixed

(h1) (**Sufficient decrease condition**). For each $k \in \mathbb{N}$,

$$f(x^{(k+1)}) + a|x^{(k+1)} - x^{(k)}|^2 \leq f(x^{(k)})\,;$$

(h2) (**Relative error condition**). For each $k \in \mathbb{N}$, there exists $w^{(k+1)} \in \partial f(x^{(k+1)})$ such that

$$|w^{(k+1)}| \leq b|x^{(k+1)} - x^{(k)}|\,;$$

(h3) (**Continuity condition**). There exists a subsequence $(x^{k_j})_{j\in\mathbb{N}}$ and $\tilde{x}$ such that

$$x^{k_j} \to \tilde{x} \quad \text{and} \quad f(x^{k_j}) \to f(\tilde{x})\,, \quad \text{as } j \to \infty\,.$$

- These properties are shared by many first-order optimization algorithms.

The following analysis is motivated by [Bolte, Sabach, Teboulle, 2013].

Lemma:

- $(f(x^{(k)}))_{k\in\mathbb{N}}$ is non-increasing and converging,
- $\sum_{j=1}^{k} |x^{(j+1)} - x^{(j)}|^2 < +\infty$ and, therefore $|x^{(k+1)} - x^{(k)}| \to 0$, as $k \to \infty$.

Direct consequences for the set of limit points

Define:

- Let ω_0 be the set of limit points of a bounded sequence $(x^{(k)})_{k\in\mathbb{N}}$.
- Subset of limit points that allow for subsequences along which f is continuous, i.e.,

$$\overline{\omega}_0 := \{\bar{x} \in \omega_0 \,|\, x^{(k_j)} \xrightarrow{f} \bar{x} \text{ for } j \to \infty\} \subset \omega_0 \,.$$

Lemma: If f is continuous on $\operatorname{dom} f$, then $\omega_0 = \overline{\omega}_0$.

From now on, let $(x^{(k)})_{k\in\mathbb{N}}$ be a bounded sequence.

Lemma:

- $\overline{\omega}_0$ is non-empty, and $\overline{\omega}_0 \subset \operatorname{crit} f$.
- ω_0 is non-empty, compact, and connected.
- It holds that $\lim_{k\to\infty} \operatorname{dist}(x^{(k)}, \omega_0) = 0$.
- F is constant and finite on $\overline{\omega}_0$.

An abstract convergence theorem

Theorem: ([Attouch et al. 2013])

If

- $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper, lsc,
- $(x^{(k)})_{k\in\mathbb{N}}$ satisfies **(h1)**, **(h2)**, and **(h3)**, and
- f has the KL property at the cluster point $\tilde{x}$,

then

- $(x^{(k)})_{k\in\mathbb{N}}$ converges to $\bar{x} = \tilde{x}$,
- $\bar{x}$ is a critical point of f,
- $(x^{(k)})_{k\in\mathbb{N}}$ has a finite length.

- However, is does not apply to inertial methods directly.

Unifying abstract convergence theorem [O., 2016]

- $(u^{(k)})_{k\in\mathbb{N}}$ be a sequence of parameters in $\mathbb{R}^P$.
- $(\varepsilon_k)_{k\in\mathbb{N}}$ be an ℓ_1-summable sequence of non-negative real numbers.
- $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$, and $(d_k)_{k\in\mathbb{N}}$ of non-negative real numbers.

