Backpropagation

Michael Taylor

Backpropagation is a means of using the chain rule to evaluate the derivative of a map defined as a multi-form composition of a certain type. The basic example is

$$G(W,x) = F_2(W_2 F_1(W_1 x)). \tag{1}$$

Here $W = (W_1, W_2)$ and W_j are matrices. We have

$$x \in \mathbb{R}^{m_0}, \quad W_1 x \in \mathbb{R}^{n_1}, \quad W_1 : \mathbb{R}^{m_0} \to \mathbb{R}^{n_1} \text{ linear.}$$

Generally,

$$\begin{aligned} F_j : \mathbb{R}^{n_j} &\longrightarrow \mathbb{R}^{m_j}, \quad \text{differentiable,} \\ W_j : \mathbb{R}^{m_{j-1}} &\longrightarrow \mathbb{R}^{n_j}, \quad \text{linear.} \end{aligned} \tag{2}$$

We write $W_j \in \mathcal{L}(\mathbb{R}^{m_{j-1}}, \mathbb{R}^{n_j})$. Also suppose

$$V_j \in \mathcal{L}(\mathbb{R}^{m_{j-1}}, \mathbb{R}^{n_j}). \tag{3}$$

We are interested in the derivative of $G(W,x)$ with respect to W. Let us use the notation

$$\begin{aligned} D_{V_2} G(W,x) &= \frac{d}{dt} G(W_1, W_2 + tV_2, x)\big|_{t=0}, \\ D_{V_1} G(W,x) &= \frac{d}{dt} G(W_1 + tV_1, W_2, x)\big|_{t=0}. \end{aligned} \tag{4}$$

Applying the chain rule gives

$$\begin{aligned} D_{V_2} G(W,x) &= DF_2(W_2 F_1(W_1 x)) V_2 F_1(W_1 x), \\ D_{V_1} G(W,x) &= DF_2(W_2 F_1(W_1 x)) W_2 DF_1(W_1 x) V_1 x. \end{aligned} \tag{5}$$

Note that the leading factors on the right sides of these formulas are identical; they are both

$$DF_2(W_2 F_1(W_1 x)). \tag{6}$$

Keeping this in mind enables us to avoid unnecessary duplication in computing the W-derivative.

For the general case, consider a k-fold composition, defined recursively by

$$G_k(W,x) = F_k(W_k G_{k-1}(W',x)), \quad W' = (W_1,\ldots,W_{k-1}). \tag{7}$$

Set

$$D_{V_j} G_k(W,x) = \frac{d}{dt} G_k(W_1,\ldots,W_j + tV_j,\ldots,W_k,x)\Big|_{t=0}. \tag{8}$$

Then

$$D_{V_k} G_k(W,x) = DF_k(W_k G_{k-1}(W',x)) V_k G_{k-1}(W',x), \tag{9}$$

and, for $j \le k-1$,

$$D_{V_j} G_k(W,x) = DF_k(W_k G_{k-1}(W',x)) W_k D_{V_j} G_{k-1}(W',x). \tag{10}$$

As before, the factors

$$DF_k(W_k G_{k-1}(W',x)) \tag{11}$$

appearing first on the right sides of (9) and (10) are identical. The factor

$$D_{V_j} G_{k-1}(W',x) \tag{12}$$

appearing on the right side of (10) is computed inductively. This is the “back” in backpropagation.