Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems

Jérôme Bolte, Tam Le, Edouard Pauwels, and Antonio Silveti-Falls




ISMP 2024 Montréal

Plan

- **Motivation**
- Conservative Gradients
- Results
- Applications
- Numerical Examples

A Motivating Example

Recall the LASSO problem:

$$\hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

A Motivating Example

Recall the LASSO problem:

$$\hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

Here, $A \in \mathbb{R}^{n \times p}$ is the design matrix for the n observations, $b \in \mathbb{R}^n$ is the associated response, and $\theta \in \mathbb{R}$ is a parameter.

A Motivating Example

Recall the LASSO problem:

$$\hat{x} \in \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \left\| Ax - b \right\|_2^2 + e^{\theta} \left\| x \right\|_1 .$$

Here, $A \in \mathbb{R}^{n \times p}$ is the design matrix for the n observations, $b \in \mathbb{R}^n$ is the associated response, and $\theta \in \mathbb{R}$ is a parameter.

Given some measure of task performance $C(\hat{x}(\theta))$, how to pick the "best" value of θ?

The problem of choosing θ becomes a bilevel optimization problem:

$$\min_{\theta \in \mathbb{R}} C(\hat{x}(\theta)) \quad \text{such that} \quad \hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

A Motivating Example

The problem of choosing θ becomes a bilevel optimization problem:

$$\min_{\theta \in \mathbb{R}} C(\hat{x}(\theta)) \quad \text{such that} \quad \hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

If C and $\hat{x}$ are smooth then we can use first-order optimization methods using the gradient:

$$\nabla C(\hat{x}(\theta)) = J_{\hat{x}}(\theta)^T \nabla_x C(\hat{x}(\theta)).$$

The problem of choosing θ becomes a bilevel optimization problem:

$$\min_{\theta \in \mathbb{R}} C(\hat{x}(\theta)) \quad \text{such that} \quad \hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

If C and $\hat{x}$ are smooth then we can use first-order optimization methods using the gradient:

$$\nabla C(\hat{x}(\theta)) = J_{\hat{x}}(\theta)^T \nabla_x C(\hat{x}(\theta)).$$

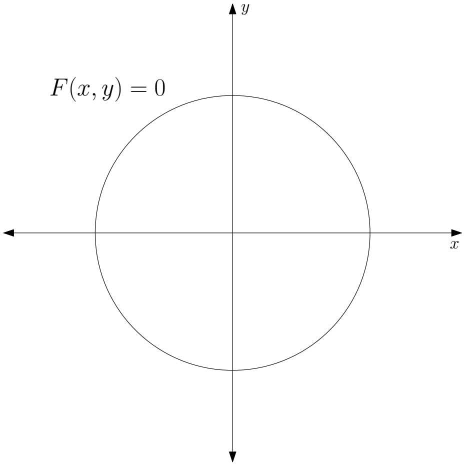
However, $\hat{x}(\cdot)$ might not be smooth (often the case in machine learning settings). We need a method to derive functions like $\hat{x}$ which are implicitly defined.

Consider the smooth function

$$F(x, y) = x^2 + y^2 - 1$$

and the equation

$$F(x, y) = 0.$$



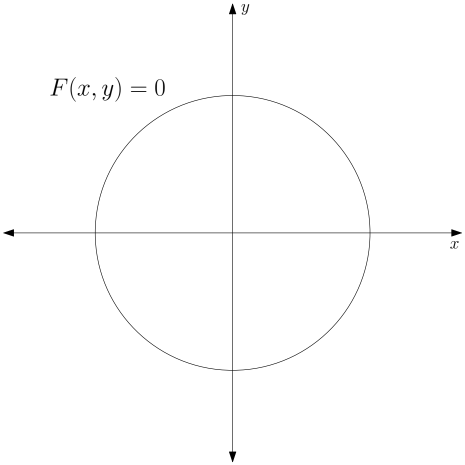
Consider the smooth function

$$F(x, y) = x^2 + y^2 - 1$$

and the equation

$$F(x, y) = 0.$$

- Can we find a function $y = G(x)$ so that $F(x, G(x)) = 0$?



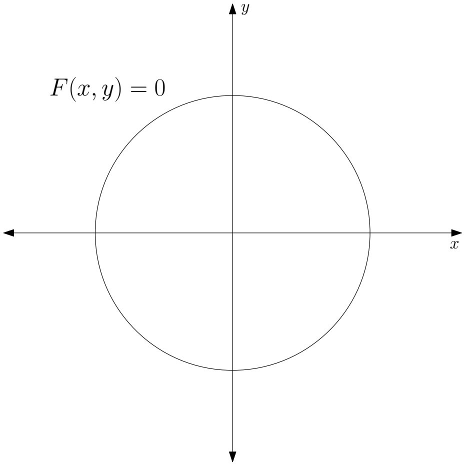
Consider the smooth function

$$F(x, y) = x^2 + y^2 - 1$$

and the equation

$$F(x, y) = 0.$$

- Can we find a function $y = G(x)$ so that $F(x, G(x)) = 0$?
- Can we compute the gradient of G ?



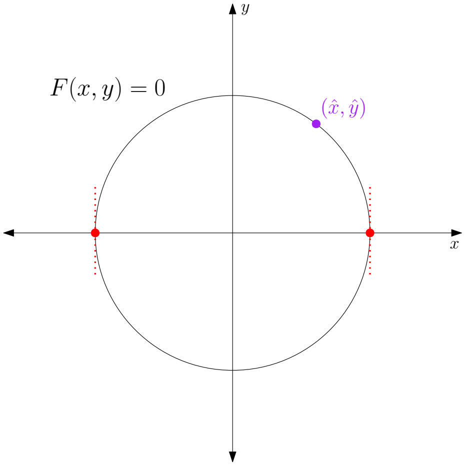
Consider the smooth function

$$F(x, y) = x^2 + y^2 - 1$$

and the equation

$$F(x, y) = 0.$$

- Can we find a function $y = G(x)$ so that $F(x, G(x)) = 0$?
- Can we compute the gradient of G ?



Consider the smooth function

$$F(x, y) = x^2 + y^2 - 1$$

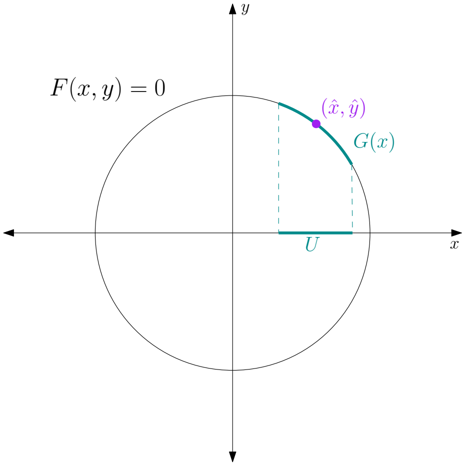
and the equation

$$F(x, y) = 0.$$

- Can we find a function $y = G(x)$ so that $F(x, G(x)) = 0$?

Existence

- Can we compute the gradient of G ?



