On Correctness of Automatic Differentiation for Non-Differentiable Functions

Wonyeol Lee
1
Stanford, USA

Hangyeol Yu
2
Riiid AI Research, South Korea

Xavier Rival
3
INRIA/ENS/CNRS, France

Hongseok Yang
4
KAIST, South Korea

Presented at NeurIPS 2020 (Spotlight)
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,
\[
D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)
\]
for every \( x \in \mathbb{R}^n \).
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** $\approx$ 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,

$$D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$
Theorem $h_i$’s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla h(x)$.

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

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

Problem For $h : \mathbb{R}^N \rightarrow \mathbb{R}$ given by $h(x) = (h_L \circ \cdots \circ h_1)(x)$, how to compute $\nabla h(x)$ correctly and efficiently?
Theorem \( h_i \)’s are differentiable everywhere \( \Rightarrow \) autodiff correctly computes \( \nabla h(x) \).
Theorem  $h_i$’s are differentiable everywhere $\implies$ autodiff correctly computes $\nabla h(x)$.

e.g., $\text{ReLU}(x) = \text{if } x \geq 0 \text{ then } x \text{ else } 0 = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$
Autodiff: Practice

Discrepancy between theory and practice.

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

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

non-differentiable on a measure-zero set

measure = generalization of length, area, ...
**Theorem** $h_i$’s are differentiable everywhere \(\Rightarrow\) autodiff correctly computes $\nabla h(x)$.

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

non-differentiable on a measure-zero set

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

measure = generalization of length, area, ...
Theorem $h_l$’s are differentiable everywhere $\implies$ autodiff correctly computes $\nabla h(x)$.

e.g., $\text{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$ non-differentiable on a measure-zero set

Belief: Measure-zero non-differentiability would not matter.
Our Questions: Part 1

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

Theorem \( h_i \)'s are differentiable everywhere \( \implies \) autodiff correctly computes \( \nabla h(x) \).

e.g., ReLU\( (x) = \text{if } x \geq 0 \text{ then } x \text{ else } 0 \) = non-differentiable on a measure-zero set

almost-everywhere = except for a measure-zero set.
Belief: Measure-zero non-differentiability would not matter.
Our Results: Part 1

Measure-zero non-differentiabilities do matter!

Theorem \( h_i \)'s are differentiable everywhere \( \implies \) autodiff correctly computes \( \nabla h(x) \).

Chain Rule For \( f : \mathbb{R}^n \to \mathbb{R}^m \) and \( g : \mathbb{R}^m \to \mathbb{R}^l \) differentiable everywhere, \( \nabla (g \circ f)(x) = Dg(f(x)) \cdot Df(x) \) for every \( x \in \mathbb{R}^n \).
Our Results: Part 1

Measure-zero non-differentiabilities do matter!

**Theorem** $h_i$'s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla h(x)$.

**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 everywhere, $D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)$ for every $x \in \mathbb{R}^n$. Almost-everywhere
Subtlety 1

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

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

$$\Rightarrow \quad (g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$

for a.e. $x \in \mathbb{R}$. 
Subtlety 1

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

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

\[
(g \circ f)'(x) = g'(f(x)) \cdot f'(x)
\]

for a.e. $x \in \mathbb{R}$.  

well-defined?
Subtlety 1

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

\[ (g \circ f)'(x) = g'(f(x)) \cdot f'(x) \]

for a.e. $x \in \mathbb{R}$. 

Well-defined?
Subtlety 1: Undefined \((g \circ f)\)'

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

\(f, g : \text{a.e.-differentiable and continuous}\)

\((g \circ f)'(x) = g'(f(x)) \cdot f'(x)\) for a.e. \(x \in \mathbb{R}\).
Claim 1 For any \( f, g : \mathbb{R} \to \mathbb{R} \),

\[
(f \circ g)'(x) = g'(f(x)) \cdot f'(x)
\]

for a.e. \( x \in \mathbb{R} \).

