On Correctness of Automatic Differentiation for Non-Differentiable Functions

Wonyeol Lee¹, *
Hangyeol Yu¹, **
Xavier Rival²
Hongseok Yang¹

¹KAIST, South Korea
*now at Stanford, USA
²INRIA/ENS/CNRS, France
**now at Riiid!, South Korea

NeurIPS 2020 (Spotlight)
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?
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?

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
**Problem** For $F : \mathbb{R}^N \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^l$ differentiable everywhere, 

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

**Theorem** $F_l$’s are differentiable everywhere $\implies$ autodiff correctly computes $\nabla F(x)$.

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


**Theorem**  

$F_l$’s are differentiable everywhere $\implies$ autodiff correctly computes $\nabla F(x)$.

**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,  

$$D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$
**Problem**  For $F : \mathbb{R}^N \rightarrow \mathbb{R}$ given by $F(x) = (F_L \circ \cdots \circ F_1)(x)$, how to compute $\nabla F(x)$ correctly and efficiently?

**Theorem**  $F_i$’s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla F(x)$.

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

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

\[
D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.
\]
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: If $F_i$'s are differentiable everywhere, then autodiff correctly computes $\nabla F(x)$.

Example: $\text{ReLU}(x) = \max\{x, 0\}$ is non-differentiable on a measure-zero set.

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

$$D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$
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 $\implies$ autodiff correctly computes $\nabla F(x)$.

- e.g., ReLU$(x) = \max\{x, 0\}$ is almost-everywhere differentiable on a measure-zero set.

**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 $x \in \mathbb{R}^n$. 
Theorem  \( F_l \)'s are differentiable everywhere \( \Rightarrow \) autodiff correctly computes \( \nabla F(x) \).

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?

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 \( x \in \mathbb{R}^n \).
**Our Results**

**Problem** For $F : \mathbb{R}^N \rightarrow \mathbb{R}$ 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 $\iff$ autodiff correctly computes $\nabla F(x)$.

No, measure-zero non-differentiabilities matter!

**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 $x \in \mathbb{R}^n$. 
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** $F_i$’s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla F(x)$.

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

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

almost-

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

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

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

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

for a.e. \( x \in \mathbb{R} \).
Subtleties in Chain Rule (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. $x \in \mathbb{R}$.  

$f, g : \text{a.e.-differentiable and continuous}$
Subtleties in Chain Rule (1)

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

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

\[ (g \circ f)'(x) = g'(f(x)) \cdot f'(x) \quad \text{for a.e. } x \in \mathbb{R}. \]

Counterexample Involves the Cantor function.
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. $x \in \mathbb{R}$. 

Subtleties in Chain Rule (2)

Claim 2  For any \( f, g : \mathbb{R} \to \mathbb{R} \),
\[ f, g : \text{a.e.-differentiable and continuous} \]
\[\Rightarrow \] \( (g \circ f)'(x) = g'(f(x)) \cdot f'(x) \) for a.e. \( x \in \mathbb{R} \).

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

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

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

for a.e. $x \in \mathbb{R}$.
Claim 2 For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$

$f, g$: a.e.-differentiable and continuous \( \cdots (*) \)

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

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

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

\[ \Rightarrow \text{easy to check that } (*) \text{ holds.} \]
Subtleties in Chain Rule (2)

**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 \( g(y) = \text{ReLU}(y). \)

\[
\Rightarrow \\
g'(f(x)) \\
= g'(0) \\
= \text{undefined for all } x
\]
Claim 2  For any $f, g : \mathbb{R} \to \mathbb{R}$, and $g \circ f$, $f, g$: a.e.-differentiable and continuous

$$\forall x \in \mathbb{R} : (g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$

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

$$\Rightarrow (g \circ f)'(x) = 0$$

$$= g'(0) = 0$$

$$= \text{undefined for all } x$$
Subtleties in Chain Rule (2)

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

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

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

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

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

well-defined?

$$
dg(y) = \begin{cases} 
7 & \text{for } y = 0 \\
g'(y) & \text{for } y \neq 0 
\end{cases}
$$
Subtleties in Chain Rule (2)

Claim 2  For any \( f, g : \mathbb{R} \rightarrow \mathbb{R} \),
and \( g \circ f \)
\( f, g \): a.e.-differentiable and continuous
\( (g \circ f)'(x) = g'(f(x)) \cdot f'(x) \)

well-defined? for a.e. \( x \in \mathbb{R} \).

Counterexample  \( f(x) = 0 \) and \( g(y) = \text{ReLU}(y) \).
\( \Rightarrow \) \( (g \circ f)'(x) = dg(f(x)) \cdot f'(x) \) 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 $f, g : \mathbb{R} \rightarrow \mathbb{R}$, and $g \circ f$:

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

\[ (g \circ f)'(x) = dg(f(x)) \cdot df(x) \quad \text{for a.e. } x \in \mathbb{R}. \]

\[ \exists df, dg : \mathbb{R} \rightarrow \mathbb{R} \text{ such that } df \equiv f', dg \equiv g', \text{ and} \]
Subtleties in Chain Rule (3)

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

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

$$\left( g \circ f \right)'(x) \mathbin{\not=} dg(f(x)) \cdot df(x) \text{ for a.e. } x \in \mathbb{R}. $$

Counterexample  Involves the Cantor function again.
Our Results

Problem: For $F : \mathbb{R}^N \rightarrow \mathbb{R}$ 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 $\Rightarrow$ autodiff correctly computes $\nabla F(x)$ almost-everywhere.

