First order Methods for Smooth Adaptable Nonconvex Composite Optimization

Shoham Sabach

Faculty of Industrial Engineering and Management

Technion - Israel Institute of Technology

Joint work with Jerome Bolte, Marc Teboulle and Yakov Vaisbourd

Games, Dynamics and Optimization (GDO2019)

Babes-Bolyai University Cluj-Napoca

April 9, 2019

Recall: A Central Pillar Underlying Analysis of FOM

$$\inf \left\{ \Phi(x) := f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

- $f : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lower semicontniuous.
- $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable.

Captures many applied problems, and the source for fundamental FOM.

A central assumption: g admits an L-Lipschitz continuous gradient on $\mathbb{R}^d$.

Recall: A Central Pillar Underlying Analysis of FOM

$$\inf \left\{ \Phi(x) := f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

- $f : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lower semicontniuous.
- $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable.

Captures many applied problems, and the source for fundamental FOM.

A central assumption: g admits an L-Lipschitz continuous gradient on $\mathbb{R}^d$.

A simple, yet a key consequence of this, is the so-called **descent Lemma**:

$$g(x) \leq g(y) + \langle \nabla g(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall \, x, y \in \mathbb{R}^d.$$

This inequality provides

1. An **upper quadratic approximation** of g.
2. A crucial pillar in the development and analysis of many FOM.

However, in many contexts and applications:

- the differentiable function g **does not** have a global L-smooth gradient.
- Hence precludes **direct use** of basic FOM methodology and schemes.

Starting Point: A Recent Approach for Convex Problems

Recently, in (Bauschke, Bolte and Teboulle (2017)) a simple framework was proposed for the **convex** composite minimization

$$\inf \left\{ \Phi\left(x\right) := f\left(x\right) + g\left(x\right) : \; x \in \mathbb{R}^{d} \right\},$$

- $f : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ is proper and lower semicontniuous.
- $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is continuously differentiable, **but does not have a globally Lipschitz continuous gradient**.

Starting Point: A Recent Approach for Convex Problems

Recently, in (Bauschke, Bolte and Teboulle (2017)) a simple framework was proposed for the **convex** composite minimization

$$\inf\left\{\Phi(x) := f(x) + g(x) : \; x \in \mathbb{R}^d\right\},$$

- $f : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lower semicontniuous.
- $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, **but does not have a globally Lipschitz continuous gradient**.

1. The idea of this framework is based on “better capturing the geometry" of the problem at hand.
2. Allows for a new Descent Lemma: the classical upper quadratic approximation of Ψ is **replaced by a more suitable approximation which can be adapted to the objective function**.
3. The corresponding emerges FOM enjoys **guaranteed complexity estimates and pointwise global convergence** results in the **convex setting**.

Starting Point: A Recent Approach for Convex Problems

Recently, in (Bauschke, Bolte and Teboulle (2017)) a simple framework was proposed for the **convex** composite minimization

$$\inf \left\{ \Phi(x) := f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

- $f : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lower semicontniuous.
- $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, **but does not have a globally Lipschitz continuous gradient**.

1. The idea of this framework is based on “better capturing the geometry" of the problem at hand.
2. Allows for a new Descent Lemma: the classical upper quadratic approximation of Ψ is **replaced by a more suitable approximation which can be adapted to the objective function**.
3. The corresponding emerges FOM enjoys **guaranteed complexity estimates and pointwise global convergence** results in the **convex setting**.

Main Goal: Extend BBT framework to analyze nonconvex composite minimization problems.

The Nonconvex Composite Model

We are focusing on the nonconvex and nonsmooth composite problem

$$(\mathrm{P}) \quad \inf \left\{ \Psi(x) \equiv f(x) + g(x) : \ x \in \overline{C} \right\},$$

Assumption 1

- C is a nonempty, convex and open subset of $\mathbb{R}^d$.
- $f : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper and lsc function with $\mathrm{dom}\, f \cap C \neq \emptyset$.
- $g : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lsc, and C^1 on C.
- $v(P) := \inf \left\{ \Psi(x) : \ x \in \overline{C} \right\} > -\infty$.

Smooth Adaptable Functions

First, we need to define a specific class of Kernel functions.

