3.6. Back-propagation

François Fleuret

https://fleuret.org/ee559/

May 8, 2020
We want to train an MLP by minimizing a loss over the training set

\[ \mathcal{L}(w, b) = \sum_n \ell(f(x_n; w, b), y_n). \]
We want to train an MLP by minimizing a loss over the training set

$$\mathcal{L}(w, b) = \sum_n \ell(f(x_n; w, b), y_n).$$

To use gradient descent, we need the expression of the gradient of the per-sample loss $\ell_n = \ell(f(x_n; w, b), y_n)$ with respect to the parameters, e.g.

$$\frac{\partial \ell_n}{\partial w_{i,j}^{(l)}} \quad \text{and} \quad \frac{\partial \ell_n}{\partial b_i^{(l)}}.$$
For clarity, we consider a single training sample $x$, and introduce $s^{(1)}, \ldots, s^{(L)}$ as the summations before activation functions.

$$x^{(0)} = x \xrightarrow{w^{(1)}, b^{(1)}} s^{(1)} \xrightarrow{\sigma} x^{(1)} \xrightarrow{w^{(2)}, b^{(2)}} s^{(2)} \xrightarrow{\sigma} \ldots \xrightarrow{w^{(L)}, b^{(L)}} s^{(L)} \xrightarrow{\sigma} x^{(L)} = f(x; w, b).$$
For clarity, we consider a single training sample $x$, and introduce $s^{(1)}, \ldots, s^{(L)}$ as the summations before activation functions.

$$x^{(0)} = x \xrightarrow{w^{(1)}, b^{(1)}} s^{(1)} \xrightarrow{\sigma} x^{(1)} \xrightarrow{w^{(2)}, b^{(2)}} s^{(2)} \xrightarrow{\sigma} \ldots \xrightarrow{w^{(L)}, b^{(L)}} s^{(L)} \xrightarrow{\sigma} x^{(L)} = f(x; w, b).$$

Formally we set $x^{(0)} = x$,

$$\forall l = 1, \ldots, L, \quad \left\{ \begin{array}{l}
  s^{(l)} = w^{(l)} x^{(l-1)} + b^{(l)} \\
  x^{(l)} = \sigma(s^{(l)})
\end{array} \right.,$$

and we set the output of the network as $f(x; w, b) = x^{(L)}$. 
For clarity, we consider a single training sample \( x \), and introduce \( s^{(1)}, \ldots, s^{(L)} \) as the summations before activation functions.

\[
x^{(0)} = x \xrightarrow{w^{(1)}, b^{(1)}} s^{(1)} \xrightarrow{\sigma} x^{(1)} \xrightarrow{w^{(2)}, b^{(2)}} s^{(2)} \xrightarrow{\sigma} \ldots \xrightarrow{w^{(L)}, b^{(L)}} s^{(L)} \xrightarrow{\sigma} x^{(L)} = f(x; w, b).
\]

Formally we set \( x^{(0)} = x \),

\[
\forall l = 1, \ldots, L, \quad \begin{cases} 
  s^{(l)} = w^{(l)}x^{(l-1)} + b^{(l)} \\
  x^{(l)} = \sigma (s^{(l)})
\end{cases}
\]

and we set the output of the network as \( f(x; w, b) = x^{(L)} \).

This is the forward pass.
The core principle of the back-propagation algorithm is the “chain rule” from differential calculus:

\[(g \circ f)' = (g' \circ f)f'.\]

The linear approximation of a composition of mappings is the product of their individual linear approximations.
The core principle of the back-propagation algorithm is the “chain rule” from differential calculus:

$$(g \circ f)' = (g' \circ f)f'.$$

The linear approximation of a composition of mappings is the product of their individual linear approximations.

This generalizes to longer compositions and higher dimensions

$$J_{f_N \circ f_{N-1} \circ \cdots \circ f_1}(x) = J_{f_N}(f_{N-1}(\cdots(x)))) \cdots J_{f_3}(f_2(f_1(x))) J_{f_2}(f_1(x)) J_{f_1}(x)$$

where $J_f(x)$ is the Jacobian of $f$ at $x$, that is the matrix of the linear approximation of $f$ in the neighborhood of $x$. 
Since $s(l)$ influences $l$ only through $x(l)$ with $x(l) = \sigma(s(l))$,
we have
\[
\frac{\partial l}{\partial s(l)} \overset{w(l), b(l)}{\to} \sigma \overset{x(l)}{\to} x(l)
\]
And since $x(l-1)$ influences $l$ only through the $s(l)$ with
$s(l) = \sum_{j} w(l)_{ij} x(l-1)_j + b(l)_i$,
we have
\[
\frac{\partial l}{\partial x(l-1)_j} = \sum_{i} \frac{\partial l}{\partial s(l)_i} \frac{\partial s(l)_i}{\partial x(l-1)_j} = \sum_{i} \frac{\partial l}{\partial s(l)_i} w(l)_{ij}
\]
Since $s_i^{(l)}$ influences $\ell$ only through $x_i^{(l)}$ with

