Newton methods for nonsmooth composite optimization

Gilles Bareilles

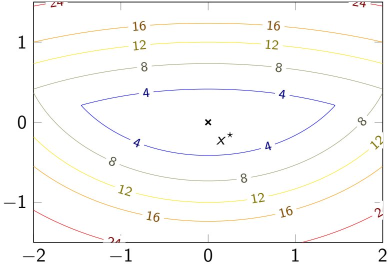
LJK, Univ. Grenoble Alpes

Journées SMAI MODE 2022
June 1-3, 2022

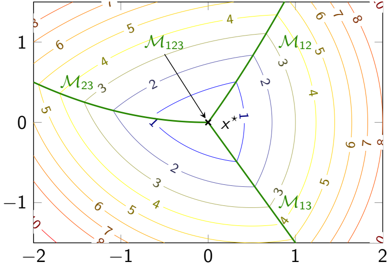
Composite optimization: $F(x) = g(c(x))$

Includes: max. of $\mathcal{C}^2$ functions, max. eigenvalue

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



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



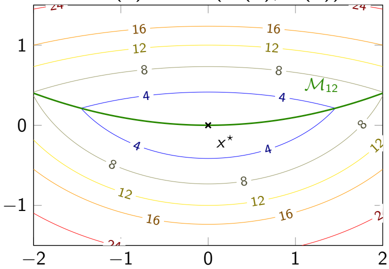
Composite optimization: $F(x) = g(c(x))$

Includes: max. of $\mathcal{C}^2$ functions, max. eigenvalue

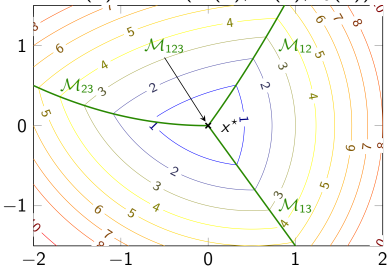
Observations

- nondifferentiability points organize in smooth manifolds
- F is smooth on them

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



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



Composite optimization: $F(x) = g(c(x))$

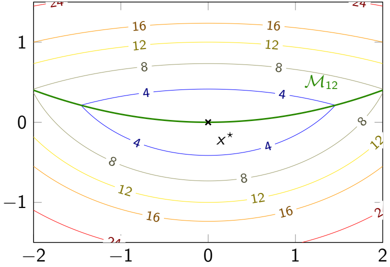
Includes: max. of $\mathcal{C}^2$ functions, max. eigenvalue

Observations

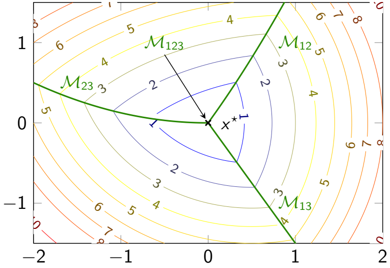
- nondifferentiability points organize in smooth manifolds
- F is smooth on them

These are ***structure manifolds***. ◇ Lewis '02

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



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



Composite optimization: $F(x) = g(c(x))$

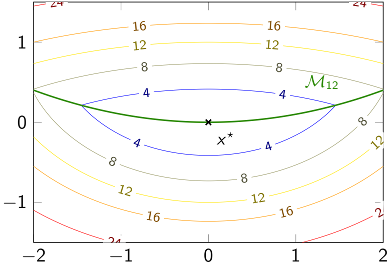
Includes: max. of $\mathcal{C}^2$ functions, max. eigenvalue

Observations

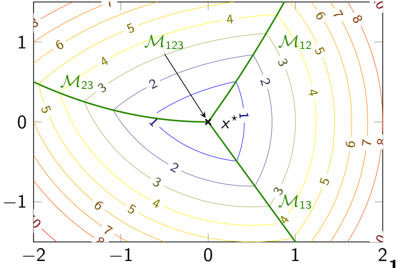
- nondifferentiability points organize in smooth manifolds
- F is smooth on them

These are ***structure manifolds***. ◇ Lewis '02

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



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



Many algorithms for nonsmooth (composite) optimization:

- prox-linear methods ◇ Lewis Wright, '16,
- bundle methods ◇ Mifflin Sagastizábal, '05,
- gradient sampling ◇ Burke Lewis Overton, '05,
- nonsmooth BFGS ◇ Lewis Overton, '13

Most algorithms are oblivious to structure, ***we try to leverage it***.

