Backpropagation

Goodfellow et al. 2016 (sec. 6.5)

Gradient descent

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$$

- SGD step requires computing the gradient of the loss

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha^{(t)} \nabla_{\boldsymbol{\theta}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_t), y_t)$$

- Use chain rule

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) = \ell'(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x})$$


BACKPROPAGANDA

History

S. Linnainmaa
1970

P. Werbos
1974

D. Rumelhart

G. Hinton

R. Williams

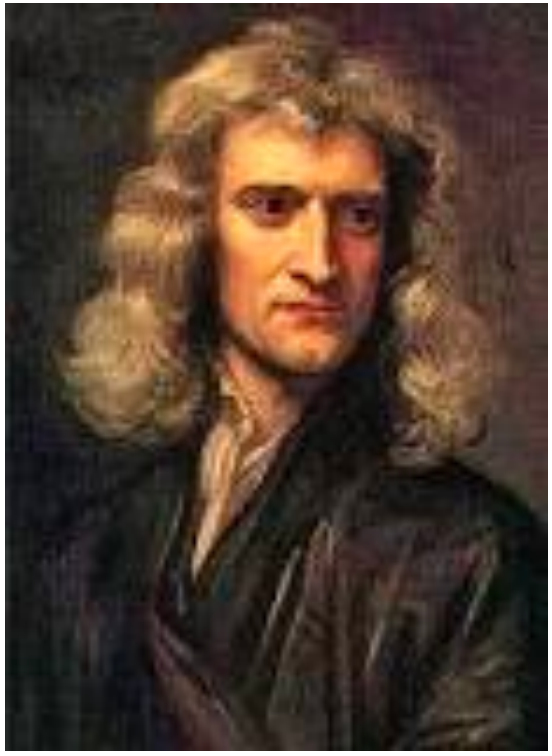
nature
International journal of science

Letter | Published: 09 October 1986

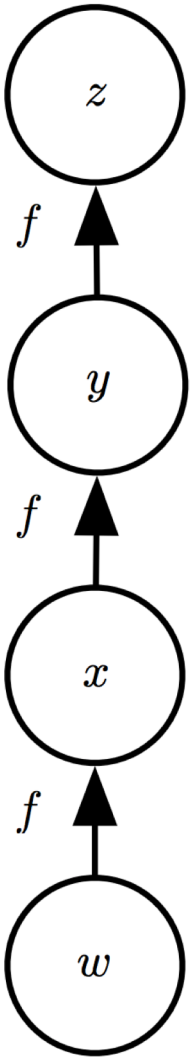
Learning representations by back-propagating errors

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams

History



Scalar example



$$\frac{\partial z}{\partial w}$$
$$= \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}\frac{\partial x}{\partial w}$$
$$= f'(y)f'(x)f'(w)$$
$$= f'(f(f(w)))f'(f(w))f'(w)$$

