A New Look at First Order Methods Lifting the Lipschitz Gradient Continuity Restriction

Marc Teboulle

School of Mathematical Sciences
Tel Aviv University

Joint work with H. Bauschke and J. Bolte

Optimization and Discrete Geometry: Theory and Practice

24-26 April, 2018, Tel Aviv University

Recall: The Basic Pillar underlying FOM

$$\inf\{\Phi(x) := f(x) + g(x) : \ x \in \mathbb{R}^d\}, f, g \text{ convex}, \text{ with } g \in C^1.$$

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

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

Recall: The Basic Pillar underlying FOM

$$\inf\{\Phi(x) := f(x) + g(x) : \ x \in \mathbb{R}^d\}, f, g \text{ convex}, \text{ with } g \in C^1.$$

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

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

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

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

This inequality naturally provides

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

However, in many contexts and applications:

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

FOM Beyond Lipschitz Gradient Continuity

Goals/Outline:

- **Circumvent the longstanding question of Lipschitz Gradient continuity imposed on FOM.**
- **Derive FOM “free” from this smoothness assumption, with guaranteed complexity estimates and convergence results.**
- **Apply our results to a broad class of important problems lacking smooth gradients.**

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 $\mathbb{R}^d$:

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

Simple algebra shows that it can be equivalently written as:

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

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 function g on $\mathbb{R}^d$

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

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 $\frac{L}{2}\|x\|^2 - g(x)$!

Thus, for a given smooth function g on $\mathbb{R}^d$

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

Capture the Geometry of Constraint/Objective Naturally suggests to replace the *squared norm* with a general convex function $h(\cdot)$ that captures the geometry of the constraint/objective.

A Lipschitz-Like Convexity Condition

Following our basic observation: Replace the $\|\cdot\|^2$ with a convex h.

- Trade L-smooth gradient of g on $\mathbb{R}^d$ with
- Convexity condition on couple (g, h), $\operatorname{dom} g \supset \operatorname{dom} h, g \in C^1(\operatorname{int} \operatorname{dom} h)$.

A Lipschitz-like/Convexity Condition

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

- Condition (LC) $\Longleftrightarrow$ **New descent Lemma** we seek for.
- It also naturally leads to the well-known **Bregman distance**.

A Descent Lemma without Lipschitz Gradient Continuity

Lemma (Descent lemma without Lipschitz Gradient Continuity)

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

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

D_h stands for the Bregman Distance associated to a convex h:

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

Proof of Descent Lemma. $D_{Lh-g}(x, y) \geq 0$ for the convex function $Lh - g$! □

A Descent Lemma without Lipschitz Gradient Continuity

Lemma (Descent lemma without Lipschitz Gradient Continuity)

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

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

D_h stands for the Bregman Distance associated to a convex h:

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

Proof of Descent Lemma. $D_{Lh-g}(x, y) \geq 0$ for the convex function $Lh - g$! □

Distance-Like Properties - For all $(x, y) \in$ **dom** $h \times \operatorname{int}$ **dom** h

- $x \to D_h(x, y)$ is convex with h convex.
- $D_h(x, y) \geq 0$ and "$= 0$" **iff** $x = y$.(h strictly convex).
- **However, note that D_h is in general not symmetric!**

The use of Bregman distances in optimization started with Bregman (67). For initial works and main results on *Proximal Bregman Algorithms:* **[Censor-Zenios (92), T. (92), Chen-T. (93), Eckstein (93), Bauschke-Borwein (97).]**

Some Useful Examples for Bregman Distances D_h

Each example is a one dimensional convex h. The corresponding 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/useful kernels h include:** Nonseparable Bregman, e.g., any convex h on $\mathbb{R}^d$ as well as for handling **matrix problems:** PSD matrices, cone constraints, etc.., **[details in Auslender and T. (2005)]**.

The Convex Model and Blanket Assumption

Our aim is to solve the composite convex problem

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

where $\overline{\mathrm{dom}}\, h$ denotes the closure of dom h,

Under the following standard assumption.

The "Hidden h" (in unconstrained case) will adapt to Nonlinear Geometry of $\mathcal{P}$

Blanket Assumption as Usual:

(i) $f : \mathbb{R}^d \to (-\infty, \infty]$ is proper lower semicontinuous (lsc) convex,

(ii) $h : \mathbb{R}^d \to (-\infty, \infty]$ is proper, lsc convex.

