Nonsmooth differential calculus and optimization, the conservative gradient approach

Edouard Pauwels (IRIT, Toulouse 3, France)

joint work with Jérôme Bolte

Ryan Boustany, Tâm Lê, Swann Marx, Béatrice Pesquet-Popescu, Antonio Silveti-Falls, Samuel Vaiter

Journées MOA, Nice, (Octobre, 2022)

- $f : \mathbb{R}^p \to \mathbb{R}$ differentiable expressed as

$$f = g_L \circ \ldots \circ g_1 \quad \text{with } g_i \text{ ``elementary'' differentiable.}$$

- $f : \mathbb{R}^p \to \mathbb{R}$ differentiable expressed as
$$f = g_L \circ \ldots \circ g_1 \quad \text{with } g_i \text{ "elementary" differentiable.}$$

- backprop : efficient algorithm to compute derivatives with the chain rule.

  In the smooth world BP outputs: $\boxed{\text{backprop}\, f = \text{Jac}\, g_L \circ \ldots \circ \text{Jac}\, g_1 = \nabla f^T}$

- $f : \mathbb{R}^p \to \mathbb{R}$ differentiable expressed as

$$f = g_L \circ \ldots \circ g_1 \quad \text{with } g_i \text{ "elementary" differentiable.}$$

- $\mathrm{backprop}$: efficient algorithm to compute derivatives with the chain rule.

  In the smooth world BP outputs: $\boxed{\mathrm{backprop}\, f = \mathrm{Jac}\, g_L \circ \ldots \circ \mathrm{Jac}\, g_1 = \nabla f^T}$

- **Baur-Strassen**: $\boxed{\text{Computing cost } (f, \nabla f) \leq 5 \text{ Computing cost } (f)}$ instead of the naive Computing cost $(f, \nabla f) \leq p$ Computing cost (f)

Smooth backpropagation

- $f : \mathbb{R}^p \to \mathbb{R}$ differentiable expressed as

$$f = g_L \circ \ldots \circ g_1 \quad \text{with } g_i \text{ ``elementary'' differentiable.}$$

- backprop : efficient algorithm to compute derivatives with the chain rule.

  In the smooth world BP outputs: $\boxed{\mathrm{backprop}\, f = \mathrm{Jac}\, g_L \circ \ldots \circ \mathrm{Jac}\, g_1 = \nabla f^T}$

- **Baur-Strassen**: $\boxed{\text{Computing cost } (f, \nabla f) \leq 5 \text{ Computing cost } (f)}$ instead of the naive Computing cost $(f, \nabla f) \leq p$ Computing cost (f)

- **Essential** element in modern AI / deep learning:

TensorFlow PyTorch

Smooth backpropagation

- $f : \mathbb{R}^p \to \mathbb{R}$ differentiable expressed as

$$f = g_L \circ \ldots \circ g_1 \quad \text{with } g_i \text{ "elementary" differentiable.}$$

- backprop : efficient algorithm to compute derivatives with the chain rule.

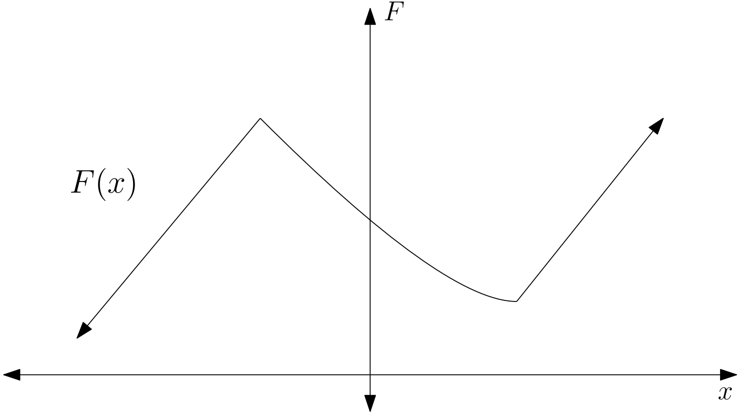
  In the smooth world BP outputs: $\boxed{\mathrm{backprop}\, f = \mathrm{Jac}\, g_L \circ \ldots \circ \mathrm{Jac}\, g_1 = \nabla f^T}$

- **Baur-Strassen**: $\boxed{\text{Computing cost } (f, \nabla f) \leq 5 \text{ Computing cost } (f)}$ instead of the naive Computing cost $(f, \nabla f) \leq p$ Computing cost (f)

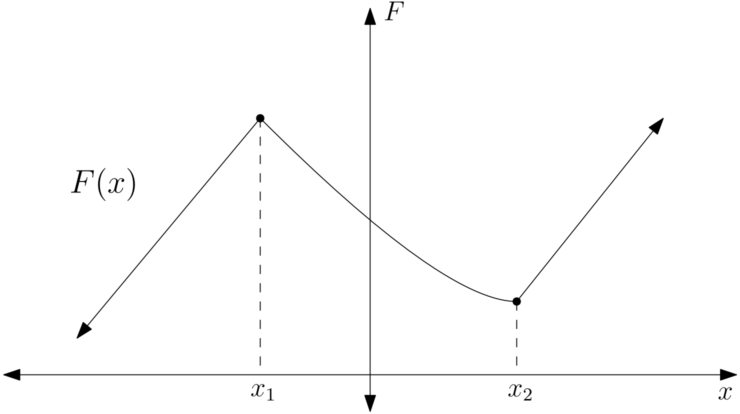
- **Essential** element in modern AI / deep learning:

  TensorFlow PyTorch

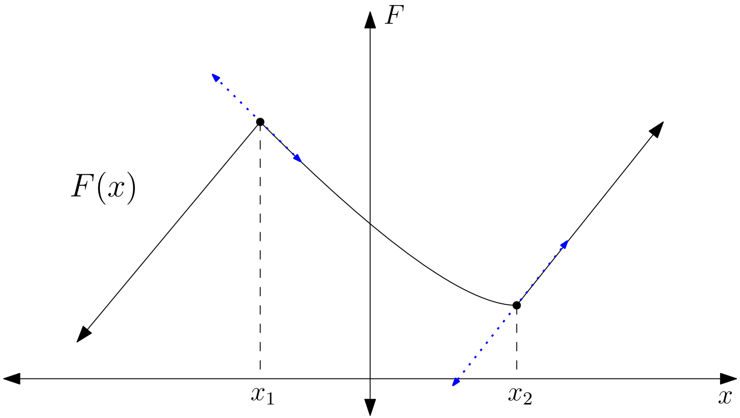
- $\boxed{\textbf{Nonsmoothness is needed:}}$ $g_i = \mathrm{relu}$, sort, maxpool, implicit layers



$F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz

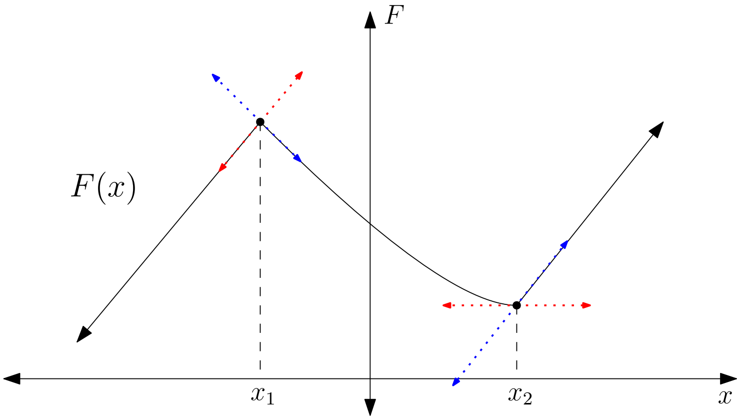


$F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz, differentiable almost everywhere (Rademacher).



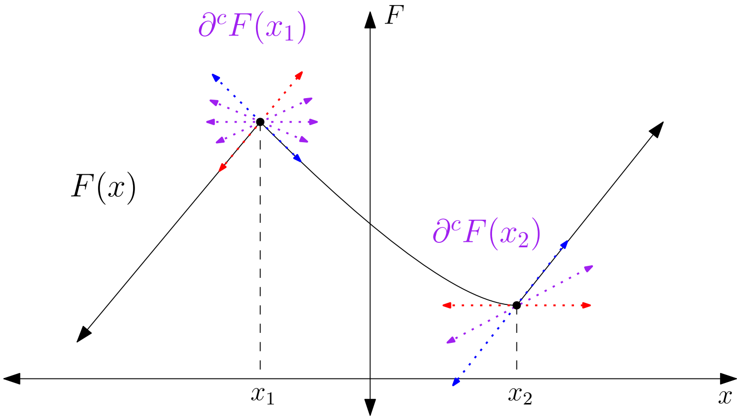
$F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz, differentiable almost everywhere (Rademacher).

$$\mathrm{Jac}^{\,c} F(x) = \mathrm{conv}\left\{ M \in \mathbb{R}^{p\times q} :\ x^k \to x,\ F \text{ diff. at } x_k,\ \mathrm{Jac}\, F(x^k) \to M \right\}$$



$F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz, differentiable almost everywhere (Rademacher).