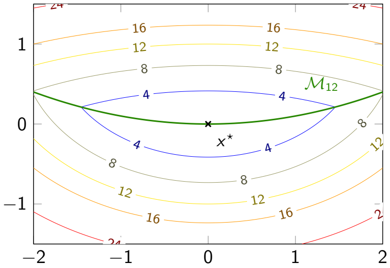
Composite problem

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

Finding a minimizer of F nonsmooth can be seen as:

- find the right structure
  e.g. which c_i are maximum
- leverage the right structure to minimize F
  e.g. solve smooth problem with smooth constraints

$\rightarrow$ We replace (nonsmooth) minimization by smooth constrained minimization.



Challenges:

1. How to detect the optimal structure $\mathcal{M}^\star \ni x^\star$?
2. How to exploit structure to better minimize F?

Prox. for finding structure

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

For *simple functions*, the proximity operator can be computed exactly

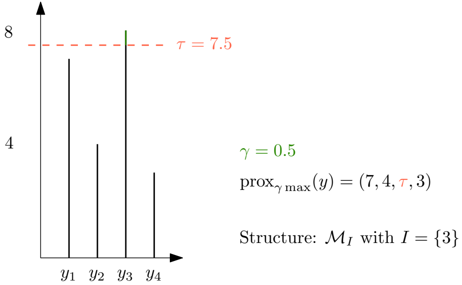
Example (Prox of max)

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

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

Structure manifold:

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



Prox. for finding structure

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

For *simple functions*, the proximity operator can be computed exactly

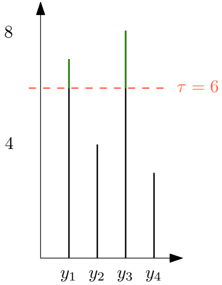
Example (Prox of max)

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

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

Structure manifold:

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



$\gamma = 5$

$\mathrm{prox}_{\gamma\max}(y) = (\tau, 4, \tau, 3)$

Structure: $\mathcal{M}_I$ with $I = \{1, 3\}$

Prox. for finding structure

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

For *simple functions*, the proximity operator can be computed exactly

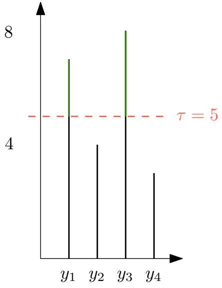
Example (Prox of max)

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

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

Structure manifold:

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



$\gamma = 5$

$\mathrm{prox}_{\gamma \max}(y) = (\tau, 4, \tau, 3)$

Structure: $\mathcal{M}_I$ with $I = \{1, 3\}$

Prox. for finding structure

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

For *simple functions*, the proximity operator can be computed exactly

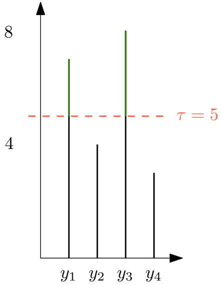
Example (Prox of max)

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

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

Structure manifold:

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



$\gamma = 5$

$\mathrm{prox}_{\gamma \max}(y) = (\tau, 4, \tau, 3)$

Structure: $\mathcal{M}_I$ with $I = \{1, 3\}$

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

Identification with explicit prox

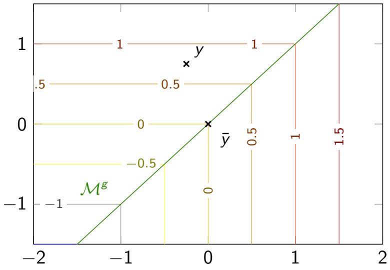
Lemma (B., Iutzeler, Malick, '22)

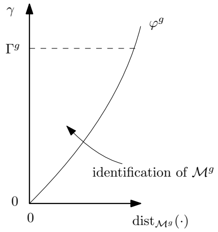
Consider a function g and point $\bar{y}$ with structure $\mathcal{M}^g$ that meet two technical assumptions. For all y near $\bar{y}$,

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

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 on the manifold.