▷ Let C be a nonempty, convex and open subset of $\mathbb{R}^d$.

▷ Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, lsc and convex such that

(i) $\mathrm{dom}\, h \subset \overline{C}$ and $\mathrm{dom}\, \partial h = C$.

(ii) h is C^1 on $\mathrm{int}\, \mathrm{dom}\, h \equiv C$.

We denote this class of functions by $\mathcal{G}(C)$.

Smooth Adaptable Functions

First, we need to define a specific class of Kernel functions.

▹ Let C be a nonempty, convex and open subset of $\mathbb{R}^d$.

▹ Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, lsc and convex such that

(i) $\mathrm{dom}\, h \subset \overline{C}$ and $\mathrm{dom}\, \partial h = C$.

(ii) h is C^1 on $\mathrm{int}\, \mathrm{dom}\, h \equiv C$.

We denote this class of functions by $\mathcal{G}(C)$.

Definition (L-smooth adaptable)

Let $h \in \mathcal{G}(C)$, and let $g : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lsc function with $\mathrm{dom}\, h \subset \mathrm{dom}\, g$, which is C^1 on $C \equiv \mathrm{int}\, \mathrm{dom}\, h$. The pair (g, h) is called ***L*-smooth adaptable** on C if there exists $L > 0$ such that $Lh - g$ and $Lh + g$ are convex on C.

Approximation without Lipschitz Gradient Continuity

Lemma (Fundamental approximation)

The pair of functions (g, h) is ***L-smooth adaptable*** *on C if and only if:*

$$\left| g(x) - g(y) - \langle \nabla g(y), x - y \rangle \right| \leq L D_h(x, y), \quad \forall\ x, y \in \operatorname{int} \operatorname{dom} h.$$

Approximation without Lipschitz Gradient Continuity

Lemma (Fundamental approximation)

The pair of functions (g, h) is **L-smooth adaptable** *on C if and only if:*

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq L D_h(x, y), \quad \forall\, x, y \in \operatorname{int}\operatorname{dom} h.$$

D_h stands for the Bregman Distance (Bregman (67)) associated to $h \in \mathcal{G}(C)$:

$$D_h(x, y) := h(x) - [h(y) + \langle \nabla h(y), x - y \rangle], \quad \forall\, x \in \operatorname{dom} h, y \in \operatorname{int}\operatorname{dom} h.$$

Distance-like properties. For all $(x, y) \in \operatorname{dom} h \times \operatorname{int}\operatorname{dom} h$ we have

(i) h is convex if and only if $D_h(x, y) \geq 0$ for all $x \in \operatorname{dom} h$ and $y \in \operatorname{int}\operatorname{dom} h$.

(ii) When h is strictly convex, $D_h(x, y) = 0$ if and only if $x = y$.

(iii) **However, note that D_h is in general not symmetric!**

Approximation without Lipschitz Gradient Continuity

Lemma (Fundamental approximation)

The pair of functions (g, h) *is* **L-smooth adaptable** *on* C *if and only if:*

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq L D_h(x, y), \quad \forall\ x, y \in \operatorname{int} \operatorname{dom} h.$$

D_h stands for the Bregman Distance (Bregman (67)) associated to $h \in \mathcal{G}(C)$:

$$D_h(x, y) := h(x) - [h(y) + \langle \nabla h(y), x - y \rangle], \quad \forall\ x \in \operatorname{dom} h, y \in \operatorname{int} \operatorname{dom} h.$$

Distance-like properties. For all $(x, y) \in \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h$ we have

(i) h is convex if and only if $D_h(x, y) \geq 0$ for all $x \in \operatorname{dom} h$ and $y \in \operatorname{int} \operatorname{dom} h$.

(ii) When h is strictly convex, $D_h(x, y) = 0$ if and only if $x = y$.

