

## Warm-up

Assume that the observation  $Y$  takes values in  $\{1, \dots, M\}$  and that  $X \in \mathbb{R}^d$ . The negative loglikelihood to be minimized to estimate the parameters of the model is given by:

$$\theta \mapsto \ell_n^{\text{multi}}(\theta) = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^M \mathbb{1}_{Y_i=k} \log \mathbb{P}_\theta(Y_i = k | X_i),$$

where  $\{(X_i, Y_i)\}_{1 \leq i \leq n}$  are i.i.d. observations with the same law as  $(X, Y)$ .

1. Explain the construction of  $\mathbb{P}_\theta(Y_i = k | X_i)$ ,  $1 \leq i \leq n$  for the following model. A feed forward neural network with a first hidden layer with dimension  $d_1$  and activation function  $\varphi_1$ , a second hidden layer with dimension  $d_2$  and activation function  $\varphi_2$ , and an output layer of dimension  $M$  and activation function given by the softmax function.

Let  $X_i$  be the input and define all layers as follows.

$$\begin{aligned} h_\theta^0(X_i) &= X_i, \\ z_\theta^1(X_i) &= b^1 + W^1 h_\theta^0(X_i), \quad b^1 \in \mathbb{R}^{d_1}, W^1 \in \mathbb{R}^{d_1 \times d}, \\ h_\theta^1(x) &= \varphi_1(z_\theta^1(X_i)), \\ z_\theta^2(X_i) &= b^2 + W^2 h_\theta^1(X_i), \quad b^2 \in \mathbb{R}^{d_2}, W^2 \in \mathbb{R}^{d_2 \times d_1}, \\ h_\theta^2(x) &= \varphi_2(z_\theta^2(X_i)), \\ z_\theta^3(X_i) &= b^3 + W^3 h_\theta^2(X_i), \quad b^3 \in \mathbb{R}^M, W^3 \in \mathbb{R}^{M \times d_2}, \\ h_\theta^3(X_i) &= \{\mathbb{P}_\theta(Y_i = k | X_i)\}_{1 \leq k \leq M} = \text{Softmax}(z_\theta^3(X_i)), \end{aligned}$$

2. What is the unknown parameter  $\theta$  of the previous model ? Explain how to estimate  $\theta$  with a stochastic gradient descent.

The unknown parameter is  $\theta = \{(b^j, W^j)\}_{1 \leq j \leq 3}$  is estimated iteratively. Let  $\theta_0$  be an initial estimate (randomly chosen). Then for each iteration  $p \geq 1$ , the new estimate is computed as follows

$$\theta_p = \theta_{p-1} - \gamma_p \nabla_{\theta=\theta_{p-1}} \left( -\frac{1}{B} \sum_{i=1}^B \sum_{k=1}^M \mathbb{1}_{Y_{I_i}=k} \log \mathbb{P}_\theta(Y_{I_i} = k | X_{I_i}) \right),$$

where  $B$  is the batch size,  $(\gamma_p)_{p \geq 1}$  are positive step-sizes, and  $(I_i)_{1 \leq i \leq B}$  are i.i.d. with uniform distribution on  $\{1, \dots, n\}$ . Of course, this elementary stochastic gradient descent algorithm can be improved (with for instance Adagrad, Adadelata, Rmsprop, Adam).

3. What is the complexity of an iteration of the previous algorithm ?

Using  $B$  randomly chosen observations to provide each update instead of using all observations allows to reduce the complexity (proportional to  $B$  instead of  $n$ ).