$$x_i^{(l)} = \sigma(s_i^{(l)})$$
Since $s_i^{(l)}$ influences $\ell$ only through $x_i^{(l)}$ with

$$x_i^{(l)} = \sigma(s_i^{(l)})$$

we have

$$\frac{\partial\ell}{\partial s_i^{(l)}} = \frac{\partial\ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}}$$
\[ X^{(l-1)} \xrightarrow{w^{(l)}, b^{(l)}} S^{(l)} \xrightarrow{\sigma} X^{(l)} \]

Since \( s_i^{(l)} \) influences \( \ell \) only through \( x_i^{(l)} \) with

\[ x_i^{(l)} = \sigma(s_i^{(l)}) , \]

we have

\[ \frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)}) , \]
Since $s_i^{(l)}$ influences $\ell$ only through $x_i^{(l)}$ with

$$x_i^{(l)} = \sigma(s_i^{(l)}),$$

we have

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)}),$$

And since $x_j^{(l-1)}$ influences $\ell$ only through the $s_i^{(l)}$ with

$$s_i^{(l)} = \sum_j w_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)},$$
Since $s^{(l)}_i$ influences $\ell$ only through $x^{(l)}_i$ with

$$x^{(l)}_i = \sigma(s^{(l)}_i),$$

we have

$$\frac{\partial \ell}{\partial s^{(l)}_i} = \frac{\partial \ell}{\partial x^{(l)}_i} \frac{\partial x^{(l)}_i}{\partial s^{(l)}_i} = \frac{\partial \ell}{\partial x^{(l)}_i} \sigma'(s^{(l)}_i),$$

And since $x^{(l-1)}_j$ influences $\ell$ only through the $s^{(l)}_i$ with

$$s^{(l)}_i = \sum_j w^{(l)}_{i,j} x^{(l-1)}_j + b^{(l)}_i,$$

we have

$$\frac{\partial \ell}{\partial x^{(l-1)}_j} = \sum_i \frac{\partial \ell}{\partial s^{(l)}_i} \frac{\partial s^{(l)}_i}{\partial x^{(l-1)}_j}.$$
Since $s_i^{(l)}$ influences $\ell$ only through $x_i^{(l)}$ with

$$x_i^{(l)} = \sigma(s_i^{(l)})$$

we have

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

And since $x_j^{(l-1)}$ influences $\ell$ only through the $s_i^{(l)}$ with

$$s_i^{(l)} = \sum_j w_{i,j} x_j^{(l-1)} + b_i^{(l)}$$

we have

$$\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_j \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}.$$
\[ \chi^{(l-1)} \xrightarrow{w^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} \chi^{(l)} \]

Since \( s^{(l)}_i \) influences \( \ell \) only through \( \chi^{(l)}_i \) with

\[ \chi^{(l)}_i = \sigma(s^{(l)}_i) , \]

we have

\[ \frac{\partial \ell}{\partial s^{(l)}_i} = \frac{\partial \ell}{\partial \chi^{(l)}_i} \frac{\partial \chi^{(l)}_i}{\partial s^{(l)}_i} = \frac{\partial \ell}{\partial \chi^{(l)}_i} \sigma'(s^{(l)}_i) , \]

And since \( \chi^{(l-1)}_j \) influences \( \ell \) only through the \( s^{(l)}_i \) with

\[ s^{(l)}_i = \sum_j w_{i,j} \chi^{(l-1)}_j + b^{(l)}_i , \]

we have

\[ \frac{\partial \ell}{\partial \chi^{(l-1)}_j} = \sum_i \frac{\partial \ell}{\partial s^{(l)}_i} \frac{\partial s^{(l)}_i}{\partial \chi^{(l-1)}_j} = \sum_i \frac{\partial \ell}{\partial s^{(l)}_i} w_{i,j} . \]
Since $w_{i,j}^{(l)}$ and $b_{i}^{(l)}$ influences $\ell$ only through $s_{i}^{(l)}$ with

$$s_{i}^{(l)} = \sum_{j} w_{i,j}^{(l)} x_{j}^{(l-1)} + b_{i}^{(l)},$$
Since $w_{i,j}^{(l)}$ and $b_i^{(l)}$ influences $\ell$ only through $s_i^{(l)}$ with