(iii) $g : \mathbb{R}^d \to (-\infty, \infty]$ is proper lsc convex with $\mathrm{dom}\, g \supset \mathrm{dom}\, h$ and $g \in C^1(\mathrm{int}\,\mathrm{dom}\, h)$

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

(v) $-\infty < v(\mathcal{P}) = \inf\{\Phi(x) : x \in \overline{\mathrm{dom}}\, h\} = \inf\{\Phi(x) : x \in \mathrm{dom}\, h\}$.

Algorithm NoLips for $\inf\{f(x)+g(x) : x \in C \equiv \overline{\text{dom}}\, h\}$

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

$$\mathbf{T}_\lambda(\mathbf{x}) := \text{argmin}\left\{\mathbf{f}(\mathbf{u}) + \mathbf{g}(\mathbf{x}) + \langle \nabla \mathbf{g}(\mathbf{x}), \mathbf{u}-\mathbf{x}\rangle + \frac{\mathbf{1}}{\lambda}\mathbf{D_h}(\mathbf{u},\mathbf{x}) : \mathbf{u} \in \overline{\text{dom}}\,\mathbf{h}\right\} \ (\lambda > \mathbf{0}).$$

NoLips Main Iteration: $x \in \text{int}\,\text{dom}\ h, \qquad x^+ = T_\lambda(x),\ (\lambda > 0).$

Algorithm NoLips for $\inf\{f(x)+g(x) : x \in C \equiv \overline{\text{dom}}\, h\}$

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

$$\mathbf{T}_\lambda(\mathbf{x}) := \operatorname{argmin}\left\{\mathbf{f}(\mathbf{u}) + \mathbf{g}(\mathbf{x}) + \langle \nabla \mathbf{g}(\mathbf{x}), \mathbf{u}-\mathbf{x}\rangle + \frac{1}{\lambda}\mathbf{D}_\mathbf{h}(\mathbf{u},\mathbf{x}) : \mathbf{u} \in \overline{\text{dom}}\,\mathbf{h}\right\} \ (\lambda > \mathbf{0}).$$

NoLips Main Iteration: $x \in \text{int dom } h, \qquad x^+ = T_\lambda(x), \ (\lambda > 0).$

Algorithm NoLips – in More Details

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

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

Main Issues / Questions for NoLips

- **Well posedness and Computation of $T_\lambda(\cdot)$?**
- **What is the complexity of NoLips?**
- **Does NoLips converge to an optimal solution?**
- **In particular: Can we identify the <u>most aggressive</u> step-size in terms of problem's data?**

NoLips is Well Defined

We assume h is a **Legendre** function **[Rockafellar 70]**.

- h is strictly convex and differentiable on $\operatorname{int}\operatorname{dom} h \neq \emptyset$ and

$$\operatorname{dom} \partial h = \operatorname{int}\operatorname{dom} h \text{ with } \partial h(x) = \{\nabla h(x)\}, \forall x \in \operatorname{int}\operatorname{dom} h.$$

- $\|\nabla h(x^k)\| \to \infty$ whenever $\{x^k\} \subset \operatorname{int}\operatorname{dom} h$, $x^k \to x \in \operatorname{Bdy}(\operatorname{dom} h.)$

With h Legengre: ∇h is a *bijection* from $\operatorname{int}\operatorname{dom} h \to \operatorname{int}\operatorname{dom} h^*$ and $(\nabla h)^{-1} = \nabla h^*$.

NoLips is Well Defined

We assume h is a **Legendre** function **[Rockafellar 70]**.

- h is strictly convex and differentiable on $\operatorname{int}\operatorname{dom} h \neq \emptyset$ and

$$\operatorname{dom}\partial h = \operatorname{int}\operatorname{dom} h \text{ with } \partial h(x) = \{\nabla h(x)\}, \forall x \in \operatorname{int}\operatorname{dom} h.$$

- $\|\nabla h(x^k)\| \to \infty$ whenever $\{x^k\} \subset \operatorname{int}\operatorname{dom} h$, $x^k \to x \in \operatorname{Bdy}(\operatorname{dom} h.)$

With h Legengre: ∇h is a *bijection* from $\operatorname{int}\operatorname{dom} h \to \operatorname{int}\operatorname{dom} h^*$ and $(\nabla h)^{-1} = \nabla h^*$.

Note:

- Legendre functions "abound" for defining useful D_h. (All previous examples and more..).
- Crucial for deriving meaningful convergence results.

Equipped with the above, one can prove (see technical details in our paper.)

Lemma (Well posedness of the method)

The proximal gradient map $T_\lambda \neq \emptyset$, is single-valued and maps $\operatorname{int}\operatorname{dom}\ h$ *in* $\operatorname{int}\operatorname{dom}\ h$.