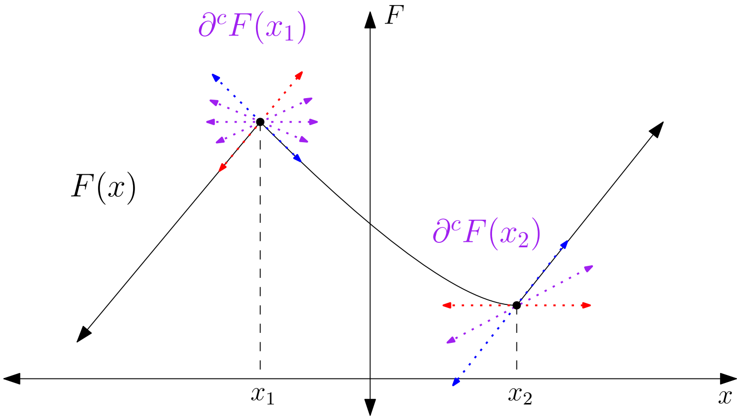
$$\mathrm{Jac}^{\,c} F(x) = \mathrm{conv}\left\{M \in \mathbb{R}^{p\times q} :\ x^k \to x,\ F \text{ diff. at } x_k,\ \mathrm{Jac}\, F(x^k) \to M\right\}$$



$F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz, differentiable almost everywhere (Rademacher).

$$\mathrm{Jac}^{\,c} F(x) = \mathrm{conv}\left\{ M \in \mathbb{R}^{p \times q} :\ x^k \to x,\ F \text{ diff. at } x_k,\ \mathrm{Jac}\, F(x^k) \to M \right\}$$

Denoted by $\partial^c F(x)$ when $q = 1$



$F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz, differentiable almost everywhere (Rademacher).

$\mathrm{Jac}^{\,c} F(x) = \mathrm{conv}\left\{ M \in \mathbb{R}^{p\times q} : \; x^k \to x, \; F \text{ diff. at } x_k, \; \mathrm{Jac}\, F(x^k) \to M \right\}$

Denoted by $\partial^c F(x)$ when $q = 1$

Set valued $\mathrm{Jac}^{\,c} F \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{q\times p}$

- Take $f\colon \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \ldots, g_L$

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu},$ sort, maxpool, output of nonsmooth numerical program.

How does nonsmooth backprop work?

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \dots, g_L$

$$f = g_L \circ \dots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu},$ sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth backprop is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \dots \circ \mathrm{Jac}^c g_1$$

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \dots, g_L$

$$f = g_L \circ \dots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu}$, sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth $\mathrm{backprop}$ is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \dots \circ \mathrm{Jac}^c g_1$$

- $\mathrm{backprop}_f : \mathbb{R}^p \to \mathbb{R}^p$ is a selection in the set valued field $\mathrm{Jac}^c g_L \circ \dots \circ \mathrm{Jac}^c g_1 : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$.

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \dots, g_L$

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu},$ sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth backprop is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1$$

- $\mathrm{backprop}_f : \mathbb{R}^p \to \mathbb{R}^p$ is a selection in the set valued field $\mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1 : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$.

- This is what common is done in:

TensorFlow PyTorch

How does nonsmooth backprop work?

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \ldots, g_L$

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu}$, sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth backprop is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1$$

- $\mathrm{backprop}_f : \mathbb{R}^p \to \mathbb{R}^p$ is a selection in the set valued field $\mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1 : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$.

- This is what common is done in:

TensorFlow PyTorch

$\boxed{\text{But what does backprop output? What sort of gradient could it be?}}$

Plan

Outputs are partly unpredictible

$$\mathrm{relu}(t) = \max\{0, t\} \qquad \mathrm{relu}_2(t) = \mathrm{relu}(-t) + t \qquad \mathrm{relu}_3(t) = 1/2(\mathrm{relu}(t) + \mathrm{relu}_2(t))$$

Outputs are partly unpredictible

$$\mathrm{relu}(t) = \max\{0, t\} \qquad \mathrm{relu}_2(t) = \mathrm{relu}(-t) + t \qquad \mathrm{relu}_3(t) = 1/2(\mathrm{relu}(t) + \mathrm{relu}_2(t))$$

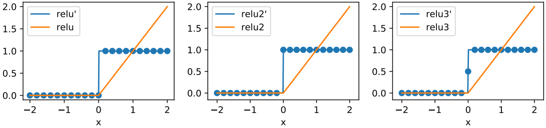
Then $\mathrm{relu} = \mathrm{relu}_2 = \mathrm{relu}_3$.

$$\mathrm{relu}(t) = \max\{0, t\} \qquad \mathrm{relu}_2(t) = \mathrm{relu}(-t) + t \qquad \mathrm{relu}_3(t) = 1/2(\mathrm{relu}(t) + \mathrm{relu}_2(t))$$

Then $\mathrm{relu} = \mathrm{relu}_2 = \mathrm{relu}_3$.

- TensorFlow (TF) set $\mathrm{backprop}\,\mathrm{relu}(0) = 0$. TF's gives

$$\mathrm{backprop}\,\mathrm{relu}_2(0) = 1 \text{ and } \mathrm{backprop}\,\mathrm{relu}_3(0) = 1/2.$$



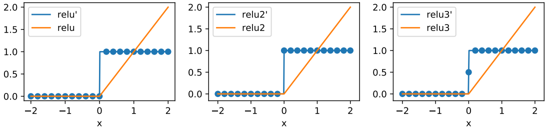
Outputs are partly unpredictible

$$\mathrm{relu}(t) = \max\{0, t\} \qquad \mathrm{relu}_2(t) = \mathrm{relu}(-t) + t \qquad \mathrm{relu}_3(t) = 1/2(\mathrm{relu}(t) + \mathrm{relu}_2(t))$$

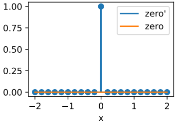
Then $\mathrm{relu} = \mathrm{relu}_2 = \mathrm{relu}_3$.

- TensorFlow (TF) set $\mathrm{backprop}\,\mathrm{relu}(0) = 0$. TF's gives

$$\mathrm{backprop}\,\mathrm{relu}_2(0) = 1 \text{ and } \mathrm{backprop}\,\mathrm{relu}_3(0) = 1/2.$$



- Artifacts: $\mathrm{zero}(x) = \mathrm{relu2}(x) - \mathrm{relu}(x) = 0$.

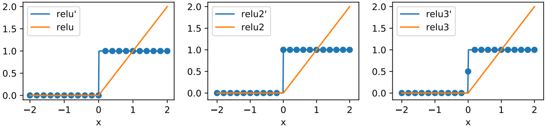


$$\mathrm{relu}(t) = \max\{0, t\} \qquad \mathrm{relu}_2(t) = \mathrm{relu}(-t) + t \qquad \mathrm{relu}_3(t) = 1/2(\mathrm{relu}(t) + \mathrm{relu}_2(t))$$

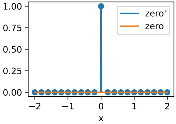
Then $\mathrm{relu} = \mathrm{relu}_2 = \mathrm{relu}_3$.

- TensorFlow (TF) set $\mathrm{backprop}\,\mathrm{relu}(0) = 0$. TF's gives

$$\mathrm{backprop}\,\mathrm{relu}_2(0) = 1 \text{ and } \mathrm{backprop}\,\mathrm{relu}_3(0) = 1/2.$$



- Artifacts: $\mathrm{zero}(x) = \mathrm{relu2}(x) - \mathrm{relu}(x) = 0$.



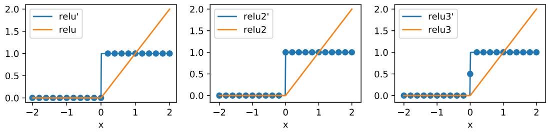
- Actually $s \times \mathrm{zero} = 0$ and $\mathrm{backprop}\,[s \times \mathrm{zero}](0) = s \in \mathbb{R}$ arbitrary

$$\mathrm{relu}(t) = \max\{0, t\} \qquad \mathrm{relu}_2(t) = \mathrm{relu}(-t) + t \qquad \mathrm{relu}_3(t) = 1/2(\mathrm{relu}(t) + \mathrm{relu}_2(t))$$

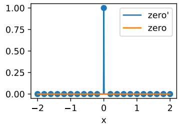
Then $\mathrm{relu} = \mathrm{relu}_2 = \mathrm{relu}_3$.