Share similarities with ◇ Lewis '02, ◇ Lewis Hare '04, ◇ Vaiter Peyré Fadili '17

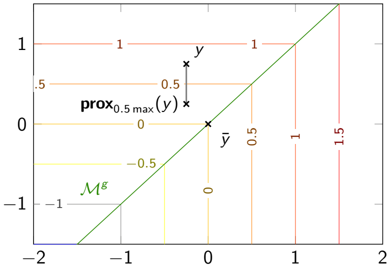
Identification with explicit prox

Lemma (B., Iutzeler, Malick, '22)

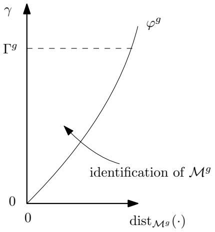
Consider a function g and point $\bar{y}$ with structure $\mathcal{M}^g$ that meet two technical assumptions. For all y near $\bar{y}$,

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

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 on the manifold.



Share similarities with ◇ Lewis '02, ◇ Lewis Hare '04, ◇ Vaiter Peyré Fadili '17

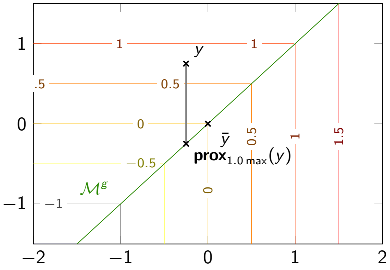
Identification with explicit prox

Lemma (B., Iutzeler, Malick, '22)

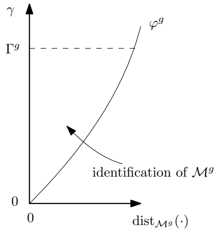
Consider a function g and point $\bar{y}$ with structure $\mathcal{M}^g$ that meet two technical assumptions. For all y near $\bar{y}$,

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

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 on the manifold.



Share similarities with ◇ Lewis '02, ◇ Lewis Hare '04, ◇ Vaiter Peyré Fadili '17

Identification with explicit prox

Lemma (B., Iutzeler, Malick, '22)

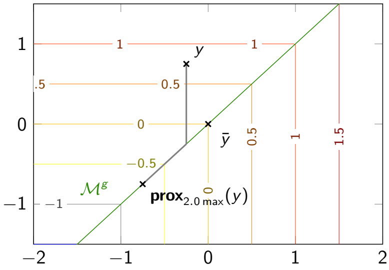
Consider a function g and point $\bar{y}$ with structure $\mathcal{M}^g$ that meet two technical assumptions. For all y near $\bar{y}$,

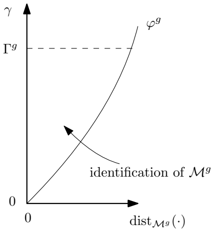
$$\mathbf{prox}_{\gamma g}(y) \in \mathcal{M}^g \quad \text{for all } \gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$$

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 on the manifold.

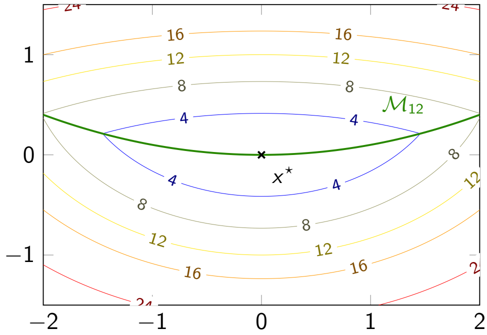
Share similarities with ◇ Lewis '02, ◇ Lewis Hare '04, ◇ Vaiter Peyré Fadili '17



