Nonsmooth implicit differentiation for optimization

Edouard Pauwels (IRIT, Toulouse 3, France)

joint work with

Jérôme Bolte, Tâm Lê, Antonio Silveti-Falls (TSE, Toulouse 1, France)

OWOS seminar (September 2021)

Observations:

- The classical implicit function theorem has two parts (existence and calculus)
- Nonsmooth generalizations essentially focused on existence.

Observations:

- The classical implicit function theorem has two parts (existence and calculus)
- Nonsmooth generalizations essentially focused on existence.

Contributions: nonsmooth generalization of the calculus part.

Observations:

- The classical implicit function theorem has two parts (existence and calculus)
- Nonsmooth generalizations essentially focused on existence.

Contributions: nonsmooth generalization of the calculus part.

- Direct generalization of calculus fails.

Observations:

- The classical implicit function theorem has two parts (existence and calculus)
- Nonsmooth generalizations essentially focused on existence.

Contributions: nonsmooth generalization of the calculus part.

- Direct generalization of calculus fails.
- Our solution: use conservative Jacobians.

Observations:

- The classical implicit function theorem has two parts (existence and calculus)
- Nonsmooth generalizations essentially focused on existence.

Contributions: nonsmooth generalization of the calculus part.

- Direct generalization of calculus fails.
- Our solution: use conservative Jacobians.
- Applications in compositional modeling (ML, DEQ), bilevel optimization, . . .

Plan

Plan

Classical implicit function theorem (Dini 1877)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with Jacobian $\mathrm{Jac}_F(x, y) = [A_x \ B_y] \in \mathbb{R}^{m\times(n+m)}$ and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

If $B_{\bar{y}}$ is invertible, then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a differentiable function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

$$\mathrm{Jac}_G(x) = -B^{-1}A, \qquad [A \ B] = \mathrm{Jac}_F(x, G(x)).$$

Classical implicit function theorem (Dini 1877)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with Jacobian $\mathrm{Jac}_F(x, y) = [A_x \ B_y] \in \mathbb{R}^{m\times(n+m)}$ and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

If $B_{\bar{y}}$ is invertible, then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a differentiable function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

$$\mathrm{Jac}_G(x) = -B^{-1}A, \qquad [A\ B] = \mathrm{Jac}_F(x, G(x)).$$

Classical implicit function theorem (Dini 1877)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with Jacobian $\mathrm{Jac}_F(x, y) = [A_x \; B_y] \in \mathbb{R}^{m\times(n+m)}$ and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

If $B_{\bar{y}}$ is invertible, then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a differentiable function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

$$\mathrm{Jac}_G(x) = -B^{-1}A, \qquad [A \; B] = \mathrm{Jac}_F(x, G(x)).$$

- **Existence:** Equation $F(x, y) = 0$ defines a functional relation $y = G(x)$ around $\bar{x}$.

Classical implicit function theorem (Dini 1877)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with Jacobian $\mathrm{Jac}_F(x,y) = [A_x \; B_y] \in \mathbb{R}^{m\times(n+m)}$ and $(\bar{x},\bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

If $B_{\bar{y}}$ is invertible, then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a differentiable function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

$$\mathrm{Jac}_G(x) = -B^{-1}A, \qquad [A \; B] = \mathrm{Jac}_F(x, G(x)).$$

- **Existence:** Equation $F(x,y) = 0$ defines a functional relation $y = G(x)$ around $\bar{x}$.

Classical implicit function theorem (Dini 1877)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with Jacobian $\mathrm{Jac}_F(x, y) = [A_x \ B_y] \in \mathbb{R}^{m\times(n+m)}$ and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

If $B_{\bar{y}}$ is invertible, then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a differentiable function $G(x)$ so that

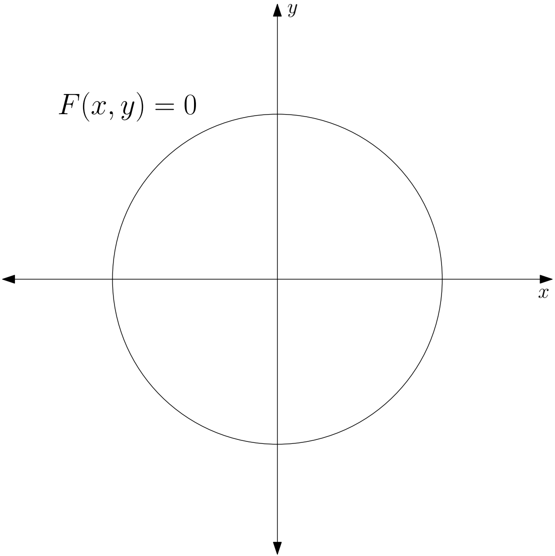
$$\forall x \in U \qquad F(x, G(x)) = 0,$$

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

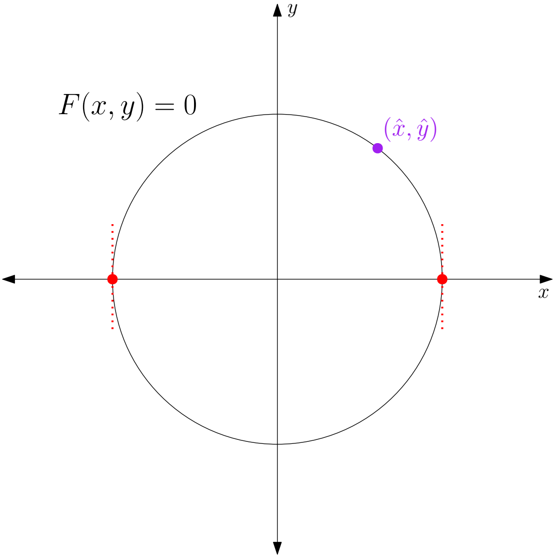
$$\mathrm{Jac}_G(x) = -B^{-1}A, \qquad [A \ B] = \mathrm{Jac}_F(x, G(x)).$$

- **Existence:** Equation $F(x, y) = 0$ defines a functional relation $y = G(x)$ around $\bar{x}$.
- **Implicit differentiation:** Calculus rule for the derivative of G.

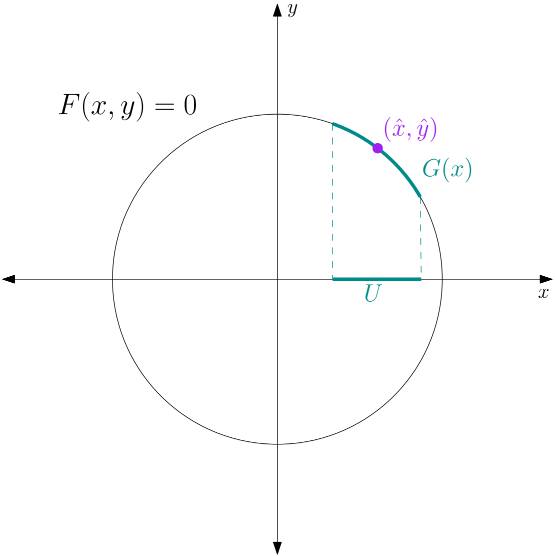
$F(x, y) = x^2 + y^2 - 1.$



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



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



Classical inverse mapping theorem

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable with Jacobian $\mathrm{Jac}_F : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $\mathrm{Jac}_F(\bar{x})$ is nonsingular. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a diffeomorphism. For all $x \in U$,

$$\mathrm{Jac}_{F^{-1}}(F(x)) = \mathrm{Jac}_F(x)^{-1}.$$

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable with Jacobian $\mathrm{Jac}_F : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $\mathrm{Jac}_F(\bar{x})$ is nonsingular. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a diffeomorphism. For all $x \in U$,

$$\mathrm{Jac}_{F^{-1}}(F(x)) = \mathrm{Jac}_F(x)^{-1}.$$

Classical inverse mapping theorem

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable with Jacobian $\mathrm{Jac}_F : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $\mathrm{Jac}_F(\bar{x})$ is nonsingular. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a diffeomorphism. For all $x \in U$,

$$\mathrm{Jac}_{F^{-1}}(F(x)) = \mathrm{Jac}_F(x)^{-1}.$$

- Existence of a functional inverse for F around $\bar{x}$.

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable with Jacobian $\mathrm{Jac}_F : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $\mathrm{Jac}_F(\bar{x})$ is nonsingular. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a diffeomorphism. For all $x \in U$,

$$\mathrm{Jac}_{F^{-1}}(F(x)) = \mathrm{Jac}_F(x)^{-1}.$$

- Existence of a functional inverse for F around $\bar{x}$.

