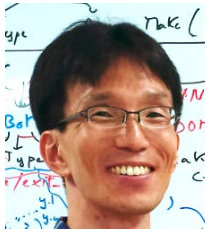
On Correctness of Automatic Differentiation for Non-Differentiable Functions



Wonyeol Lee[1] Hangyeol Yu[2] Xavier Rival[3] Hongseok Yang[4]

[1]Stanford, USA [2]Riiid AI Research, South Korea

[3]INRIA/ENS/CNRS, France [4]KAIST, South Korea

Presented at NeurIPS 2020 (Spotlight)

Autodiff: Theory

Problem For $h : \mathbb{R}^N \to \mathbb{R}$ given by $h(x) = (h_L \circ \cdots \circ h_1)(x)$, how to compute $\nabla h(x)$ correctly and efficiently?

Autodiff: Theory

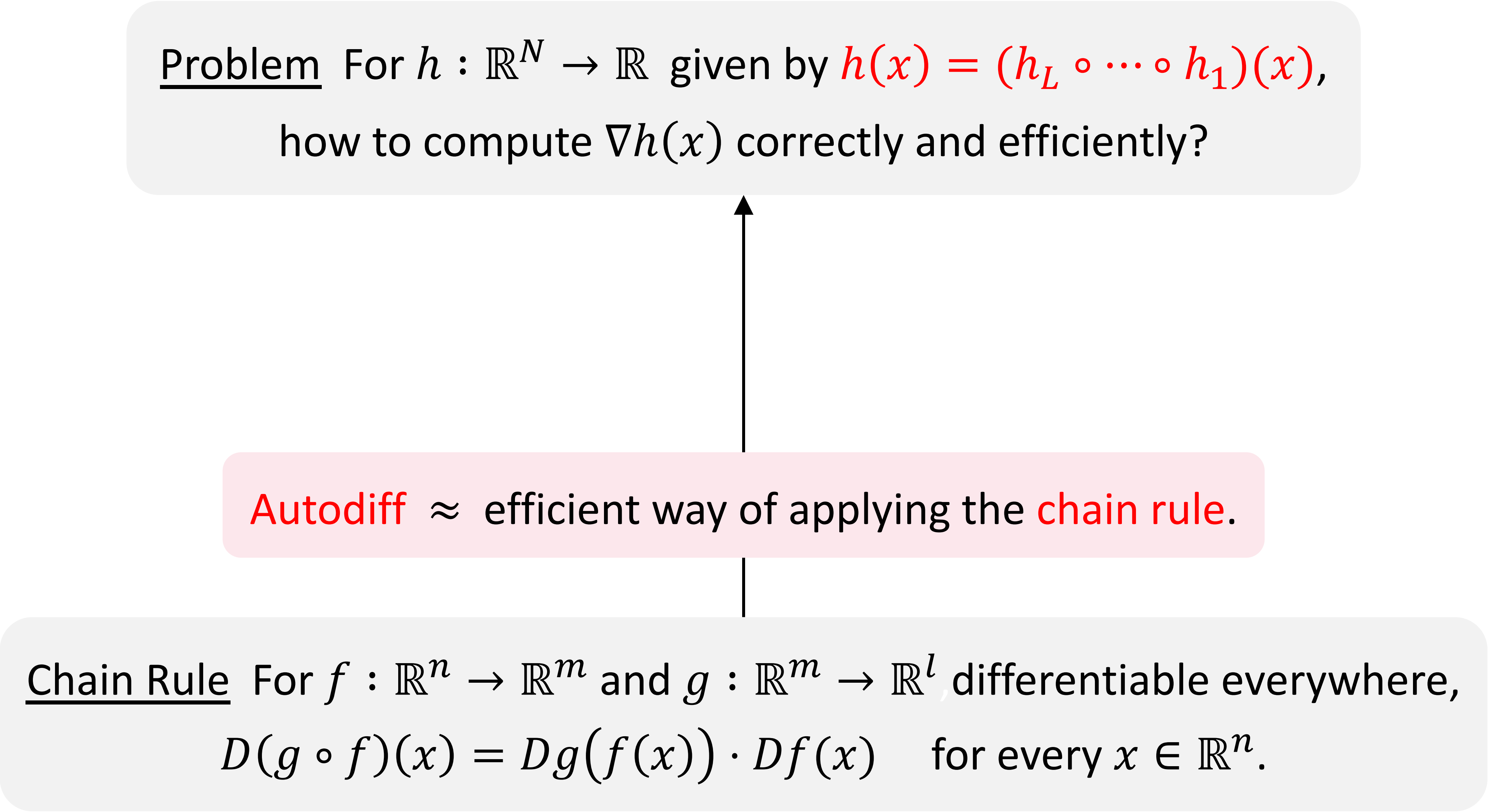
Problem For $h : \mathbb{R}^N \to \mathbb{R}$ given by $h(x) = (h_L \circ \cdots \circ h_1)(x)$,

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

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: Theory



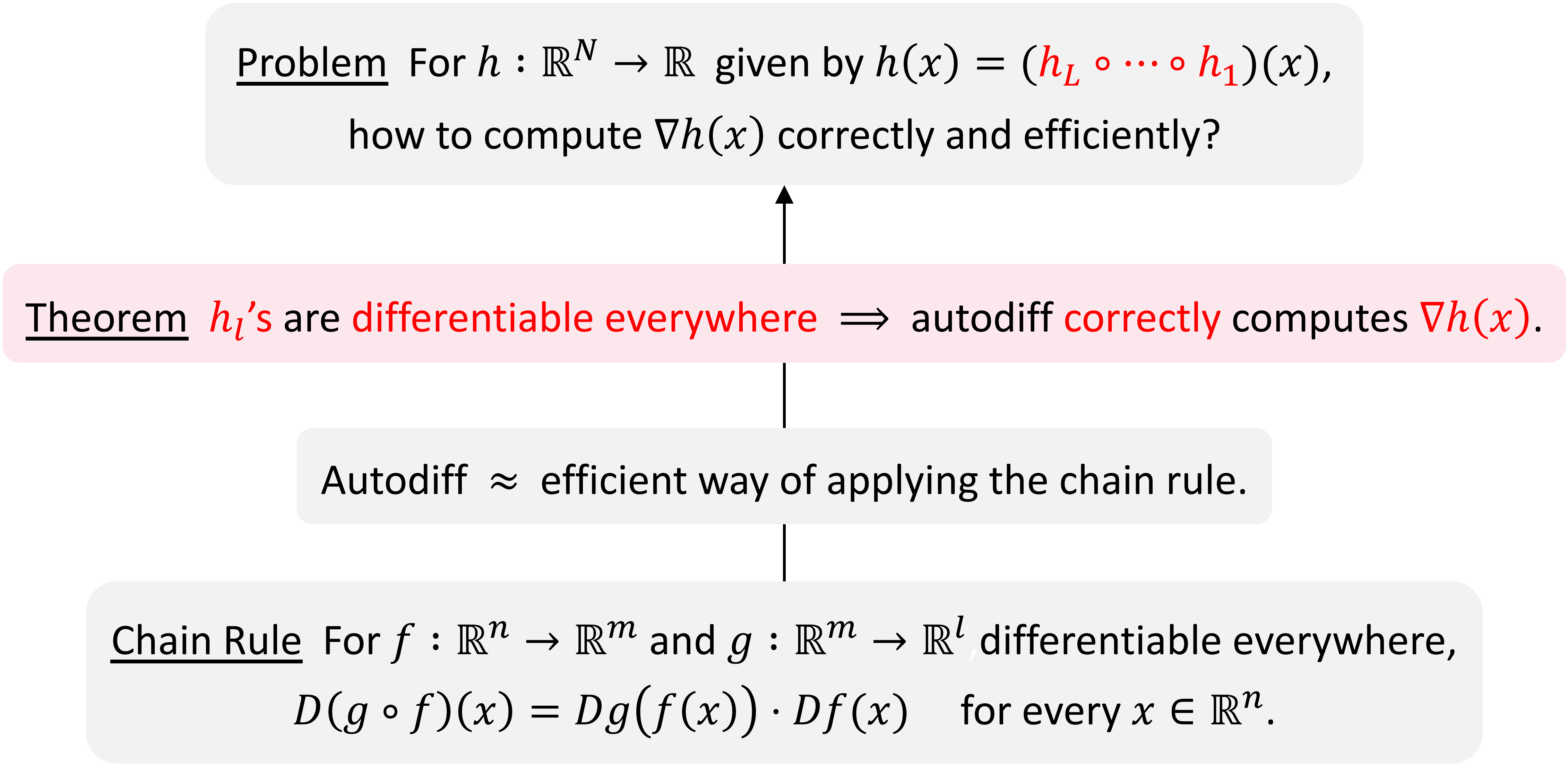
<u>Problem</u> For $h : \mathbb{R}^N \to \mathbb{R}$ given by $h(x) = (h_L \circ \cdots \circ h_1)(x)$, how to compute $\nabla h(x)$ correctly and efficiently?