Back-prop avoids computing this twice

Scalar example

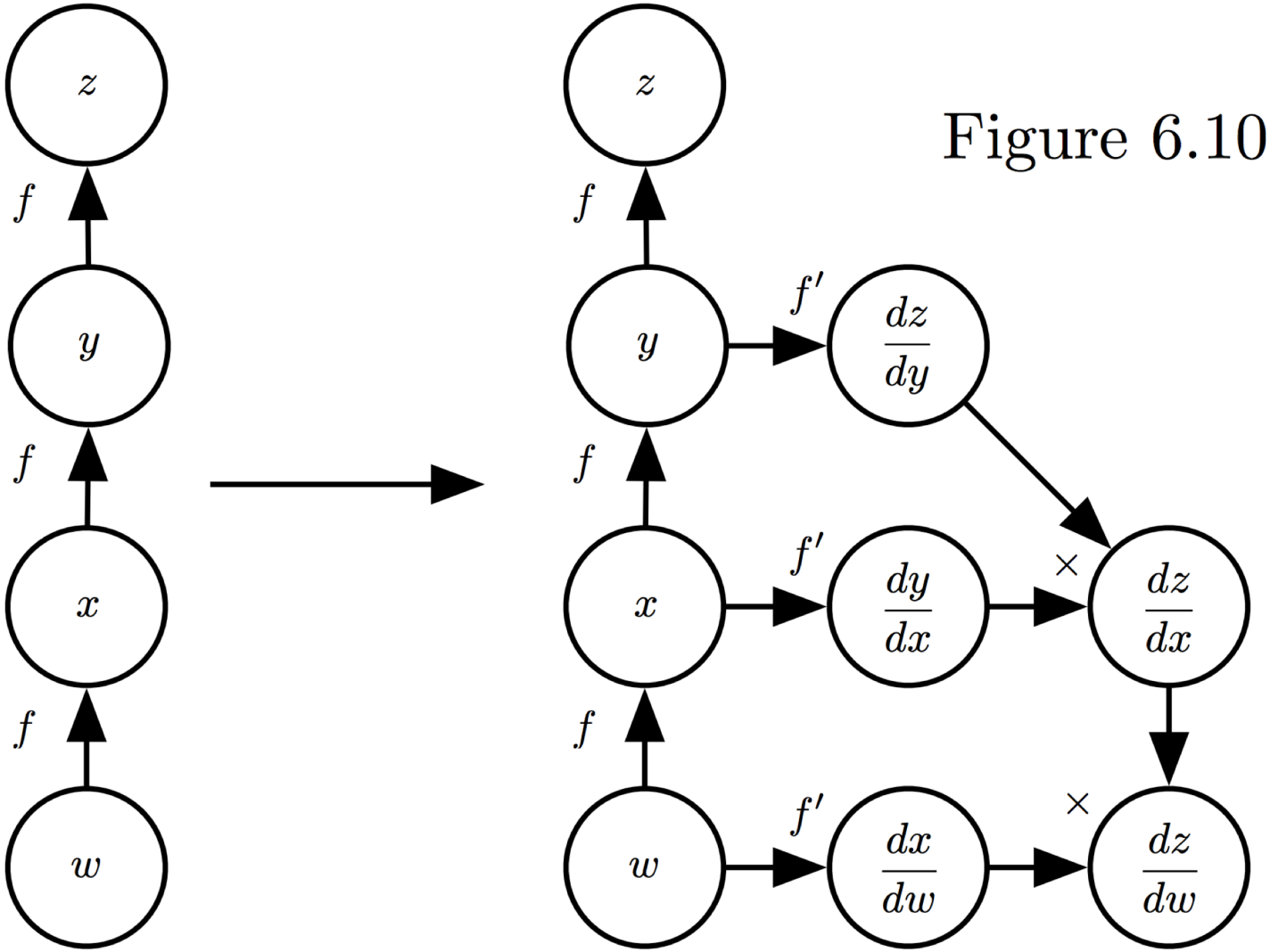


Figure 6.10

Goodfellow et al. 2016

Vector case

$$\ell(\psi(\mathbf{A}\,\varphi(\mathbf{B}\mathbf{x})))$$

$$\begin{aligned}
&& \mathbf{x} &\in \mathbb{R}^m \\
\mathbf{z}' &= \mathbf{B}\mathbf{x} & \mathbf{z}' &\in \mathbb{R}^n \\
\mathbf{y}' &= \varphi(\mathbf{z}') & \mathbf{y}' &\in \mathbb{R}^n \\
\mathbf{z} &= \mathbf{A}\mathbf{y}' & \mathbf{z} &\in \mathbb{R}^k \\
\mathbf{y} &= \psi(\mathbf{z}) & \mathbf{y} &\in \mathbb{R}^k \\
l &= \ell(\mathbf{y}) & l &\in \mathbb{R}
\end{aligned}$$