Consider the smooth function

$$F(x, y) = x^2 + y^2 - 1$$

and the equation

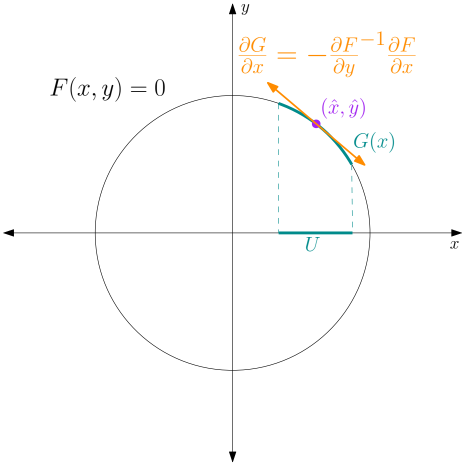
$$F(x, y) = 0.$$

- Can we find a function $y = G(x)$ so that $F(x, G(x)) = 0$?

Existence

- Can we compute the gradient of G ?

Calculus



In the nonsmooth world...

Existence: can we find implicit functions in the nonsmooth setting?

Calculus: can we compute the corresponding "gradient-like" objects in this setting?

In the nonsmooth world. . .

Existence: can we find implicit functions in the nonsmooth setting?
Yes! See (Clarke 1976, Hiriart-Urruty 1979, etc.) for locally Lipschitz functions.

Calculus: can we compute the corresponding "gradient-like" objects in this setting?

Motivation

In the nonsmooth world. . .

Existence: can we find implicit functions in the nonsmooth setting?
Yes! See (Clarke 1976, Hiriart-Urruty 1979, etc.) for locally Lipschitz functions.

Calculus: can we compute the corresponding "gradient-like" objects in this setting?
Not with past theorems (Clarke, etc) - possible with conservative Jacobians [Bolte, Pauwels 2021] and path differentiable functions.

In the nonsmooth world...

Existence: can we find implicit functions in the nonsmooth setting?
Yes! See (Clarke 1976, Hiriart-Urruty 1979, etc.) for locally Lipschitz functions.

Calculus: can we compute the corresponding "gradient-like" objects in this setting?
Not with past theorems (Clarke, etc) - possible with conservative Jacobians [Bolte, Pauwels 2021] and path differentiable functions.

In practice one hopes for an algorithm of the form

$$x^+ = x - \gamma d(x)$$

where $d(x)$ is some descent direction or surrogate "gradient".

Nonsmooth Implicit Function Theorem with Clarke Subdifferential

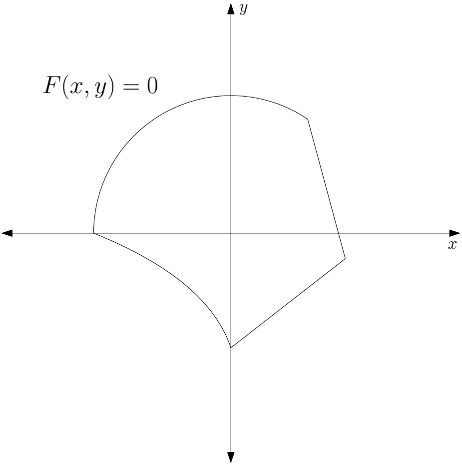
Theorem (Clarke 1976, Hiriart-Urruty 1979)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$F(\hat{x}, \hat{y}) = 0.$$

If, $\forall [A\ B] \in \partial^c F(\hat{x}, \hat{y})$, B is invertible, then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a locally Lipschitz function $G(x)$ so that

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



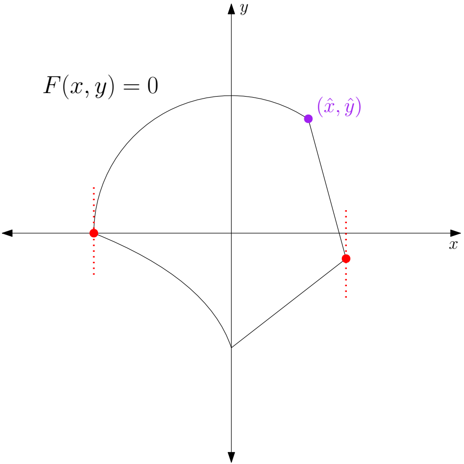
Theorem (Clarke 1976, Hiriart-Urruty 1979)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$F(\hat{x}, \hat{y}) = 0.$$

If, $\forall [A \; B] \in \partial^c F(\hat{x}, \hat{y})$, B is invertible, then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a locally Lipschitz function $G(x)$ so that

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



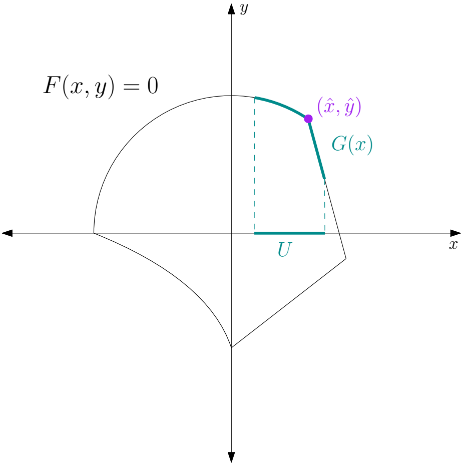
Theorem (Clarke 1976, Hiriart-Urruty 1979)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$F(\hat{x}, \hat{y}) = 0.$$

If, $\forall [A\ B] \in \partial^c F(\hat{x}, \hat{y})$, B is invertible, then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a locally Lipschitz function $G(x)$ so that

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



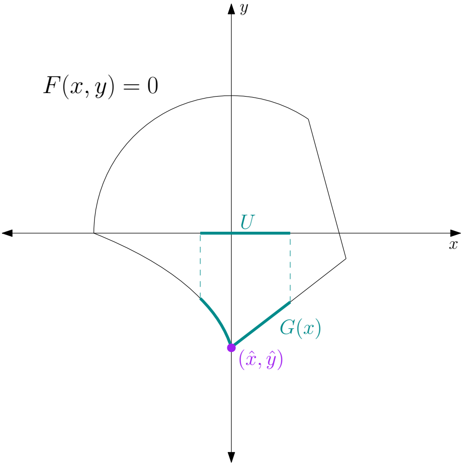
Theorem (Clarke 1976, Hiriart-Urruty 1979)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$F(\hat{x}, \hat{y}) = 0.$$

If, $\forall [A\ B] \in \partial^c F(\hat{x}, \hat{y})$, B is invertible, then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a locally Lipschitz function $G(x)$ so that

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that Clarke's IFT holds with implicit function $G(x)$.

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that Clarke's IFT holds with implicit function $G(x)$.

Recall from smooth IFT: $J_G(x) = -B^{-1}A \quad [A\ B] = J_F(x, G(x))$.

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that Clarke's IFT holds with implicit function $G(x)$.

Recall from smooth IFT: $J_G(x) = -B^{-1}A \quad [A\ B] = J_F(x, G(x))$.