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: Theory



<u>Problem</u> For $h : \mathbb{R}^N \to \mathbb{R}$ given by $h(x) = (h_L \circ \cdots \circ h_1)(x)$, how to compute $\nabla h(x)$ correctly and efficiently?

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

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: Practice

What about in practice?

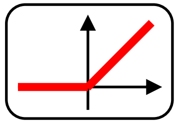
?

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

Autodiff: Practice

Discrepancy between theory and practice.

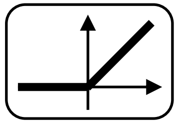
Theorem h_l's are ~~differentiable everywhere~~ $\Longrightarrow$ autodiff correctly computes $\nabla h(x)$.

e.g., $\mathrm{ReLU}(x) = \texttt{if}\ x \geq 0\ \texttt{then}\ x\ \texttt{else}\ 0 =$ 

Autodiff: Practice

Discrepancy between theory and practice.

Theorem h_l's are ~~differentiable everywhere~~ $\Longrightarrow$ autodiff correctly computes $\nabla h(x)$.

e.g., $\mathrm{ReLU}(x) = \texttt{if}\ x \geq 0\ \texttt{then}\ x\ \texttt{else}\ 0 =$ 

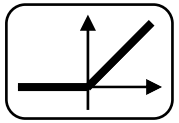
non-differentiable on a measure-zero set

measure = generalization of length, area, …

Autodiff: Practice

Belief: Measure-zero non-differentiability would not matter.

Theorem h_l's are ~~differentiable everywhere~~ $\Longrightarrow$ autodiff correctly computes $\nabla h(x)$.

e.g., $\mathrm{ReLU}(x) = \texttt{if}\ x \geq 0\ \texttt{then}\ x\ \texttt{else}\ 0 =$ 

non-differentiable on a measure-zero set

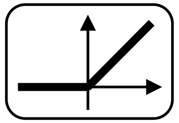
measure = generalization of length, area, …

Our Questions: Part 1

?

Belief: Measure-zero non-differentiability would not matter.

Theorem h_l's are ~~differentiable everywhere~~ $\Longrightarrow$ autodiff correctly computes $\nabla h(x)$.

e.g., $\mathrm{ReLU}(x) = \texttt{if}\ x \geq 0\ \texttt{then}\ x\ \texttt{else}\ 0 =$ 

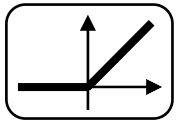
non-differentiable on a measure-zero set

Our Questions: Part 1

?

Belief: Measure-zero non-differentiability would not matter.

<ins>Theorem</ins> h_l's are differentiable $\underset{\text{almost-}}{\wedge}$ everywhere $\overset{?}{\boxed{\Longrightarrow}}$ autodiff correctly computes $\nabla h(x)\underset{\text{almost-everywhere}}{\wedge}$.

e.g., $\mathrm{ReLU}(x) = \texttt{if}\ x \geq 0\ \texttt{then}\ x\ \texttt{else}\ 0 =$ 

non-differentiable on a measure-zero set

almost-everywhere = except for a measure-zero set.

Our Questions: Part 1

?

Belief: Measure-zero non-differentiability would not matter.

Theorem h_l's are differentiable $\underbrace{\ }_{\text{almost-}}$everywhere $\boxed{\overset{?}{\Longrightarrow}}$ autodiff correctly computes $\nabla h(x)\underbrace{\ }_{\text{almost-everywhere}}$.

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

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

Our Results: Part 1

Measure-zero non-differentiabilities do matter!

<u>Theorem</u> h_l’s are differentiable $\underbrace{}_{\text{almost-}}$everywhere $\xcancel{\Longrightarrow}$ autodiff correctly computes $\nabla h(x)\underbrace{}_{\text{almost-everywhere}}$.

↑

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

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

Our Results: Part 1

Measure-zero non-differentiabilities do matter!

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

↑

Our Result This and related claims are false!

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

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

Subtlety 1

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

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

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

Subtlety 1

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

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

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

well-defined?

Subtlety 1

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

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

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

well-defined?

Subtlety 1: Undefined $(g \circ f)'$

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

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

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

well-defined?

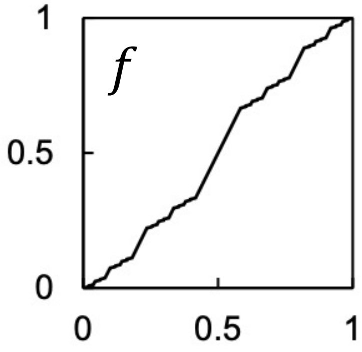
Subtlety 1: Undefined $(g \circ f)'$

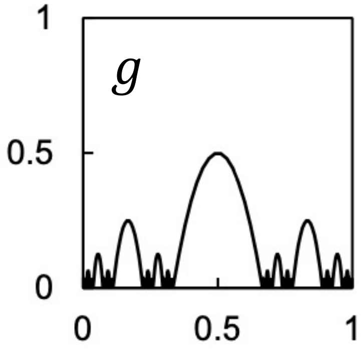
Claim 1 For any $f, g : \mathbb{R} \to \mathbb{R}$,

$$f, g : \text{a.e.-differentiable and continuous} \;\not\Longrightarrow\; \underset{\times}{\boxed{(g \circ f)'(x)}} = \boxed{g'(f(x))} \cdot \underset{\bigcirc}{\boxed{f'(x)}} \quad \text{for a.e. } x \in \mathbb{R}.$$

well-defined?