- TensorFlow (TF) set $\mathrm{backprop}\,\mathrm{relu}(0) = 0$. TF's gives

$$\mathrm{backprop}\,\mathrm{relu}_2(0) = 1 \text{ and } \mathrm{backprop}\,\mathrm{relu}_3(0) = 1/2.$$



- Artifacts: $\mathrm{zero}(x) = \mathrm{relu2}(x) - \mathrm{relu}(x) = 0$.



- Actually $s \times \mathrm{zero} = 0$ and $\mathrm{backprop}\,[s \times \mathrm{zero}](0) = s \in \mathbb{R}$ arbitrary
- Spurious critical point: $\mathrm{identity}(x) := x - \mathrm{zero}(x) = x$ but $\mathrm{backprop}\,\mathrm{identity}(0) = 0$

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c (g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c (g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.

Subgradient calculus

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c(g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c(g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.
- holds with equality if g_1 and g_2 are subdifferentially regular.

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c(g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.
- holds with equality if g_1 and g_2 are subdifferentially regular.
- no equality in general: $g \colon x \mapsto |x|$

$$\partial^c(g - g) = \partial^c(x \mapsto 0) = \{0\} \subset \quad \partial^c(g) + \partial^c(-g) = \quad \begin{cases} 0 & \text{if } x \neq 0 \\ [-2, 2] & \text{if } x = 0 \end{cases}.$$

No convexity, no calculus: $g_1 : \mathbb{R}^p \to \mathbb{R}$, $g_2 : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c(g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.
- holds with equality if g_1 and g_2 are subdifferentially regular.
- no equality in general: $g : x \mapsto |x|$

$$\partial^c(g - g) = \partial^c(x \mapsto 0) = \{0\} \subset \quad \partial^c(g) + \partial^c(-g) = \quad \begin{cases} 0 & \text{if } x \neq 0 \\ [-2, 2] & \text{if } x = 0 \end{cases}.$$

Deep learning: no convexity, no smoothness. Calculus rules?

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c (g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.
- holds with equality if g_1 and g_2 are subdifferentially regular.
- no equality in general: $g \colon x \mapsto |x|$

$$\partial^c (g - g) = \partial^c (x \mapsto 0) = \{0\} \subset \quad \partial^c (g) + \partial^c (-g) = \quad \begin{cases} 0 & \text{if } x \neq 0 \\ [-2, 2] & \text{if } x = 0 \end{cases}.$$

Deep learning: no convexity, no smoothness. Calculus rules?

- backprop : selection in enlarged "subgradient", artifacts

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c (g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.
- holds with equality if g_1 and g_2 are subdifferentially regular.
- no equality in general: $g \colon x \mapsto |x|$

$$\partial^c(g - g) = \partial^c(x \mapsto 0) = \{0\} \subset \quad \partial^c(g) + \partial^c(-g) = \quad \begin{cases} 0 & \text{if } x \neq 0 \\ [-2, 2] & \text{if } x = 0 \end{cases}.$$

Deep learning: no convexity, no smoothness. Calculus rules?

- backprop : selection in enlarged "subgradient", artifacts
- **Non uniqueness:** Different programs may implement the same function.

No convexity, no calculus: $g_1 \colon \mathbb{R}^p \to \mathbb{R}$, $g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz.

$$\partial^c (g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2.$$

- holds with equality if g_1 and g_2 are continuously differentiable.
- holds with equality if g_1 and g_2 are convex.
- holds with equality if g_1 and g_2 are subdifferentially regular.
- no equality in general: $g \colon x \mapsto |x|$

$$\partial^c (g - g) = \partial^c (x \mapsto 0) = \{0\} \subset \quad \partial^c (g) + \partial^c (-g) = \quad \begin{cases} 0 & \text{if } x \neq 0 \\ [-2, 2] & \text{if } x = 0 \end{cases}.$$

Deep learning: no convexity, no smoothness. Calculus rules?

- backprop : selection in enlarged "subgradient", artifacts
- **Non uniqueness:** Different programs may implement the same function.
- **Stochastic approximation:** $\partial^c \left(\frac{1}{n} \sum_{i=1}^n \ell_i \right) \subset \frac{1}{n} \sum_{i=1}^n \partial^c \ell_i$.

Plan

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- Lipschitz $F\colon \mathbb{R}^n \to \mathbb{R}^m$ has none or multiple conservative Jacobians $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$. Notation D_F if $m = 1$ for conservative gradients.
- If conservative Jacobians exist, F is called **path-differentiable**.
- **Solve calculus issue:** compatible with **compositional calculus** rules
- Conservative gradients have a **minimizing behavior** similar to subgradients in optimization.

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- Lipschitz $F\colon \mathbb{R}^n \to \mathbb{R}^m$ has none or multiple conservative Jacobians $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$. Notation D_F if $m = 1$ for conservative gradients.
- If conservative Jacobians exist, F is called **path-differentiable**.
- **Solve calculus issue:** compatible with **compositional calculus** rules
- Conservative gradients have a **minimizing behavior** similar to subgradients in optimization.

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- Lipschitz $F\colon \mathbb{R}^n \to \mathbb{R}^m$ has none or multiple conservative Jacobians $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$. Notation D_F if $m = 1$ for conservative gradients.
- If conservative Jacobians exist, F is called **path-differentiable**.
- **Solve calculus issue:** compatible with **compositional calculus** rules
- Conservative gradients have a **minimizing behavior** similar to subgradients in optimization.

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- Lipschitz $F\colon \mathbb{R}^n \to \mathbb{R}^m$ has none or multiple conservative Jacobians $J_F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$. Notation D_F if $m = 1$ for conservative gradients.
- If conservative Jacobians exist, F is called **path-differentiable**.
- **Solve calculus issue:** compatible with **compositional calculus** rules
- Conservative gradients have a **minimizing behavior** similar to subgradients in optimization.

- Objects akin to Clarke's subgradient / Jacobian (for locally Lipschitz functions).
- Lipschitz $F \colon \mathbb{R}^n \to \mathbb{R}^m$ has none or multiple conservative Jacobians $J_F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$. Notation D_F if $m = 1$ for conservative gradients.
- If conservative Jacobians exist, F is called **path-differentiable**.
- **Solve calculus issue:** compatible with **compositional calculus** rules
- Conservative gradients have a **minimizing behavior** similar to subgradients in optimization.

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

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

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

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

$$v_k \in \partial^c f(\theta_k).$$

$f \colon \mathbb{R}^p \to \mathbb{R}$ 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 Lipschitz curves (Brézis, Valadier).
Hypothesis: Fix any Lipschitz 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, $\quad f(\theta_{k+1}) \leq f(\theta_k)$?

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

$$v_k \in \partial^c f(\theta_k).$$

Chain rule along Lipschitz curves (Brézis, Valadier).
Hypothesis: Fix any Lipschitz 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 $t \in [0,1]$,

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

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

$$v_k \in \partial^c f(\theta_k).$$

Chain rule along Lipschitz curves (Brézis, Valadier).
Hypothesis: Fix any Lipschitz 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 $t \in [0,1]$,

Under the carpet: $\alpha_k \to 0$, small step limit $\to$ solutions to the differential inclusion.

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

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

$$v_k \in \partial^c f(\theta_k).$$

Chain rule along Lipschitz curves (Brézis, Valadier).
Hypothesis: Fix any Lipschitz 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 $t \in [0,1]$,

Under the carpet: $\alpha_k \to 0$, small step limit $\to$ solutions to the differential inclusion.

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

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

Chain rule along Lipschitz curves (Brézis, Valadier).
Hypothesis: Fix any Lipschitz 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 $0 \notin \partial^c f(\gamma(t))$.

Under the carpet: $\alpha_k \to 0$, small step limit $\to$ solutions to the differential inclusion.

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

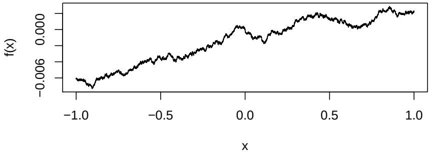
- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).

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.

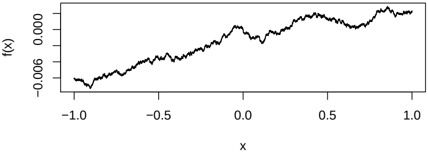
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ 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.

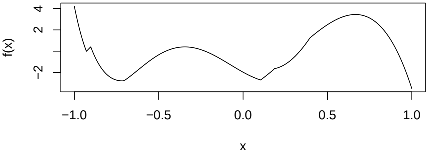


Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ 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.

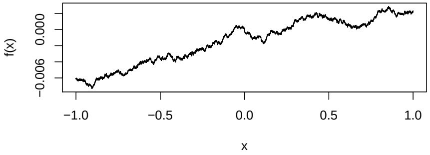


Let f be a *tame* locally Lipschitz function ("generic" in applications),



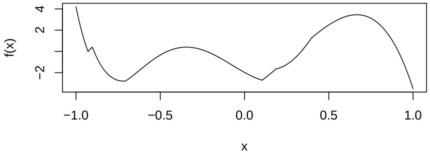
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ 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.



