Local Linear Convergence of First-order Proximal Splitting Methods

Jingwei Liang

University of Cambridge

Joint work with: Clarice Poon (Bath)

Carola Schönlieb (Cambridge)

Jalal Fadili (CNRS, ENSICAEN)

Gabriel Peyré (CNRS, ENS-Paris)

Adrian Lewis (Cornell)

INRIA, February 2019

Outline

A brief overview of first-order methods

Geometry of non-smooth regularisation

Local convergence analysis

- Finite activity identification
- Local linear convergence

Numerical Experiments

Part I

First-order Proximal Splitting Algorithms

Example: data science

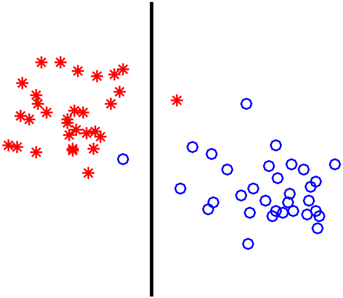
Sparse logistic regression [Friedman et al, 2001]

$(z_i, y_i) \in \mathbb{R}^n \times \{\pm 1\}, i = 1, ..., m,$

$$\min_{(b,x)\in\mathbb{R}\times\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^m f(\langle x, z_i\rangle + b, y_i),$$

where $f(u_i, y_i) = \log(1 + e^{-u_i y_i})$.

$$\|x\|_1 = \sum_{\ell=1}^n |x_\ell|.$$



Example: image processing

TV based Image deblur [Rudin et al, 1992]

$$w = Hx_{\text{ob}} + \omega,$$

where $H \in \mathbb{R}^{m \times n}$ is blur kernel, $\omega \in \mathbb{R}^m$ is additive noise.

$$\text{TV}(x) = \|\nabla x\|_1$$

x_{ob}

w

recovered x

Example: computer vision

Principal component pursuit [Candès et al, 2011]

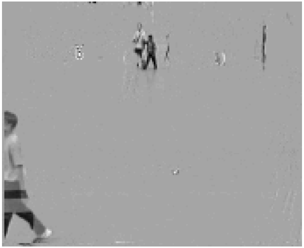
$$w = x_{\mathrm{ob},l} + x_{\mathrm{ob},s} + \omega,$$

$x_{\mathrm{ob},l} \in \mathbb{R}^{m\times n}$ is low-rank, $x_{\mathrm{ob},s} \in \mathbb{R}^{m\times n}$ is sparse and $\omega \in \mathbb{R}^{m\times n}$ is noise.

$$\|x\|_* = \sum_{\ell=1}^{\mathrm{rank}(x)} \sigma_\ell(x).$$

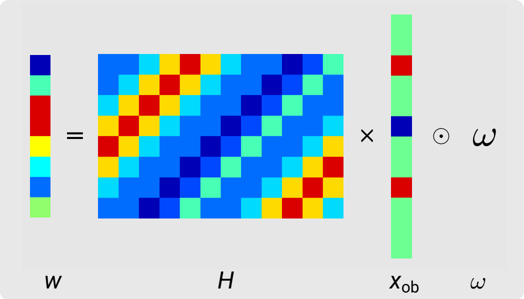
w

x_l



x_s

Example: inverse problems



Forward model:

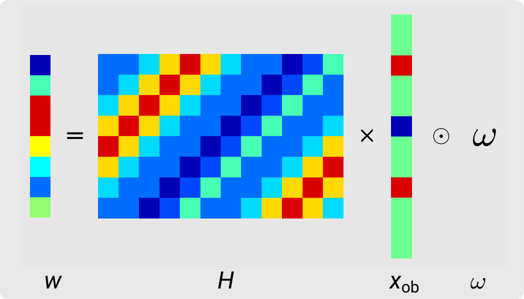
$$w = (Hx_{ob}) \odot \omega.$$

Goal : recover x_{ob}

Challenge : ill-posed

Hope : prior knowledge of x_{ob}

Example: inverse problems



Forward model:

$$w = (Hx_{\text{ob}}) \odot \omega.$$

Goal : recover x_{ob}

Challenge : ill-posed

Hope : prior knowledge of x_{ob}

- Regularisation: promoting low-complexity structure to the solution...
- Examples:

Sparsity	ℓ_1-norm, $\ell_{1,2}$-norm, ℓ_p-norm, ℓ_0 pseudo-norm
Analysis sparsity	total variation, wavelet, dictionary...
Low-rank	nuclear norm, rank function
Constraints	simplex, non-negativity...

Optimization problem

Non-smooth optimisation problem

$$\min_{x \in \mathbb{R}^n} \left\{ \Phi(x) \stackrel{\text{def}}{=} F(x) + \sum_{i=1}^{r} R_i(x) \right\}.$$

F: data fidelity term...

R_i: non-smooth regularisation terms...

Optimization problem

Non-smooth optimisation problem

$$\min_{x\in\mathbb{R}^n} \left\{\Phi(x) \stackrel{\text{def}}{=} F(x) + \sum_{i=1}^r R_i(x)\right\}.$$

F: data fidelity term...

R_i: non-smooth regularisation terms...

Image deblur

$$\min_{x\in\mathbb{R}^n} \mu\|\nabla x\|_1 + \frac{1}{2}\|Hx - w\|^2.$$

Sparse logistic regression

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^m f(\langle x,\, z_i\rangle + b, y_i).$$

Principal component pursuit

$$\min_{x_{\mathrm{l}}, x_{\mathrm{s}}\in\mathbb{R}^{m\times n}} \mu_1\|x_{\mathrm{s}}\|_1 + \mu_2\|x_{\mathrm{l}}\|_* + \frac{1}{2}\|w - x_{\mathrm{l}} - x_{\mathrm{s}}\|^2.$$

Optimization problem

Non-smooth optimisation problem

$$\min_{x\in\mathbb{R}^n} \left\{\Phi(x) \stackrel{\text{def}}{=} F(x) + \textstyle\sum_{i=1}^{r} R_i(x)\right\}.$$

F: data fidelity term...

R_i: non-smooth regularisation terms...

Image deblur

$$\min_{x\in\mathbb{R}^n} \mu\|\nabla x\|_1 + \frac{1}{2}\|Hx - w\|^2.$$

Sparse logistic regression

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\textstyle\sum_{i=1}^{m} f(\langle x, z_i\rangle + b, y_i).$$

Principal component pursuit

$$\min_{x_l, x_s\in\mathbb{R}^{m\times n}} \mu_1\|x_s\|_1 + \mu_2\|x_l\|_* + \frac{1}{2}\|w - x_l - x_s\|^2.$$

Non-smooth, (non-convex), composite, high dimension

First-order methods: two basic ingredients

Gradient descent

$$\min_{x \in \mathbb{R}^n} F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz

$$x_{k+1} = x_k - \gamma \nabla F(x_k), \ \ \gamma_k \in]0, 2/L[.$$

First-order methods: two basic ingredients

Gradient descent

$$\min_{x \in \mathbb{R}^n} F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz

$$x_{k+1} = x_k - \gamma \nabla F(x_k), \quad \gamma_k \in]0, 2/L[.$$

Proximal point algorithm [Rockafellar, 1976]

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

with R being proper convex and l.s.c.. Define "proximity operator" by

$$\mathrm{prox}_{\gamma R}(v) \stackrel{\mathrm{def}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n} \gamma R(x) + \frac{1}{2}\|x - v\|^2.$$

Proximal point algorithm

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_k), \quad \gamma_k > 0.$$

First-order methods: from structure to splitting

FoM [Bauschke and Combettes, 2011]...