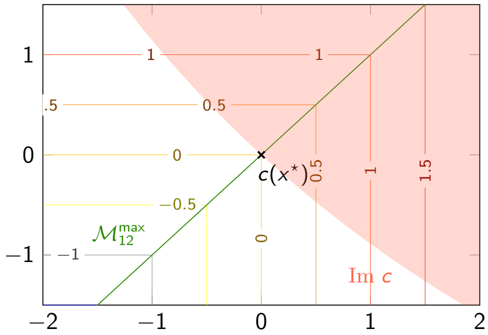


No prox. of F

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



$F(x) = \max(c_1(x), \ldots, c_m(x))$



$g(y) = \max(y_1, \ldots, y_m)$

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

Back to the optimization space

Theorem (B., Iutzeler, Malick, '22)

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

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

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

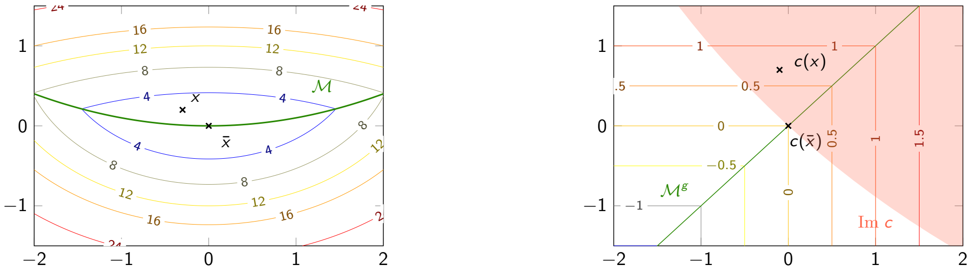
Back to the optimization space

Theorem (B., Iutzeler, Malick, '22)

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

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

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



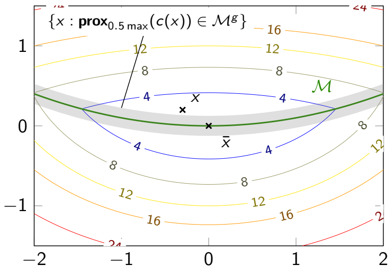
Back to the optimization space

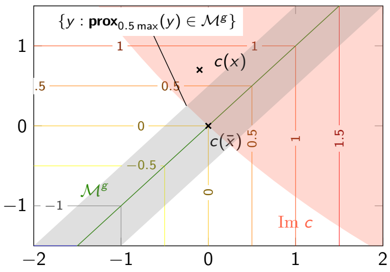
Theorem (B., Iutzeler, Malick, '22)

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

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

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





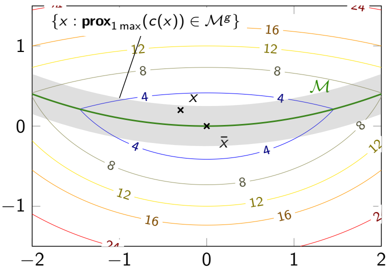
Back to the optimization space

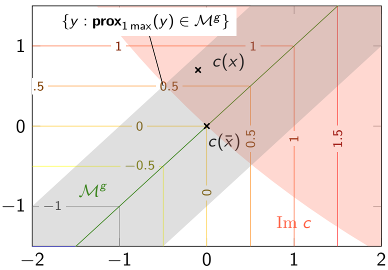
Theorem (B., Iutzeler, Malick, '22)

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

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

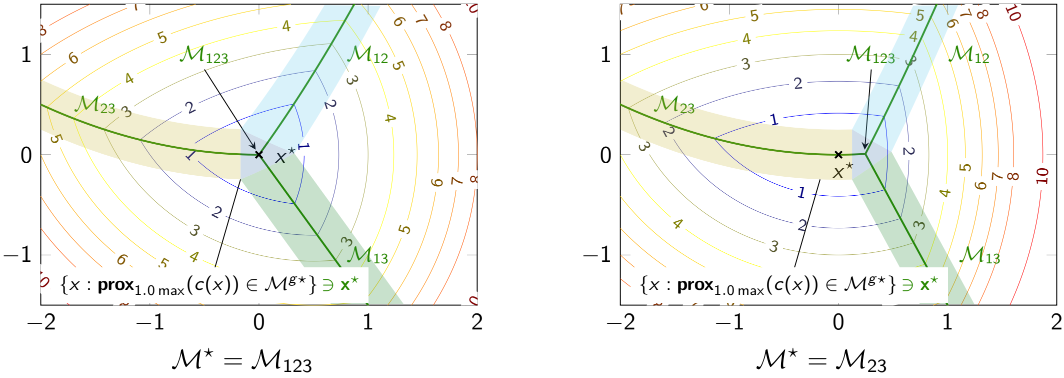
where $\Gamma > 0$ and $\varphi(t) = \frac{c_{map}}{c_{ri}} t + \mathcal{O}(t^2)$. Furthermore, $\mathcal{M} = c^{-1}(\mathcal{M}^g)$.