Counterexample Involves the Cantor function.
Subtlety 1: Undefined \((g \circ f)'\)

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

\[ (g \circ f)'(x) = g'(f(x)) \cdot f'(x) \]

for a.e. \(x \in \mathbb{R}\).

Counterexample  Involves the **Cantor function**. has pathological properties
Subtlety 1: Undefined \((g \circ f)'\) 

**Claim 1** For any \(f, g : \mathbb{R} \to \mathbb{R}\), \(f, g : \text{a.e.-differentiable and continuous}\)

\[
(g \circ f)'(x) = g'(f(x)) \cdot f'(x)
\]

for a.e. \(x \in \mathbb{R}\).

**Counterexample** Involves the Cantor function.

\(f\) is a bijection:
- continuous, a.e.-diff’l.
- positive-measure set \(\Leftrightarrow\) measure-zero set.

\[
\begin{align*}
\text{has pathological properties} & \\
\text{Involves the Cantor function.} & \\
\end{align*}
\]
Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$, $f, g$ a.e.-differentiable and continuous

$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$ for a.e. $x \in \mathbb{R}$. 
Subtlety 2

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

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

$\Rightarrow$

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

well-defined?
Claim 2  For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$:

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

$$f' \quad ? \quad \Rightarrow \quad (g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$

well-defined? for a.e. $x \in \mathbb{R}$. 
Subtlety 2

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

$f, g: \text{a.e.-differentiable and continuous} \ldots (*)$

$\Rightarrow (g \circ f)'(x) = g'(f(x)) \cdot f'(x)$

for a.e. $x \in \mathbb{R}$.

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

$\Rightarrow$ easy to check that (*) holds.
Subtlety 2: Undefined $g'$

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

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

$$ (g \circ f)'(x) = g'(f(x)) \cdot f'(x) $$

for a.e. $x \in \mathbb{R}$.

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

$$ g'(f(x)) $$

$$ = g'(0) $$

$$ = \text{undefined for all } x $$
Subtlety 2: Undefined $g'$

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

$f, g$ a.e.-differentiable and continuous, then

$$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$

for a.e. $x \in \mathbb{R}$.

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

$$\implies (g \circ f)'(x) = 0 = g'(0) = 0 = \text{undefined for all } x$$
Subtlety 2: Undefined $g'$

Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$:
- $f, g : \text{a.e.-differentiable and continuous}$

$$
(g \circ f)'(x) = g'(f(x)) \cdot f'(x)
$$

for a.e. $x \in \mathbb{R}$.

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

$$
(g \circ f)'(x) = 0 \quad dg(f(x)) = 0 \\
\Rightarrow \quad dg(y) = \begin{cases} 
7 & \text{for } y = 0 \\
g'(y) & \text{for } y \neq 0
\end{cases}
$$
Subtlety 2: Undefined $g'$

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

- $f, g$: a.e.-differentiable and continuous
- $$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$ for a.e. $x \in \mathbb{R}$.

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

$$(g \circ f)'(x) = dg(f(x)) \times f'(x)$$ for all $x \in \mathbb{R}$.

$$d g(y) = \begin{cases} 7 & \text{for } y = 0 \\ g'(y) & \text{for } y \neq 0 \end{cases}$$
Claim 3  For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$,

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

exists $df, dg : \mathbb{R} \to \mathbb{R}$ such that $df \overset{\text{a.e.}}{=} f'$, $dg \overset{\text{a.e.}}{=} g'$, and $\frac{d}{dx} (g \circ f) = dg(f(x)) \cdot df(x)$ for a.e. $x \in \mathbb{R}$. 

Subtlety 3
Subtlety 3

Claim 3 For any \( f, g : \mathbb{R} \to \mathbb{R} \), and \( g \circ f \)
\( f, g \) : a.e.-differentiable and continuous
\( \Rightarrow \)
\[ (g \circ f)'(x) = dg(f(x)) \cdot df(x) \]
for a.e. \( x \in \mathbb{R} \).