Question

Can we have a calculus of the form:

$$\left\{-B^{-1}A : [A\ B] \in \partial^c F(\hat{x}, \hat{y})\right\} = \partial^c G(x)$$

for the Clarke Jacobian?

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that Clarke's IFT holds with implicit function $G(x)$.

Recall from smooth IFT: $J_G(x) = -B^{-1}A \quad [A\ B] = J_F(x, G(x))$.

Question

Can we have a calculus of the form:

$$\left\{-B^{-1}A : [A\ B] \in \partial^c F(\hat{x}, \hat{y})\right\} = \partial^c G(x)$$

for the Clarke Jacobian?

No - need something beyond ∂^c.

Lack of Calculus

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that Clarke's IFT holds with implicit function $G(x)$.

Recall from smooth IFT: $J_G(x) = -B^{-1}A \quad [A\ B] = J_F(x, G(x))$.

Question

Can we have a calculus of the form:

$$\left\{-B^{-1}A : [A\ B] \in \partial^c F(\hat{x}, \hat{y})\right\} = \partial^c G(x)$$

for the Clarke Jacobian?

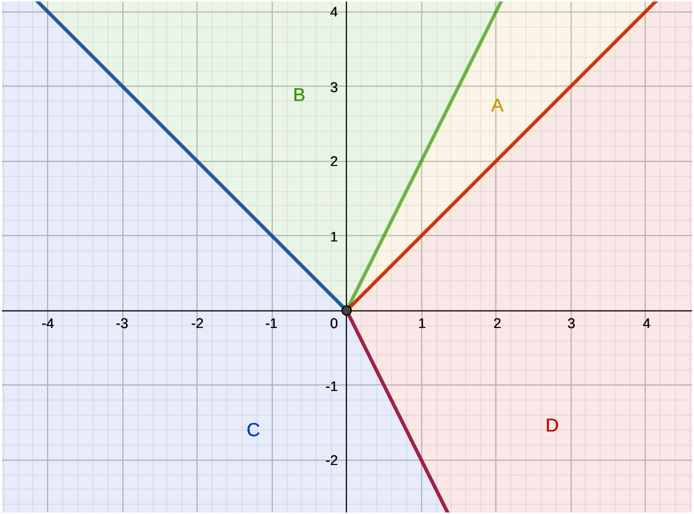
No - need something beyond ∂^c.

$\exists$ piecewise linear $F : \mathbb{R}^2 \to \mathbb{R}^2$ for which Clarke's inverse mapping theorem fails:

$$\exists M \in J^c_F(0,0) \quad \text{such that} \quad M^{-1} \notin J^c_{F^{-1}}(0,0)$$

Counterexample

$$F(x, y) = (|x| + y, 2x + |y|)$$



F^{-1} is linear on each region $A, B, C, \& D$.

Plan

- Motivation
- **Conservative Gradients**
- Results
- Applications
- Numerical Examples

Definition (Conservative field (Bolte, Pauwels 2019))

A set valued mapping $D_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field (or conservative Jacobian) for $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz if:

Definition (Conservative field (Bolte, Pauwels 2019))

A set valued mapping $D_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field (or conservative Jacobian) for $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz if:

1. For all $x \in \mathbb{R}^n$, $D_F(x)$ is nonempty (ideally convex!).

Definition (Conservative field (Bolte, Pauwels 2019))

A set valued mapping $D_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field (or conservative Jacobian) for $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz if:

1. For all $x \in \mathbb{R}^n$, $D_F(x)$ is nonempty (ideally convex!).
2. D_F has a closed graph and is locally bounded.

Definition (Conservative field (Bolte, Pauwels 2019))

A set valued mapping $D_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field (or conservative Jacobian) for $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz if:

1. For all $x \in \mathbb{R}^n$, $D_F(x)$ is nonempty (ideally convex!).
2. D_F has a closed graph and is locally bounded.
3. For any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F(\gamma(t)) = \langle u, \dot{\gamma}(t)\rangle \qquad \forall u \in D_F(\gamma(t))$$

   for almost all $t \in [0,1]$.

We call F path differentiable.

Definition (Conservative field (Bolte, Pauwels 2019))

A set valued mapping $D_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field (or conservative Jacobian) for $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz if:

1. For all $x \in \mathbb{R}^n$, $D_F(x)$ is nonempty (ideally convex!).
2. D_F has a closed graph and is locally bounded.
3. For any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F\left(\gamma\left(t\right)\right) = \langle u, \dot{\gamma}(t) \rangle \qquad \forall u \in D_F\left(\gamma(t)\right)$$

   for almost all $t \in [0,1]$.

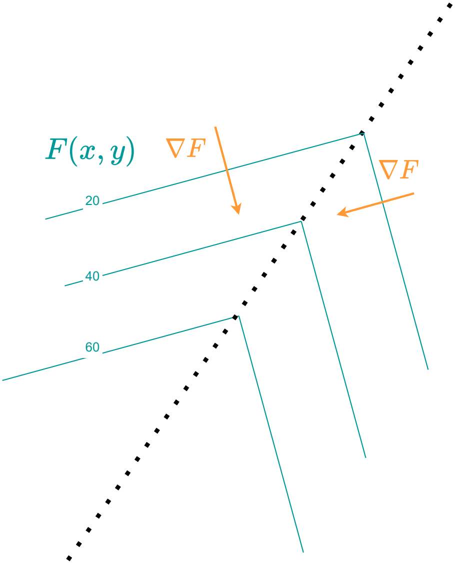
We call F path differentiable.

Take home message: conservative fields/Jacobians faithfully model what is computed by backpropagation in practice.

For any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F(\gamma(t)) = \langle u, \dot{\gamma}(t) \rangle$$

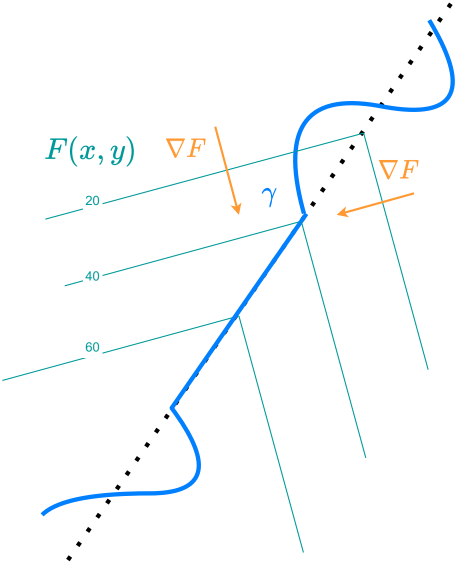
for all $u \in D_F(\gamma(t))$,
for almost all $t \in [0,1]$.