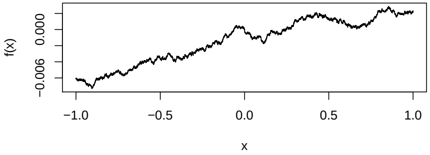
Let f be a *tame* locally Lipschitz function ("generic" in applications),

- piecewise polynomial.



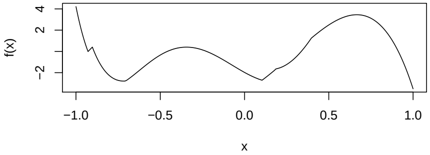
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ 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.



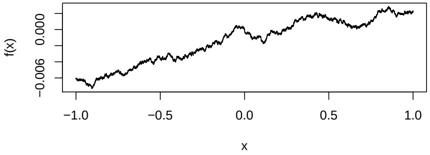
Let f be a *tame* locally Lipschitz function ("generic" in applications),

- piecewise polynomial.
- semi-algebraic.



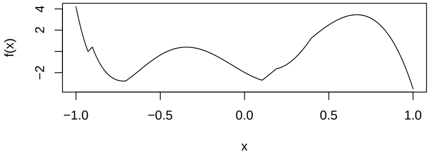
Borwein-Moors (2000), Loewen-Wang (2000): Let f be a typical/generic 1-Lipschitz function (in sup norm), then

- $\partial^c f$ 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.



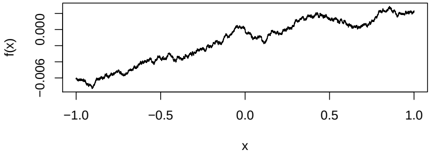
Let f be a *tame* locally Lipschitz function ("generic" in applications),

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



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

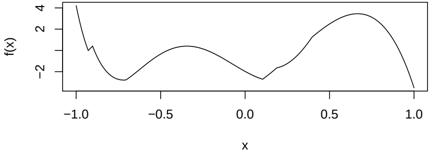
- $\partial^c f$ is the unit ball everywhere (no chain rule, no subgradient algorithm).
- local minimizers are dense: there is a local minimizer arbitrarily close to any argument.



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 Lipschitz curves.



Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$,
For any Lipschitz curve $\gamma : [0,1] \mapsto \mathbb{R}^p$

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

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$,
For any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$

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

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$

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

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz 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 (could be many).
  Conservative Jacobians defined similarly

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz 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 (could be many).
  Conservative Jacobians defined similarly

- **Gradient a.e.:** $D(x) = \{\nabla f(x)\}$ for *almost all* $x \in \mathbb{R}^p$.

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz 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 (could be many).
  Conservative Jacobians defined similarly

- **Gradient a.e.:** $D(x) = \{\nabla f(x)\}$ for *almost all* $x \in \mathbb{R}^p$.
- **Minimal convex conservative gradient:** $\partial^c f(x) \subset \mathrm{conv}(D(x))$ for all $x \in \mathbb{R}^p$.

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz 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 (could be many).
  Conservative Jacobians defined similarly

- **Gradient a.e.:** $D(x) = \{\nabla f(x)\}$ for *almost all* $x \in \mathbb{R}^p$.
- **Minimal convex conservative gradient:** $\partial^c f(x) \subset \mathrm{conv}(D(x))$ for all $x \in \mathbb{R}^p$.
- **Fermat rule:** $0 \in \mathrm{conv}(D(x))$ for all local minima $x \in \mathbb{R}^p$.

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz 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 (could be many). Conservative Jacobians defined similarly

- **Gradient a.e.:** $D(x) = \{\nabla f(x)\}$ for *almost all* $x \in \mathbb{R}^p$.
- **Minimal convex conservative gradient:** $\partial^c f(x) \subset \mathrm{conv}(D(x))$ for all $x \in \mathbb{R}^p$.
- **Fermat rule:** $0 \in \mathrm{conv}(D(x))$ for all local minima $x \in \mathbb{R}^p$.
- **Equivalent caracterization:** f is path-differentiable, if and only if ∂f^c is conservative.

Summary:

- Clarke's subdifferential / Jacobian not compatible with differential calculus.
- Chain rule along Lipschitz curves ensures optimization behavior.

Definition [Conservative gradient] (Bolte-Pauwels 2019):
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz
$D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty valued, locally bounded,
For any Lipschitz 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 (could be many). Conservative Jacobians defined similarly

- **Gradient a.e.:** $D(x) = \{\nabla f(x)\}$ for *almost all* $x \in \mathbb{R}^p$.
- **Minimal convex conservative gradient:** $\partial^c f(x) \subset \mathrm{conv}(D(x))$ for all $x \in \mathbb{R}^p$.
- **Fermat rule:** $0 \in \mathrm{conv}(D(x))$ for all local minima $x \in \mathbb{R}^p$.
- **Equivalent caracterization:** f is path-differentiable, if and only if ∂f^c is conservative.
- **Tame functions are path-differentiable (generic in applications):** chain rule for ∂^c.

Conservative calculus example: finite sums

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t)\rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t)\rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot\gamma(t)\rangle = \left\langle \sum_{i=1}^n v_i, \dot\gamma(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \ldots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot\gamma(t)\rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t)\rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t)\rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot\gamma(t)\rangle = \left\langle \sum_{i=1}^n v_i, \dot\gamma(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \ldots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot\gamma(t)\rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot{\gamma}(t) \rangle = \left\langle \sum_{i=1}^n v_i, \dot{\gamma}(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \ldots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative calculus example: finite sums

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot{\gamma}(t) \rangle = \left\langle \sum_{i=1}^n v_i, \dot{\gamma}(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \ldots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot\gamma(t) \rangle = \left\langle \sum_{i=1}^n v_i, \dot\gamma(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \ldots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot\gamma(t) \rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot\gamma(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot\gamma(t) \rangle = \left\langle \sum_{i=1}^n v_i, \dot\gamma(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \ldots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot\gamma(t) \rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative calculus example: finite sums

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \ldots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \ldots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \ldots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \ldots, n.$$

$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot{\gamma}(t) \rangle = \left\langle \sum_{i=1}^n v_i, \dot{\gamma}(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \, \forall i = 1, ..$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Conservative (outer) sum rule (Bolte-Pauwels 2019):
$f_i \colon \mathbb{R}^p \to \mathbb{R}$ path differentiable (locally Lipschitz), for $i = 1, \dots, n$. Then $D = \sum_i \partial^c f_i$ is conservative for $f = \sum_i f_i$.

Fix any Lipschitz curve $\gamma \colon [0,1] \mapsto \mathbb{R}^p$, for any $i = 1, \dots, n$,

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall t \in E_i, \quad \lambda(E_i^c) = 0$$

Set $E = \cap_i E_i$, we have $\lambda(E^c) = \lambda(\cup_i E_i^c) = 0$.

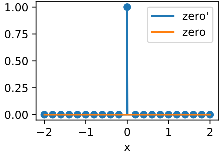
Inversion of quantifiers: for all t in E, $t \in E_i$ for all $i = 1, \dots, n$, that is

$$\frac{d}{dt} f_i(\gamma(t)) = \langle v_i, \dot{\gamma}(t) \rangle \qquad \forall v_i \in \partial^c f_i(\gamma(t)), \qquad \forall i = 1, \dots, n.$$

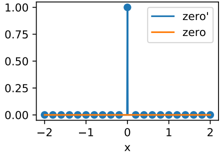
$$\sum_{i=1}^n \frac{d}{dt} f_i(\gamma(t)) = \sum_{i=1}^n \langle v_i, \dot{\gamma}(t) \rangle = \left\langle \sum_{i=1}^n v_i, \dot{\gamma}(t) \right\rangle \quad \forall v_i \in \partial^c f_i(\gamma(t)), \forall i = 1, \dots$$

$$\frac{d}{dt} \sum_{i=1}^n f_i(\gamma(t)) = \frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in \sum_{i=1}^n \partial^c f_i(\gamma(t)) = D(\gamma(t)).$$

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



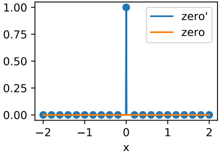
Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



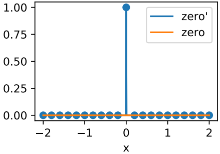
Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0, 1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0, 1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. ($\mathrm{relu}(t) = \max\{0, t\}$).



Calculus,

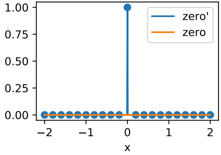
$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0,1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0,1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

More calculus on calculus and chain rule

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. ($\mathrm{relu}(t) = \max\{0, t\}$).