Slide credit: A. Bronstein

Chain rule

$$\frac{\partial l}{\partial \mathbf{A}} = \frac{\partial \mathbf{y}}{\partial \mathbf{A}} \frac{\partial l}{\partial \mathbf{y}} = \sum_{i=1}^{k} \frac{\partial y_i}{\partial \mathbf{A}} \frac{\partial l}{\partial y_i}$$

$$\delta \mathbf{A} = \frac{\partial \mathbf{y}}{\partial \mathbf{A}} \delta \mathbf{y} = \sum_{i=1}^{k} \frac{\partial y_i}{\partial \mathbf{A}} \delta y_i$$

$$\delta \mathbf{y} = \frac{\partial l}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial l}{\partial y_1} \\ \vdots \\ \frac{\partial l}{\partial y_k} \end{pmatrix}$$

$$\delta \mathbf{A} = \frac{\partial l}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial l}{\partial a_{11}} & \frac{\partial l}{\partial a_{1n}} \\ \frac{\partial l}{\partial a_{k1}} & \frac{\partial l}{\partial a_{kn}} \end{pmatrix}$$

Chain rule

$$y_i = \psi\left(\sum_{j=1}^{n} a_{ij} y_j'\right) = \psi(z_i)$$

$$\frac{\partial y_i}{\partial \mathbf{A}} = \begin{pmatrix} - & -\mathbf{0}- & - \\ \psi'(z_i)y_1' & \cdots & \psi'(z_i)y_n' \\ - & -\mathbf{0}- & - \end{pmatrix}$$

Chain rule

$$\frac{\partial y_i}{\partial \mathbf{A}} = \begin{pmatrix} - & -\mathbf{0}- & - \\ \psi'(z_i)y_1' & \cdots & \psi'(z_i)y_n' \\ - & -\mathbf{0}- & - \end{pmatrix}$$

$$\delta\mathbf{A} = \sum_{i=1}^{k} \frac{\partial y_i}{\partial \mathbf{A}} \delta y_i = \begin{pmatrix} \delta y_1 \psi'(z_1) y_1' & \cdots & \delta y_1 \psi'(z_1) y_n' \\ \vdots & \ddots & \vdots \\ \delta y_k \psi'(z_k) y_1' & \cdots & \delta y_k \psi'(z_k) y_n' \end{pmatrix}$$

$$= \begin{pmatrix} \delta y_1 & & \\ & \ddots & \\ & & \delta y_k \end{pmatrix} \begin{pmatrix} \psi'(z_1) & & \\ & \ddots & \\ & & \psi'(z_k) \end{pmatrix} \begin{pmatrix} -\mathbf{y}'^T - \\ \vdots \\ -\mathbf{y}'^T - \end{pmatrix}$$

Chain rule

$$\frac{\partial l}{\partial \mathbf{B}} = \frac{\partial \mathbf{y}'}{\partial \mathbf{B}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}'} \frac{\partial l}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}'}{\partial \mathbf{B}} \delta \mathbf{y}' = \sum_{i=1}^{n} \frac{\partial y_i'}{\partial \mathbf{B}} \delta y_i' \qquad \delta \mathbf{y}' = \frac{\partial \mathbf{y}}{\partial \mathbf{y}'} \frac{\partial l}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}}{\partial \mathbf{y}'} \delta \mathbf{y}$$

Chain rule

$$y_i' = \varphi\left(\sum_{j=1}^{m} b_{ij} x_j\right) = \varphi(z_i')$$

$$\frac{\partial y_i'}{\partial \mathbf{B}} = \begin{pmatrix} - & - & - \\ \varphi'(z_i')x_1 & \cdots & \varphi'(z_i')x_m \\ - & - & - \end{pmatrix}$$

Chain rule

$$\frac{\partial y_i'}{\partial \mathbf{B}} = \begin{pmatrix} - & - & - \\ \varphi'(z_i')x_1 & \cdots & \varphi'(z_i')x_m \\ - & - & - \end{pmatrix}$$

$$\delta \mathbf{B} = \sum_{i=1}^{n} \frac{\partial y_i'}{\partial \mathbf{B}} \delta y_i'$$

$$= \begin{pmatrix} \delta y_i' & & \\ & \ddots & \\ & & \delta y_n' \end{pmatrix} \begin{pmatrix} \varphi'(z_1') & & \\ & \ddots & \\ & & \varphi'(z_n') \end{pmatrix} \begin{pmatrix} - & \mathbf{x}^T & - \\ & \vdots & \\ - & \mathbf{x}^T & - \end{pmatrix}$$

