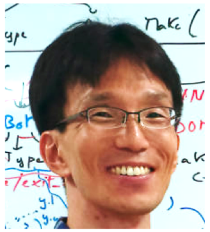
On Correctness of Automatic Differentiation for Non-Differentiable Functions

<u>Wonyeol Lee</u>[1,*]   Hangyeol Yu[1,**]   Xavier Rival[2]   Hongseok Yang[1]

[1]KAIST, South Korea   [2]INRIA/ENS/CNRS, France

[*]now at Stanford, USA   [**]now at Riiid!, South Korea

NeurIPS 2020 (Spotlight)

Autodiff

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,

how to compute $\nabla F(x)$ correctly and efficiently?

Autodiff

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,

how to compute $\nabla F(x)$ correctly and efficiently?

<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x)$ for every $x \in \mathbb{R}^n$.

Autodiff

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,

how to compute $\nabla F(x)$ correctly and efficiently?

Autodiff $\approx$ efficient way of applying the chain rule.

Chain Rule For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

$$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$

Autodiff

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,
how to compute $\nabla F(x)$ correctly and efficiently?

$\uparrow$

<u>Theorem</u> F_l's are differentiable everywhere $\Longrightarrow$ autodiff correctly computes $\nabla F(x)$.

Autodiff $\approx$ efficient way of applying the chain rule.

<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

$$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$

Autodiff in Practice

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,
how to compute $\nabla F(x)$ correctly and efficiently?

?

<u>Theorem</u> F_l's are $\boxed{\text{differentiable everywhere}}$ $\implies$ autodiff correctly computes $\nabla F(x)$.

Autodiff $\approx$ efficient way of applying the chain rule.

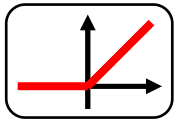
<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,
$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x)$ for every $x \in \mathbb{R}^n$.

Autodiff in Practice

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,

how to compute $\nabla F(x)$ correctly and efficiently?

<u>Theorem</u> F_l's are ~~differentiable everywhere~~ $\Longrightarrow$ autodiff correctly computes $\nabla F(x)$.

e.g., $\mathrm{ReLU}(x) = \max\{x, 0\} =$ 

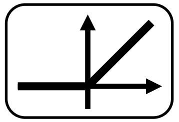
<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x)$ for every $x \in \mathbb{R}^n$.

Autodiff in Practice

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$, how to compute $\nabla F(x)$ correctly and efficiently?

<u>Theorem</u> F_l's are ~~differentiable everywhere~~ $\implies$ autodiff correctly computes $\nabla F(x)$.

e.g., $\mathrm{ReLU}(x) = \max\{x, 0\} =$ 

non-differentiable on a measure-zero set

<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

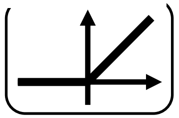
$$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$

Autodiff in Practice

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,

how to compute $\nabla F(x)$ correctly and efficiently?

<u>Theorem</u> F_l's are differentiable $\underbrace{\text{everywhere}}_{\text{almost-}}$ $\boxed{\overset{?}{\Longrightarrow}}$ $\underbrace{\text{autodiff correctly computes } \nabla F(x)}_{\text{almost-everywhere}}$.

e.g., $\mathrm{ReLU}(x) = \max\{x, 0\} =$ 

non-differentiable on a measure-zero set

<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

$$D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$

Autodiff in Practice

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$, how to compute $\nabla F(x)$ correctly and efficiently?

Theorem F_l's are differentiable everywhere $\boxed{\overset{?}{\Longrightarrow}}$ autodiff correctly computes $\nabla F(x)$.

almost-

almost-everywhere

almost-

Chain Rule For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable everywhere,

$$\boxed{D(g \circ f)(x) \overset{?}{=} Dg\big(f(x)\big) \cdot Df(x)} \quad \text{for every } x \in \mathbb{R}^n.$$

almost-

Our Results

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,
how to compute $\nabla F(x)$ correctly and efficiently?

Theorem F_l's are $\underbrace{\text{differentiable everywhere}}_{\text{almost-}}$ $\cancel{\Longrightarrow}$ autodiff correctly computes $\underbrace{\nabla F(x)}_{\text{almost-everywhere}}$.