For any absolutely continuous curve $\gamma : [0, 1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F(\gamma(t)) = \langle u, \dot{\gamma}(t) \rangle$$

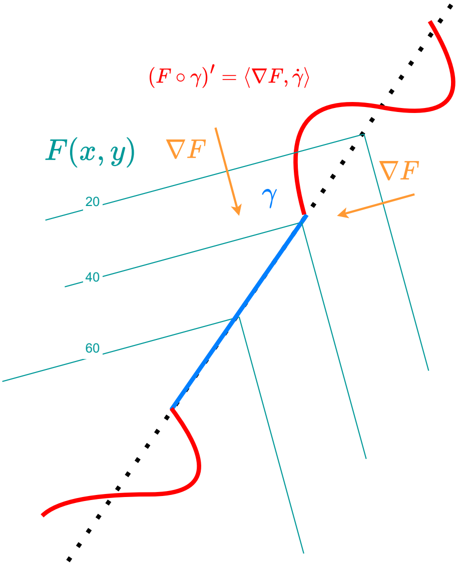
for all $u \in D_F(\gamma(t))$,
for almost all $t \in [0, 1]$.



For any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F(\gamma(t)) = \langle u, \dot{\gamma}(t) \rangle$$

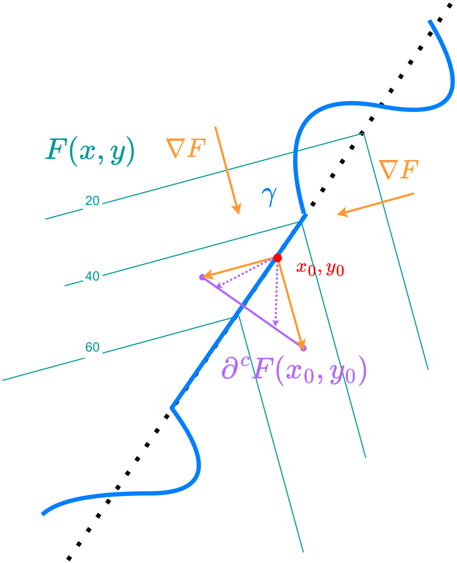
for all $u \in D_F(\gamma(t))$,
for almost all $t \in [0,1]$.



For any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F(\gamma(t)) = \langle u, \dot{\gamma}(t) \rangle$$

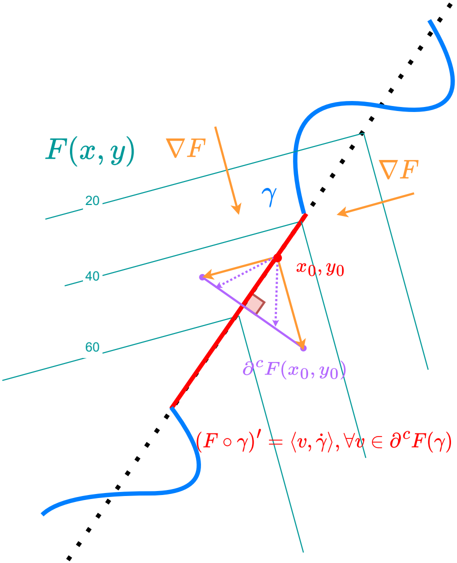
for all $u \in D_F(\gamma(t))$,
for almost all $t \in [0,1]$.



For any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt} F(\gamma(t)) = \langle u, \dot{\gamma}(t) \rangle$$

for all $u \in D_F(\gamma(t))$,
for almost all $t \in [0,1]$.



Plan

- Motivation
- Conservative Gradients
- **Results**
- Applications
- Numerical Examples

Theorem (Bolte, Le, Pauwels, S.F. 2021)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path diff. and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that

$$F(\hat{x}, \hat{y}) = 0.$$

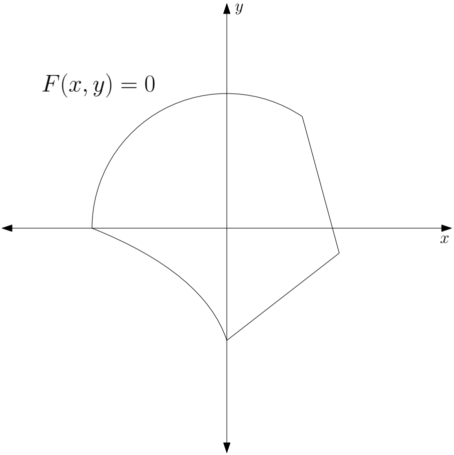
Assume $D_F(\hat{x}, \hat{y})$ is convex and $\forall [A\ B] \in D_F(\hat{x}, \hat{y})$, B is invertible. Then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a path diff. function G such that

$$\forall x \in U \qquad F(x, G(x)) = 0.$$

A conservative Jacobian of G is

$$D_G(x) = \left\{ -B^{-1}A : [A\ B] \in D_F(x, G(x)) \right\}$$

and is compatible with backprop.



Path Differentiable Nonsmooth Implicit Function Theorem

Theorem (Bolte, Le, Pauwels, S.F. 2021)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path diff. and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that

$$F(\hat{x}, \hat{y}) = 0.$$

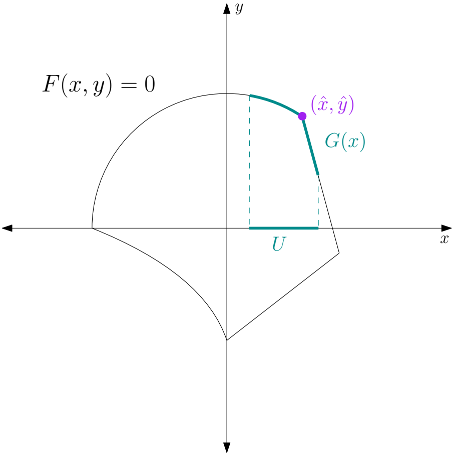
Assume $D_F(\hat{x}, \hat{y})$ is convex and $\forall [A\ B] \in D_F(\hat{x}, \hat{y})$, B is invertible. Then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a path diff. function G such that

$$\forall x \in U \qquad F(x, G(x)) = 0.$$

A conservative Jacobian of G is

$$D_G(x) = \left\{ -B^{-1}A : [A\ B] \in D_F(x, G(x)) \right\}$$

and is compatible with backprop.



Theorem (Bolte, Le, Pauwels, S.F. 2021)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path diff. and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that

$$F(\hat{x}, \hat{y}) = 0.$$

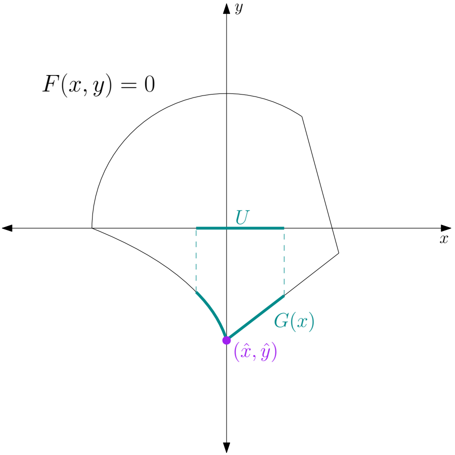
Assume $D_F(\hat{x}, \hat{y})$ is convex and $\forall [A\ B] \in D_F(\hat{x}, \hat{y})$, B is invertible. Then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a path diff. function G such that

$$\forall x \in U \qquad F(x, G(x)) = 0.$$

A conservative Jacobian of G is

$$D_G(x) = \left\{ -B^{-1}A : [A\ B] \in D_F(x, G(x)) \right\}$$

and is compatible with backprop.



