First Order Optimization Methods

Lecture 7 FOM Beyond Lipschitz Gradient Continuity

Marc Teboulle

School of Mathematical Sciences
Tel Aviv University

PGMO Lecture Series

January 25-26, 2017 Ecole Polytechnique, Paris

Recall: The Basic Pillar underlying FOM

$X = \mathbb{R}^d$ Euclidean with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

$$\inf\{\Phi(x) := f(x) + g(x) : \ x \in X\}, f, g \text{ convex, with } g \text{ smooth.}$$

Key assumption: g admits L-Lipschitz continuous gradient on $\mathbb{R}^d$

A simple, yet crucial consequence of this is the so-called descent Lemma:

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

This inequality naturally provides

1. The upper quadratic approximation of g
2. A crucial pillar in the analysis of **any** current FOM.

Recall: The Basic Pillar underlying FOM

$X = \mathbb{R}^d$ Euclidean with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

$$\inf\{\Phi(x) := f(x) + g(x) : \ x \in X\}, f, g \text{ convex, with } g \text{ smooth.}$$

Key assumption: g admits L-Lipschitz continuous gradient on $\mathbb{R}^d$

A simple, yet crucial consequence of this is the so-called descent Lemma:

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

This inequality naturally provides

1. The upper quadratic approximation of g
2. A crucial pillar in the analysis of **any** current FOM.

However, in many contexts and applications:

⊖ **the differentiable function g does not have a L-smooth gradient[e.g., in the broad class of Poisson inverse problems].**

⊖ **Hence precludes the use of basic FOM methodology and schemes.**

Recall: The Basic Pillar underlying FOM

$X = \mathbb{R}^d$ Euclidean with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

$$\inf\{\Phi(x) := f(x) + g(x) : \ x \in X\}, f, g \text{ convex, with } g \text{ smooth.}$$

Key assumption: g admits L-Lipschitz continuous gradient on $\mathbb{R}^d$

A simple, yet crucial consequence of this is the so-called descent Lemma:

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

This inequality naturally provides

1. The upper quadratic approximation of g
2. A crucial pillar in the analysis of **any** current FOM.

However, in many contexts and applications:

⊖ **the differentiable function g does not have a L-smooth gradient[e.g., in the broad class of Poisson inverse problems].**

⊖ **Hence precludes the use of basic FOM methodology and schemes.**

Lecture 7 FOM without Lipschitz Gradient Continuity

Goal: Circumvent this longstanding and intricate question of Lipschitz continuity required in gradient based methods.

Lecture 7 FOM without Lipschitz Gradient Continuity

Goal: Circumvent this longstanding and intricate question of Lipschitz continuity required in gradient based methods.

- A New Descent Lemma without Lipschitz Gradient Continuity
- Non Euclidean Proximal Distances
- Proximal Gradient Algorithm free of Lipschitz Gradient Assmuption
- Convergence and Complexity
- Examples and Applications

Main Observation: An Elementary Fact

Main Observation: An Elementary Fact

Consider the descent Lemma for the smooth $g \in C_L^{1,1}$ on $\mathbb{R}^d$:

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

Main Observation: An Elementary Fact

Consider the descent Lemma for the smooth $g \in C_L^{1,1}$ on $\mathbb{R}^d$:

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

Simple algebra shows that it can be equivalently written as:

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

Main Observation: An Elementary Fact

Consider the descent Lemma for the smooth $g \in C_L^{1,1}$ on $\mathbb{R}^d$:

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

Simple algebra shows that it can be equivalently written as:

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

Nothing else but the gradient inequality for the convex $\frac{L}{2}\|x\|^2 - g(x)$!

Main Observation: An Elementary Fact

Consider the descent Lemma for the smooth $g \in C_L^{1,1}$ on $\mathbb{R}^d$:

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

Simple algebra shows that it can be equivalently written as:

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

Nothing else but the gradient inequality for the convex $\frac{L}{2}\|x\|^2 - g(x)$!
Thus, for a given smooth convex function g on $\mathbb{R}^d$

$$\text{Descent Lemma} \quad \Longleftrightarrow \quad \frac{\mathbf{L}}{\mathbf{2}}\|\mathbf{x}\|^2 - \mathbf{g}(\mathbf{x}) \text{ is convex on } \mathbb{R}^{\mathbf{d}}.$$

Main Observation: An Elementary Fact

Consider the descent Lemma for the smooth $g \in C_L^{1,1}$ on $\mathbb{R}^d$:

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

Simple algebra shows that it can be equivalently written as:

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

Nothing else but the gradient inequality for the convex $\frac{L}{2}\|x\|^2 - g(x)$!
Thus, for a given smooth convex function g on $\mathbb{R}^d$

$$\text{Descent Lemma} \quad \Longleftrightarrow \quad \frac{\mathbf{L}}{\mathbf{2}}\|\mathbf{x}\|^{\mathbf{2}} - \mathbf{g}(\mathbf{x}) \text{ is convex on } \mathbb{R}^{\mathbf{d}}.$$

To Capture the Geometry of a Constraint set C Naturally suggests to consider - instead of the *squared norm* used for the unconstrained case $C = \mathbb{R}^d$ - a more general convex function that captures the geometry of the constraint.

Trading Gradient Lipschitz Continuity with Convexity

Capturing in a very simple way the geometry of the constraints

Following our basic observation: A convexity condition on the couple (g, h) replaces the usual Lipschitz continuity property required on the gradient of g.

Trading Gradient Lipschitz Continuity with Convexity

Capturing in a very simple way the geometry of the constraints

Following our basic observation: A convexity condition on the couple (g, h) replaces the usual Lipschitz continuity property required on the gradient of g.

A Lipschitz-like/Convexity Condition

$$\text{(LC)} \qquad \exists L > 0 \quad \textbf{with} \quad Lh - g \ \textbf{convex on} \ \operatorname{int} \textbf{dom}\, h,$$

As just seen, when $h(x) = \frac{1}{2}\|x\|^2$, (LC) translates to the Descent Lemma.

Since g is assumed convex, this is equivalent to: ∇g is L-Lipschitz continuous.

Trading Gradient Lipschitz Continuity with Convexity

Capturing in a very simple way the geometry of the constraints

Following our basic observation: A convexity condition on the couple (g, h) replaces the usual Lipschitz continuity property required on the gradient of g.

A Lipschitz-like/Convexity Condition

$$\text{(LC)} \qquad \exists L > 0 \quad \textbf{with} \quad Lh - g \ \textbf{convex on} \ \operatorname{int} \mathbf{dom}\, h,$$

As just seen, when $h(x) = \frac{1}{2}\|x\|^2$, (LC) translates to the Descent Lemma.

Since g is assumed convex, this is equivalent to: ∇g is L-Lipschitz continuous.

- We shall see, that the mere translation of condition (LC) into its first-order characterization immediately yields **the new descent Lemma** we seek for.
- It naturally leads to the **Non Euclidean Proximal Bregman distance**, we introduce next.

Bregman Proximal Distance

Defintion: Bregman distance [Bregman (67)] Let $h : X \to (-\infty, \infty]$ be a closed proper strictly convex function, differentiable on $\operatorname{int}\operatorname{dom} h$. The Bregman distance associated to h (or with kernel h) is defined by

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

Bregman Proximal Distance

Defintion: Bregman distance [Bregman (67)] Let $h : X \to (-\infty, \infty]$ be a closed proper strictly convex function, differentiable on $\operatorname{int}\operatorname{dom} h$. The Bregman distance associated to h (or with kernel h) is defined by

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

Geometrically, it measures the vertical difference between $h(x)$, the value at x of a linearized approximation of h around y.

Proposition: Distance-Like Properties

- D_h is strictly convex with respect to its first argument.
- $D_h(x, y) \geq 0$ and "$= 0$" iff $x = y$.

Proof. Immediate by the gradient inequality. □

Thus, D_h provides a natural distance measure .

However, note that D_h is in general asymmetric.

First Examples

- **Example 1** The choice $h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|^2, \operatorname{dom} h = \mathbb{R}^d$ yields the usual squared Euclidean norm distance $D_h(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$.
- **Example 2** The entropy-like distance defined on the simplex,

$$h(\mathbf{z}) = \sum_{j=1}^{d} z_j \ln z_j, \text{ for } \mathbf{z} \in \operatorname{dom} h := \Delta_d = \{\mathbf{z} \in \mathbb{R}^d : \sum_{j=1}^{d} z_j = 1, \mathbf{z} \geq \mathbf{0}\}.$$

- In that case, $D_h(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} x_j \ln \frac{x_j}{y_j}$.

More examples soon...

Legendre Functions - Useful Device to Handle constraints

Strategy to handle a constraint set is standard: Pick a Legendre function on C.

Definition (Legendre functions)[Rockafellar '70]. $h : X \to (-\infty, \infty]$, lsc proper convex is called *Legendre type* if h is essentially smooth and strictly convex on $\operatorname{int}\operatorname{dom} h$.

Recall

- *Essentially smooth:* if h is differentiable on $\operatorname{int}\operatorname{dom} h$, with $\|\nabla h(x^k)\| \to \infty$ for every sequence $\{x^k\}_{k\in\mathbb{N}} \subset \operatorname{int}\operatorname{dom} h$ converging to a boundary point of $\operatorname{dom} h$ as $k \to +\infty$.
- ∇h is a *bijection* from $\operatorname{int}\operatorname{dom} h \to \operatorname{int}\operatorname{dom} h^*$ and

$$(\nabla h)^{-1} = \nabla h^*$$

where $h^*(u) := \sup_v\{\langle u, v\rangle - h(v)\}$ is the Fenchel conjugate of h.

A Descent Lemma without Lipschitz Gradient Continuity

Lemma[Descent lemma without Lipschitz Gradient Continuity]
Let $h : X \to (-\infty, \infty]$ be a Legendre function, and $g : X \to (-\infty, \infty]$ be convex function with $\operatorname{dom} g \supset \operatorname{dom} h$ which is C^1 on $\operatorname{int} \operatorname{dom} h$.