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

well-defined!
Subtlety 3

Claim 3 For any \( f, g : \mathbb{R} \to \mathbb{R} \),
and \( g \circ f \)
\( f, g \): a.e.-differentiable and continuous
\[ (g \circ f)'(x) = dg(f(x)) \cdot df(x) \]
for a.e. \( x \in \mathbb{R} \).

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

well-defined!
Subtlety 3: Wrong Equation for $(g \circ f)'$

Claim 3 For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$, $f, g$ : a.e.-differentiable and continuous

\[ (g \circ f)'(x) \neq dg(f(x)) \cdot df(x) \]

for a.e. $x \in \mathbb{R}$.

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

Counterexample Involves the Cantor function again.
Subtlety 3: Wrong Equation for \((g \circ f)'\)

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

and \(g \circ f\): a.e.-differentiable and continuous

\((g \circ f)'(x) \neq 0\) and \(f'(x) = 0\) for positive-measure \(x\).

Counterexample: Involves the Cantor function again.
Our Results: Part 1

**Theorem** $h_l$’s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla h(x)$. Almost-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 everywhere, $\forall x \in \mathbb{R}^n$

\[
D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)
\]
Our Results: Part 1

Our Result  Autodiff has been used without correctness guarantee!

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

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 everywhere, \( D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \) for every \( x \in \mathbb{R}^n \).
Can we recover the correctness theorem?
Chain Rule

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

$$D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)$$

for every $x \in \mathbb{R}^n$.

Our Result

This and related claims are false!

Our Questions: Part 2

Can we recover the correctness theorem?

What do the outputs of autodiff even mean?

(e.g., ReLU'(0) = 0 in TensorFlow, PyTorch, ...)
Our Questions: Part 2

Can we recover the correctness theorem?

What do the outputs of autodiff even mean?
(e.g., ReLU' (0) = 0 in TensorFlow, PyTorch, ...)

They are not Clarke-subdifferentials [KL18]:

• $\partial^c f(x) := \text{conv} \left\{ \lim_{n \to 0} Df(x_n) \mid x_n \to x \text{ and } \exists Df(x_n) \right\}$. 
Our Questions: Part 2

Can we recover the correctness theorem?

What do the outputs of autodiff even mean?
(e.g., ReLU’(0) = 0 in TensorFlow, PyTorch, ...)

They are not Clarke-subdifferentials [KL18]:
• \( \partial^c f(x) := \text{conv} \{ \lim_{n \to 0} Df(x_n) \mid x_n \to x \text{ and } \exists Df(x_n) \} \).
• \( f(x) = \text{ReLU}(x) - \text{ReLU}(-x) : \partial^c f(0) = \{1\} \not\supseteq 0 = f'(0) \) (by autodiff).
Our Results: Part 2

Theorem $h_i$’s are differentiable everywhere

autodiff correctly computes $\nabla h(x)$

almost-everywhere
Our Results: Part 2

Theorem \( h_i \)'s are differentiable everywhere \( \Rightarrow \) autodiff correctly computes \( \nabla h(x) \) almost-everywhere so-called “PAP”

a.e.-differentiable “PAP”

new property we propose
Theorem

\[ h_i \text{'s are differentiable everywhere} \Rightarrow \text{autodiff correctly computes } \nabla h(x). \]

almost-\text{almost-}\text{so-called “PAP”}

new property we propose

a.e.-differentiable

pathological func’s
(Cantor func, ⋯)

“PAP”

func’s used in practice
(ReLU, ⋯)
Our Results: Part 2

Our Result: Prove the claim for PAP functions $h_i$'s.

Theorem: $h_i$'s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla h(x)$, almost-everywhere.

- so-called “PAP”
- new property we propose

a.e.-differentiable

- pathological func’s (Cantor func, ⋯)
- func’s used in practice (ReLU, ⋯)
Theorem \( h_1 \)’s are differentiable everywhere \( \Rightarrow \) autodiff correctly computes \( \nabla h(x) \).

