Newton methods for structured nonsmooth optimization

proximal identification and fast local convergence

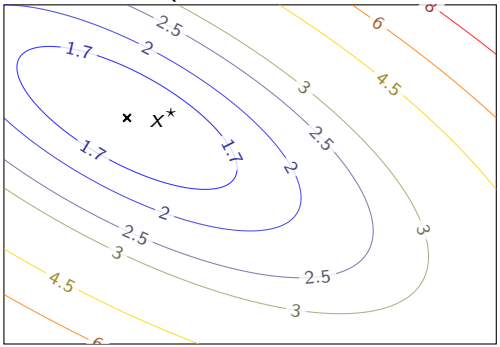
Gilles Bareilles
LJK, Univ. Grenoble Alpes
gbareilles.fr
December 2, 2022

Chair	Nadia Brauner
Reviewer	Jalal Fadili
Reviewer	Claudia Sagastizábal
Examiner	Jean-Charles Gilbert
Examiner	Mathurin Massias
Advisor	Franck Iutzeler
Advisor	Jérôme Malick
Guest	Claude Lemaréchal

What this PhD is about

Optimization of functions: $\min_x F(x)$

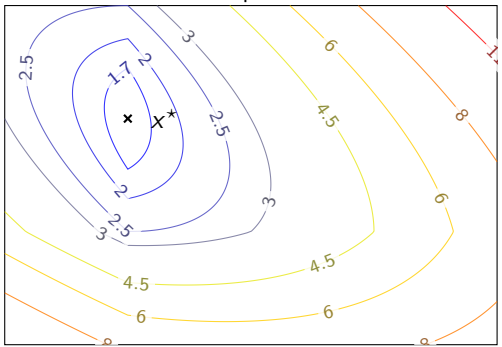
Quadratic function



$F(x) = \|Ax - b\|_2^2$

Smooth

Lasso problem



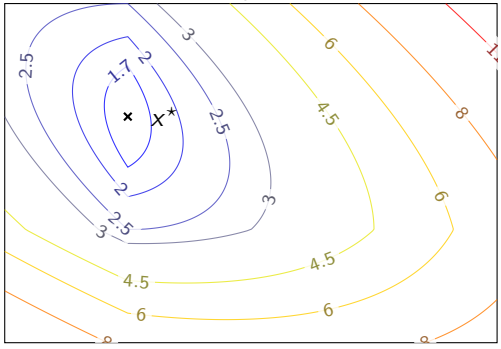
$F(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1$

Nonsmooth

What this PhD is about

Optimization of nonsmooth functions: $\min_x F(x)$

Lasso problem



$$F(x) = \|Ax - b\|_2^2 + \lambda \|x\|_1$$

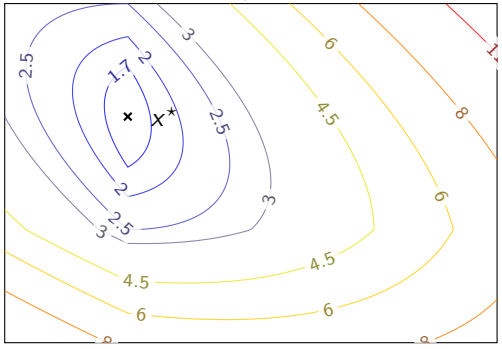
What this PhD is about

Optimization of **structured** nonsmooth functions: $\min_x F(x)$

▹ nonsmooth points are "well-behaved"

on x & y axes

Lasso problem

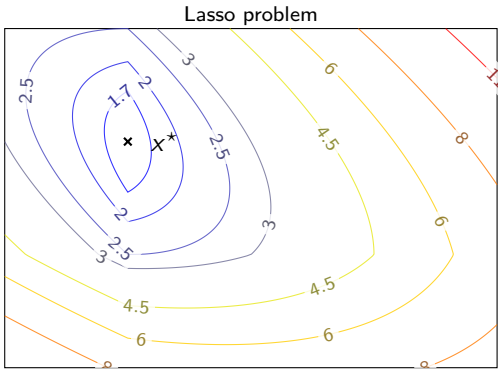


$$F(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1$$

What this PhD is about

Optimization of **structured** nonsmooth functions: $\min_x F(x)$

▹ nonsmooth points are "well-behaved"
on x & y axes



$$F(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1$$

Ok for lasso.
What about **structure in general nonsmooth functions**?

Sources of nonsmoothness 1

▷ **Implicit** nonsmoothness

$$F(x) = \sup_{u \in U} h(x, u)$$

Examples

- Robust optimization ◇ Ben-Tal *et al* '09
- Decomposition methods *e.g.*, Lagrangian relaxation, Benders decomposition ◇ BGLS '06
- Risk-averse optimization... advertising my team:
  - superquantile optim. ◇ Laguel Malick Harchaoui '22
  - distributionally robust optim. ◇ Azizian Iutzeler Malick '22

Sources of nonsmoothness 1

▹ **Implicit** nonsmoothness

$$F(x) = \sup_{u \in U} h(x, u)$$

Examples

- Robust optimization ◇ Ben-Tal *et al* '09
- Decomposition methods *e.g.*, Lagrangian relaxation, Benders decomposition ◇ BGLS '06
- Risk-averse optimization... advertising my team:
  - superquantile optim. ◇ Laguel Malick Harchaoui '22
  - distributionally robust optim. ◇ Azizian Iutzeler Malick '22

▹ *Full* **first-order** description of F near x *e.g.*, $\partial F(x)$?

Sources of nonsmoothness 1

▷ **Implicit** nonsmoothness

$$F(x) = \sup_{u \in U} h(x, u)$$

Examples

- Robust optimization ◇ Ben-Tal *et al* '09
- Decomposition methods *e.g.*, Lagrangian relaxation, Benders decomposition ◇ BGLS '06
- Risk-averse optimization... advertising my team:
  - superquantile optim. ◇ Laguel Malick Harchaoui '22
  - distributionally robust optim. ◇ Azizian Iutzeler Malick '22

▷ *Full* **first-order** description of F near x *e.g.*, $\partial F(x)$?

This would require the *full* set $\arg\max_{u \in U} h(x, u)$.

Sources of nonsmoothness 2

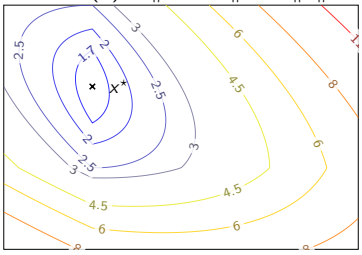
▷ **Chosen** nonsmoothness to attract minimizers

$$F(x) = f(x) + g(x), \quad \text{with } f \text{ smooth, } g \text{ nonsmooth}$$

Examples inverse problems, sparse regression *e.g.* lasso

◇ Scherzer *et al* '09, Vaiter *et al* '15

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



Sources of nonsmoothness 2

- **Chosen** nonsmoothness to attract minimizers

  $F(x) = f(x) + g(x)$, with f smooth, g nonsmooth

  Examples inverse problems, sparse regression *e.g.* lasso

  ◇ Scherzer *et al* '09, Vaiter *et al* '15

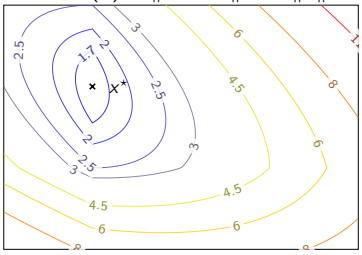
- **In-between** nonsmoothness

  $F(x) = g \circ c(x)$, with c smooth map, g nonsmooth

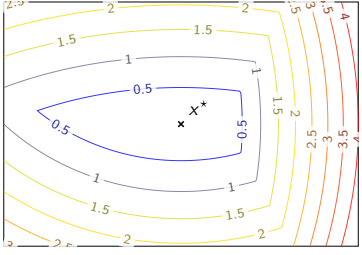
  Examples Robust regression, optimal control, SDP

  ◇ Shapiro '03, Noll '05, Lewis Wright '16

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$



Sources of nonsmoothness 2

- **Chosen** nonsmoothness to attract minimizers

  $F(x) = f(x) + g(x)$, with f smooth, g nonsmooth

  Examples inverse problems, sparse regression *e.g.* lasso

  ◇ Scherzer *et al* '09, Vaiter *et al* '15

- **In-between** nonsmoothness

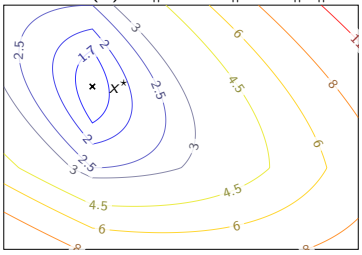
  $F(x) = g \circ c(x)$, with c smooth map, g nonsmooth

  Examples Robust regression, optimal control, SDP

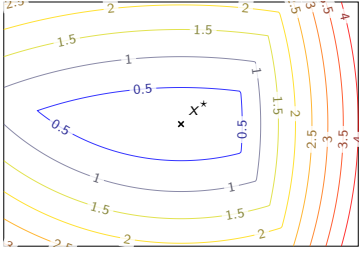
  ◇ Shapiro '03, Noll '05, Lewis Wright '16

- In these two cases, we know

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$



Sources of nonsmoothness 2

- **Chosen** nonsmoothness to attract minimizers

  $F(x) = f(x) + g(x)$, with f smooth, g nonsmooth

  Examples inverse problems, sparse regression *e.g.* lasso

  ◇ Scherzer *et al* '09, Vaiter *et al* '15

- **In-between** nonsmoothness

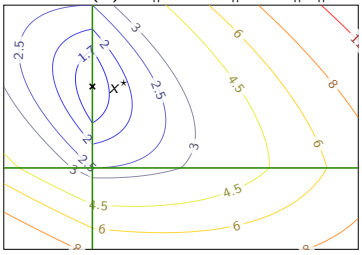
  $F(x) = g \circ c(x)$, with c smooth map, g nonsmooth

  Examples Robust regression, optimal control, SDP

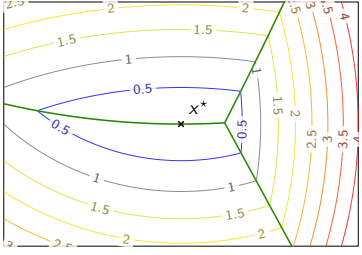
  ◇ Shapiro '03, Noll '05, Lewis Wright '16

- In these two cases, we know
  - where F is nonsmooth

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$



Sources of nonsmoothness 2

- **Chosen** nonsmoothness to attract minimizers

  $F(x) = f(x) + g(x)$, with f smooth, g nonsmooth

  Examples inverse problems, sparse regression *e.g.* lasso

  ◇ Scherzer *et al* '09, Vaiter *et al* '15

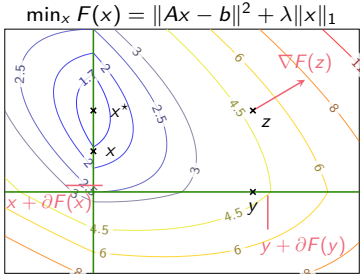
- **In-between** nonsmoothness

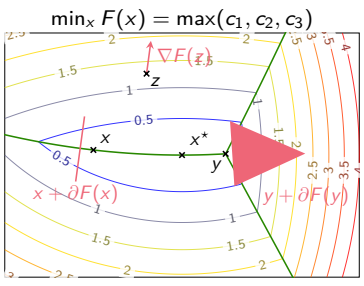
  $F(x) = g \circ c(x)$, with c smooth map, g nonsmooth

  Examples Robust regression, optimal control, SDP

  ◇ Shapiro '03, Noll '05, Lewis Wright '16

- In these two cases, we know
  - where F is nonsmooth
  - the full first-order description $\partial F(x)$ and more





