Two-Layer Neural Network Gradient Derivation

Terrence Alsup

August 2023

1 Set-up

A two-layer neural network is a function $f : \mathbb{R}^{m_0} \to \mathbb{R}^{m_2}$ defined by

$$f(\mathbf{x}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$$

where

$$\mathbf{W}_1 \in \mathbb{R}^{m_1 \times m_0}, \quad \mathbf{W}_2 \in \mathbb{R}^{m_2 \times m_1}$$

are the weights,

$$\mathbf{b}_1 \in \mathbb{R}^{m_1}, \quad \mathbf{b}_2 \in \mathbb{R}^{m_2}$$

are the biases, and $\sigma(x) = \max(x, 0)$ is the rectified linear unit (ReLU) activation function. Here we overload notation so that σ applies element-wise to each entry in a matrix or vector as well.

Given paired training data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the loss function is the mean-squared error

$$\ell(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2; \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \|f(\mathbf{x}_i) - \mathbf{y}_i\|^2 .$$

To vectorize over a batch of data points define the matrices

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} \in \mathbb{R}^{m_0 \times n}, \quad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n \end{bmatrix} \in \mathbb{R}^{m_2 \times n},$$

and the vector $\mathbf{1}_n = [1, \ldots, 1] \in \mathbb{R}^n$ as vector of all ones. A batch evaluation of data points is given by

$$f(\mathbf{X}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\mathbf{1}_n^\top) + \mathbf{b}_2\mathbf{1}_n^\top .$$

The outer-product $\mathbf{b}_1\mathbf{1}_n^\top$ ensures that $\mathbf{b}_1$ is added to each column of the matrix $\mathbf{W}_1\mathbf{X}$.

2 Gradient derivations

To derive each of the gradients of the loss function with respect to the parameters $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$, we use the chain rule. For a scalar parameter θ,

$$\frac{\partial \ell}{\partial \theta} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \frac{\partial f}{\partial \theta}$$

and $\frac{\partial f}{\partial \theta} \in \mathbb{R}^{m_2}$.

2.1 Gradient w.r.t. $\mathbf{b}_2$

Starting with the easiest gradient to compute, let $b_2^{(j)}$ be a single component of the vector $\mathbf{b}_2$ with $1 \leq j \leq m_2$, then

$$\frac{\partial \ell}{\partial b_2^{(j)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \frac{\partial f}{\partial b_2^{(j)}} \,.$$

Since $f(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_{m_2}(\mathbf{x})] \in \mathbb{R}^{m_2}$ we have

$$\frac{\partial f}{\partial b_2^{(j)}} = \begin{bmatrix} \frac{\partial f_1}{\partial b_2^{(j)}} & \cdots & \frac{\partial f_{m_2}}{\partial b_2^{(j)}} \end{bmatrix} ,$$

so that for $1 \leq k \leq m_2$

$$\frac{\partial f_k}{\partial b_2^{(j)}} = \delta_{kj}$$

which is 1 if $k = j$ and 0 otherwise. Thus, the gradient is the coordinate vector

$$\frac{\partial f}{\partial b_2^{(j)}} = \mathbf{e}_j \in \mathbb{R}^{m_2} \,.$$

Plugging this in gives

$$\frac{\partial \ell}{\partial b_2^{(j)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{e}_j = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^{(j)} \,,$$

which is the j-th component of the vector

$$\frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i) \,.$$

Therefore,

$$\frac{\partial \ell}{\partial \mathbf{b}_2} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i) \,.$$

2.2 Gradient w.r.t. $\mathbf{W}_2$

Now let $W_2^{(jk)}$ be an entry in the matrix $\mathbf{W}_2$ for $1 \leq j \leq m_2$ and $1 \leq k \leq m_1$. We have

$$\frac{\partial \ell}{\partial W_2^{(jk)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \frac{\partial f}{\partial W_2^{(jk)}} ,$$

and

$$f_i(\mathbf{x}) = b_2^{(i)} + \sum_{p=1}^{m_1} W_2^{(ip)} \sigma^{(p)} .$$

Notice that we have changed the use of the index i here and are considering an arbitrary $\mathbf{x}$. Again $\sigma^{(p)}$ is the p-th component of the vector $\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$. Then,

$$\frac{\partial f_i}{\partial W_2^{(jk)}} = \delta_{ij} \sigma^{(k)} ,$$

so that

$$\frac{\partial f}{\partial W_2^{(jk)}} = \sigma^{(k)} \mathbf{e}_j .$$

Plugging this in gives

$$\frac{\partial \ell}{\partial W_2^{(jk)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^{(j)} \sigma^{(k)} .$$