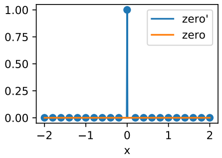
Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0,1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0,1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. ($\mathrm{relu}(t) = \max\{0, t\}$).



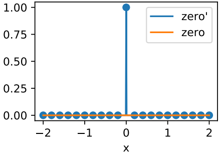
Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0, 1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0, 1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



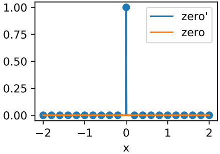
Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0, 1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0, 1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



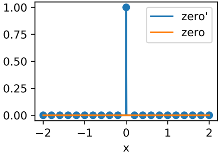
Calculus,

$$D \colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma \colon [0, 1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0, 1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



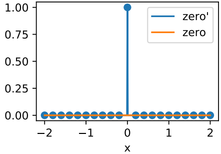
Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0,1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$:
- the set $\{t \in [0,1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



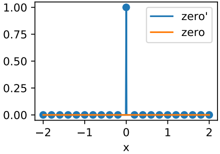
Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0,1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$: for some $\epsilon > 0$, $\gamma(s) \neq 0$ for $s \neq t$ and $s \in (t - \epsilon, t + \epsilon)$.
- the set $\{t \in [0,1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

Artifacts: $\mathrm{zero}(x) = \mathrm{relu}(-x) - \mathrm{relu}(x) + x = 0$. $(\mathrm{relu}(t) = \max\{0, t\})$.



Calculus,

$$D\colon x \rightrightarrows -\partial^c \mathrm{relu}(-x) - \partial^c \mathrm{relu}(x) + \partial^c(x) = \begin{cases} 0 - 1 + 1 = 0 & x > 0 \\ -1 + 0 + 1 = 0 & x < 0 \\ [-1, 0] - [0, 1] + 1 = [-1, 1] & x = 0. \end{cases}$$

Chain rule intuition: $\gamma\colon [0, 1] \to \mathbb{R}$ Lipschitz, differentiable *a.e.*,
Need to check $\frac{d}{dt}\mathrm{zero}(\gamma(t)) = \gamma'(t) \times D(\gamma(t)) = 0$ for almost all t.
Suppose γ differentiable at t:

- $\gamma(t) \neq 0$: $\gamma'(t) \times D(\gamma(t)) = \gamma'(t) \times 0 = 0$. Suppose in addition $\gamma(t) = 0$.
- $\gamma(t) = 0$, $\gamma'(t) = 0$: $\gamma'(t) \times D(\gamma(t)) = 0 \times [-1, 1] = 0$.
- $\gamma(t) = 0$, $\gamma'(t) \neq 0$: for some $\epsilon > 0$, $\gamma(s) \neq 0$ for $s \neq t$ and $s \in (t - \epsilon, t + \epsilon)$.
- the set $\{t \in [0, 1], \gamma(t) = 0, \gamma'(t) \neq 0\}$ is denumerable (zero measure).

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \ldots, g_L$,

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu},$ sort, maxpool, output of nonsmooth numerical program.

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \ldots, g_L$,

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu},$ sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth backprop is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1$$

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \ldots, g_L$,

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu}$, sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth $\mathrm{backprop}$ is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1$$

- **Conservative chain rule:** if $g_1, \ldots, g_L$ are path differentiable, then the set valued field $\mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1 : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is conservative for f.

- Take $f : \mathbb{R}^p \to \mathbb{R}$ Lipschitz expressed from elementary blocks $g_1, \ldots, g_L$,

$$f = g_L \circ \ldots \circ g_1$$

$\boxed{\text{Ex}}$ $g_i = \mathrm{relu}$, sort, maxpool, output of nonsmooth numerical program.

- Nonsmooth $\mathrm{backprop}$ is **formal chain rule:**

$$\mathrm{backprop}_f \in \mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1$$

- **Conservative chain rule:** if $g_1, \ldots, g_L$ are path differentiable, then the set valued field $\mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1 : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is conservative for f.

- Widespread "conservative gradients oracles":

TensorFlow PyTorch JAX

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k \in \mathbb{N}}$ positive sequence

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} = \mathrm{backprop}\,\ell(\theta_k) \in \left(\mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1\right)(\theta_k).$$

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k \in \mathbb{N}}$ positive sequence

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} = \mathrm{backprop}\, \ell(\theta_k) \in \left(\mathrm{Jac}\,^c g_L \circ \ldots \circ \mathrm{Jac}\,^c g_1\right)(\theta_k).$$

Theorem (Bolte-Pauwels 2019-2020):

- **Step size condition:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k \to 0$.
- Accumulation points satisfy $0 \in \mathrm{conv}\left(\mathrm{Jac}\,^c g_L \circ \ldots \circ \mathrm{Jac}\,^c g_1\right)(\theta)$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.
- Same result for any definable conservative gradient instead of $\mathrm{Jac}\,^c g_L \circ \ldots \circ \mathrm{Jac}\,^c g_1$.

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k \in \mathbb{N}}$ positive sequence

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} = \mathrm{backprop}\, \ell(\theta_k) \in \left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta_k).$$

Theorem (Bolte-Pauwels 2019-2020):

- **Step size condition:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k \to 0$.
- Accumulation points satisfy $0 \in \mathrm{conv}\left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta)$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.
- Same result for any definable conservative gradient instead of $\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1$.

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k \in \mathbb{N}}$ positive sequence

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} = \mathrm{backprop}\, \ell(\theta_k) \in \left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta_k).$$

Theorem (Bolte-Pauwels 2019-2020):

- **Step size condition:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k \to 0$.
- Accumulation points satisfy $0 \in \mathrm{conv}\left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta)$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.
- Same result for any definable conservative gradient instead of $\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1$.

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k\in\mathbb{N}}$ positive sequence

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} = \mathrm{backprop}\, \ell(\theta_k) \in \left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta_k).$$

Theorem (Bolte-Pauwels 2019-2020):

- **Step size condition:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k \to 0$.
- Accumulation points satisfy $0 \in \mathrm{conv}\left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta)$
- For "most" such sequences, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.
- Same result for any definable conservative gradient instead of $\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1$.

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

Qualitatively similar results under appropriate assumptions.

- **Subsampling:** at step k sample $i_k \subset \{1, \ldots, n\}$ uniformly at random.

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^c g_{i_k,L} \circ \ldots \circ \mathrm{Jac}^c g_{i_k,1} \right) (\theta_k).$$

- **Incremental:** cycle through each element of the sum, for $i = 1, \ldots, n$

$$\theta_{k,i+1} \in \theta_{k,i} - \alpha_k \left(\mathrm{Jac}^c g_{i,L} \circ \ldots \circ \mathrm{Jac}^c g_{i,1} \right) (\theta_{k,i}).$$

- **Step size:** scalar adaptive step size (adagrad).
- **Algorithms:** discretization of continuous time dynamics with Lyapunov functions (second order INNA, Castera et.al. 2019).

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

Qualitatively similar results under appropriate assumptions.

- **Subsampling:** at step k sample $i_k \subset \{1, \ldots, n\}$ uniformly at random.

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^c g_{i_k,L} \circ \ldots \circ \mathrm{Jac}^c g_{i_k,1}\right)(\theta_k).$$

- **Incremental:** cycle through each element of the sum, for $i = 1, \ldots, n$

$$\theta_{k,i+1} \in \theta_{k,i} - \alpha_k \left(\mathrm{Jac}^c g_{i,L} \circ \ldots \circ \mathrm{Jac}^c g_{i,1}\right)(\theta_{k,i}).$$

- **Step size:** scalar adaptive step size (adagrad).
- **Algorithms:** discretization of continuous time dynamics with Lyapunov functions (second order INNA, Castera et.al. 2019).

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

Qualitatively similar results under appropriate assumptions.

- **Subsampling:** at step k sample $i_k \subset \{1, \ldots, n\}$ uniformly at random.

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^c g_{i_k,L} \circ \ldots \circ \mathrm{Jac}^c g_{i_k,1}\right)(\theta_k).$$

- **Incremental:** cycle through each element of the sum, for $i = 1, \ldots, n$

$$\theta_{k,i+1} \in \theta_{k,i} - \alpha_k \left(\mathrm{Jac}^c g_{i,L} \circ \ldots \circ \mathrm{Jac}^c g_{i,1}\right)(\theta_{k,i}).$$

- **Step size:** scalar adaptive step size (adagrad).
- **Algorithms:** discretization of continuous time dynamics with Lyapunov functions (second order INNA, Castera et.al. 2019).

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

Qualitatively similar results under appropriate assumptions.

- **Subsampling:** at step k sample $i_k \subset \{1, \ldots, n\}$ uniformly at random.