$F + R$ Forward–Backward splitting (FB), inertial FB, Nesterov, FISTA

$F = \frac{1}{m}\sum_i f_i$: stochastic gradient methods

First-order methods: from structure to splitting

FoM [Bauschke and Combettes, 2011]...

$F + R$	Forward–Backward splitting (FB), inertial FB, Nesterov, FISTA
	$F = \frac{1}{m}\sum_i f_i$: stochastic gradient methods
$R_1 + R_2$	Douglas–Rachford splitting

First-order methods: from structure to splitting

FoM [Bauschke and Combettes, 2011]...

$F + R$	Forward–Backward splitting (FB), inertial FB, Nesterov, FISTA
	$F = \frac{1}{m}\sum_i f_i$: stochastic gradient methods
$R_1 + R_2$	Douglas–Rachford splitting
$F + R(\mathcal{W}\cdot)$	Class of Primal–Dual splitting
	Alternating Direction Method of Multipliers (ADMM)

First-order methods: from structure to splitting

FoM [Bauschke and Combettes, 2011]...

- $F + R$ Forward–Backward splitting (FB), inertial FB, Nesterov, FISTA
  - $F = \frac{1}{m}\sum_i f_i$: stochastic gradient methods
- $R_1 + R_2$ Douglas–Rachford splitting
- $F + R(\mathcal{W}\cdot)$ Class of Primal–Dual splitting
  - Alternating Direction Method of Multipliers (ADMM)
- $F + \sum_{i=1}^{r} R_i$ Three-operator splitting ($r = 2$)
  - Forward–Douglas–Rachford ($r = 2, R_2 = \iota_{\mathcal{V}}(\cdot)$)
  - Generalized Forward–Backward splitting ($r \geq 2$)
- ...

First-order methods: from structure to splitting

FoM [Bauschke and Combettes, 2011]...

$F + R$	Forward–Backward splitting (FB), inertial FB, Nesterov, FISTA
	$F = \frac{1}{m}\sum_i f_i$: stochastic gradient methods
$R_1 + R_2$	Douglas–Rachford splitting
$F + R(\mathcal{W}\cdot)$	Class of Primal–Dual splitting
	Alternating Direction Method of Multipliers (ADMM)
$F + \sum_{i=1}^{r} R_i$	Three-operator splitting ($r = 2$)
	Forward–Douglas–Rachford ($r = 2, R_2 = \iota_{\mathcal{V}}(\cdot)$)
	Generalized Forward–Backward splitting ($r \geq 2$)
–	...

Dates back to 1950s for numerical PDE, now ubiquitous in signal/image processing, inverse problems, data science, statistics and machine learning, game theory...

Part II

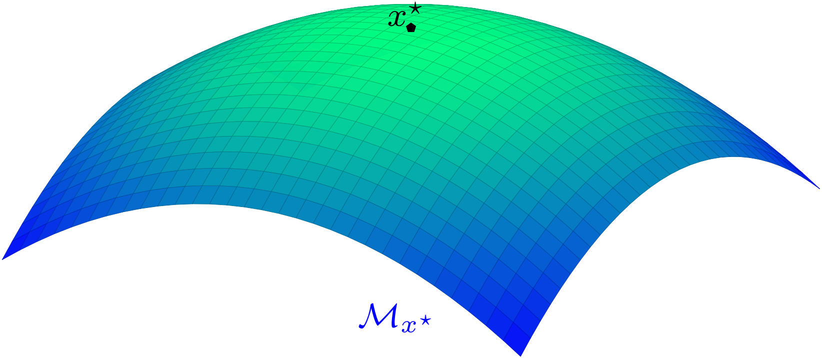
A Geometric Perspective of Non-smooth Regularisation

What happens when regularises...

Goal: find $x^\star$ which has low-complexity, *e.g.* $x^\star \in \mathcal{M}_{x^\star}$.

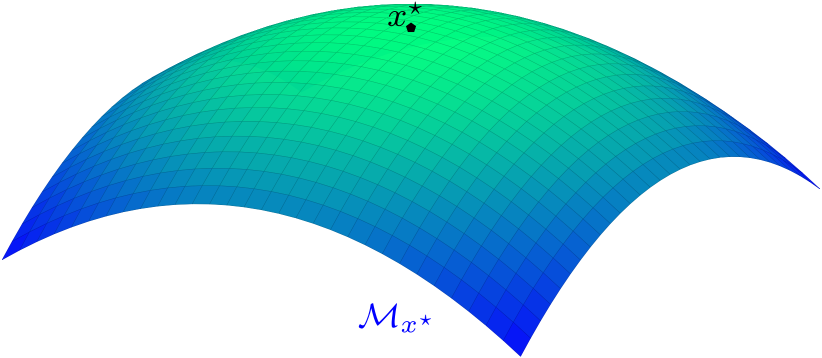
What happens when regularises...

Goal: find $x^\star$ which has low-complexity, *e.g.* $x^\star \in \mathcal{M}_{x^\star}$.



What happens when regularises...

Goal: find $x^\star$ which has low-complexity, *e.g.* $x^\star \in \mathcal{M}_{x^\star}$.



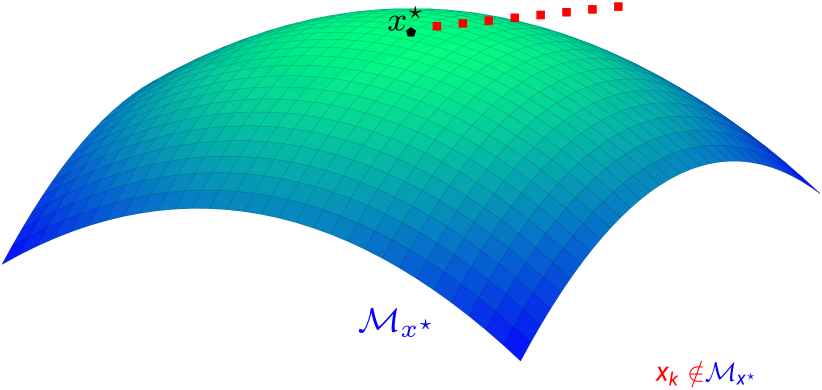
FoM: generates x_k that converges to $x^\star$.

How about

$$\boxed{x_k \,?\, \mathcal{M}_{x^\star}}$$

What happens when regularises...

Goal: find $x^\star$ which has low-complexity, *e.g.* $x^\star \in \mathcal{M}_{x^\star}$.



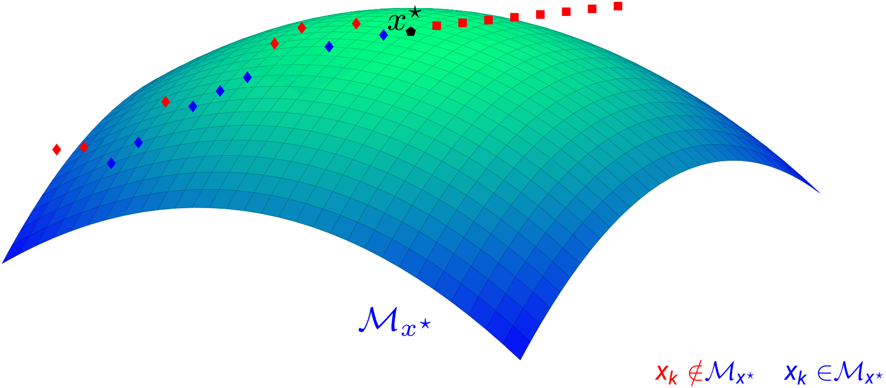
FoM: generates x_k that converges to $x^\star$.