Detection with multiple manifolds

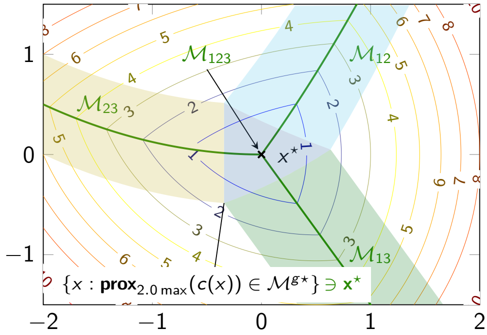
Generally, there are more than one manifolds near $x^\star$.



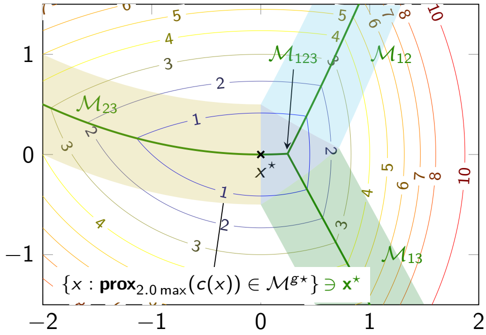
Importance of γ: too small, detection of $\mathcal{M}^\star$ only near $x^\star$; too large, no detection near $x^\star$.

Detection with multiple manifolds

Generally, there are more than one manifolds near $x^\star$.



$\mathcal{M}^\star = \mathcal{M}_{123}$

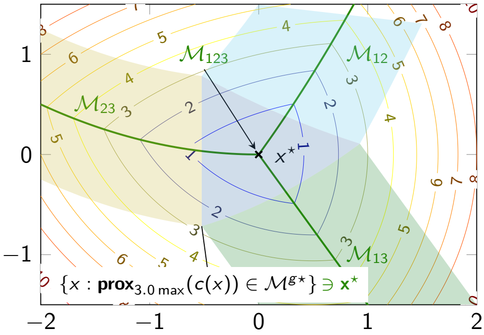


$\mathcal{M}^\star = \mathcal{M}_{23}$

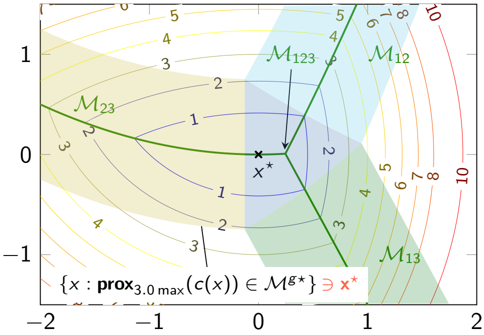
Importance of γ: too small, detection of $\mathcal{M}^\star$ only near $x^\star$; too large, no detection near $x^\star$.

Detection with multiple manifolds

Generally, there are more than one manifolds near $x^\star$.