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^{\,c} g_{i_k,L} \circ \ldots \circ \mathrm{Jac}^{\,c} g_{i_k,1}\right)(\theta_k).$$

- **Incremental:** cycle through each element of the sum, for $i = 1, \ldots, n$

$$\theta_{k,i+1} \in \theta_{k,i} - \alpha_k \left(\mathrm{Jac}^{\,c} g_{i,L} \circ \ldots \circ \mathrm{Jac}^{\,c} g_{i,1}\right)(\theta_{k,i}).$$

- **Step size:** scalar adaptive step size (adagrad).
- **Algorithms:** discretization of continuous time dynamics with Lyapunov functions (second order INNA, Castera et.al. 2019).

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

Qualitatively similar results under appropriate assumptions.

- **Subsampling:** at step k sample $i_k \subset \{1, \ldots, n\}$ uniformly at random.

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^c g_{i_k,L} \circ \ldots \circ \mathrm{Jac}^c g_{i_k,1}\right)(\theta_k).$$

- **Incremental:** cycle through each element of the sum, for $i = 1, \ldots, n$

$$\theta_{k,i+1} \in \theta_{k,i} - \alpha_k \left(\mathrm{Jac}^c g_{i,L} \circ \ldots \circ \mathrm{Jac}^c g_{i,1}\right)(\theta_{k,i}).$$

- **Step size:** scalar adaptive step size (adagrad).
- **Algorithms:** discretization of continuous time dynamics with Lyapunov functions (second order INNA, Castera et.al. 2019).

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

Qualitatively similar results under appropriate assumptions.

- **Subsampling:** at step k sample $i_k \subset \{1, \ldots, n\}$ uniformly at random.

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^c g_{i_k,L} \circ \ldots \circ \mathrm{Jac}^c g_{i_k,1}\right)(\theta_k).$$

- **Incremental:** cycle through each element of the sum, for $i = 1, \ldots, n$

$$\theta_{k,i+1} \in \theta_{k,i} - \alpha_k \left(\mathrm{Jac}^c g_{i,L} \circ \ldots \circ \mathrm{Jac}^c g_{i,1}\right)(\theta_{k,i}).$$

- **Step size:** scalar adaptive step size (adagrad).
- **Algorithms:** discretization of continuous time dynamics with Lyapunov functions (second order INNA, Castera et.al. 2019).

Conservative gradients / Jacobians:

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

Conservative gradients / Jacobians:

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

Despite differential calculus artifacts, optimization works with nonsmooth autodiff:

TensorFlow PyTorch

Plan

$f : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$
μ measure on $\mathbb{R}^m$, $f(x, \cdot)$ μ-integrable for all x.
$F : x \mapsto \int_{\mathbb{R}^m} f(x,s) d\mu(s)$.

$f \colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$
μ measure on $\mathbb{R}^m$, $f(x, \cdot)$ μ-integrable for all x.
$F \colon x \mapsto \int_{\mathbb{R}^m} f(x,s) d\mu(s)$.

Inversion integral / derivative:

$x \mapsto f(x,s)$, **smooth**, for all s,

$$\forall (x,s),\ \|\nabla_x f(x,s)\| \leq \kappa(s)$$

for $\kappa : \mathbb{R}^m \to \mathbb{R}_+$, μ integrable.

Gradient of F

$$x \mapsto \int_{\mathbb{R}^m} \nabla f(x,s) d\mu(s)$$

$f\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$
μ measure on $\mathbb{R}^m$, $f(x,\cdot)$ μ-integrable for all x.
$F\colon x \mapsto \int_{\mathbb{R}^m} f(x,s)d\mu(s)$.

Inversion integral / derivative:

$x \mapsto f(x,s)$, **smooth**, for all s,

$$\forall (x,s),\ \|\nabla_x f(x,s)\| \leq \kappa(s)$$

for $\kappa : \mathbb{R}^m \to \mathbb{R}_+$, μ integrable.

Gradient of F

$$x \mapsto \int_{\mathbb{R}^m} \nabla f(x,s)d\mu(s)$$

$f\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$
μ measure on $\mathbb{R}^m$, $f(x,\cdot)$ μ-integrable for all x.
$F\colon x \mapsto \int_{\mathbb{R}^m} f(x,s)d\mu(s)$.

Inversion integral / derivative:

$x \mapsto f(x,s)$, **smooth**, for all s,

$$\forall (x,s),\ \|\nabla_x f(x,s)\| \leq \kappa(s)$$

for $\kappa : \mathbb{R}^m \to \mathbb{R}_+$, μ integrable.

Gradient of F

$$x \mapsto \int_{\mathbb{R}^m} \nabla f(x,s)d\mu(s)$$

Nonsmooth inversion:

$x \mapsto f(x,s)$, **path-differentiable**,

$$\forall (x,s),\ \forall v \in \partial_x^c f(x,s), \quad \|v\| \leq \kappa(s)$$

for $\kappa : \mathbb{R}^m \to \mathbb{R}_+$, μ integrable.

Conservative gradient of F

$$x \rightrightarrows \int_{\mathbb{R}^m} \partial_x^c f(x,s)d\mu(s).$$

Abstract integrals (with Bolte, Le, 2021)

$f : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$
μ measure on $\mathbb{R}^m$, $f(x, \cdot)$ μ-integrable for all x.
$F : x \mapsto \int_{\mathbb{R}^m} f(x, s) d\mu(s)$.

Inversion integral / derivative:

$x \mapsto f(x, s)$, **smooth**, for all s,

$$\forall (x, s),\ \|\nabla_x f(x, s)\| \leq \kappa(s)$$

for $\kappa : \mathbb{R}^m \to \mathbb{R}_+$, μ integrable.

Gradient of F

$$x \mapsto \int_{\mathbb{R}^m} \nabla f(x, s) d\mu(s)$$

Nonsmooth inversion:

$x \mapsto f(x, s)$, **path-differentiable**,

$$\forall (x, s),\ \forall v \in \partial_x^c f(x, s), \quad \|v\| \leq \kappa(s)$$

for $\kappa : \mathbb{R}^m \to \mathbb{R}_+$, μ integrable.

Conservative gradient of F

$$x \rightrightarrows \int_{\mathbb{R}^m} \partial_x^c f(x, s) d\mu(s).$$

Applications: Stochastic optimization, chain rule for parametric integrals (assumption).

Ordinary differential equations (with Marx, 2022)

$F\colon \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz

$$\frac{d}{dt}X(t,\theta) = F(X(t,\theta))$$
$$X(0) = \theta \in \mathbb{R}^m.$$

$F\colon \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz

$$\frac{d}{dt}X(t,\theta) = F(X(t,\theta))$$
$$X(0) = \theta \in \mathbb{R}^m.$$

Sensitivity equation:

F, **smooth**.

$$\frac{d}{dt}M(t,\theta) = \mathrm{Jac}\, F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (1)$$

$\theta \mapsto X(t,\theta)$ is **smooth**, **Jacobian**:

$$\theta \mapsto M(t,\theta), \quad \text{s.t.} \quad M \text{ solution to (1)}.$$

$F \colon \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz

$$\frac{d}{dt}X(t,\theta) = F(X(t,\theta))$$
$$X(0) = \theta \in \mathbb{R}^m.$$

Sensitivity equation:

F, **smooth**.

$$\frac{d}{dt}M(t,\theta) = \mathrm{Jac}\, F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (1)$$

$\theta \mapsto X(t,\theta)$ is **smooth**, **Jacobian**:

$\theta \mapsto M(t,\theta), \quad \text{s.t.} \quad M$ solution to (1).

$F\colon \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz

$$\frac{d}{dt}X(t,\theta) = F(X(t,\theta))$$
$$X(0) = \theta \in \mathbb{R}^m.$$

Sensitivity equation:

F, **smooth**.

$$\frac{d}{dt}M(t,\theta) = \mathrm{Jac}\, F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (1)$$

$\theta \mapsto X(t,\theta)$ is **smooth**, **Jacobian**:

$\theta \mapsto M(t,\theta), \quad \text{s.t.} \quad M$ solution to (1).

$F\colon \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz

$$\frac{d}{dt}X(t,\theta) = F(X(t,\theta))$$
$$X(0) = \theta \in \mathbb{R}^m.$$

Sensitivity equation:

F, **smooth**.

$$\frac{d}{dt}M(t,\theta) = \mathrm{Jac}\, F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (1)$$

$\theta \mapsto X(t,\theta)$ is **smooth**, **Jacobian**:

$\theta \mapsto M(t,\theta), \quad \text{s.t.} \quad M$ solution to (1).