Then, the condition **(LC):** $Lh - g$ **convex on** $\operatorname{int} \operatorname{dom}\ h$ **is equivalent to**

$$g(x) \leq g(y) + \langle \nabla g(y), x - y \rangle + LD_h(x, y),\ \forall (x, y) \in \operatorname{int} \operatorname{dom}\ h \times \operatorname{int} \operatorname{dom}\ h$$

where, D_h stands for the Bregman Distance associated to h.

A Descent Lemma <u>without</u> Lipschitz Gradient Continuity

Lemma[Descent lemma without Lipschitz Gradient Continuity]
Let $h : X \to (-\infty, \infty]$ be a Legendre function, and $g : X \to (-\infty, \infty]$ be convex function with $\operatorname{dom} g \supset \operatorname{dom} h$ which is C^1 on $\operatorname{int} \operatorname{dom} h$.

Then, the condition **(LC):** $Lh - g$ **convex on** $\operatorname{int} \operatorname{dom} h$ **is equivalent to**

$$g(x) \le g(y) + \langle \nabla g(y), x - y \rangle + LD_h(x, y), \ \forall (x, y) \in \operatorname{int} \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h$$

where, D_h stands for the Bregman Distance associated to h.

Proof. Simply apply the gradient inequality for the convex function $Lh - g$:

- $Lh(y) - g(y) - (Lh(x) - g(x)) \le \langle L\nabla h(y) - \nabla g(y), y - x \rangle$
- $g(x) - g(y) - \langle \nabla g(y), x - y \rangle \le L(h(x) - h(y) - \langle \nabla h(y), x - y \rangle) = LD_h(x, y).$

A Descent Lemma without Lipschitz Gradient Continuity

Lemma[Descent lemma without Lipschitz Gradient Continuity]
Let $h : X \to (-\infty, \infty]$ be a Legendre function, and $g : X \to (-\infty, \infty]$ be convex function with $\operatorname{dom} g \supset \operatorname{dom} h$ which is C^1 on $\operatorname{int}\operatorname{dom} h$.

Then, the condition **(LC):** $Lh - g$ **convex on** $\operatorname{int}\operatorname{dom}\ h$ **is equivalent to**

$$g(x) \leq g(y) + \langle \nabla g(y), x - y \rangle + LD_h(x, y),\ \forall (x, y) \in \operatorname{int}\operatorname{dom}\ h \times \operatorname{int}\operatorname{dom}\ h$$

where, D_h stands for the Bregman Distance associated to h.

Proof. Simply apply the gradient inequality for the convex function $Lh - g$:

- $Lh(y) - g(y) - (Lh(x) - g(x)) \leq \langle L\nabla h(y) - \nabla g(y), y - x \rangle$
- $g(x) - g(y) - \langle \nabla g(y), x - y \rangle \leq L(h(x) - h(y) - \langle \nabla h(y), x - y \rangle) = LD_h(x, y)$.

Compactly, $\forall (x, y) \in \operatorname{int}\operatorname{dom}\ h \times \operatorname{int}\operatorname{dom}\ h$

$$Lh - g \text{ **convex** } \iff D_g(x, y) \leq LD_h(x, y) \iff D_{Lh-g} \geq 0.$$

□

Some Useful Examples for Bregman Distances D_h

Each example is a one dimensional h which is Legendre. The corresponding Legendre function $\tilde{h}$ and Bregman distance in $\mathbb{R}^d$ simply use the formulae

$$\tilde{h}(x) = \sum_{j=1}^{n} h(x_j) \text{ and } D_{\tilde{h}}(x, y) = \sum_{j=1}^{n} D_h(x_j, y_j).$$

Name	h	dom h
Energy	$\frac{1}{2}x^2$	$\mathbb{R}$
Boltzmann-Shannon entropy	$x \log x$	$[0, \infty]$
Burg's entropy	$-\log x$	$(0, \infty)$
Fermi-Dirac entropy	$x \log x + (1-x)\log(1-x)$	$[0, 1]$
Hellinger	$-(1-x^2)^{1/2}$	$[-1, 1]$
Fractional Power	$(px - x^p)/(1-p), p \in (0,1)$	$[0, \infty)$

- **Other possible kernels** h**:** Nonseparable Bregman, and for handling cone constraints e.g., PSD matrices, Lorentz cone etc.., see refs. for details.

(LC) There exists $L > 0$: $Lh - g$ Convex - First Examples

(LC) admits alternative reformulations which facilitates its checking; (see paper).

A useful one, is in the 1D case, with h is C^2, $h'' > 0$ on $\operatorname{int}\operatorname{dom} h$. In this case :

$$(LC) \qquad \text{is equivalent to} \qquad \sup\left\{ \frac{g''(x)}{h''(x)} : \; x \in \operatorname{int}\operatorname{dom} h \right\} < \infty.$$

(LC) There exists $L > 0$: $Lh - g$ Convex - First Examples

(LC) admits alternative reformulations which facilitates its checking; (see paper).

A useful one, is in the 1D case, with h is C^2, $h'' > 0$ on $\operatorname{int}\operatorname{dom} h$. In this case :

$$(LC) \qquad \text{is equivalent to} \qquad \sup\left\{\frac{g''(x)}{h''(x)} : \; x \in \operatorname{int}\operatorname{dom} h\right\} < \infty.$$

Two examples with g is C^2 which **does not** have a classical L-smooth gradient, yet where (LC) holds.

- Let h be the Fermi-Dirac entropy. Then, (LC) reads

$$\sup_{0<x<1} x(1-x)g''(x) < \infty,$$

  which clearly holds when $[0, 1] \subseteq \operatorname{int}\operatorname{dom} g$.
  For instance, this holds with $g(x) = x \log x$ which *does not* have a Lipschitz gradient.

- Let h be the Burg's entropy, and $g(x) = -\log x$ which *does not* have a Lipschitz gradient. Then, (LC) trivially holds!

More examples in important applications soon...

The Problem and Blanket Assumption

Our aim is to solve the composite convex problem

$$v(\mathcal{P}) = \inf\{\Phi(x) := f(x) + g(x) \mid x \in \overline{\text{dom}}\, h\},$$

where $\overline{\text{dom}}\, h \equiv C$ denotes the closure of $\text{dom}\, h$.

The following is our blanket assumption.

Standard..but now the "Hidden h" will handle constraint C...

Blanket Assumption

(i) $g : X \to (-\infty, \infty]$ is proper lower semicontinuous (lsc) convex,

(ii) $h : X \to (-\infty, \infty]$ is proper, lsc convex, and Legendre.

(iii) $f : X \to (-\infty, \infty]$ is proper lsc convex with $\text{dom}\, g \supset \text{dom}\, h$, which is differentiable on $\text{int}\, \text{dom}\, h$,

(iv) $\text{dom}\, f \cap \text{int}\, \text{dom}\, h \neq \emptyset$,

(v) Solution set $\mathcal{S}_* := \text{argmin}\{\Phi(x) : x \in C = \overline{\text{dom}}\, h\} \neq \emptyset$.

Algorithm NoLips for $\inf\{f(x)+g(x) : x \in C\}$

Main Algorithmic Operator– **[Reduces to classical prox-grad, when h quadratic]**

$$\mathbf{T}_\lambda(\mathbf{x}) := \operatorname{argmin}\left\{\mathbf{f}(\mathbf{u}) + \mathbf{g}(\mathbf{x}) + \langle \nabla \mathbf{g}(\mathbf{x}), \mathbf{u}-\mathbf{x}\rangle + \frac{1}{\lambda}\mathbf{D_h}(\mathbf{u},\mathbf{x}) : \mathbf{u} \in \mathbf{X}\right\}.$$

Algorithm – NoLips

0. **Input.** Choose a Legendre function h with $C = \overline{\operatorname{dom}}\, h$ such that there exists $L > 0$ with $Lh - g$ convex on $\operatorname{int}\operatorname{dom} h$.
1. **Initialization.** Start with any $x^0 \in \operatorname{int}\operatorname{dom} h$.
2. **Recursion.** For each $k \geq 1$ with $\lambda_k > 0$, generate $\left\{x^k\right\}_{k\in\mathbb{N}} \in \operatorname{int}\operatorname{dom} h$ via

$$x^k = T_{\lambda_k}(x^{k-1}) = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{f(x) + \left\langle \nabla g(x^{k-1}), x - x^{k-1}\right\rangle + \frac{1}{\lambda_k}D_h(x, x^{k-1})\right\}$$

We shall systematically assume that $T_\lambda \neq \emptyset$, single-valued and maps $\operatorname{int}\operatorname{dom}\, h$ **in** $\operatorname{int}\operatorname{dom}\, h$**.**

More precise technical details, see our paper.

Main Issues / Questions for NoLips

- **Computation of $T_\lambda(\cdot)$?**
- **What is the complexity of NoLips?**
- **Does it converge? What is the step size λ_k?**

NoLips – Decomposition of $T_\lambda(\cdot)$ into Elementary Steps

T_λ shares the same structural decomposition as the usual proximal gradient.
It splits into *"elementary" steps* useful for computational purposes.

NoLips – Decomposition of $T_\lambda(\cdot)$ into Elementary Steps

T_λ shares the same structural decomposition as the usual proximal gradient.
It splits into *"elementary" steps* useful for computational purposes.

⊕ **Define Bregman gradient step**

$$p_\lambda(x) := \text{argmin}\left\{\langle \nabla g(x), u\rangle + \frac{1}{\lambda} D_h(u,x) : u \in X\right\} \equiv \nabla h^*(\nabla h(x) - \lambda \nabla g(x))$$

Clearly reduces to the usual explicit gradient step when $h = \frac{1}{2}\|\cdot\|^2$.

⊕ **Define the proximal Bregman operator**

$$\text{prox}^h_{\lambda f}(y) := \text{argmin}\left\{\lambda f(u) + D_h(u,y) : u \in \mathbb{R}^d\right\},\ y \in \text{int dom } h$$

Then, one can show (simply write optimality condition) that **NoLips** simply reduces to the
composition of a Bregman proximal step with a Bregman gradient step:

NoLips Main Iteration: $x \in \text{int dom } h, \qquad x^+ = \text{prox}^h_{\lambda f} \circ p_\lambda(x) \ \ (\lambda > 0)$

Examples for Bregman Gradient Step $p_\lambda(x) = \nabla h^*(v(x))$

Let $v(x) := \nabla h(x) - \lambda \nabla g(x)$.