## Backpropagation

Let  $x \in \mathbb{R}^d$  be the input of a MLP with  $L$  layers and define all layers as follows.

$$\begin{aligned} h_\theta^0(x) &= x, \\ z_\theta^k(x) &= b^k + W^k h_\theta^{k-1}(x) \quad \text{for all } 1 \leq k \leq L, \\ h_\theta^k(x) &= \varphi_k(z_\theta^k(x)) \quad \text{for all } 1 \leq k \leq L, \end{aligned}$$

where  $b^1 \in \mathbb{R}^{d_1}$ ,  $W^1 \in \mathbb{R}^{d_1 \times d}$  and for all  $2 \leq k \leq L$ ,  $b^k \in \mathbb{R}^{d_k}$ ,  $W^k \in \mathbb{R}^{d_k \times d_{k-1}}$ . For all  $1 \leq k \leq L$ ,  $\varphi_k : \mathbb{R}^{d_k} \rightarrow \mathbb{R}^{d_k}$  is a nonlinear activation function. Let  $\theta = \{b^1, W^1, \dots, b^L, W^L\}$  be the unknown parameters of the MLP and

$$f_\theta(x) = h_\theta^L(x)$$

be the output layer of the MLP. As there is no modelling assumptions anymore, virtually any activation functions  $\varphi^m$ ,  $1 \leq m \leq L-1$  may be used. In this section, it is assumed that these intermediate activation functions apply elementwise and, with a minor abuse of notations, we write for all  $1 \leq m \leq L-1$  and all  $z \in \mathbb{R}^{d_m}$ ,

$$\varphi_m(z) = (\varphi_m(z_1), \dots, \varphi_m(z_{d_m})),$$

with  $\varphi_m : \mathbb{R} \rightarrow \mathbb{R}$  the selected scalar activation function.

In a classification setting, the output  $h_\theta^L(x)$  is the estimate of the probability that the class is  $k$  for all  $1 \leq k \leq M$ , given the input  $x$ . The common choice in this case is the softmax function: for all  $1 \leq i \leq M$

$$\varphi_L(z)_i = \text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^M e^{z_j}}.$$

In this case  $d_L = M$  and each component  $k$  of  $h_\theta^L(x)$  contains  $\mathbb{P}(Y = k|X)$ .

1. Prove that for all  $1 \leq i, j \leq M$ ,

$$\partial_{z_i}(\varphi_L(z))_j = \begin{cases} \text{softmax}(z)_i(1 - \text{softmax}(z)_i) & \text{if } i = j, \\ -\text{softmax}(z)_i \text{softmax}(z)_j & \text{otherwise.} \end{cases}$$

It is enough to write for all  $1 \leq j \leq M$ ,

$$\varphi_L(z)_j = \frac{e^{z_j}}{\sum_{j=1}^M e^{z_j}}.$$

Therefore,

$$\partial_{z_j}(\varphi_L(z))_j = \frac{e^{z_j} \sum_{j=1}^M e^{z_j} - e^{z_j} e^{z_j}}{\left(\sum_{\ell=1}^M e^{z_\ell}\right)^2} = \varphi_L(z)_j - \varphi_L^2(z)_j = \varphi_L(z)_j(1 - \varphi_L(z)_j).$$

The case  $i \neq j$  can be dealt with similarly.

2. Write  $\ell_\theta(X, Y) = -\sum_{k=1}^M \mathbb{1}_{Y=k} \log f_\theta(X)_k$  so that

$$\ell_n : \theta \mapsto \frac{1}{n} \sum_{i=1}^n \ell_\theta(X_i, Y_i).$$

Prove that the gradient with respect to all parameters can be computed as follows.

$$\begin{aligned} \nabla_{W^L} \ell_\theta(X, Y) &= (f_\theta(X) - \mathbb{1}_Y)(h_\theta^{L-1}(X))^\top, \\ \nabla_{b^L} \ell_\theta(X, Y) &= f_\theta(X) - \mathbb{1}_Y, \end{aligned}$$

where  $\mathbb{1}_Y$  is the vector where all entries equal to 0 except the entry with index  $Y$  which equals 1.