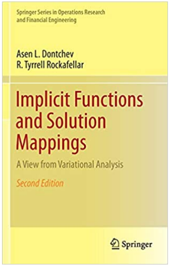
Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable with Jacobian $\mathrm{Jac}_F : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $\mathrm{Jac}_F(\bar{x})$ is nonsingular. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a diffeomorphism. For all $x \in U$,

$$\mathrm{Jac}_{F^{-1}}(F(x)) = \mathrm{Jac}_F(x)^{-1}.$$

- Existence of a functional inverse for F around $\bar{x}$.
- Calculus rule for the derivative of F^{-1}.

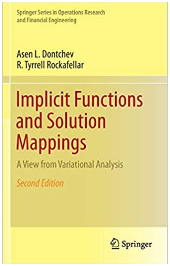
- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.

- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.



In nonsmooth analysis:

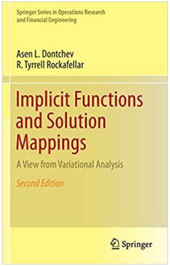
- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.



In nonsmooth analysis:

- **Strict differentiability:** Leach (1961), Nijenhuis (1974).

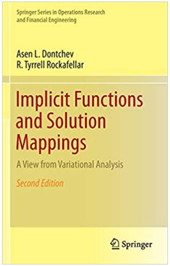
- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.



In nonsmooth analysis:

- **Strict differentiability:** Leach (1961), Nijenhuis (1974).
- **Inclusions, set valued:** Robinson (1980), Dontchev-Rockafellar (2009).

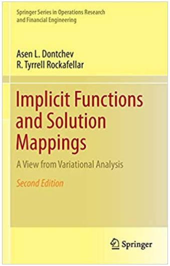
- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.



In nonsmooth analysis:

- **Strict differentiability:** Leach (1961), Nijenhuis (1974).
- **Inclusions, set valued:** Robinson (1980), Dontchev-Rockafellar (2009).
- **Inverse, set valued:** Aubin (1982), Rockafellar (1985), Aubin-Frankowka (1984), Dontchev-Hager (1994).

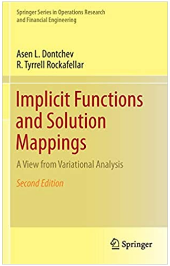
- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.



In nonsmooth analysis:

- **Strict differentiability:** Leach (1961), Nijenhuis (1974).
- **Inclusions, set valued:** Robinson (1980), Dontchev-Rockafellar (2009).
- **Inverse, set valued:** Aubin (1982), Rockafellar (1985), Aubin-Frankowka (1984), Dontchev-Hager (1994).
- **Locally Lipschitz equations:** Clarke (1976), Hiriart Urruty (1979), Clarke (1983).

- $F(x, y) = 0$.
- Euclidean space.
- Continuously differentiable.
- Block invertible Jacobian.



In nonsmooth analysis:

- **Strict differentiability:** Leach (1961), Nijenhuis (1974).
- **Inclusions, set valued:** Robinson (1980), Dontchev-Rockafellar (2009).
- **Inverse, set valued:** Aubin (1982), Rockafellar (1985), Aubin-Frankowka (1984), Dontchev-Hager (1994).
- **Locally Lipschitz equations:** Clarke (1976), Hiriart Urruty (1979), Clarke (1983).
-
    - Robinson (1991) directional derivatives with calculus (restricted subclass).
    - Sun (2001), semismoothness.
    - Fukui, Kurdyka, Paunescu (2007), subanalytic / tame.

Implicit function theorem:

- **Existence:** Locally implicitely defined functional relation.
- **Calculus:** Jacobians from matrix inversion.

Implicit function theorem:

- **Existence:** Locally implicitely defined functional relation.
- **Calculus:** Jacobians from matrix inversion.

Context of this presentation:

- **Lipschitz equations:** possibly nonsmooth, finite dimension.
- **Implicit differentiation:** Calculus part

Implicit function theorem:

- **Existence:** Locally implicitely defined functional relation.
- **Calculus:** Jacobians from matrix inversion.

Context of this presentation:

- **Lipschitz equations:** possibly nonsmooth, finite dimension.
- **Implicit differentiation:** Calculus part

Motivation and applications:

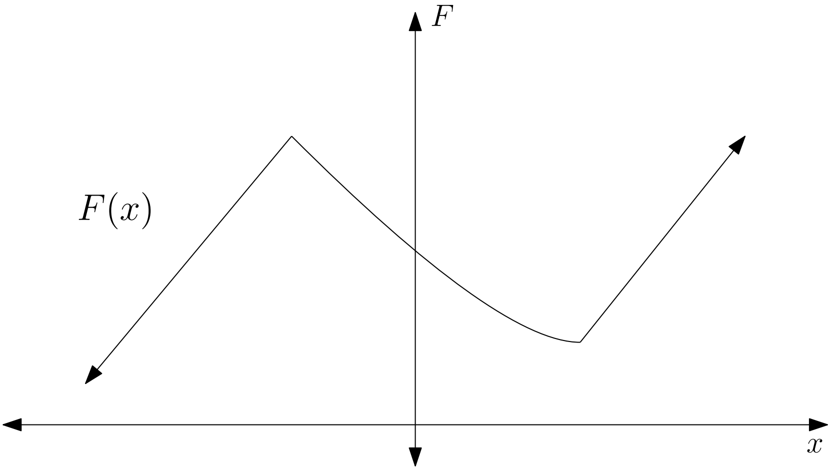
- Generalizations focused on the existence / regularity part.
- Applications:
  - Bilevel optimization: differentiate solutions of optimization problems.
  - Implicit compositional modeling: equilibrium models, declarative networks . . .

Plan

Clarke's generalized derivatives: Given a locally Lipschitz function $F : \mathbb{R}^n \to \mathbb{R}^m$, the Clarke Jacobian at a point $x \in \mathbb{R}^n$ is

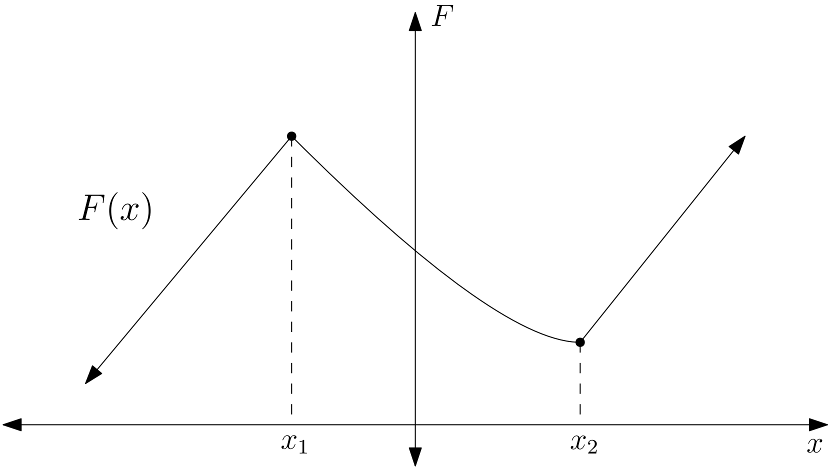
$$J_F^c(x) = \mathrm{conv}\left(\left\{\lim_{k\to\infty} \mathrm{Jac}\,_F(x_k) : x_k \in \mathrm{diff}_F \text{ and } x_k \to x\right\}\right),$$

where diff_F is the set of differentiability point of F (Rademacher: full measure).

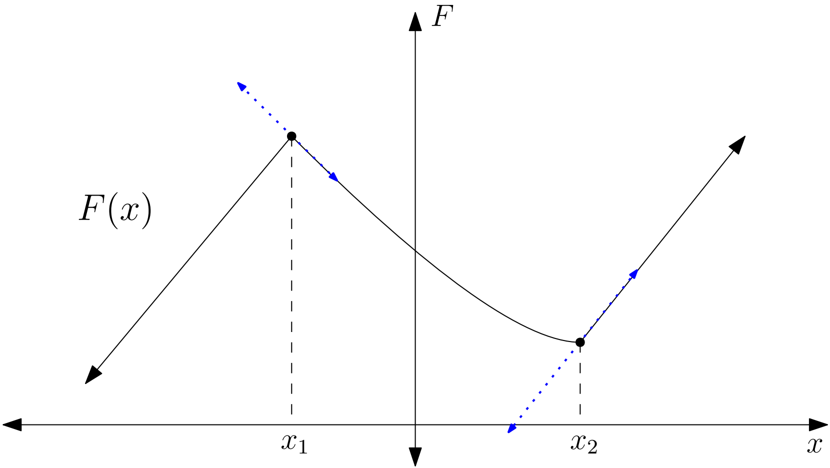


$$J_F^c(x) = \mathrm{conv}\left(\left\{\lim_{k\to\infty} \mathrm{Jac}\, _F(x_k) : x_k \in \mathrm{diff}_F \text{ and } x_k \to x\right\}\right).$$