Sources of nonsmoothness 2

- **Chosen** nonsmoothness to attract minimizers

  $F(x) = f(x) + g(x)$, with f smooth, g nonsmooth

  Examples inverse problems, sparse regression *e.g.* lasso

  ◇ Scherzer *et al* '09, Vaiter *et al* '15

- **In-between** nonsmoothness

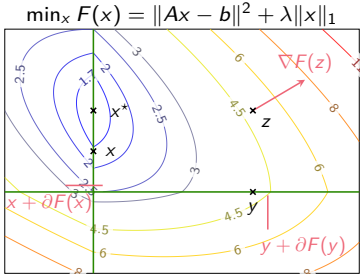
  $F(x) = g \circ c(x)$, with c smooth map, g nonsmooth

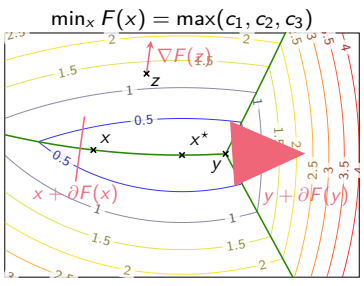
  Examples Robust regression, optimal control, SDP

  ◇ Shapiro '03, Noll '05, Lewis Wright '16

- In these two cases, we know
  - where F is nonsmooth
  - the full first-order description $\partial F(x)$ and more

  Here **nonsmoothness is explicit**. Focus of this talk

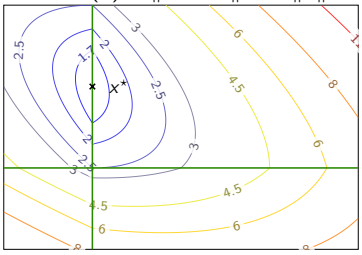




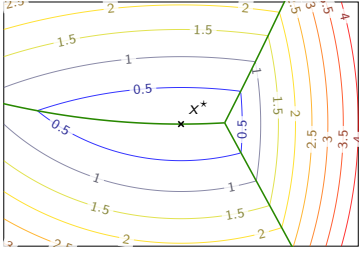
Structure in nonsmoothness

In most target applications, we **observe** that:

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$

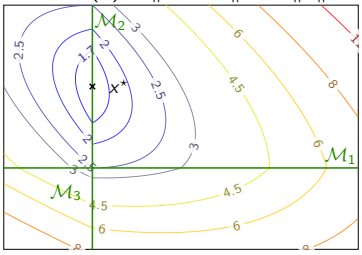


Structure in nonsmoothness

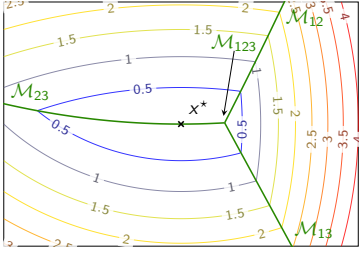
In most target applications, we **observe** that:

- nondiff. points organize in smooth manifolds $\mathcal{M}$

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$

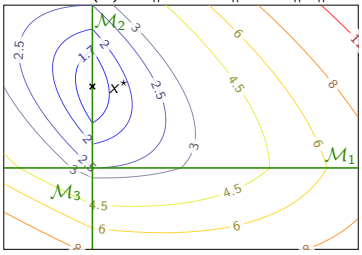


Structure in nonsmoothness

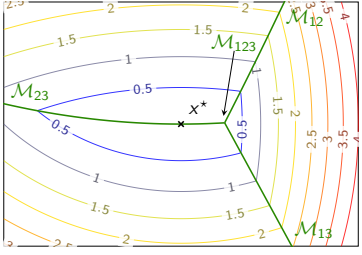
In most target applications, we **observe** that:

- nondiff. points organize in smooth manifolds $\mathcal{M}$
- locally, F is smooth along and nonsmooth across $\mathcal{M}$

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$



Structure in nonsmoothness

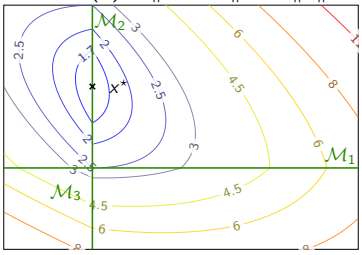
In most target applications, we **observe** that:

- nondiff. points organize in smooth manifolds $\mathcal{M}$
- locally, F is smooth along and nonsmooth across $\mathcal{M}$

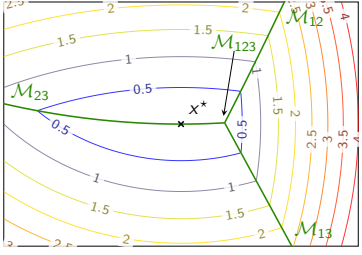
These are **structure manifolds**.

◇ Lewis '02, Fadili Malick Peyré '18, Davis Drusviatsky '19

$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$



Structure in nonsmoothness

In most target applications, we **observe** that:

- nondiff. points organize in smooth manifolds $\mathcal{M}$
- locally, F is smooth along and nonsmooth across $\mathcal{M}$

These are **structure manifolds**.

◇ Lewis '02, Fadili Malick Peyré '18, Davis Drusviatsky '19

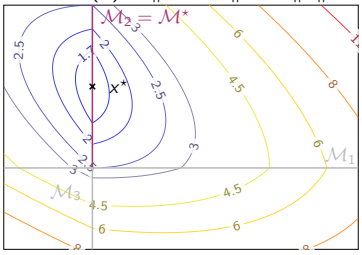
If $x^\star$ is nonsmooth for F, there is an

optimal manifold $\mathcal{M}^\star \ni x^\star$

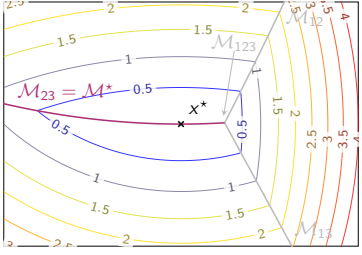
→ This PhD explores the following question

Can structure help optimization?

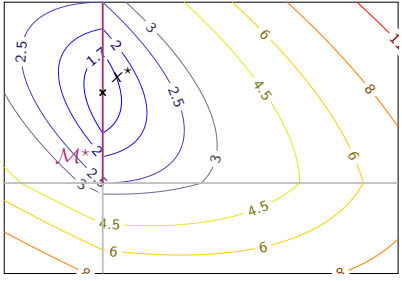
$\min_x F(x) = \|Ax - b\|^2 + \lambda\|x\|_1$



$\min_x F(x) = \max(c_1, c_2, c_3)$

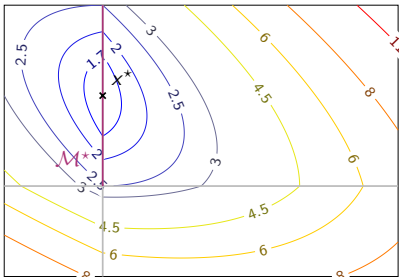


1. Detecting structure



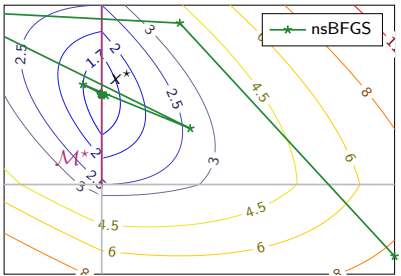
1. Detecting structure

- **Implicit** nonsmoothness: $x \mapsto F(x)$, $v \in \partial F(x)$
  - *nonsmooth BFGS* ◇ Lewis Overton '13
  - gradient sampling ◇ Lewis *et al* '05
  - bundle methods ◇ HULL '93



1. Detecting structure

- **Implicit** nonsmoothness: $x \mapsto F(x)$, $v \in \partial F(x)$
  - *nonsmooth BFGS* ◇ Lewis Overton '13
  - gradient sampling ◇ Lewis *et al* '05
  - bundle methods ◇ HULL '93



- nsBFGS: $x_k \notin \mathcal{M}^\star$, oscillations

→ **no detection** of $\mathcal{M}^\star$, suffered nonsmoothness

1. Detecting structure

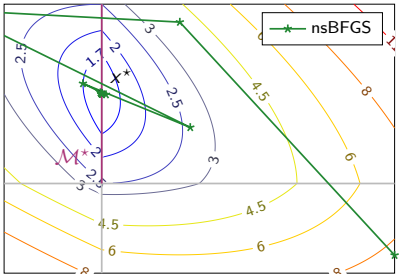
- **Implicit** nonsmoothness: $x \mapsto F(x)$, $v \in \partial F(x)$
  - *nonsmooth BFGS* ◇ Lewis Overton '13
  - gradient sampling ◇ Lewis *et al* '05
  - bundle methods ◇ HULL '93

- **Explicit** nonsmoothness: $\mathbf{prox}_{\gamma g}$
  - *splitting methods* ◇ Combettes Pesquet, '11
  - prox-linear methods ◇ Lewis Wright, '16

$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Implicit first-order step

Closed-form for simple nonsmooth functions



- nsBFGS: $x_k \notin \mathcal{M}^\star$, oscillations

$\rightarrow$ **no detection** of $\mathcal{M}^\star$, suffered nonsmoothness

1. Detecting structure

▹ **Implicit** nonsmoothness: $x \mapsto F(x)$, $v \in \partial F(x)$

- *nonsmooth BFGS* ◇ Lewis Overton '13
- gradient sampling ◇ Lewis *et al* '05
- bundle methods ◇ HULL '93

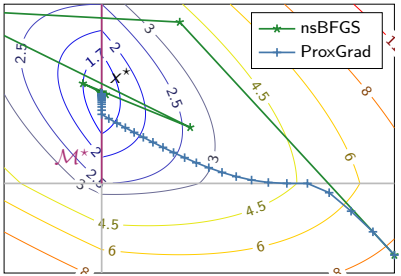
▹ **Explicit** nonsmoothness: $\mathbf{prox}_{\gamma g}$

- *splitting methods* ◇ Combettes Pesquet, '11
- prox-linear methods ◇ Lewis Wright, '16

$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Implicit first-order step

Closed-form for simple nonsmooth functions



▹ nsBFGS: $x_k \notin \mathcal{M}^\star$, oscillations
→ **no detection** of $\mathcal{M}^\star$, suffered nonsmoothness

▹ ProxGrad: after some time $x_k \in \mathcal{M}^\star$
→ **detection** of $\mathcal{M}^\star$, but **no exploitation**

2. Exploiting structure

▷ With knowledge of $\mathcal{M}^\star$, nonsmooth min. turns into smooth constrained opt.

$$\min_{x \in \mathbb{R}^n} F(x) \qquad \overset{\text{with } \mathcal{M}^\star}{\longrightarrow} \qquad \min_{x \in \mathcal{M}^\star} F|_{\mathcal{M}^\star}(x)$$

2. Exploiting structure

▷ With knowledge of $\mathcal{M}^\star$, nonsmooth min. turns into smooth constrained opt.

$$\min_{x \in \mathbb{R}^n} F(x) \quad \overset{\text{with } \mathcal{M}^\star}{\longrightarrow} \quad \min_{x \in \mathcal{M}^\star} F|_{\mathcal{M}^\star}(x)$$

- Tune splitting method parameters $\to$ linear rate ◇ Liang Fadili Peyré '17
- $\mathcal{VU}$-algorithm $\to$ superlinear steps ◇ Mifflin Sagastizábal '05
  Approximate Newton on smooth directions, approximate prox

These algorithms **exploit** knowledge of $\mathcal{M}^\star$

2. Exploiting structure

▷ With knowledge of $\mathcal{M}^\star$, nonsmooth min. turns into smooth constrained opt.

$$\min_{x \in \mathbb{R}^n} F(x) \quad \overset{\text{with } \mathcal{M}^\star}{\longrightarrow} \quad \min_{x \in \mathcal{M}^\star} F|_{\mathcal{M}^\star}(x)$$

- Tune splitting method parameters $\to$ linear rate ◇ Liang Fadili Peyré '17
- $\mathcal{VU}$-algorithm $\to$ superlinear steps ◇ Mifflin Sagastizábal '05
  Approximate Newton on smooth directions, approximate prox