Counterexample Involves the Cantor function.





Subtlety 1: Undefined $(g \circ f)'$

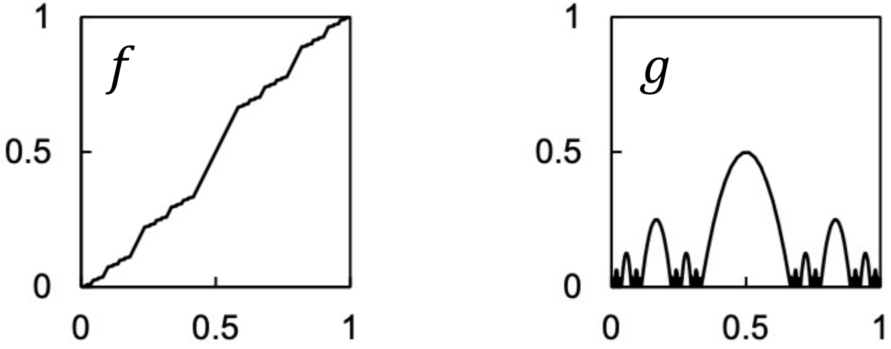
Claim 1 For any $f, g : \mathbb{R} \to \mathbb{R}$,

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

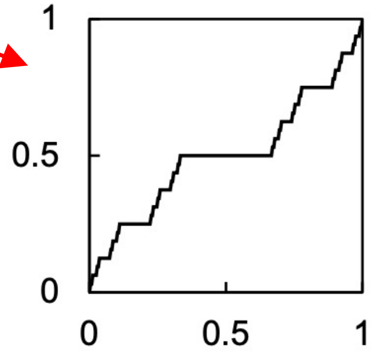
$$\not\Longrightarrow \quad \underset{\times}{\boxed{(g \circ f)'(x)}} = \underset{\blacktriangle}{\boxed{g'\big(f(x)\big)}} \cdot \underset{\bigcirc}{\boxed{f'(x)}} \quad \text{for a.e. } x \in \mathbb{R}.$$

well-defined?

Counterexample Involves the Cantor function.



has pathological properties



Subtlety 1: Undefined $(g \circ f)'$

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

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

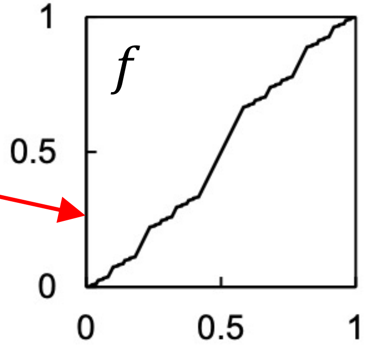
$$\not\Longrightarrow \quad \underset{\times}{\boxed{(g \circ f)'(x)}} = \underset{\blacktriangle}{\boxed{g'\big(f(x)\big)}} \cdot \underset{\bigcirc}{\boxed{f'(x)}} \quad \text{for a.e. } x \in \mathbb{R}.$$

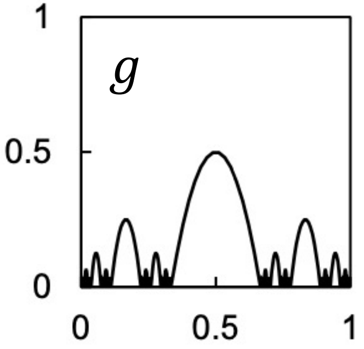
well-defined?

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

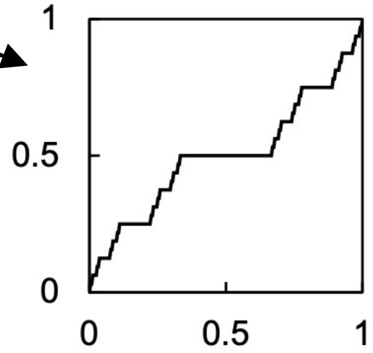
f is a bijection:

- continuous, a.e.-diff’l.
- positive-measure set ⇄ measure-zero set.





has pathological properties



Subtlety 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

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

Subtlety 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

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

well-defined?

Subtlety 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

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

well-defined?

Subtlety 2

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

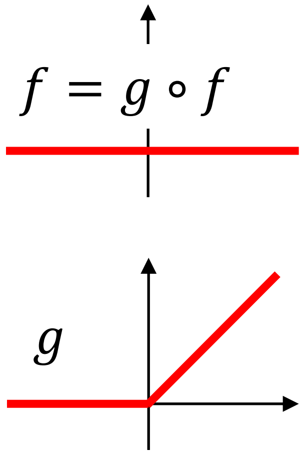
$$\overbrace{f, g}^{\text{and } g\circ f} : \text{a.e.-differentiable and continuous} \cdots (*)$$

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

well-defined?

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

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



Subtlety 2: Undefined g'

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

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

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

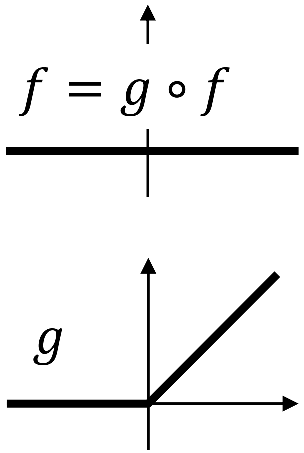
well-defined?

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

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

$= g'(0)$

$=$ undefined for all x



Subtlety 2: Undefined g'

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

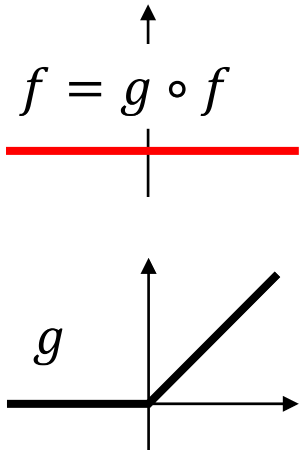
$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

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

well-defined?

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

$$\Longrightarrow \quad \underset{=\,0}{\underset{\uparrow}{(g \circ f)'(x)}} \qquad \underset{\substack{=\, g'(0) \\ =\, \text{undefined for all } x}}{\underset{\uparrow}{g'\big(f(x)\big)}} \qquad \underset{=\,0}{\underset{\uparrow}{f'(x)}}$$



Subtlety 2: Undefined g'

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

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

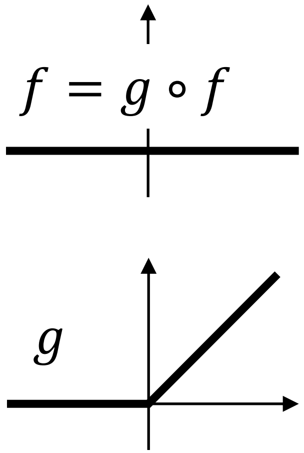
$$\cancel{\implies} \quad \underset{\triangle}{\boxed{(g \circ f)'(x)}} = \underset{\times}{\boxed{g'\big(f(x)\big)}} \cdot \underset{\triangle}{\boxed{f'(x)}} \qquad \text{for a.e. } x \in \mathbb{R}.$$

well-defined?