Our Results: Part 2

**Our Result** Autodiff computes so-called “intensional derivatives” of \( h \).

**Our Result** Prove the claim for PAP functions \( h_1 \)’s.

- New property we propose
- Pathological func’s (Cantor func, …)
- Func’s used in practice (ReLU, …)
- A.e.-differentiable

\[ \text{PAP} \]
PAP Functions

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

\[
\left. f_1 \right|_{A_1}, \left. f_2 \right|_{A_2}, \ldots
\]

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

$$f_1|_{A_1}, f_2|_{A_2}, \ldots$$

such that

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

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

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

\[
\left. f_1 \right|_{A_1}, \left. f_2 \right|_{A_2}, \ldots
\]

such that

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

**Example** \( f(x) = \text{ReLU}(x) \).

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

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

$$f_1|_{A_1}, f_2|_{A_2}, \ldots$$

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

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

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

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

\[
\left. f \right|_{A_1}, \left. f \right|_{A_2}, \ldots
\]

such that

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

Example \( f(x) = \text{ReLU}(x) \).

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

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

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

$$f_1|_{A_1}, f_2|_{A_2}, \ldots$$

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

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

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

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

\[ 0 \]
PAP Functions

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

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

such that

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

Example

PAP functions include all functions used in practice.

Proposition PAP functions are a.e.-differentiable.
PAP Functions

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

$$f_1|_{A_1}, f_2|_{A_2}, \ldots$$

such that

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

**Observation** PAP functions include all functions used in practice.

**Proposition** PAP functions are a.e.-differentiable.

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

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

$$f_1|_{A_1}, f_2|_{A_2}, \ldots$$

such that

$$f_i : \mathbb{R}^n \to \mathbb{R}^m$$

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

**Example**

- $(f \cdot x = 0, A = \{x \in \mathbb{R} : x \leq 0\})$
- $(f' \cdot x = x, A' = \{x \in \mathbb{R} : x > 0\})$
- $(f' \cdot x = 0, A = \{x \in \mathbb{R} : x < 0\})$

**Observation** PAP functions include all functions used in practice.

**Proposition** PAP functions are a.e.-differentiable.

**Definition** PAP functions have “intensional derivatives”.

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

$$f_1|_{A_1}, f_2|_{A_2}, \ldots$$

such that

$$f_i : \mathbb{R}^n \to \mathbb{R}^m$$

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

Observation PAP functions include all functions used in practice.

Proposition PAP functions are a.e.-differentiable.

Definition PAP functions have “intensional derivatives”.

54
Intensional Derivatives

Example \( f(x) = \text{ReLU}(x) \).

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

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

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

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

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

- $f_3(x) = 7x, A_3 = \{x \in \mathbb{R} : x = 0\}$.
Intensional Derivatives

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

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

- (Conclusion)

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

  \( \int f'(x) \text{ d}x = E \) for \( x \leq 0 \)
  \( \int f'(x) \text{ d}x = 1 \) for \( x > 0 \)

analytic functions
Intensional Derivatives

Example \( f(x) = \text{ReLU}(x) \).

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

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

\( \frac{df}{dx} = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0 \end{cases} \)
**Intensional Derivatives**

**Proposition** Intensional derivative is a **total function**.

**Proposition** Intensional derivatives *always* satisfy the **chain rule**.

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

- $(f_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\})$, $(f_2(x) = x, A_2 = \{x \in \mathbb{R} : x > 0\})$.
  
  $df(x) = \begin{cases} 
  0 & \text{for } x \leq 0 \\
  1 & \text{for } x > 0 
  \end{cases}$

- $(f_1(x) = 0, A_1 = \{x \in \mathbb{R} : x < 0\})$, $(f_2(x) = x, A_2 = \{x \in \mathbb{R} : x > 0\})$.
  