These algorithms **exploit** knowledge of $\mathcal{M}^\star$

Problem: $\mathcal{M}^\star$ is not known, only approximated numerically
Except in specific cases ◇ Daniilidis Sagastizábal Solodov '09

This thesis: detecting & exploiting structure

This thesis: detecting & exploiting structure

- Smooth reduced problem

$$\min_{x \in \mathbb{R}^n} F(x) \qquad \overset{\text{with } \mathcal{M}^\star}{\longrightarrow} \qquad \min_{x \in \mathcal{M}^\star} F|_{\mathcal{M}^\star}(x)$$

If $\mathcal{M}^\star$ is known, we can use **fast Newton-type methods** relative to $\mathcal{M}^\star$

This thesis: detecting & exploiting structure

- Smooth reduced problem

$$\min_{x \in \mathbb{R}^n} F(x) \qquad \xrightarrow{\text{with } \mathcal{M}^\star} \qquad \min_{x \in \mathcal{M}^\star} F|_{\mathcal{M}^\star}(x)$$

If $\mathcal{M}^\star$ is known, we can use **fast Newton-type methods** relative to $\mathcal{M}^\star$

Challenge: $\mathcal{M}^\star$ is never known; but near point x_k, there is relevant structure $\mathcal{M}_k$

This thesis: detecting & exploiting structure

▷ Smooth reduced problem

$$\min_{x \in \mathbb{R}^n} F(x) \qquad \overset{\text{with } \mathcal{M}^\star}{\longrightarrow} \qquad \min_{x \in \mathcal{M}^\star} F|_{\mathcal{M}^\star}(x)$$

If $\mathcal{M}^\star$ is known, we can use **fast Newton-type methods** relative to $\mathcal{M}^\star$

Challenge: $\mathcal{M}^\star$ is never known; but near point x_k, there is relevant structure $\mathcal{M}_k$

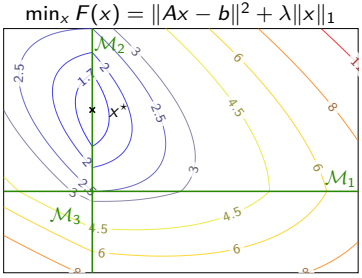
▷ We must use structure **adaptively**, which raises two difficulties:

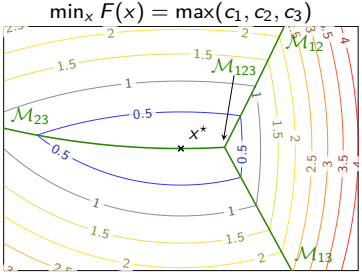
- How to *provably* **detect** relevant structure near a point x_k? → $\mathbf{prox}_{\gamma g}$
- How to **exploit** the detected structure $\mathcal{M}_k$? → Newton method

Outline

1. Additive nonsmoothness: $\min_x f(x) + g(x)$
   Iterates naturally live on $\mathcal{M}$, feasible optim. methods

2. Composite nonsmoothness: $\min_x g \circ c(x)$
   Iterates don't live on $\mathcal{M}$, infeasible optim. methods





Outline

Additive problems

$$\text{Find } x^\star \in \arg\min_{x \in \mathbb{R}^n} F(x) = \underbrace{f(x)}_{\text{smooth}} + \underbrace{g(x)}_{\text{non smooth}}$$

Example:

In inverse problems / learning, g is **chosen** to encode **prior knowledge**:

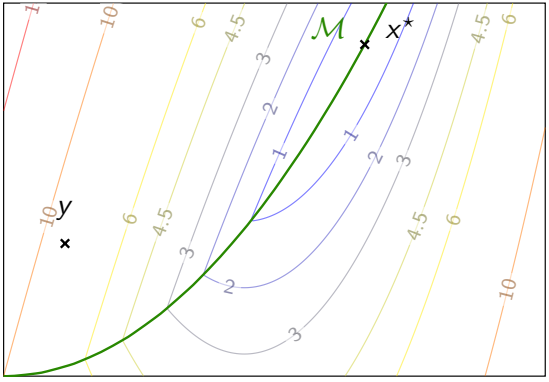
- $x^\star$ sparse $\longrightarrow g(x) \propto \|x\|_1 = \sum_{i=1}^n |x_i|$
- $X^\star$ low-rank $\longrightarrow g(X) \propto \|X\|_* = \sum_{i=1}^{\mathrm{rank}(X)} \sigma_i(X)$

◇ Candès Romberg Tao '06; Candès Recht '13; Vaiter Peyré Fadili '15

Back on the proximal gradient

- Proximal gradient algorithm: iterate

$$x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$$



$$\min_{x \in \mathbb{R}^2} F(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

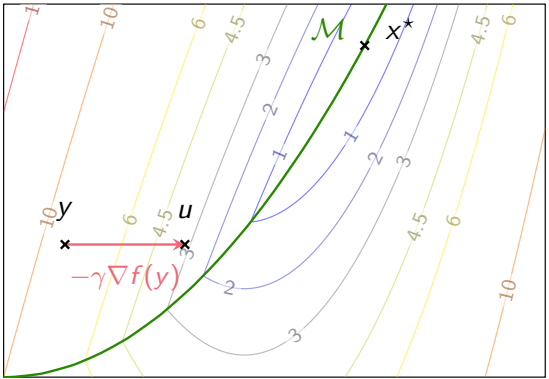
Back on the proximal gradient

- Proximal gradient algorithm: iterate

$$x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$$

1. Explicit step on f

$$u = y - \gamma \nabla f(y)$$



$$\min_{x \in \mathbb{R}^2} F(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Back on the proximal gradient

▹ Proximal gradient algorithm: iterate

$$x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$$

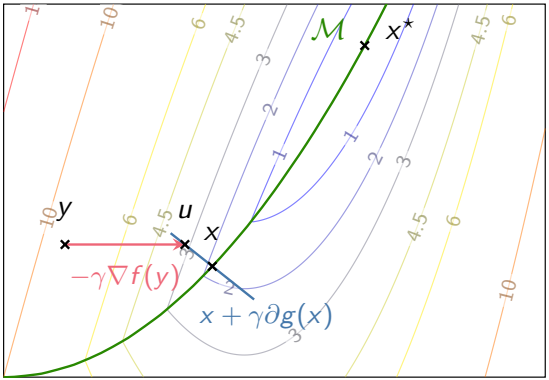
1. Explicit step on f

$$u = y - \gamma \nabla f(y)$$

2. Implicit step on g

$$x = \mathbf{prox}_{\gamma g}(u)$$
$$\Leftrightarrow \quad u \in x + \gamma \partial g(x)$$

$\rightarrow$ $\mathbf{prox}_{\gamma g}$ can send points to $\mathcal{M}$



$$\min_{x \in \mathbb{R}^2} F(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Back on the proximal gradient

▹ Proximal gradient algorithm: iterate

$$x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$$

1. Explicit step on f

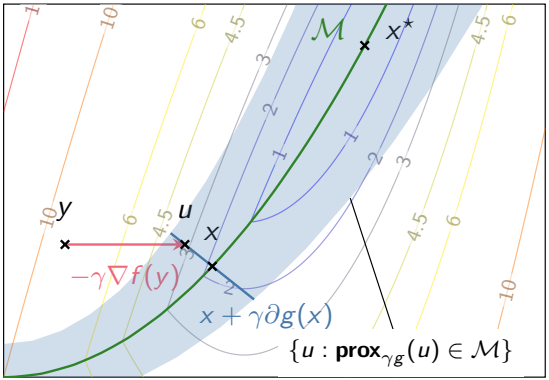
$$u = y - \gamma \nabla f(y)$$

2. Implicit step on g

$$\begin{aligned} & x = \mathbf{prox}_{\gamma g}(u) \\ \Leftrightarrow \quad & u \in x + \gamma \partial g(x) \end{aligned}$$

$\rightarrow \mathbf{prox}_{\gamma g}$ can send points to $\mathcal{M}$

More precisely...



$$\min_{x \in \mathbb{R}^2} F(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Back on the proximal gradient

▹ Proximal gradient algorithm: iterate

$$x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$$

1. Explicit step on f

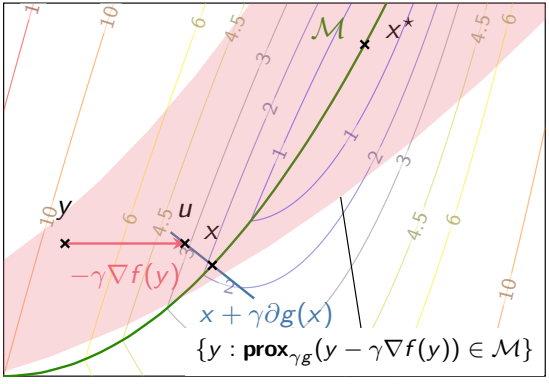
$$u = y - \gamma \nabla f(y)$$

2. Implicit step on g

$$x = \mathbf{prox}_{\gamma g}(u)$$
$$\Leftrightarrow \quad u \in x + \gamma \partial g(x)$$

$\rightarrow$ $\mathbf{prox}_{\gamma g}$ can send points to $\mathcal{M}$

More precisely...



$$\min_{x \in \mathbb{R}^2} F(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Proximal gradient: a structure detector

▷ Near manifolds, the proximal gradient sends points to the manifold, smoothly:

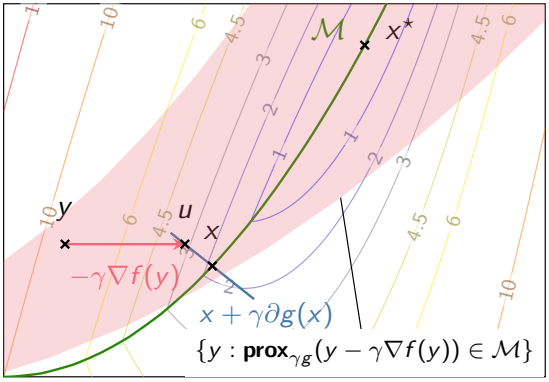
Theorem

Take y and $x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$ s.t.

- $x \in \mathcal{M}$ *partial smoothness*
- $y - \gamma \nabla f(y) - x \in \mathrm{ri}\ \gamma \partial g(x)$

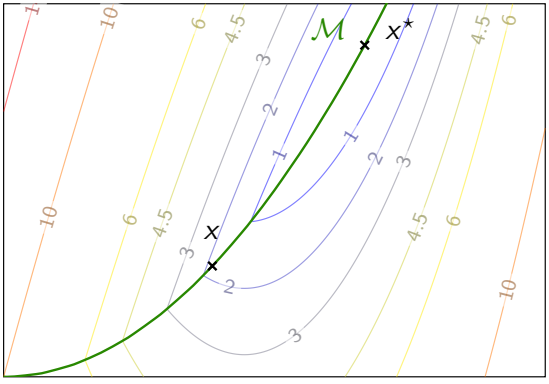
then, on a neighborhood $\mathcal{N}_y$ of y the proximal gradient is

- ***$\mathcal{M}$-valued***
- *a $\mathcal{C}^1$ operator*



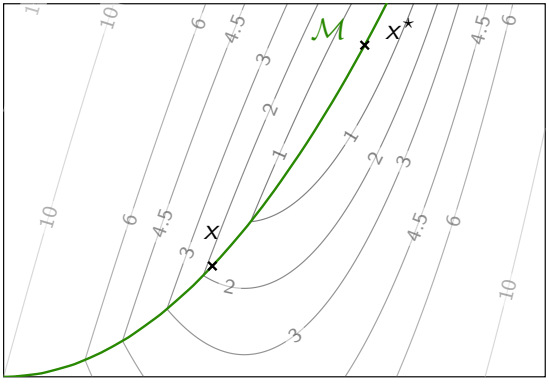
$$\min_{x \in \mathbb{R}^2} F(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Riemannian optimization, in a nutshell