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 map**

$$p_\lambda(x) := \operatorname{argmin}\left\{\lambda\langle\nabla g(x), u\rangle + D_h(u,x) : u \in \mathbb{R}^d\right\} \equiv \nabla h^*(\nabla h(x) - \lambda\nabla g(x))$$

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

⊕ **Define the proximal Bregman map**

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

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 map**

$$p_\lambda(x) := \operatorname{argmin}\left\{\lambda\langle\nabla g(x), u\rangle + D_h(u,x) : u \in \mathbb{R}^d\right\} \equiv \nabla h^*(\nabla h(x) - \lambda\nabla g(x))$$

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

⊕ **Define the proximal Bregman map**

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

One can show **NoLips $\equiv$ Composition of these two Bregman maps:**

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

For Specific and Useful Examples, see the paper.

The Key Estimation Inequality for Analyzing NoLips

Lemma (Descent inequality for NoLips)

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

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

The Key Estimation Inequality for Analyzing NoLips

Lemma (Descent inequality for NoLips)

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

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

Proof simply combines the *NoLips Descent Lemma* with **known old results:**

[Lemma 3.1 and Lemma 3.2 – Chen and T. (1993)].

1. **(The three points identity)** For any $x, y \in \operatorname{int}(\operatorname{dom} h)$ and $u \in \operatorname{dom} h$:

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

2. **(Bregman Based Proximal Inequality)** Given $z \in \operatorname{int} \operatorname{dom} h$, define

$$u^+ := \operatorname{argmin}\{\varphi(u) + t^{-1} D_h(u, z) : \ u \in X\}; \ \varphi \text{ convex}, \quad t > 0.$$

Then, for any $u \in \operatorname{dom} h$,

$$t(\varphi(u^+) - \varphi(u)) \leq D_h(u, z) - D_h(u, u^+) - D_h(u^+, z).$$

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*

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

(ii) **(Complexity for h with closed domain)** *Assume in addition, that* $\mathrm{dom}\, h = \overline{\mathrm{dom}}\, h$ *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 $\diamondsuit$ The entropies of Boltzmann-Shannon, Fermi-Dirac and Hellinger are non trivial examples for which the assumption ($\overline{\mathrm{dom}}\, h = \mathrm{dom}\, h$) holds.

$\diamondsuit$ 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.

Does NoLips Globally Converge to a Minimizer?

Does NoLips Globally Converge to a Minimizer?

Yes! by introducing an interesting notion of symmetry for D_h.

Bregman distances are in general <u>not symmetric</u>, except when h is the energy.

Does NoLips Globally Converge to a Minimizer?

Yes! by introducing an interesting notion of symmetry for D_h.

Bregman distances are in general <u>not symmetric</u>, except when h is the energy.

Definition (Symmetry coefficient-Measures Lack of Symmetry)

Let $h : \mathbb{R}^d \to (-\infty, \infty]$ be a Legendre function. Its symmetry coefficient is defined by

$$\alpha(h) := \inf \left\{ \frac{D_h(x,y)}{D_h(y,x)} : (x,y) \in \operatorname{int} dom\, h \times \operatorname{int} dom\, h,\ x \neq y \right\}.$$

Does NoLips Globally Converge to a Minimizer?

Yes! by introducing an interesting notion of symmetry for D_h.

Bregman distances are in general not symmetric, except when h is the energy.

Definition (Symmetry coefficient-Measures Lack of Symmetry)

Let $h : \mathbb{R}^d \to (-\infty, \infty]$ be a Legendre function. Its symmetry coefficient is defined by

$$\alpha(h) := \inf \left\{ \frac{D_h(x, y)}{D_h(y, x)} : (x, y) \in \operatorname{int} dom\, h \times \operatorname{int} dom\, h, \ x \neq y \right\}.$$

Properties of the Symmetry Coefficient $\alpha(h)$:

- $\alpha(h) \in [0, 1]$. *The closer is $\alpha(h)$ to 1 the more symmetric D_h is.*
- $\alpha(h) = 1$ Perfect symmetry! when h is the energy.
- $\alpha(h) = 0$ Total lack of symmetry, e.g., $h(x) = x \log x$ and $h(x) = -\log x$.
- $\alpha(h) > 0$ Some symmetry.., e.g., $h(x) = x^4$, $\alpha(h) = 2 - \sqrt{3}$.

The symmetry coefficient allows to determine the best step size of NoLips for pointwise convergence of the generated sequence $\{x^k\}$.