No, measure-zero non-differentiabilities matter!

Chain Rule For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable $\overbrace{\text{everywhere}}^{\text{almost-}}$,
$\cancel{D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)}$ for $\underbrace{\text{every}}_{\text{almost-}}$ $x \in \mathbb{R}^n$.

Our Results

<u>Problem</u> For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$, how to compute $\nabla F(x)$ correctly and efficiently?

<u>Theorem</u> F_l's are differentiable $\underbrace{\text{everywhere}}_{\text{almost-}}$ $\cancel{\Longrightarrow}$ autodiff correctly computes $\underbrace{\nabla F(x)}_{\text{almost-everywhere}}$.

<u>Our Result</u> Disprove this and related claims.

<u>Chain Rule</u> For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable $\overbrace{\text{everywhere}}^{\text{almost-}}$,

$$\cancel{D(g \circ f)(x) = Dg\big(f(x)\big) \cdot Df(x)} \quad \text{for } \underbrace{\text{every}}_{\text{almost-}} \ x \in \mathbb{R}^n.$$

Subtleties in Chain Rule (1)

Claim 1 For any $f, g : \mathbb{R} \to \mathbb{R}$,

f, g : a.e.-differentiable and continuous

$\boxed{\Longrightarrow}$? $(g \circ f)'(x) = g'\big(f(x)\big) \cdot f'(x)$ for a.e. $x \in \mathbb{R}$.

Subtleties in Chain Rule (1)

<u>Claim 1</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

f, g : a.e.-differentiable and continuous

$\boxed{\Longrightarrow}$? $\boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)}$ for a.e. $x \in \mathbb{R}$.

well-defined?

Subtleties in Chain Rule (1)

Claim 1 For any $f, g : \mathbb{R} \to \mathbb{R}$,

f, g : a.e.-differentiable and continuous

$\boxed{\Longrightarrow}$? $\quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)}$ for a.e. $x \in \mathbb{R}$.

? well-defined? O

Subtleties in Chain Rule (1)

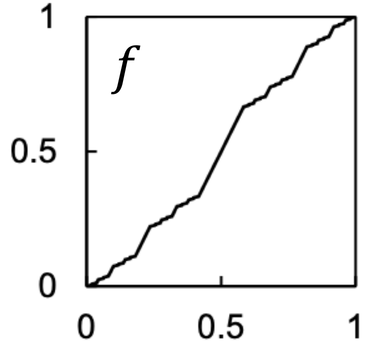
<u>Claim 1</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

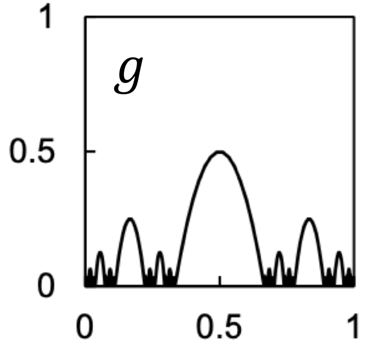
f, g : a.e.-differentiable and continuous

$$\not\Longrightarrow \quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \qquad \text{for a.e. } x \in \mathbb{R}.$$

well-defined?

<u>Counterexample</u> Involves the Cantor function.





Subtleties in Chain Rule (2)

Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

$$\boxed{\Longrightarrow}^{?} \qquad (g \circ f)'(x) = g'\big(f(x)\big) \cdot f'(x) \qquad \text{for a.e. } x \in \mathbb{R}.$$

Subtleties in Chain Rule (2)

Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$,

$$\underbrace{f, g}_{\text{and } g \circ f}\text{: a.e.-differentiable and continuous}$$

$$\boxed{\Longrightarrow}\,? \quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \quad \text{for a.e. } x \in \mathbb{R}.$$

well-defined?

Subtleties in Chain Rule (2)

<u>Claim 2</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

$\boxed{\Longrightarrow}$? $\quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \quad$ for a.e. $x \in \mathbb{R}$.

?

well-defined?

