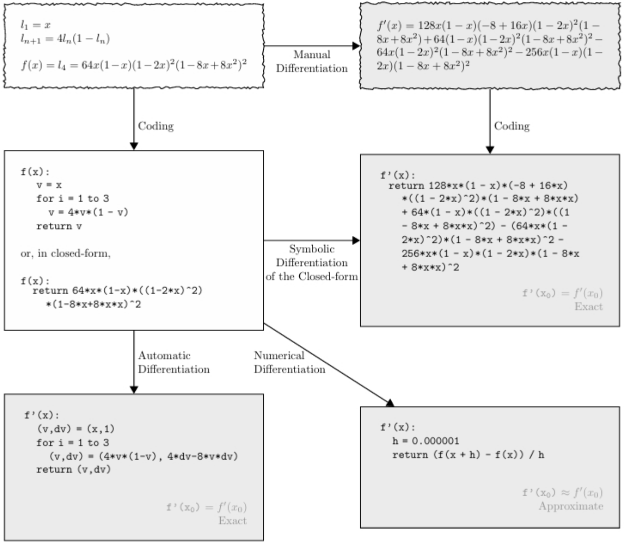
Modelos de Deep Learning

Automatic Differentiation

Universidad ORT Uruguay

8 de Setiembre, 2025

Técnicas de diferenciación



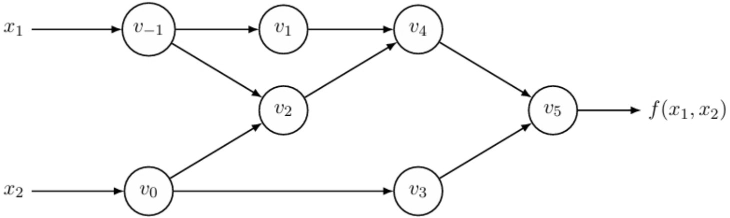
Idea central de AD

- Propagar derivadas mediante la regla de la cadena.
- Toda computación numérica se compone de operaciones elementales
  - suma
  - producto
  - exponencial
  - logaritmo
  - trigonométricas
  - etc

  cuyas derivadas son conocidas.
- Multiplicando derivadas se obtiene la derivada de la composición total.
- AD opera sobre trazas de evaluación y sobre su grafo computacional.

Grafo computacional

Sea $y = f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$



Se introducen variables intermedias v_i:

- $v_{-1} = x_1, \quad v_0 = x_2$
- $v_1 = \ln v_{-1}, \quad v_2 = v_{-1} \times v_0, \quad v_3 = \sin v_0, \quad v_4 = v_1 + v_2,$
- $v_5 = v_4 - v_3, \quad y = v_5$

AD Forward mode

- Para derivar respecto de x_1, asociar a cada v_i su tangente $\dot{v}_i = \partial v_i / \partial x_1$
- Propagar por la regla de la cadena en la *traza de evaluación*

$$\text{Inicialización: } \dot{x}_1 = 1, \;\; \dot{x}_2 = 0.$$

$$\text{Resultado: } \dot{y} = \frac{\partial y}{\partial x_1}.$$

AD Forward mode

Forward Primal Trace		
v_{-1}	$= x_1$	$= 2$
v_0	$= x_2$	$= 5$
v_1	$= \ln v_{-1}$	$= \ln 2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
v_3	$= \sin v_0$	$= \sin 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$
y	$= v_5$	$= 11.652$

Forward Tangent (Derivative) Trace		
$\dot{v}_{-1}$	$= \dot{x}_1$	$= 1$
$\dot{v}_0$	$= \dot{x}_2$	$= 0$
$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$	$= 1/2$
$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$	$= 1 \times 5 + 0 \times 2$
$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$	$= 0 \times \cos 5$
$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$	$= 0.5 + 5$
$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$	$= 5.5 - 0$
$\boldsymbol{\dot{y}}$	$\boldsymbol{= \dot{v}_5}$	$\mathbf{= 5.5}$

Números duales

- Representación: $v + \dot{v}\,\varepsilon$, con $\varepsilon^2 = 0$ y $\varepsilon \neq 0$.
- Funciones con números duales: $f(v + \dot{v}\,\varepsilon) = f(v) + f'(v)\,\dot{v}\,\varepsilon$.
- La regla del producto

$$\begin{aligned} f(v + \dot{v}\,\varepsilon) \cdot g(v + \dot{v}\,\varepsilon) &= [f(v) + f'(v)\,\dot{v}\,\varepsilon] \cdot [g(v) + g'(v)\,\dot{v}\,\varepsilon] \\ &= f(v)g(v) + [f'(v)g(v) + f(v)g'(v)]\,\dot{v}\,\varepsilon \end{aligned}$$

- La regla de la cadena se conserva:

$$f(g(v + \dot{v}\,\varepsilon)) = f(g(v)) + f'(g(v))g'(v)\,\dot{v}\,\varepsilon.$$

AD Reverse mode

- Dos fases:

  1. Forward pass original (se almacenan v_i y dependencias);
  2. Backward pass de *adjuntos* $\bar{v}_i = \partial y / \partial v_i$ desde las salidas a las entradas.

$$\text{Inicialización: } \bar{y} = 1.$$

$$\text{Resultado: } \bar{x}_1 = \frac{\partial y}{\partial x_1} \quad \bar{x}_2 = \frac{\partial y}{\partial x_2}.$$