(H1) (**Sufficient decrease condition**) For each $k\in\mathbb{N}$, it holds that

$$F(x^{(k+1)},u^{(k+1)})+a_k d_k^2\le F(x^{(k)},u^{(k)})\,.$$

(H2) (**Relative error condition**) For each $k\in\mathbb{N}$, the following holds:

$$b_{k+1}\|\partial F(x^{(k+1)},u^{(k+1)})\|_-\le \frac{b}{2}(d_{k+1}+d_k)+\varepsilon_{k+1}\,.$$

(H3) (**Continuity condition**) $\exists((x^{(k_j)},u^{(k_j)}))_{j\in\mathbb{N}}$ and $(\tilde{x},\tilde{u})\in\mathbb{R}^N\times\mathbb{R}^P$ such that

$$(x^{(k_j)},u^{(k_j)})\xrightarrow{F}(\tilde{x},\tilde{u})\quad\text{as}\quad j\to\infty\,.$$

(H4) (**Contraction condition**) It holds that

$$|x^{(k+1)}-x^{(k)}|_2\in\mathrm{o}(d_k)\quad\text{and}\quad(b_k)_{k\in\mathbb{N}}\notin\ell_1\,,\quad\sup_{k\in\mathbb{N}}b_k a_k<\infty\,,\quad\inf_k a_k=:\underline{a}>0\,.$$

Theorem:

If

- F is a proper, lsc, bounded from below, and has the KL property,
- $(x^{(k)})_{k\in\mathbb{N}}$ be a bounded sequence generated by an abstract parametrized algorithm,
- with a sequence of parameter $(u^{(k)})_{k\in\mathbb{N}}$,
- $\omega(x^{(0)}, u^{(0)}) = \overline{\omega}(x^{(0)}, u^{(0)})$,

then

- $(x^{(k)})_{k\in\mathbb{N}}$ satisfies

$$\sum_{k=0}^{\infty} |x^{(k+1)} - x^{(k)}| < +\infty\,,$$

  and $(x^{(k)})_{k\in\mathbb{N}}$ converges to some $\tilde{x}$.
- Moreover, if $(u^{(k)})_{k\in\mathbb{N}}$ is a converging sequence, then $((x^{(k)}, u^{(k)}))_{k\in\mathbb{N}}$ F-converges to $(\tilde{x}, \tilde{u})$, and $(\tilde{x}, \tilde{u})$ is a critical point of F.

Non-smoth non-convex optimization problem: (f smooth, g non-smooth)

$$\min_{x\in\mathbb{R}^N} h(x) = \min_{x\in\mathbb{R}^N} f(x) + g(x)$$

Algorithm. (variable metric iPiano, [O., 2016])

- **Initialization**: Choose a starting point $x^{(0)} \in \operatorname{dom} h$ and set $x^{(-1)} = x^{(0)}$.
- **Iterations** $(k \geq 0)$: Choose $A_k \in \mathbb{S}(N)$, $0 \prec A_k \preceq \mathrm{Id}$, and update:

$$x^{(k+1)} \in (\mathrm{Id} + \alpha_k A_k^{-1}\partial g)^{-1}\left(x^{(k)} - \alpha_k A_k^{-1}\nabla f(x^{(k)}) + \beta_k(x^{(k)} - x^{(k-1)})\right),$$

  where α_k, β_k, γ_k, and δ_k are as in the base variant of iPiano and the following monotonicity condition holds:

$$\delta_{k+1}|x^{(k+1)} - x^{(k)}|^2_{A_{k+1}} \leq \delta_k|x^{(k+1)} - x^{(k)}|^2_{A_k}.$$

- **Convergence**: Same as in the Abstract Convergence Theorem from before.

- Lipschitz constant can be estimated with backtracking.
- Algorithm can be extended to block coordinate version.

- **Heavy-ball dynamical system**:

$$\ddot{X}(t) = -\gamma \dot{X}(t) - \nabla f(X(t))$$

- The (time-discrete) **Heavy-ball method** has the update rule

$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\,.$$

- Develop algorithms for special classes of structured non-convex problems:

$$\min \; \boxed{\textbf{smooth, non-convex}} + \boxed{\textbf{non-smooth, non-convex, simple}}$$

- **iPiano**:

$$x^{(k+1)} \in \operatorname{prox}_{\alpha g}\left(x^{(k)} - \alpha \nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\right)$$

- Convergence analysis of iPiano and abstract descent methods.