Subtleties in Chain Rule (2)

Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$,

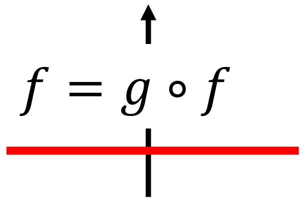
$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous $\cdots (*)$

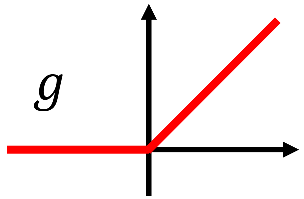
$\boxed{\Longrightarrow}$? $\boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)}$ for a.e. $x \in \mathbb{R}$.

? well-defined?

Counterexample $f(x) = 0$ and $g(y) = \mathrm{ReLU}(y)$.

$\Longrightarrow$ easy to check that $(*)$ holds.





Subtleties in Chain Rule (2)

<u>Claim 2</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

$$\cancel{\Longrightarrow} \quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \qquad \text{for a.e. } x \in \mathbb{R}.$$

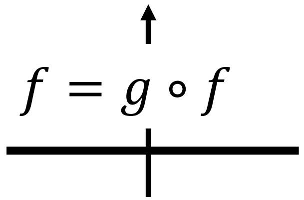
well-defined?

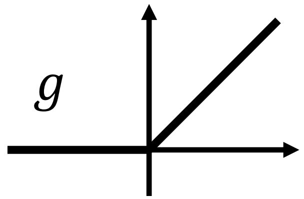
<u>Counterexample</u> $f(x) = 0$ and $g(y) = \mathrm{ReLU}(y)$.

$\Longrightarrow$ $g'\big(f(x)\big)$

$= g'(0)$

$=$ undefined for all x





Subtleties in Chain Rule (2)

Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

$$\not\Longrightarrow \quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \quad \text{for a.e. } x \in \mathbb{R}.$$

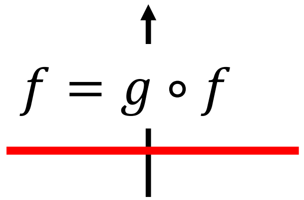
well-defined?

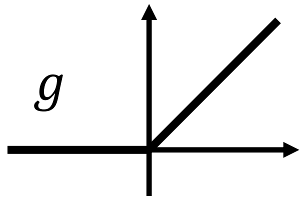
Counterexample $f(x) = 0$ and $g(y) = \mathrm{ReLU}(y)$.

$\Longrightarrow$ $(g \circ f)'(x)$ $\quad g'\big(f(x)\big)$ $\quad f'(x)$

$= 0$ $\quad = g'(0)$ $\quad = 0$

$= \text{undefined for all } x$





Subtleties in Chain Rule (2)

<u>Claim 2</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

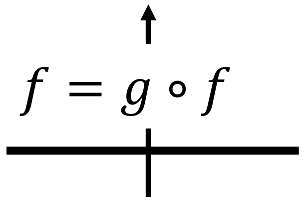
$$\cancel{\Longrightarrow} \quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \quad \text{for a.e. } x \in \mathbb{R}.$$

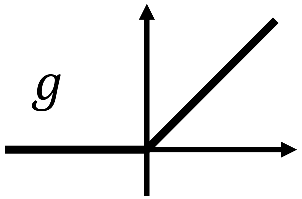
well-defined?

<u>Counterexample</u> $f(x) = 0$ and $g(y) = \mathrm{ReLU}(y)$.

$$\Longrightarrow \quad \underbrace{(g \circ f)'(x)}_{= 0} \qquad {\color{red} dg}\big(f(x)\big) \qquad \underbrace{f'(x)}_{= 0}$$

$${\color{red} dg(y)} = \begin{cases} {\color{red} 7} & {\color{red} \text{for } y = 0} \\ g'(y) & \text{for } y \neq 0 \end{cases}$$





Subtleties in Chain Rule (2)

Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

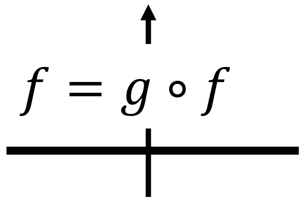
$$\not\Longrightarrow \quad \boxed{(g \circ f)'(x)} = \boxed{g'\big(f(x)\big)} \cdot \boxed{f'(x)} \quad \text{for a.e. } x \in \mathbb{R}.$$

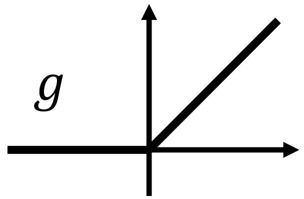
well-defined?

