On Correctness of Automatic Differentiation for Non-Differentiable Functions

Wonyeol Lee — KAIST, South Korea (now at Stanford, USA)
Hangyeol Yu — KAIST, South Korea (now at Riiid, South Korea)
Xavier Rival — INRIA/ENS/CNRS, France
Hongseok Yang — KAIST, South Korea

NEURAL INFORMATION PROCESSING SYSTEMS

Automatic Differentiation

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 $\nabla F(x)$.

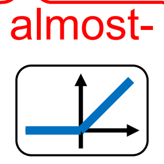
Autodiff $\approx$ efficient way of applying the chain rule.

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

Automatic Differentiation 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 $\boxed{\overset{?}{\implies}}$ autodiff correctly computes $\nabla F(x)$.
almost- almost-everywhere

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

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

Our Main Result (1)

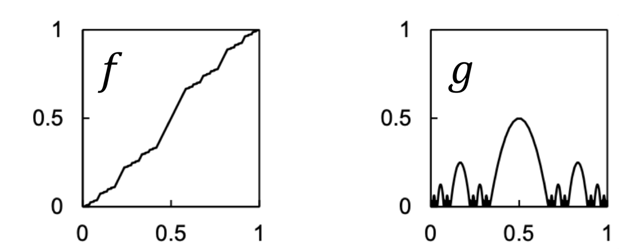
$\boxed{?}$'s do NOT hold: measure-zero non-differentiabilities do matter!

Subtleties in Chain Rule

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

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

  Proposition Claim 1 does NOT hold since $(g \circ f)'(x)$ can be undefined for a.e x.
  Counterexample Involves the Cantor function.



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

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

  Proposition Claim 2 does NOT hold since $g'(f(x))$ can be undefined for a.e x.
  Counterexample $f(x) = 0$ and $g(y) = \mathrm{ReLU}(y)$.

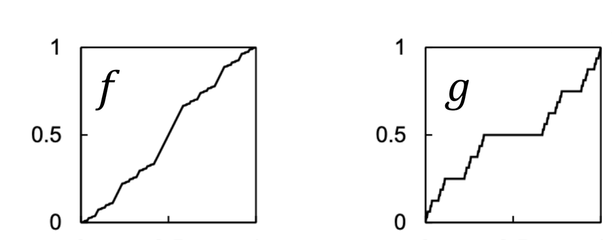
  Observation For the counterexample, an extended derivative dg of g satisfies:

  $$\boxed{(g \circ f)'(x) = dg(f(x)) \cdot f'(x)} \text{ for all } x \in \mathbb{R}.$$
  $$= 0 \qquad\qquad = 0$$
  $$dg(y) = \begin{cases} 7 & \text{for } y = 0 \\ g'(y) & \text{for } y \neq 0 \end{cases}$$

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

  $\overbrace{f, g}^{\text{and } g \circ f}$: a.e.-differentiable and continuous
  $\implies (g \circ f)'(x) = dg(f(x)) \cdot df(x)$ for a.e. $x \in \mathbb{R}$.
  $\exists\, df, dg : \mathbb{R} \to \mathbb{R}$ such that $df \overset{\text{a.e.}}{=} f'$, $dg \overset{\text{a.e.}}{=} g'$, and

  Proposition Claim 3 does NOT hold since the equality can fail to hold for a.e x.
  Counterexample Involves the Cantor function.



Our Main Result (2)

F_l's are so-called "PAP" $\boxed{\implies}$ autodiff correctly computes $\nabla F(x)$ a.e.

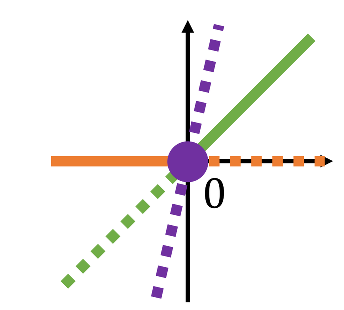
PAP Functions

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".

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

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



Proposition PAP implies a.e.-differentiability.
Observation Virtually all functions used in practice are PAP.

Intensional Derivatives

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

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

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

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

Proposition Intensional derivatives satisfy the chain rule.
Proposition Any intensional derivative $\overset{\text{a.e.}}{=}$ standard derivative.

Main Theorem

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