AD Reverse Mode

Forward Primal Trace		
v_{-1}	$= x_1$	$= 2$
v_0	$= x_2$	$= 5$
v_1	$= \ln v_{-1}$	$= \ln 2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
v_3	$= \sin v_0$	$= \sin 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$
y	$= v_5$	$= 11.652$

Reverse Adjoint (Derivative) Trace			
$\boldsymbol{\bar{x}_1}$	$\boldsymbol{= \bar{v}_{-1}}$		$\mathbf{= 5.5}$
$\boldsymbol{\bar{x}_2}$	$\boldsymbol{= \bar{v}_0}$		$\mathbf{= 1.716}$
$\bar{v}_{-1}$	$= \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}}$	$= \bar{v}_{-1} + \bar{v}_1 / v_{-1}$	$= 5.5$
$\bar{v}_0$	$= \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0}$	$= \bar{v}_0 + \bar{v}_2 \times v_{-1}$	$= 1.716$
$\bar{v}_{-1}$	$= \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}}$	$= \bar{v}_2 \times v_0$	$= 5$
$\bar{v}_0$	$= \bar{v}_3 \frac{\partial v_3}{\partial v_0}$	$= \bar{v}_3 \times \cos v_0$	$= -0.284$
$\bar{v}_2$	$= \bar{v}_4 \frac{\partial v_4}{\partial v_2}$	$= \bar{v}_4 \times 1$	$= 1$
$\bar{v}_1$	$= \bar{v}_4 \frac{\partial v_4}{\partial v_1}$	$= \bar{v}_4 \times 1$	$= 1$
$\bar{v}_3$	$= \bar{v}_5 \frac{\partial v_5}{\partial v_3}$	$= \bar{v}_5 \times (-1)$	$= -1$
$\bar{v}_4$	$= \bar{v}_5 \frac{\partial v_5}{\partial v_4}$	$= \bar{v}_5 \times 1$	$= 1$
$\bar{v}_5$	$= \bar{y}$	$= 1$	

Forward mode vs Reverse mode

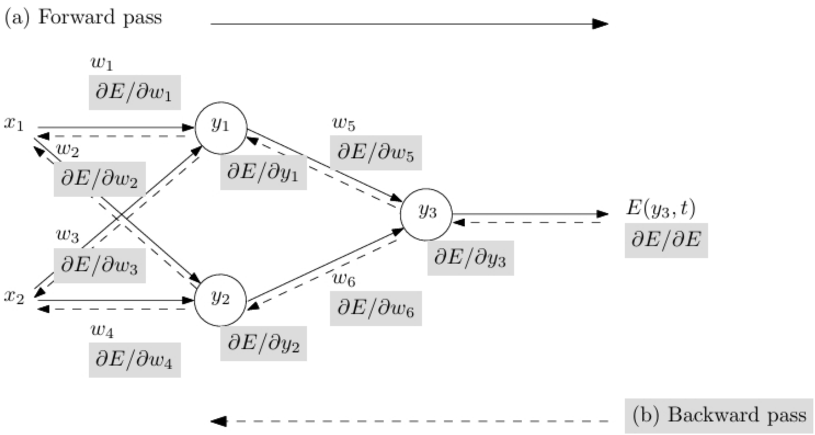
- Para $f : \mathbb{R}^n \to \mathbb{R}$, una sola pasada reverse produce $\nabla f = \left(\partial y / \partial x_i\right)_{i=1}^{n}$.

  Para $f : \mathbb{R}^n \to \mathbb{R}^m$: si $\text{ops}(f)$ es el costo de evaluar f, el Jacobiano cuesta

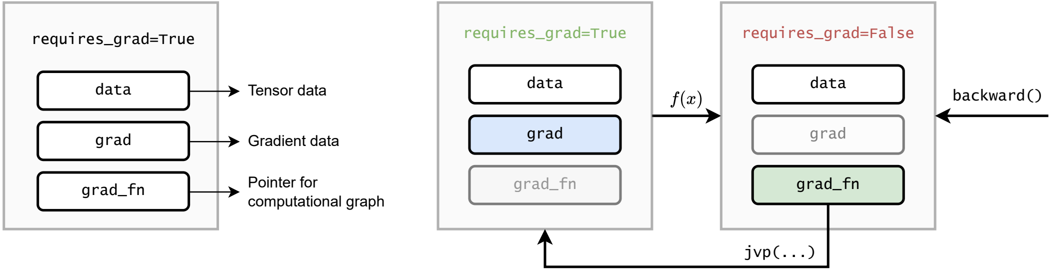
  - $n \cdot c \cdot \text{ops}(f)$ en AD forward mode
  - $m \cdot c \cdot \text{ops}(f)$ en AD reverse mode.

- El reverse es ventajoso cuando $m \ll n$ (es el caso de ML).
- Desventaja: costo en memoria, requiere almacenar muchos intermedios.

Backpropagation como caso especial de AD



AD en PyTorch



Referencias

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018).
Automatic differentiation in machine learning: a survey.
Journal of machine learning research, 18(153):1–43.

Scardapane, S. (2024).
Alice's adventures in a differentiable wonderland–volume i, a tour of the land.
arXiv preprint arXiv:2404.17625.