Chain rule

$$y_i = \psi\left(\sum_{j=1}^{n} a_{ij} y_j'\right) = \psi(z_i)$$

$$\frac{\partial y_i}{\partial y_j'} = \psi'(z_i)\, a_{ij}$$

Chain rule

$$\frac{\partial y_i}{\partial y'_j} = \psi'(z_i)\, a_{ij}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{y}'} = \begin{pmatrix} \frac{\partial y_1}{\partial y'_1} & \cdots & \frac{\partial y_k}{\partial y'_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial y'_n} & \cdots & \frac{\partial y_k}{\partial y'_n} \end{pmatrix} = \begin{pmatrix} \psi'(z_1)a_{11} & \cdots & \psi'(z_k)a_{k1} \\ \vdots & \ddots & \vdots \\ \psi'(z_1)a_{1n} & \cdots & \psi'(z_k)a_{kn} \end{pmatrix}$$

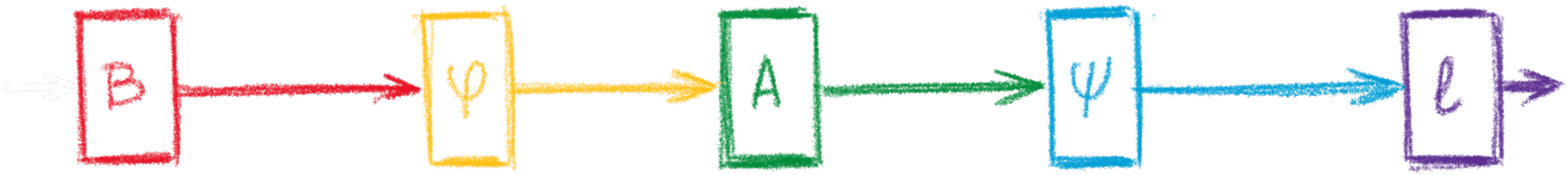
Chain rule

$$\frac{\partial \mathbf{y}}{\partial \mathbf{y}'} = \begin{pmatrix} \psi'(z_1) & & \\ & \ddots & \\ & & \psi'(z_k) \end{pmatrix} \mathbf{A}^T$$