  $df(x) = \begin{cases} 
  0 & \text{for } x < 0 \\
  1 & \text{for } x > 0 \\
  7 & \text{for } x = 0 
  \end{cases}$

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

$\frac{df}{dx} = \begin{cases} 
  0 & \text{for } x < 0 \\
  1 & \text{for } x > 0 \\
  7 & \text{for } x = 0 
  \end{cases}$
Intensional Derivatives

**Proposition**  Intensional derivative is a total function.

**Proposition**  Intensional derivatives always satisfy the chain rule.

**Proposition**  Intensional derivative $\equiv$ standard derivative $\ a.e.$

Example $f(x)$ = Red Bar$x$ = Green Bar

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

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

**Proposition** Intensional derivative is a total function.

**Proposition** Intensional derivatives always satisfy the chain rule.

**Proposition** Intensional derivative $\equiv$ standard derivative.

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

Theorem  For any $h = h_L \circ \cdots \circ h_1$ with PAP $h_L$, if autodiff uses an intensional derivative of $h_L$ for "$D^{-1}h_L$", autodiff computes an intensional derivative of $h$, and thus computes the correct gradient of $h$ a.e.
In TensorFlow and PyTorch,

- \( D\text{relu}(x) = 0 \) for \( x \leq 0 \); \( 1 \) for \( x > 0 \).

**Theorem** For any \( h = h_L \circ \cdots \circ h_1 \) with PAP \( h_L \),

if autodiff uses an intensional derivative of \( h_L \) for “\( D\)” \( h_L \),

autodiff computes an intensional derivative of \( h \),
and thus computes the correct gradient of \( h \) a.e.
In TensorFlow and PyTorch,

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

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

if autodiff uses an intensional derivative of $h_L$ for "$D$"$h_L$,

autodiff computes an intensional derivative of $h$,

and thus computes the correct gradient of $h$ a.e.
Correctness of Autodiff

In TensorFlow and PyTorch,

- \(D\)\(\text{relu}(x) = 0\) for \(x \leq 0\); 1 for \(x > 0\). √
- \(D\)\(\text{sqrt}(x) = \infty\) for \(x = 0\); \(1/2\sqrt{x}\) for \(x > 0\). ×

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

**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.
In TensorFlow and PyTorch,

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

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

**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 $\rightarrow$ higher-order.

• (First-order) intensional derivative = PAP function.

• Extended to higher-order derivatives. Enjoy the same properties.
Intensional Derivatives: Remarks

First-order $\rightarrow$ 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: $\partial^c f \in [\mathbb{R}^n \to \mathcal{P}(\mathbb{R}^{m \times n})]$.
  $\rightarrow$ 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.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autodiff</td>
<td>differentiable func’s</td>
<td>a.e.-differentiable func’s</td>
</tr>
</tbody>
</table>
# High-Level Messages

We often have discrepancy between theory and practice of ML algorithms. But our theoretical understanding on such discrepancy is still limited.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autodiff and many more</td>
<td>differentiable func’s</td>
<td>a.e.-differentiable func’s</td>
</tr>
</tbody>
</table>

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 Korea  
{wonyeol, yhk1344, hongseok}@kaist.ac.kr  

[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.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autodiff and many more</td>
<td>differentiable func’s</td>
<td>a.e.-differentiable func’s</td>
</tr>
<tr>
<td>Variational inference, ...</td>
<td>func’s <strong>with finite integrals</strong> (and other nice properties)</td>
<td>func’s <strong>with infinite integrals</strong> (or some bad properties)</td>
</tr>
</tbody>
</table>

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 U.
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.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autodiff and many more</td>
<td>differentiable func’s</td>
<td>a.e.-differentiable func’s</td>
</tr>
<tr>
<td>Variational inference, ...</td>
<td>func’s with finite integrals (and other nice properties)</td>
<td>func’s with infinite integrals (or some bad properties)</td>
</tr>
<tr>
<td>Most algorithms</td>
<td>func’s on reals</td>
<td>func’s on floating-points</td>
</tr>
</tbody>
</table>

Verifying Bit-Manipulations of Floating-Point

Wonyeol Lee  Rahul Sharma  Alex Aiken
Stanford University, USA
{wonyeol, sharmar, aiken}@cs.stanford.edu

On Automatically Proving the Correctness of math.h Implementations

WONYEOL LEE*, Stanford University, USA
RAHUL SHARMA, Microsoft Research, India
ALEX AIKEN, Stanford University, USA
Comments? Questions?