How about

$$\boxed{x_k \,?\, \mathcal{M}_{x^\star}}$$

What happens when regularises...

Goal: find $x^\star$ which has low-complexity, *e.g.* $x^\star \in \mathcal{M}_{x^\star}$.



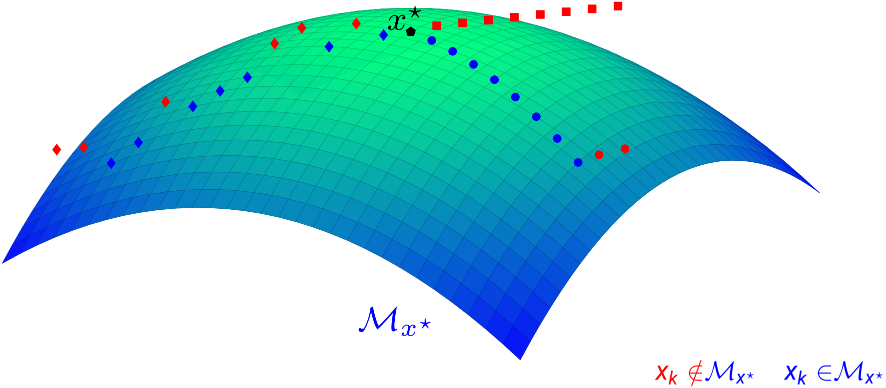
FoM: generates x_k that converges to $x^\star$.

How about

$$\boxed{x_k \,?\, \mathcal{M}_{x^\star}}$$

What happens when regularises...

Goal: find $x^\star$ which has low-complexity, *e.g.* $x^\star \in \mathcal{M}_{x^\star}$.



FoM: generates x_k that converges to $x^\star$.

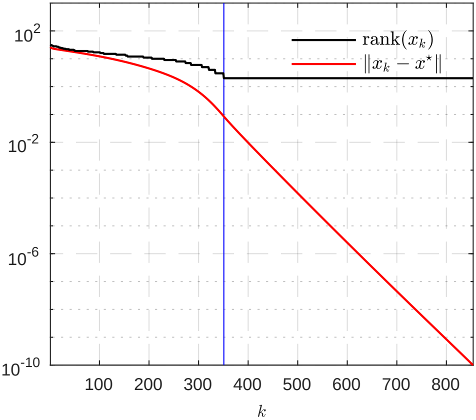
How about

$$\boxed{x_k \,?\, \mathcal{M}_{x^\star}}$$

Typical observation

Low-rank recovery

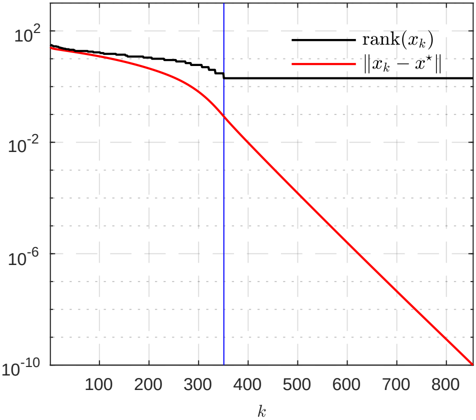
$$\min_{x \in \mathbb{R}^{n \times n}} \mu \|x\|_* + \frac{1}{2} \|Hx - w\|^2.$$



Typical observation

Low-rank recovery

$$\min_{x \in \mathbb{R}^{n\times n}} \mu\|x\|_* + \frac{1}{2}\|Hx - w\|^2.$$



"Activity" of x_k : $q = \mathrm{rank}(x^\star)$

- $\mathrm{rank}(x) \in]q, n[$: $k \leq K$
- $\mathrm{rank}(x) = q$: $k \geq K$

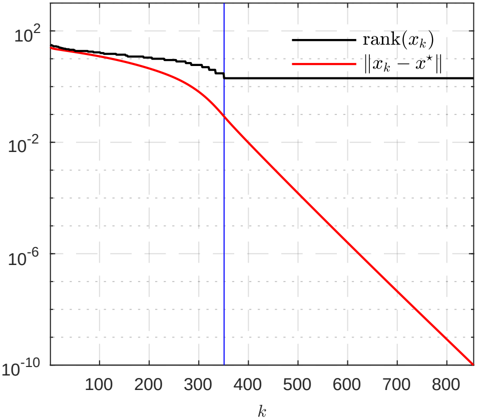
Rate of convergence:

- Sub-linear: $k \leq K$
- Linear: $k \geq K$

Typical observation

Low-rank recovery

$$\min_{x \in \mathbb{R}^{n \times n}} \mu \|x\|_* + \frac{1}{2}\|Hx - w\|^2.$$



"Activity" of x_k : $q = \mathrm{rank}(x^\star)$

- $\mathrm{rank}(x) \in]q, n[$: $k \leq K$
- $\mathrm{rank}(x) = q$: $k \geq K$

Rate of convergence:

- Sub-linear: $k \leq K$
- Linear: $k \geq K$

Phase transition of convergence rate coincides with that of "activity".

Open questions

- What are the possible mechanisms underlying the identification of "activity"?
- How fast is the global sub-linear convergence rate?
- How to explain the local linear convergence?
- What is the relation between local linear convergence and the identification of "activity"?
- Can we accelerate
  - the local convergence rate?
  - higher-order methods?

Contributions

Specific problems (*e.g.* ℓ_1-norm)

Specific algorithms (*e.g.* FB)

Cannot explain "phase transition"

A unified framework is missing!

Contributions

Specific problems (*e.g.* ℓ_1-norm)

Specific algorithms (*e.g.* FB)

Cannot explain "phase transition"

A unified framework is missing!

- ~~Global $o(1/\sqrt{k})$ sub-linear convergence rate for $\|x_k - x_{k-1}\|$.~~

Contributions

Specific problems (*e.g.* ℓ_1-norm)

Specific algorithms (*e.g.* FB)

Cannot explain "phase transition"

A unified framework is missing!

- ~~Global $o(1/\sqrt{k})$ sub-linear convergence rate for $\|x_k - x_{k-1}\|$.~~
- A unified framework for:
  - Finite time activity identification
  - Local linear convergence
  - Relation between the two...

  Covers both ***deterministic*** and ***stochastic*** setting.