$\mathcal{M}^\star = \mathcal{M}_{123}$

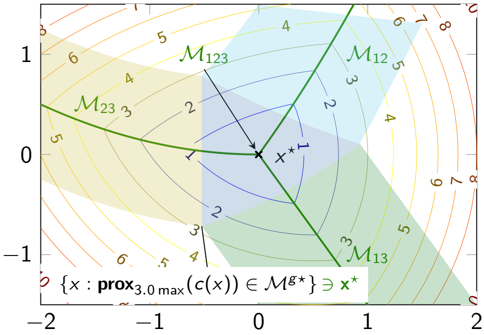


$\mathcal{M}^\star = \mathcal{M}_{23}$

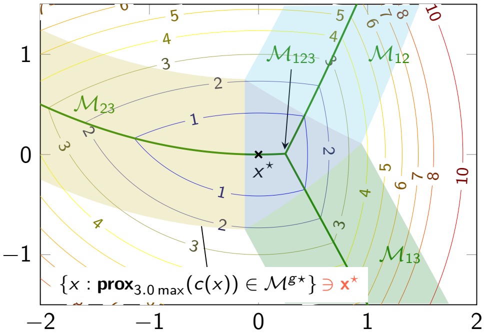
Importance of γ: too small, detection of $\mathcal{M}^\star$ only near $x^\star$; too large, no detection near $x^\star$.

Detection with multiple manifolds

Generally, there are more than one manifolds near $x^\star$.



$\mathcal{M}^\star = \mathcal{M}_{123}$



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

Importance of γ: too small, detection of $\mathcal{M}^\star$ only near $x^\star$; too large, no detection near $x^\star$.

Take-away: We detect $\mathcal{M}^\star \ni x^\star$ with $\mathbf{prox}_{\gamma g} \circ c(\cdot)$ with the right range of steps.

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

Nonsmooth to smooth

- Structure manifolds provide second order models of the nonsmooth F:

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

$$\min_x F(x) \quad \text{and} \quad \mathcal{M} \qquad \textit{turns into} \qquad \min_x \widetilde{F}(x) \text{ s.t. } h(x) = 0.$$

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

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, ...

Newton step and algorithm

Iteration k:

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

- With structure candidate $\mathcal{M}_k$: SQP step on $\min_x \widetilde{F}_k(x)$ s.t. $h_k(x) = 0$.

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

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

- $\gamma_{k+1} = \frac{\gamma_k}{2}$

Similar works with *heuristic* structure detection: ⋄ Womersley Fletcher '86 for max, ⋄ Noll Apkarian, '05 for $\lambda_{\max}$.

Local exact structure identification and quadratic convergence

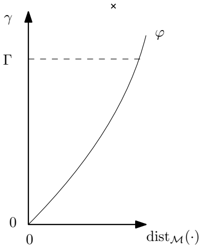
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough *and no Maratos effect happens*, *then there exists $C > 0$ such that:*

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_\mathcal{M}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Local exact structure identification and quadratic convergence

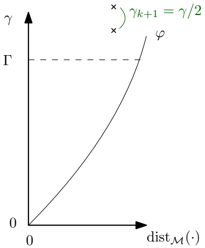
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and no Maratos effect happens, then there exists $C > 0$ such that:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_\mathcal{M}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Local exact structure identification and quadratic convergence

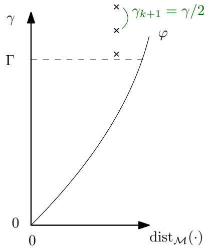
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and no Maratos effect happens, then there exists $C > 0$ such that:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_{\mathcal{M}}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Local exact structure identification and quadratic convergence

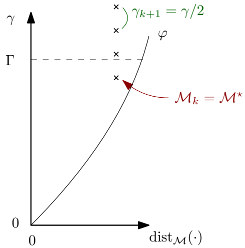
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and no Maratos effect happens, then there exists $C > 0$ such that:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_{\mathcal{M}}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Local exact structure identification and quadratic convergence

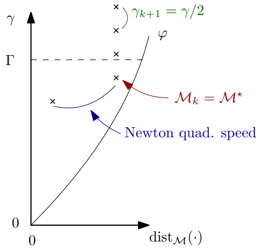
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and no Maratos effect happens, then there exists $C > 0$ such that:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_{\mathcal{M}}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Local exact structure identification and quadratic convergence

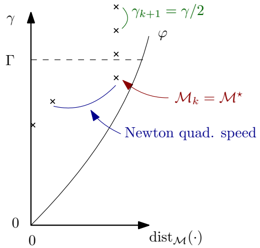
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and no Maratos effect happens, then there exists $C > 0$ such that:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \textit{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \textit{for all } k \textit{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_{\mathcal{M}}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Local exact structure identification and quadratic convergence