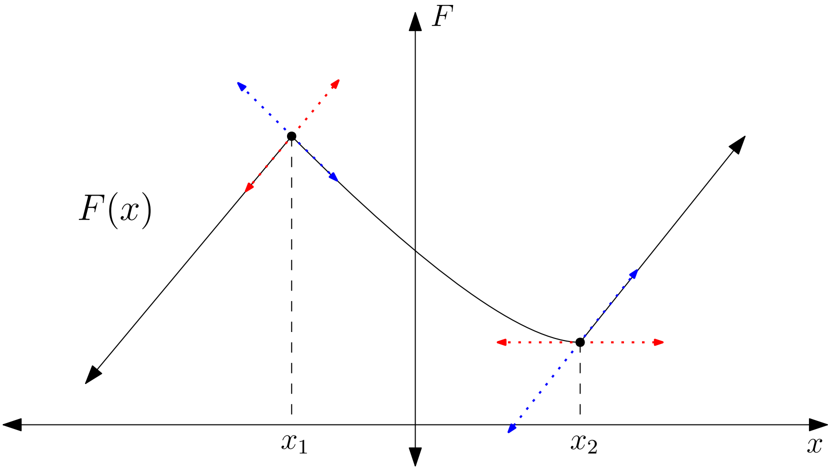
Generalized derivative



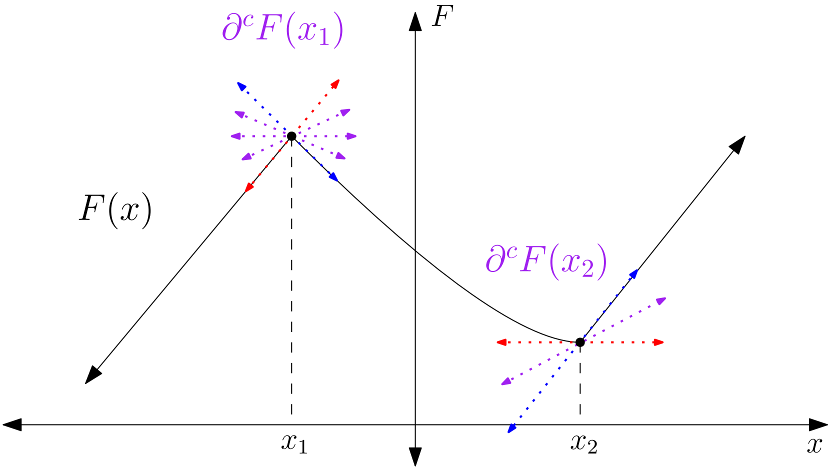
$$J_F^c(x) = \operatorname{conv}\left(\left\{\lim_{k\to\infty} \operatorname{Jac}_F(x_k) : x_k \in \operatorname{diff}_F \text{ and } x_k \to x\right\}\right).$$



$$J_F^c(x) = \operatorname{conv}\left(\left\{\lim_{k\to\infty} \operatorname{Jac}_F(x_k) : x_k \in \operatorname{diff}_F \text{ and } x_k \to x\right\}\right).$$



$$J_F^c(x) = \operatorname{conv}\left(\left\{\lim_{k\to\infty} \operatorname{Jac}_F(x_k) : x_k \in \operatorname{diff}_F \text{ and } x_k \to x\right\}\right).$$



$$J_F^c(x) = \mathrm{conv}\left(\left\{\lim_{k\to\infty} \mathrm{Jac}\,_F(x_k) : x_k \in \mathrm{diff}_F \text{ and } x_k \to x\right\}\right).$$

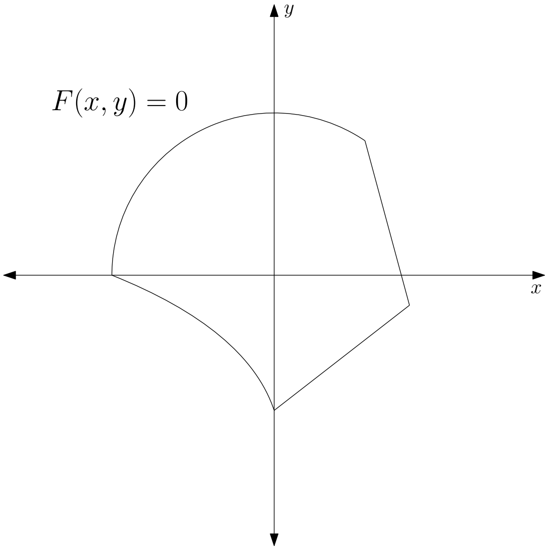
Let $F:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^m$ be locally Lipschitz and $(\bar{x},\bar{y})\in\mathbb{R}^n\times\mathbb{R}^m$ such that

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

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

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

and $y=G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.



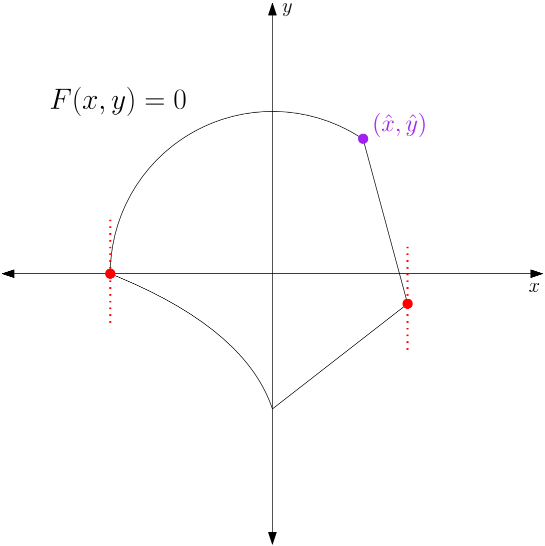
Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

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

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.



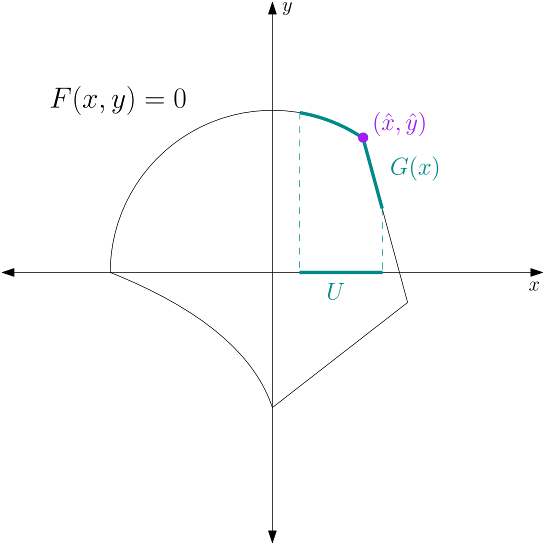
Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

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

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.



Lipschitz implicit function theorem (Hiriart Urruty 1979, Clarke 1976)

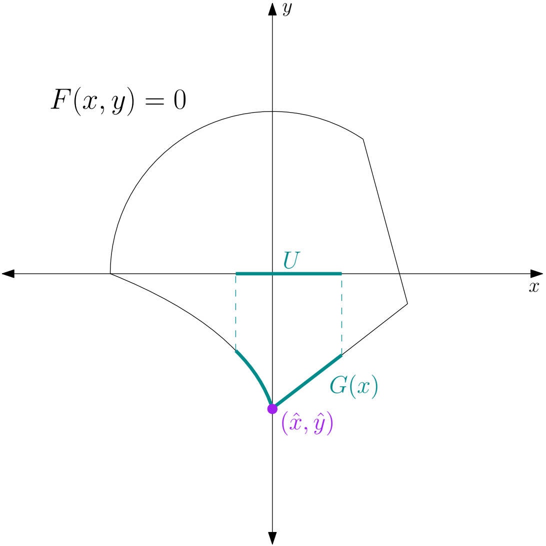
Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

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

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

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.