$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

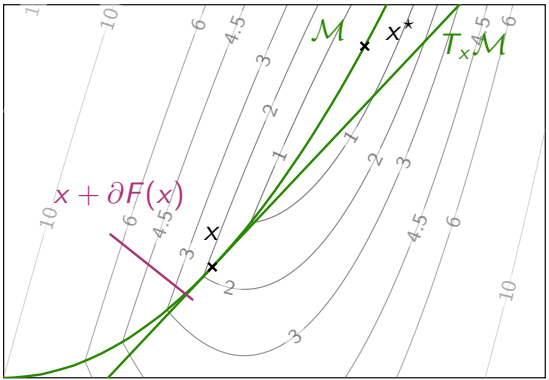
Riemannian optimization, in a nutshell



$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

Riemannian optimization, in a nutshell

Elementary tools:

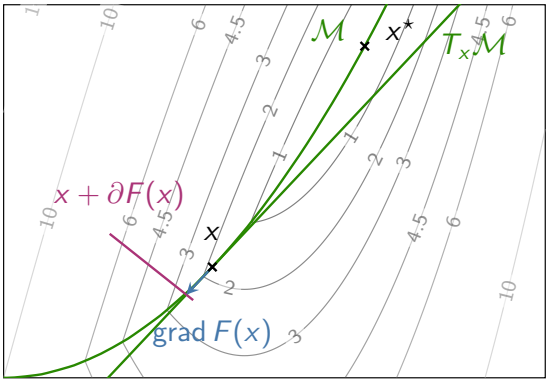


$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

Riemannian optimization, in a nutshell

Elementary tools:

- Riemannian gradient, Hessian



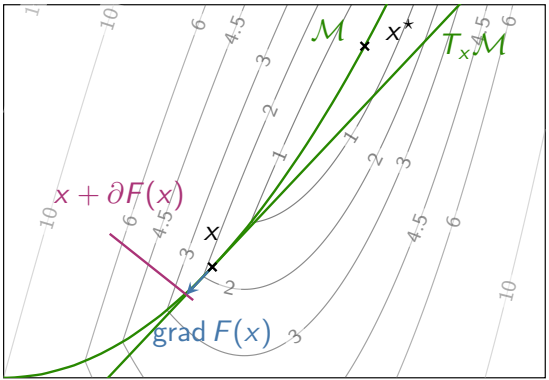
$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

Riemannian optimization, in a nutshell

Elementary tools:

- Riemannian gradient, Hessian

Typical Riemannian step



$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

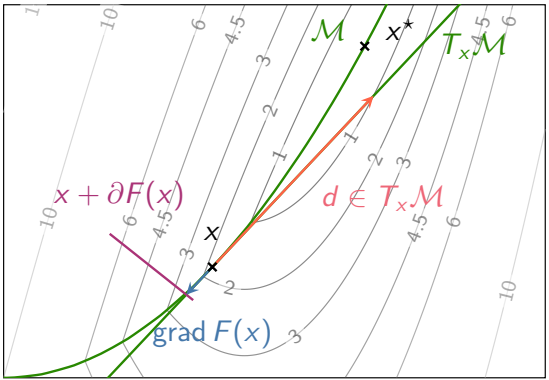
Riemannian optimization, in a nutshell

Elementary tools:

- Riemannian gradient, Hessian

Typical Riemannian step

- find $d \in T_x\mathcal{M}$ *e.g.,* $d = -\alpha \operatorname{grad} F(x)$



$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

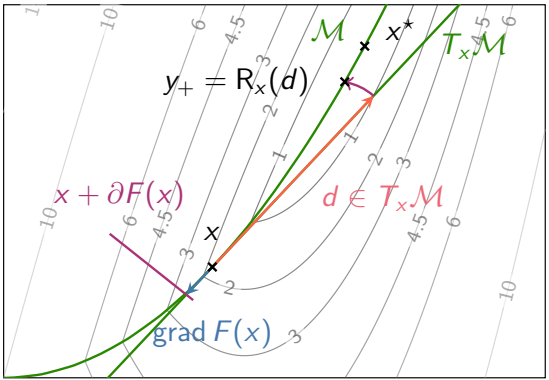
Riemannian optimization, in a nutshell

Elementary tools:

- Riemannian gradient, Hessian

Typical Riemannian step

- find $d \in T_x\mathcal{M}$ *e.g.*, $d = -\alpha \operatorname{grad} F(x)$
- send d to $\mathcal{M}$ *e.g.*, retraction $\mathrm{R}_x(d)$



$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

Riemannian optimization, in a nutshell

Elementary tools:

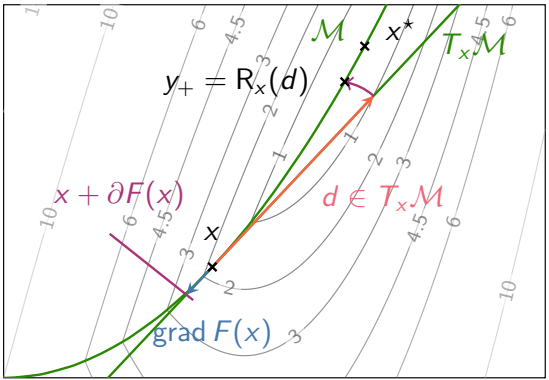
- Riemannian gradient, Hessian

Typical Riemannian step

- find $d \in T_x\mathcal{M}$ *e.g.*, $d = -\alpha \operatorname{grad} F(x)$
- send d to $\mathcal{M}$ *e.g.*, retraction $\mathrm{R}_x(d)$

$\rightarrow$ Many methods of smooth optim carry over manifold optimization.

◇ Absil *et al* '09, Boumal '22



$$\min_{x \in \mathbb{R}^2} 10(1 - x_1)^2 + 5|x_1^2 - x_2| \text{ s.t. } x \in \mathcal{M}$$

Proposed algorithm

Iteration k:

- Compute *point* $x_k = \mathbf{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$ and *manifold* $\mathcal{M}_k \ni x_k$
- Compute y_k from point x_k by a manifold update on $\mathcal{M}_k$ *e.g.*, Riemannian Newton step

Proposed algorithm

Iteration k:

- Compute *point* $x_k = \mathbf{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$ and *manifold* $\mathcal{M}_k \ni x_k$
- Compute y_k from point x_k by a manifold update on $\mathcal{M}_k$ *e.g.*, Riemannian Newton step

Theorem

- *If* $\gamma < 1/L$ *and the manifold update decreases function value, then any limit point* $\bar{x}$ *of* (x_k) *is a critical point:* $0 \in \partial F(\bar{x})$

Proposed algorithm

Iteration k:

- Compute *point* $x_k = \mathbf{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$ and *manifold* $\mathcal{M}_k \ni x_k$
- Compute y_k from point x_k by a manifold update on $\mathcal{M}_k$ *e.g.*, Riemannian Newton step

Theorem

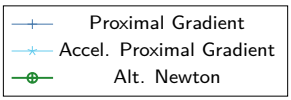
- *If $\gamma < 1/L$ and the manifold update decreases function value, then any limit point $\bar{x}$ of (x_k) is a critical point: $0 \in \partial F(\bar{x})$*
- *Assume that a Riemannian Newton method is used, and that one limit point $x^\star$ is such that*
  - *g has structure $\mathcal{M}^\star$ at $x^\star$*
  - *$0 \in \operatorname{ri} \partial F(x^\star)$, $\operatorname{Hess}_{\mathcal{M}^\star} F(x^\star) \succ 0$ and $\operatorname{Hess}_{\mathcal{M}^\star}$ is locally Lipschitz around $x^\star$*

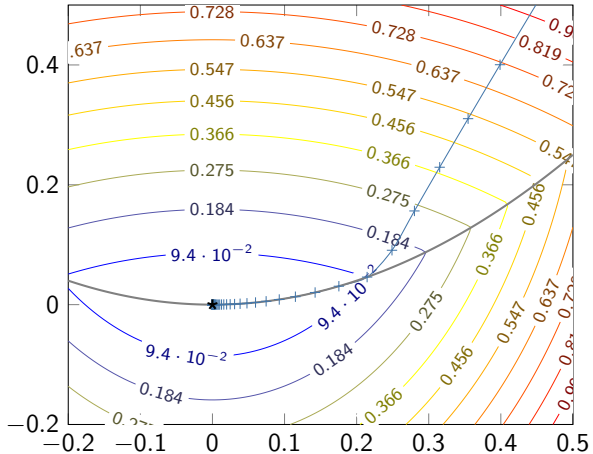
Then, after some finite time

- *$x_k \in \mathcal{M}^\star$*
- *x_k converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Illustration: toy example

$$\min_{x \in \mathbb{R}^2} \underbrace{2x_1^2 + x_2^2}_{f(x)} + \underbrace{|x_1^2 - x_2|}_{g(x)}$$

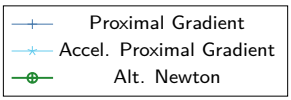


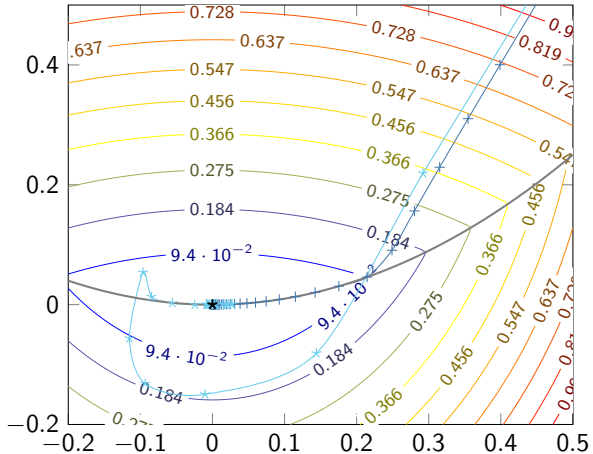


◇ Beck Teboulle '09, Chambolle Dossal '15, B Iutzeler '20

Illustration: toy example

$$\min_{x \in \mathbb{R}^2} \underbrace{2x_1^2 + x_2^2}_{f(x)} + \underbrace{|x_1^2 - x_2|}_{g(x)}$$

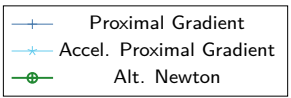


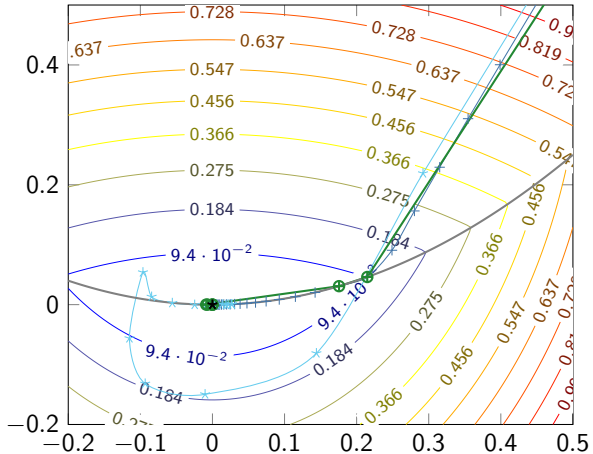


◇ Beck Teboulle '09, Chambolle Dossal '15, B Iutzeler '20

Illustration: toy example

$$\min_{x \in \mathbb{R}^2} \underbrace{2x_1^2 + x_2^2}_{f(x)} + \underbrace{|x_1^2 - x_2|}_{g(x)}$$





◇ Beck Teboulle '09, Chambolle Dossal '15, B Iutzeler '20

Illustration 1: logistic regression

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \log(y_i \sigma(\langle A_i, x \rangle)) + \lambda \|x\|_1$$