However, $(f(\mathbf{x}_i) - \mathbf{y}_i)^{(j)} \sigma^{(k)}$ is the (j,k)-th entry of the rank-1 matrix given by the outer product

$$(f(\mathbf{x}_i) - \mathbf{y}_i) \sigma(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1)^\top ,$$

and therefore

$$\frac{\partial \ell}{\partial \mathbf{W}_2} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i) \sigma(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1)^\top \in \mathbb{R}^{m_2 \times m_1} .$$

2.3 Gradient w.r.t. $\mathbf{b}_1$

Proceeding as before

$$\frac{\partial \ell}{\partial b_1^{(j)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \frac{\partial f}{\partial b_1^{(j)}} .$$

Since

$$f_i(\mathbf{x}) = b_2^{(i)} + \sum_{p=1}^{m_1} W_2^{(ip)} \sigma^{(p)}$$

we have by the chain rule that

$$\frac{\partial f_i}{\partial b_1^{(j)}} = \sum_{p=1}^{m_1} W_2^{(ip)} \frac{\partial \sigma^{(p)}}{\partial b_1^{(j)}} ,$$

which is the i-th component of the vector $\mathbf{W}_2 \frac{\partial \sigma}{\partial b_1^{(j)}}$ and therefore

$$\frac{\partial f}{\partial b_1^{(j)}} = \mathbf{W}_2 \frac{\partial \sigma}{\partial b_1^{(j)}} .$$

We have

$$\sigma^{(i)} = \sigma \left(\sum_{k=1}^{m_0} W_1^{(ik)} x^{(k)} + b_1^{(i)} \right) ,$$

so that

$$\frac{\partial \sigma^{(i)}}{\partial b_1^{(j)}} = \delta_{ij} \sigma' \left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1 \right)^{(j)} ,$$

and therefore

$$\frac{\partial \sigma}{\partial b_1^{(j)}} = \mathbf{e}_j \sigma' \left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1 \right)^{(j)} .$$

Plugging this in gives

$$\frac{\partial \ell}{\partial b_1^{(j)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{W}_2 \mathbf{e}_j \sigma' \left(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1 \right)^{(j)} .$$

Notice that $(f(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{W}_2 \mathbf{e}_j$ gives the j-th component of the vector

$$\mathbf{W}_2^\top (f(\mathbf{x}_i) - \mathbf{y}_i) ,$$

and thus

$$\frac{\partial \ell}{\partial \mathbf{b}_1} = \frac{2}{n} \sum_{i=1}^{n} \mathbf{W}_2^\top (f(\mathbf{x}_i) - \mathbf{y}_i) \odot \sigma' \left(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1 \right) ,$$

where $\odot$ represents element-wise multiplication.

2.4 Gradient w.r.t. $\mathbf{W}_1$

Again, proceeding by the chain rule and for $1 \leq j \leq m_1$ and $1 \leq k \leq m_0$ we have

$$\frac{\partial \ell}{\partial W_1^{(jk)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{W}_2 \frac{\partial \sigma}{\partial W_1^{(jk)}}$$

We have

$$\frac{\partial \sigma^{(i)}}{\partial W_1^{(jk)}} = \delta_{ij} \sigma' (\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)^{(j)} x^{(k)} ,$$

so that

$$\frac{\partial \sigma}{\partial W_1^{(jk)}} = \mathbf{e}_j \sigma' (\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)^{(j)} x^{(k)} ,$$

and

$$\frac{\partial \ell}{\partial W_1^{(jk)}} = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{W}_2 \mathbf{e}_j \sigma' (\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1)^{(j)} \mathbf{x}_i^{(k)}$$

and

$$\frac{\partial \ell}{\partial \mathbf{W}_1} = \frac{2}{n}\sum_{i=1}^{n} \left(\mathbf{W}_2^\top \left(f(\mathbf{x}_i) - \mathbf{y}_i\right) \odot \sigma'\left(\mathbf{W}_1\mathbf{x}_i + \mathbf{b}_1\right)\right)\mathbf{x}_i^\top .$$