The Step-Size Choice λ for Global Convergence of NoLips

Defining Step Size in Terms of Problem's Data

$$0 < \lambda \leq \frac{(1 + \alpha(h)) - \delta}{L} \quad \textbf{for some} \quad \delta \in (0, 1 + \alpha(h)),$$

- $[0, 1] \ni \alpha(h)$ **is the symmetry coefficient of** h**.**
- $L > 0$ **is the constant in condition (LC)** $Lh - g$ **convex.**

- When $h(\cdot) := \|\cdot\|^2/2$, then $\alpha(h) = 1$, L is usual Lipchitz constant for ∇g and above reduces to

$$0 < \lambda \leq \frac{2 - \delta}{L}$$

  recovers the classical step size allowed for pointwise convergence of the classical proximal gradient method **[Combettes-Wajs 05]**.
- With the above step-size choice, we can establish global convergence of the sequence $\{x^k\}$ generated by NoLips.

Pointwise Convergence for NoLips

Theorem (NoLips: Point convergence - With $\lambda \in (0, L^{-1}(1+\alpha(h)))$)

Assume that the solution set $\mathcal{S}^*$ of $(\mathcal{P})$ is nonempty. Then, the following holds.

(i) **(Subsequential convergence)** *If $\mathcal{S}^*$ is compact, any limit point of $\{x^k\}_{k\in\mathbb{N}}$ is a solution to $(\mathcal{P})$.*

(ii) **(Global convergence)** *Assume that $\overline{\mathrm{dom}}\, h = dom\, h$ and that $(\mathbf{H})$ is satisfied. Then the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to some solution x^* of $(\mathcal{P})$.*

Note Nontrivial examples: Boltzmann-Shannon, Fermi-Dirac and Hellinger entropies satisfy the set of assumptions in **H** and $\overline{\mathrm{dom}}\, h = \mathrm{dom}\, h$.

Additional assumption on D_h is to ensure separation properties of D_h at the boundary.

Assumption H:

(i) For every $x \in \mathrm{dom}\, h$ and $\beta \in \mathbb{R}$, the level set $\{y \in \mathrm{int\,dom}\, h : \; D_h(x, y) \leq \beta\}$ is bounded.

(ii) If $\{x^k\}_{k\in\mathbb{N}}$ converges to some x in $\mathrm{dom}\, h$ then $D_h(x, x^k) \to 0$.

(iii) Reciprocally, if x is in $\mathrm{dom}\, h$ and if $\{x^k\}_{k\in\mathbb{N}}$ is such that $D_h(x, x^k) \to 0$, then $x^k \to x$.

Applications - A Prototype: Linear Inverse 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.

Huge amount of literature: **astronomy, nuclear medicine (PET), electronic microscropy, statistical estimation (EM), image deconvolution, denoising speckle (multiplicative) noise, ect....**

Problem: Given a matrix $A \in \mathbb{R}_+^{m \times n}$ and $b \in \mathbb{R}_{++}^m$ the goal is to reconstruct the signal/image $x \in \mathbb{R}_+^n$ from the noisy measurements b such that $Ax \simeq b$.

Applications - A Prototype: Linear Inverse 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.

Huge amount of literature: **astronomy, nuclear medicine (PET), electronic microscropy, statistical estimation (EM), image deconvolution, denoising speckle (multiplicative) noise, ect....**

Problem: Given a matrix $A \in \mathbb{R}_+^{m \times n}$ and $b \in \mathbb{R}_{++}^m$ the goal is to reconstruct the signal/image $x \in \mathbb{R}_+^n$ from the noisy measurements b such that $Ax \simeq b$.

A natural proximity measure in $\mathbb{R}_+^n$ - (Kullback-Liebler Divergence):

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

which (up to some const.) is the negative Poisson log-likelihood function.

- The optimization problem: $(\mathbb{E})$ minimize $\{\mu f(x) + g(x) : \; x \in \mathbb{R}_+^n\}$
- $g(x) \equiv \mathcal{D}(d, Ax)$, f a regularizer – smooth or nonsmooth, $\mu > 0$
- $x \to \mathcal{D}(b, Ax)$ convex, **but does not admit a globally Lipschitz continuous gradient.**

NoLips in Action : New Simple Schemes for Many Problems

The optimization problem will be of the form:

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

where $g(x) := \mathcal{D}_\phi(b, Ax)$ for some convex ϕ, and $f(x)$ some convex regularizer.

NoLips in Action : New Simple Schemes for Many Problems

The optimization problem will be of the form:

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

where $g(x) := \mathcal{D}_\phi(b, Ax)$ for some convex ϕ, and $f(x)$ some convex regularizer.