Contributions

Specific problems (*e.g.* ℓ_1-norm)

Specific algorithms (*e.g.* FB)

Cannot explain "phase transition"

$\left.\begin{array}{l} \\ \\ \\ \end{array}\right\}$ **A unified framework is missing!**

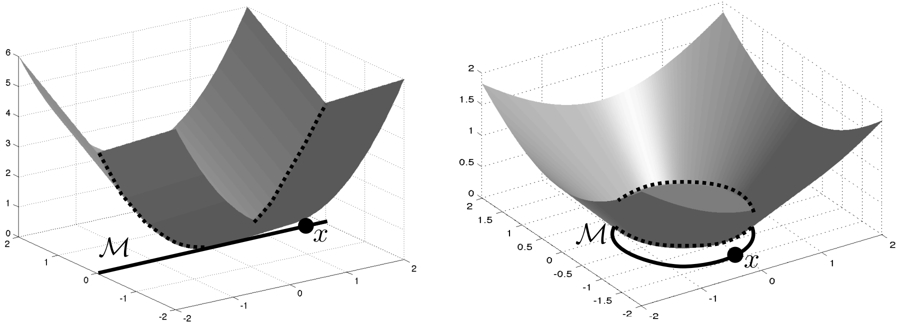
- ~~Global $o(1/\sqrt{k})$ sub-linear convergence rate for $\|x_k - x_{k-1}\|$.~~
- A unified framework for:
  - Finite time activity identification
  - Local linear convergence
  - Relation between the two...

  Covers both ***deterministic*** and ***stochastic*** setting.
- Geometry based acceleration which bridges 1st-order and 2nd-order methods.

Part III

Partial Smoothness and a Unified Framework

Partial smoothness



- A partly smooth function behaves smoothly along a manifold $\mathcal{M}$, and sharply normal to it.
- The behaviours of the function and its minimizers depend essentially on their restrictions to the manifold.
- It offers a powerful framework for algorithmic and sensitivity analysis.

Partial smoothness

Partly smooth function [Lewis, 2003]

Let $R \in \Gamma_0(\mathbb{R}^n)$, R is *partly smooth at x relative to a set $\mathcal{M}_x$* containing x if $\partial R(x) \neq \emptyset$

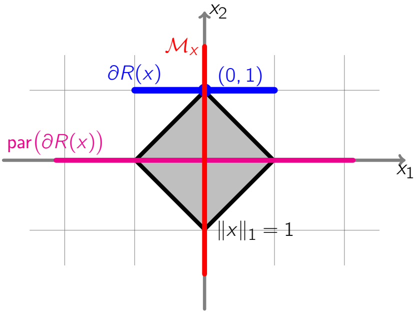
$$
\begin{aligned}
\textbf{Smoothness}: &\ \mathcal{M}_x \text{ is a } C^2\text{-manifold, } R|_{\mathcal{M}_x} \text{ is } C^2 \text{ near } x. \\
\textbf{Sharpness}: &\ \text{Tangent space } \mathcal{T}_{\mathcal{M}_x}(x) \text{ is } T_x \stackrel{\text{def}}{=} \mathrm{par}\big(\partial R(x)\big)^{\perp}. \\
\textbf{Continuity}: &\ \partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n \text{ is continuous along } \mathcal{M}_x \text{ near } x.
\end{aligned}
$$

Partial smoothness

Partly smooth function [Lewis, 2003]

Let $R \in \Gamma_0(\mathbb{R}^n)$, R is *partly smooth at x relative to a set $\mathcal{M}_x$* containing x if $\partial R(x) \neq \emptyset$

$$\begin{aligned} \textbf{Smoothness}:&\ \mathcal{M}_x \text{ is a } C^2\text{-manifold, } R|_{\mathcal{M}_x} \text{ is } C^2 \text{ near } x. \\ \textbf{Sharpness}:&\ \text{Tangent space } \mathcal{T}_{\mathcal{M}_x}(x) \text{ is } T_x \overset{\text{def}}{=} \mathrm{par}\big(\partial R(x)\big)^{\perp}. \\ \textbf{Continuity}:&\ \partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n \text{ is continuous along } \mathcal{M}_x \text{ near } x. \end{aligned}$$



Calculus rules:

- Sum and composition
- Smooth perturbation
- Spectral lifting

par(C): sub-space parallel to C, where $C \subset \mathbb{R}^n$ is a non-empty convex set.

Partial smoothness

Partly smooth function [Lewis, 2003]

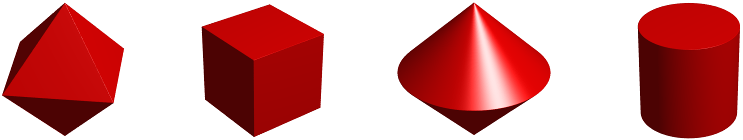
Let $R \in \Gamma_0(\mathbb{R}^n)$, R is *partly smooth at x relative to a set $\mathcal{M}_x$* containing x if $\partial R(x) \neq \emptyset$

Smoothness : $\mathcal{M}_x$ is a C^2-manifold, $R|_{\mathcal{M}_x}$ is C^2 near x.

Sharpness : Tangent space $\mathcal{T}_{\mathcal{M}_x}(x)$ is $T_x \stackrel{\text{def}}{=} \mathrm{par}\big(\partial R(x)\big)^{\perp}$.

Continuity : $\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is continuous along $\mathcal{M}_x$ near x.

$\mathrm{PSF}_x(\mathcal{M}_x)$: partly smooth function at x relative to $\mathcal{M}_x$



ℓ_1-norm | ℓ_∞-norm | $\ell_{1,2}$-norm | Nuclear norm

A unified framework

$\boxed{\text{Proximal splitting algorithm}}$

A unified framework

$\boxed{\textbf{Proximal splitting algorithm}}$

$$\Downarrow$$

Convergence of objective function, sequence

A unified framework

$\boxed{\textbf{Proximal splitting algorithm}}$

$\Downarrow$

Convergence of objective function, sequence

$\Downarrow$

$\boxed{\textbf{Non-degeneracy condition: finite activity identification}}$

A unified framework

$\boxed{\text{Proximal splitting algorithm}}$

$\Downarrow$

Convergence of objective function, sequence

$\Downarrow$

$\boxed{\text{Non-degeneracy condition: finite activity identification}}$

$\Downarrow$

Local linearised iteration: matrix M

A unified framework

$\boxed{\textbf{Proximal splitting algorithm}}$

$\Downarrow$

Convergence of objective function, sequence

$\Downarrow$

$\boxed{\textbf{Non-degeneracy condition: finite activity identification}}$

$\Downarrow$

Local linearised iteration: matrix M

$\Downarrow$

Spectral properties of M

A unified framework

$\boxed{\textbf{Proximal splitting algorithm}}$

$\Downarrow$

Convergence of objective function, sequence

$\Downarrow$

$\boxed{\textbf{Non-degeneracy condition: finite activity identification}}$

$\Downarrow$

Local linearised iteration: matrix M

$\Downarrow$

Spectral properties of M

$\Downarrow$

$\boxed{\textbf{Local linear convergence}}$

A unified framework

$\boxed{\text{Proximal splitting algorithm}}$ **(non-linear)**

$\Downarrow$

Convergence of objective function, sequence

$\Downarrow$

$\boxed{\text{Non-degeneracy condition: finite activity identification}}$

$\Downarrow$

Local linearised iteration: matrix M (**linear**)

$\Downarrow$

Spectral properties of M

$\Downarrow$

$\boxed{\text{Local linear convergence}}$

A unified framework

- Forward–Backward-type:
  - FB, inertial FB, FISTA [L, Jalal & Peyré, 14, 17]
  - Stochastic variants [Poon, L & Schönlieb, 18]
  - Non-convex [L, Jalal & Peyré, 16]
- Douglas–Rachford splitting, ADMM [L, Jalal & Peyré, 16]
- Class of Primal–Dual splitting methods [L, Jalal & Peyré, 18]
- Forward–Douglas–Rachford/Generalized Forward–Backward splitting, Three-operator splitting [Molinari, L & Jalal, 18]