1. *Regularized Burg's Entropy - Nonnegative Constraints*. Here all computations are 1-D. $h(t) := \frac{\sigma}{2}t^2 - \mu \log t$ with $\operatorname{dom} h = (0, \infty)$, $(\sigma, \mu > 0)$. Then, on can show that $\operatorname{dom} h^* = \mathbb{R}$,

$$\nabla h^*(s) = (\sigma\rho^2(s) + \mu)(s^2 + 4\mu\sigma)^{-1/2},\ \rho(s) := \frac{s + \sqrt{s^2 + 4\mu\sigma}}{2\sigma} > 0.$$

Examples for Bregman Gradient Step $p_\lambda(x) = \nabla h^*(v(x))$

Let $v(x) := \nabla h(x) - \lambda \nabla g(x)$.

1. *Regularized Burg's Entropy - Nonnegative Constraints*. Here all computations are 1-D. $h(t) := \frac{\sigma}{2}t^2 - \mu \log t$ with $\operatorname{dom} h = (0, \infty)$, $(\sigma, \mu > 0)$. Then, on can show that $\operatorname{dom} h^* = \mathbb{R}$,

$$\nabla h^*(s) = (\sigma \rho^2(s) + \mu)(s^2 + 4\mu\sigma)^{-1/2},\ \rho(s) := \frac{s + \sqrt{s^2 + 4\mu\sigma}}{2\sigma} > 0.$$

2. *Hellinger-Like function - Ball Constraints.*
$h(x) = -\sqrt{1 - \|x\|^2}$; $\operatorname{dom} h = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ yields a nonseparable Bregman distance which is relevant for ball constraints. We then obtain,

$$p_\lambda(x) = (1 + v^2(x))^{-1/2} v(x);\ \operatorname{dom} h^* = \mathbb{R}^n.$$

Examples for Bregman Gradient Step $p_\lambda(x) = \nabla h^*(v(x))$

Let $v(x) := \nabla h(x) - \lambda \nabla g(x)$.

1. *Regularized Burg's Entropy - Nonnegative Constraints*. Here all computations are 1-D. $h(t) := \frac{\sigma}{2} t^2 - \mu \log t$ with $\operatorname{dom} h = (0, \infty)$, $(\sigma, \mu > 0)$. Then, on can show that $\operatorname{dom} h^* = \mathbb{R}$,

$$\nabla h^*(s) = (\sigma \rho^2(s) + \mu)(s^2 + 4\mu\sigma)^{-1/2},\ \rho(s) := \frac{s + \sqrt{s^2 + 4\mu\sigma}}{2\sigma} > 0.$$

2. *Hellinger-Like function - Ball Constraints.*
$h(x) = -\sqrt{1 - \|x\|^2}$; $\operatorname{dom} h = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ yields a nonseparable Bregman distance which is relevant for ball constraints. We then obtain,

$$p_\lambda(x) = (1 + v^2(x))^{-1/2} v(x);\ \operatorname{dom} h^* = \mathbb{R}^n.$$

3. *Conic constraints.* Bregman distances can be defined on S^d.
⊕ Example 1 – SDP Constraints: $h(x) = -\log\det(x)$, $\operatorname{dom} h = S^d_{++}$. Then we obtain,

$$p_\lambda(x) = v(x)^{-1},\ v(x),\ x \in S^d_{++}.$$

⊕ Example 2 – SOC Constraints: can be similarly handled with adequate h.

Some Examples for $\mathrm{prox}_{\lambda f}^{h}(y)$

1. **Entropic thresholding.** Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = x \log x, \ \mathrm{dom}\, h = [0, \infty)$. Then,

$$\mathrm{prox}_{\lambda f}^{h}(y) = \begin{cases} \exp(\lambda) y & \text{if } y < \exp(-\lambda) a, \\ a & \text{if } y \in [\exp(-\lambda) a, \exp(\lambda) a], \\ \exp(-\lambda) y & \text{if } y > \exp(\lambda) a. \end{cases}$$

Some Examples for $\mathrm{prox}_{\lambda f}^{h}(y)$

1. **Entropic thresholding.** Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = x \log x, \; \mathrm{dom}\, h = [0, \infty)$. Then,

$$\mathrm{prox}_{\lambda f}^{h}(y) = \begin{cases} \exp(\lambda) y & \text{if } y < \exp(-\lambda) a, \\ a & \text{if } y \in [\exp(-\lambda) a, \exp(\lambda) a], \\ \exp(-\lambda) y & \text{if } y > \exp(\lambda) a. \end{cases}$$

2. **Log thresholding.** Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = -\log x, \; \mathrm{dom}\, h = (0, \infty)$. Assume $\lambda a < 1$. Then,

$$\mathrm{prox}_{\lambda f}^{h}(y) = \begin{cases} \frac{y}{1+\lambda y} & \text{if } y < \frac{a}{1-\lambda a}, \\ a & \text{if } y \in \left[\frac{a}{1-\lambda a}, \frac{a}{1+\lambda a}\right], \\ \frac{y}{1-\lambda y} & \text{if } y > \frac{a}{1+\lambda a}. \end{cases}$$

Similar formulas may be derived when $\lambda a > 1$.

Some Examples for $\mathrm{prox}^h_{\lambda f}(y)$

1. **Entropic thresholding.** Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = x \log x,$ dom $h = [0, \infty).$ Then,

$$\mathrm{prox}^h_{\lambda f}(y) = \begin{cases} \exp(\lambda) y & \text{if } y < \exp(-\lambda) a, \\ a & \text{if } y \in [\exp(-\lambda) a, \exp(\lambda) a], \\ \exp(-\lambda) y & \text{if } y > \exp(\lambda) a. \end{cases}$$

2. **Log thresholding.** Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = -\log x,$ dom $h = (0, \infty).$ Assume $\lambda a < 1.$ Then,

$$\mathrm{prox}^h_{\lambda f}(y) = \begin{cases} \frac{y}{1+\lambda y} & \text{if } y < \frac{a}{1-\lambda a}, \\ a & \text{if } y \in \left[\frac{a}{1-\lambda a}, \frac{a}{1+\lambda a}\right], \\ \frac{y}{1-\lambda y} & \text{if } y > \frac{a}{1+\lambda a}. \end{cases}$$

   Similar formulas may be derived when $\lambda a > 1.$

3. **Exponential.** Let $f(u) = c e^u,$ $c > 0$, and take $h(x) = e^x,$ dom $h = \mathbb{R}.$ Then $\mathrm{prox}^h_{\lambda f}(y) = y - \log(1 + \lambda c).$

Analysis of NoLips: Relies on 3 Basic Results

A Key Property for D_h : Pythagoras...Without Squares!

- A very simple, but key property of Bregman distances.
- Plays a crucial role in the analysis of any optimization method based on Bregman distances.

Lemma (The three points identity)

For any three points $\mathbf{x}, \mathbf{y} \in int(dom\, h)$ and $\mathbf{u} \in dom\, h$, the following three points identity holds true

$$D_h(\mathbf{u}, \mathbf{y}) - D_h(\mathbf{u}, \mathbf{x}) - D_h(\mathbf{x}, \mathbf{y}) = \langle \nabla h(\mathbf{y}) - \nabla h(\mathbf{x}), \mathbf{x} - \mathbf{u} \rangle.$$

Proof. Simply follows by using the definition of D_h! □

With $h(\mathbf{u}) := \|\mathbf{u}\|^2/2$ we recover the classical Pythagoras/Triangle identity:

$$\|\mathbf{z} - \mathbf{y}\|^2 - \|\mathbf{z} - \mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 = 2\langle \mathbf{z} - \mathbf{x}, \mathbf{x} - \mathbf{y} \rangle.$$

Bregman Based Proximal Inequality

Extends a similar property of the Euclidean squared prox.

Lemma. Let $\varphi : X \to (-\infty, \infty]$ be a closed proper convex function. Given $t > 0$, and $\mathbf{z} \in \operatorname{int}\operatorname{dom} h$, define:

$$\mathbf{u}^+ := \underset{\mathbf{u}\in\mathbb{E}}{\operatorname{argmin}} \left\{ \varphi(\mathbf{u}) + \frac{1}{t} D_h(\mathbf{u}, \mathbf{z}) \right\}.$$

Then, $t(\varphi(\mathbf{u}^+) - (\mathbf{u})) \leq [D_h(\mathbf{u}, \mathbf{z}) - D_h(\mathbf{u}, \mathbf{u}^+) - D_h(\mathbf{u}^+, \mathbf{z})], \forall \mathbf{u} \in \operatorname{dom} h$.

Bregman Based Proximal Inequality

Extends a similar property of the Euclidean squared prox.

Lemma. Let $\varphi : X \to (-\infty, \infty]$ be a closed proper convex function. Given $t > 0$, and $\mathbf{z} \in \operatorname{int} \operatorname{dom} h$, define:

$$\mathbf{u}^+ := \underset{\mathbf{u} \in \mathbb{E}}{\operatorname{argmin}} \left\{ \varphi(\mathbf{u}) + \frac{1}{t} D_h(\mathbf{u}, \mathbf{z}) \right\}.$$

Then, $t(\varphi(\mathbf{u}^+) - (\mathbf{u})) \leq [D_h(\mathbf{u}, \mathbf{z}) - D_h(\mathbf{u}, \mathbf{u}^+) - D_h(\mathbf{u}^+, \mathbf{z})], \forall \mathbf{u} \in \operatorname{dom} h.$

Proof. $\mathbf{u} \mapsto t\varphi(\mathbf{u}) + D_h(\mathbf{u}, \mathbf{z})$ is strictly convex with unique minimizer $\mathbf{u}^+$ characterized via optimality condition. For any $\mathbf{u} \in \operatorname{dom} h$:

$$\langle t\boldsymbol{\omega} + \nabla_{\mathbf{u}} D_h(\mathbf{u}^+, \mathbf{z}), \mathbf{u} - \mathbf{u}^+ \rangle \geq 0, \; \boldsymbol{\omega} \in \partial\varphi(\mathbf{u}^+).$$

Bregman Based Proximal Inequality

Extends a similar property of the Euclidean squared prox.