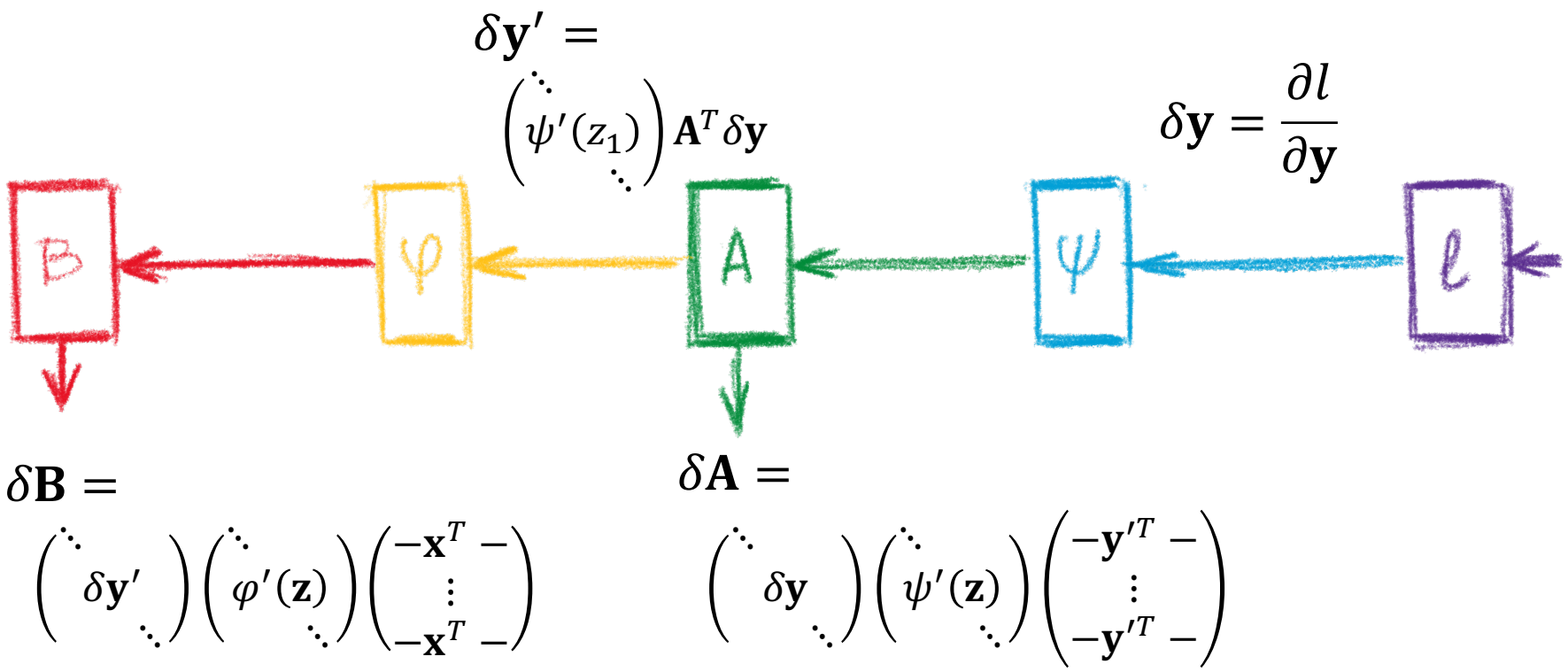
$$\delta \mathbf{y}' = \frac{\partial \mathbf{y}}{\partial \mathbf{y}'} \delta \mathbf{y} = \begin{pmatrix} \psi'(z_1) & & \\ & \ddots & \\ & & \psi'(z_k) \end{pmatrix} \mathbf{A}^T \delta \mathbf{y}$$

Chain rule

$\mathbf{z}' = \mathbf{B}\mathbf{x}$ $\quad \mathbf{y}' = \varphi(\mathbf{z}')$ $\quad \mathbf{z} = \mathbf{A}\mathbf{y}'$ $\quad \mathbf{y} = \psi(\mathbf{z})$ $\quad l = \ell(\mathbf{y})$



Chain rule


$\delta \mathbf{y}' = \begin{pmatrix} \ddots & \\ & \psi'(z_1) \\ & & \ddots \end{pmatrix} \mathbf{A}^T \delta \mathbf{y}$
$\delta \mathbf{y} = \frac{\partial l}{\partial \mathbf{y}}$
B
φ
A
ψ
ℓ
$\delta \mathbf{B} = \begin{pmatrix} \ddots & \\ & \delta \mathbf{y}' \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & \\ & \varphi'(\mathbf{z}) \\ & & \ddots \end{pmatrix} \begin{pmatrix} -\mathbf{x}^T - \\ \vdots \\ -\mathbf{x}^T - \end{pmatrix}$
$\delta \mathbf{A} = \begin{pmatrix} \ddots & \\ & \delta \mathbf{y} \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & \\ & \psi'(\mathbf{z}) \\ & & \ddots \end{pmatrix} \begin{pmatrix} -\mathbf{y}'^T - \\ \vdots \\ -\mathbf{y}'^T - \end{pmatrix}$