Counterexample $f(x) = 0$ and $g(y) = \mathrm{ReLU}(y)$.

$$\Longrightarrow \quad \boxed{\underbrace{(g \circ f)'(x)}_{=0} = dg\big(f(x)\big) \cdot \underbrace{f'(x)}_{=0}} \quad \text{for all } x \in \mathbb{R}.$$

$$dg(y) = \begin{cases} 7 & \text{for } y = 0 \\ g'(y) & \text{for } y \neq 0 \end{cases}$$





Subtleties in Chain Rule (3)

Claim 3 For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\underbrace{f, g}_{}$ and $g \circ f$: a.e.-differentiable and continuous

$\boxed{\Longrightarrow}$?

$$(g \circ f)'(x) = dg\big(f(x)\big) \cdot df(x) \qquad \text{for a.e. } x \in \mathbb{R}.$$

$\exists\, df, dg : \mathbb{R} \to \mathbb{R}$ such that $df \overset{\text{a.e.}}{=} f'$, $dg \overset{\text{a.e.}}{=} g'$, and

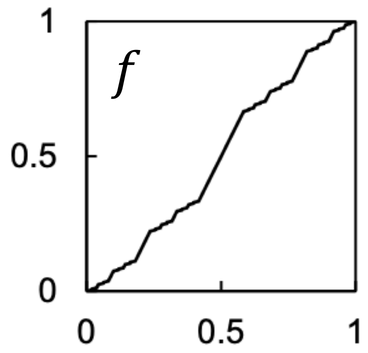
Subtleties in Chain Rule (3)

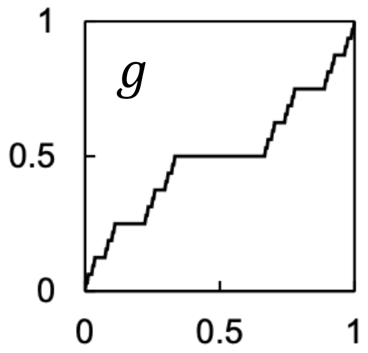
<u>Claim 3</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

$$\cancel{\Longrightarrow} \quad \underbrace{(g \circ f)'(x) \cancel{=} dg\big(f(x)\big) \cdot df(x)}_{\exists\, df, dg \,:\, \mathbb{R} \to \mathbb{R} \text{ such that } df \overset{\text{a.e.}}{=} f',\; dg \overset{\text{a.e.}}{=} g', \text{ and}} \qquad \text{for a.e. } x \in \mathbb{R}.$$

<u>Counterexample</u> Involves the Cantor function again.





Our Results

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,

how to compute $\nabla F(x)$ correctly and efficiently?

Theorem F_l's are $\underbrace{\text{differentiable everywhere}}_{\text{almost-}}$ $\cancel{\Longrightarrow}$ $\underbrace{\text{autodiff correctly computes } \nabla F(x)}_{\text{almost-everywhere}}$.

Our Result Disprove this and related claims.

Chain Rule For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ differentiable $\overbrace{\text{everywhere}}^{\text{almost-}}$,

$\cancel{D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)}$ for $\underbrace{\text{every } x \in \mathbb{R}^n}_{\text{almost-}}$.

Our Results

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$,
how to compute $\nabla F(x)$ correctly and efficiently?

Question How to recover this claim?

Theorem F_l's are differentiable everywhere $\cancel{\Longrightarrow}$ autodiff correctly computes $\nabla F(x)$.

? almost-

almost-everywhere

✓

Our Result Disprove this and related claims.

almost-

Chain Rule For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$, differentiable everywhere,

$\cancel{D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)}$ for every $x \in \mathbb{R}^n$.

almost-

Our Results

Problem For $F : \mathbb{R}^N \to \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$, how to compute $\nabla F(x)$ correctly and efficiently?

Our Result Prove this claim for a wide class of F_l's.

Theorem F_l's are ~~differentiable everywhere~~ $\boxed{\Rightarrow}$ autodiff correctly computes $\nabla F(x)$.

~~almost-~~

so-called "PAP"

almost-everywhere

Our Result Disprove this and related claims.

