Deep learning

3.6. Back-propagation

François Fleuret

https://fleuret.org/dlc/
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

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

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.

$$\frac{\partial \ell_n}{\partial w^{(l)}_{i,j}} \quad \text{and} \quad \frac{\partial \ell_n}{\partial b^{(l)}_i}.$$
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 $x^{(0)} = x$,

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

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

This is the forward pass.
For clarity, we consider a single training sample $x$, and introduce $s^{(1)}, \ldots, s^{(L)}$ as the summations before activation functions.

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

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 \(x\).
\[ x^{(l-1)} \xrightarrow{w^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)} \]
Since $s_i^{(l)}$ influences $\ell$ only through $x_i^{(l)}$ with

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

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

we have

$$\frac{\partial \ell}{\partial s^{(l)}_i} = \frac{\partial \ell}{\partial x^{(l)}_i} \frac{\partial x^{(l)}_i}{\partial s^{(l)}_i}$$
Since \( s_i^{(l)} \) influences \( \ell \) only through \( x_i^{(l)} \) with
\[
x_i^{(l)} = \sigma(s_i^{(l)}),
\]
we have
\[
\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)}),
\]
\[ x^{(l-1)} \xrightarrow{w^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)} \]

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

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

we have

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

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

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

$$\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial x_j^{(l-1)}}$$
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

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

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_{i}^{(l)}} \frac{\partial s_{i}^{(l)}}{\partial w_{i,j}^{(l)}},$$

$$\frac{\partial \ell}{\partial b_{i}^{(l)}} = \frac{\partial \ell}{\partial s_{i}^{(l)}} \frac{\partial s_{i}^{(l)}}{\partial b_{i}^{(l)}}.$$
\[
\begin{align*}
x^{(l-1)} \xrightarrow{w^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}
\end{align*}
\]

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

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_{i}^{(l)}} \frac{\partial s_{i}^{(l)}}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_{i}^{(l)}} x_{j}^{(l-1)},$$
\[
\begin{align*}
\mathbf{x}^{(l-1)} &\xrightarrow{w^{(l)}, b^{(l)}} \mathbf{s}^{(l)} \xrightarrow{\sigma} \mathbf{x}^{(l)}
\end{align*}
\]

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
\[
\begin{align*}
\frac{\partial \ell}{\partial w_{i,j}^{(l)}} &= \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)}, \\
\frac{\partial \ell}{\partial b_i^{(l)}} &= \frac{\partial \ell}{\partial s_i^{(l)}}.
\end{align*}
\]
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

$$\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} .$$

This is the **backward pass**.
To write in tensorial form we will use the following notation for the gradient of a loss $\ell : \mathbb{R}^N \rightarrow \mathbb{R}$,

$$\left[ \frac{\partial \ell}{\partial x} \right] = \begin{pmatrix} \frac{\partial \ell}{\partial x_1} \\ \vdots \\ \frac{\partial \ell}{\partial x_N} \end{pmatrix},$$

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

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

\[
x^{(l-1)} \times \quad + \quad s^{(l)} \quad \sigma \quad x^{(l)}
\]
\[
\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma' \left( s_i^{(l)} \right)
\]
\[
\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i w_{i,j}^{(l)} \frac{\partial \ell}{\partial s_i^{(l)}}
\]
\[
\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}
\]
\[
\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)}
\]
\( x(l-1) \) \( \times \) \( \frac{\partial \ell}{\partial x(l-1)} \) \( \times \cdot \top \times \) \( \frac{\partial \ell}{\partial s(l)} \) \( \sigma \) \( s(l) \) \( \sigma' \) \( \frac{\partial \ell}{\partial b(l)} \) \( \frac{\partial \ell}{\partial w(l)} \) \( \frac{\partial \ell}{\partial x(l)} \)
Forward pass

Compute the activations.

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

Backward pass
Compute the derivatives of the loss w.r.t. the activations.
\[
\begin{cases}
\frac{\partial \ell}{\partial x^{(l)}} & \text{from the definition of } \ell \\
\text{if } l < L, \frac{\partial \ell}{\partial x^{(l)}} &= (w^{(l+1)})^\top \left[ \frac{\partial \ell}{\partial s^{(l+1)}} \right]
\end{cases}
\]
Compute the derivatives of the loss w.r.t. the parameters.
\[
\begin{align*}
\frac{\partial \ell}{\partial w^{(l)}} &= \left[ \frac{\partial \ell}{\partial s^{(l)}} \right] (x^{(l-1)})^\top \\
\frac{\partial \ell}{\partial b^{(l)}} &= \left[ \frac{\partial \ell}{\partial s^{(l)}} \right].
\end{align*}
\]
Forward pass

Compute the activations.

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

Backward pass

Compute the derivatives of the loss w.r.t. the activations.

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

\[
\frac{\partial \ell}{\partial s^{(l)}} = \left( \frac{\partial \ell}{\partial x^{(l)}} \right) \circ \sigma'(s^{(l)})
\]

Compute the derivatives of the loss w.r.t. the parameters.

\[
\begin{bmatrix}
\frac{\partial \ell}{\partial w^{(l)}} \\
\frac{\partial \ell}{\partial b^{(l)}}
\end{bmatrix} = \begin{bmatrix}
\frac{\partial \ell}{\partial s^{(l)}} \\
\frac{\partial \ell}{\partial s^{(l)}}
\end{bmatrix} (x^{(l-1)})^\top
\]

Gradient step

Update the parameters.

\[
w^{(l)} \leftarrow w^{(l)} - \eta \left( \frac{\partial \ell}{\partial w^{(l)}} \right)
\]

\[
b^{(l)} \leftarrow b^{(l)} - \eta \left( \frac{\partial \ell}{\partial b^{(l)}} \right)
\]
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.
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.

**Without tricks, we have to keep in memory all the activations computed during the forward pass.**
Regarding computation, since the costly operation for the forward pass is

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

and for the backward

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

and

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

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