Lemma. Let $\varphi : X \to (-\infty, \infty]$ be a closed proper convex function. Given $t > 0$, and $\mathbf{z} \in \operatorname{int} \operatorname{dom} h$, define:

$$\mathbf{u}^+ := \underset{\mathbf{u}\in\mathbb{E}}{\operatorname{argmin}} \left\{ \varphi(\mathbf{u}) + \frac{1}{t} D_h(\mathbf{u}, \mathbf{z}) \right\}.$$

Then, $t(\varphi(\mathbf{u}^+) - (\mathbf{u})) \leq [D_h(\mathbf{u}, \mathbf{z}) - D_h(\mathbf{u}, \mathbf{u}^+) - D_h(\mathbf{u}^+, \mathbf{z})], \forall \mathbf{u} \in \operatorname{dom} h.$

Proof. $\mathbf{u} \mapsto t\varphi(\mathbf{u}) + D_h(\mathbf{u}, \mathbf{z})$ is strictly convex with unique minimizer $\mathbf{u}^+$ characterized via optimality condition. For any $\mathbf{u} \in \operatorname{dom} h$:

$$\langle t\boldsymbol{\omega} + \nabla_{\mathbf{u}} D_h(\mathbf{u}^+, \mathbf{z}), \mathbf{u} - \mathbf{u}^+ \rangle \geq 0, \ \boldsymbol{\omega} \in \partial\varphi(\mathbf{u}^+).$$

Since $\nabla_{\mathbf{u}} D_h(\mathbf{u}^+, \mathbf{z}) = \nabla h(\mathbf{u}^+) - \nabla h(\mathbf{z})$, rearranging above reads as:

- $t\langle \boldsymbol{\omega}, \mathbf{u}^+ - \mathbf{u} \rangle \leq \langle \nabla h(\mathbf{u}^+) - \nabla h(\mathbf{z}), \mathbf{u} - \mathbf{u}^+ \rangle,$
- φ is convex: $\Rightarrow \ t(\varphi(\mathbf{u}^+) - \varphi(\mathbf{u})) \leq t\langle \boldsymbol{\omega}, \mathbf{u}^+ - \mathbf{u} \rangle.$
- Combine above: $t(\varphi(\mathbf{u}^+) - \varphi(\mathbf{u})) \leq \langle \nabla h(\mathbf{z}) - \nabla h(\mathbf{u}^+), \mathbf{u}^+ - \mathbf{u} \rangle$
- Invoke the three points identity for D_h gives the desired result. □

Key Estimation Inequality for $\Phi = f + g$

Lemma (Descent inequality for NoLips)

Let $\lambda > 0$. For all x in $\operatorname{int} dom\, h$, let $x^+ := T_\lambda(x)$. Then,

$$\lambda \left(\Phi(x^+) - \Phi(u)\right) \le D_h(u,x) - D_h(u,x^+) - (1-\lambda L) D_h(x^+,x), \ \forall u \in dom\, h.$$

Key Estimation Inequality for $\Phi = f + g$

Lemma (Descent inequality for NoLips)

Let $\lambda > 0$. For all x in $\text{int}\, dom\, h$, let $x^+ := T_\lambda(x)$. Then,

$$\lambda \left(\Phi(x^+) - \Phi(u)\right) \leq D_h(u,x) - D_h(u,x^+) - (1-\lambda L) D_h(x^+,x), \ \forall u \in dom\, h.$$

Proof. Fix any $x \in \text{int}\,\text{dom}\, h$. With $(x^+, u, x) \in \text{int}\,\text{dom}\ h \times \text{dom}\, h \times \text{int}\,\text{dom}\ h)$, we apply Appy the B-prox inequality to

$$u \to \varphi(u) := f(u) + g(x) + \langle \nabla g(x), u - x \rangle,$$

, followed by the NL-Lemma, and the convexity of g to obtain for every $u \in \text{dom}\, h$:

$$\begin{aligned}
\lambda(f(x^+) - f(u)) &\leq \lambda \langle \nabla g(x), u - x^+ \rangle + D_h(u,x) - D_h(u,x^+) - D_h(x^+,x) \\
\lambda(g(x^+) - g(x)) &\leq \lambda \langle \nabla g(x), x^+ - x \rangle + \lambda L D_h(x^+,x) \\
\lambda(g(x) - g(u)) &\leq \lambda \langle \nabla g(x), x - u \rangle.
\end{aligned}$$

Add the 3 inequalities, recalling that $\Phi(x) = f(x) + g(x)$, we thus obtain

$$\lambda \left(\Phi(x^+) - \Phi(u)\right) \leq D_h(u,x) - D_h(u,x^+) - (1 - \lambda L) D_h(x^+,x). \quad \square$$

Complexity for NoLips: $O(1/k)$

Theorem (NoLips: Complexity)

(i) **(Global estimate in function values)** *Let* $\{x^k\}_{k\in\mathbb{N}}$ *be the sequence generated by* NoLips *with* $\lambda \in (0, 1/L]$. *Then*

$$\Phi(x^k) - \Phi(u) \leq \frac{LD_h(u, x^0)}{k} \qquad \forall u \in dom\, h.$$

(ii) **(Complexity for h with closed domain)** *Assume in addition, that* $dom\, h = \overline{\mathrm{dom}}\, h$ *and that* $(\mathcal{P})$ *has at least a solution. Then for any solution* $\bar{x}$ *of* $(\mathcal{P})$,

$$\Phi(x^k) - \min_C \Phi \leq \frac{LD_h(\bar{x}, x^0)}{k}$$

Notes ♢ When $h(x) = \frac{1}{2}\|x\|^2, g \in C_L^{1,1}$, and we thus recover the classical sublinear global rate of the usual proximal gradient method.

♢ The entropies of Boltzmann-Shannon, Fermi-Dirac and Hellinger are non trivial examples for which the assumption ($\overline{\mathrm{dom}}\, h = \mathrm{dom}\, h$) is obviously satisfied.

Proof of $O(1/k)$ Complexity for NoLips

Fix $k \geq 1$. Using our Descent inequality Lemma with $x^k = T_\lambda(x^{k-1})$, and $\lambda \leq 1/L$, we obtain, for all $u \in \operatorname{dom} h$,

$$\Phi(x^k) - \Phi(u) \leq LD_h(u, x^{k-1}) - LD_h(u, x^k) \tag{1}$$

The claims easily follow from this inequality. Set $u = x^{k-1}$ in (1) we get

- $\Phi(x^k) - \Phi(x^{k-1}) \leq 0 \;\Rightarrow\; \sum_{k=1}^{n}(k-1)\{\Phi(x^k) - \Phi(x^{k-1})\} \leq 0$
- which reads $-\sum_{k=1}^{n} \Phi(x^k) + \sum_{k=1}^{n} k\Phi(x^k) - (k-1)\Phi(x^{k-1}) \leq 0$
- and hence, $-\sum_{k=1}^{n} \Phi(x^k) + n\Phi(x^n) \leq 0$.
- Sum (1) $\sum_{k=1}^{n} \Phi(x^k) - n\Phi(u) \leq LD_h(u, x^0) - LD_h(u, x^n) \leq LD(u, x^0)$.
- Add the above, proves (a), and when $\operatorname{dom} h = \overline{\operatorname{dom}}\, h$, plug $u = x^*$ yields (b). □

Note: One can also deduce *pointwise convergence* for NoLips:

$$\{x^k\}_{k \in \mathbb{N}} \text{ converges to some solution } x^* \text{ of } (\mathcal{P})$$

via a more precise analysis, and with dynamic step-size λ_k expressed in terms of a symmetry measure for D_h, see the paper for details.

Applications: A Prototype Broad Class of Problems with Poisson Noise

A very large class of problems arising in Statistical and Image Sciences areas: inverse problems where data measurements are collected by counting discrete events (e.g., photons, electrons) contaminated by noise described by a Poisson process.

One then needs to recover a nonnegative signal/image for the given problem.

Huge amount of literature in many contexts:

- Astronomy,
- Nuclear medicine (PET)-Positron Emission Tomography; electronic microscropy,
- Statistical estimation (EM)-Expectation Maximization,
- Image deconvolution, denoising speckle (multiplicative) noise, etc...

Linear Inverse Problems - The Optimization Model

Problem:

- **Given a matrix $A \in \mathbb{R}_+^{m \times n}$ describing the experimental protocol.**
- $b \in \mathbb{R}_{++}^m$ **is given vector of measurements.**
- **The goal is to reconstruct the signal $x \in \mathbb{R}_+^n$ from the noisy measurements b such that**

$$Ax \simeq b.$$

Moreover, there is often a need to regularize the problem through an appropriate choice of a regularizer f reflecting desired features of the solution.

Optimization Model to Recover x

$$(\mathbb{E}) \qquad \textbf{minimize} \quad \{\mathcal{D}(b, Ax) + \mu f(x) : \; x \in \mathbb{R}_+^n\}$$

$\oplus$ $\mathcal{D}(\cdot,\cdot)$ a convex proximity measure that quantifies the "error" between b and Ax

$\oplus$ $\mu > 0$ controls the tradeoff between matching the data fidelity criteria and the weight given to its regularizer. ($\mu = 0$ when no regularizer needed.)

NoLips in Action : New Simple Schemes for Many Problems

The optimization problem will be of the form:

$$(\mathbb{E}) \qquad \min_{x}\{f(x) + \mathcal{D}_\phi(b, Ax)\} \;\text{ or }\; \min_{x}\{f(x) + \mathcal{D}_\phi(Ax, b)\}$$

for some convex ϕ, and $f(x)$ some nonsmooth convex regularizer.