(iii) **However, note that D_h is in general not symmetric!**

Proof of Lemma. $Lh \pm g$ is convex on $C = \operatorname{int} \operatorname{dom} h$ is equivalent to:

$$D_{Lh \pm g}(x, y) \geq 0 \quad \Longleftrightarrow \quad L D_h(x, y) \pm D_g(x, y) \geq 0.$$

Smooth Adaptable Functions - Few Comments

- When $C = \mathbb{R}^d$ and $h(\cdot) = (1/2)\|\cdot\|^2$, we recover the fundamental:

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq \frac{L}{2}\|x - y\|^2, \quad \forall\ x, y \in \mathbb{R}^d.$$

- When g is assumed convex, the condition $Lh + g$ is convex, trivially holds! And we recover the **Nolips Descent Lemma of BBT**, *i.e.*,

$$D_g(x, y) \leq LD_h(x, y).$$

- The convexity of $Lh + g$ can be written with respect to a different parameter $\ell \leq L$.

Smooth Adaptable Functions - Few Comments

- When $C = \mathbb{R}^d$ and $h(\cdot) = (1/2)\|\cdot\|^2$, we recover the fundamental:

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq \frac{L}{2}\|x - y\|^2, \quad \forall\ x, y \in \mathbb{R}^d.$$

- When g is assumed convex, the condition $Lh + g$ is convex, trivially holds! And we recover the **Nolips Descent Lemma of BBT**, *i.e.*,

$$D_g(x, y) \leq LD_h(x, y).$$

- The convexity of $Lh + g$ can be written with respect to a different parameter $\ell \leq L$.

Note 1: Here g **is not convex.**

Smooth Adaptable Functions - Few Comments

- When $C = \mathbb{R}^d$ and $h(\cdot) = (1/2)\|\cdot\|^2$, we recover the fundamental:

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq \frac{L}{2}\|x - y\|^2, \quad \forall\ x, y \in \mathbb{R}^d.$$

- When g is assumed convex, the condition $Lh + g$ is convex, trivially holds! And we recover the **Nolips Descent Lemma of BBT**, *i.e.*,

$$D_g(x, y) \leq LD_h(x, y).$$

- The convexity of $Lh + g$ can be written with respect to a different parameter $\ell \leq L$.

Note 1: Here g **is not convex.**
Note 2: We can always assume that h is σ-strongly convex:

$$Lh - g = L\left(h - (\sigma/2)\|\cdot\|^2\right) - \left(g - (L\sigma/2)\|\cdot\|^2\right) := L\bar{h} - \bar{g}.$$

Smooth Adaptable Functions - Few Comments

- When $C = \mathbb{R}^d$ and $h(\cdot) = (1/2)\|\cdot\|^2$, we recover the fundamental:

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq \frac{L}{2}\|x - y\|^2, \quad \forall\, x, y \in \mathbb{R}^d.$$

- When g is assumed convex, the condition $Lh + g$ is convex, trivially holds! And we recover the **Nolips Descent Lemma of BBT**, *i.e.*,

$$D_g(x, y) \leq L D_h(x, y).$$

- The convexity of $Lh + g$ can be written with respect to a different parameter $\ell \leq L$.

Note 1: Here g **is not convex.**
Note 2: We can always assume that h is σ-strongly convex:

$$Lh - g = L\left(h - (\sigma/2)\|\cdot\|^2\right) - \left(g - (L\sigma/2)\|\cdot\|^2\right) := L\bar{h} - \bar{g}.$$

For our purposes, it will be enough to consider **only the condition that $Lh - g$ is convex on C.**

The Nonconvex Composite Model

We are focusing on the nonconvex nonsmooth composite problem

$$\text{(P)} \quad \inf\left\{\Psi\left(x\right) \equiv f\left(x\right) + g\left(x\right) : \ x \in \overline{C}\right\},$$

Assumption 1

(i) $h \in \mathcal{G}(C)$ such that $Lh - g$ convex on $C = \operatorname{int}\operatorname{dom} h$

(ii) $f : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper and lsc function with $\operatorname{dom} f \cap C \neq \emptyset$.

(iii) $g : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lsc with $\operatorname{dom} h \subset \operatorname{dom} g$, and C^1 on C.

(iv) $v(P) := \inf\left\{\Psi\left(x\right) : \ x \in \overline{C}\right\} > -\infty$.

Bregman Proximal Gradient Map

For all $x \in \operatorname{int}\operatorname{dom} h$ and any $\lambda > 0$, we define

$$T_\lambda(x) := \operatorname{argmin}_u \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) \right\}.$$

The map emerges from the usual approach:

- Linearize the differentiable part g around x and regularize it.
- Leave untouched the nonsmooth function f.
- Since f is **nonconvex**, the mapping T_λ is **not**, in general, single-valued.

