TP4

Thibault Duhamel 18026048, Heng Shi 18171434

April 12, 2019

Question 1

The cross-entropy function is given by:

$$E_D(W) = -\sum_n \sum_k t_{kn} ln(\frac{e^{a_k}}{\sum_c e^{a_c}})$$

In order to calculate its derivative with respect to a_i, we first need to split the sums and isolate a_i, when $k = i$ or $c = i$:

$$E_D(W) = -\sum_n \sum_{k \neq i} t_{kn} ln(\frac{e^{a_k}}{\sum_{c \neq i} e^{a_c} + e^{a_i}}) - \sum_n t_{in} ln(\frac{e^{a_i}}{\sum_{c \neq i} e^{a_c} + e^{a_i}})$$

Now, it is possible to compute the gradient of the above expression:

$$\frac{dE_D}{da_i} = \sum_n \sum_{k \neq i} t_{kn} \frac{e^{a_i}}{\sum_{c \neq i} e^{a_c} + e^{a_i}} - \sum_n t_{in}(1 - \frac{e^{a_i}}{\sum_{c \neq i} e^{a_c} + e^{a_i}})$$

The sums of indices c can be reassembled, for $c = i$:

$$\frac{dE_D}{da_i} = \sum_n \sum_{k \neq i} t_{kn} \frac{e^{a_i}}{\sum_c e^{a_c}} - \sum_n t_{in}(1 - \frac{e^{a_i}}{\sum_c e^{a_c}})$$

$$\frac{dE_D}{da_i} = \sum_n \sum_{k \neq i} t_{kn} \frac{e^{a_i}}{\sum_c e^{a_c}} - \sum_n t_{in} + \sum_n t_{in} \frac{e^{a_i}}{\sum_c e^{a_c}}$$

The sum depending on k can be reassembled too, for $k = i$:

$$\frac{dE_D}{da_i} = \sum_n \sum_k t_{kn} \frac{e^{a_i}}{\sum_c e^{a_c}} - \sum_n t_{in}$$

$$\frac{dE_D}{da_i} = \sum_n \sum_k t_{kn} \frac{e^{a_i}}{\sum_c e^{a_c}} - t_{in}$$

As the target t_n is in a one-hot vector, its sum over every classes is 1:

$$\frac{dE_D}{da_i} = \sum_n \frac{e^{a_i}}{\sum_c e^{a_c}} - t_{in}$$

Which is, eventually:

$$\boxed{\frac{dE_D}{da_i} = \sum_n y_{w_i}(x_n) - t_{in}}$$

Question 2

As $k(x, x')$ is a valid kernel, there exists a function ϕ, such as:

$$k(x, x') = \phi(x)^T.\phi(x')$$

To prove that $exp(k(x, x'))$ is valid, we use a Taylor development around zero:

$$exp(x) = e^0 + e^0 x + \frac{e^0}{2!}x^2 + ... = 1 + x + \frac{x^2}{2!} + ...$$

$$exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

$$exp(k(x, x')) = \sum_{i=0}^{\infty} \frac{k(x, x')^i}{i!}$$

As any linear combination and product of two valid kernels are also valid, $exp(k(x, x'))$ is valid too.

Question 3

The momentum gradient descent is based on a speed vector v, such that:

$$v_{t+1} = \rho v_t - \nabla E$$

$$w_{t+1} = w_t - \eta v_{t+1}$$

Let us prove that there indeed exists an link with physics. To do so, we start with Newton's second law of motion, and write:

$$ma = f$$

where a and f can either be scalars or vectors of same dimension. a is the acceleration ($m.s^{-2}$) of an object and f is a force ($kg.m.s^{-2}$) applied to this object. We also now, by definition, that:

$$a = \frac{dv}{dt}$$

$$v = \frac{dp}{dt}$$

where v is the speed ($m.s^{-1}$) and p is the position (m).
In a discrete context of time steps $\{t_i\}$, we can write:

$$a_{t+1} = v_{t+1} - v_t$$

$$v_{t+1} = p_{t+1} - p_t$$

Using this expression of a in Newton's second law of motion:

$$v_{t+1} = v_t + \frac{1}{m}f$$

$$p_{t+1} = p_t + v_{t+1}$$

This is exactly the momentum expression, with $p = w$, $\rho = 1$, $\eta = -1$ and $\nabla E = -\frac{1}{m}f$. Hence, according to this comparison, the gradient of a loss can be seen as a force "pushing" the parameters in a direction. The higher this gradient is, the quicker the parameters will "move".