A unified framework

- Forward–Backward-type:
  - FB, inertial FB, FISTA [L, Jalal & Peyré, 14, 17]
  - Stochastic variants [Poon, L & Schönlieb, 18]
  - Non-convex [L, Jalal & Peyré, 16]
- Douglas–Rachford splitting, ADMM [L, Jalal & Peyré, 16]
- Class of Primal–Dual splitting methods [L, Jalal & Peyré, 18]
- Forward–Douglas–Rachford/Generalized Forward–Backward splitting, Three-operator splitting [Molinari, L & Jalal, 18]

Forward–Backward splitting

Recall the optimisation problem

$$\min_{x \in \mathbb{R}^n} \left\{ \Phi(x) = R(x) + F(x) \right\}.$$

Forward–Backward splitting

Recall the optimisation problem

$$\min_{x\in\mathbb{R}^n} \big\{\Phi(x) = R(x) + F(x)\big\}.$$

Inexact Forward–Backward (iFB)

Let $\gamma_k \in]0, 2/L[$ and $\epsilon_k \in \mathbb{R}^n$:

$$x_{k+1} = \operatorname{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \epsilon_k)\big)$$

Forward–Backward splitting

Recall the optimisation problem

$$\min_{x \in \mathbb{R}^n} \big\{ \Phi(x) = R(x) + F(x) \big\}.$$

Inexact Forward–Backward (iFB)

Let $\gamma_k \in]0, 2/L[$ and $\epsilon_k \in \mathbb{R}^n$:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \epsilon_k)\big)$$

Remark For $F = \frac{1}{m} \sum_i f_i$: $i_k \in \{1, \cdots, n\}$

Forward–Backward splitting

Recall the optimisation problem

$$\min_{x \in \mathbb{R}^n} \; \big\{ \Phi(x) = R(x) + F(x) \big\}.$$

Inexact Forward–Backward (iFB)

Let $\gamma_k \in]0, 2/L[$ and $\epsilon_k \in \mathbb{R}^n$:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \epsilon_k)\big)$$

Remark For $F = \frac{1}{m}\sum_i f_i$: $i_k \in \{1, \cdots, n\}$

- Stochastic gradient descent (SGD):

$$\epsilon_k = \nabla f_{i_k}(x_k) - \nabla F(x_k).$$

Forward–Backward splitting

Recall the optimisation problem

$$\min_{x\in\mathbb{R}^n} \ \big\{\Phi(x) = R(x) + F(x)\big\}.$$

Inexact Forward–Backward (iFB)

Let $\gamma_k \in]0, 2/L[$ and $\epsilon_k \in \mathbb{R}^n$:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \epsilon_k)\big)$$

Remark For $F = \frac{1}{m}\sum_i f_i$: $i_k \in \{1, \cdots, n\}$

- Stochastic gradient descent (SGD):

$$\epsilon_k = \nabla f_{i_k}(x_k) - \nabla F(x_k).$$

- SAGA [Defazio et al, 14]:

$$\epsilon_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{k-1}) + \tfrac{1}{m}\textstyle\sum_{j=1}^m \nabla f_{i_{k-j}}(x_{k-j}) - \nabla F(x_k).$$

Forward–Backward splitting

Recall the optimisation problem

$$\min_{x\in\mathbb{R}^n} \ \big\{\Phi(x) = R(x) + F(x)\big\}.$$

Inexact Forward–Backward (iFB)

Let $\gamma_k \in]0, 2/L[$ and $\epsilon_k \in \mathbb{R}^n$:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \epsilon_k)\big)$$

Remark For $F = \frac{1}{m}\sum_i f_i$: $i_k \in \{1, \cdots, n\}$

- Stochastic gradient descent (SGD):

$$\epsilon_k = \nabla f_{i_k}(x_k) - \nabla F(x_k).$$

- SAGA [Defazio et al, 14]:

$$\epsilon_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{k-1}) + \tfrac{1}{m}\textstyle\sum_{j=1}^m \nabla f_{i_{k-j}}(x_{k-j}) - \nabla F(x_k).$$

- Prox-SVRG [Xiao & Zhang, 14]

$$\epsilon_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}_\ell) + \nabla F(\tilde{x}_\ell) - \nabla F(x_k).$$

Step 1 - Convergence properties

Let $x^\star \in \mathrm{Argmin}(\Phi)$ be a global minimiser.

Convergence of iFB [L, Jalal & Peyré, 16, Poon, L & Schönlieb, 18]

Deterministic: $x_k \to x^\star$ if

$$\gamma_k \in]0, 2/L[\quad \text{and} \quad \textstyle\sum_k \|\epsilon_k\| < +\infty.$$

Stochastic: $x_k \to x^\star$ almost surely if

$$\gamma_k \equiv \gamma \in]0, 1/L[\quad \text{and} \quad \textstyle\sum_k \mathbb{E}[\|\epsilon_k\|^2] < +\infty.$$

Step 1 - Convergence properties

Let $x^\star \in \operatorname{Argmin}(\Phi)$ be a global minimiser.

Convergence of iFB [L, Jalal & Peyré, 16, Poon, L & Schönlieb, 18]

Deterministic: $x_k \to x^\star$ if

$$\gamma_k \in]0, 2/L[\quad \text{and} \quad \sum_k \|\epsilon_k\| < +\infty.$$

Stochastic: $x_k \to x^\star$ almost surely if

$$\gamma_k \equiv \gamma \in]0, 1/L[\quad \text{and} \quad \sum_k \mathbb{E}[\|\epsilon_k\|^2] < +\infty.$$

Remark

- The convergence of $\Phi(x_k)$ follows that of x_k

Step 1 - Convergence properties

Let $x^\star \in \mathrm{Argmin}(\Phi)$ be a global minimiser.

Convergence of iFB [L, Jalal & Peyré, 16, Poon, L & Schönlieb, 18]

Deterministic: $x_k \to x^\star$ if

$$\gamma_k \in]0, 2/L[\quad \text{and} \quad \sum_k \|\epsilon_k\| < +\infty.$$

Stochastic: $x_k \to x^\star$ almost surely if

$$\gamma_k \equiv \gamma \in]0, 1/L[\quad \text{and} \quad \sum_k \mathbb{E}[\|\epsilon_k\|^2] < +\infty.$$

Remark

- The convergence of $\Phi(x_k)$ follows that of x_k
- SPG:

$$\mathbb{E}[\|\epsilon_k\|] \in]0, +\infty[.$$

Step 1 - Convergence properties

Let $x^\star \in \text{Argmin}(\Phi)$ be a global minimiser.