Classical case: With $h(\cdot) := \frac{1}{2}\|\cdot\|^2$ nothing else but the classical proximal gradient (forward-backward) map:

$$T_\lambda(x) := \operatorname{argmin}_u \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{2\lambda}\|u - x\|^2 \right\}.$$

Bregman Proximal Gradient Map

For all $x \in \operatorname{int} \operatorname{dom} h$ and any $\lambda > 0$, we define

$$T_\lambda(x) := \operatorname{argmin}_u \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) \right\}.$$

The map emerges from the usual approach:

- Linearize the differentiable part g around x and regularize it.
- Leave untouched the nonsmooth function f.
- Since f is **nonconvex**, the mapping T_λ is **not**, in general, single-valued.

Classical case: With $h(\cdot) := \frac{1}{2}\|\cdot\|^2$ nothing else but the classical proximal gradient (forward-backward) map:

$$T_\lambda(x) := \operatorname{argmin}_u \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{2\lambda} \|u - x\|^2 \right\}.$$

- Mostly studied in the **convex setting**: f and g both convex, with **Lipschitz** ∇g.
- Further extended to find the zero of maximal monotone inclusions.

(Bruck (77), Passty (79), Lions-Mercier (79), Fukushima-Milne (81)...)

Well-Posedness of T_λ

Assumption 2

(i) The function $h + \lambda f$ is supercoercive for all $\lambda > 0$, that is,

$$\lim_{\|u\| \to \infty} \frac{h(u) + \lambda f(u)}{\|u\|} = \infty.$$

(ii) For all $x \in C$ we have $T_\lambda(x) \subset C, \quad \forall\, x \in C$.

Well-Posedness of T_λ

Assumption 2

(i) The function $h + \lambda f$ is supercoercive for all $\lambda > 0$, that is,

$$\lim_{\|u\| \to \infty} \frac{h(u) + \lambda f(u)}{\|u\|} = \infty.$$

(ii) For all $x \in C$ we have $T_\lambda(x) \subset C, \quad \forall\, x \in C$.

- Item (i) is a quite standard coercivity condition, e.g., automatically satisfied when $\overline{C}$ is compact.
- Item (ii) can be shown to hold under a classical constraint qualification condition:

$$\partial^\infty f(x) \cap (-\partial^\infty h(x)) = \{0\}, \quad \forall\, x \in \mathbb{R}^d.$$

- Item (ii) also holds automatically when $C = \mathbb{R}^d$ or f is convex (note: problem (P) **remains nonconvex!**).

  In later case, T_λ reduces to minimize a strictly convex function, and T_λ is single-valued from $\mathrm{int\,dom}\, h$ to $\mathrm{int\,dom}\, h$.

The Bregman Proximal Gradient Algorithm

Bregman Proximal Gradient - BPG
Input. A function $h \in \mathcal{G}(C)$ with $C = \operatorname{int}\operatorname{dom} h$ such that $Lh - g$ convex on C.
Initialization. $x^0 \in \operatorname{int}\operatorname{dom} h$ and let $\lambda > 0$.
General Step. For $k = 1, 2, \ldots$, compute

$$x^k \in \operatorname{argmin}\left\{f(x) + \left\langle x - x^{k-1}, \nabla g\left(x^{k-1}\right)\right\rangle + \frac{1}{\lambda} D_h\left(x, x^{k-1}\right) : \; x \in \overline{C}\right\}.$$

Under our standing Assumption 1 and 2 the algorithm is well-defined.

The Bregman Proximal Gradient Algorithm

Bregman Proximal Gradient - BPG
Input. A function $h \in \mathcal{G}(C)$ with $C = \operatorname{int}\operatorname{dom} h$ such that $Lh - g$ convex on C.
Initialization. $x^0 \in \operatorname{int}\operatorname{dom} h$ and let $\lambda > 0$.
General Step. For $k = 1, 2, \ldots$, compute

$$x^k \in \operatorname{argmin}\left\{f(x) + \left\langle x - x^{k-1}, \nabla g\left(x^{k-1}\right)\right\rangle + \frac{1}{\lambda} D_h\left(x, x^{k-1}\right) : \; x \in \overline{C}\right\}.$$

Under our standing Assumption 1 and 2 the algorithm is well-defined.