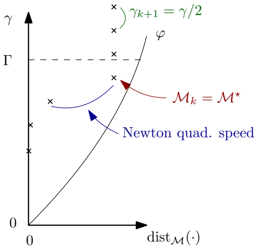
Theorem (B., Iutzeler, Malick, '22)

Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer with structure manifold $\mathcal{M}^\star$ that meets the technical assumptions.
If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and no Maratos effect happens, then there exists $C > 0$ such that:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

Proof idea

- if $\mathcal{M}_k = \mathcal{M}^\star$, the SQP step brings quadratic improvement
- since γ_k decreases, at some point $\gamma_k \in [\varphi(\mathrm{dist}_\mathcal{M}(x_k)), \Gamma]$
- to stay in that region, decrease γ not too fast



Quadratic convergence

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

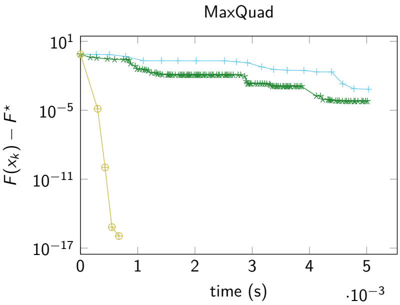
$$\mathcal{M}^\star = \{x : c_2(x) = \cdots = c_5(x)\}$$

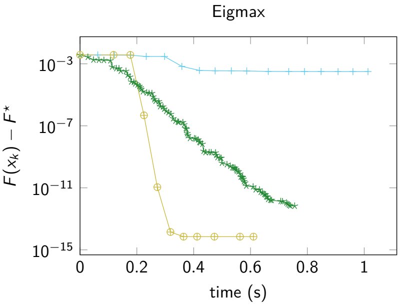
Historical maxquad problem ◇ HULL '93

$$\min_{x \in \mathbb{R}^{25}} \lambda_{\max} \left(A_0 + \sum_{i=1}^{n} x_i A_i \right)$$

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

Matrices are symmetric, 50×50





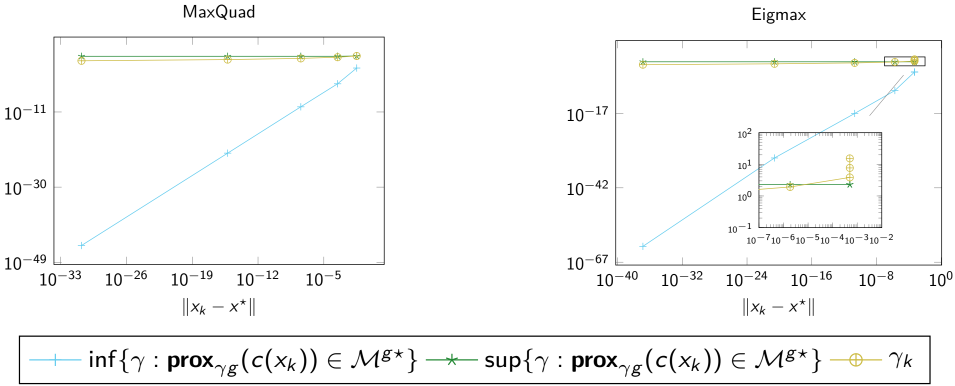
Gradient Sampling nsBFGS LocalNewton

Proximal identification

Corollary: There exists $L > 0$, $\epsilon > 0$ such that

$$\|x - x^\star\| \leq \epsilon \text{ and } L\|x - x^\star\| \leq \gamma \leq \Gamma \quad \Longrightarrow \quad \mathbf{prox}_{\gamma g}(c(x)) \in \mathcal{M}^{g\star}.$$

This checks out in practice:



Conclusion

Take-away messages

- Proximal methods identify smooth structure in nonsmooth composite problems
- We show local ***exact*** identification and quadratic rate for $g \circ c$, where g is prox-simple, no convexity required

B. & Iutzeler & Malick: Harnessing structure in composite nonsmooth minimization

soon

Work in progress and perspectives

- Drop the locality: i) need more information to identify, ii) globalize constrained Newton

Thank you!

