APPENDIX : PROF. FRANCESCA A.N. PALMIERI UNIV. DELLA CAMPANIA Nov. 2023 ASF. 1

THE SOFTMAX FUNCTION

Given a real N-dimensional column vector $\mathbf{x}$, a Softmax function is defined as

$$S(\mathbf{x}) = \frac{e^{\mathbf{x}}}{\sum_{i=1}^{N} e^{x_i}},$$

where the notation $e^{\mathbf{x}}$ denotes the vector $[e^{x_1} e^{x_2} ... e^{x_N}]^T$. More compactly

$$S(\mathbf{x}) = \frac{e^{\mathbf{x}}}{\mathbf{1}^T e^{\mathbf{x}}},$$

where $\mathbf{1}$ is a vector with all ones.

B.1 Property 1 : SHIFT INVARIANCE

$$S(\mathbf{x} + c\mathbf{1}) = S(\mathbf{x}), \quad \text{for any constant } c. \tag{34}$$

Proof: immediate because $\frac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \frac{e^{x_i}}{\sum_j e^{x_j}}$.

B.2 Property 2 : THE INVERSE

Given $\mathbf{p} = S(\mathbf{x})$, the inverse function is

$$\mathbf{x} = \log \mathbf{p} + c\mathbf{1}, \tag{35}$$

where c is an undetermined constant.

This property means that the softmax function is invertible, except for a constant that cannot be determined from $\mathbf{p}$. Constraints on $\mathbf{x}$ may help to obtain the constant. Note that this is also a consequence of Property 1, i.e. any constant added to the input is definitely lost.

B.3 Property 3 ; SUPERPOSITION AND SCALING

$$S(\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + ... + \alpha_N \mathbf{x}_N + c\mathbf{1}) \propto S(\mathbf{x}_1)^{\alpha_1} \odot S(\mathbf{x}_2)^{\alpha_2} \odot \cdot \odot S(\mathbf{x}_N)^{\alpha_N}, \tag{36}$$

where α_1, α_2, ...α_N are any real numbers.

Proof: From Property B.1 the constant term is irrelevant. Observing that $\frac{e^{\alpha x_i}}{\sum_{j=1}^{N} e^{\alpha x_j}} \propto (e^{x_i})^{\alpha}$ and that $\frac{e^{x_i+y_i}}{\sum_j e^{x_j+y_j}} \propto e^{x_j} e^{y_j}$, we have the result.

Note that product in the right hand side of (36) is a distribution, ~~but it must be~~ and may need to be normalized. More specifically, Property 2 implies that the sum becomes

$$S(\mathbf{x}_1 + \mathbf{x}_2) \propto S(\mathbf{x}_1) \odot S(\mathbf{x}_1). \tag{37}$$

Similarly a difference becomes

$$S(\mathbf{x}_1 - \mathbf{x}_2) \propto S(\mathbf{x}_1)./S(\mathbf{x}_1), \tag{38}$$

where ./ denotes element-by-element division. Clearly the values that appear at the denominator must be non null. Scaling the input gives

$$S(\alpha\mathbf{x}) \propto S(\mathbf{x})^{\alpha}, \tag{39}$$

(see Property 5)

which represent a sharper or smother distribution with respect to $S(\mathbf{x})$ if $\alpha > 1$ or $0 < \alpha < 1$ respectively. Changing the sign to the input gives

$$S(-\mathbf{x}) \propto 1./S(\mathbf{x}), \tag{40}$$

where obviously the numbers at the denominator must be non null.

~~properties of the derivatives~~

PROPERTY 4: THE JACOBIAN [12]

The softmax is a vector function and every output depends on all inputs, because of the denominator. Therefore we must consider all the derivatives $\frac{\partial y_i}{\partial z_l}$ $\forall i,l = 1,\dots,M$, i.e. the Jacobian

$$\underline{\underline{J}} = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} & \cdots & \frac{\partial y_1}{\partial z_M} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} & \cdots & \frac{\partial y_2}{\partial z_2} \\ \frac{\partial y_M}{\partial z_1} & \frac{\partial y_M}{\partial z_2} & \cdots & \frac{\partial y_M}{\partial z_M} \end{bmatrix}$$

The Jacobian for the softmax function is

$$\underline{\underline{J}} = \begin{bmatrix} y_1(1-y_1) & -y_1y_2 & \cdots & -y_1y_M \\ -y_2y_1 & y_2(1-y_2) & \cdots & -y_2y_M \\ & & & \\ -y_My_1 & -y_My_2 & \cdots & y_M(1-y_M) \end{bmatrix} = \mathrm{diag}(\underline{y}) - \underline{y}\,\underline{y}^T$$

or for a generic element

$$\frac{\partial y_i}{\partial z_l} = \delta_{il} y_i - y_i y_l$$

(δ_{il} is the Kronecker delta $= 0$ if $i \neq l$ and $= 1$ for $i = l$)

PROOF :

$$\frac{\partial y_i}{\partial z_k} = \frac{\partial}{\partial z_k}\left(\frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}}\right)$$

For $k = i$

$$\frac{\partial y_i}{\partial z_i} = \frac{\partial}{\partial z_i}\frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}} = \frac{e^{z_i}\left(\sum_{j=1}^{M} e^{z_j}\right) - e^{z_i} e^{z_i}}{\left(\sum_{j=1}^{M} e^{z_j}\right)^2} = \frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}} - \left(\frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}}\right)^2$$

$$= y_i(1 - y_i)$$

For $k \neq i$

$$\frac{\partial y_i}{\partial z_k} = \frac{\partial}{\partial z_k}\left(\frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}}\right) = \frac{-e^{z_i} e^{z_k}}{\left(\sum_{j=1}^{M} e^{z_j}\right)^2} = -\frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}} \cdot \frac{e^{z_k}}{\sum_{j=1}^{M} e^{z_j}}$$

$$= -y_i y_k$$

THAT DO MAX BETWEEN

PROPERTY 5 Sharpness of the softmax.

