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

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

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.

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

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

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

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

\[J_{f_N \circ f_{N-1} \circ \ldots \circ f_1}(x) = J_{f_N}(f_{N-1}(\ldots (x)))) \ldots J_{f_3}(f_2(f_1(x))) J_{f_2}(f_1(x)) J_{f_1}(x)\]

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 $s(l)_i$ influences $l$ only through $x(l)_i$, we have

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

And since $x(l-1)_j$ influences $l$ only through $s(l)_i$ with

$$
s(l)_i = \sum_j w(l)_i j x(l-1)_j + b(l)_i,
$$

we have

$$
\frac{\partial l}{\partial x(l-1)_j} = \sum_i \frac{\partial l}{\partial s(l)_i} \frac{\partial s(l)_i}{\partial x(l-1)_j} = \sum_i \frac{\partial l}{\partial s(l)_i} w(l)_i j.
$$
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

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

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 \( s_i^{(l)} = \sum_j w_{i,j} x_j^{(l-1)} + b_i^{(l)} \), 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)}}.
Since \( w_i^{(l)} \) and \( b_i^{(l)} \) influences \( \ell \) only through \( s_i^{(l)} \) with

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

we have

\[
\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial w_{i,j}^{(l)}}
\]
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

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)},$$
\[
\begin{align*}
&x^{(l-1)} \xrightarrow{w^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)} \\
&\text{Since } w^{(l)}_{i,j} \text{ and } b^{(l)}_i \text{ influences } \ell \text{ only through } s_i^{(l)} \text{ with } \\
&s_i^{(l)} = \sum_j w^{(l)}_{i,j} x_{j}^{(l-1)} + b^{(l)}_i, \\
&\text{we have} \\
\frac{\partial \ell}{\partial w^{(l)}_{i,j}} &= \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial w^{(l)}_{i,j}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_{j}^{(l-1)}, \\
\frac{\partial \ell}{\partial b^{(l)}_i} &= \frac{\partial \ell}{\partial s_i^{(l)}}.
\end{align*}
\]
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

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

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

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

\[\partial \ell / \partial \mathbf{w}^{(l)}\]

\[\partial \ell / \partial \mathbf{b}^{(l)}\]
\[
\begin{bmatrix}
\frac{\partial \ell}{\partial x(l-1)}
\end{bmatrix}
\]
\[
\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})
\]
\[
\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i w_{i,j}^{(l)} \frac{\partial \ell}{\partial s_i^{(l)}}
\]
\[
\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}
\]
\[
\frac{\partial \ell}{\partial w_{i,j}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)}
\]
Forward pass

Compute the activations.

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

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

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

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

Compute the activations.

\[
x^{(0)} = x, \quad \forall l = 1, \ldots, L,
\]

\[
\begin{align*}
    s^{(l)} &= w^{(l)} x^{(l-1)} + b^{(l)} \\
    x^{(l)} &= \sigma(s^{(l)})
\end{align*}
\]

Backward pass

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

\[
\begin{aligned}
    \frac{\partial \ell}{\partial x^{(L)}} & \quad \text{from the definition of } \ell \\
    \frac{\partial \ell}{\partial x^{(l)}} & = (w^{(l+1)})^\top \left[ \frac{\partial \ell}{\partial s^{(l+1)}} \right] \\
    \left[ \frac{\partial \ell}{\partial s^{(l)}} \right] & = \left[ \frac{\partial \ell}{\partial x^{(l)}} \right] \circ \sigma'(s^{(l)})
\end{aligned}
\]

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

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

Gradient step

Update the parameters.

\[
\begin{aligned}
    w^{(l)} & \leftarrow w^{(l)} - \eta \left[ \frac{\partial \ell}{\partial w^{(l)}} \right] \\
    b^{(l)} & \leftarrow b^{(l)} - \eta \left[ \frac{\partial \ell}{\partial b^{(l)}} \right]
\end{aligned}
\]
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

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

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