Layer

$$\begin{aligned}\text{input}: \quad x \in \mathbb{R}^m \quad &\rightarrow \quad Wx + b \ \in \mathbb{R}^n \\ &\rightarrow \quad f(Wx + b) \ \in \mathbb{R}^n \quad : \text{output}\end{aligned} \tag{1}$$

Network

$$\text{input}: \quad x_1 \quad \rightarrow \quad y_1 = f_1(W_1 x_1 + b_1) \tag{2}$$

$$x_2 = y_1 \quad \rightarrow \quad y_2 = f_2(W_2 x_2 + b_2) \tag{3}$$

$$\vdots$$

$$x_N = y_{N-1} \quad \rightarrow \quad y_N = f_N(W_N x_N + b_N) \quad : \text{output} \tag{4}$$

Training

Training data: a collection of pairs (x, y_{true})

$$\text{Loss function:} \qquad L = \frac{1}{N_{\text{samples}}} \sum_{x \in \text{samples}} ||y_N(x) - y_{\text{true}}||^2 \quad \text{(for example)} \tag{5}$$

Parameters repeatedly updated according to

$$W \ \rightarrow \ W - r \frac{\partial L}{\partial W}\bigg|_{\text{batch}} \tag{6}$$

$$b \ \rightarrow \ b - r \frac{\partial L}{\partial b}\bigg|_{\text{batch}} \tag{7}$$

where "batch" is a randomly selected subset of the training data.

Gradient

Let λ_n denote a parameter (W or b) in the n-th layer. Then

$$\frac{\partial L}{\partial \lambda_n} = \frac{\partial L}{\partial y_N^i} \frac{\partial y_N^i}{\partial x_N^j} \frac{\partial y_{N-1}^j}{\partial x_{N-1}^k} \cdots \frac{\partial y_n^l}{\partial \lambda_n} \tag{8}$$

and introducing the notation

$$(J_{f_m})^i{}_j := \frac{\partial f_m^i(x)}{\partial x^j}\bigg|_{x = W_m x_m + b_m} \tag{9}$$

the derivatives in Eq. (8) are given by

$$\frac{\partial y_m^i}{\partial x_m^j} = (J_{f_m})^i{}_k (W_m)^k{}_j \tag{10}$$

$$\frac{\partial y_n^i}{\partial (W_n)^j{}_k} = (J_{f_n})^i{}_j x_n^k \tag{11}$$

$$\frac{\partial y_n^i}{\partial b_n^j} = (J_{f_n})^i{}_j \tag{12}$$