Chain Rule For $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$, differentiable everywhere,

almost-

$\cancel{D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)}$ for every $x \in \mathbb{R}^n$.

almost-

PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition)

roughly iff f can be "decomposed" into $f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$ such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is "analytic".

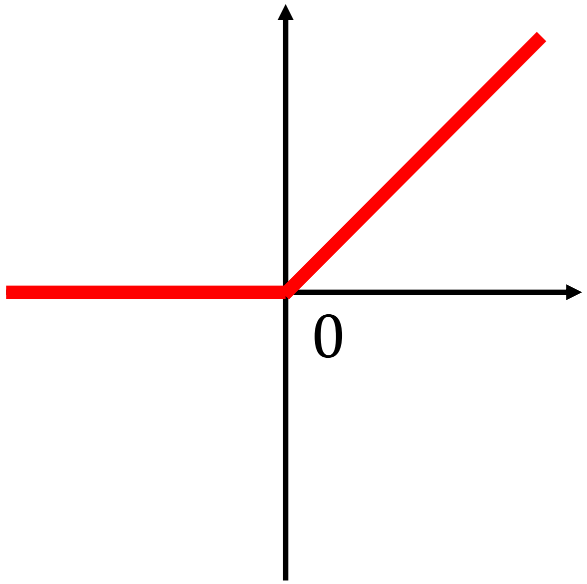
PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition)

roughly iff f can be "decomposed" into $f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$ such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is "analytic".

Example $f(x) = \mathrm{ReLU}(x)$.



PAP Functions

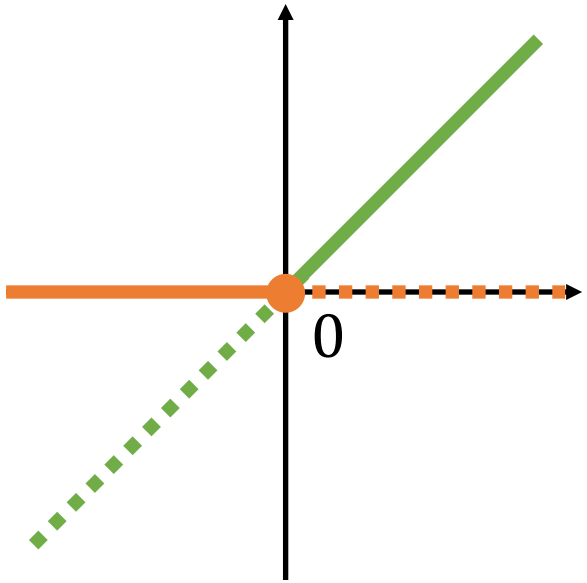
Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition)

roughly iff f can be "decomposed" into $f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$ such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is "analytic".

Example $f(x) = \mathrm{ReLU}(x)$.

- $(f_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \le 0\})$,
  $(f_2(x) = x, A_2 = \{x \in \mathbb{R} : x > 0\})$.



PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition)

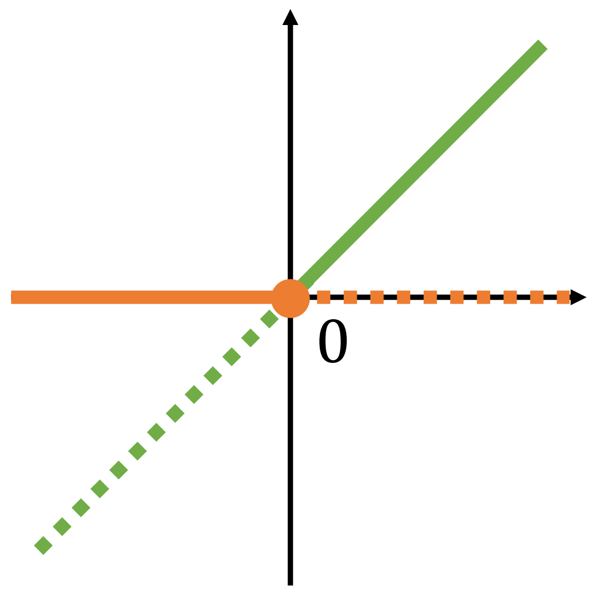
roughly iff f can be "decomposed" into $f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$ such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is "analytic".