$$s_i^{(l)} = \sum_j w_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)},$$

we have

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial w_{i,j}^{(l)}}$$
Since $w_{i,j}^{(l)}$ and $b_{i}^{(l)}$ influences $\ell$ only through $s_{i}^{(l)}$ with

$$s_{i}^{(l)} = \sum_{j} w_{i,j}^{(l)} x_{j}^{(l-1)} + b_{i}^{(l)},$$

we have

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_{i}^{(l)}} \frac{\partial s_{i}^{(l)}}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_{i}^{(l)}} x_{j}^{(l-1)},$$
\[
\begin{align*}
x^{(l-1)} & \xrightarrow{w^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)} \\
\text{Since } w_{i,j}^{(l)} \text{ and } b_i^{(l)} \text{ influences } \ell \text{ only through } s_i^{(l)} \text{ with } \\
s_i^{(l)} &= \sum_j w_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}, \\
\text{we have} \\
\frac{\partial \ell}{\partial w_{i,j}^{(l)}} &= \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)}, \\
\frac{\partial \ell}{\partial b_i^{(l)}} &= \frac{\partial \ell}{\partial s_i^{(l)}}.
\end{align*}
\]
To summarize: we can compute \( \frac{\partial \ell}{\partial x_i^{(L)}} \) from the definition of \( \ell \), and recursively **propagate backward** the derivatives of the loss w.r.t the activations with

\[
\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})
\]

and

\[
\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}^{(l)}.
\]
To summarize: we can compute $\frac{\partial \ell}{\partial x_i^{(l)}}$ from the definition of $\ell$, and recursively **propagate backward** the derivatives of the loss w.r.t the activations with

$$
\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})
$$

and

$$
\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}^{(l)}.
$$

And then compute the derivatives w.r.t the parameters with

$$
\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)},
$$

and

$$
\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}.
$$
To write in tensorial form we will use a notation for the Jacobian to make explicit the variable wrt which the derivatives are computed. For $\psi : \mathbb{R}^N \rightarrow \mathbb{R}^M$, 

$$
\left[ \frac{\partial \psi}{\partial x} \right] = \begin{pmatrix}
\frac{\partial \psi_1}{\partial x_1} & \cdots & \frac{\partial \psi_1}{\partial x_N} \\
\vdots & \ddots & \vdots \\
\frac{\partial \psi_M}{\partial x_1} & \cdots & \frac{\partial \psi_M}{\partial x_N}
\end{pmatrix},
$$

and if $\psi : \mathbb{R}^{N\times M} \rightarrow \mathbb{R}$, we will use the notation

$$
\left[ \frac{\partial \psi}{\partial w} \right] = \begin{pmatrix}
\frac{\partial \psi}{\partial w_{1,1}} & \cdots & \frac{\partial \psi}{\partial w_{1,M}} \\
\vdots & \ddots & \vdots \\
\frac{\partial \psi}{\partial w_{N,1}} & \cdots & \frac{\partial \psi}{\partial w_{N,M}}
\end{pmatrix}.
$$
\[
\sigma(x^{(l-1)} \cdot w^{(l)} + b^{(l)})
\]

\[
\frac{\partial l}{\partial x} = \sigma'(s^{(l)}) \cdot \sigma' \cdot x^{(l-1)}
\]

\[
\frac{\partial l}{\partial b} = \sigma'(s^{(l)}) \cdot x^{(l-1)}
\]

\[
\frac{\partial l}{\partial w} = \sigma'(s^{(l)}) \cdot x^{(l-1)}
\]
\[
\frac{\partial \ell}{\partial x(l-1)} \times \left[ \begin{array}{c}
\frac{\partial \ell}{\partial w(l)} \\
\frac{\partial \ell}{\partial b(l)}
\end{array} \right] \odot \sigma' \cdot ^\top \times \left[ \begin{array}{c}
\frac{\partial \ell}{\partial x(l-1)} \\
\frac{\partial \ell}{\partial b(l)} \\
\frac{\partial \ell}{\partial w(l)}
\end{array} \right]
\]
\[
\frac{\partial \ell}{\partial x(l)} = \sigma^T \sum_{j=1}^{H} \sigma_{j} \left( \sum_{k=1}^{L} w(l)_{kj} \frac{\partial \ell}{\partial s(l)} + b(l)_j \right) + \frac{\partial \ell}{\partial x(l-1)}
\]
\[
\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma' \left( s_i^{(l)} \right)
\]
$\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i w_{i,j}^{(l)} \frac{\partial \ell}{\partial s_i^{(l)}}$
\[
\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}
\]
\[
\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_{j}^{(l-1)}
\]
\[
\frac{\partial \ell}{\partial x(l-1)} \cdot T \times \left[ \frac{\partial \ell}{\partial x(l)} \right] = \sigma' \cdot \left[ \frac{\partial \ell}{\partial s(l)} \right] = \left[ \frac{\partial \ell}{\partial x(l)} \right]
\]
Forward pass