Clarke's inverse mapping theorem: Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz with Clarke Jacobian $J_F^c \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n\times n}$ only contains nonsingular matrices. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Formal inverse differentiation? For all $x \in U$,

$$J_{F^{-1}}^c(F(x)) = J_F^c(x)^{-1} := \left\{ M^{-1},\ M \in J_F^c(x) \right\} ?$$

Clarke's inverse mapping theorem: Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz with Clarke Jacobian $J_F^c : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n \times n}$ only contains nonsingular matrices. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Formal inverse differentiation? For all $x \in U$,

$$J_{F^{-1}}^c(F(x)) = J_F^c(x)^{-1} := \left\{ M^{-1},\ M \in J_F^c(x) \right\} ?$$

Clarke's inverse mapping theorem: Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz with Clarke Jacobian $J_F^c : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n\times n}$ only contains nonsingular matrices. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Formal inverse differentiation? For all $x \in U$,

$$J^c_{F^{-1}}(F(x)) = J^c_F(x)^{-1} := \left\{M^{-1},\, M \in J^c_F(x)\right\} ?$$

Clarke's inverse mapping theorem: Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz with Clarke Jacobian $J_F^c \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ and $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n\times n}$ only contains nonsingular matrices. Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Formal inverse differentiation? For all $x \in U$,

$$J_{F^{-1}}^c(F(x)) = J_F^c(x)^{-1} := \left\{M^{-1},\, M \in J_F^c(x)\right\}?$$

Failure of formal implicit differentiation

From Clarke's book: consider the function $F\colon \mathbb{R}^2 \to \mathbb{R}^2$

$$F\colon \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} |x| + y \\ 2x + |y| \end{pmatrix}$$

- $F(0) = 0$
- $J^c F(0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 2 & \beta \end{pmatrix}, \alpha, \beta \in [-1, 1] \right\}$
- Complies with hypotheses of Clarke's inverse mapping theorem

Failure of formal implicit differentiation

From Clarke's book: consider the function $F\colon \mathbb{R}^2 \to \mathbb{R}^2$

$$F\colon \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} |x| + y \\ 2x + |y| \end{pmatrix}$$

- $F(0) = 0$
- $J^c F(0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 2 & \beta \end{pmatrix}, \alpha, \beta \in [-1, 1] \right\}$
- Complies with hypotheses of Clarke's inverse mapping theorem

Failure of formal implicit differentiation

From Clarke's book: consider the function $F\colon \mathbb{R}^2 \to \mathbb{R}^2$

$$F\colon \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} |x| + y \\ 2x + |y| \end{pmatrix}$$

- $F(0) = 0$
- $J^c F(0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 2 & \beta \end{pmatrix}, \alpha, \beta \in [-1, 1] \right\}$
- Complies with hypotheses of Clarke's inverse mapping theorem

Failure of formal implicit differentiation

From Clarke's book: consider the function $F\colon \mathbb{R}^2 \to \mathbb{R}^2$

$$F\colon \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} |x| + y \\ 2x + |y| \end{pmatrix}$$

- $F(0) = 0$
- $J^c F(0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 2 & \beta \end{pmatrix}, \alpha, \beta \in [-1, 1] \right\}$
- Complies with hypotheses of Clarke's inverse mapping theorem

Failure of formal implicit differentiation

From Clarke's book: consider the function $F \colon \mathbb{R}^2 \to \mathbb{R}^2$

$$F \colon \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} |x| + y \\ 2x + |y| \end{pmatrix}$$

- $F(0) = 0$
- $J^c F(0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 2 & \beta \end{pmatrix}, \alpha, \beta \in [-1, 1] \right\}$
- Complies with hypotheses of Clarke's inverse mapping theorem

Failure of Jacobian inversion rule:

- $\dim(J^c_F(0)) = 2$
- $\dim(J^c_{F^{-1}}(0)) = 3$
- There exists $M \in J^c_{F^{-1}}(0)$ such that $M^{-1} \notin J^c_F(0)$

From Clarke's book: consider the function $F\colon \mathbb{R}^2 \to \mathbb{R}^2$

$$F\colon \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} |x| + y \\ 2x + |y| \end{pmatrix}$$

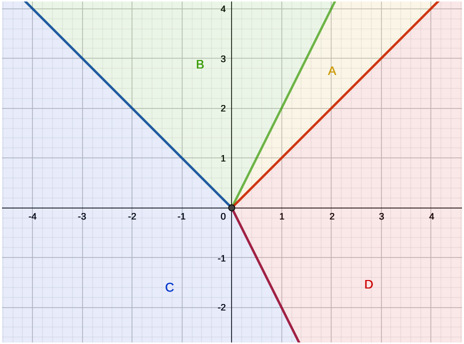
- $F(0) = 0$
- $J^c F(0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 2 & \beta \end{pmatrix}, \alpha, \beta \in [-1, 1] \right\}$
- Complies with hypotheses of Clarke's inverse mapping theorem

Failure of Jacobian inversion rule:

- $\dim(J^c_F(0)) = 2$
- $\dim(J^c_{F^{-1}}(0)) = 3$
- There exists $M \in J^c_{F^{-1}}(0)$ such that $M^{-1} \notin J^c_F(0)$

Explicit piecewise affine inverse.

$$
\begin{aligned}
F^{-1}(u,v) &= (v-u, 2u-v) && \text{for } (u,v)\in A,\\
F^{-1}(u,v) &= \frac{1}{3}(u+v, 2u-v) && \text{for } (u,v)\in B,\\
F^{-1}(u,v) &= (u+v, 2u+v) && \text{for } (u,v)\in C,\\
F^{-1}(u,v) &= \frac{1}{3}(v-u, 2u+v) && \text{for } (u,v)\in D,
\end{aligned}
$$



$F\colon \mathbb{R}^2 \to \mathbb{R}^2$ complies with Clarke's inverse mapping theorem.

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

Plan

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F : \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$.
- Compatible with compositional calculus rules
  - $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ conservative for $F : \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G : \mathbb{R}^m \rightrightarrows \mathbb{R}^{p \times m}$ conservative for $G : \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

In a nutshell

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$.
- Compatible with compositional calculus rules
  - $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ conservative for $F\colon \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{p\times m}$ conservative for $G\colon \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

In a nutshell

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F \colon \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$.
- Compatible with compositional calculus rules
  - $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ conservative for $F \colon \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{p \times m}$ conservative for $G \colon \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F \colon \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$.
- Compatible with compositional calculus rules
  - $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ conservative for $F \colon \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{p \times m}$ conservative for $G \colon \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F \colon \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$.
- Compatible with compositional calculus rules
  - $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ conservative for $F \colon \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{p \times m}$ conservative for $G \colon \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

In a nutshell

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$.
- Compatible with compositional calculus rules
  - $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ conservative for $F\colon \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{p\times m}$ conservative for $G\colon \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

In a nutshell

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- A given function $F : \mathbb{R}^n \to \mathbb{R}^m$ has multiple conservative Jacobians $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$.
- Compatible with compositional calculus rules
  - $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ conservative for $F : \mathbb{R}^n \to \mathbb{R}^m$.
  - $J_G : \mathbb{R}^m \rightrightarrows \mathbb{R}^{p\times m}$ conservative for $G : \mathbb{R}^m \to \mathbb{R}^p$.
  - Then $x \rightrightarrows J_G(F(x)) \times J_F(x)$ is conservative for $G \circ F$.
  - Sum rule, product rule, ...
- Conservative gradients have a minimizing behavior similar to subgradients in optimization.

Bibliography:

- **Introduction / nonsmooth algorithmic differentiation:** Bolte-Pauwels (2020).
- **Lazy gradient oracle:** Bianchi-Hachem-Schechtman (2020).
- **Structure / residual:** Lewis-Tian (2021).
- **Semi-smoothness:** Davis-Drusvyatskiy (2021).

$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz,

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Descent mechanism: chain rule along Lipschitz curves

$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Descent mechanism: chain rule along Lipschitz curves

$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Chain rule along Absolutely Continuous (AC) curves (Brézis, Valadier).
Hypothesis: For any AC curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \partial^c f(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1]$$

$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Chain rule along Absolutely Continuous (AC) curves (Brézis, Valadier).
Hypothesis: For any AC curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \partial^c f(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1]$$

Suppose: $\dot{\gamma}(t) \in -\partial^c f(\gamma(t))$ for almost all $t \in [0,1]$,

$f \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Chain rule along Absolutely Continuous (AC) curves (Brézis, Valadier).
Hypothesis: For any AC curve $\gamma \colon [0, 1] \mapsto \mathbb{R}^p$

$$\begin{aligned} \frac{d}{dt} f(\gamma(t)) &= \langle v, \dot\gamma(t)\rangle \qquad \forall v \in \partial^c f(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1] \\ &= -\|\dot\gamma(t)\|^2, \qquad \text{a.e.} \quad t \in [0,1] \end{aligned}$$

Suppose: $\dot\gamma(t) \in -\partial^c f(\gamma(t))$ for almost all $t \in [0, 1]$,

Descent mechanism: chain rule along Lipschitz curves

$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Chain rule along Absolutely Continuous (AC) curves (Brézis, Valadier).
Hypothesis: For any AC curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$

$$\begin{aligned} \frac{d}{dt} f(\gamma(t)) &= \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \partial^c f(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1] \\ &= -\|\dot{\gamma}(t)\|^2, \qquad \text{a.e.} \quad t \in [0,1] \end{aligned}$$

Suppose: $\dot{\gamma}(t) \in -\partial^c f(\gamma(t))$ for almost all $t \in [0,1]$,
then $t \mapsto f(\gamma(t))$ decreases, strictly if $0 \notin \partial^c f(\gamma(t))$.