NoLips in Action : New Simple Schemes for Many Problems

The optimization problem will be of the form:

$$(\mathbb{E}) \qquad \min_x \{f(x) + \mathcal{D}_\phi(b, Ax)\} \text{ or } \qquad \min_x \{f(x) + \mathcal{D}_\phi(Ax, b)\}$$

for some convex ϕ, and $f(x)$ some nonsmooth convex regularizer.

To apply NoLips :

1. Pick an h, to warrant an L in terms of problem's data, s.t. $Lh - g$ convex.
2. In turns, this determines the step-size λ defined through $\lambda \in (0, L^{-1}]$.
3. Compute $p_\lambda(\cdot)$ and $\mathrm{prox}^h_{\lambda f}(\cdot))$ – Bregman-like [gradient and proximal] steps.

Resulting algorithms for which our results can be applied lead to

Simple schemes via explicit map $M_j(\cdot)$:

$$x > 0, \qquad x_j^+ = M_j(b, A, x) \cdot x_j, \qquad j = 1, \ldots, n,$$

with (λ, L) determined in terms of the problem data (A, b).

A Typical Linear Inverse Problem with Poisson Noise

A natural proximity measure in $\mathbb{R}^n_+$ - Kullback-Liebler Relative Entropy:

$$D_\phi(b, Ax) \equiv \mathcal{D}(b, Ax) := \sum_{i=1}^{m} \{b_i \log \frac{b_i}{(Ax)_i} + (Ax)_i - b_i\}, \ (\phi(u) = \sum_{i=1}^{m} u_i \log u_i)$$

which (up to some constants) corresponds to the negative Poisson log-likelihood function.

A Typical Linear Inverse Problem with Poisson Noise

A natural proximity measure in $\mathbb{R}^n_+$ - Kullback-Liebler Relative Entropy:

$$D_\phi(b, Ax) \equiv \mathcal{D}(b, Ax) := \sum_{i=1}^{m} \{b_i \log \frac{b_i}{(Ax)_i} + (Ax)_i - b_i\}, \ (\phi(u) = \sum_{i=1}^{m} u_i \log u_i)$$

which (up to some constants) corresponds to the negative Poisson log-likelihood function.

- The optimization problem:

$$(\mathbb{E}) \qquad \text{minimize} \ \ \{g(x) + \mu f(x) : \ x \in \mathbb{R}^n_+\}$$

- $g(x) \equiv \mathcal{D}(d, Ax)$, and f a regularizer, possibly nonsmooth
- $x \to \mathcal{D}(b, Ax)$ convex, **but does not admit a globally Lipschitz continuous gradient.**

Two Simple Algorithms for Poisson Linear Inverse Problems

Given $g(x) := D_\phi(b, Ax)$ ($\phi(u) = u \log u$), **to apply NoLips**, we need to identify an adequate h.

- We take the Burg's entropy $h(x) = -\sum_{j=1}^{n} \log x_j,$ $\operatorname{dom} h = \mathbb{R}^n_{++}$.
- We need to find $L > 0$ such that $Lh - g$ is convex in $\mathbb{R}^n_{++}$.

Two Simple Algorithms for Poisson Linear Inverse Problems

Given $g(x) := D_\phi(b, Ax)$ ($\phi(u) = u \log u$), **to apply NoLips**, we need to identify an adequate h.

- We take the Burg's entropy $h(x) = -\sum_{j=1}^n \log x_j,\ \mathrm{dom}\, h = \mathbb{R}^n_{++}$.
- We need to find $L > 0$ such that $Lh - g$ is convex in $\mathbb{R}^n_{++}$.

Lemma. Let $g(x) = D_\phi(b, Ax)$ and $h(x)$ as defined above. Then,

$$\text{for any } L \geq \|b\|_1 = \sum_{i=1}^m b_i, \text{ the function } Lh - g \text{ is convex on } \mathbb{R}^n_{++}.$$

Two Simple Algorithms for Poisson Linear Inverse Problems

Given $g(x) := D_\phi(b, Ax)$ ($\phi(u) = u \log u$), **to apply NoLips**, we need to identify an adequate h.

- We take the Burg's entropy $h(x) = -\sum_{j=1}^n \log x_j$, $\operatorname{dom} h = \mathbb{R}^n_{++}$.
- We need to find $L > 0$ such that $Lh - g$ is convex in $\mathbb{R}^n_{++}$.

Lemma. Let $g(x) = D_\phi(b, Ax)$ and $h(x)$ as defined above. Then,

$$\text{for any } L \geq \|b\|_1 = \sum_{i=1}^m b_i, \text{ the function } Lh - g \text{ is convex on } \mathbb{R}^n_{++}.$$

Thus, we can take $\lambda = L^{-1} = \|b\|_1^{-1}$.

Applying NoLips, given $x \in \mathbb{R}^n_{++}$, the main algorithmic step $x^+ = T_\lambda(x)$ is then:

$$x^+ = \operatorname{argmin} \left\{ \mu f(u) + \langle \nabla g(x), u \rangle + \frac{1}{\lambda} \sum_{j=1}^n \left(\frac{u_j}{x_j} - \log \frac{u_j}{x_j} - 1 \right) : u > 0 \right\}.$$

We now show that the above abstract iterative process yields closed form algorithms for Poisson reconstruction problems with two typical regularizers used in applications.

Example 1 – Sparse Poisson Linear Inverse Problem

Sparse regularization. Let $f(x) := \|x\|_1$, which is known to promote sparsity. Define,

$$c_j(x) := \sum_{i=1}^{m} b_i \frac{a_{ij}}{\langle a_i, x \rangle}, \; r_j := \sum_i a_{ij} > 0.$$

Then, NoLips yields the following explicit iteration to solve $(\mathbb{E})$ with $\lambda = \|b\|_1^{-1}$:

$$x_j^+ = \frac{x_j}{1 + \lambda \left(\mu x_j + x_j (r_j - c_j(x)) \right)}, \; j = 1, \ldots n$$

Example 1 – Sparse Poisson Linear Inverse Problem

Sparse regularization. Let $f(x) := \|x\|_1$, which is known to promote sparsity. Define,

$$c_j(x) := \sum_{i=1}^{m} b_i \frac{a_{ij}}{\langle a_i, x \rangle}, \; r_j := \sum_i a_{ij} > 0.$$

Then, NoLips yields the following explicit iteration to solve $(\mathbb{E})$ with $\lambda = \|b\|_1^{-1}$:

$$x_j^+ = \frac{x_j}{1 + \lambda \left(\mu x_j + x_j (r_j - c_j(x)) \right)}, \; j = 1, \dots n$$

Special Case: A New Scheme for the Poisson MLE problem

For $\mu = 0$ problem $(\mathbb{E})$ is the Poisson Maximum Likelihood Estimation Problem. In that particular case the iterates of NoLips simply become

$$x_j^+ = \frac{x_j}{1 + \lambda x_j (r_j - c_j(x))}, \; j = 1, \dots n.$$

In contrast to the standard EM algorithm given by the iteration:

$$x_j^+ = \frac{x_j}{r_j} c_j(x), \; j = 1, \dots, n.$$

Example 2 - Thikhonov - Poisson Linear Inverse Problems

Tikhonov regularization. Let $f(x) := \frac{1}{2}\|x\|^2$. We recall that this term is used as a penalty in order to promote solutions of $Ax = b$ with *small Euclidean norms*.

Example 2 - Thikhonov - Poisson Linear Inverse Problems

Tikhonov regularization. Let $f(x) := \frac{1}{2}\|x\|^2$. We recall that this term is used as a penalty in order to promote solutions of $Ax = b$ with *small Euclidean norms*.

Using previous notation, NoLips yields a

" A log-Thikonov method" : Set $\lambda = \|b\|_1^{-1}$ and start with $x \in \mathbb{R}^n_{++}$

$$x_j^+ = \frac{\sqrt{\rho_j^2(x) + 4\mu\lambda x_j^2} - \rho_j(x)}{2\mu\lambda x_j},\ j = 1, \ldots, n.$$

where

$$\rho_j(x) := 1 + \lambda x_j\left(r_j - c_j(x)\right),\ j = 1, \ldots, n.$$

As just mentioned, many other interesting methods can be considered

- By choosing different kernels for ϕ, or
- By reversing the order of the arguments in the proximity measure (which is not symmetric!..hence defining different problems.)

References

- Lecture is based on [1].
- Results on Bregman-prox [5].
- On the Subgradient/Mirror Descent [4]
- Much more.. on NonEuclidean prox in [2,3].

1. Bauschke H., Bolte J., and Teboulle, M. A Descent Lemma beyond Lipshitz Gradient Continuity: First Order Methods Revisited and Applications. *Mathematics of Operations Research*, (2016), 1–19. Available Online.
2. A. Auslender and M. Teboulle. Interior gradient and proximal methods in convex and conic optimization. *SIAM J. Optimization*, **16**, (2006), 697–725.
3. A. Auslender, M. Teboulle Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, **104**, (2005), 39–68.
4. A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters,* 31 (2003), 167–175.
5. G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization* 3 (1993), 538–543.

First Order Optimization Methods
Lecture 8 - FOM beyond Convexity

Marc Teboulle

School of Mathematical Sciences
Tel Aviv University

PGMO Lecture Series

January 25-26, 2017 Ecole Polytechnique, Paris

Lecture 8 - FOM Beyond Convexity

Goal: Derive a simple self-contained convergence analysis framework for a broad class of nonconvex and nonsmooth minimization problems.

- A "Recipe" for proving global convergence to a critical point.
- A prototype of a simple/useful Algorithm: PALM.
- Many Applications: phase retrieval for diffractive imaging, dictionary learning,... .... Sparse nonnegative matrix factorization ... Regularized Structured Total Least Squares....

The Problem : An Abstract Formulation

Let $F : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, lsc and bounded from below function.

$$(P) \qquad \inf \left\{ F(z) : \; z \in \mathbb{R}^d \right\}.$$

Suppose $\mathcal{A}$ is a generic algorithm which generates a sequence $\left\{ z^k \right\}_{k \in \mathbb{N}}$ via:

$$z^0 \in \mathbb{R}^d, z^{k+1} \in \mathcal{A}(z^k), \; k = 0, 1, \ldots .$$

Goal: Prove that the whole sequence $\left\{ z^k \right\}_{k \in \mathbb{N}}$ converges to a critical point of F.