Applying NoLips requires:

1. To pick an adequate h, so that $Lh - g$ convex; L in terms of problem's data.
2. In turns, this determines the step-size λ defined through $(L, \alpha(h))$.
3. Compute $p_\lambda(\cdot)$ and $\mathrm{prox}^h_{\lambda f}(\cdot))$ – Bregman - gradient and proximal steps.

Our convergence/complexity results hold and produce new simple algorithms:

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

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

Two Simple Algorithms for Poisson Linear Inverse Problems

Given $g(x) := D_\phi(b, Ax)$ ($\phi(u) = u \log u$), **to apply NoLips**:

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

Given $g(x) := D_\phi(b, Ax)$ ($\phi(u) = u \log u$), **to apply NoLips**:

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

Lemma. With (g, h) above, $Lh - g$ is convex on $\mathbb{R}^n_{++}$ for any $L \geq \|b\|_1 := \sum_{i=1}^m b_i$.

Two Simple Algorithms for Poisson Linear Inverse Problems

Given $g(x) := D_\phi(b, Ax)$ ($\phi(u) = u \log u$), **to apply NoLips**:

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

Lemma. With (g, h) above, $Lh - g$ is convex on $\mathbb{R}^n_{++}$ for any $L \geq \|b\|_1 := \sum_{i=1}^{m} b_i$.

Thus, we can take $\lambda = L^{-1} = \|b\|_1^{-1}$, and applying NoLips with $x \in \mathbb{R}^n_{++}$ reads:

$$x^+ = \text{argmin} \left\{ \mu f(u) + \langle \nabla g(x), u \rangle + \|b\|_1 \sum_{j=1}^{n} \left(\frac{u_j}{x_j} - \log \frac{u_j}{x_j} - 1 \right) : u > 0 \right\}.$$

The above yields closed form algorithms for Poisson reconstruction problems with two typical regularizers.

Example 1 – Sparse Poisson Linear Inverse Problem

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

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

NoLips for Sparse Poisson Linear Inverse Problems

$$x_j > 0, \;\; x_j^+ = \frac{\|b\|_1 x_j}{\|b\|_1 + (\mu x_j + x_j(r_j - c_j(x)))}, \quad j = 1, \ldots n$$

Example 1 – Sparse Poisson Linear Inverse Problem

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

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

NoLips for Sparse Poisson Linear Inverse Problems

$$x_j > 0, \;\; x_j^+ = \frac{\|b\|_1 x_j}{\|b\|_1 + (\mu x_j + x_j(r_j - c_j(x)))}, \quad j = 1, \ldots n$$

Special Case: $\mu = 0$, $(\mathbb{E})$ is the *Poisson Maximum Likelihood Estimation*.

NoLips yields in that case: A New Scheme for Poisson MLE

$$x_j > 0, \;\; x_j^+ = \frac{\|b\|_1 x_j}{\|b\|_1 + x_j(r_j - c_j(x))}, \; j = 1, \ldots n.$$

Example 2 - Thikhonov - Poisson Linear Inverse Problems

Tikhonov regularization. Let $f(x) := \mu\|x\|^2/2$. 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 $f(x) := \mu\|x\|^2/2$. 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 Poisson-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, see the paper.)

Conclusion and More Details/Results on NoLips

Proposed framework offers a new paragdim for FOM

- **Breaks the longstanding question asking for L-smooth gradient.**
- **Proven Complexity and Pointwise Convergence as Classical case.**
- **Allows to derive new FOM without Lipschitz gradient.**

Details and More Results: Bauschke H., Bolte J., and Teboulle M.

"A Descent Lemma beyond Lipshitz Gradient Continuity: First Order Methods Revisited and Applications". *Mathematics of Operations Research*, (2017), 330–348.

Available Online `http://dx.doi.org/10.1287/moor.2016.0817`

Conclusion and More Details/Results on NoLips

Proposed framework offers a new paragdim for FOM

- **Breaks the longstanding question asking for L-smooth gradient.**
- **Proven Complexity and Pointwise Convergence as Classical case.**
- **Allows to derive new FOM without Lipschitz gradient.**

Details and More Results: Bauschke H., Bolte J., and Teboulle M.

"A Descent Lemma beyond Lipshitz Gradient Continuity: First Order Methods Revisited and Applications". *Mathematics of Operations Research*, (2017), 330–348.

Available Online `http://dx.doi.org/10.1287/moor.2016.0817`

THANK YOU FOR LISTENING!

NoLips (Red) Versus a FAST NoLips (Blue)...