Example $f(x) = \mathrm{ReLU}(x)$.

- $(f_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\})$,
  $(f_2(x) = x, A_2 = \{x \in \mathbb{R} : x > 0\})$.

analytic functions



PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition)

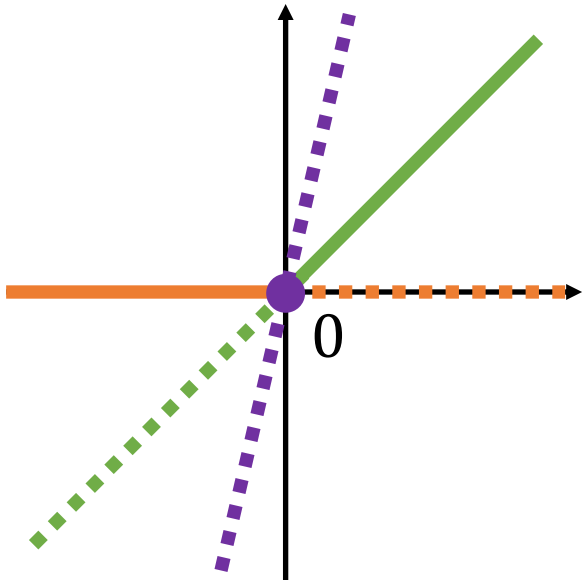
roughly iff f can be "decomposed" into $f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$ such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is "analytic".

Example $f(x) = \mathrm{ReLU}(x)$.

- $(f_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\})$, $(f_2(x) = x, A_2 = \{x \in \mathbb{R} : x > 0\})$.

- $(f_1(x) = 0, \quad A_1 = \{x \in \mathbb{R} : x < 0\})$, $(f_2(x) = x, \quad A_2 = \{x \in \mathbb{R} : x > 0\})$, $(f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\})$.



PAP Functions

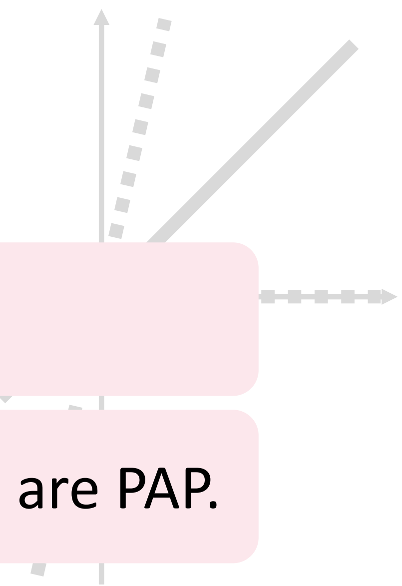
Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition)

roughly iff f can be "decomposed" into $f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$ such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is "analytic".

Example $f(x) = \mathrm{ReLU}(x)$.

- $(f_1(x) = 0,\ A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
  $(f_2(x) =$
- $(f_1(x) =$
  $(f_2(x) =$
  $(f_3(x) =$



Proposition PAP implies a.e.-differentiability.

Observation Virtually all functions used in practice are PAP.

Intensional Derivatives

analytic functions

<u>Example</u> $f(x) = \mathrm{ReLU}(x)$.

- $(\boxed{f_1(x) = 0}, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
  $(\boxed{f_2(x) = x}, A_2 = \{x \in \mathbb{R} : x > 0\}).$
- $(f_1(x) = 0, \quad A_1 = \{x \in \mathbb{R} : x < 0\}),$
  $(f_2(x) = x, \quad A_2 = \{x \in \mathbb{R} : x > 0\}),$
  $(f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\}).$

Intensional Derivatives

analytic functions

$(f_1'(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}).$

Example $f(x) = \mathrm{ReLU}(x).$

- $(\boxed{f_1(x) = 0}, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
  $(\boxed{f_2(x) = x}, A_2 = \{x \in \mathbb{R} : x > 0\}).$
- $(f_1(x) = 0, \quad A_1 = \{x \in \mathbb{R} : x < 0\}),$
  $(f_2(x) = x, \quad A_2 = \{x \in \mathbb{R} : x > 0\}),$
  $(f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\}).$

Intensional Derivatives

analytic functions

<u>Example</u> $f(x) = \mathrm{ReLU}(x)$.