In this instance, $n = m = 4000$

$$\dim(\mathcal{M}^\star) = 249$$

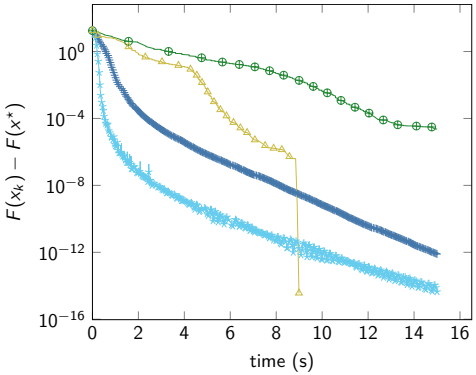
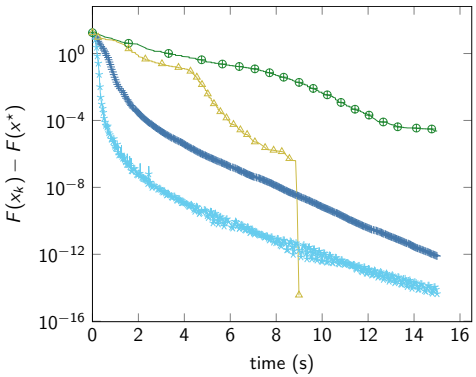


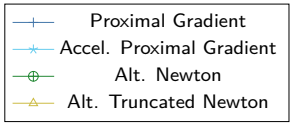
Illustration 1: logistic regression

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \log(y_i \sigma(\langle A_i, x \rangle)) + \lambda \|x\|_1$$

In this instance, $n = m = 4000$

$$\dim(\mathcal{M}^\star) = 249$$





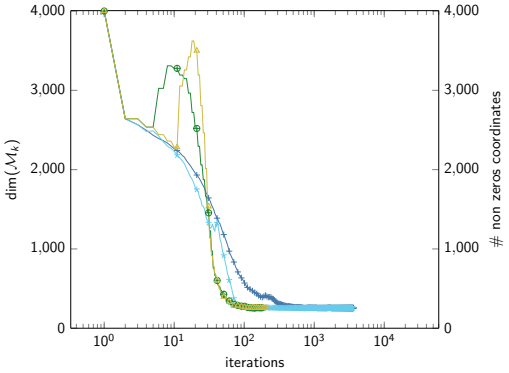
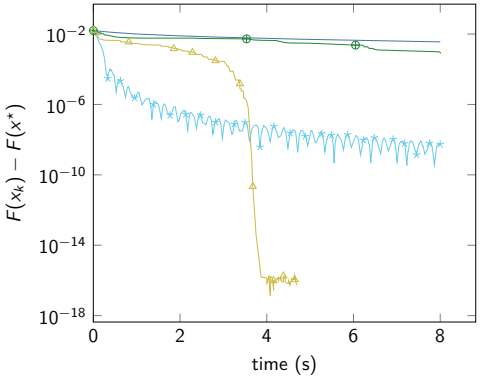


Illustration 2: tracenorm regression

$$\min_{X \in \mathbb{R}^{10\times 12}} \sum_{i=1}^{m} \left(\langle A_i, X\rangle - y_i\right)^2 + \lambda \|X\|_*$$

In this instance, $m = 60$

$$\mathcal{M}^\star = \{X : \operatorname{rank}(X) = 6\}$$



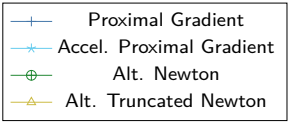
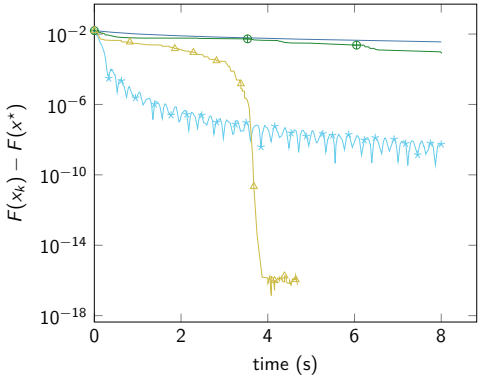


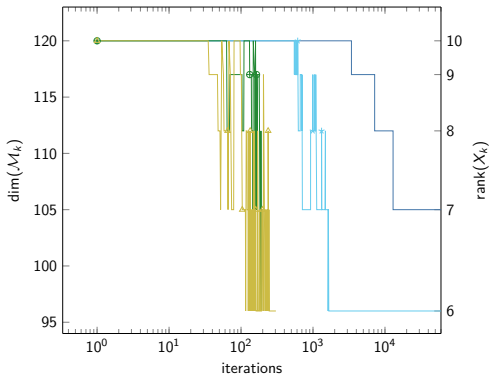
Illustration 2: tracenorm regression

$$\min_{X \in \mathbb{R}^{10\times 12}} \sum_{i=1}^{m} \left(\langle A_i, X\rangle - y_i\right)^2 + \lambda \|X\|_*$$

In this instance, $m = 60$

$$\mathcal{M}^\star = \{X : \operatorname{rank}(X) = 6\}$$

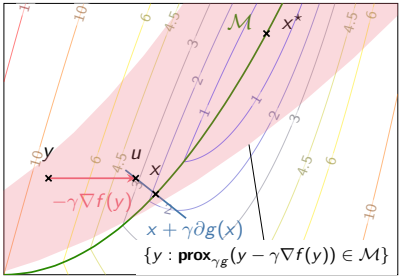




Summary on additive problems

Main messages:

- The proximal-gradient **identifies structure**
- With Riemannian Newton steps, we propose a global algorithm that
  - identifies $\mathcal{M}^\star$
  - converges locally quadratically
- We observe these results numerically



Proximal gradient identification

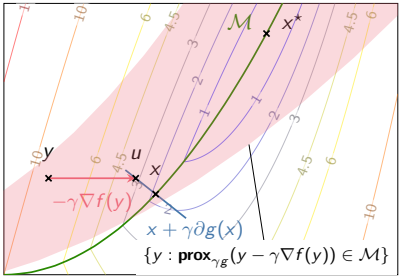
Summary on additive problems

Main messages:

- The proximal-gradient **identifies structure**
- With Riemannian Newton steps, we propose a global algorithm that
  - identifies $\mathcal{M}^\star$
  - converges locally quadratically
- We observe these results numerically

Extensions: This is the simplest scheme

- other Riemannian method cubic regularization of Newton
- other interleaving



Proximal gradient identification

Outline

Composite problem

$$\text{Find } x^\star \in \arg\min_{x \in \mathbb{R}^n} F(x) = g \circ c(x), \text{ with } g \text{ nonsmooth and } c \text{ a smooth mapping}$$

Example:

- Maximum of smooth functions: $F(x) = \max(c_1(x), \ldots, c_m(x))$ ◇ Womersley '86
- Spectral maximum: $F(x) = \lambda_{\max}(c(x))$, where $c(x) \in \mathbb{S}_m$ ◇ Oustry '00, Noll '05

Again, nonsmoothness is **explicit**: $\mathbf{prox}_{\gamma g}$ is available

Composite problem

$$\text{Find } x^\star \in \arg\min_{x \in \mathbb{R}^n} F(x) = g \circ c(x), \text{ with } g \text{ nonsmooth and } c \text{ a smooth mapping}$$

Example:

- Maximum of smooth functions: $F(x) = \max(c_1(x), \ldots, c_m(x))$ ◇ Womersley '86
- Spectral maximum: $F(x) = \lambda_{\max}(c(x))$, where $c(x) \in \mathbb{S}_m$ ◇ Oustry '00, Noll '05

Again, nonsmoothness is **explicit**: $\mathbf{prox}_{\gamma g}$ is available

▷ Structure detection: iterates of optimization methods **do not belong to manifolds** anymore

- Composite bundle ◇ Sagastizábal '13
- proxlinear methods ◇ Lewis Wright '16

→ What structure near point x?

Back to the prox

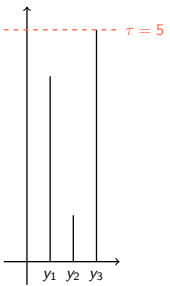
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i: y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



$y = (4, 1, 5)$

$\gamma = 0$

$\mathbf{prox}_{\gamma \max}(y) = (4, 1, 5)$

Back to the prox

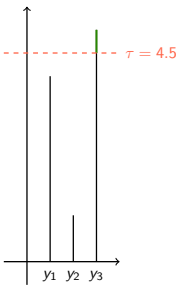
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i: y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



$y = (4, 1, 5)$

$\gamma = 0.5$

$\mathbf{prox}_{\gamma \max}(y) = (4, 1, 4.5)$

Back to the prox

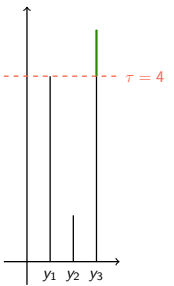
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i: y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



$y = (4, 1, 5)$

$\gamma = 1$

$\mathbf{prox}_{\gamma \max}(y) = (4, 1, 4)$

Back to the prox

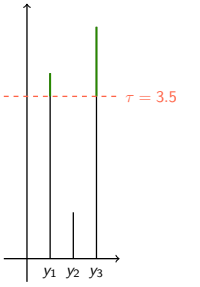
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i: y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



$y = (4, 1, 5)$

$\gamma = 2$

$\mathbf{prox}_{\gamma \max}(y) = (3.5, 1, 3.5)$

Back to the prox

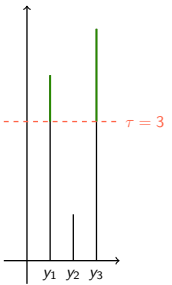
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i: y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



$y = (4, 1, 5)$

$\gamma = 3$

$\mathbf{prox}_{\gamma \max}(y) = (3, 1, 3)$

Back to the prox

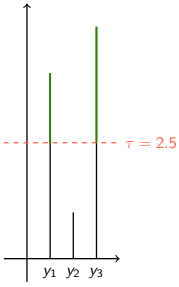
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i:y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



$y = (4, 1, 5)$

$\gamma = 4$

$\mathbf{prox}_{\gamma \max}(y) = (2.5, 1, 2.5)$

Back to the prox

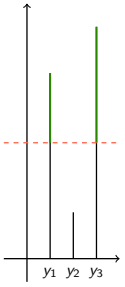
$$\mathbf{prox}_{\gamma g}(y) = \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}$$

Example: Prox of max function

$$[\mathbf{prox}_{\gamma \max}(y)]_i = \begin{cases} \tau & \text{if } y_i \geq \tau \\ y_i & \text{else} \end{cases}$$

where τ solves $\sum_{\{i: y_i > \tau\}} (y_i - \tau) = \gamma$

$$\mathcal{M}_I = \{y : y_i = \max(y) \text{ for } i \in I\}$$



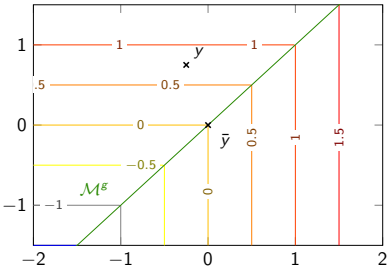
$y = (4, 1, 5)$

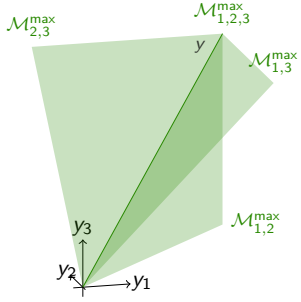
$\gamma = 4$

$\mathbf{prox}_{\gamma \max}(y) = (2.5, 1, 2.5)$

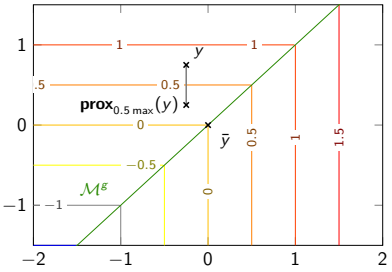
$\rightarrow$ Computing $\mathbf{prox}_{\gamma g}(y)$ also gives **structure information** $\mathcal{M} \ni \mathbf{prox}_{\gamma g}(y)$.