Quick Recall

- (Limiting) Subdifferential $\partial \Psi(x)$:

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

- $x \in \mathbb{R}^d$ is a critical point of F if $\partial F(x) \ni 0$.

A General Recipe in 3 Main Steps for Descent Methods

A sequence z^k is called *a descent sequence* for $F : \mathbb{R}^n \to (-\infty, +\infty]$ if

C1. Sufficient decrease property

$$\exists \rho_1 > 0 \quad \text{with} \quad \rho_1 \|z^{k+1} - z^k\|^2 \leq F(z^k) - F(z^{k+1}), \quad \forall k \geq 0$$

C2. Iterates gap For each k there exists $w^k \in \partial F(z^k)$ such that:

$$\exists \rho_2 > 0 \quad \text{with} \quad \left\| w^{k+1} \right\| \leq \rho_2 \|z^{k+1} - z^k\|, \forall k \geq 0.$$

A General Recipe in 3 Main Steps for Descent Methods

A sequence z^k is called *a descent sequence* for $F : \mathbb{R}^n \to (-\infty, +\infty]$ if

C1. Sufficient decrease property

$$\exists \rho_1 > 0 \quad \text{with} \quad \rho_1 \|z^{k+1} - z^k\|^2 \leq F(z^k) - F(z^{k+1}), \quad \forall k \geq 0$$

C2. Iterates gap For each k there exists $w^k \in \partial F(z^k)$ such that:

$$\exists \rho_2 > 0 \quad \text{with} \quad \left\|w^{k+1}\right\| \leq \rho_2 \|z^{k+1} - z^k\|, \forall k \geq 0.$$

- These two steps are typical for **any descent** type algorithms but lead **only to subsequential convergence**.

A General Recipe in 3 Main Steps for Descent Methods

A sequence z^k is called *a descent sequence* for $F : \mathbb{R}^n \to (-\infty, +\infty]$ if

C1. Sufficient decrease property

$$\exists \rho_1 > 0 \quad \text{with} \quad \rho_1 \|z^{k+1} - z^k\|^2 \leq F(z^k) - F(z^{k+1}), \quad \forall k \geq 0$$

C2. Iterates gap For each k there exists $w^k \in \partial F(z^k)$ such that:

$$\exists \rho_2 > 0 \quad \text{with} \quad \left\|w^{k+1}\right\| \leq \rho_2 \|z^{k+1} - z^k\|, \forall k \geq 0.$$

- These two steps are typical for **any descent** type algorithms but lead **only to subsequential convergence**.
- To get **global convergence** to a critical point, we need a deep mathematical tool.[Łojasiewicz (68), Kurdyka (98)]

The Third Main Step of our Recipe

C3. The Kurdyka-Łojasiewicz property: Assume that F is a KL function. Use this property to prove that the generated sequence $\{z^k\}_{k \in \mathbb{N}}$ is a *Cauchy sequence*, and thus converges!

The Third Main Step of our Recipe

C3. The Kurdyka-Łojasiewicz property: Assume that F is a KL function. Use this property to prove that the generated sequence $\{z^k\}_{k\in\mathbb{N}}$ is a *Cauchy sequence*, and thus converges!

This general recipe

- Singles out the 3 main ingredients at play to derive global convergence in the nonconvex and nonsmooth setting.
- **Applicable to any descent algorithm**.

Main Convergence Result

Theorem - Abstract Global Convergence

- Let F be a KL function – namely condition C3 holds.
- z^k is a descent sequence for F – namely conditions C1 and C2 hold.

If z^k is bounded, it converges to a critical point of F.

What is a KL function?

The KL Property – Informal

Let $\bar{z}$ be critical, with $F(\bar{z}) = 0$ (true up to translation); $\mathcal{L} := \{z \in \mathbb{R}^d : 0 < F(z) < \eta\}$

Definition [Sharpness] A function $F : \mathbb{R}^d \to (-\infty, +\infty]$ is called sharp on $\mathcal{L}$ if there exists $c > 0$ such that

$$\mathrm{dist}\,(0, \partial F(z)) := \min\{\|\xi\| : \ \xi \in \partial F(z)\} \geq c > 0 \quad \forall\, z \in \mathcal{L}.$$

KL expresses the fact that a function can be made "sharp" by re-parametrization of its values.

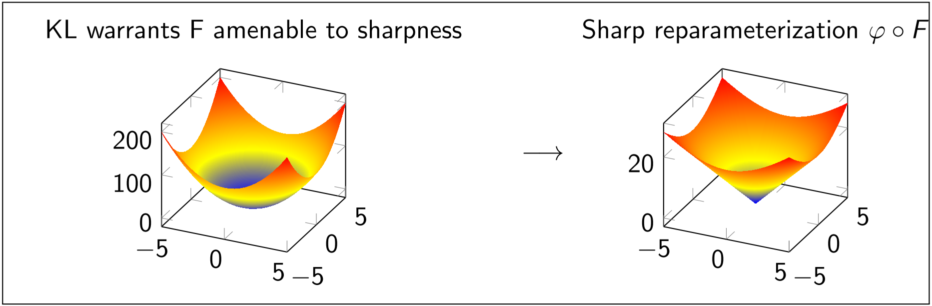
The KL Property – Informal

Let $\bar{z}$ be critical, with $F(\bar{z}) = 0$ (true up to translation); $\mathcal{L} := \{z \in \mathbb{R}^d : 0 < F(z) < \eta\}$

Definition [Sharpness] A function $F : \mathbb{R}^d \to (-\infty, +\infty]$ is called sharp on $\mathcal{L}$ if there exists $c > 0$ such that

$$\operatorname{dist}(0, \partial F(z)) := \min\{\|\xi\| : \xi \in \partial F(z)\} \geq c > 0 \quad \forall z \in \mathcal{L}.$$

KL expresses the fact that a function can be made "sharp" by re-parametrization of its values.



The KL Property: (Łojasiewicz (68), Kurdyka (98))

Desingularizing functions on $(0, \eta)$. **Let** $\eta > 0$.

$$\Phi_\eta := \{\varphi \in C[0,\eta) \cap C^1(0,\eta) : , \textbf{ concave with } \varphi' > 0, \varphi(0) = 0.\}$$

For $\bar{x} \in \mathbf{dom}\, \partial F,\ \mathcal{L} := \{x \in \mathbb{R}^d : F(\bar{x}) < F(x) < F(\bar{x}) + \eta\}$

The KL Property F has the KL property on $\mathcal{L}$ if there exists a desingularizing function φ such that

$$\varphi'\left(F(x) - F(\bar{x})\right) \operatorname{dist}\left(0, \partial F(x)\right) \geq 1, \quad \forall x \in \mathcal{L}$$

Local version: KL at $\bar{x} \in \operatorname{dom} F$, replace $\mathcal{L}$ with: its intersection with a closed ball $B(\bar{x}, \varepsilon)$ for some $\varepsilon > 0$.

Meaning: Subgradients of $x \to \varphi \circ (F(x) - F(\bar{x}))$ have a norm greater than 1, no matter how close is x to the critical point $\bar{x}$ (provided $F(x) > F(\bar{x})$) – **This is sharpness.**

Are there many functions satisfying KL? How we verify KL?

Are there Many Functions Satisfying KL?

Are there Many Functions Satisfying KL?

YES! Semi Algebraic Functions

Theorem

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lsc function. If σ is semi-algebraic then it satisfies the KL property at any point of $\operatorname{dom}\sigma$.

Recall: Semi-algebraic sets and functions

(i) A semialgebraic subset of $\mathbb{R}^d$ is a finite union of sets

$$\{x \in \mathbb{R}^d : \; p_i(x) = 0, \; q_j(x) < 0, \; i \in I, \; j \in J\}$$

where $p_i, q_j : \mathbb{R}^d \to \mathbb{R}$ are real polynomial (analytic) functions and I, J are finite.

(ii) A function σ is semi-algebraic if its graph

$$\left\{(u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t\right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

Operations on Semi-Algebraic Objects

Semi-Algebraic Property is Preserved under Many Operations

- If S is semi-algebraic, so is the closure $\overline{S}$.
- Unions/intersections of semi-algebraic sets are semi-algebraic.
- Indicator of a semi-algebraic set is semi-algebraic.
- Finite sums and product of semi-algebraic functions
- Composition of semi-algebraic functions;
- Sup/Inf type function, *e.g.*, $\sup\left\{g\left(u,v\right):\ v \in C\right\}$ is semi-algebraic when g is a semi-algebraic function and C a semi-algebraic set.

There is a Wealth of Semi-Algebraic Functions!

Semi-Algebraic Sets/Functions "Starring" in Optimization/Applications

- Real polynomial functions: $\|Ax - b\|^2, (A, B) \rightarrow \|AB - M\|_F^2$
- Any Polyhedral set is semi-algebraic
- In matrix theory: cone of PSD matrices, constant rank matrices, Stiefel manifolds...
- The function $x \rightarrow \operatorname{dist}(x, S)^2$ is semi-algebraic whenever S is a nonempty semi-algebraic subset of $\mathbb{R}^n$.
- The l_1-norm $\|x\|_1$ is semi-algebraic, as sum of absolute values function. For example, to show that $\sigma(u) := |u|$ is semi-algebraic note that $\mathrm{Graph}(\sigma) = \overline{\mathrm{S}}$, where

$$S = \{(u, s) : \ u + s = 0, -u > 0\} \cup \{(u, s) : \ u - s = 0, u > 0\}.$$

- $\|\cdot\|_0$ is semi-algebraic. Its graph can be shown to be a finite union of product sets.