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

$$\implies \quad \underset{= 0}{\underset{\uparrow}{(g \circ f)'(x)}} \qquad \underset{\uparrow}{dg\big(f(x)\big)} \qquad \underset{= 0}{\underset{\uparrow}{f'(x)}}$$

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



Subtlety 2: Undefined g'

<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

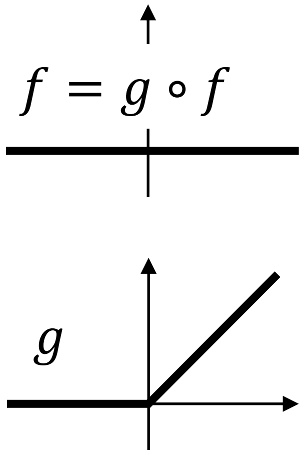
$$\xcancel{\Longrightarrow} \quad \underset{\triangle}{\boxed{(g \circ f)'(x)}} = \underset{\times}{\boxed{g'\big(f(x)\big)}} \cdot \underset{\triangle}{\boxed{f'(x)}} \quad \text{for a.e. } x \in \mathbb{R}.$$

well-defined?

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

$$\Longrightarrow \quad \boxed{\underset{=\,0}{(g \circ f)'(x)} = \underset{\uparrow}{dg\big(f(x)\big)} \times \underset{=\,0}{f'(x)}} \quad \text{for all } x \in \mathbb{R}.$$

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



Subtlety 3

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

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

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

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

Subtlety 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

well-defined!

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

Subtlety 3

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

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

well-defined!

$$\underset{?}{\boxed{\implies}} \quad \underbrace{\boxed{(g \circ f)'(x)} \underset{?}{=} \boxed{dg\big(f(x)\big)} \cdot \boxed{df(x)}}_{\left(\exists\, df, dg : \mathbb{R} \to \mathbb{R} \text{ such that } df \overset{\text{a.e.}}{=} f',\ dg \overset{\text{a.e.}}{=} g', \text{ and}\right)} \quad \text{for a.e. } x \in \mathbb{R}.$$

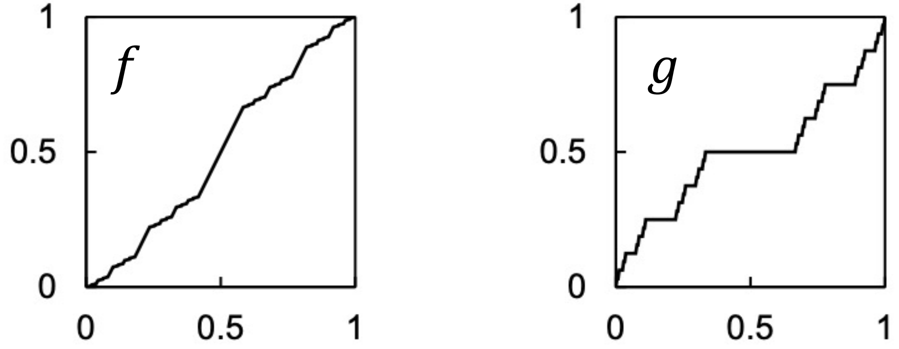
Subtlety 3: Wrong Equation for $(g \circ f)'$

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

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

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

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



Subtlety 3: Wrong Equation for $(g \circ f)'$

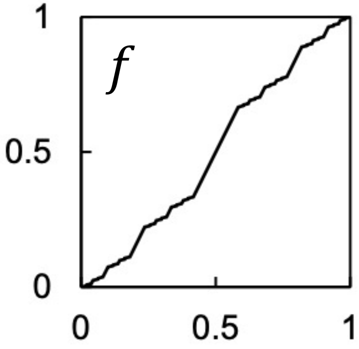
<u>Claim 3</u> For any $f, g : \mathbb{R} \to \mathbb{R}$,

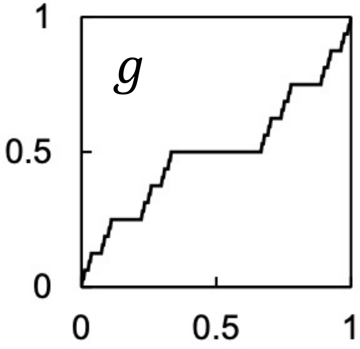
$\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous

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

<u>Show</u> $\boxed{(g \circ f)'(x) \neq 0}$ and $\boxed{f'(x) = 0}$ for positive-measure x.

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





Our Results: Part 1

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

✓ Our Result This and related claims are false!

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

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

Our Results: Part 1

<u>Our Result</u> Autodiff has been used without correctness guarantee!

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

✓ <u>Our Result</u> This and related claims are false!

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

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

Our Questions: Part 2

Can we recover the correctness theorem?

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

✓ Our Result This and related claims are false!

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

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

Our Questions: Part 2

Can we recover the correctness theorem?

What do the outputs of autodiff even mean?

(e.g., $\mathrm{ReLU}'(0) = 0$ in TensorFlow, PyTorch, …)

Our Result This and related claims are false!

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

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

Our Questions: Part 2

Can we recover the correctness theorem?

What do the outputs of autodiff even mean?

(e.g., $\mathrm{ReLU}'(0) = 0$ in TensorFlow, PyTorch, …)

They are not Clarke-subdifferentials [KL18]:

- $\partial^c f(x) := \mathrm{conv}\left\{ \lim_{n\to 0} Df(x_n) \,\middle|\, x_n \to x \text{ and } \exists Df(x_n) \right\}.$

Our Questions: Part 2

Can we recover the correctness theorem?

What do the outputs of autodiff even mean?

(e.g., $\mathrm{ReLU}'(0) = 0$ in TensorFlow, PyTorch, …)

They are not Clarke-subdifferentials [KL18]:

- $\partial^c f(x) \coloneqq \mathrm{conv}\left\{ \lim_{n\to 0} Df(x_n) \,\middle|\, x_n \to x \text{ and } \exists Df(x_n) \right\}$.
- $f(x) = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$: $\partial^c f(0) = \{1\} \not\ni 0 = f'(0)$ (by autodiff).

Our Results: Part 2

Theorem h_l's are differentiable everywhere $\cancel{\implies}$ autodiff correctly computes $\nabla h(x)$.

? almost-

almost-everywhere

Our Results: Part 2

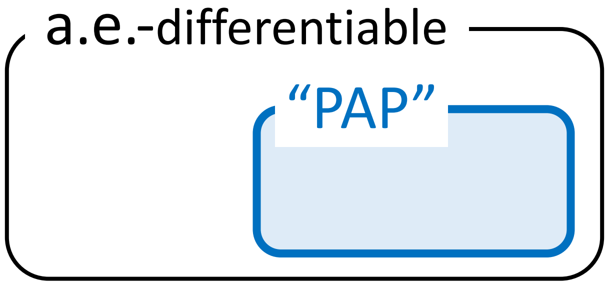
Theorem h_l's are ~~differentiable everywhere~~ ⟹ autodiff correctly computes $\nabla h(x)$.