Theorem (Bolte, Le, Pauwels, S.F. 2021)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path diff. and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that

$$F(\hat{x}, \hat{y}) = 0.$$

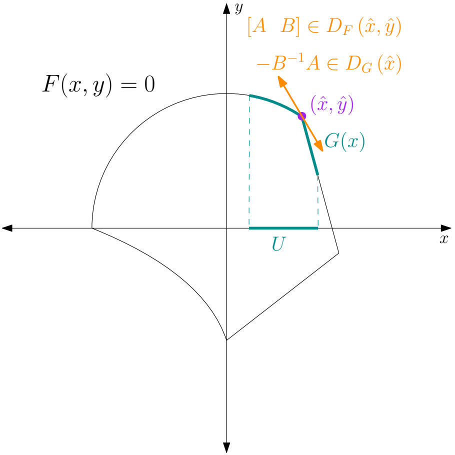
Assume $D_F(\hat{x}, \hat{y})$ is convex and $\forall [A \ B] \in D_F(\hat{x}, \hat{y})$, B is invertible. Then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a path diff. function G such that

$$\forall x \in U \qquad F(x, G(x)) = 0.$$

A conservative Jacobian of G is

$$D_G(x) = \left\{ -B^{-1}A : [A \ B] \in D_F(x, G(x)) \right\}$$

and is compatible with backprop.



Path Differentiable Nonsmooth Implicit Function Theorem

Theorem (Bolte, Le, Pauwels, S.F. 2021)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path diff. and $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that

$$F(\hat{x}, \hat{y}) = 0.$$

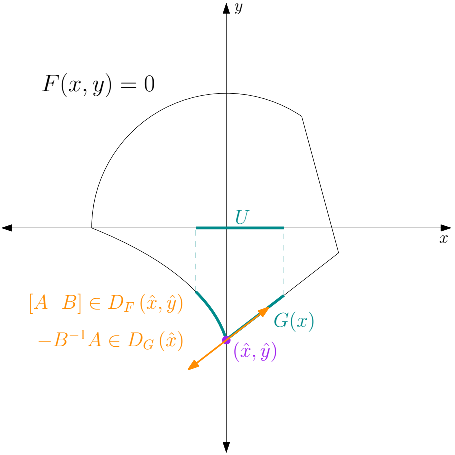
Assume $D_F(\hat{x}, \hat{y})$ is convex and $\forall [A\ B] \in D_F(\hat{x}, \hat{y})$, B is invertible. Then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a path diff. function G such that

$$\forall x \in U \qquad F(x, G(x)) = 0.$$

A conservative Jacobian of G is

$$D_G(x) = \left\{ -B^{-1}A : [A\ B] \in D_F(x, G(x)) \right\}$$

and is compatible with backprop.



Plan

- Motivation
- Conservative Gradients
- Results
- **Applications**
- Numerical Examples

With our new theorem we can answer the question:

how to differentiate the solution to a nonsmooth convex optimization problem ?

$$\hat{x}(\theta) := \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x, \theta)$$

With our new theorem we can answer the question:

how to differentiate the solution to a nonsmooth convex optimization problem ?

$$\hat{x}(\theta) := \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x, \theta)$$

Solution: find a necessary and sufficient path differentiable optimality condition and apply our implicit function theorem.

The IFT and Optimization

With our new theorem we can answer the question:

how to differentiate the solution to a nonsmooth convex optimization problem ?

$$\hat{x}(\theta) := \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x, \theta)$$

Solution: find a necessary and sufficient path differentiable optimality condition and apply our implicit function theorem.

- Hyperparameter tuning of the LASSO [Bertrand, Klopfenstein, Blondel, Vaiter, Gramfort, Salmon 2020].

The IFT and Optimization

With our new theorem we can answer the question:

how to differentiate the solution to a nonsmooth convex optimization problem ?

$$\hat{x}(\theta) := \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x, \theta)$$

Solution: find a necessary and sufficient path differentiable optimality condition and apply our implicit function theorem.

- Hyperparameter tuning of the LASSO [Bertrand, Klopfenstein, Blondel, Vaiter, Gramfort, Salmon 2020].
- Differentiating monotone inclusions with protodifferentiability [Adly, Rockafellar 2021].

The IFT and Optimization

With our new theorem we can answer the question:

how to differentiate the solution to a nonsmooth convex optimization problem ?

$$\hat{x}(\theta) := \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x, \theta)$$

Solution: find a necessary and sufficient path differentiable optimality condition and apply our implicit function theorem.

- Hyperparameter tuning of the LASSO [Bertrand, Klopfenstein, Blondel, Vaiter, Gramfort, Salmon 2020].
- Differentiating monotone inclusions with protodifferentiability [Adly, Rockafellar 2021].
- Set-valued implicit function theorems + semismooth localizations [Gferer, Outrata 2024].
- etc.

We recall the LASSO problem:

$$\hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

We recall the LASSO problem:

$$\hat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + e^{\theta} \|x\|_1 .$$

A fixed point condition for optimality:

$$\underbrace{\mathrm{prox}_{e^{\theta}\|\cdot\|_1}(\hat{x} - A^T(A\hat{x} - b)) - \hat{x}}_{F(\theta, x)} = 0.$$

the proximal mapping here is simply the "soft thresholding" operator, which is path differentiable. Thus, the function F is path differentiable.

Verification of Invertibility - Example With LASSO

It is necessary to verify the invertibility of the conservative Jacobian of F with respect to x evaluated at $(\theta, \hat{x})$.

Verification of Invertibility - Example With LASSO

It is necessary to verify the invertibility of the conservative Jacobian of F with respect to x evaluated at $(\theta, \hat{x})$.
We define the *equicorrelation* set:

$$\mathcal{E} := \left\{ j \in \{1, \ldots, p\} : \left| A_j^T (b - A\hat{x}(\theta)) \right| = e^{\theta} \right\}.$$

It is necessary to verify the invertibility of the conservative Jacobian of F with respect to x evaluated at $(\theta, \hat{x})$.
We define the *equicorrelation* set:

$$\mathcal{E} := \left\{ j \in \{1, \ldots, p\} : \left| A_j^T (b - A\hat{x}(\theta)) \right| = e^{\theta} \right\}.$$

- The set $\mathcal{E}$ does NOT depend on the choice of solution $\hat{x}(\theta)$ [Tibshirani 2013].

Verification of Invertibility - Example With LASSO

It is necessary to verify the invertibility of the conservative Jacobian of F with respect to x evaluated at $(\theta, \hat{x})$.
We define the *equicorrelation* set:

$$\mathcal{E} := \left\{ j \in \{1, \ldots, p\} : \left| A_j^T (b - A\hat{x}(\theta)) \right| = e^{\theta} \right\}.$$

- The set $\mathcal{E}$ does NOT depend on the choice of solution $\hat{x}(\theta)$ [Tibshirani 2013].
- The set $\mathcal{E}$ contains the *support* set $\mathcal{S} := \{i \in \{1, \ldots, p\} : \hat{x}_i(\theta) \neq 0\}$.