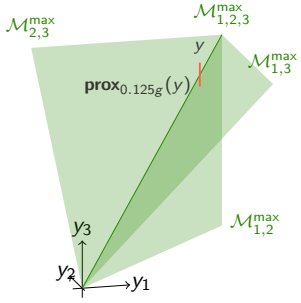
Identification for g, explicit $\mathrm{prox}_{\gamma g}$



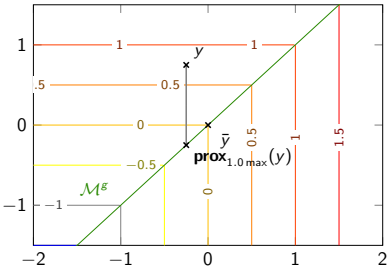


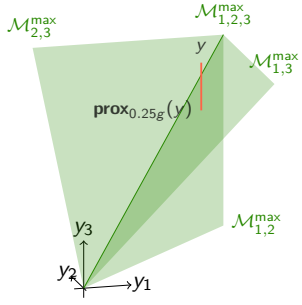
Identification for g, explicit $\mathrm{prox}_{\gamma g}$



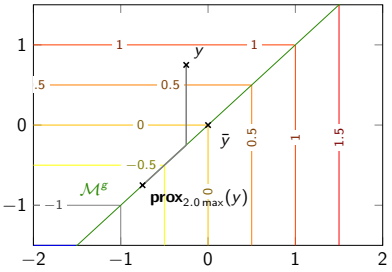


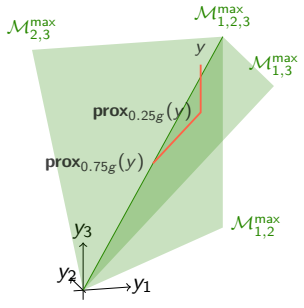
Identification for g, explicit $\mathrm{prox}_{\gamma g}$



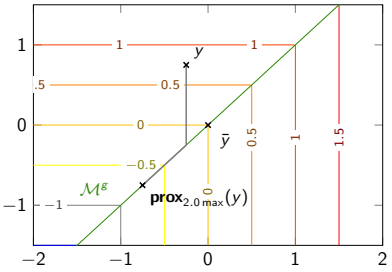


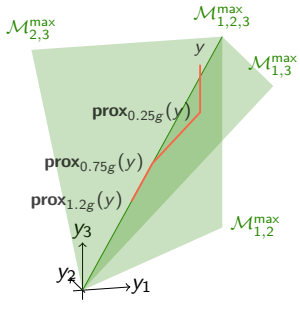
Identification for g, explicit $\mathrm{prox}_{\gamma g}$





Identification for g, explicit $\mathrm{prox}_{\gamma g}$





Identification for g, explicit $\mathbf{prox}_{\gamma g}$

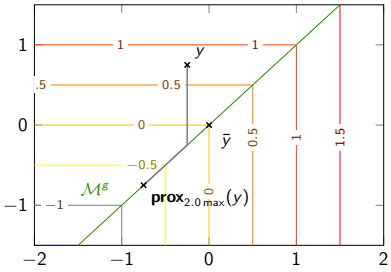
Lemma

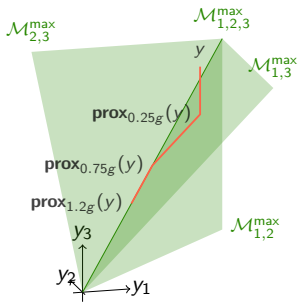
Consider a function g and point $\bar{y}$ with structure $\mathcal{M}^g$ that meet two technical assumptions. For all y near $\bar{y}$,

$$\mathbf{prox}_{\gamma g}(y) \in \mathcal{M}^g \qquad \text{for all } \gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$$

where $\Gamma^g > 0$ and $\varphi^g(t) = \frac{1}{c_{ri}} t + \mathcal{O}(t^2)$.

Technical assumptions: normal ascent, control on projection curves





Identification for g, explicit $\mathbf{prox}_{\gamma g}$

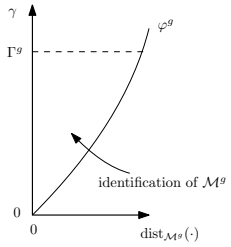
Lemma

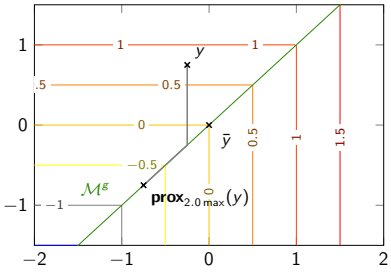
Consider a function g and point $\bar{y}$ with structure $\mathcal{M}^g$ that meet two technical assumptions. For all y near $\bar{y}$,

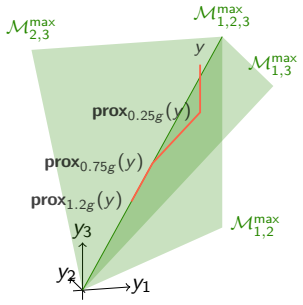
$$\mathbf{prox}_{\gamma g}(y) \in \mathcal{M}^g \qquad \text{for all } \gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$$

where $\Gamma^g > 0$ and $\varphi^g(t) = \frac{1}{c_{ri}} t + \mathcal{O}(t^2)$.

Technical assumptions: normal ascent, control on projection curves

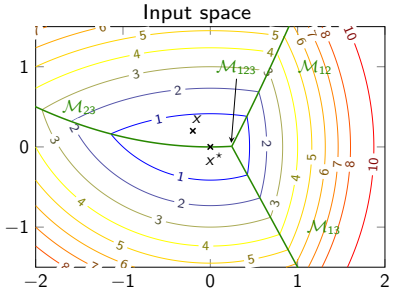




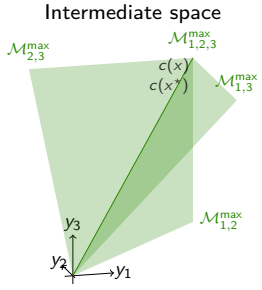


Identification for $g \circ c$, no $\text{prox}_{\gamma g \circ c}$

The prox of $F = g \circ c$ is **not available**, but **we do have** $\text{prox}_{\gamma g}$.



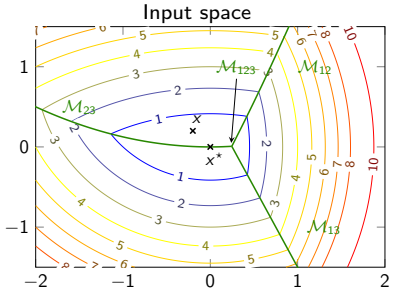
$F(x) = \max(c_1(x), c_2(x), c_3(x))$



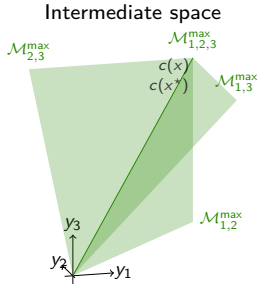
$g(y) = \max(y_1, y_2, y_3)$

Identification for $g \circ c$, no $\mathbf{prox}_{\gamma g \circ c}$

The prox of $F = g \circ c$ is **not available**, but **we do have** $\mathbf{prox}_{\gamma g}$.



$F(x) = \max(c_1(x), c_2(x), c_3(x))$



$g(y) = \max(y_1, y_2, y_3)$

Observation: $\mathbf{prox}_{\gamma g}$ can map points to $\mathcal{M}^g$.
The structure naturally lies in the intermediate space ... but that's ok!

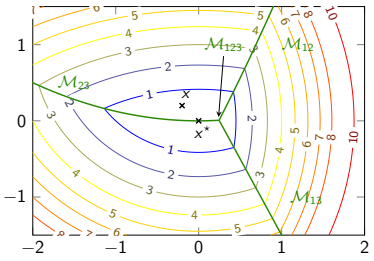
Identification for $g \circ c$

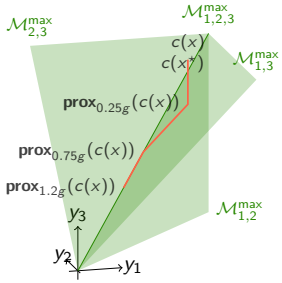
Theorem

Consider g, c smooth and a point $\bar{x}$ such that $c(\bar{x})$ has structure manifold $\mathcal{M}^g$ with c and $\mathcal{M}^g$ are transversal at $c(\bar{x})$. For all x near $\bar{x}$,

$$\mathbf{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g \quad \text{for all } \gamma \in [\varphi(\mathrm{dist}_{\mathcal{M}}(x)), \Gamma]$$

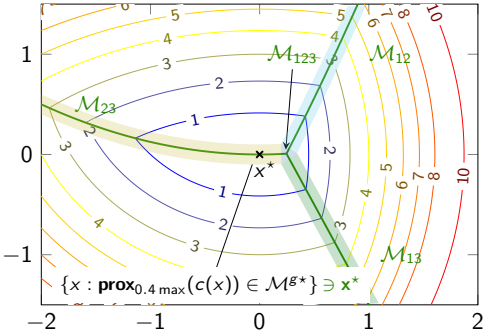
where $\Gamma > 0$ and $\varphi(t) = \frac{c_{map}}{c_{ri}} t + \mathcal{O}(t^2)$. Furthermore, $\mathcal{M} = c^{-1}(\mathcal{M}^g)$.





$[\varphi(\mathrm{dist}_{\mathcal{M}}(x)), \Gamma] \subset [0.25, 0.75]$

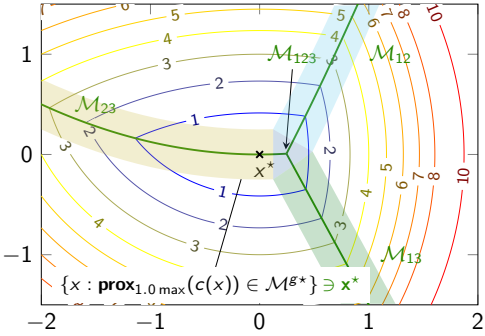
Towards an algorithm: i. detecting structure



γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

$$\mathcal{M}^\star = \mathcal{M}_{23}$$

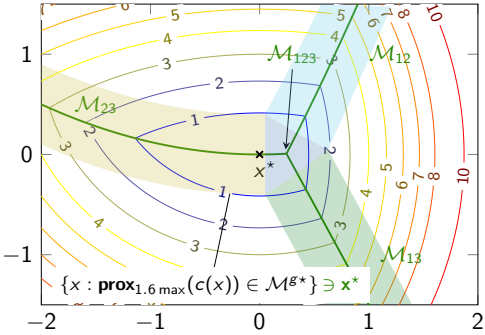
Towards an algorithm: i. detecting structure



γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

$$\mathcal{M}^\star = \mathcal{M}_{23}$$

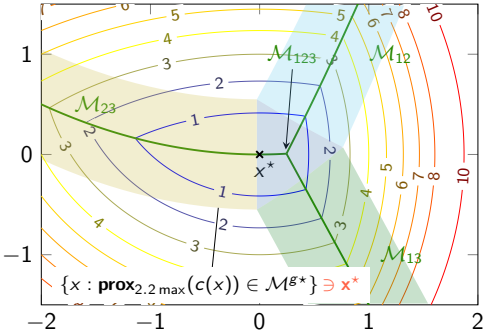
Towards an algorithm: i. detecting structure



γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

$$\mathcal{M}^\star = \mathcal{M}_{23}$$

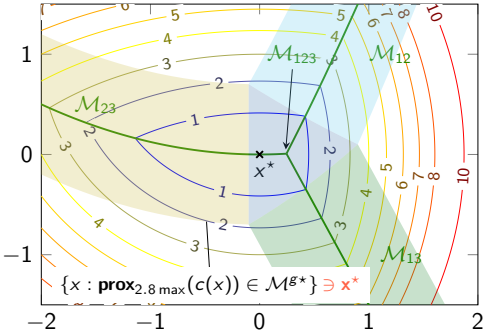
Towards an algorithm: i. detecting structure



γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

$$\mathcal{M}^\star = \mathcal{M}_{23}$$

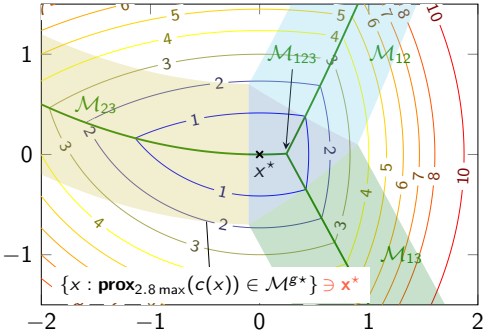
Towards an algorithm: i. detecting structure



γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

$$\mathcal{M}^\star = \mathcal{M}_{23}$$

Towards an algorithm: i. detecting structure

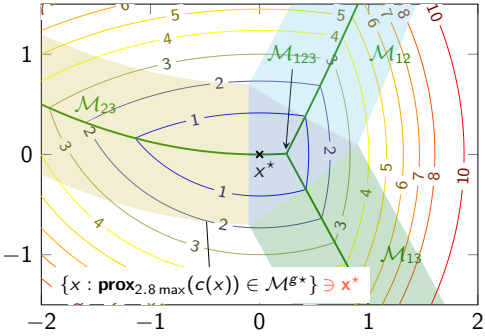


$\mathcal{M}^\star = \mathcal{M}_{23}$

γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

γ too large $\Rightarrow$ no detection of $\mathcal{M}^\star$ near $x^\star$.