3 Vectorized implementation

Let

$$\mathbf{R} = f(\mathbf{X}) - \mathbf{Y} = \begin{bmatrix}\mathbf{r}_1 & \cdots & \mathbf{r}_n\end{bmatrix} \in \mathbb{R}^{m_2 \times n}$$

be the matrix of residuals with

$$\mathbf{r}_i = f(\mathbf{x}_i) - \mathbf{y}_i \, .$$

Also define the matrices $\mathbf{S}, \mathbf{D} \in \mathbb{R}^{m_1 \times n}$ by

$$\mathbf{S} = \sigma(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\mathbf{1}_n^\top) \, , \quad \mathbf{D} = \sigma'(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\mathbf{1}_n^\top) \, .$$

Also, for brevity and to avoid doing the multiplication twice set

$$\mathbf{V} = \mathbf{W}_2^\top \mathbf{R} = \mathbf{W}_2^\top \left(f(\mathbf{X}) - \mathbf{Y}\right) \in \mathbb{R}^{m_1 \times n} \, .$$

3.1 Implementation of gradient w.r.t. $\mathbf{b}_2$

Since

$$\frac{\partial \ell}{\partial \mathbf{b}_2} = \frac{2}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - \mathbf{y}_i) = \frac{2}{n}\sum_{i=1}^{n}\mathbf{r}_i$$

is a sum across the columns of the matrix $\mathbf{R}$. We may write

$$\frac{\partial \ell}{\partial \mathbf{b}_2} = \frac{2}{n}\texttt{sum}\left(\mathbf{R},\ \texttt{columns}\right)$$

to denote the operation.

3.2 Implementation of gradient w.r.t. $\mathbf{W}_2$

We may write

$$\frac{\partial \ell}{\partial \mathbf{W}_2} = \frac{2}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - \mathbf{y}_i)\sigma(\mathbf{W}_1\mathbf{x}_i + \mathbf{b}_1)^\top = \frac{2}{n}\sum_{i=1}^{n}\mathbf{r}_i\mathbf{s}_i^\top \, ,$$

where

$$\mathbf{S} = \begin{bmatrix}\mathbf{s}_1 & \cdots & \mathbf{s}_n\end{bmatrix} \, .$$

The (j,k)-th entry is

$$\left(\sum_{i=1}^{n}\mathbf{r}_i\mathbf{s}_i^\top\right)^{(jk)} = \sum_{i=1}^{n}\mathbf{R}^{(ji)}\mathbf{S}^{(ki)} = (\mathbf{R}\mathbf{S}^\top)^{(jk)}$$

and therefore

$$\frac{\partial \ell}{\partial \mathbf{W}_2} = \frac{2}{n}\mathbf{R}\mathbf{S}^\top \, .$$

3.3 Implementation of gradient w.r.t. $\mathbf{b}_1$

Writing

$$\frac{\partial \ell}{\partial \mathbf{b}_1} = \frac{2}{n} \sum_{i=1}^{n} \mathbf{W}_2^\top \left(f(\mathbf{x}_i) - \mathbf{y}_i \right) \odot \sigma' \left(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1 \right) = \frac{2}{n} \sum_{i=1}^{n} \mathbf{v}_i \odot \mathbf{d}_i \,,$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} \,, \quad \mathbf{D} = \begin{bmatrix} \mathbf{d}_1 & \cdots & \mathbf{d}_n \end{bmatrix} \,.$$

Thus,

$$\frac{\partial \ell}{\partial \mathbf{b}_1} = \frac{2}{n} \texttt{sum} \left(\mathbf{V} \odot \mathbf{D}, \ \texttt{columns} \right) \,.$$

3.4 Implementation of gradient w.r.t. $\mathbf{W}_1$

Following the steps of the previous derivation

$$\begin{aligned} \frac{\partial \ell}{\partial \mathbf{W}_1} &= \frac{2}{n} \sum_{i=1}^{n} \left(\mathbf{W}_2^\top \left(f(\mathbf{x}_i) - \mathbf{y}_i \right) \odot \sigma' \left(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1 \right) \right) \mathbf{x}_i^\top \\ &= \frac{2}{n} \sum_{i=1}^{n} \left(\mathbf{v}_i \odot \mathbf{d}_i \right) \mathbf{x}_i^\top \,. \end{aligned}$$

By looking at the (j,k)-th component as before we see that

$$\frac{\partial \ell}{\partial \mathbf{W}_1} = \frac{2}{n} \left(\mathbf{V} \odot \mathbf{D} \right) \mathbf{X}^\top \,.$$