Nonsmooth sensitivity equation:

F, **path differentiable**.

$$\frac{d}{dt}M(t,\theta) \in \mathrm{Jac}^{\,c} F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (2)$$

Conservative jacobian of $\theta \mapsto X(t,\theta)$

$\theta \rightrightarrows \{M(t,\theta), \quad \forall M$ solution to (2).$\}$

$F\colon \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz

$$\frac{d}{dt}X(t,\theta) = F(X(t,\theta))$$
$$X(0) = \theta \in \mathbb{R}^m.$$

Sensitivity equation:

F, **smooth**.

$$\frac{d}{dt}M(t,\theta) = \mathrm{Jac}\, F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (1)$$

$\theta \mapsto X(t,\theta)$ is **smooth**, **Jacobian**:

$\theta \mapsto M(t,\theta), \quad \text{s.t.} \quad M$ solution to (1).

Nonsmooth sensitivity equation:

F, **path differentiable**.

$$\frac{d}{dt}M(t,\theta) \in \mathrm{Jac}^{\,c} F(X(t,\theta))M(t,\theta)$$
$$M(0) = I \in \mathbb{R}^{m\times m}. \qquad (2)$$

Conservative jacobian of $\theta \mapsto X(t,\theta)$

$\theta \rightrightarrows \{M(t,\theta), \quad \forall M$ solution to (2).$\}$

Applications: Neural ODE, adjoint method, optimization under ODE constraints.

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $F(\hat{\theta}, \hat{x}) = 0$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $F(\hat{\theta}, \hat{x}) = 0$

Classical implicit differentiation:

F **smooth,** assume

$$[A, B] = \mathrm{Jac}\, F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions to $F(\theta, x) = 0$ locally parametrized by $G : U \to \mathbb{R}^n$, **smooth:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **jacobian** of G:

$$\theta \to -B^{-1}A : [A, B] = \mathrm{Jac}\, F(\theta, G(\theta)).$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $F(\hat{\theta}, \hat{x}) = 0$

Classical implicit differentiation:

F **smooth,** assume

$$[A, B] = \mathrm{Jac}\, F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions to $F(\theta, x) = 0$ locally parametrized by $G : U \to \mathbb{R}^n$, **smooth:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **jacobian** of G:

$$\theta \to -B^{-1}A : [A, B] = \mathrm{Jac}\, F(\theta, G(\theta)).$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $F(\hat{\theta}, \hat{x}) = 0$

Classical implicit differentiation:

F **smooth,** assume

$$[A, B] = \mathrm{Jac}\, F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions to $F(\theta, x) = 0$ locally parametrized by $G : U \to \mathbb{R}^n$, **smooth:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **jacobian** of G:

$$\theta \to -B^{-1}A : [A, B] = \mathrm{Jac}\, F(\theta, G(\theta)).$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $F(\hat{\theta}, \hat{x}) = 0$

Classical implicit differentiation:

F **smooth,** assume

$$[A, B] = \mathrm{Jac}\, F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions to $F(\theta, x) = 0$ locally parametrized by $G : U \to \mathbb{R}^n$, **smooth:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **jacobian** of G:

$$\theta \to -B^{-1}A : [A, B] = \mathrm{Jac}\, F(\theta, G(\theta)).$$

Nonsmooth implicit differentiation:

F **path differentiable**, assume

$$\forall [A, B] \in \mathrm{Jac}^{\,c} F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions locally parametrized by $G : U \to \mathbb{R}^n$, **path-differentiable:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **conservative jacobian** for G:

$$\theta \rightrightarrows \left\{ -B^{-1}A : [A\ B] \in \mathrm{Jac}^c_F(\theta, G(\theta)) \right\}.$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $F(\hat{\theta}, \hat{x}) = 0$

Classical implicit differentiation:

F **smooth,** assume

$$[A, B] = \mathrm{Jac}\, F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions to $F(\theta, x) = 0$ locally parametrized by $G : U \to \mathbb{R}^n$, **smooth:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **jacobian** of G:

$$\theta \to -B^{-1}A : [A, B] = \mathrm{Jac}\, F(\theta, G(\theta)).$$

Nonsmooth implicit differentiation:

F **path differentiable**, assume

$$\forall [A, B] \in \mathrm{Jac}^{\,c} F(\hat{\theta}, \hat{x}), \quad B \text{ invertible.}$$

Solutions locally parametrized by $G : U \to \mathbb{R}^n$, **path-differentiable:**

$$F(\theta, G(\theta)) = 0.$$

Implicit **conservative jacobian** for G:

$$\theta \rightrightarrows \left\{ -B^{-1}A : [A \ B] \in \mathrm{Jac}^c_F(\theta, G(\theta)) \right\}.$$

Applications: Differentiate $G(x)$ uniquely defined as $F(x, G(x)) = 0$.
parametric optimization, bilevel optimization, implicit modeling, hyperparameter tuning.

Algorithmic unrolling (with Bolte, Vaiter, 2022)

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

For all θ, $x \to F(x, \theta)$ is ρ Lipschitz, $\rho < 1$: $\quad x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta).$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

For all θ, $x \to F(x, \theta)$ is ρ Lipschitz, $\rho < 1$: $x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta)$.

Classical asymptotics (Gilbert 92):

F **smooth**.

Forward **jacobian** propagation:

$$\mathrm{Jac}\, x_{k+1}(\theta) = B \mathrm{Jac}\, x_k(\theta) + A$$
$$[A, B] = \mathrm{Jac}\, F(\theta, x_k(\theta))$$

Limiting **jacobian.**

$$\mathrm{Jac}\, x_k(\theta) \underset{k\to\infty}{\to} \mathrm{Jac}\, \bar{x}(\theta)$$
$$= (I - B)^{-1} A,\ [A, B] = \mathrm{Jac}\, F(\theta, \bar{x}(\theta))$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

For all θ, $x \to F(x, \theta)$ is ρ Lipschitz, $\rho < 1$: $\quad x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta).$

Classical asymptotics (Gilbert 92):

F **smooth**.

Forward **jacobian** propagation:

$$\begin{aligned} \mathrm{Jac}\, x_{k+1}(\theta) &= B \mathrm{Jac}\, x_k(\theta) + A \\ [A, B] &= \mathrm{Jac}\, F(\theta, x_k(\theta)) \end{aligned}$$

Limiting **jacobian.**

$$\mathrm{Jac}\, x_k(\theta) \underset{k\to\infty}{\to} \mathrm{Jac}\, \bar{x}(\theta)$$

$$= (I - B)^{-1} A,\ [A, B] = \mathrm{Jac}\, F(\theta, \bar{x}(\theta))$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

For all θ, $x \to F(x, \theta)$ is ρ Lipschitz, $\rho < 1$: $\qquad x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta).$

Classical asymptotics (Gilbert 92):

F **smooth**.

Forward **jacobian** propagation:

$$\begin{aligned} \mathrm{Jac}\, x_{k+1}(\theta) &= B \mathrm{Jac}\, x_k(\theta) + A \\ [A, B] &= \mathrm{Jac}\, F(\theta, x_k(\theta)) \end{aligned}$$

Limiting **jacobian.**

$$\mathrm{Jac}\, x_k(\theta) \underset{k\to\infty}{\to} \mathrm{Jac}\, \bar{x}(\theta)$$
$$= (I - B)^{-1} A,\ [A, B] = \mathrm{Jac}\, F(\theta, \bar{x}(\theta))$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

For all θ, $x \to F(x, \theta)$ is ρ Lipschitz, $\rho < 1$: $\quad x_k(\theta) \underset{k \to \infty}{\to} \bar{x}(\theta).$

Classical asymptotics (Gilbert 92):

F **smooth**.

Forward **jacobian** propagation:

$$\begin{aligned} \mathrm{Jac}\, x_{k+1}(\theta) &= B \mathrm{Jac}\, x_k(\theta) + A \\ [A, B] &= \mathrm{Jac}\, F(\theta, x_k(\theta)) \end{aligned}$$

Limiting **jacobian.**

$$\mathrm{Jac}\, x_k(\theta) \underset{k \to \infty}{\to} \mathrm{Jac}\, \bar{x}(\theta)$$
$$= (I - B)^{-1} A,\ [A, B] = \mathrm{Jac}\, F(\theta, \bar{x}(\theta))$$

Nonsmooth unrolling :

F **path-differentiable**.

Conservative jacobian propagation:

$$D_{k+1}(\theta) = \big\{ B D_k(\theta) + A \quad [A, B] \in \mathrm{Jac}^{\,c} F(\theta, x_k(\theta)) \big\}$$

Limiting **conservative jacobian:**

$$D_k(\theta) \underset{k \to \infty}{\to} \bar{D}(\theta)$$
$$\supset \left\{ (I - B)^{-1} A, \quad [A, B] \in \mathrm{Jac}^{\,c} F(\theta, \bar{x}(\theta)) \right\}$$

$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^n$

$$x_{k+1}(\theta) = F(\theta, x_k(\theta)).$$

For all θ, $x \to F(x, \theta)$ is ρ Lipschitz, $\rho < 1$: $\quad x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta)$.

Classical asymptotics (Gilbert 92):

F **smooth**.