Main computational step. For any $x \in C$, needs to solve

$$x^+ = T_\lambda(x) := \operatorname{argmin}_u \left\{\lambda f(u) + h(u) + \langle u, c(x)\rangle\right\}.$$

Properties and Rate of Convergence for NonConvex-BPG Algorithm

Theorem (Properties and rate of Ncvx-BPG)

Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by BPG. Then, with $\lambda L \in (0,1)$ we have

(i) *(**Sufficient decrease**) The sequence $\{\Psi(x^k)\}_{k\in\mathbb{N}}$ is nonincreasing.*

(ii) *(**Summability**) $\sum_{k=1}^{\infty} D_h(x^k, x^{k-1}) < \infty$.*

(iii) *(**Rate**) $\min_{1\leq k\leq n} D_h(x^k, x^{k-1}) \leq \frac{\lambda}{n}\left(\frac{\Psi(x^0)-\Psi_*}{1-\lambda L}\right)$.*

Properties and Rate of Convergence for NonConvex-BPG Algorithm

Theorem (Properties and rate of Ncvx-BPG)

Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by BPG. Then, with $\lambda L \in (0,1)$ we have

(i) *(**Sufficient decrease**) The sequence $\{\Psi(x^k)\}_{k\in\mathbb{N}}$ is nonincreasing.*

(ii) *(**Summability**) $\sum_{k=1}^{\infty} D_h(x^k, x^{k-1}) < \infty$.*

(iii) *(**Rate**) $\min_{1\le k\le n} D_h(x^k, x^{k-1}) \le \frac{\lambda}{n}\left(\frac{\Psi(x^0)-\Psi_*}{1-\lambda L}\right)$.*

- Recall: we can assume that h is σ-strongly convex on C, and we immediately get:

$$\min_{1\le k\le n} \operatorname{dist}^2\left(x^{k-1}, T_\lambda\left(x^{k-1}\right)\right) \le \min_{1\le k\le n}\left\|x^k - x^{k-1}\right\|^2 \le \frac{\lambda}{n}\cdot\frac{\Psi\left(x^0\right)-\Psi_*}{\sigma(1-\lambda L)}.$$

- **Special case:** $h(u) = (1/2)\|u\|^2$, the classical $O(n^{-1/2})$ rate for proximal gradient in the nonconvex setting is recovered:

$$\gamma_n := \min_{1\le k\le n}\left\|x^k - x^{k-1}\right\| \le \frac{1}{\sqrt{n}}\left(\frac{2\left(\Psi\left(x^0\right)-\Psi_*\right)}{L}\right)^{1/2}.$$

Global Convergence Analysis of Nonconvex BPG

From now on, we consider the same nonconvex model, **but** with $C \equiv \mathbb{R}^d$.

$$\text{(P)} \quad \inf \left\{ \Psi(x) \equiv f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

Global Convergence Analysis of Nonconvex BPG

From now on, we consider the same nonconvex model, **but** with $C \equiv \mathbb{R}^d$.

$$(\text{P}) \quad \inf \left\{ \Psi(x) \equiv f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

Assumption 3

(i) $\mathrm{dom}\, h = \mathbb{R}^d$.

(ii) h is strongly convex on $\mathbb{R}^d$.

(iii) ∇h and ∇g are Lipschitz continuous on any bounded subset of $\mathbb{R}^d$.

In this case, the set of critical points of Ψ is simply given by:

$$\mathrm{crit}\, \Psi = \left\{ x \in \mathbb{R}^d : \; 0 \in \partial \Psi(x) \equiv \partial f(x) + \nabla g(x) \right\}.$$

Notes: Item (iii) is harmless. Item (ii) can always be enforced, without impairing the convexity of $Lh - g$!.