Descent mechanism: chain rule along Lipschitz curves

$f \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Chain rule along Absolutely Continuous (AC) curves (Brézis, Valadier).
Hypothesis: For any AC curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$

$$\begin{aligned} \frac{d}{dt} f(\gamma(t)) &= \langle v, \dot\gamma(t) \rangle \qquad \forall v \in \partial^c f(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1] \\ &= -\|\dot\gamma(t)\|^2, \qquad \text{a.e.} \quad t \in [0,1] \end{aligned}$$

Suppose: $\dot\gamma(t) \in -\partial^c f(\gamma(t))$ for almost all $t \in [0,1]$,
then $t \mapsto f(\gamma(t))$ decreases, strictly if $0 \notin \partial^c f(\gamma(t))$.

Benaim-Haufbauer-Sorin (2005) subgradient plus zero mean noise, under proper assumptions:

Descent mechanism: chain rule along Lipschitz curves

$f \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k v_k \\ v_k &\in \partial^c f(\theta_k). \end{aligned} \qquad \Leftrightarrow \qquad \frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -\partial^c f(\theta_k)$$

Chain rule along Absolutely Continuous (AC) curves (Brézis, Valadier).
Hypothesis: For any AC curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$

$$\begin{aligned} \frac{d}{dt} f(\gamma(t)) &= \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \partial^c f(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1] \\ &= -\|\dot{\gamma}(t)\|^2, \qquad \text{a.e.} \quad t \in [0,1] \end{aligned}$$

Suppose: $\dot{\gamma}(t) \in -\partial^c f(\gamma(t))$ for almost all $t \in [0,1]$,
then $t \mapsto f(\gamma(t))$ decreases, strictly if $0 \notin \partial^c f(\gamma(t))$.

Benaim-Haufbauer-Sorin (2005) subgradient plus zero mean noise, under proper assumptions:

Vanishing step sizes, almost surely all accumulation points are critical points: $0 \in \partial^c f(\bar{\theta})$.

Generic triviality, generic rigidity

Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).

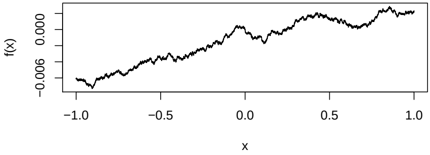
Generic triviality, generic rigidity

Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.

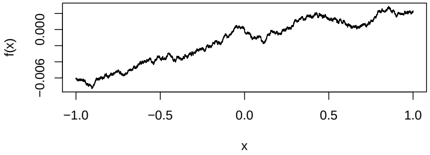
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.

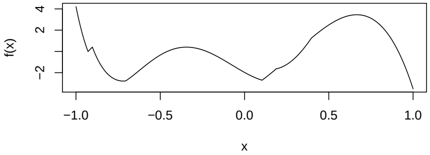


Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.

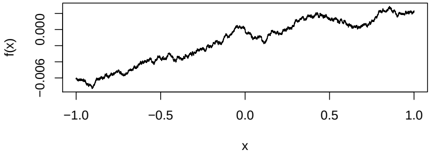


Let f be a *tame* locally Lipschitz function ("generic" in applications),



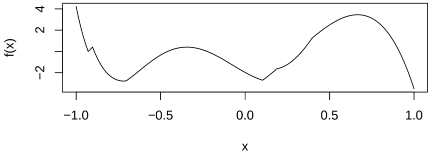
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.



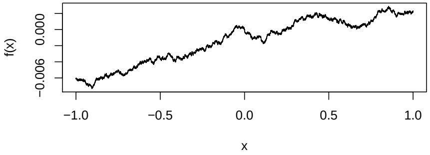
Let f be a *tame* locally Lipschitz function ("generic" in applications),

- piecewise polynomial.



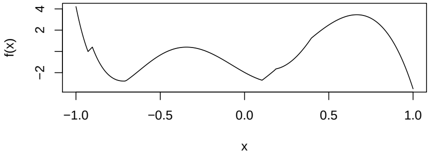
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.



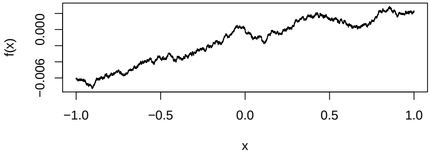
Let f be a *tame* locally Lipschitz function ("generic" in applications),

- piecewise polynomial.
- semi-algebraic.



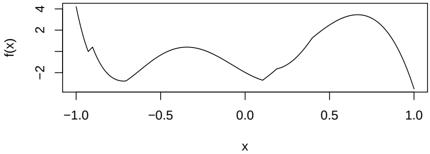
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.



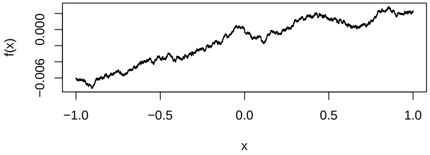
Let f be a *tame* locally Lipschitz function ("generic" in applications),

- piecewise polynomial.
- semi-algebraic.
- definable.



Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

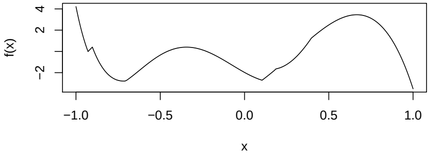
- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.



Let f be a *tame* locally Lipschitz function ("generic" in applications),

- piecewise polynomial.
- semi-algebraic.
- definable.

Davis et .al. 2019, **Bolte et. al. 2007:** Subgradient projection formula implies chain rule along AC curves.



Conservative gradient (Bolte-Pauwels):
$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$,
For any AC curve $\gamma : [0,1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in D(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1]$$

Conservative gradient (Bolte-Pauwels):
$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$,
For any AC curve $\gamma : [0,1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in D(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1]$$

Conservative gradient (Bolte-Pauwels):
$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any AC curve $\gamma : [0, 1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in D(\gamma(t)), \qquad \text{a.e.} \quad t \in [0, 1]$$

Conservative gradient (Bolte-Pauwels):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any AC curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t)\rangle \qquad \forall v \in D(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1]$$

- f is path differentiable.
- D is a conservative gradient for f.
- Conservative Jacobians defined similarly

Conservative gradient (Bolte-Pauwels):
$f \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any AC curve $\gamma \colon [0, 1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot\gamma(t)\rangle \qquad \forall v \in D(\gamma(t)), \qquad \text{a.e.} \quad t \in [0, 1]$$

- f is path differentiable.
- D is a conservative gradient for f.
- Conservative Jacobians defined similarly

Results:

- $D(x) = \{\nabla f(x)\}$ for *almost all* $x \in \mathbb{R}^p$.
- $\partial^c f(x) \subset \mathrm{conv}(D(x))$ for all $x \in \mathbb{R}^p$.
- Sum, linear combinations, compositions of conservative Jacobians are conservative.

Conservative gradient and optimization

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{j=1}^{L} J_{i,j}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

Conservative gradient and optimization

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{l=1}^{L} J_{i,l}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{l=1}^{L} J_{i,l}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{l=1}^{L} J_{i,l}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

Algorithmic differentiation + stochastic approximation: fix $\theta_0 \in \mathbb{R}^p$, $(I_k)_{k \in \mathbb{N}}$ *i.i.d.* uniform in $\{1, \ldots, N\}$,

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -D_{I_k}(\theta_k)$$

Conservative gradient and optimization

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{l=1}^{L} J_{i,l}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

Algorithmic differentiation + stochastic approximation: fix $\theta_0 \in \mathbb{R}^p$, $(I_k)_{k \in \mathbb{N}}$ *i.i.d.* uniform in $\{1, \ldots, N\}$,

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -D_{I_k}(\theta_k)$$

- **Step size:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k = o(1/\log(k))$.
- **Boundedness:** there exists $M > 0$, $\|\theta_k\| \leq M$ almost surely.
- Almost surely, $\ell(\theta_k)$ converges, accumulation points satisfy $0 \in \sum_{i=1}^{N} \mathrm{conv}(D_i(\bar{\theta}))$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{l=1}^{L} J_{i,l}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

Algorithmic differentiation + stochastic approximation: fix $\theta_0 \in \mathbb{R}^p$, $(I_k)_{k \in \mathbb{N}}$ *i.i.d.* uniform in $\{1, \ldots, N\}$,

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -D_{I_k}(\theta_k)$$

- **Step size:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k = o(1/\log(k))$.
- **Boundedness:** there exists $M > 0$, $\|\theta_k\| \leq M$ almost surely.
- Almost surely, $\ell(\theta_k)$ converges, accumulation points satisfy $0 \in \sum_{i=1}^{N} \mathrm{conv}(D_i(\bar{\theta}))$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

For $i \in \{1, \ldots, N\}$, set $D_i = \prod_{l=1}^{L} J_{i,l}$.

- D_i is a conservative gradient for ℓ_i.
- Algorithmic differentiation is an oracle for D_i.