For all  $1 \leq j \leq M$ ,

$$\begin{aligned}\partial_{(z_\theta^L(X))_j} \ell_\theta(X, Y) &= - \sum_{k=1}^M \mathbb{1}_{Y=k} \partial_{(z_\theta^L(X))_j} \log f_\theta(X)_k, \\ &= - \sum_{k=1}^M \mathbb{1}_{Y=k} \partial_{(z_\theta^L(X))_j} \log \varphi_L(z_\theta^L(X))_k, \\ &= - \sum_{k=1}^M \mathbb{1}_{Y=k} \frac{\varphi_L(z_\theta^L(X))_j (1 - \varphi_L(z_\theta^L(X))_j) \mathbb{1}_{j=k} - \varphi_L(z_\theta^L(X))_j \varphi_L(z_\theta^L(X))_k \mathbb{1}_{j \neq k}}{\varphi_L(z_\theta^L(X))_k}, \\ &= - \sum_{k=1}^M \mathbb{1}_{Y=k} \left\{ (1 - \varphi_L(z_\theta^L(X))_k) \mathbb{1}_{j=k} - \varphi_L(z_\theta^L(X))_k \mathbb{1}_{j \neq k} \right\}.\end{aligned}$$

Therefore,

$$\nabla_{z_\theta^L(X)} \ell_\theta(X, Y) = f_\theta(X) - \mathbb{1}_Y.$$

Then, for all  $1 \leq i \leq M$  and all  $1 \leq j \leq d_{L-1}$ , by the chain rule, and using that  $z_\theta^L(X) = b^L + W^L h_\theta^{L-1}(X)$ ,

$$\begin{aligned}\partial_{W_{i,j}^L} \ell_\theta(X, Y) &= \sum_{k=1}^M \partial_{(z_\theta^L(X))_k} \ell_\theta(X, Y) \partial_{W_{i,j}^L} (z_\theta^L(X))_k, \\ &= \sum_{k=1}^M (\ell_\theta(X, Y) - \mathbb{1}_Y)_k \mathbb{1}_{i=k} (h_\theta^{L-1}(X))_j, \\ &= (f_\theta(X) - \mathbb{1}_Y)_i (h_\theta^{L-1}(X))_j.\end{aligned}$$

Therefore,

$$\nabla_{W^L} \ell_\theta(X, Y) = (f_\theta(X) - \mathbb{1}_Y) (h_\theta^{L-1}(X))^\top.$$

Similarly, for all  $1 \leq i \leq M$ , using that  $z_\theta^L(X) = b^L + W^L h_\theta^{L-1}(X)$ ,

$$\begin{aligned}\partial_{b_i^L} \ell_\theta(X, Y) &= \sum_{k=1}^M \partial_{(z_\theta^L(X))_k} \ell_\theta(X, Y) \partial_{b_i^L} (z_\theta^L(X))_k, \\ &= \sum_{k=1}^M (f_\theta(X) - \mathbb{1}_Y)_k \mathbb{1}_{i=k}, \\ &= (f_\theta(X) - \mathbb{1}_Y)_i.\end{aligned}$$

Therefore,

$$\nabla_{b^L} \ell_\theta(X, Y) = f_\theta(X) - \mathbb{1}_Y.$$

3. Prove that for all  $1 \leq m \leq L-1$ ,

$$\begin{aligned}\nabla_{W^m} \ell_\theta(X, Y) &= \nabla_{z_\theta^m(X)} \ell_\theta(X, Y) (h_\theta^{m-1}(X))^\top, \\ \nabla_{b^m} \ell_\theta(X, Y) &= \nabla_{z_\theta^m(X)} \ell_\theta(X, Y),\end{aligned}$$

where  $\nabla_{z_\theta^m(X)}$  is computed recursively as follows.

$$\begin{aligned}\nabla_{z^L(X)} \ell_\theta(X, Y) &= \ell_\theta(X, Y) - \mathbb{1}_Y, \\ \nabla_{h_\theta^m(X)} \ell_\theta(X, Y) &= (W^{m+1})^\top \nabla_{z_\theta^{m+1}(X)} \ell_\theta(X, Y), \\ \nabla_{z_\theta^m(X)} \ell_\theta(X, Y) &= \nabla_{h_\theta^m(X)} \ell_\theta(X, Y) \odot \varphi'_m(z_\theta^m(X)),\end{aligned}$$

where  $\odot$  is the elementwise multiplication.