Compute the activations.

\[ x^{(0)} = x, \quad \forall l = 1, \ldots, L, \quad \left\{ \begin{array}{l} s^{(l)} = w^{(l)}x^{(l-1)} + b^{(l)} \\ x^{(l)} = \sigma (s^{(l)}) \end{array} \right. \]
\textbf{Forward pass}

Compute the activations.

\[ x^{(0)} = x, \quad \forall l = 1, \ldots, L, \]
\[ \begin{aligned}
    s^{(l)} &= w^{(l)} x^{(l-1)} + b^{(l)} \\
    x^{(l)} &= \sigma(s^{(l)})
\end{aligned} \]

\textbf{Backward pass}

Compute the derivatives of the loss wrt the activations.

\[ \begin{aligned}
    &\begin{cases}
        \left[ \frac{\partial \ell}{\partial x^{(l)}} \right] \text{ from the definition of } \ell \\
        \text{if } l < L, \left[ \frac{\partial \ell}{\partial x^{(l)}} \right] = (w^{(l+1)})^T \left[ \frac{\partial \ell}{\partial s^{(l+1)}} \right]
    \end{cases} \\
    \left[ \frac{\partial \ell}{\partial s^{(l)}} \right] &= \left[ \frac{\partial \ell}{\partial x^{(l)}} \right] \odot \sigma'(s^{(l)})
\end{aligned} \]

Compute the derivatives of the loss wrt the parameters.

\[ \begin{aligned}
    \left[ \frac{\partial \ell}{\partial w^{(l)}} \right] &= \left[ \frac{\partial \ell}{\partial s^{(l)}} \right] (x^{(l-1)})^T \\
    \left[ \frac{\partial \ell}{\partial b^{(l)}} \right] &= \left[ \frac{\partial \ell}{\partial s^{(l)}} \right]
\end{aligned} \]
Forward pass

Compute the activations.

\[ x^{(0)} = x, \quad \forall l = 1, \ldots, L, \quad \left\{ \begin{array}{l} s^{(l)} = w^{(l)} x^{(l-1)} + b^{(l)} \\ x^{(l)} = \sigma (s^{(l)}) \end{array} \right. \]

Backward pass

Compute the derivatives of the loss wrt the activations.

\[
\begin{cases}
\frac{\partial \ell}{\partial x^{(l)}} & \text{from the definition of } \ell \\
\frac{\partial \ell}{\partial x^{(l)}} = (w^{(l+1)})^T \frac{\partial \ell}{\partial s^{(l+1)}} & \text{if } l < L
\end{cases}
\]

Compute the derivatives of the loss wrt the parameters.

\[
\begin{align*}
\frac{\partial \ell}{\partial w^{(l)}} &= \left( \frac{\partial \ell}{\partial s^{(l)}} \right) (x^{(l-1)})^T \\
\frac{\partial \ell}{\partial b^{(l)}} &= \left( \frac{\partial \ell}{\partial s^{(l)}} \right)
\end{align*}
\]

Gradient step

Update the parameters.

\[
\begin{align*}
w^{(l)} &\leftarrow w^{(l)} - \eta \left( \frac{\partial \ell}{\partial w^{(l)}} \right) \\
b^{(l)} &\leftarrow b^{(l)} - \eta \left( \frac{\partial \ell}{\partial b^{(l)}} \right)
\end{align*}
\]
In spite of its hairy formalization, the backward pass is a simple algorithm: apply the chain rule again and again.

As for the forward pass, it can be expressed in tensorial form. Heavy computation is concentrated in linear operations, and all the non-linearities go into component-wise operations.
Regarding computation, since the costly operation for the forward pass is

\[ s^{(l)} = w^{(l)} x^{(l-1)} + b^{(l)} \]

and for the backward

\[
\left[ \frac{\partial \ell}{\partial x^{(l)}} \right] = \left( w^{(l+1)} \right)^T \left[ \frac{\partial \ell}{\partial s^{(l+1)}} \right]
\]

and

\[
\left[ \frac{\partial \ell}{\partial w^{(l)}} \right] = \left[ \frac{\partial \ell}{\partial s^{(l)}} \right] \left( x^{(l-1)} \right)^T,
\]

the rule of thumb is that the backward pass is twice more expensive than the forward one.
The end