Verification of Invertibility - Example With LASSO

It is necessary to verify the invertibility of the conservative Jacobian of F with respect to x evaluated at $(\theta, \hat{x})$.
We define the *equicorrelation* set:

$$\mathcal{E} := \left\{ j \in \{1, \ldots, p\} : \left| A_j^T (b - A\hat{x}(\theta)) \right| = e^{\theta} \right\}.$$

- The set $\mathcal{E}$ does NOT depend on the choice of solution $\hat{x}(\theta)$ [Tibshirani 2013].
- The set $\mathcal{E}$ contains the *support* set $\mathcal{S} := \{i \in \{1, \ldots, p\} : \hat{x}_i(\theta) \neq 0\}$.

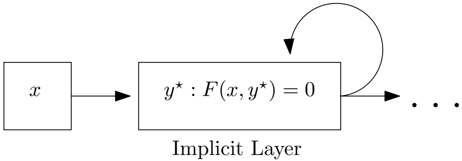
Proposition (Prop. 5 [Bolte, Le, Pauwels, S.F. 21])

Define, $\forall \theta \in \mathbb{R}$, the matrix $A_{\mathcal{E}}$ by taking the columns of A indexed by $\mathcal{E}$. If, $\forall \theta \in \mathbb{R}$, the matrix $A_{\mathcal{E}}^T A_{\mathcal{E}}$ is full rank, then $\hat{x}(\cdot)$ is a path differentiable function with a conservative Jacobian given by

$$D_{\hat{x}} : \theta \rightrightarrows \left\{ \left[-e^{\theta} (\mathrm{Id}_p - \mathrm{diag}(q)(\mathrm{Id}_p - A^T A))^{-1} \mathrm{diag}(q) \mathrm{sign}(\hat{x} - A^T (A\hat{x} - b)) \right] : q \in \mathcal{M}(\theta) \right\}$$

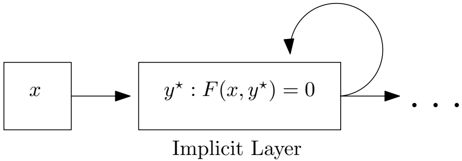
where $$\mathcal{M}(\theta) = \left\{ q : q_i \in \begin{cases} \{1\} & \text{if } i \in \mathcal{S} \\ [0,1] & \text{if } i \in \mathcal{E} \setminus \mathcal{S} \\ \{0\} & \text{if } i \notin \mathcal{E} \end{cases} \right\}.$$

Deep Learning with Implicit Layers



- Deep equilibrium networks [Bai, Kolter, Koltun 2019].
- Implicit networks [El Ghaoui, Gu, Travacca, Askari, Tsai 2019].
- Declarative networks [Gould, Hartley, Campbell 2019].
- Monotone deep equilibrium networks [Winston, Kolter 2020].
- Optimization layers (OptNET) [Amos, Kolter 2017].
- General convex optimization layers [Agrawal, Amos, Barratt, Boyd, Diamond, Kolter 2019].

Deep Learning with Implicit Layers



- Deep equilibrium networks [Bai, Kolter, Koltun 2019].
- Implicit networks [El Ghaoui, Gu, Travacca, Askari, Tsai 2019].
- Declarative networks [Gould, Hartley, Campbell 2019].
- Monotone deep equilibrium networks [Winston, Kolter 2020].
- Optimization layers (OptNET) [Amos, Kolter 2017].
- General convex optimization layers [Agrawal, Amos, Barratt, Boyd, Diamond, Kolter 2019].

conservative Jacobians $+$ path differentiable implicit function theorem

$\Longrightarrow$ convergence guarantees (every acc. point is a Clarke stationary point almost surely, objective values converge) for these network types.

Consider two parametric maximal monotone operators $\mathcal{A}_\theta$ and $\mathcal{B}_\theta$ and the inclusion

$$0 \in \mathcal{A}_\theta(x^\star) + \mathcal{B}_\theta(x^\star)$$

where $\mathcal{A}_\theta$ is set-valued but $\mathcal{B}_\theta$ is Lipschitz continuous. Note: We assume that $x^\star(\theta)$ is unique for each θ.

Generalizing to Monotone Inclusions/Generalized Equations

Consider two parametric maximal monotone operators $\mathcal{A}_\theta$ and $\mathcal{B}_\theta$ and the inclusion

$$0 \in \mathcal{A}_\theta(x^\star) + \mathcal{B}_\theta(x^\star)$$

where $\mathcal{A}_\theta$ is set-valued but $\mathcal{B}_\theta$ is Lipschitz continuous. Note: We assume that $x^\star(\theta)$ is unique for each θ. Fixed point equation for the monotone inclusion

$$\underbrace{R_{\gamma\mathcal{A}_\theta}(x - \gamma\mathcal{B}_\theta x)}_{H(\theta,x)} = x$$

We call H the Forward-Backward mapping. Formally we denote $T(\theta, x) := R_{\gamma\mathcal{A}_\theta}(x)$ and $S(\theta, x) := x - \gamma\mathcal{B}_\theta(x)$ the forward and backward maps which gives an equation we can apply the IFT to:

$$F(\theta, x) := H(\theta, x) - x = 0.$$

Consider two parametric maximal monotone operators $\mathcal{A}_\theta$ and $\mathcal{B}_\theta$ and the inclusion

$$0 \in \mathcal{A}_\theta(x^\star) + \mathcal{B}_\theta(x^\star)$$

where $\mathcal{A}_\theta$ is set-valued but $\mathcal{B}_\theta$ is Lipschitz continuous. Note: We assume that $x^\star(\theta)$ is unique for each θ. Fixed point equation for the monotone inclusion

$$\underbrace{R_{\gamma\mathcal{A}_\theta}(x - \gamma\mathcal{B}_\theta x)}_{H(\theta,x)} = x$$

We call H the Forward-Backward mapping. Formally we denote $T(\theta, x) := R_{\gamma\mathcal{A}_\theta}(x)$ and $S(\theta, x) := x - \gamma\mathcal{B}_\theta(x)$ the forward and backward maps which gives an equation we can apply the IFT to:

$$F(\theta, x) := H(\theta, x) - x = 0.$$

We will assume that F is path differentiable jointly in (θ, x).

Choice of Conservative Jacobian

Beware: conservative Jacobians are not unique and not defined pointwise!

$$\text{Example: path differentiable } f : \mathbb{R} \to \mathbb{R}, \quad \tilde{\mathcal{J}}_f(x) = \begin{cases} \mathcal{J}_f(x) \cup \{1\} & x \in \mathbb{N} \\ \mathcal{J}_f(x) & x \notin \mathbb{N} \end{cases}$$

Choice of Conservative Jacobian

Beware: conservative Jacobians are not unique and not defined pointwise!

$$\text{Example: path differentiable } f : \mathbb{R} \to \mathbb{R}, \quad \tilde{\mathcal{J}}_f(x) = \begin{cases} \mathcal{J}_f(x) \cup \{1\} & x \in \mathbb{N} \\ \mathcal{J}_f(x) & x \notin \mathbb{N} \end{cases}$$

We must make a choice for which conservative Jacobian to consider - it should reflect what is computed in practice and also be theoretically accessible.