Convergence of iFB [L, Jalal & Peyré, 16, Poon, L & Schönlieb, 18]

Deterministic: $x_k \to x^\star$ if

$$\gamma_k \in]0, 2/L[\quad \text{and} \quad \textstyle\sum_k \|\epsilon_k\| < +\infty.$$

Stochastic: $x_k \to x^\star$ almost surely if

$$\gamma_k \equiv \gamma \in]0, 1/L[\quad \text{and} \quad \textstyle\sum_k \mathbb{E}[\|\epsilon_k\|^2] < +\infty.$$

Remark

- The convergence of $\Phi(x_k)$ follows that of x_k
- SPG:

$$\mathbb{E}[\|\epsilon_k\|] \in]0, +\infty[.$$

- SAGA/Prox-SVRG:

$$\mathbb{E}[\|\epsilon_k\|] \to 0.$$

Step 2 - Finite activity identification

Let $x^\star \in \operatorname{Argmin}(\Phi)$, then

$$0 \in \nabla F(x^\star) + \partial R(x^\star).$$

Step 2 - Finite activity identification

Let $x^\star \in \mathrm{Argmin}(\Phi)$, then

$$0 \in \nabla F(x^\star) + \partial R(x^\star).$$

Finite identification [L, Jalal & Peyré, 17, Poon, L & Schönlieb, 18]

Let the convergence of iFB hold. Suppose that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and the non-degeneracy condition

$$0 \in \mathrm{ri}\big(\nabla F(x^\star) + \partial R(x^\star)\big), \tag{ND}$$

holds. Then, there exists a $K \geq 0$ such that for all $k \geq K$:

$$x_k \in \mathcal{M}_{x^\star}.$$

Step 2 - Finite activity identification

Let $x^\star \in \mathrm{Argmin}(\Phi)$, then

$$0 \in \nabla F(x^\star) + \partial R(x^\star).$$

Finite identification [L, Jalal & Peyré, 17, Poon, L & Schönlieb, 18]

Let the convergence of iFB hold. Suppose that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and the non-degeneracy condition

$$0 \in \mathrm{ri}\big(\nabla F(x^\star) + \partial R(x^\star)\big), \tag{ND}$$

holds. Then, there exists a $K \geq 0$ such that for all $k \geq K$:

$$x_k \in \mathcal{M}_{x^\star}.$$

Remark

- A bound on K can be provided.

Step 2 - Finite activity identification

Let $x^\star \in \mathrm{Argmin}(\Phi)$, then

$$0 \in \nabla F(x^\star) + \partial R(x^\star).$$

Finite identification [L, Jalal & Peyré, 17, Poon, L & Schönlieb, 18]

Let the convergence of iFB hold. Suppose that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and the non-degeneracy condition

$$0 \in \mathrm{ri}\big(\nabla F(x^\star) + \partial R(x^\star)\big), \tag{ND}$$

holds. Then, there exists a $K \geq 0$ such that for all $k \geq K$:

$$x_k \in \mathcal{M}_{x^\star}.$$

Remark

- A bound on K can be provided.
- Stochastic proximal gradient does **NOT** have finite identification.

Step 2 - Finite activity identification

Let $x^\star \in \mathrm{Argmin}(\Phi)$, then

$$0 \in \nabla F(x^\star) + \partial R(x^\star).$$

Finite identification [L, Jalal & Peyré, 17, Poon, L & Schönlieb, 18]

Let the convergence of iFB hold. Suppose that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and the non-degeneracy condition

$$0 \in \mathrm{ri}\big(\nabla F(x^\star) + \partial R(x^\star)\big), \tag{ND}$$

holds. Then, there exists a $K \geq 0$ such that for all $k \geq K$:

$$x_k \in \mathcal{M}_{x^\star}.$$

Remark

- A bound on K can be provided.
- Stochastic proximal gradient does NOT have finite identification.
- The identification of SAGA/Prox-SVRG is almost surely.

Step 3 - Local linearisation: deterministic

Local linearisation [L, Jalal & Peyré, 17]

For the iFB iteration, suppose the Identification theorem holds. If F is locally C^2 around $x^\star$,

$$\gamma_k \to \gamma \in]0, 2/L[,$$

then for all k large enough, there exist a matrix M such that

$$x_{k+1} - x^\star = M(x_k - x^\star) + o(\|x_k - x^\star\|) + \epsilon_k.$$

Step 3 - Local linearisation: deterministic

Local linearisation [L, Jalal & Peyré, 17]

For the iFB iteration, suppose the Identification theorem holds. If F is locally C^2 around $x^\star$,

$$\gamma_k \to \gamma \in]0, 2/L[,$$

then for all k large enough, there exist a matrix M such that

$$x_{k+1} - x^\star = M(x_k - x^\star) + o(\|x_k - x^\star\|) + \epsilon_k.$$

Remark

- $o(\|x_k - x^\star\|)$ vanishes if R is locally polyhedral around $x^\star$, and

$$\gamma_k \equiv \gamma.$$

Step 3 - Local linearisation: deterministic

Local linearisation [L, Jalal & Peyré, 17]

For the iFB iteration, suppose the Identification theorem holds. If F is locally C^2 around $x^\star$,

$$\gamma_k \to \gamma \in]0, 2/L[,$$

then for all k large enough, there exist a matrix M such that

$$x_{k+1} - x^\star = M(x_k - x^\star) + o(\|x_k - x^\star\|) + \epsilon_k.$$

Remark

- $o(\|x_k - x^\star\|)$ vanishes if R is locally polyhedral around $x^\star$, and

$$\gamma_k \equiv \gamma.$$

- M is similar to a symmetric positive semidefinite matrix.

Step 4 - Local linear convergence: deterministic

Restricted injectivity : $\exists \alpha > 0$ such that $\forall h \in T_{x^\star}$,

$$\langle h,\, \nabla^2 F(x^\star) h \rangle \geq \alpha \|h\|^2. \tag{RI}$$

Step 4 - Local linear convergence: deterministic

Restricted injectivity : $\exists \alpha > 0$ such that $\forall h \in T_{x^\star}$,

$$\langle h,\, \nabla^2 F(x^\star) h \rangle \geq \alpha \|h\|^2. \tag{RI}$$

Spectral radius of M [L, Jalal & Peyré, 17]

For matrix M, suppose (RI) holds, then $\rho(M) < 1$ as long as

$$\gamma \in]0, 2/L[,$$

and $\rho(M)$ can be given explicitly.

Step 4 - Local linear convergence: deterministic

Restricted injectivity : $\exists \alpha > 0$ such that $\forall h \in T_{x^\star}$,

$$\langle h,\, \nabla^2 F(x^\star) h \rangle \geq \alpha \|h\|^2. \tag{RI}$$

Spectral radius of M [L, Jalal & Peyré, 17]

For matrix M, suppose (RI) holds, then $\rho(M) < 1$ as long as

$$\gamma \in]0, 2/L[,$$

and $\rho(M)$ can be given explicitly.