(Limiting) Subdifferential $\partial \Psi(x)$ (Rockafellar-Wets (93)):

$$\begin{aligned} x^* \in \partial \Psi(x) \quad & \text{iff} \quad (x_k, x^*) \to (x, x^*) \text{ s.t. } \Psi(x_k) \to \Psi(x) \text{ and} \\ F(u) \quad & \geq \quad F(x_k) + \langle x_k^*, u - x_k \rangle + o(\|u - x_k\|) \end{aligned}$$

Global Convergence of the BPG

$$\text{(P)} \quad \inf \left\{ \Psi(x) \equiv f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

Bregman Proximal Gradient - BPG

Input. A function $h \in \mathcal{G}(C)$ with $C = \operatorname{int} \operatorname{dom} h$ such that $Lh - g$ convex on C.

Initialization. $x^0 \in \operatorname{int} \operatorname{dom} h$ and let $\lambda > 0$.

General Step. For $k = 1, 2, \ldots$, compute

$$x^k \in \operatorname{argmin} \left\{ f(x) + \left\langle x - x^{k-1}, \nabla g\left(x^{k-1}\right) \right\rangle + \frac{1}{\lambda} D_h\left(x, x^{k-1}\right) : \; x \in \overline{C} \right\}.$$

Theorem (Convergence of BPG)

Let $\left\{x^k\right\}_{k \in \mathbb{N}}$ *be a bounded sequence generated by BPG and* $0 < \lambda L < 1$.

(i) **Subsequential convergence.** *Any limit point of* $\left\{x^k\right\}_{k \in \mathbb{N}}$ *is a critical point of* Ψ.

(ii) **Global convergence.** *Assume* Ψ *is* **real semi-algebraic**. *Then the sequence* $\left\{x^k\right\}_{k \in \mathbb{N}}$ *has finite length and converges to a critical point* x^* *of* Ψ.

Global Convergence of the BPG

$$(\mathrm{P}) \quad \inf \left\{ \Psi(x) \equiv f(x) + g(x) : \; x \in \mathbb{R}^d \right\},$$

Bregman Proximal Gradient - BPG

Input. A function $h \in \mathcal{G}(C)$ with $C = \operatorname{int} \operatorname{dom} h$ such that $Lh - g$ convex on C.

Initialization. $x^0 \in \operatorname{int} \operatorname{dom} h$ and let $\lambda > 0$.

General Step. For $k = 1, 2, \ldots$, compute

$$x^k \in \operatorname{argmin} \left\{ f(x) + \left\langle x - x^{k-1}, \nabla g\left(x^{k-1}\right) \right\rangle + \frac{1}{\lambda} D_h\left(x, x^{k-1}\right) : \; x \in \overline{C} \right\}.$$

Theorem (Convergence of BPG)

Let $\left\{x^k\right\}_{k \in \mathbb{N}}$ be a bounded sequence generated by BPG and $0 < \lambda L < 1$.

(i) **Subsequential convergence.** *Any limit point of $\left\{x^k\right\}_{k \in \mathbb{N}}$ is a critical point of Ψ.*

(ii) **Global convergence.** *Assume Ψ is* **real semi-algebraic**. *Then the sequence $\left\{x^k\right\}_{k \in \mathbb{N}}$ has finite length and converges to a critical point x^* of Ψ.*

Open Question: Convergence when $\operatorname{dom} h \neq \mathbb{R}^d$?

Application: Quadratic Inverse Problems

- $A_i \in \mathbb{R}^{d \times d}$, $i = 1, 2, \ldots, m$ symmetric matrices.
- $b \in \mathbb{R}^m$ vector of noisy measurements.

Goal: find $x \in \mathbb{R}^d$, that solves the following system

$$x^T A_i x \simeq b_i, \quad i = 1, 2, \ldots, m.$$

- Natural extension of classical **linear inverse problems**.
- Includes the class of **phase retrieval problems**, which is fundamental in physical/engineering sciences. See ***Phase retrieval, what's new ?*** (Luke (17)).

Application: Quadratic Inverse Problems

- $A_i \in \mathbb{R}^{d \times d}$, $i = 1, 2, \ldots, m$ symmetric matrices.
- $b \in \mathbb{R}^m$ vector of noisy measurements.

Goal: find $x \in \mathbb{R}^d$, that solves the following system

$$x^T A_i x \simeq b_i, \quad i = 1, 2, \ldots, m.$$

- Natural extension of classical **linear inverse problems**.
- Includes the class of **phase retrieval problems**, which is fundamental in physical/engineering sciences. See ***Phase retrieval, what's new ?*** (Luke (17)).