Choice of Conservative Jacobian

Beware: conservative Jacobians are not unique and not defined pointwise!

$$\text{Example: path differentiable } f : \mathbb{R} \to \mathbb{R}, \quad \tilde{\mathcal{J}}_f(x) = \begin{cases} \mathcal{J}_f(x) \cup \{1\} & x \in \mathbb{N} \\ \mathcal{J}_f(x) & x \notin \mathbb{N} \end{cases}$$

We must make a choice for which conservative Jacobian to consider - it should reflect what is computed in practice and also be theoretically accessible.

We take the product of Clarke Jacobians of the forward and backward maps giving

$$\begin{aligned}
\mathcal{J}_{H_\theta}(\theta, x) &= \mathrm{Jac}^c_{\mathcal{T}}(S(\theta, x)) \times \mathrm{Jac}^c_{S}(\theta, x) \\
&= \Big\{ \begin{bmatrix} A & B \end{bmatrix} \times \begin{bmatrix} \mathrm{Id}_p & 0 \\ -C & \mathrm{Id}_n - \gamma D \end{bmatrix} : [A \ B] \in \mathrm{Jac}^c_{\mathcal{T}}(\theta, x - \gamma \mathcal{B}_\theta(x)), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad [C \ D] \in \mathrm{Jac}^c_{\mathcal{B}}(\theta, x) \Big\} \\
&= \Big\{ \begin{bmatrix} A - BC & B(\mathrm{Id}_n - \gamma D) \end{bmatrix} : [A \ B] \in \mathrm{Jac}^c_{\mathcal{T}}(\theta, x - \gamma \mathcal{B}_\theta(x)), [C \ D] \in \mathrm{Jac}^c_{\mathcal{B}}(\theta, x) \Big\}
\end{aligned}$$

which is a conservative Jacobian for H.

Strong Monotonicity is All You Need™

Theorem (Bolte, Pauwels, S.F. 2024)

Assume that $\mathcal{B}_\theta$ is β-Lipschitz continuous and that either $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone, for some $\alpha, \beta > 0$, uniformly in θ. For $\gamma \in (0, \frac{2\alpha}{(\alpha+\beta)^2})$, the invertibility condition holds and $x^\star$ is path differentiable with a conservative Jacobian whose formula is computable from $\mathcal{J}_H(x^\star(\theta))$.

Strong Monotonicity is All You Need™

Theorem (Bolte, Pauwels, S.F. 2024)

Assume that $\mathcal{B}_\theta$ is β-Lipschitz continuous and that either $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone, for some $\alpha, \beta > 0$, uniformly in θ. For $\gamma \in (0, \frac{2\alpha}{(\alpha+\beta)^2})$, the invertibility condition holds and $x^\star$ is path differentiable with a conservative Jacobian whose formula is computable from $\mathcal{J}_H(x^\star(\theta))$.

Proof.

If $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone, then either T or S is a strict contraction, and we can choose γ to ensure that the composition H is a strict contraction. Then, the product of Clarke Jacobians will have norm < 1. □

Strong Monotonicity is All You Need™

Theorem (Bolte, Pauwels, S.F. 2024)

Assume that $\mathcal{B}_\theta$ is β-Lipschitz continuous and that either $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone, for some $\alpha, \beta > 0$, uniformly in θ. For $\gamma \in (0, \frac{2\alpha}{(\alpha+\beta)^2})$, the invertibility condition holds and $x^\star$ is path differentiable with a conservative Jacobian whose formula is computable from $\mathcal{J}_H(x^\star(\theta))$.

Proof.

If $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone, then either T or S is a strict contraction, and we can choose γ to ensure that the composition H is a strict contraction. Then, the product of Clarke Jacobians will have norm < 1. □

Corollary

The solution to the optimization problem

$$\min_{x \in \mathbb{R}^p} f_\theta(x) + g_\theta(x),$$

where f_θ is β-Lipschitz smooth and g_θ is nonsmooth, is path differentiable if either f_θ or g_θ is α-strongly convex.

Plan

- Motivation
- Conservative Gradients
- Results
- Applications
- **Numerical Examples**

Piecewise quadratic objective function posed as a bilevel problem:

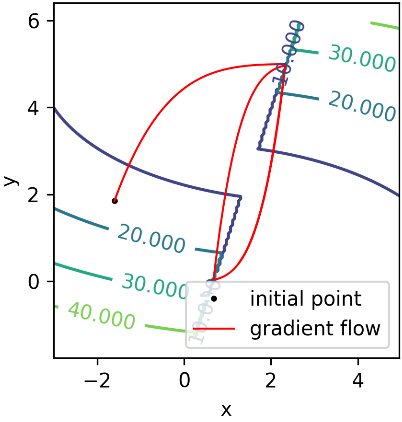
$$\min_{x,y,s} \quad (x - s_1)^2 + 4\,(y - s_2)^2 \qquad \text{such that}$$

$$s \in \arg\max \left\{ \, (a + b)\,(-2x + y + 2) : a \in [0, 3], b \in [0, 5] \, \right\}$$

Piecewise quadratic objective function posed as a bilevel problem:

$$\min_{x,y,s} \quad (x - s_1)^2 + 4\,(y - s_2)^2 \qquad \text{such that}$$

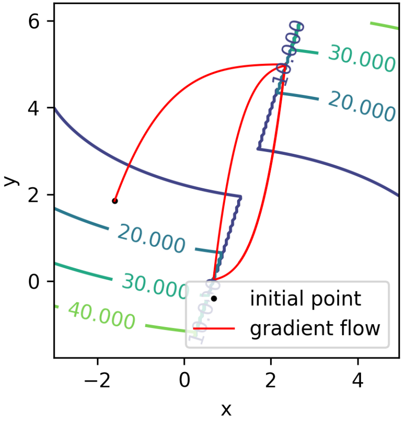
$$s \in \arg\max \left\{ (a+b)\,(-2x+y+2) : a \in [0,3], b \in [0,5] \right\}$$

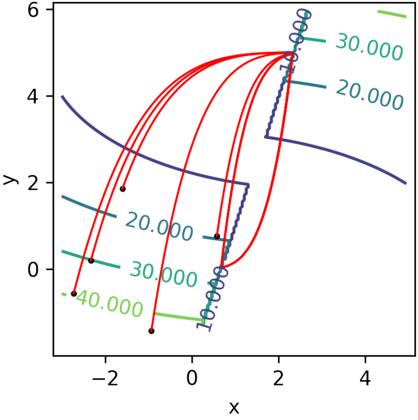


Piecewise quadratic objective function posed as a bilevel problem:

$$\min_{x,y,s} \quad (x - s_1)^2 + 4\,(y - s_2)^2 \qquad \text{such that}$$

$$s \in \arg\max \left\{ \, (a+b)\,(-2x + y + 2) : a \in [0,3], b \in [0,5] \, \right\}$$