Local linear convergence [L, Jalal & Peyré, 17]

Suppose iFB creates a sequence $x_k \to x^\star \in \mathrm{Argmin}(\Phi)$ such that the Identification, Linearisation and Spectral radius theorems hold, and $\|\epsilon_k\|$ decays fast enough. Then given any $\rho \in [\rho(M), 1[$, there is K large enough such that for all $k \geq K$,

$$\|x_k - x^\star\| = O(\rho^k).$$

Step 4 - Local linear convergence: stochastic

Quadratic growth [L, Jalal & Peyré, 17]

Let $x^\star \in \mathrm{Argmin}(\Phi)$ be such that (ND) and (RI) are fulfilled and $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, then $x^\star$ is the unique minimiser of Φ and there exist $\alpha > 0$ and $r > 0$ such that

$$\Phi(x) - \Phi(x^\star) \geq \alpha \|x - x^\star\|^2 : \ \forall x \ \ \text{s.t.} \ \ \|x - x^\star\| \leq r.$$

Step 4 - Local linear convergence: stochastic

Quadratic growth [L, Jalal & Peyré, 17]

Let $x^\star \in \mathrm{Argmin}(\Phi)$ be such that (ND) and (RI) are fulfilled and $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, then $x^\star$ is the unique minimiser of Φ and there exist $\alpha > 0$ and $r > 0$ such that

$$\Phi(x) - \Phi(x^\star) \geq \alpha\|x - x^\star\|^2 : \ \forall x \ \text{ s.t. } \ \|x - x^\star\| \leq r.$$

Local linear convergence [Poon, L & Schönlieb, 18]

Suppose iFB creates a sequence $x_k \to x^\star \in \mathrm{Argmin}(\Phi)$ such that the Identification theorem and condition (RI) hold. Then there exists $\rho < 1$ such that for all k large enough,

$$\mathbb{E}[\|x_k - x^\star\|] = O(\rho^k).$$

Step 4 - Local linear convergence: stochastic

Quadratic growth [L, Jalal & Peyré, 17]

Let $x^\star \in \text{Argmin}(\Phi)$ be such that (ND) and (RI) are fulfilled and $R \in \text{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, then $x^\star$ is the unique minimiser of Φ and there exist $\alpha > 0$ and $r > 0$ such that

$$\Phi(x) - \Phi(x^\star) \geq \alpha\|x - x^\star\|^2 : \ \forall x \ \text{ s.t. } \ \|x - x^\star\| \leq r.$$

Local linear convergence [Poon, L & Schönlieb, 18]

Suppose iFB creates a sequence $x_k \to x^\star \in \text{Argmin}(\Phi)$ such that the Identification theorem and condition (RI) hold. Then there exists $\rho < 1$ such that for all k large enough,

$$\mathbb{E}[\|x_k - x^\star\|] = O(\rho^k).$$

Remark The theoretical rate estimation of in general is not as tight as their deterministic counterparts.

Higher-order acceleration

$$\begin{array}{ccc} \text{global} & & \text{local} \\ \text{non-smooth}\,(\mathbb{R}^n) & \xrightarrow[]{\textit{finite activity iden.}} & C^2\text{-smooth}\,(\mathcal{M}) \end{array}$$

Local condition	better Lipschitz constant along $\mathcal{M}_{x^\star}$
Locally polyhedral	finite termination if F is quadratic
General manifold	Newton-like, Conjugate gradient, Manifold based optimisation methods

Part V

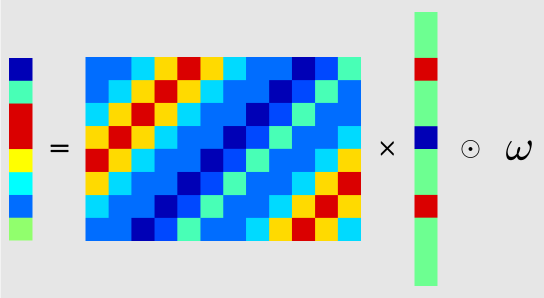
Numerical Experiments

Examples

Sparse LR $(z_i, y_i) \in \mathbb{R}^n \times \{\pm 1\}, m = 64, n = 96$

$$\min_{(b,x)\in\mathbb{R}\times\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} f(\langle x, z_i\rangle + b, y_i) + \mu\|x\|_1,$$

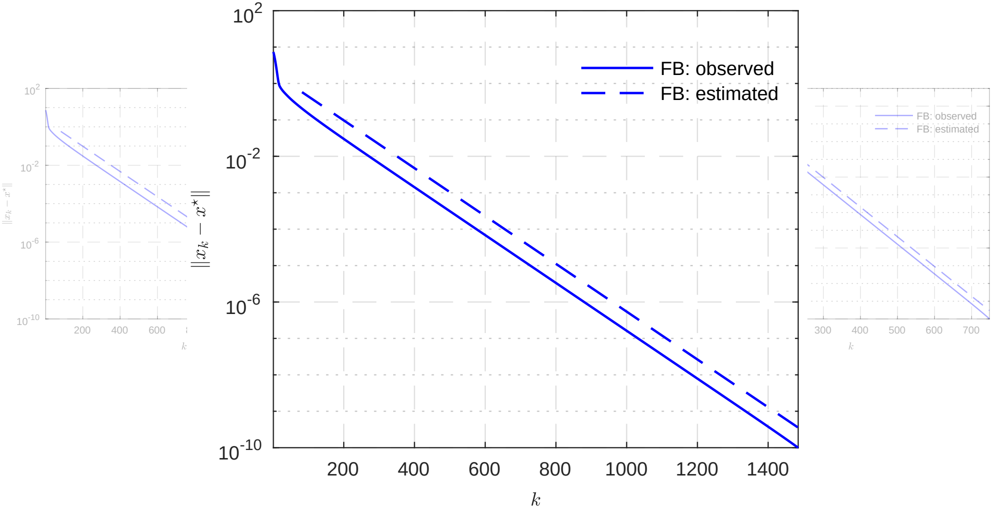
where $f(w_i, y_i) = \log(1 + e^{-w_i y_i})$.



$\min_x \mu R(x) + \frac{1}{2}\|Hx - w\|^2$

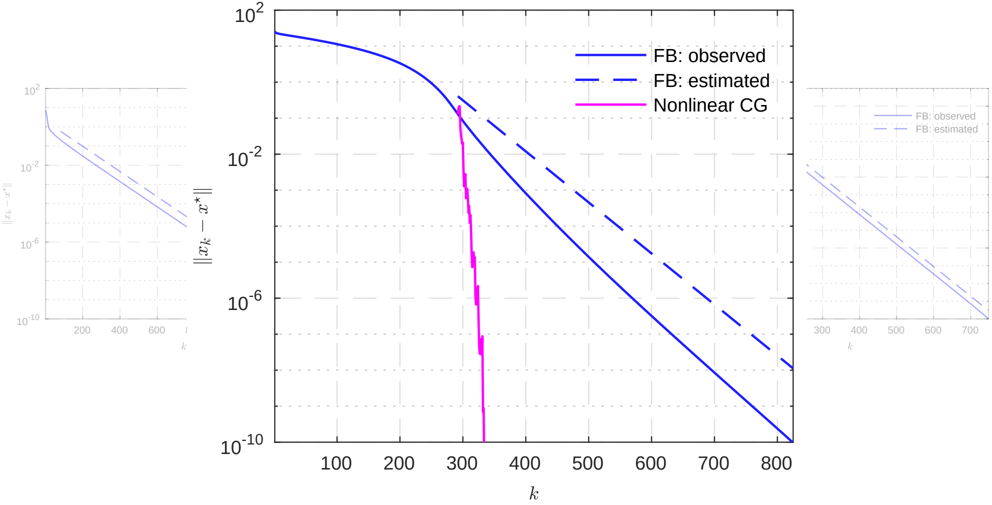
$\min_x \mu\|\nabla x\|_1 + \frac{1}{2}\|Hx - w\|^2$