A Broad Class of Nonsmooth Nonconvex Problems

A Useful Block Optimization Model

$$(B) \qquad \text{minimize}_{x,y} \Psi(x,y) := f(x) + g(y) + H(x,y)$$

- $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^m \to (-\infty, +\infty]$ proper and lsc.
- $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function with gradient Lipschitz continuous on bounded subsets of $\mathbb{R}^n \times \mathbb{R}^m$ (e.g., true when $H \in C^2$).
- Partial gradients of H are $C^{1,1}$: $H(\cdot, y) \in C^{1,1}_{L(y)}$ and $H(x, \cdot) \in C^{1,1}_{L(x)}$.

♠ **NO convexity** assumed in the objective and the constraints (built-in through f and g extended valued).

Two blocks is only for the sake of simplicity. Same for the p-blocks case:

$$\text{minimize}_{x_1,\ldots,x_p} H(x_1, x_2, \ldots, x_p) + \sum_{i=1}^{p} f_i(x_i), \; x_i \in \mathbb{R}^{n_i}, n = \sum_{i=1}^{p} n_i$$

This optimization model covers many applications: signal/image processing, blind deconvolution, dictionary learning, matrix factorization, etc....Vast Literature...

PALM: Proximal Alternating Linearized Minimization

Cocktail Time! PALM "blends" old spices:

$\oplus$ **Space decomposition [á la Gauss-Seidel]**

$\oplus$ **Composite decomposition [á la Prox-Gradient].**

PALM: Proximal Alternating Linearized Minimization

Cocktail Time! PALM "blends" old spices:

$\oplus$ **Space decomposition [á la Gauss-Seidel]**

$\oplus$ **Composite decomposition [á la Prox-Gradient].**

PALM Algorithm

1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 L_1\left(y^k\right)$ and compute

$$x^{k+1} \in \operatorname{prox}^{f}_{c_k}\left(x^k - \frac{1}{c_k}\nabla_x H\left(x^k, y^k\right)\right).$$

2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 L_2\left(x^{k+1}\right)$ and compute

$$y^{k+1} \in \operatorname{prox}^{g}_{d_k}\left(y^k - \frac{1}{d_k}\nabla_y H\left(x^{k+1}, y^k\right)\right).$$

Stepsizes c_k^{-1}, d_k^{-1} are in $\left]0, 1/L_2(y^k)\right[\quad \& \quad \left]0, 1/L_1(x^{k+1})\right[$.

Main computational step: Computing the prox of a nonconvex function.

Convergence of PALM

Theorem [Global convergence to critical point]. Assume f, g, H semi-algebraic. Any bounded PALM sequence $\{z^k\}_{k\in\mathbb{N}}$ converges to a critical point $z^* = (x^*, y^*)$ of Ψ.

Note: The boundedness assumption on the generated sequence $\{z^k\}_{k\in\mathbb{N}}$ holds in several scenarios, e.g., when f, g have bounded level sets, or follows from the structure of the problem at hand.

- **I will outline the 3 key building blocks for the analysis and proof of Theorem.**
- **But, first it is instructive to see how KL works for simple smooth descent methods.**

Smooth case $f \in C_L^{1,1}$ - KL and Descent Methods.

Illustrating the Recipe for Sequences with Smooth Gradient.

- **C1. Sufficient desc.:** $\exists a > 0, f(x^{k+1}) \leq f(x^k) - a\|x^{k+1} - x^k\|^2$ (proved)
- **Assume Iterates:** $\exists b > 0 : b\|\nabla f(x^k)\| \leq \|x^{k+1} - x^k\|$.
  (f L-smooth, $\Rightarrow$ **C2 holds:** $\exists \rho > 0 : \|\nabla f(x^{k+1})\| \leq \rho\|x^{k+1} - x^k\|$, $(\rho = b^{-1} + L)$.)
- **C3. Assume KL:** $\varphi'(f(x) - f_*)\|\nabla f(x)\| \geq 1$, φ **concave**, $\varphi' > 0$

For convenience let $v^k := f(x^k) - f_*$. Using the above we then get:

$$\varphi(v^{k+1}) - \varphi(v^k) \quad \leq \quad \varphi'(v^k)(v^{k+1} - v^k), \quad (\varphi \text{ concave})$$

Smooth case $f \in C_L^{1,1}$ - KL and Descent Methods.

Illustrating the Recipe for Sequences with Smooth Gradient.

- **C1. Sufficient desc.:** $\exists a > 0, f(x^{k+1}) \le f(x^k) - a\|x^{k+1} - x^k\|^2$ (proved)
- **Assume Iterates:** $\exists b > 0 : b\|\nabla f(x^k)\| \le \|x^{k+1} - x^k\|$.
  (f L-smooth, $\Rightarrow$ **C2 holds:** $\exists \rho > 0 : \|\nabla f(x^{k+1})\| \le \rho\|x^{k+1} - x^k\|$, $(\rho = b^{-1} + L)$.)
- **C3. Assume KL:** $\varphi'(f(x) - f_*)\|\nabla f(x)\| \ge 1$, φ **concave**, $\varphi' > 0$

For convenience let $v^k := f(x^k) - f_*$. Using the above we then get:

$$\begin{aligned}
\varphi(v^{k+1}) - \varphi(v^k) &\le \varphi'(v^k)(v^{k+1} - v^k), \ \ (\varphi \text{ concave}) \\
v^{k+1} - v^k &\le -a\|x^{k+1} - x^k\|^2 \le -ab\|x^{k+1} - x^k\| \cdot \|\nabla f(x^k)\| \\
\varphi'(v^k)(v^{k+1} - v^k) &\le -ab\|x^{k+1} - x^k\|\varphi'(v^k)\|\nabla f(x^k)\| \ \ (\varphi' > 0) \\
&\le -ab\|x^{k+1} - x^k\|, \text{ (by KL), and hence} \\
\varphi(v^{k+1}) - \varphi(v^k) &\le -ab\|x^{k+1} - x^k\|.
\end{aligned}$$

Smooth case $f \in C^{1,1}_L$ - KL and Descent Methods.

Illustrating the Recipe for Sequences with Smooth Gradient.

- **C1. Sufficient desc.:** $\exists a > 0, f(x^{k+1}) \leq f(x^k) - a\|x^{k+1} - x^k\|^2$ (proved)
- **Assume Iterates:** $\exists b > 0 : b\|\nabla f(x^k)\| \leq \|x^{k+1} - x^k\|$.
  (f L-smooth, $\Rightarrow$ **C2 holds:** $\exists \rho > 0 : \|\nabla f(x^{k+1})\| \leq \rho\|x^{k+1} - x^k\|$, $(\rho = b^{-1} + L)$.)
- **C3. Assume KL:** $\varphi'(f(x) - f_*)\|\nabla f(x)\| \geq 1$, φ **concave**, $\varphi' > 0$

For convenience let $v^k := f(x^k) - f_*$. Using the above we then get:

$$\begin{aligned}
\varphi(v^{k+1}) - \varphi(v^k) &\leq \varphi'(v^k)(v^{k+1} - v^k), \ \ (\varphi \text{ concave}) \\
v^{k+1} - v^k &\leq -a\|x^{k+1} - x^k\|^2 \leq -ab\|x^{k+1} - x^k\| \cdot \|\nabla f(x^k)\| \\
\varphi'(v^k)(v^{k+1} - v^k) &\leq -ab\|x^{k+1} - x^k\|\varphi'(v^k)\|\nabla f(x^k)\| \ \ (\varphi' > 0) \\
&\leq -ab\|x^{k+1} - x^k\|, \text{ (by KL)}, \ \text{ and hence} \\
\varphi(v^{k+1}) - \varphi(v^k) &\leq -ab\|x^{k+1} - x^k\|.
\end{aligned}$$

- Therefore, $\|x^{k+1} - x^k\| \leq (ab)^{-1}\left(\varphi(v^k) - \varphi(v^{k+1})\right)$, and by telescoping
- we get finite length $\sum_k \|x^{k+1} - x^k\|$, and x^k Cauchy and converges.

Proximal Map for Nonconvex Functions

Let $\sigma : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lsc function. Given $x \in \mathbb{R}^n$ and $t > 0$, the proximal map defined by:

$$\mathrm{prox}_t^\sigma(x) := \operatorname{argmin}\left\{\sigma(u) + \frac{t}{2}\|u - x\|^2 : \ u \in \mathbb{R}^n\right\}.$$

Proposition [Well-definedness of proximal maps] If $\inf_{\mathbb{R}^n} \sigma > -\infty$, then, for every $t \in (0, \infty)$, the set $\mathrm{prox}_{1/t}^\sigma(x)$ is nonempty and compact.

Here prox_t^σ is a set-valued map. When $\sigma := \delta_X$, for a nonempty and closed set X, the proximal map reduces to the set-valued projection operator onto X.

Thanks to the prox properties, since PALM is defined by two proximal computations, all we need to assume is:

$$\inf_{\mathbb{R}^n \times \mathbb{R}^m} \Psi > -\infty, \quad \inf_{\mathbb{R}^n} f > -\infty \quad \text{and} \quad \inf_{\mathbb{R}^m} g > -\infty.$$

Thus, Problem (M) is inf-bounded and **PALM is well defined**.

1. A Key Nonconvex Proximal-Gradient Inequality

It extends to the nonconvex case the convex prox-gradient inequality.

Lemma [Sufficient decrease property]

(i) $h : \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$ with L_h-Lipschitz gradient.

(ii) $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper and lsc function with $\inf_{\mathbb{R}^d} \sigma > -\infty$.

Then, for any $u \in \operatorname{dom} \sigma$ and any $u^+ \in \mathbb{R}^d$ defined by

$$u^+ \in \operatorname{prox}_t^\sigma \left(u - \frac{1}{t} \nabla h(u) \right), \quad t > L_h,$$

we have

$$h(u^+) + \sigma(u^+) \leq h(u) + \sigma(u) - \frac{1}{2}(t - L_h) \left\| u^+ - u \right\|^2 .$$

Proof. Follows along the same line of analysis as in the convex case. □

2. PALM Properties: Standard Subsequences Convergence

From now on we assume that the sequence $\{z^k\}_{k\in\mathbb{N}} := \{(x^k, y^k)\}$ generated by PALM is bounded.