Gilles Bareilles – gbareilles.fr

Technical assumptions

Normal ascent: g increases at $\bar{y}$ on normal directions:

$$0 \in \mathrm{ri}\ \mathbf{proj}_{N_{\bar{y}}\mathcal{M}^g} \partial g(\bar{y})$$

Manifold curves: A function g with structure $\mathcal{M}^g$ at $\bar{y}$ satisfies the *curve property* if there exists a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ and $T > 0$ such that, for any smooth application $e : \mathcal{N}_{\bar{y}} \times [0, T] \to \mathcal{M}^g$ verifying $e(y, 0) = \mathbf{proj}_{\mathcal{M}^g}(y)$ and $\frac{d}{dt} e(y, 0) = -\operatorname{grad} g(\mathbf{proj}_{\mathcal{M}^g}(y))$, there holds

$$\| \mathbf{proj}_{N_{e(y,t)}\mathcal{M}^g}(e(y,t) - y)\| \leq \mathrm{dist}_{\mathcal{M}^g}(y) + \tilde{L}\, t^2 \quad \text{for all } y \in \mathcal{N}_{\bar{y}}, t \in [0, T],$$

where $\operatorname{grad} g(p) \in T_p\mathcal{M}^g$ denotes the Riemannian gradient of g, obtained as $\mathbf{proj}_{T_p\mathcal{M}^g}(\partial g(p))$.

No Maratos: near a minimizer $x^\star$, a step d that makes $x + d$ quadratically closer to $x^\star$ than x implies descent $F(x + d) \leq F(x)$.

Transversality: the mapping $c : \mathbb{R}^n \to \mathbb{R}^m$ is transversal to manifold $\mathcal{M} \subset \mathbb{R}^m$ at $c(x)$ if:

$$\ker\left(\mathrm{Jac}_c(x)^\top\right) \cap N_{c(x)}\mathcal{M}^g = \{0\}$$

$\Rightarrow$ if $\mathrm{Jac}_h(c(x))$ is full rank, then $\mathrm{Jac}_{h \circ c}(x)$ is also full-rank.

Maximum structure and initial stepsize

In the generated instance, the multiplicity of the maximum eigenvalue at optimum is $r = 3$. The maximum structure of a point, useful in setting γ_0, is $\mathcal{M}_r$, with $r = 6$, and not the matrix size $m = 50$. Indeed, the codimension of $\mathcal{M}_r$, that is the dimension of its normal spaces, should be lower than that of $\mathbb{R}^n$: $r(r+1)/2 - 1 \leq 25$, that is $r \leq 6$ (see the discussion in [**?**, pp. 555-556, Eq. 2.5]).

Quadratic convergence, BigFloat precision

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

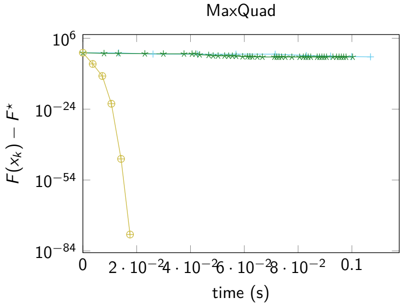
$$\mathcal{M}^\star = \{x : c_2(x) = \cdots = c_5(x)\}$$

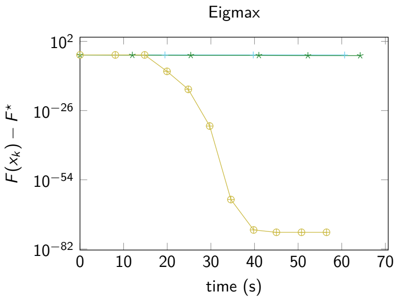
Historical maxquad problem

$$\min_{x \in \mathbb{R}^{25}} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right)$$

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

Matrices are symmetric, 50×50.





Gradient Sampling — nsBFGS — LocalNewton