Nonconvex optimization formulation. Adopting the usual least-squares model:

$$\text{(QIP)} \qquad \min \left\{ \Psi(x) := \frac{1}{4} \sum_{i=1}^{m} \left(x^T A_i x - b_i \right)^2 + f(x) : \ x \in \mathbb{R}^d \right\},$$

where f (can be **nonconvex**) describes relevant constraints or is a regularizer (nonsmooth) to ensure well-posedness.

Applying Ncvx BPG on Quadratic Inverse Problems

The function $g(x) := \frac{1}{4} \sum_{i=1}^{m} \left(x^T A_i x - b_i \right)^2$

- is nonconvex and C^1 on $\mathbb{R}^d$,
- **but does not** admit a global Lipschitz continuous gradient.

Applying Ncvx BPG on Quadratic Inverse Problems

The function $g(x) := \frac{1}{4}\sum_{i=1}^{m}\left(x^T A_i x - b_i\right)^2$

- is nonconvex and C^1 on $\mathbb{R}^d$,
- **but does not** admit a global Lipschitz continuous gradient.

We can activate Ncvx PBG, that is:

- find an h and **an explicit** L in terms of (A_i, b_i) such that $Lh - g$ is convex on $\mathbb{R}^d$.
- Explicitly computable $T_\lambda(\cdot)$ in the following two important cases for QIP:
  - (a) **A convex ℓ_1-norm regularization.** With $f(x) = \|x\|_1$.
  - (b) **A nonconvex sparsity constraint.** With $f(x) = \delta_{\mathbb{B}_0^s}(x)$:

$$\mathbb{B}_0^s \equiv \left\{x : \ \|x\|_0 \leq s\right\}, \ (\ell_0 \text{ quasi-ball } 0 < s < d).$$

Applying Ncvx BPG on Quadratic Inverse Problems

The function $g(x) := \frac{1}{4}\sum_{i=1}^{m}\left(x^T A_i x - b_i\right)^2$

- is nonconvex and C^1 on $\mathbb{R}^d$,
- **but does not** admit a global Lipschitz continuous gradient.

We can activate Ncvx PBG, that is:

- find an h and **an explicit** L in terms of (A_i, b_i) such that $Lh - g$ is convex on $\mathbb{R}^d$.
- Explicitly computable $T_\lambda(\cdot)$ in the following two important cases for QIP:
  (a) **A convex ℓ_1-norm regularization.** With $f(x) = \|x\|_1$.

  (b) **A nonconvex sparsity constraint.** With $f(x) = \delta_{\mathbb{B}_0^s}(x)$:

$$\mathbb{B}_0^s \equiv \left\{x : \ \|x\|_0 \leq s\right\}, \ (\ell_0 \text{ quasi-ball } 0 < s < d).$$

In both cases, the data is semi-algebraic, and the main computational step:

$$T_\lambda(x) = \operatorname{argmin}\left\{f(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{\lambda}D_h(u, x) : \ u \in \mathbb{R}^d\right\} \ (\lambda > 0).$$

produces a **new, simple and explicit schemes**, proven to generate globally convergent sequences (details in (Bolte-Sabach-T.-Vaisbourd (18))).

A Smooth Adaptable (h, g) for the QIP

Clearly, the nonconvex function $g : \mathbb{R}^d \to (-\infty, +\infty]$ defined by

$$g(x) := \frac{1}{4} \sum_{i=1}^{m} \left(x^T A_i x - b_i \right)^2,$$

is C^1 on $\mathbb{R}^d$, but **does not** admit a global Lipschitz continuous gradient.

A Smooth Adaptable (h, g) for the QIP

Clearly, the nonconvex function $g : \mathbb{R}^d \to (-\infty, +\infty]$ defined by

$$g(x) := \frac{1}{4} \sum_{i=1}^{m} \left(x^T A_i x - b_i \right)^2,$$

is C^1 on $\mathbb{R}^d$, but **does not** admit a global Lipschitz continuous gradient.