Algorithmic differentiation + stochastic approximation: fix $\theta_0 \in \mathbb{R}^p$, $(I_k)_{k \in \mathbb{N}}$ *i.i.d.* uniform in $\{1, \ldots, N\}$,

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} \in -D_{I_k}(\theta_k)$$

- **Step size:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k = o(1/\log(k))$.
- **Boundedness:** there exists $M > 0$, $\|\theta_k\| \leq M$ almost surely.
- Almost surely, $\ell(\theta_k)$ converges, accumulation points satisfy $0 \in \sum_{i=1}^{N} \mathrm{conv}(D_i(\bar{\theta}))$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.

Conservative gradients / Jacobians:

- Objects akin to Clarke's subgradient / Jacobian.
- Compatible with compositional calculus rules
- Have a minimizing behavior similar to subgradients in optimization.

Plan

Clarke's inverse mapping theorem:

- $F : \mathbb{R}^n \to \mathbb{R}^n$ locally Lipschitz
- Clarke Jacobian $J_F^c : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$
- $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n \times n}$ only contains nonsingular matrices.

Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Failure of formal differentiation

$$J_{F^{-1}}^c(y) \neq J_F^c(F^{-1}(y))^{-1} := \left\{ M^{-1},\ M \in J_F^c(F^{-1}(y)) \right\}$$

Clarke's inverse mapping theorem:

- $F : \mathbb{R}^n \to \mathbb{R}^n$ locally Lipschitz
- Clarke Jacobian $J_F^c : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$
- $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n\times n}$ only contains nonsingular matrices.

Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Failure of formal differentiation

$$J_{F^{-1}}^c(y) \neq J_F^c(F^{-1}(y))^{-1} := \left\{ M^{-1},\, M \in J_F^c(F^{-1}(y)) \right\}$$

Clarke's inverse mapping theorem:

- $F : \mathbb{R}^n \to \mathbb{R}^n$ path differentiable
- Clarke Jacobian $J_F^c : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$
- $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n \times n}$ only contains nonsingular matrices.

Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Failure of formal differentiation

$$J_{F^{-1}}^c(y) \neq J_F^c(F^{-1}(y))^{-1} := \left\{ M^{-1},\, M \in J_F^c(F^{-1}(y)) \right\}$$

Clarke's inverse mapping theorem:

- $F : \mathbb{R}^n \to \mathbb{R}^n$ path differentiable
- Clarke Jacobian $J_F^c : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$
- $\bar{x} \in \mathbb{R}^n$ such that $J_F^c(\bar{x}) \subset \mathbb{R}^{n\times n}$ only contains nonsingular matrices.

Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Failure of formal differentiation

$$J_{F^{-1}}^c(y) \neq J_F^c(F^{-1}(y))^{-1} := \left\{M^{-1},\, M \in J_F^c(F^{-1}(y))\right\}$$

Conservative calculus:

$$y \rightrightarrows J_F^c(F^{-1}(y))^{-1} := \left\{M^{-1},\, M \in J_F^c(F^{-1}(y))\right\}$$

is a conservative Jacobian for F^{-1} (in a neighborhood of $F(\bar{x})$).

Clarke's inverse mapping theorem:

- $F : \mathbb{R}^n \to \mathbb{R}^n$ path differentiable
- Conservative Jacobian $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ convex valued
- $\bar{x} \in \mathbb{R}^n$ such that $J_F(\bar{x}) \subset \mathbb{R}^{n\times n}$ only contains nonsingular matrices.

Then there exists $U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ such that F_U is a bi-Lipschitz homeomorphism.

Failure of formal differentiation

$$J^c_{F^{-1}}(y) \neq J^c_F(F^{-1}(y))^{-1} := \left\{M^{-1},\ M \in J^c_F(F^{-1}(y))\right\}$$

Conservative calculus:

$$y \rightrightarrows J_F(F^{-1}(y))^{-1} := \left\{M^{-1},\ M \in J_F(F^{-1}(y))\right\}$$

is a conservative Jacobian for F^{-1} (in a neighborhood of $F(\bar{x})$).

- $F \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ locally Lipschitz
- Clarke Jacobian $J_F^c \colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^{m \times (n+m)}$
- $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $F(\bar{x}, \bar{y}) = 0$.
- $\forall [A\ B] \in J_F^c(\bar{x}, \bar{y})$, B is invertible

then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a locally Lipschitz function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

Lipschitz implicit function theorem (Hiriart Urruty 1979, Clarke 1976)

- $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ path differentiable
- Clarke Jacobian $J_F^c\colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^{m\times(n+m)}$
- $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $F(\bar{x}, \bar{y}) = 0$.
- $\forall [A\ B] \in J_F^c(\bar{x}, \bar{y})$, B is invertible

then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a locally Lipschitz function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

Lipschitz implicit function theorem (Hiriart Urruty 1979, Clarke 1976)

- $F: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ path differentiable
- Clarke Jacobian $J_F^c: \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^{m\times(n+m)}$
- $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $F(\bar{x}, \bar{y}) = 0$.
- $\forall [A\ B] \in J_F^c(\bar{x}, \bar{y})$, B is invertible

then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a locally Lipschitz function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

Nonsmooth implicit differentiation:

$$x \rightrightarrows \left\{ -B^{-1}A : [A\ B] \in J_F^c(x, G(x)) \right\}.$$

is a conservative Jacobian for G in a neighborhood of $\bar{x}$.

Lipschitz implicit function theorem (Hiriart Urruty 1979, Clarke 1976)

- $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ path differentiable
- Conservative Jacobian $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ convex valued
- $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $F(\bar{x}, \bar{y}) = 0$.
- $\forall [A\ B] \in J_F(\bar{x}, \bar{y})$, B is invertible

then $\exists U \subset \mathbb{R}^n$ a neighborhood of $\bar{x}$ and a locally Lipschitz function $G(x)$ so that

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

and $y = G(x)$ is the unique such solution in a neighborhood of $\bar{y}$.

Nonsmooth implicit differentiation:

$$x \rightrightarrows \left\{ -B^{-1}A : [A\ B] \in J_F(x, G(x)) \right\}.$$

is a conservative Jacobian for G in a neighborhood of $\bar{x}$.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

- Extends to implicitely defined input output relations.
- Preserved by inversion / implicit definition.

$\Rightarrow$ convergence of small step first order methods.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

- Extends to implicitely defined input output relations.
- Preserved by inversion / implicit definition.

$\Rightarrow$ convergence of small step first order methods.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

- Extends to implicitely defined input output relations.
- Preserved by inversion / implicit definition.

$\Rightarrow$ convergence of small step first order methods.

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

Assumption: For $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, L\}$,

- $g_{i,j}$ locally Lipschitz, conservative Jacobian $J_{i,j}$, semialgebraic (or definable).

- Extends to implicitely defined input output relations.
- Preserved by inversion / implicit definition.

$\Rightarrow$ convergence of small step first order methods.

Plan

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{s.t.} \quad & x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{s.t.} \quad & x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{s.t.} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{where} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

How to differentiate the solution of an optimization problem?

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{where} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

How to differentiate the solution of an optimization problem?

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{where} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Example: Lasso hyperparameter optimization (Bertrand *et. al.* 2020).

$$x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} \frac{1}{2}\|Ax - b\|_2^2 + \theta\|x\|_1, \qquad \theta > 0$$

$(A, b) \in \mathbb{R}^{n \times m} \times \mathbb{R}^n$, training data , ℓ loss on held out data

$$x = \operatorname{prox}_{s\theta\|\cdot\|_1}(x - sA^T(Ax - b)) \qquad s > 0$$

How to differentiate the solution of an optimization problem?

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{where} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Example: Lasso hyperparameter optimization (Bertrand *et. al.* 2020).

$$x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} \frac{1}{2}\|Ax - b\|_2^2 + \theta\|x\|_1, \qquad \theta > 0$$

$(A, b) \in \mathbb{R}^{n \times m} \times \mathbb{R}^n$, training data , ℓ loss on held out data

$$x = \mathrm{prox}_{s\theta\|\cdot\|_1}(x - sA^T(Ax - b)) \qquad s > 0$$

How to differentiate the solution of an optimization problem?

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{where} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Example: Lasso hyperparameter optimization (Bertrand *et. al.* 2020).

$$x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} \frac{1}{2}\|Ax - b\|_2^2 + \theta\|x\|_1, \qquad \theta > 0$$

$(A, b) \in \mathbb{R}^{n \times m} \times \mathbb{R}^n$, training data , ℓ loss on held out data

$$x = \operatorname{prox}_{s\theta\|\cdot\|_1}(x - sA^T(Ax - b)) \qquad s > 0$$

Equicorrelation set: $\mathcal{E} := \{j \in \{1, \ldots, m\} : |A_j^T(b - Ax(\theta))| = \theta\}$.

How to differentiate the solution of an optimization problem?