~~almost-~~

almost-everywhere

so-called "PAP"

new property we propose



Our Results: Part 2

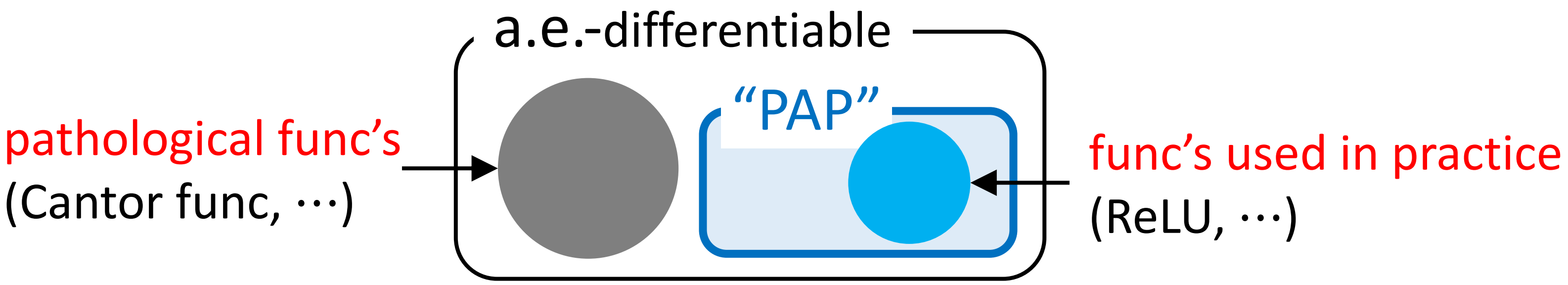
<u>Theorem</u> h_l's are ~~differentiable everywhere~~ $\Longrightarrow$ autodiff correctly computes $\nabla h(x)$.

~~almost-~~

so-called "PAP"

new property we propose

almost-everywhere



a.e.-differentiable

"PAP"

pathological func's
(Cantor func, ⋯)

func's used in practice
(ReLU, ⋯)

Our Results: Part 2

Our Result Prove the claim for PAP functions h_l's.

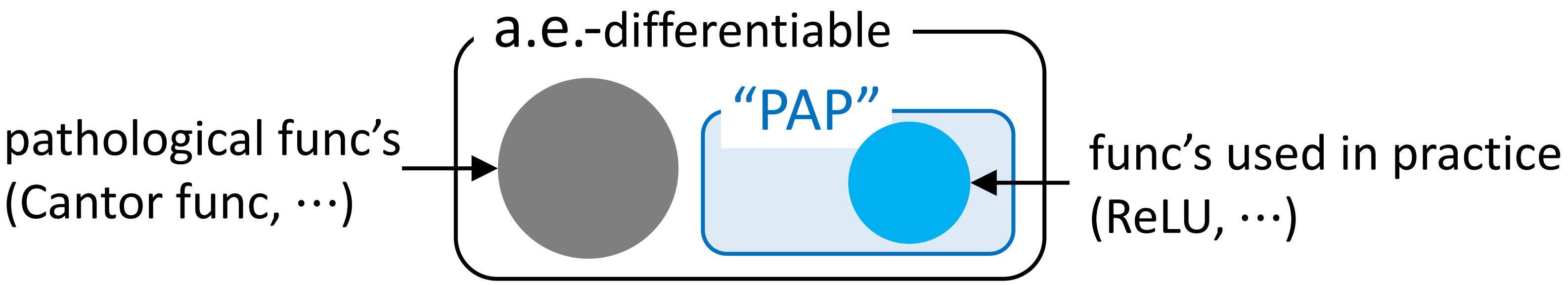
Theorem h_l's are ~~differentiable everywhere~~ ⇒ autodiff correctly computes $\nabla h(x)$.

~~almost-~~

so-called "PAP"

new property we propose

almost-everywhere



a.e.-differentiable

"PAP"

pathological func's (Cantor func, ⋯)

func's used in practice (ReLU, ⋯)

Our Results: Part 2

Our Result Autodiff computes so-called “intensional derivatives” of h.

Our Result Prove the claim for PAP functions h_l’s.

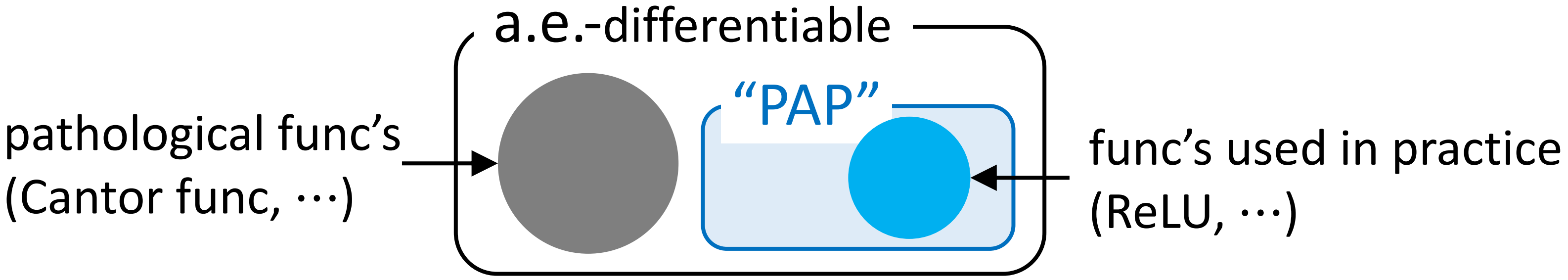
Theorem h_l’s are ~~differentiable everywhere~~ ⇒ autodiff correctly computes $\nabla h(x)$.

~~almost-~~

so-called “PAP”

new property we propose

almost-everywhere



PAP Functions

piecewise analytic under analytic partition

$\downarrow$

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

such that

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

analytic = has derivatives of all orders that are bounded nicely.

PAP Functions

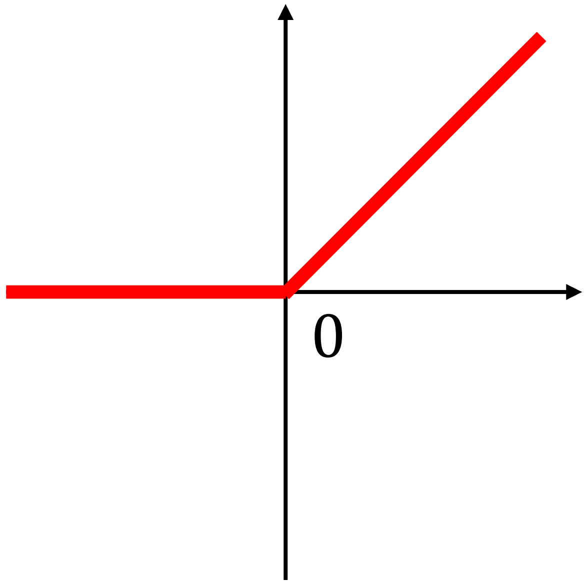
Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

such that

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

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



analytic = has derivatives of all orders that are bounded nicely.

PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

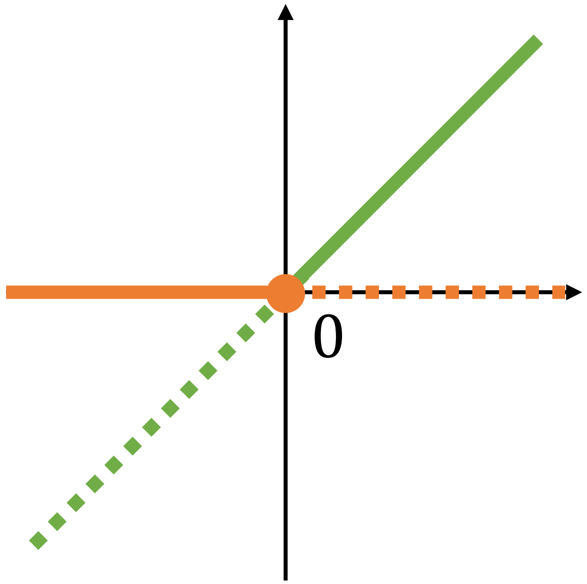
$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

such that

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

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 = has derivatives of all orders that are bounded nicely.

PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

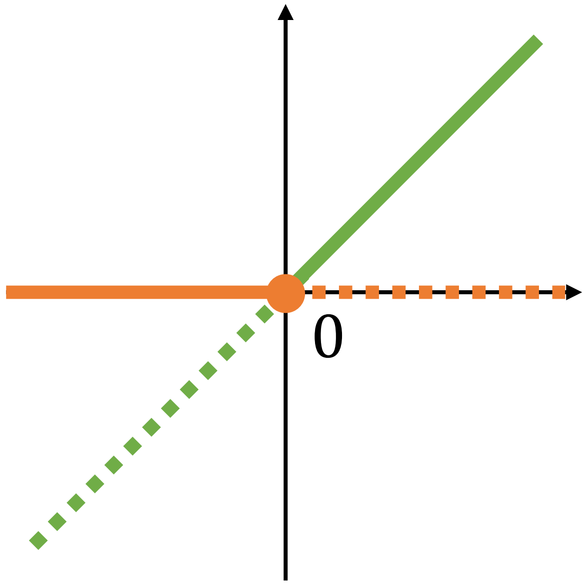
such that

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

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

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

analytic functions



analytic = has derivatives of all orders that are bounded nicely.

PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be “decomposed” into

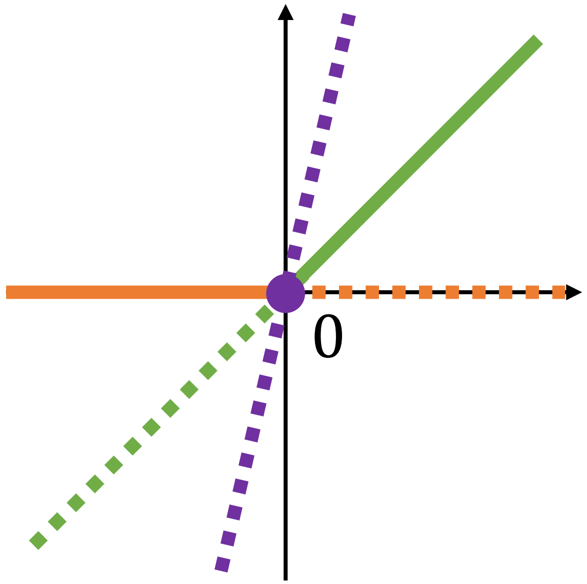
$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

such that

$f_i : \mathbb{R}^n \to \mathbb{R}^m$ and $A_i \subseteq \mathbb{R}^n$ are “analytic”.

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

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



PAP Functions

can be a subset of $\mathbb{R}^n$

<u>Definition</u> $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

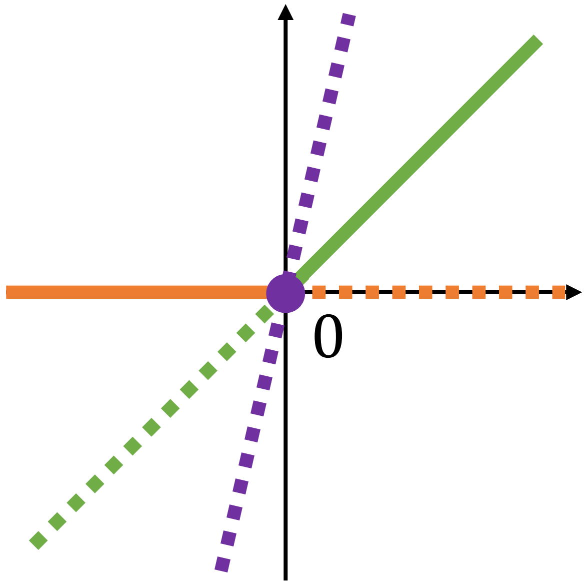
such that

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

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

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

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



PAP Functions

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

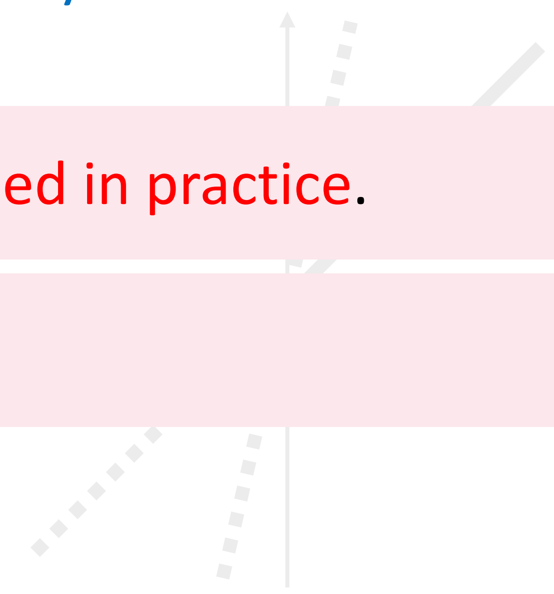
such that

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

Examp

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

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



Observation PAP functions include all functions used in practice.

Proposition PAP functions are a.e.-differentiable.

PAP Functions

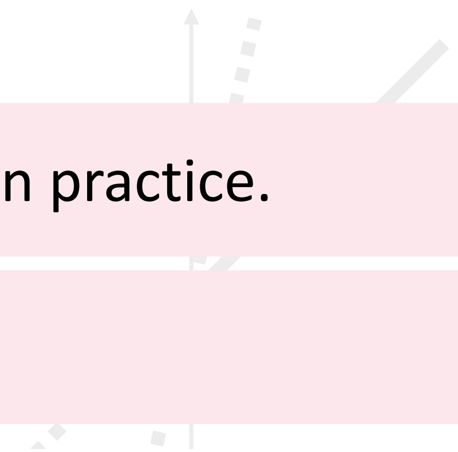
Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

such that

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

Examp



Observation PAP functions include all functions used in practice.

Proposition PAP functions are a.e.-differentiable.

For any non-constant, analytic function $g : \mathbb{R}^n \to \mathbb{R}$, $\{x \in \mathbb{R}^n |\, g(x) = 0\}$ has measure zero.