Numerical result: deterministic



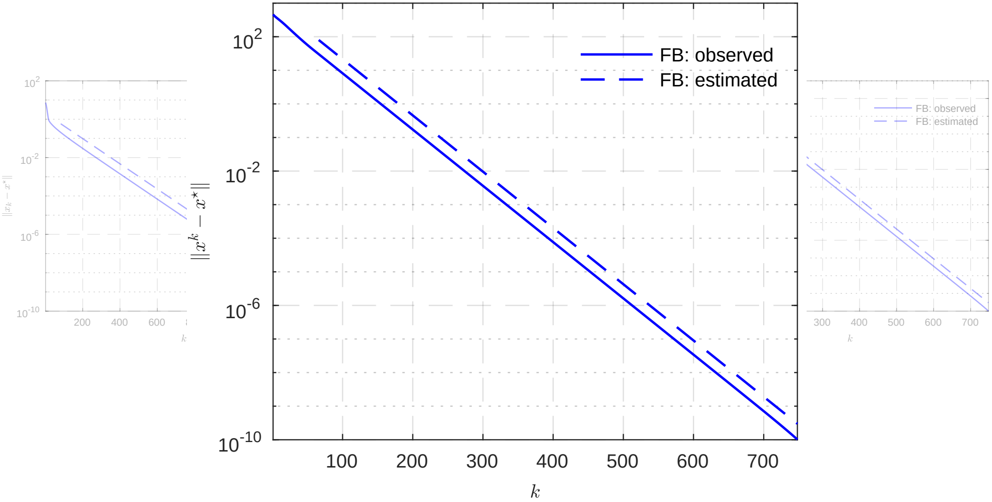
Sparse LR

Numerical result: deterministic



Low-rank recovery

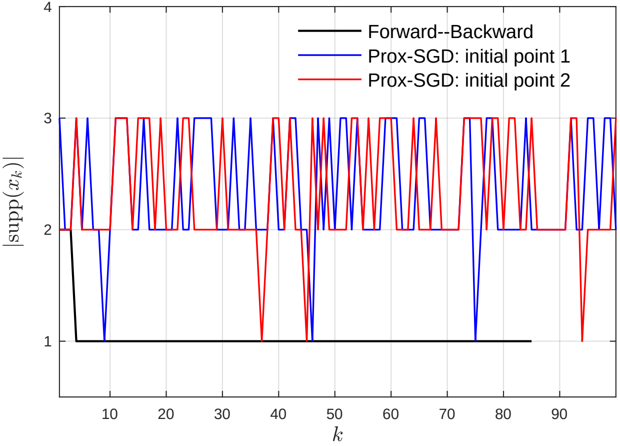
Numerical result: deterministic



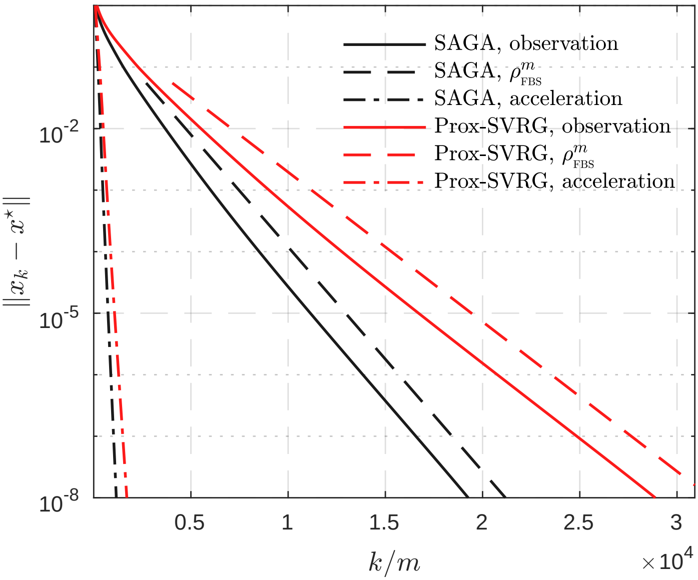
TV deblur

Numerical result: stochastic

$$\min_{x\in\mathbb{R}^3} \frac{1}{3}\|x\|_1 + \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}\|H_i x - b_i\|^2,\quad H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{3} \end{bmatrix} \text{ and } b = \begin{pmatrix} 2 \\ \sqrt{2}/3 \\ \sqrt{3}/4 \end{pmatrix}.$$



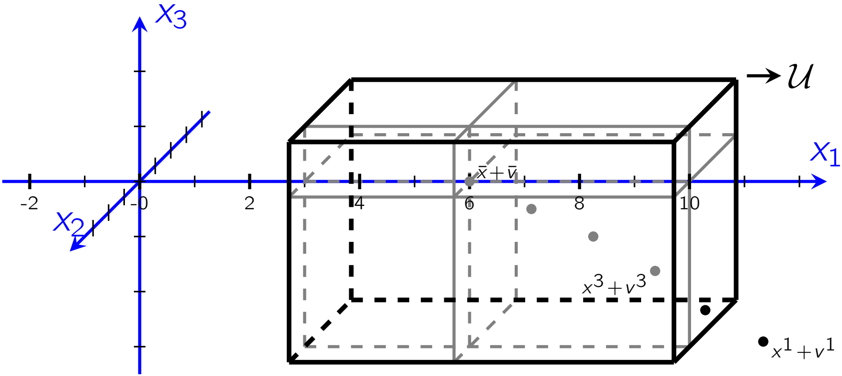
Numerical result: stochastic



Extension of partial smoothness

Extensions

- Beyond non-degeneracy: enlarged manifold (with J. Fadili)
- Beyond optimization: set-valued operators (with A. Lewis)



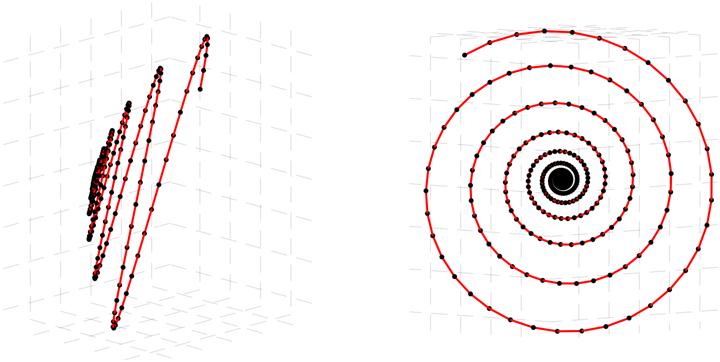
Let $\bar{x} = (5; 0; 0)$, $\mathcal{M}_x = [\mathbb{R}; 0; 0]$, $A = \partial\|\cdot\|_1$ and $\bar{v} \in \mathrm{ri}(A(\bar{x}))$:

$$\mathcal{U} = \cup_{x \in \mathcal{M}_x} \big(x + \partial\|\cdot\|_1(x)\big) \to \bar{x} + \bar{v} \in \mathrm{int}(\mathcal{U})$$

Adaptive acceleration

Non-smooth opt. $\xRightarrow{\text{structure}}$ FoM $\xRightarrow{\text{geometry?}}$ Tra. of Seq. $\Longrightarrow$ Acceleration?

Power iteration in 3D: 2nd biggest eigenvalue is complex; the trajectory of eigenvector of the biggest eigenvalue



Ada-acceleration (with C. Poon): adaptive acceleration based on geometry

Takeaway messages

Partial smoothness builds an elegant connection between functions and the underlying Riemannian geometry

A unified framework for local analysis

Higher-order acceleration

Better understanding of existing algorithms

Steer new direction for designing accelerated schemes

Thank you very much!

https://jliang993.github.io/