- $\boxed{f_1(x) = 0}, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
  $\boxed{f_2(x) = x}, A_2 = \{x \in \mathbb{R} : x > 0\}).$

$(f_1'(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}).$

- $(f_1(x) = 0, \quad A_1 = \{x \in \mathbb{R} : x < 0\}),$
  $(f_2(x) = x, \quad A_2 = \{x \in \mathbb{R} : x > 0\}),$
  $(f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\}).$

$$\boxed{df(x)} = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0 \end{cases}$$

intensional derivative of f

Intensional Derivatives

analytic functions

$(f_1'(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}).$

Example $f(x) = \mathrm{ReLU}(x).$

- $(f_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\}),$
  $(f_2(x) = x, A_2 = \{x \in \mathbb{R} : x > 0\}).$

$$df(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0 \end{cases}$$

- $(f_1(x) = 0, \quad A_1 = \{x \in \mathbb{R} : x < 0\}),$
  $(f_2(x) = x, \quad A_2 = \{x \in \mathbb{R} : x > 0\}),$
  $(f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\}).$

$$\boxed{df(x)} = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > 0 \\ 7 & \text{for } x = 0 \end{cases}$$

$(f_1'(x) = 0, A_1 = \{x \in \mathbb{R} : x < 0\}),$
$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}),$
$(f_3'(x) = 7, A_3 = \{x \in \mathbb{R} : x = 0\}).$

Intensional Derivatives

<u>Proposition</u> Intensional derivatives satisfy the chain rule.

<u>Proposition</u> Any intensional derivative $\overset{\text{a.e.}}{=}$ standard derivative.

$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\})$.

<u>Example</u> $f(x) = \text{ReLU}(x)$.

- $\boxed{(f_1(x) = 0}, A_1 = \{x \in \mathbb{R} : x \le 0\})$,
  $\boxed{(f_2(x) = x}, A_2 = \{x \in \mathbb{R} : x > 0\})$.

$$df(x) = \begin{cases} 0 & \text{for } x \le 0 \\ 1 & \text{for } x > 0 \end{cases}$$

- $(f_1(x) = 0, \quad A_1 = \{x \in \mathbb{R} : x < 0\})$,
  $(f_2(x) = x, \quad A_2 = \{x \in \mathbb{R} : x > 0\})$,
  $(f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\})$.

$$df(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > 0 \\ 7 & \text{for } x = 0 \end{cases}$$

$(f_1'(x) = 0, A_1 = \{x \in \mathbb{R} : x < 0\})$,
$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\})$,
$(f_3'(x) = 7, A_3 = \{x \in \mathbb{R} : x = 0\})$.

Intensional Derivatives

<u>Proposition</u> Intensional derivatives satisfy the chain rule.

<u>Proposition</u> Any intensional derivative $\overset{\text{a.e.}}{=}$ standard derivative.

<u>Theorem</u> For PAP functions,

what autodiff computes is an intensional derivative,

and thus autodiff correctly computes gradients a.e.

<u>Example</u> $f(x) = \mathrm{ReLU}(x)$.

- $\boxed{f_1(x) = 0}, A_1 = \{x \in \mathbb{R} : x \le 0\}),$
  $\boxed{f_2(x) = x}, A_2 = \{x \in \mathbb{R} : x > 0\}).$

$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}).$

$$df(x) = \begin{cases} 0 & \text{for } x \le 0 \\ 1 & \text{for } x > 0 \end{cases}$$

- $(f_1(x) = 0$
  $(f_2(x) = x$
  $(f_3(x) = 7$

or $x < 0$
or $x > 0$
or $x = 0$

$(f_2'(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}),$
$(f_3'(x) = 7, A_3 = \{x \in \mathbb{R} : x = 0\}).$

High-Level Messages

- Measure-zero non-differentiabilities often bring us unexpected subtleties, when we try to establish formal correctness of ML algorithms (e.g., autodiff).

High-Level Messages

- Measure-zero non-differentiabilities often bring us unexpected subtleties, when we try to establish formal correctness of ML algorithms (e.g., autodiff).
- PAP functions and intensional derivatives would play an important role, when we try to deal with such subtleties (e.g., arising from other ML algorithms).