PAP Functions

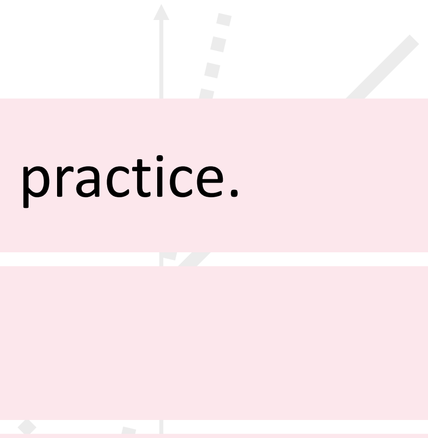
Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is called PAP if f can be "decomposed" into

$$f_1\big|_{A_1}, f_2\big|_{A_2}, \cdots$$

such that

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

Examp



Observation PAP functions include all functions used in practice.

Proposition PAP functions are a.e.-differentiable.

Definition PAP functions have "intensional derivatives".

Intensional Derivatives

analytic functions

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

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

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

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

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

Intensional Derivatives

analytic functions

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

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

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

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

intensional derivative of f

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

Intensional Derivatives

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

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

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

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

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

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

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

Intensional Derivatives

Proposition Intensional derivative is a total function.

Proposition Intensional derivatives always satisfy the chain rule.

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

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

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

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

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

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

Intensional Derivatives

Proposition Intensional derivative is a total function.

Proposition Intensional derivatives always satisfy the chain rule.

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

Proposition Intensional derivative $\overset{\text{a.e.}}{=}$ standard derivative.

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

$\{x \in \mathbb{R}^n \mid df(x) \neq Df(x)\}$ is contained in a countable union of the zero-sets of (non-const) analytic func's.

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

Correctness of Autodiff

Proposition Intensional derivative is a total function.

Proposition Intensional derivatives always satisfy the chain rule.

Proposition Intensional derivative $\overset{\text{a.e.}}{=}$ standard derivative.

Theorem For any $h = h_L \circ \cdots \circ h_1$ with PAP h_l,

autodiff computes an intensional derivative of h,

and thus computes the correct gradient of h a.e.

Correctness of Autodiff

if autodiff uses an intensional derivative of h_l for "D"h_l,

Theorem For any $h = h_L \circ \cdots \circ h_1$ with PAP h_l,

autodiff computes an intensional derivative of h,

and thus computes the correct gradient of h a.e.

Correctness of Autodiff

In TensorFlow and PyTorch,

- "D"$\texttt{relu}(x) = 0$ for $x \leq 0$; 1 for $x > 0$. ✓

if autodiff uses an intensional derivative of h_l for "D"h_l,

<u>Theorem</u> For any $h = h_L \circ \cdots \circ h_1$ with PAP h_l,

autodiff computes an intensional derivative of h,

and thus computes the correct gradient of h a.e.

Correctness of Autodiff

In TensorFlow and PyTorch,

- "D"$\texttt{relu}(x) = 0$ for $x \le 0$; 1 for $x > 0$. ✓
- "D"$\texttt{sqrt}(x) = \infty$ for $x = 0$; $1/2\sqrt{x}$ for $x > 0$. ✗

if autodiff uses an intensional derivative of h_l for "D"h_l,

Theorem For any $h = h_L \circ \cdots \circ h_1$ with PAP h_l,

autodiff computes an intensional derivative of h,

and thus computes the correct gradient of h a.e.

Correctness of Autodiff

In TensorFlow and PyTorch,

- $\text{“}D\text{”}\texttt{relu}(x) = 0$ for $x \leq 0;\ 1$ for $x > 0$. ✓
- $\text{“}D\text{”}\texttt{sqrt}(x) = \infty$ for $x = 0;\ 1/2\sqrt{x}$ for $x > 0$. ✗

  For $f(x) = \texttt{sqrt}(\texttt{mult}(x, 0))$, they compute $f'(x) =$ NaN for all x.

if autodiff uses an intensional derivative of h_l for $\text{“}D\text{”}h_l$,

<u>Theorem</u> For any $h = h_L \circ \cdots \circ h_1$ with PAP h_l,

autodiff computes an intensional derivative of h,

and thus computes the correct gradient of h a.e.

Correctness of Autodiff

In TensorFlow and PyTorch,

- $\text{“}D\text{”}\texttt{relu}(x) = 0 \text{ for } x \leq 0;\ 1 \text{ for } x > 0.$ ✓
- $\text{“}D\text{”}\texttt{sqrt}(x) = \overset{7}{\infty} \text{ for } x = 0;\ 1/2\sqrt{x} \text{ for } x > 0.$ ✓

  For $f(x) = \texttt{sqrt}(\texttt{mult}(x, 0))$, they compute $f'(x) = \overset{0}{\text{NaN}}$ for all x.

if autodiff uses an intensional derivative of h_l for “D”h_l,

Theorem For any $h = h_L \circ \cdots \circ h_1$ with PAP h_l,
autodiff computes an intensional derivative of h,
and thus computes the correct gradient of h a.e.

Intensional Derivatives: Remarks

First-order → higher-order.

- (First-order) intensional derivative = PAP function.
- Extended to higher-order derivatives. Enjoy the same properties.

Intensional Derivatives: Remarks

First-order → higher-order.

- (First-order) intensional derivative = PAP function.
- Extended to higher-order derivatives. Enjoy the same properties.

Difference from Clarke-subdifferentials.

- Intentional derivative: $\partial^i f \in \mathcal{P}([\mathbb{R}^n \to \mathbb{R}^{m\times n}])$.
- Clarke-subdifferential: $\partial^c f \in [\mathbb{R}^n \to \mathcal{P}(\mathbb{R}^{m\times n})]$.

  → Difficult to extend to higher-order derivatives.

High-Level Messages

We often have discrepancy between theory and practice of ML algorithms.

But our theoretical understanding on such discrepancy is still limited.

ML Algorithm	Theory	Practice
Autodiff	differentiable func’s	a.e.-differentiable func’s

High-Level Messages

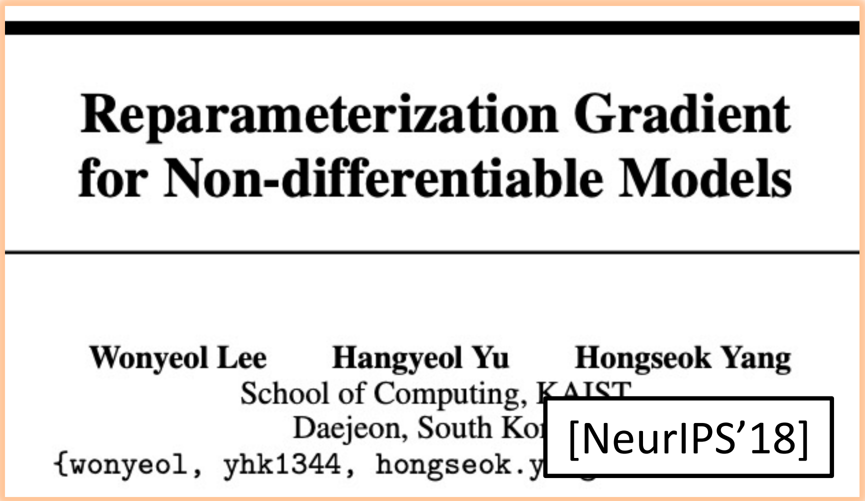
We often have discrepancy between theory and practice of ML algorithms.

