FEED FORWARD NEURAL NETWORKS

1 Warm-up

Assume that the observation Y takes values in $\{1, ..., M\}$ 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 $\{(X_i, Y_i)\}_{1 \le i \le n}$ 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 φ_1 , a second hidden layer with dimension d_2 and activation function φ_2 , and an output layer of dimension M and activation function given by the softmax function.
- 2. What is the unknown parameter θ of the previous model? Explain how to estimate θ with a stochastic gradient descent.
- 3. What is the complexity of an iteration of the previous algorithm?

2 Backpropagation

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

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

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 $1 \leq k \leq L$, $\varphi_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ 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 modeling assumptions anymore, virtually any activation functions φ^m , $1 \leqslant m \leqslant L-1$ 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 $1 \leqslant m \leqslant L-1$ and all $z \in \mathbb{R}^{d_m}$,

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

with $\varphi_m: \mathbb{R} \to \mathbb{R}$ 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 $1 \le k \le M$, given the input x. The common choice in this case is the softmax function: for all $1 \le k \le M$

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

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 $1 \leq i, j \leq M$,

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

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.

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

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

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

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

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

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

$$\begin{split} & \nabla_{z^L(X)} \ell_{\theta}(X,Y) = \ell_{\theta}(X,Y) - \mathbbm{1}_Y \,, \\ & \nabla_{h^m_{\theta}(X)} \ell_{\theta}(X,Y) = (W^{m+1})^\top \nabla_{z^{m+1}_{\theta}(X)} \ell_{\theta}(X,Y) \,, \\ & \nabla_{z^m_{\theta}(X)} \ell_{\theta}(X,Y) = \nabla_{h^m_{\theta}(X)} \ell_{\theta}(X,Y) \odot \varphi'_m(z^m_{\theta}(X)) \,, \end{split}$$

where \odot is the elementwise multiplication.