Towards an algorithm: i. detecting structure



γ too small $\Rightarrow$ detection of $\mathcal{M}^\star$ only near $x^\star$;

γ too large $\Rightarrow$ no detection of $\mathcal{M}^\star$ near $x^\star$.

$$\mathcal{M}^\star = \mathcal{M}_{23}$$

Bottom line: $\mathbf{prox}_{\gamma g} \circ c(\cdot)$ detects $\mathcal{M}^\star$ from any x near $x^\star$ when $\gamma \in [\frac{c_{map}}{c_{ri}} \operatorname{dist}_{\mathcal{M}}(x), \Gamma]$.

$\rightarrow$ How to choose the step in practice?

Towards an algorithm: ii. exploiting structure

$$\begin{array}{lll} \mathcal{M} \text{ is smooth} & \exists h \text{ smooth s.t.} & x \in \mathcal{M} \Leftrightarrow h(x) = 0 \\ F \text{ smooth on } \mathcal{M} & \exists \widetilde{F} \text{ smooth s.t.} & F|_{\mathcal{M}} \equiv \widetilde{F} \text{ on } \mathcal{M} \end{array}$$

$$\min_{x \in \mathbb{R}^n} F(x) \quad \xrightarrow{\text{with } \mathcal{M}} \quad \min_{x} \widetilde{F}(x) \text{ s.t. } h(x) = 0.$$

Towards an algorithm: ii. exploiting structure

$$\begin{array}{lll} \mathcal{M} \text{ is smooth} & \exists h \text{ smooth s.t.} & x \in \mathcal{M} \Leftrightarrow h(x) = 0 \\ F \text{ smooth on } \mathcal{M} & \exists \widetilde{F} \text{ smooth s.t.} & F|_{\mathcal{M}} \equiv \widetilde{F} \text{ on } \mathcal{M} \end{array}$$

$$\min_{x \in \mathbb{R}^n} F(x) \quad \xrightarrow{\text{with } \mathcal{M}} \quad \min_x \widetilde{F}(x) \text{ s.t. } h(x) = 0.$$

Example ($F = \max(c_1, c_2, c_3)$)

For structure $\mathcal{M}_{12}$,

- $h = c_1 - c_2$
- $\widetilde{F}(x) = (c_1 + c_2)/2$

Towards an algorithm: ii. exploiting structure

$$\begin{array}{ll} \mathcal{M} \text{ is smooth} & \exists h \text{ smooth s.t. } \quad x \in \mathcal{M} \Leftrightarrow h(x) = 0 \\ F \text{ smooth on } \mathcal{M} & \exists \widetilde{F} \text{ smooth s.t. } \quad F|_{\mathcal{M}} \equiv \widetilde{F} \text{ on } \mathcal{M} \end{array}$$

$$\min_{x \in \mathbb{R}^n} F(x) \quad \xrightarrow{\text{with } \mathcal{M}} \quad \min_{x} \widetilde{F}(x) \text{ s.t. } h(x) = 0.$$

Example ($F = \max(c_1, c_2, c_3)$)

For structure $\mathcal{M}_{12}$,

- $h = c_1 - c_2$
- $\widetilde{F}(x) = (c_1 + c_2)/2$

▹ Many tools for smooth constrained optimization: Interior Point Methods, **Sequential Quadratic Programming**, Augmented Lagrangian Methods, *etc*

Algorithm: detection & exploitation

Iteration k:

Algorithm: detection & exploitation

Iteration k:

- Compute $\mathbf{prox}_{\gamma_k g}(c(x_k))$ and obtain $\mathcal{M}_k$ and h_k, $\widetilde{F}_k$

Algorithm: detection & exploitation

Iteration k:

- Compute $\mathbf{prox}_{\gamma_k g}(c(x_k))$ and obtain $\mathcal{M}_k$ and h_k, $\widetilde{F}_k$
- **Newton-SQP step** relative to $\mathcal{M}_k$:

$$d_k^{\mathrm{SQP}}(x_k) = \underset{d \in \mathbb{R}^n}{\arg\min} \quad \langle \nabla \widetilde{F}_k(x_k), d \rangle + \frac{1}{2} \langle \nabla^2_{xx} L_k(x_k, \lambda_k(x_k)) d, d \rangle$$
$$\text{s.t.} \quad h_k(x_k) + \mathrm{D}\, h_k(x_k) d = 0$$

where $L_k(x, \lambda) = \widetilde{F}_k(x) + \langle \lambda, h_k(x) \rangle$, and $\lambda_k(x_k) = \arg\min_{\lambda \in \mathbb{R}^r} \left\| \nabla \widetilde{F}_k(x_k) + \sum_{i=1}^m \lambda_i \nabla h_{k,i}(x_k) \right\|^2$

Set $x_{k+1} = x_k + d_k^{\mathrm{SQP}}(x_k)$ if $F(x_k + d_k^{\mathrm{SQP}}(x_k)) < F(x_k)$

Algorithm: detection & exploitation

Iteration k:

▷ Compute $\mathbf{prox}_{\gamma_k g}(c(x_k))$ and obtain $\mathcal{M}_k$ and h_k, $\widetilde{F}_k$

▷ **Newton-SQP step** relative to $\mathcal{M}_k$:

$$d_k^{\mathrm{SQP}}(x_k) = \operatorname*{arg\,min}_{d \in \mathbb{R}^n} \quad \langle \nabla \widetilde{F}_k(x_k), d \rangle + \frac{1}{2} \langle \nabla^2_{xx} L_k(x_k, \lambda_k(x_k)) d, d \rangle$$
$$\text{s.t.} \quad h_k(x_k) + \mathrm{D}\, h_k(x_k) d = 0$$

where $L_k(x, \lambda) = \widetilde{F}_k(x) + \langle \lambda, h_k(x) \rangle$, and $\lambda_k(x_k) = \arg\min_{\lambda \in \mathbb{R}^r} \left\| \nabla \widetilde{F}_k(x_k) + \sum_{i=1}^m \lambda_i \nabla h_{k,i}(x_k) \right\|^2$

Set $x_{k+1} = x_k + d_k^{\mathrm{SQP}}(x_k)$ if $F(x_k + d_k^{\mathrm{SQP}}(x_k)) < F(x_k)$

▷ $\gamma_{k+1} = \frac{\gamma_k}{2}$

Local exact structure identification and quadratic convergence

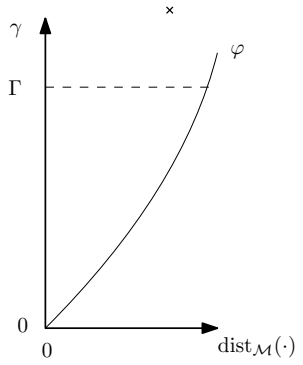
Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions

Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:



Local exact structure identification and quadratic convergence

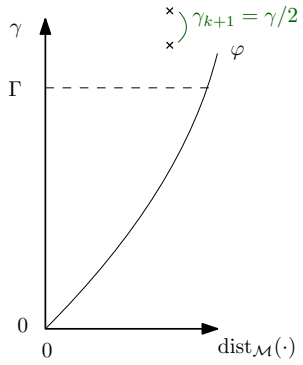
Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions

Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:



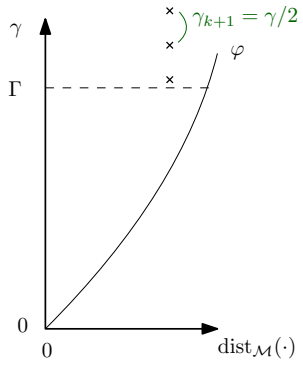
Local exact structure identification and quadratic convergence

Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions
Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:



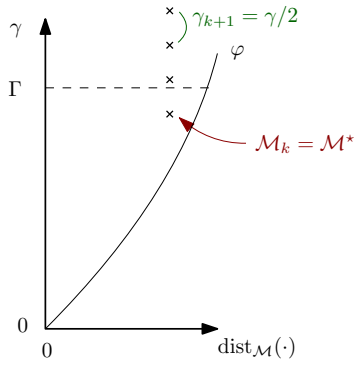
Local exact structure identification and quadratic convergence

Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions
Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:



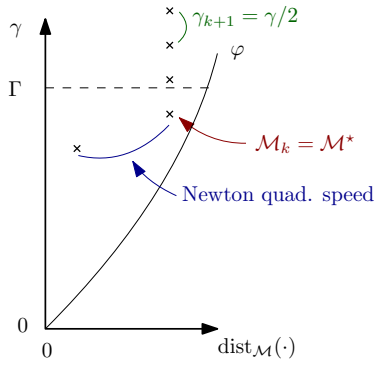
Local exact structure identification and quadratic convergence

Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions
Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:



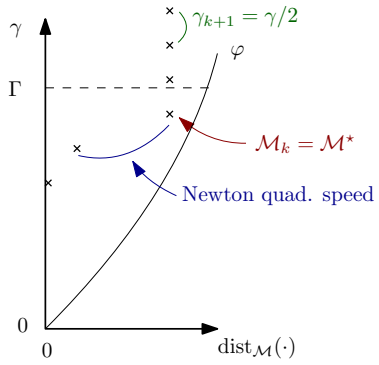
Local exact structure identification and quadratic convergence

Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions
Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:



Local exact structure identification and quadratic convergence

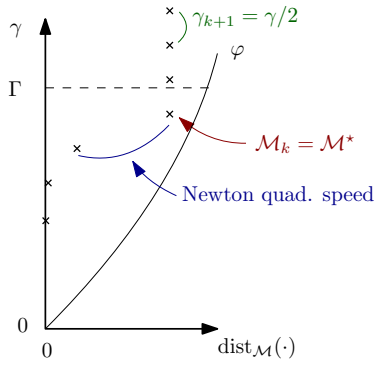
Theorem

Consider $F = g \circ c$ and a minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions

Then, if x_0 is close enough to $x^\star$ and γ_0 is large enough, after some finite time

- $\mathcal{M}_k = \mathcal{M}^\star$
- x_k *converges to $x^\star$ at a* ***quadratic rate****:* $\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

Sketch of proof, on a picture:

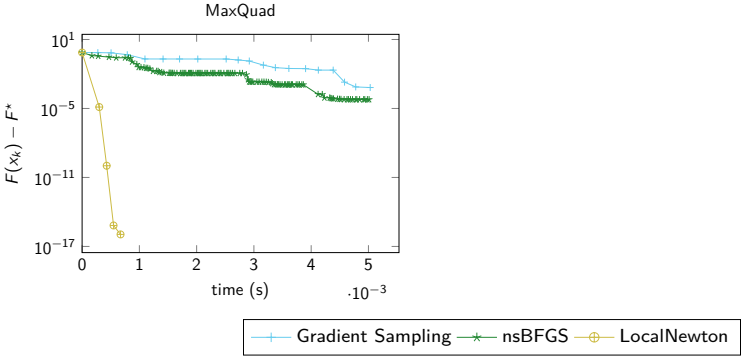


Illustrations: local behavior

$$\min_{x \in \mathbb{R}^{10}} \max_{i=1,\cdots,5} (c_i(x))$$

In this historical instance ◇ HULL '93

$$\mathcal{M}^\star = \{x : c_2(x) = \ldots = c_5(x)\}$$

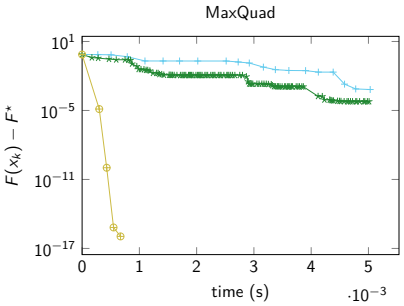


Illustrations: local behavior

$$\min_{x \in \mathbb{R}^{10}} \max_{i=1,\cdots,5} (c_i(x))$$

In this historical instance ◇ HULL '93

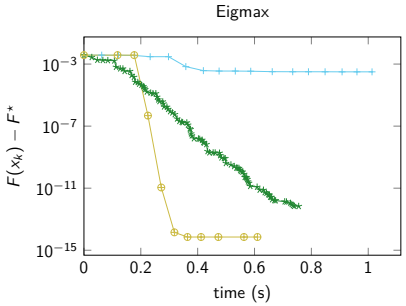
$$\mathcal{M}^\star = \{x : c_2(x) = \ldots = c_5(x)\}$$



$$\min_{x \in \mathbb{R}^n} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right)$$

In this instance, $n = 25$, $A_i \in \mathbb{S}_{50}$

$$\mathcal{M}^\star = \{x : \lambda_{\max}(c(x)) \text{ has multiplicity } 3\}$$



Gradient Sampling — nsBFGS — LocalNewton

So far, on composite problems

Main messages:

- The proximal operator **identifies structure** in composite problems
- With Newton SQP steps, we propose a algorithm that **locally**
  - identifies $\mathcal{M}^\star$
  - converges quadratically
- We observe these results numerically

So far, on composite problems

Main messages:

- The proximal operator **identifies structure** in composite problems
- With Newton SQP steps, we propose a algorithm that **locally**
  - identifies $\mathcal{M}^\star$
  - converges quadratically
- We observe these results numerically

So far, on composite problems

Main messages:

- The proximal operator **identifies structure** in composite problems
- With Newton SQP steps, we propose a algorithm that **locally**
  - identifies $\mathcal{M}^\star$
  - converges quadratically
- We observe these results numerically

Do we really need to start near $x^\star$?

Towards a global nonsmooth Newton method

▷ **Difficulty # 1:** Newton's direction may fail to provide functional descent.

→ linesearch: find α such that $F(x + \alpha d) \leq F(x) + \alpha m F'(x; d)$.

- Is d a descent direction $F'(x; d) < 0$?
- The linesearch may jeopardizes the final quadratic rate? Maratos effect, *etc*

Towards a global nonsmooth Newton method

▷ **Difficulty # 1:** Newton's direction may fail to provide functional descent.

$\rightarrow$ linesearch: find α such that $F(x + \alpha d) \leq F(x) + \alpha m F'(x; d)$.

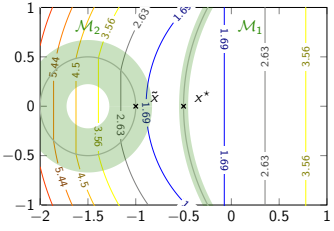
- Is d a descent direction $F'(x; d) < 0$?
- The linesearch may jeopardizes the final quadratic rate? Maratos effect, *etc*

▷ **Difficulty # 2:** Detection of relevant structure, far from $x^\star$

Example In green, structure detection of $\mathbf{prox}_{\gamma g} \circ c$

Point $\tilde{x}$

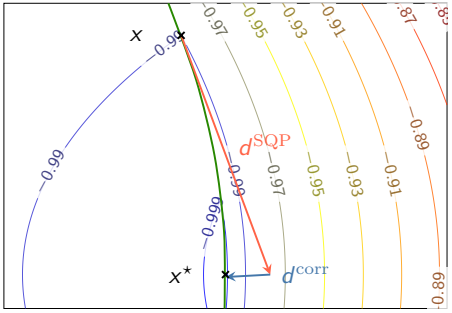
- is not a minimizer
- but stable structure detection by $\mathbf{prox}_{\gamma g} \circ c$.



No Maratos effect with correction

Maratos effect: $F(x + d^{\mathrm{SQP}}) > F(x)$
$\rightarrow$ Newton direction is not accepted

Second-order correction: from $x + d^{\mathrm{SQP}}$, go towards $\mathcal{M}$ Newton-Raphson step



Theorem (Local admissibility of unit stepsize)

Consider $F = g \circ c$ and a strong minimizer $x^\star \in \mathcal{M}^\star$ that meet some assumptions
Then if x is near $x^\star$ and $m \in (0, 1/2)$,

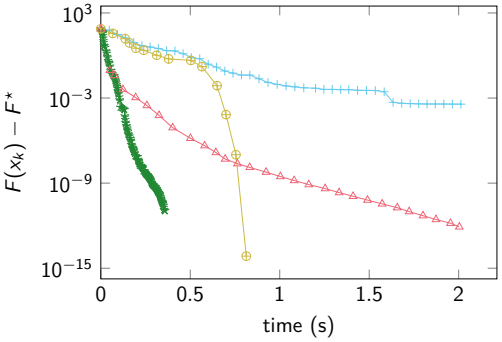
$$F\left(x + d^{\mathrm{SQP}} + d^{\mathrm{corr}}\right) \leq F(x) + mF'(x; d^{\mathrm{SQP}}).$$

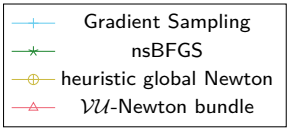
Heuristic algorithm illustrated

$$\min_{x \in \mathbb{R}^n} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right)$$

In this instance, $n = 25$, $A_i \in \mathbb{S}_{50}$

$$\mathcal{M}^\star = \{x : \lambda_{\max}(c(x)) \text{ has multiplicity 3}\}$$





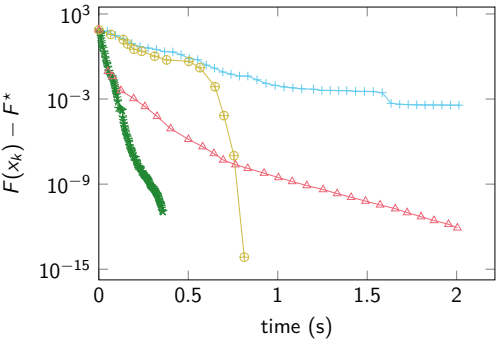
(our implementation of ◇ Mifflin Sagastizábal '05)

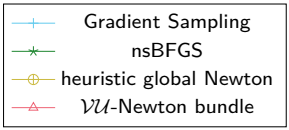
Heuristic algorithm illustrated

$$\min_{x \in \mathbb{R}^n} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right)$$

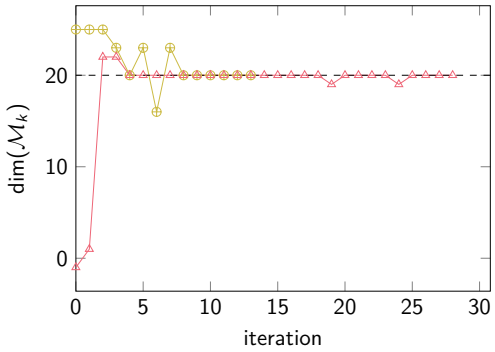
In this instance, $n = 25$, $A_i \in \mathbb{S}_{50}$

$$\mathcal{M}^\star = \{x : \lambda_{\max}(c(x)) \text{ has multiplicity 3}\}$$





(our implementation of ◇ Mifflin Sagastizábal '05)

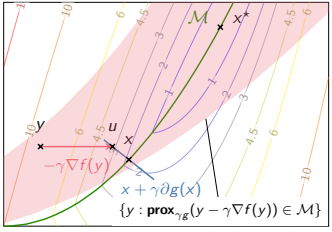


Outline

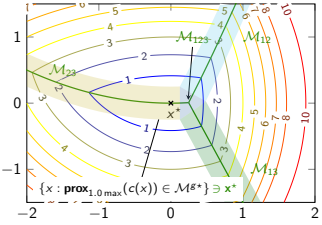
Three years, Three years, in one slide

We have made some steps towards

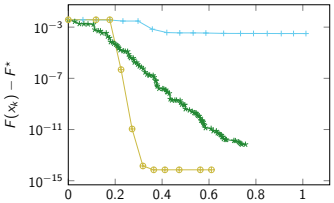
detecting and **exploiting** structure in nonsmooth optimization



Proximal gradient identification



Proximal identification



Quadratic convergence
on nonsmooth functions

Other aspects

Reproducibility

All experiments have online open source implementation in Julia

`gbareilles.fr/software`

Not discussed but important & interesting

- *active-set* solver on spectraplex
- implementation of $\mathcal{VU}$-bundle algorithm
- second derivatives of $\|\cdot\|_*$ and $\lambda_{\max}$ near nonsmooth points!
- nonconvex setting

Perspectives

Extensions

- Globalize local Newton method for $g \circ c$ – work in progress
- Smarter second-order updates: large-scale settings, efficient even away from minimizers
  $\rightarrow$ Investigate quasi-Newton (BFGS), cubic regularization of Newton
- Benchmarking: further numerical comparison $\diamond$ Massias *et al* '22

Perspectives

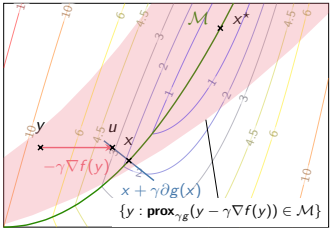
Extensions

- Globalize local Newton method for $g \circ c$ – work in progress
- Smarter second-order updates: large-scale settings, efficient even away from minimizers
  $\rightarrow$ Investigate quasi-Newton (BFGS), cubic regularization of Newton
- Benchmarking: further numerical comparison $\diamond$ Massias *et al* '22

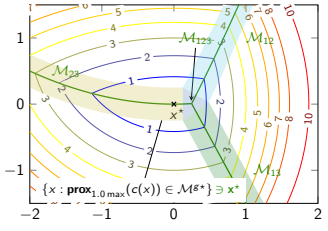
Applications

- Richer additive structures $f + g + h(L\cdot)$, write as $g \circ c$ with g prox-simple
  $\rightarrow$ Apply Local Newton Algorithm
- Smooth minimization on structured sets: $\min_{x \in \mathcal{S}} f(x) \rightarrow \min_x f(x) + \imath_{\mathcal{S}}(x)$
  $\rightarrow$ Newton acceleration of projected gradient

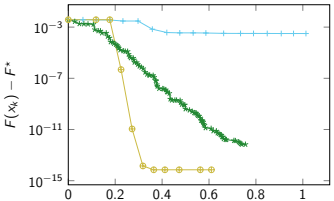
Thank you for your *presence* and *attention*!



Proximal gradient identification



Proximal identification



Quadratic rates
on nonsmooth functions