$\omega(z^0)$ denotes the set of all limit points.

Lemma. [Properties of the limit point set $\omega(z^0)$] Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence generated by PALM. Then

(i) $\emptyset \neq \omega(z^0) \subset \mathrm{crit}\,\Psi$.

(ii) $\lim_{k\to\infty} \mathrm{dist}\left(z^k, \omega(z^0)\right) = 0$.

(iii) $\omega(z^0)$ is a nonempty, compact and connected set.

(iv) The objective function Ψ is finite and constant on $\omega(z^0)$.

Proof. Deduced by showing that **C1, C2** hold for the sequence $\{z^k\}_{k\in\mathbb{N}}$ + standard analysis arguments, see paper [4]. □

3. A Uniformization of KL

Lemma [Uniformized KL property]

- Let $\sigma : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lower semicontinuous function.
- Let Ω be a compact set.
- Assume σ is constant on Ω and satisfies the KL property at each point of Ω.

Then, there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in \Phi_\eta$ such that for all $\overline{u}$ in Ω and all u in the following intersection

$$\mathbb{W} := \left\{u \in \mathbb{R}^d : \ \mathrm{dist}\,(u, \Omega) < \varepsilon\right\} \cap [\sigma\,(\overline{u}) < \sigma\,(u) < \sigma\,(\overline{u}) + \eta] \tag{1}$$

one has,

$$\varphi'\,(\sigma\,(u) - \sigma\,(\overline{u}))\,\mathrm{dist}\,(0, \partial\sigma\,(u)) \geq 1. \tag{2}$$

Proof. See reference [4]. □

Recall: Let $\eta \in (0, +\infty]$. Φ_η is the class of all concave C^1 functions s.t.: $\varphi\,(0) = 0$ and $\varphi'\,(s) > 0$ for all $s \in (0, \eta)$.

Sketch of Proof for Global Convergence of PALM

Using the three described results, on can proceed as follows.

- Use sufficient decrease property and $\lim_{k\to\infty} \mathrm{dist}\left(z^k, \omega\left(z^0\right)\right) = 0$ to verify that there exists l such that $z^k \in \mathbb{W}$ for all $k > l$.
- Use the established facts: $\emptyset \neq \omega\left(z^0\right)$ and compact + Ψ finite and constant on $\omega\left(z^0\right)$, so that UKL Lemma can be applied with $\Omega \equiv \omega\left(z^0\right)$.

Sketch of Proof for Global Convergence of PALM

Using the three described results, on can proceed as follows.

- Use sufficient decrease property and $\lim_{k\to\infty} \operatorname{dist}\left(z^k, \omega\left(z^0\right)\right) = 0$ to verify that there exists l such that $z^k \in \mathbb{W}$ for all $k > l$.
- Use the established facts: $\emptyset \neq \omega\left(z^0\right)$ and compact + Ψ finite and constant on $\omega\left(z^0\right)$, so that UKL Lemma can be applied with $\Omega \equiv \omega\left(z^0\right)$.
- Use property of φ (concave inequality) and KL inequality 2 of the Lemma to show that $\left\{z^k\right\}_{k\in\mathbb{N}}$ has finite length, that is

$$\sum_{k=1}^{\infty} \left\|z^{k+1} - z^k\right\| < \infty.$$

Sketch of Proof for Global Convergence of PALM

Using the three described results, on can proceed as follows.

- Use sufficient decrease property and $\lim_{k\to\infty} \operatorname{dist}\left(z^k, \omega\left(z^0\right)\right) = 0$ to verify that there exists l such that $z^k \in \mathbb{W}$ for all $k > l$.
- Use the established facts: $\emptyset \neq \omega\left(z^0\right)$ and compact + Ψ finite and constant on $\omega\left(z^0\right)$, so that UKL Lemma can be applied with $\Omega \equiv \omega\left(z^0\right)$.
- Use property of φ (concave inequality) and KL inequality 2 of the Lemma to show that $\left\{z^k\right\}_{k\in\mathbb{N}}$ has finite length, that is

$$\sum_{k=1}^{\infty} \left\|z^{k+1} - z^k\right\| < \infty.$$

- Then, it follows that $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a Cauchy sequence and hence is a convergent sequence.
- The result follows immediately from the previous fact $\emptyset \neq \omega\left(z^0\right) \subset \operatorname{crit}\Psi$. □

Rate of Convergence Results

Theorem - Rate of Convergence for the sequence $\{z^k\}$ - Generic
Let F be a function which satisfies the KL property with

$$\varphi(s) = cs^{1-\theta}, \quad , c > 0, \theta \in [0,1),$$

and z^k a descent sequence for F. Then,

(i) If $\theta = 0$ then the sequence z^k converges in a finite number of steps.

(ii) If $\theta \in (0, 1/2]$ $\exists b > 0$ and $\tau \in [0,1)$ such that $\left\|z^k - \bar{z}\right\| \leq b\,\tau^k$.

(iii) If $\theta \in (1/2, 1)$ $\exists b > 0$ such that

$$\left\|z^k - \bar{z}\right\| \leq b\,k^{-\frac{1-\theta}{2\theta-1}}.$$

Finding θ can be difficult....

Applications: Nonnegative Matrix Factorization Problems

The NMF Problem: Given $A \in \mathbb{R}^{m \times n}$ and $r \ll \min\{m, n\}$.
Find $X \in \mathbb{R}^{m \times r}$ and $Y \in \mathbb{R}^{r \times n}$ such that

$$A \approx XY, \;\; X \in \mathcal{K}_{m,r} \cap \mathcal{F}, \;\; Y \in \mathcal{K}_{r,n} \cap \mathcal{G},$$

$$\begin{aligned} \mathcal{K}_{p,q} &= \left\{ M \in \mathbb{R}^{p \times q} : \; M \geq 0 \right\} \\ \mathcal{F} &= \left\{ X \in \mathbb{R}^{m \times r} : \; R_1(X) \leq \alpha \right\} \\ \mathcal{G} &= \left\{ Y \in \mathbb{R}^{r \times n} : \; R_2(Y) \leq \beta \right\}. \end{aligned}$$

$R_1(\cdot)$ and $R_2(\cdot)$ are functions used to describe some additional/required features of X, Y.

(NMF) covers a very large number of problems in applications: Text Mining (data clusters in documents); Audio-Denoising (speech dictionnary); Bio-informatics (clustering gene expression); Medical Imaging,...Vast Literature.

The Optimization Approach

We adopt the Constrained Nonconvex Nonsmooth Formulation

$$(MF) \qquad \min\left\{\frac{1}{2}\|A - XY\|_F^2 : \; X \in \mathcal{K}_{m,r} \cap \mathcal{F}, Y \in \mathcal{K}_{r,n} \cap \mathcal{G}\right\},$$

This formulation fits our general nonsmooth nonconvex model (M) with obvious identifications for H, f, g.

We now illustrate with semi-algebraic data on two important cases.

Example: Applying PALM on NMF Problems

I. Nonnegative Matrix Factorization (NMF): $\mathcal{F} \equiv \mathbb{R}^{m \times r}$; $\mathcal{G} \equiv \mathbb{R}^{r \times n}$.

$$\min \left\{ \frac{1}{2} \left\| A - XY \right\|_F^2 : X \geq 0, \, Y \geq 0 \right\}.$$

Example: Applying PALM on NMF Problems

I. Nonnegative Matrix Factorization (NMF): $\mathcal{F} \equiv \mathbb{R}^{m \times r}$; $\mathcal{G} \equiv \mathbb{R}^{r \times n}$.

$$\min \left\{ \frac{1}{2} \|A - XY\|_F^2 : X \geq 0, Y \geq 0 \right\}.$$

II. Sparsity Constrained (SNMF): Useful in many applications

$$\min \left\{ \frac{1}{2} \|A - XY\|_F^2 : \|X\|_0 \leq \alpha, \|Y\|_0 \leq \beta,\ X \geq 0, Y \geq 0 \right\}.$$

Sparsity measure of matrix: $\|X\|_0 := \sum_i \|x_i\|_0$, ($x_i$ column vector of X).

Example: Applying PALM on NMF Problems

I. Nonnegative Matrix Factorization (NMF): $\mathcal{F} \equiv \mathbb{R}^{m \times r}$; $\mathcal{G} \equiv \mathbb{R}^{r \times n}$.

$$\min \left\{ \frac{1}{2} \|A - XY\|_F^2 : X \geq 0, Y \geq 0 \right\}.$$

II. Sparsity Constrained (SNMF): Useful in many applications

$$\min \left\{ \frac{1}{2} \|A - XY\|_F^2 : \|X\|_0 \leq \alpha, \|Y\|_0 \leq \beta,\ X \geq 0, Y \geq 0 \right\}.$$

Sparsity measure of matrix: $\|X\|_0 := \sum_i \|x_i\|_0$, ($x_i$ column vector of X).

For Both models the data is semi-algebraic, and fit our block model (M):

- For NMF f, g are indicator of the form $\delta_{U \geq 0}$. Trivial projection on nonnegative cone.
- For SNMF: f and $g \equiv \delta_{U \geq 0} + \delta_{\|U\|_0 \leq s}$. Also admit explict prox formula.
- **PALM** produces very simple practical schemes, proven to globally converge.

References - Lecture based on [4]

1. Attouch, H. and Bolte, J., On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Mathematical Programming* **116** (2009), 5–16.
2. Attouch, H., Bolte, J. and Svaiter, B. F., Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Mathematical Programming*, Ser. A **137** (2013), 91–129.
3. Bolte, J., Daniilidis, A. and Lewis, A., The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM Journal on Optimization* **17** (2006), 1205–1223.
4. J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming, Series A*, **146**, 459–494, (2014).
5. Kurdyka, K., On gradients of functions definable in o-minimal structures, *Annales de l'institut Fourier* **48** (1998), 769–783.
6. Łojasiewicz, S., Une propriété topologique des sous-ensembles analytiques réels, Les Équations aux Dérivées Partielles. Éditions du centre National de la Recherche Scientifique, Paris, 87-89, (1963).