To obtain the recursive formulation of the gradient computations, known as the back propagation of the gradient, write, for all  $1 \leq m \leq L-1$  and all  $1 \leq j \leq d_m$ , using that  $z_\theta^{m+1}(X) = b^{m+1} + W^{m+1}h_\theta^m(X)$ ,

$$\begin{aligned}\partial_{(h_\theta^m(X))_j} \ell_\theta(X, Y) &= \sum_{i=1}^{d_{m+1}} \partial_{(z_\theta^{m+1}(X))_i} \ell_\theta(X, Y) \partial_{(h_\theta^m(X))_j} (z_\theta^{m+1}(X))_i, \\ &= \sum_{i=1}^{d_{m+1}} \partial_{(z_\theta^{m+1}(X))_i} \ell_\theta(X, Y) W_{i,j}^{m+1}.\end{aligned}$$

Therefore,

$$\nabla_{h_\theta^m(X)} \ell_\theta(X, Y) = (W^{m+1})^\top \nabla_{z_\theta^{m+1}(X)} \ell_\theta(X, Y).$$

Then, for all  $1 \leq m \leq L-1$  and all  $1 \leq j \leq d_m$ , using that  $h_\theta^m(X)_j = \varphi_m(z_\theta^m(X)_j)$ ,

$$\begin{aligned}\partial_{(z_\theta^m(X))_j} \ell_\theta(X, Y) &= \sum_{i=1}^{d_m} \partial_{(h_\theta^m(X))_i} \ell_\theta(X, Y) \partial_{(z_\theta^m(X))_j} (h_\theta^m(X))_i, \\ &= \sum_{i=1}^{d_m} \partial_{(h_\theta^m(X))_i} \ell_\theta(X, Y) \mathbb{1}_{i=j} \varphi'_m(z_\theta^m(X)_i), \\ &= \partial_{(h_\theta^m(X))_j} \ell_\theta(X, Y) \varphi'_m(z_\theta^m(X)_j).\end{aligned}$$

Therefore,

$$\nabla_{z_\theta^m(X)} \ell_\theta(X, Y) = \nabla_{h_\theta^m(X)} \ell_\theta(X, Y) \odot \varphi'_m(z_\theta^m(X)).$$

Then, for all  $1 \leq i \leq d_m$  and all  $1 \leq j \leq d_{m-1}$ , and using that  $z_\theta^m(X) = b^m + W^m h_\theta^{m-1}(X)$ ,

$$\begin{aligned}\partial_{W_{i,j}^m} \ell_\theta(X, Y) &= \sum_{k=1}^{d_m} \partial_{(z_\theta^m(X))_k} \ell_\theta(X, Y) \partial_{W_{i,j}^m} (z_\theta^m(X))_k, \\ &= \sum_{k=1}^{d_m} \partial_{(z_\theta^m(X))_k} \ell_\theta(X, Y) \mathbb{1}_{i=k} (h_\theta^{m-1}(X))_j, \\ &= \partial_{(z_\theta^m(X))_i} \ell_\theta(X, Y) (h_\theta^{m-1}(X))_j.\end{aligned}$$

Therefore,

$$\nabla_{W^m} \ell_\theta(X, Y) = \nabla_{z_\theta^m(X)} \ell_\theta(X, Y) (h_\theta^{m-1}(X))^\top.$$

Similarly, for all  $1 \leq i \leq d_m$ , using that  $z_\theta^m(X) = b^m + W^m h_\theta^{m-1}(X)$ ,

$$\begin{aligned}\partial_{b_i^m} \ell_\theta(X, Y) &= \sum_{k=1}^{d_m} \partial_{(z_\theta^m(X))_k} \ell_\theta(X, Y) \partial_{b_i^m} (z_\theta^m(X))_k, \\ &= \sum_{k=1}^{d_m} \partial_{(z_\theta^m(X))_k} \ell_\theta(X, Y) \mathbb{1}_{i=k}, \\ &= \partial_{(z_\theta^m(X))_i} \ell_\theta(X, Y).\end{aligned}$$

Therefore,

$$\nabla_{b^m} \ell_\theta(X, Y) = \nabla_{z_\theta^m(X)} \ell_\theta(X, Y).$$