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \ell(x(\theta)) \\ \text{where} \quad & x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Example: Lasso hyperparameter optimization (Bertrand *et. al.* 2020).

$$x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} \frac{1}{2}\|Ax - b\|_2^2 + \theta\|x\|_1, \qquad \theta > 0$$

$(A, b) \in \mathbb{R}^{n \times m} \times \mathbb{R}^n$, training data , ℓ loss on held out data

$$x = \mathrm{prox}_{s\theta\|\cdot\|_1}(x - sA^T(Ax - b)) \qquad s > 0$$

Equicorrelation set: $\mathcal{E} := \{j \in \{1, \dots, m\} : |A_j^T(b - Ax(\theta))| = \theta\}$.
If $A_{\mathcal{E}}^T A_{\mathcal{E}}$ has full rank, nonsmooth implicit differentiation applies.

How to differentiate the solution of an optimization problem?

Assumption: ℓ and f locally Lipschitz. For any θ,

- the inner argmin is a singleton

$$\begin{aligned} &\min_{\theta \in \mathbb{R}^p} && \ell(x(\theta)) \\ &\text{where} && x(\theta) = \operatorname*{argmin}_{x \in \mathbb{R}^m} f(x, \theta) \end{aligned}$$

Example: Lasso hyperparameter optimization (Bertrand *et. al.* 2020).

$$x(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^m} \frac{1}{2} \|Ax - b\|_2^2 + \theta \|x\|_1, \qquad \theta > 0$$

$(A, b) \in \mathbb{R}^{n \times m} \times \mathbb{R}^n$, training data , ℓ loss on held out data

$$x = \mathrm{prox}_{s\theta \|\cdot\|_1}(x - sA^T(Ax - b)) \qquad s > 0$$

Equicorrelation set: $\mathcal{E} := \{j \in \{1, \dots, m\} : |A_j^T(b - Ax(\theta))| = \theta\}$.
If $A_{\mathcal{E}}^T A_{\mathcal{E}}$ has full rank, nonsmooth implicit differentiation applies.

$\Rightarrow$ recover LARS algorithm + convergence of small step first order methods.

Neural networks:
Compositional models,

Neural networks:

Compositional models,

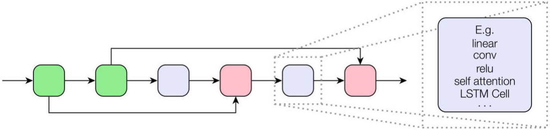


Image courtesy: `implicit-layers-tutorial.org`

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

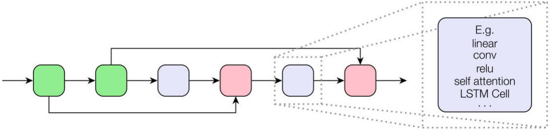


Image courtesy: `implicit-layers-tutorial.org`

Neural networks:

Compositional models, elementary blocks called *layers* (parametric functions).

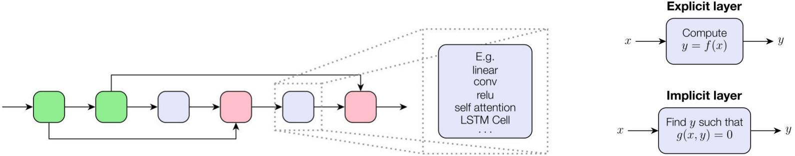


Image courtesy: `implicit-layers-tutorial.org`

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

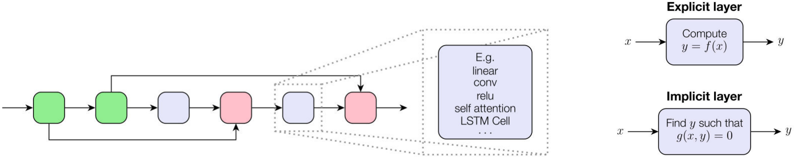


Image courtesy: `implicit-layers-tutorial.org`

Examples: Equilibrium networks (Bai *et. al.* 2019), implicit networks (El Ghaoui *et. al.* 2019) declarative networks (Gould *et. al.* 2019), optimization layers (Agrawal *et. al.* 2019)

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

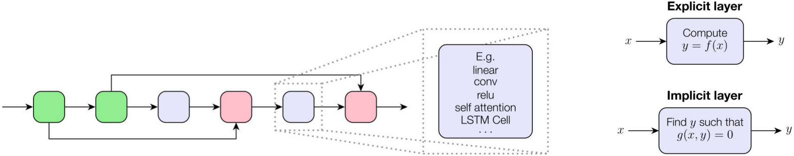


Image courtesy: `implicit-layers-tutorial.org`

Examples: Equilibrium networks (Bai *et. al.* 2019), implicit networks (El Ghaoui *et. al.* 2019) declarative networks (Gould *et. al.* 2019), optimization layers (Agrawal *et. al.* 2019)

Monotone operator DEQs: Winston *et. al.* 2020,
$\sigma : \mathbb{R}^m \to \mathbb{R}^m$ proximal operator (convex function), $W \in \mathbb{R}^{m \times m}$ such that $W + W^T \succ I$.

$$z = \sigma(Wz + b) \qquad\qquad \forall b \in \mathbb{R}^m, \text{ unique solution } z(b).$$

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

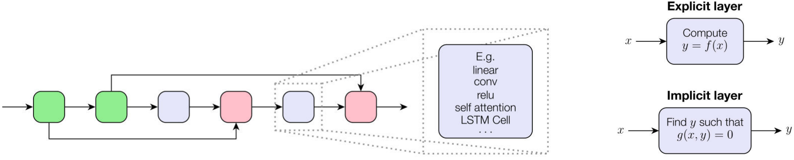


Image courtesy: `implicit-layers-tutorial.org`

Examples: Equilibrium networks (Bai *et. al.* 2019), implicit networks (El Ghaoui *et. al.* 2019) declarative networks (Gould *et. al.* 2019), optimization layers (Agrawal *et. al.* 2019)

Monotone operator DEQs: Winston *et. al.* 2020,
$\sigma \colon \mathbb{R}^m \to \mathbb{R}^m$ proximal operator (convex function), $W \in \mathbb{R}^{m \times m}$ such that $W + W^T \succ I$.

$$z = \sigma(Wz + b) \qquad \forall b \in \mathbb{R}^m, \text{ unique solution } z(b).$$

$J_\sigma^c \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{m \times m}$ Clarke Jacobian for σ (assumed path differentiable).

Compositional models

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

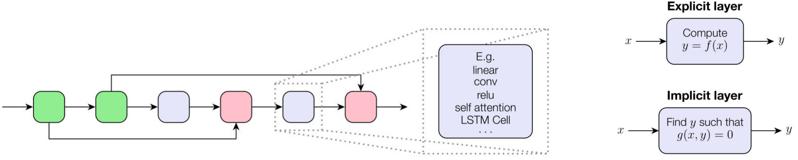


Image courtesy: `implicit-layers-tutorial.org`

Examples: Equilibrium networks (Bai *et. al.* 2019), implicit networks (El Ghaoui *et. al.* 2019) declarative networks (Gould *et. al.* 2019), optimization layers (Agrawal *et. al.* 2019)

Monotone operator DEQs: Winston *et. al.* 2020,
$\sigma \colon \mathbb{R}^m \to \mathbb{R}^m$ proximal operator (convex function), $W \in \mathbb{R}^{m \times m}$ such that $W + W^T \succ I$.

$$z = \sigma(Wz + b) \qquad\qquad \forall b \in \mathbb{R}^m, \text{ unique solution } z(b).$$

$J_\sigma^c \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{m \times m}$ Clarke Jacobian for σ (assumed path differentiable).

then $(I - JW)$ invertible for all $J \in J_\sigma^c(Wz + b)$

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

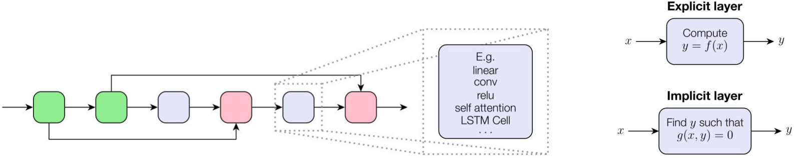


Image courtesy: `implicit-layers-tutorial.org`

Examples: Equilibrium networks (Bai *et. al.* 2019), implicit networks (El Ghaoui *et. al.* 2019) declarative networks (Gould *et. al.* 2019), optimization layers (Agrawal *et. al.* 2019)

Monotone operator DEQs: Winston *et. al.* 2020,
$\sigma\colon \mathbb{R}^m \to \mathbb{R}^m$ proximal operator (convex function), $W \in \mathbb{R}^{m\times m}$ such that $W + W^T \succ I$.

$$z = \sigma(Wz + b) \qquad \forall b \in \mathbb{R}^m, \text{ unique solution } z(b).$$

$J_\sigma^c\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{m\times m}$ Clarke Jacobian for σ (assumed path differentiable).

then $(I - JW)$ invertible for all $J \in J_\sigma^c(Wz + b)$
$\Rightarrow$ invertibility condition for nonsmooth implicit differentiation

Neural networks:
Compositional models, elementary blocks called *layers* (parametric functions).