Now we need to **identify a suitable function** $h \in \mathcal{G}(\mathbb{R}^d)$ such that $Lh - g$ convex holds for the pair (g, h). Here, we show that the following $h : \mathbb{R}^d \to \mathbb{R}$ does the job:

$$h(x) = \frac{1}{4} \|x\|_2^4 + \frac{1}{2} \|x\|_2^2.$$

A Smooth Adaptable (h, g) for the QIP

Clearly, the nonconvex function $g : \mathbb{R}^d \to (-\infty, +\infty]$ defined by

$$g(x) := \frac{1}{4} \sum_{i=1}^{m} \left(x^T A_i x - b_i \right)^2 ,$$

is C^1 on $\mathbb{R}^d$, but **does not** admit a global Lipschitz continuous gradient.

Now we need to **identify a suitable function** $h \in \mathcal{G}(\mathbb{R}^d)$ such that $Lh - g$ convex holds for the pair (g, h). Here, we show that the following $h : \mathbb{R}^d \to \mathbb{R}$ does the job:

$$h(x) = \frac{1}{4} \|x\|_2^4 + \frac{1}{2} \|x\|_2^2 .$$

Lemma

Let g and h as defined above. Then, $Lh - g$ is convex on $\mathbb{R}^d$ for any L satisfying

$$L \geq \sum_{i=1}^{m} \left(3 \|A_i\|^2 + \|A_i\| \, |b_i| \right) .$$

Explicit Formula for The case $f = \|\cdot\|_1$

Proposition (Bregman Proximal Formula for the ℓ_1-Norm)

Let $x \in \mathbb{R}^d$, let $v(x) := \mathbb{S}_{\lambda\theta}\left(\lambda \nabla g(x) - \nabla h(x)\right)$. Then, $x^+ = T_\lambda(x)$ is given by

$$x^+ = -t^* v(x) = t^* \mathbb{S}_{\lambda\theta}\left(\nabla h(x) - \lambda \nabla g(x)\right),$$

where t^* is the **unique positive real root** *of*

$$t^3 \|v(x)\|_2^2 + t - 1 = 0,$$

which admits an explicit formula (Cardano (1545))

Soft-thresholding (with parameter τ). For any $y \in \mathbb{R}^d$,

$$\mathbb{S}_\tau(y) = \mathrm{argmin}_{x \in \mathbb{R}^d} \left\{ \tau \|x\|_1 + \frac{1}{2}\|x - y\|^2 \right\} = \max\{|y| - \tau, 0\} \,\mathrm{sgn}(y),$$

with the absolute value understood to be component-wise.

Explicit Formula for the Sparsity Constrained Case $f = \delta_{\mathbb{B}_0^s}$

Proposition (Bregman Proximal Formula T_λ for the ℓ_0-Ball)

Let $x \in \mathbb{R}^d, \lambda > 0$ and $p_\lambda(x) := \nabla h(x) - \lambda \nabla g(x)$. Then,

$$x^+ \equiv T_\lambda(x) = -\sqrt{t^*} \left\| \mathcal{H}_s(p_\lambda(x)) \right\|_2^{-1} \mathcal{H}_s(p_\lambda(x)),$$

where $\sqrt{t^*} \equiv \eta^*$ is the **unique positive real root** *of the cubic equation*

$$\eta^3 + \eta - \left\| \mathcal{H}_s(p_\lambda(x)) \right\|_2 = 0,$$

which admits an explicit formula (Cardano (1545)).

Hard-thresholding (with parameter τ). For any $y \in \mathbb{R}^d$,

$$\mathcal{H}_\tau(y) = \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \|x - y\|^2 : \ x \in \mathbb{B}_0^\tau \right\} = \begin{cases} y_i, & i \leq \tau, \\ 0, & \text{otherwise}, \end{cases}$$

where we assumed, without the loss of generality, that $|y_1| \geq |y_2| \geq \cdots \geq |y_n|$.

More Details/Results

For more information and results see

Bolte, J., Sabach, S., Teboulle, M. and Vaisbourd, Y.: **First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems**. SIAM J. Optim., 28(3), 2131-2151. (2018).

Thanks for your attention!

Email: ssabach@ie.technion.ac.il

Website: http://ssabach.net.technion.ac.il/