Forward **jacobian** propagation:

$$\begin{aligned} \mathrm{Jac}\, x_{k+1}(\theta) &= B \mathrm{Jac}\, x_k(\theta) + A \\ [A, B] &= \mathrm{Jac}\, F(\theta, x_k(\theta)) \end{aligned}$$

Limiting **jacobian.**

$$\mathrm{Jac}\, x_k(\theta) \underset{k\to\infty}{\to} \mathrm{Jac}\, \bar{x}(\theta)$$

$$= (I - B)^{-1} A,\ [A, B] = \mathrm{Jac}\, F(\theta, \bar{x}(\theta))$$

Nonsmooth unrolling :

F **path-differentiable**.

Conservative jacobian propagation:

$$\begin{aligned} D_{k+1}(\theta) = \big\{ & B D_k(\theta) + A \\ & [A, B] \in \mathrm{Jac}^c F(\theta, x_k(\theta)) \big\} \end{aligned}$$

Limiting **conservative jacobian:**

$$D_k(\theta) \underset{k\to\infty}{\to} \bar{D}(\theta)$$

$$\supset \left\{ (I - B)^{-1} A, \quad [A, B] \in \mathrm{Jac}^c F(\theta, \bar{x}(\theta)) \right\}$$

Applications: Differentiation of forward-backward, Douglas-Rachford, ADMM).

Plan

Initial motivation an results:

- study nonsmooth automatic differentiation.
- compositional calculus rules: sum, product, composition.
- require chain rule along Lipschitz curves: ubiquitous in applications.
- optimization: qualitative convergence of first order methods.

Extensions:

- Optimization algorithm variations.
- Extensions of conservative calculus.

Not presented

- Proof details.
- Parametric optimality for max structured functions.
- Complexity considerations (with Bolte, Boustany, Pesquet-Popescu)

Thanks.

Initial motivation an results:

- study nonsmooth automatic differentiation.
- compositional calculus rules: sum, product, composition.
- require chain rule along Lipschitz curves: ubiquitous in applications.
- optimization: qualitative convergence of first order methods.

Extensions:

- Optimization algorithm variations.
- Extensions of conservative calculus.

Not presented

- Proof details.
- Parametric optimality for max structured functions.
- Complexity considerations (with Bolte, Boustany, Pesquet-Popescu)

Thanks.

Initial motivation an results:

- study nonsmooth automatic differentiation.
- compositional calculus rules: sum, product, composition.
- require chain rule along Lipschitz curves: ubiquitous in applications.
- optimization: qualitative convergence of first order methods.

Extensions:

- Optimization algorithm variations.
- Extensions of conservative calculus.

Not presented

- Proof details.
- Parametric optimality for max structured functions.
- Complexity considerations (with Bolte, Boustany, Pesquet-Popescu)

Thanks.

Ongoing work on "conservative calculus"

Initial motivation an results:

- study nonsmooth automatic differentiation.
- compositional calculus rules: sum, product, composition.
- require chain rule along Lipschitz curves: ubiquitous in applications.
- optimization: qualitative convergence of first order methods.

Extensions:

- Optimization algorithm variations.
- Extensions of conservative calculus.

Not presented

- Proof details.
- Parametric optimality for max structured functions.
- Complexity considerations (with Bolte, Boustany, Pesquet-Popescu)

Thanks.

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k\in\mathbb{N}}$ positive sequence

$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta_k).$$

Theorem (Bolte-Pauwels 2020):

- **Step size condition:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k \to 0$.
- Accumulation points satisfy $0 \in \mathrm{conv}\left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta)$
- There is a meagre Lebesgue null set X_0 and finite set $\Lambda \in \mathbb{R}_+$ such that if $\theta_0 \notin X_0$ and $\alpha_k \notin \Lambda$, $k \in \mathbb{N}$, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := g_L \circ \ldots \circ g_1(\theta)$$

Assumption:

- g_i is locally Lipschitz tame (piecewise polynomial, semi-algebraic, definable).

First order algorithm: fix $\theta_0 \in \mathbb{R}^p$, $(\alpha_k)_{k\in\mathbb{N}}$ positive sequence

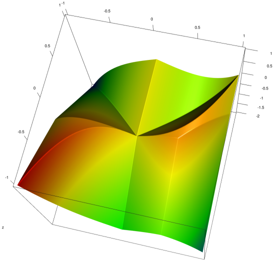
$$\theta_{k+1} \in \theta_k - \alpha_k \left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta_k).$$

Theorem (Bolte-Pauwels 2020):

- **Step size condition:** $\sum_{k=1}^{+\infty} \alpha_k = +\infty$ and $\alpha_k \to 0$.
- Accumulation points satisfy $0 \in \mathrm{conv}\left(\mathrm{Jac}^{\,c} g_L \circ \ldots \circ \mathrm{Jac}^{\,c} g_1\right)(\theta)$
- There is a meagre Lebesgue null set X_0 and finite set $\Lambda \in \mathbb{R}_+$ such that if $\theta_0 \notin X_0$ and $\alpha_k \notin \Lambda$, $k \in \mathbb{N}$, accumulation points are Clarke critical $0 \in \partial^c \ell(\theta)$.

Semi-algebraic?

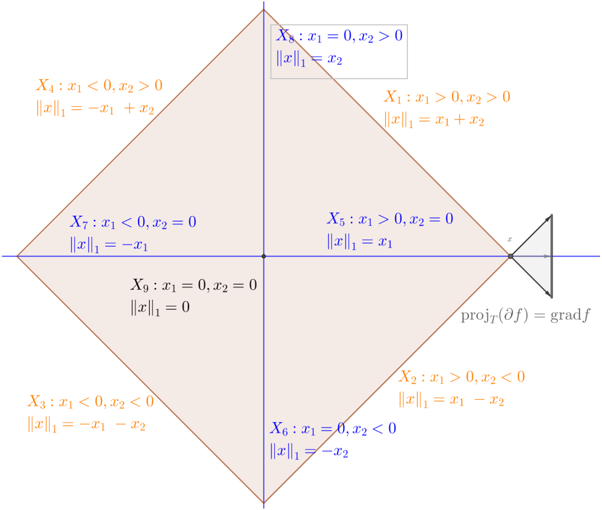
Basic set: Solution set of finitely many polynomial inequalities.
Set: Finite union of Basic semi-algebraic sets.
Function, set valued map: Semi-algebraic graph.
Examples: polynomials, square root, quotients, norm, relu, rank ...



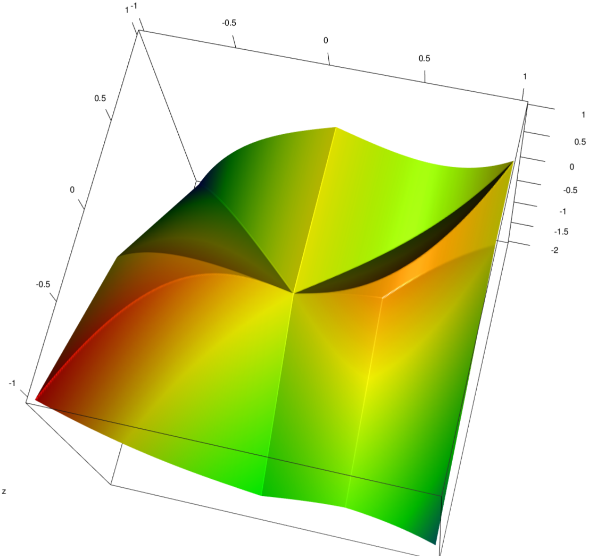
Tarski Seidenberg: first order formula involving semi-algebraic sets $\rightarrow$ semi-algebraic.

- gradient / subgradient of semi-algebraic function, partial minima, composition ....

Variational stratification: [Bolte-Daniilidis-Lewis (2007)]
Example: Projection formula .



Variational stratification: [Bolte-Daniilidis-Lewis (2007)]
Example: Projection formula $f(x_1, x_2) = |x_1| + |x_2|$.



Let $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a semi-algebraic (or definable), graph closed, locally bounded and $f \colon \mathbb{R}^p \to \mathbb{R}$, $r \in \mathbb{N}^*$. Then the following are equivalent

- D is a conservative field for f.
- (f, D) has a C^r variational stratification: there exists a stratification $\{M_i\}_{i \in I}$ of $\mathbb{R}^p$ such that
  - The restriction f_{M_i} of f to M_i is C^r for all $i \in I$.
  - For all $x \in \mathbb{R}^p$, set M_x the active stratum, T_x its tangent space at x.

$$P_{T_x} D(x) = \{\mathrm{grad}\ f_{M_x}(x)\}.$$

Whitney stratification: finite partition of $\mathbb{R}^p$ into C^r embedded manifolds (+ technical condition).

Applies to backprop:

- Morse-Sard condition.
- artefacts are "negligible" in a geometric sense.