Forward and backward pass

Input: $\mathbf{y}_0 = \mathbf{x}$

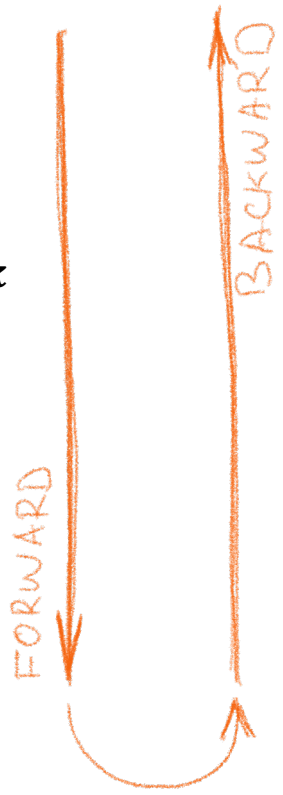
For $k = 1, \ldots, L$

$$\mathbf{z}_k = \mathbf{W}_k \mathbf{y}_{k-1} + \mathbf{b}_k$$

$$\mathbf{y}_k = \varphi_k(\mathbf{z}_k)$$

Output: $\mathbf{y} = \mathbf{y}_L$

Loss: $l = \ell(\mathbf{y})$



For $k = L, L-1, \ldots, 1$

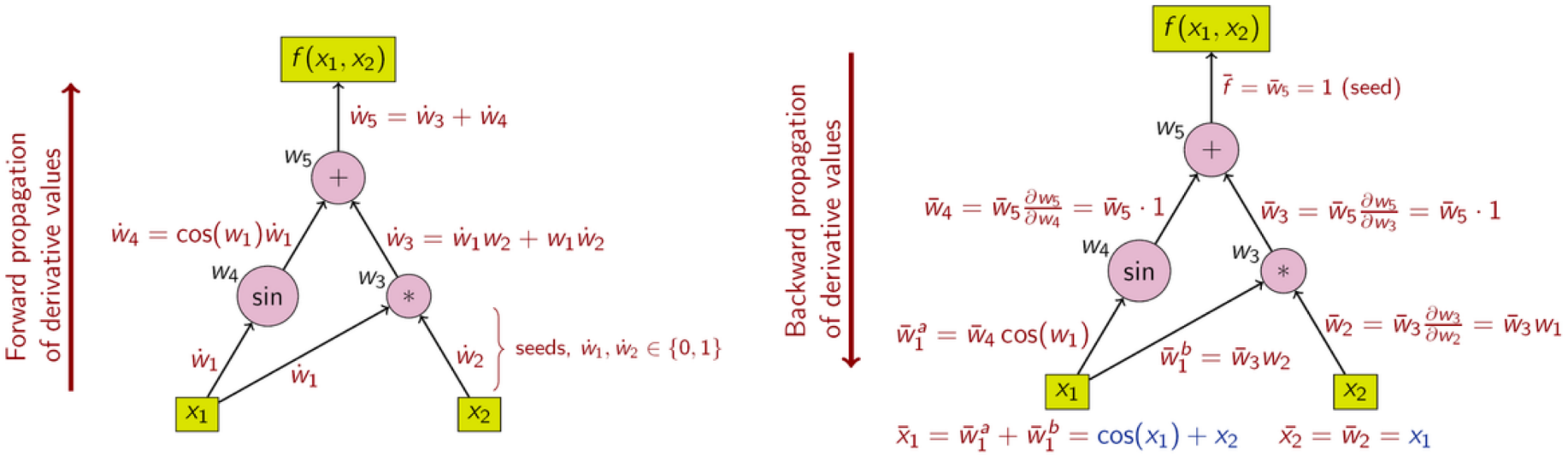
$$\delta \mathbf{W}_k = \begin{pmatrix} \ddots & & \\ & \delta \mathbf{y}_k & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & \varphi_k'(\mathbf{z}_k) & \\ & & \ddots \end{pmatrix} \mathbf{1} \mathbf{y}_{k-1}^T$$

$$\delta \mathbf{b}_k = \begin{pmatrix} \ddots & & \\ & \delta \mathbf{y}_k & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & \varphi_k'(\mathbf{z}_k) & \\ & & \ddots \end{pmatrix}$$

$$\delta \mathbf{y}_{k-1} = \begin{pmatrix} \ddots & & \\ & \varphi_k'(\mathbf{z}_k) & \\ & & \ddots \end{pmatrix} \mathbf{W}_k^T \delta \mathbf{y}_k$$

Loss grad: $\delta \mathbf{y}_L = \nabla \ell(\mathbf{y}_L)$

Automatic differentiation



"Differentiable programming"