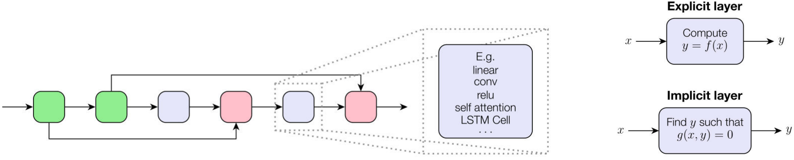


Image courtesy: `implicit-layers-tutorial.org`

Examples: Equilibrium networks (Bai *et. al.* 2019), implicit networks (El Ghaoui *et. al.* 2019) declarative networks (Gould *et. al.* 2019), optimization layers (Agrawal *et. al.* 2019)

Monotone operator DEQs: Winston *et. al.* 2020,
$\sigma\colon \mathbb{R}^m \to \mathbb{R}^m$ proximal operator (convex function), $W \in \mathbb{R}^{m\times m}$ such that $W + W^T \succ I$.

$$z = \sigma(Wz + b) \qquad\qquad \forall b \in \mathbb{R}^m, \text{ unique solution } z(b).$$

$J_\sigma^c\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{m\times m}$ Clarke Jacobian for σ (assumed path differentiable).

then $(I - JW)$ invertible for all $J \in J_\sigma^c(Wz + b)$
$\Rightarrow$ invertibility condition for nonsmooth implicit differentiation
$\Rightarrow$ convergence of small steps training algorithms.

Plan

Implicit differentiation applied to:

$$\begin{aligned} \min_{x,y,s} \quad & \ell(x, y, s) := (x - s_1)^2 + 4(y - s_2)^2 \\ \text{s.t.} \quad & s \in \arg\max\{(a + b)(-2x + y + 2) : a \in [0, 3], b \in [0, 5]\}. \end{aligned}$$

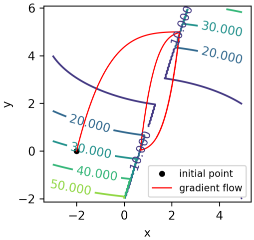
Implicit differentiation applied to:

$$\begin{aligned} \min_{x,y,s} \quad & \ell(x,y,s) := (x-s_1)^2 + 4(y-s_2)^2 \\ \text{s.t.} \quad & s \in \arg\max\{(a+b)(-2x+y+2) : a \in [0,3], b \in [0,5]\}. \end{aligned}$$

- Fixed point of projected gradient (linear over a box)

Implicit differentiation applied to:

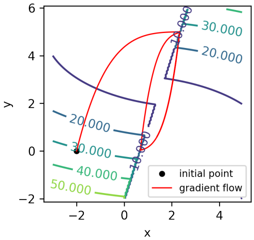
$$\min_{x,y,s} \quad \ell(x,y,s) := (x - s_1)^2 + 4(y - s_2)^2$$
$$\text{s.t.} \quad s \in \arg\max\{(a+b)(-2x+y+2) : a \in [0,3], b \in [0,5]\}.$$



- Fixed point of projected gradient (linear over a box)

Implicit differentiation applied to:

$$\min_{x,y,s} \quad \ell(x,y,s) := (x - s_1)^2 + 4(y - s_2)^2$$
$$\text{s.t.} \quad s \in \arg\max\{(a+b)(-2x+y+2) : a \in [0,3], b \in [0,5]\}.$$

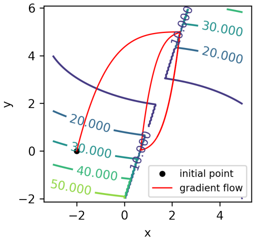


- Fixed point of projected gradient (linear over a box)
- Invertibility condition outside of the line $y = 2x - 2$.

Pathological Examples - Cyclic "*gradient*" orbits

Implicit differentiation applied to:

$$\min_{x,y,s} \quad \ell(x,y,s) := (x - s_1)^2 + 4(y - s_2)^2$$
$$\text{s.t.} \quad s \in \arg\max\{(a+b)(-2x+y+2) : a \in [0,3], b \in [0,5]\}.$$

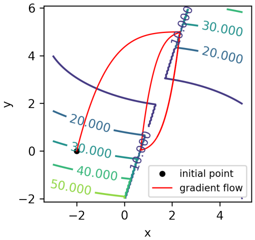


- Fixed point of projected gradient (linear over a box)
- Invertibility condition outside of the line $y = 2x - 2$.
- Discontinuity of the solution map.

Pathological Examples - Cyclic "*gradient*" orbits

Implicit differentiation applied to:

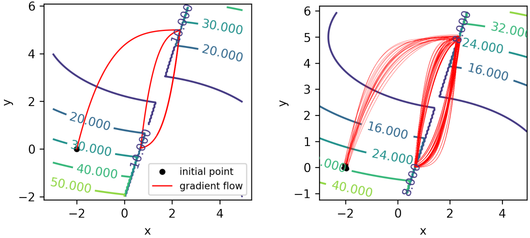
$$\min_{x,y,s} \quad \ell(x,y,s) := (x - s_1)^2 + 4(y - s_2)^2$$
$$\text{s.t.} \quad s \in \arg\max\{(a+b)(-2x+y+2) : a \in [0,3], b \in [0,5]\}.$$



- Fixed point of projected gradient (linear over a box)
- Invertibility condition outside of the line $y = 2x - 2$.
- Discontinuity of the solution map.
- Globally affects dynamics (not of gradient type) although line never met.

Implicit differentiation applied to:

$$\min_{x,y,s} \quad \ell(x,y,s) := (x - s_1)^2 + 4(y - s_2)^2$$
$$\text{s.t.} \quad s \in \arg\max\{(a+b)(-2x+y+2) : a \in [0,3], b \in [0,5]\}.$$



- Fixed point of projected gradient (linear over a box)
- Invertibility condition outside of the line $y = 2x - 2$.
- Discontinuity of the solution map.
- Globally affects dynamics (not of gradient type) although line never met.
- Generic: robust to perturbation of problem data.

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?

Conclusion

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?
- Using Clarke's Jacobian: No.
  - Inverses of Clarke Jacobians are not Clarke Jacobians

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?
- Using Clarke's Jacobian: No.
  - Inverses of Clarke Jacobians are not Clarke Jacobians
- Using Conservative Jacobian: Yes.
  - Inverses of Conservative Jacobians are conservative Jacobians

Conclusion

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?
- Using Clarke's Jacobian: No.
  - Inverses of Clarke Jacobians are not Clarke Jacobians
- Using Conservative Jacobian: Yes.
  - Inverses of Conservative Jacobians are conservative Jacobians

Practical implications:

- Extends the domain of validity of stochastic learning algorithm / compositional modeling.
- Applications in ML (bilevel hyperparameter tuning, implicit neural networks . . .).

Conclusion

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?
- Using Clarke's Jacobian: No.
  - Inverses of Clarke Jacobians are not Clarke Jacobians
- Using Conservative Jacobian: Yes.
  - Inverses of Conservative Jacobians are conservative Jacobians

Practical implications:

- Extends the domain of validity of stochastic learning algorithm / compositional modeling.
- Applications in ML (bilevel hyperparameter tuning, implicit neural networks . . .).

Improvements:

- Do pathologies occur in practice? How to check?
- How to check invertibility condition?

Conclusion

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?
- Using Clarke's Jacobian: No.
  - Inverses of Clarke Jacobians are not Clarke Jacobians
- Using Conservative Jacobian: Yes.
  - Inverses of Conservative Jacobians are conservative Jacobians

Practical implications:

- Extends the domain of validity of stochastic learning algorithm / compositional modeling.
- Applications in ML (bilevel hyperparameter tuning, implicit neural networks . . .).

Improvements:

- Do pathologies occur in practice? How to check?
- How to check invertibility condition?

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

`https://arxiv.org/abs/2106.04350`

Conclusion

Nonsmooth implicit differentiation

- Does Lipschitz implicit function theorem come with a calculus?
- Using Clarke's Jacobian: No.
  - Inverses of Clarke Jacobians are not Clarke Jacobians
- Using Conservative Jacobian: Yes.
  - Inverses of Conservative Jacobians are conservative Jacobians

Practical implications:

- Extends the domain of validity of stochastic learning algorithm / compositional modeling.
- Applications in ML (bilevel hyperparameter tuning, implicit neural networks . . .).

Improvements:

- Do pathologies occur in practice? How to check?
- How to check invertibility condition?

Jérôme Bolte, Tâm Lê, Edouard Pauwels, Antonio Silveti-Falls
`https://arxiv.org/abs/2106.04350`

Thanks.