Our Result: Disprove this and related claims.

Chain Rule: For $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^l$ differentiable everywhere, $D(g \circ f)(x) = Dg(f(x)) \cdot Df(x)$ almost-everywhere for every $x \in \mathbb{R}^n$. 
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 $F_i$’s are differentiable everywhere $\Rightarrow$ autodiff correctly computes $\nabla F(x)$.

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,

$$D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \quad \text{for every } x \in \mathbb{R}^n.$$
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 $\implies$ autodiff correctly computes $\nabla F(x)$ almost-everywhere.

Almost-called “PAP”

**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,

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

for every $x \in \mathbb{R}^n$. Almost-everywhere
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 $f_1|_{A_1}, f_2|_{A_2}, \cdots$ 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 \( f_1|_{A_1}, f_2|_{A_2}, \ldots \) such that

\[
f_i : \mathbb{R}^n \to \mathbb{R}^m \text{ is analytic and } A_i \subseteq \mathbb{R}^n \text{ is “analytic”}.
\]

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

![Graph of ReLU function](image)
Definition  \( f : \mathbb{R}^n \to \mathbb{R}^m \) is PAP (Piecewise Analytic under Analytic Partition) roughly iff \( f \) can be “decomposed” into \( f_1|_{A_1}, f_2|_{A_2}, \ldots \) such that

\[
f_i : \mathbb{R}^n \to \mathbb{R}^m \text{ is analytic and } A_i \subseteq \mathbb{R}^n \text{ is “analytic”}.
\]

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

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

\[
f_i : \mathbb{R}^n \to \mathbb{R}^m \text{ is analytic and } A_i \subseteq \mathbb{R}^n \text{ is “analytic”}.
\]

**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 functions
**Definition**  
$f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (= Piecewise Analytic under Analytic Partition) roughly iff $f$ can be “decomposed” into $f_1\big|_{A_1}, f_2\big|_{A_2}, \ldots$ such that $f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is “analytic”.

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

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

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

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is PAP (\textit{Piecewise Analytic under Analytic Partition}) roughly iff $f$ can be “decomposed” into $f_1|_{A_1}, f_2|_{A_2}, \cdots$ such that $f_i : \mathbb{R}^n \to \mathbb{R}^m$ is analytic and $A_i \subseteq \mathbb{R}^n$ is “analytic”.

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

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

Proposition PAP implies a.e.-differentiability.

Observation Virtually all functions used in practice are PAP.
Intensional Derivatives

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

- \( \begin{cases} f_1(x) = 0, & A_1 = \{x \in \mathbb{R} : x \leq 0\}, \\ f_2(x) = x, & A_2 = \{x \in \mathbb{R} : x > 0\} \end{cases} \)
- \( \begin{cases} f_1(x) = 0, & A_1 = \{x \in \mathbb{R} : x < 0\}, \\ f_2(x) = x, & A_2 = \{x \in \mathbb{R} : x > 0\}, \\ f_3(x) = 7x, & A_3 = \{x \in \mathbb{R} : x = 0\} \end{cases} \)
Example \( f(x) = \text{ReLU}(x) \).

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

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

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

- \( \begin{cases} f_1(x) = 0 \quad A_1 = \{x \in \mathbb{R} : x < 0\} \\ f_2(x) = x \quad A_2 = \{x \in \mathbb{R} : x > 0\} \\ f_3(x) = 7x \quad A_3 = \{x \in \mathbb{R} : x = 0\} \end{cases} \)

\( (f'_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\}), (f'_2(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}) \).

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

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

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

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

\[
\begin{align*}
    (f'_1(x) = 0, A_1 = \{x \in \mathbb{R} : x \leq 0\}), \\
    (f'_2(x) = 1, A_2 = \{x \in \mathbb{R} : x > 0\}), \\
    (f'_3(x) = 7, A_3 = \{x \in \mathbb{R} : x = 0\}).
\end{align*}
\]
Intensional Derivatives

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

**Proposition** Any intensional derivative \( a.e. \) standard derivative.

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

\[
\begin{align*}
(f'_1(x) = 0, A_1 &= \{x \in \mathbb{R} : x < 0\}), \\
(f'_2(x) = 1, A_2 &= \{x \in \mathbb{R} : x > 0\}), \\
(f'_3(x) = 7, A_3 &= \{x \in \mathbb{R} : x = 0\}).
\end{align*}
\]
Intensional Derivatives

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

**Proposition** Any intensional derivative a.e. standard derivative.

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

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

Theorem For PAP functions, what autodiff computes is an intensional derivative, and thus autodiff correctly computes gradients a.e.
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).