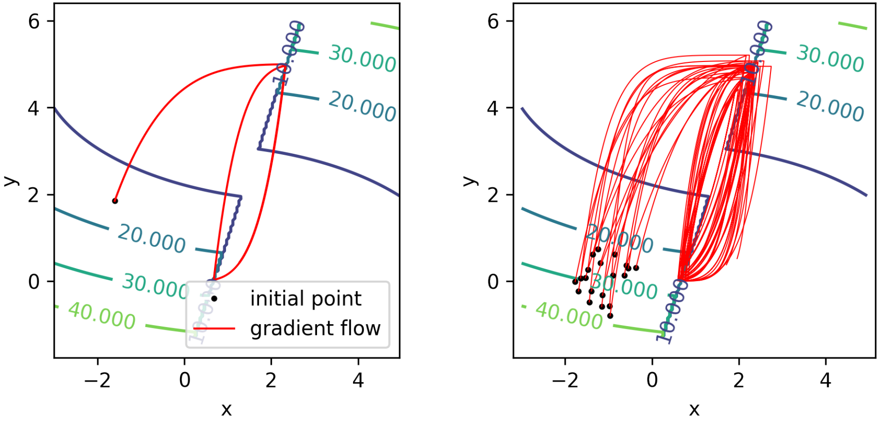




Piecewise quadratic objective function posed as a bilevel problem:

$$\min_{x,y,s} \quad (x - s_1)^2 + 4\,(y - s_2)^2 \qquad \text{such that}$$

$$s \in \arg\max \left\{ (a+b)\,(-2x + y + 2) : a \in [0,3], b \in [0,5] \right\}$$



Pathological Examples - Optimizing a Quadratic Two Ways

$$\text{Let } L(u) = L(x, y, z) = \left(10(y-x), x(28-z)-y, xy-\tfrac{8}{3}z\right)$$

Explicit formulation

$$\max_{u \in \mathbb{R}^3} \quad u^T L(u)$$

$$\iff$$

Implicit formulation

$$\max_{u \in \mathbb{R}^3} \quad u^T z \quad \text{such that}$$

$$z \in \operatorname*{argmin}_{s \in \mathbb{R}^3} \|s - L(u)\|^4$$

Let $L(u) = L(x, y, z) = \left(10(y - x), x(28 - z) - y, xy - \frac{8}{3}z\right)$

Explicit formulation

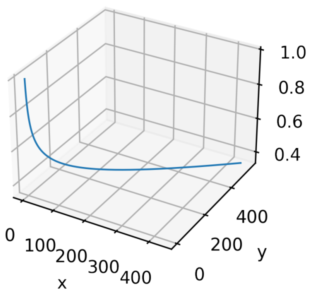
$$\max_{u \in \mathbb{R}^3} \quad u^T L(u)$$

$$\iff$$

Implicit formulation

$$\max_{u \in \mathbb{R}^3} \quad u^T z \quad \text{such that}$$

$$z \in \operatorname*{argmin}_{s \in \mathbb{R}^3} \|s - L(u)\|^4$$



Let $L(u) = L(x, y, z) = \left(10(y - x), x(28 - z) - y, xy - \frac{8}{3}z\right)$

Explicit formulation

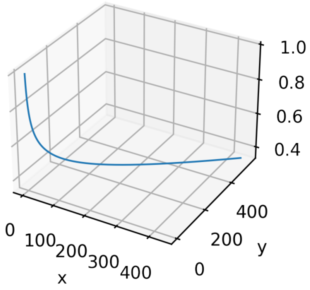
$$\max_{u \in \mathbb{R}^3} \quad u^T L(u)$$

$\iff$

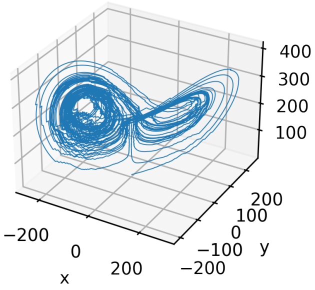
Implicit formulation

$$\max_{u \in \mathbb{R}^3} \quad u^T z \quad \text{such that}$$

$$z \in \operatorname{argmin}_{s \in \mathbb{R}^3} \|s - L(u)\|^4$$



$\not\Leftrightarrow$



Nonsmooth implicit differentiation:

- Can we have a calculus for implicitly defined functions in the locally Lipschitz setting?

Nonsmooth implicit differentiation:

- Can we have a calculus for implicitly defined functions in the locally Lipschitz setting?
- Using the Clarke subdifferential: No.
  - The inverse of a Clarke Jacobian is not necessarily a Clarke Jacobian.

Nonsmooth implicit differentiation:

- Can we have a calculus for implicitly defined functions in the locally Lipschitz setting?
- Using the Clarke subdifferential: No.
  - The inverse of a Clarke Jacobian is not necessarily a Clarke Jacobian.
- Using conservative gradients Yes.
  - The inverse of a conservative Jacobian is again a conservative Jacobian.

Nonsmooth implicit differentiation:

- Can we have a calculus for implicitly defined functions in the locally Lipschitz setting?
- Using the Clarke subdifferential: No.
  - The inverse of a Clarke Jacobian is not necessarily a Clarke Jacobian.
- Using conservative gradients Yes.
  - The inverse of a conservative Jacobian is again a conservative Jacobian.

Practical implications:

- Method to compute the gradient of solutions to convex optimization problems.
- Applications in machine learning (bilevel hyperparameter tuning, implicit neural networks, . . .).

Nonsmooth Implicit Differentiation for Machine Learning (NeurIPS, 2021)

Jérôme Bolte, Tâm Lê, Edouard Pauwels, Antonio Silveti-Falls

https://arxiv.org/abs/2106.04350

Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems (SIAM Optimization, 2024)

Jérôme Bolte, Edouard Pauwels, Antonio Silveti-Falls

https://arxiv.org/abs/2212.07844

N data points, L layers:

$$\min_{w \in \mathbb{R}^p} \ell(w) := \frac{1}{N} \sum_{i=1}^{N} \ell_i(w) \quad \text{with} \quad \ell_i := g_{i,L} \circ g_{i,L-1} \circ \ldots \circ g_{i,1}$$

Each layer $g_{i,j}$ is semialgebraic (or definable) and path differentiable - can be explicit or implicit.
$N = 1, L = 2$ recovers bilevel optimization problem setting.
Define

$$w_{k+1} = w_k - s\alpha_k v_k \qquad v_k \in J_{I_k}(w_k)$$

for $(\alpha_k)_{k \in \mathbb{N}} \in \ell^1 \setminus \ell^2$
For almost all w_0, for almost all $s \in (s_{\min}, s_{\max})$, $\ell(w_k)$ converges and all acc. points of $(w_k)_{k \in \mathbb{N}}$ are clarke critical.