But our theoretical understanding on such discrepancy is still limited.

ML Algorithm	Theory	Practice
Autodiff and many more	differentiable func’s	a.e.-differentiable func’s

Algorithm for estimating

$$\nabla_\theta \int f_\theta(z)dz$$



Reparameterization Gradient for Non-differentiable Models

Wonyeol Lee **Hangyeol Yu** **Hongseok Yang**

School of Computing, KAIST

Daejeon, South Ko

{wonyeol, yhk1344, hongseok.y

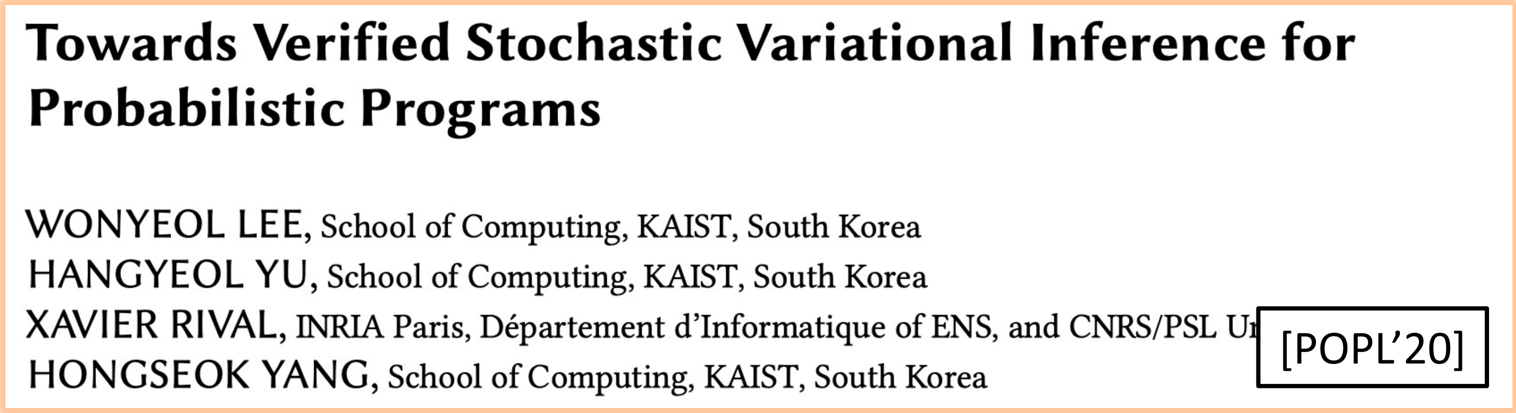
[NeurIPS’18]

High-Level Messages

We often have discrepancy between theory and practice of ML algorithms.

But our theoretical understanding on such discrepancy is still limited.

ML Algorithm	Theory	Practice
Autodiff and many more	differentiable func's	a.e.-differentiable func's
Variational inference, …	func's with finite integrals (and other nice properties)	func's with infinite integrals (or some bad properties)



Towards Verified Stochastic Variational Inference for Probabilistic Programs

WONYEOL LEE, School of Computing, KAIST, South Korea
HANGYEOL YU, School of Computing, KAIST, South Korea
XAVIER RIVAL, INRIA Paris, Département d'Informatique of ENS, and CNRS/PSL Ur
HONGSEOK YANG, School of Computing, KAIST, South Korea

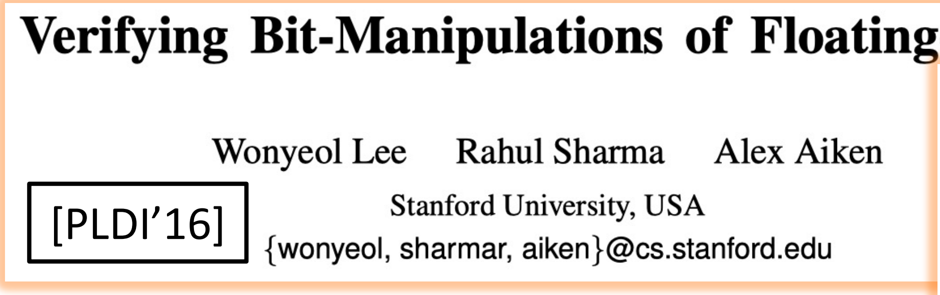
[POPL'20]

High-Level Messages

We often have discrepancy between theory and practice of ML algorithms.

But our theoretical understanding on such discrepancy is still limited.

ML Algorithm	Theory	Practice
Autodiff and many more	differentiable func's	a.e.-differentiable func's
Variational inference, …	func's with finite integrals (and other nice properties)	func's with infinite integrals (or some bad properties)
Most algorithms	func's on reals	func's on floating-points



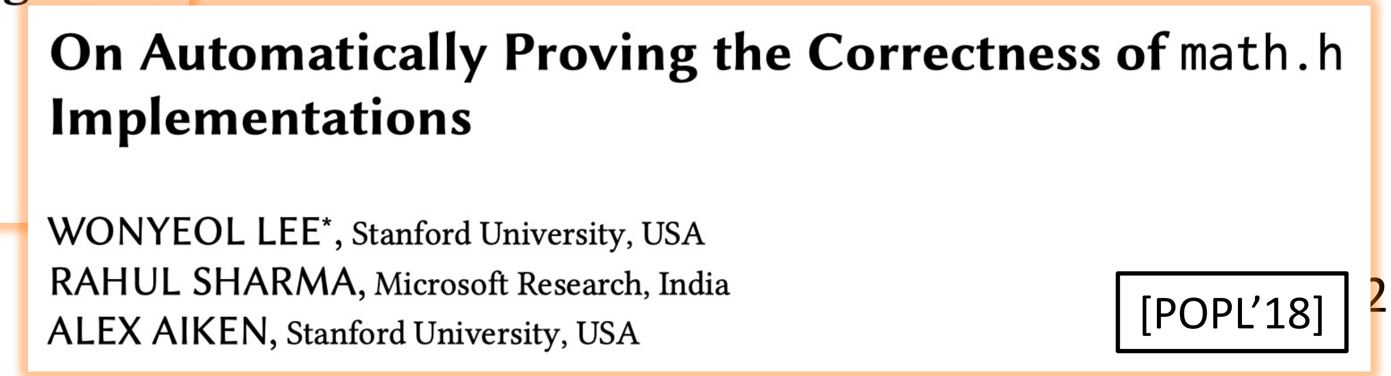
Verifying Bit-Manipulations of Floating-Point

Wonyeol Lee Rahul Sharma Alex Aiken

Stanford University, USA

{wonyeol, sharmar, aiken}@cs.stanford.edu

[PLDI'16]



On Automatically Proving the Correctness of `math.h` Implementations

WONYEOL LEE*, Stanford University, USA

RAHUL SHARMA, Microsoft Research, India

ALEX AIKEN, Stanford University, USA

[POPL'18]

Comments? Questions?