From the scaling property, we have

$$S(\alpha \underline{x}) \propto S(\underline{x})^{\alpha}$$

which means that we can control the sharpness of the output by changing α. More specifically for $\alpha \to \infty$ $\underline{y} = S(\alpha \underline{x})$ tends to a delta on the max value.

Property (5a)

Proof. Without loss of generality, assume that $x_1 x_2 \dots x_N$ are ranked in descending order, $x_1 \geq x_2 \geq \dots \geq x_N$ so that x_1 is the maximum. The ith output can be written as

$$\frac{e^{\alpha x_i}}{\sum_{j=1}^{N} e^{\alpha x_j}} = \frac{e^{\alpha x_i}}{e^{\alpha x_1}\left(1 + e^{\alpha(x_2 - x_1)} + \dots + e^{\alpha(x_N - x_1)}\right)} =$$

$$= \frac{e^{\alpha(x_i - x_1)}}{1 + e^{\alpha(x_2 - x_1)} + \dots + e^{\alpha(x_N - x_1)}} \qquad (*)$$

For $i \neq 1$ all the exponentials go to zero because $x_i - x_1 < 0$ and $y_i \to 0 \quad \forall i \neq 1$

For $i = 1$ instead we have $y_1 = \frac{1}{1 + e^{\alpha(x_2 - x_1)} + \dots + e^{\alpha(x_N - x_1)}} \to 1$ ☒

Property 5(b)

For $\alpha \to 0^+$ $S(\alpha \underline{x}) \to$ Uniform distribution. $\left(y_i = \frac{1}{N} \ \forall i\right)$

Proof

For $\alpha \to 0^+$ all the exponentials tend to zero

and $\frac{e^{\alpha x_i}}{\sum_{j=1}^{N} e^{\alpha x_j}} \to \frac{1}{N}$ ☒

A corollary of 5(a) is that if we have multiple maximums, i.e.

$$x_1 = x_2 = \dots = x_r > x_{r+1} > \dots > x_N$$

$$y_1, y_2, \dots y_r \to \frac{1}{r} \qquad y_{r+1}, \dots y_N \to 0$$

Proof: Immediate from (*)

Property (5c) (dual to 5(a))

Where $\alpha \to -\infty$ $S(\underline{x}) \to$ Delta concentrated on the minimum

Proof: Assume that the inputs are ranked

$$x_1 > x_2 > \dots > x_N$$

x_N is the minimum. Now the ith output can be written as

$$y_i = \frac{e^{\alpha x_i}}{e^{\alpha x_N}\left(e^{\alpha(x_1 - x_N)} + e^{\alpha(x_2 - x_N)} + \dots + e^{\alpha(x_{N-1} - x_N)} + 1\right)}$$

$$= \frac{e^{\alpha(x_i - x_N)}}{e^{\alpha(x_1 - x_N)} + e^{\alpha(x_2 - x_N)} + \dots + e^{\alpha(x_{N-1} - x_N)} + 1}$$

All the exponentials go to zero for $i \neq N$ because $x_i - x_N > 0$ and α negative.

Therefore for $i = N \to y_i \to 1$

Observation:

The sign of α can be used to make the softmax a softmin.

A NUMERICALLY STABLE SOFTMAX:

Since

$$S(z+c) = S(z)$$

and the outputs sum to one, the system is "overparametrized". This is because the inputs are unconstrained and may be arbitrarily biased to have small or large values.
A numerically more stable version (Restricted) for the softmax function is often used by fixing one of the inputs to a given value, for example $z_i = 0$. The input fixed to zero may also be different at every sample taking

$$S(z - \max_i z_i)$$

- In this way all exponents in

$$\frac{e^{z_j - \max_i z_i}}{\sum_{k=1}^{N} e^{z_k - \max_i z_i}}$$

are negative and therefore the exponentials are in $[0,1]$. This prevents overflow (still prone to underflow)

- At least one of the exponents is 0, and therefore at least one of the exponentials is one. This means that at least one value does not underflow. Furthermore the denominator will be ≥ 1 preventing division by zero. We have at least one non-zero numerator avoid a all-zero output caused by numerical issues.

The restricted version of the softmax rarely has an effect on learning.

The softmax for 2 outputs.

Note that the logistic function can be viewed as equivalent to a two-class classifier with the softmax at the output

$$X \rightarrow \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \rightarrow \begin{bmatrix} \text{SOFTMAX} \end{bmatrix} \begin{matrix} \rightarrow y_1 = p(a_1|x) = \dfrac{e^{z_1}}{e^{z_1}+e^{z_2}} = \dfrac{1}{1+e^{z_2-z_1}} \\ \rightarrow y_2 = p(a_2|x) = \dfrac{e^{z_2}}{e^{z_1}+e^{z_2}} = \dfrac{1}{1+e^{z_1-z_2}} \end{matrix}$$

$$\rightarrow \begin{bmatrix} z_1 \\ - \\ z_2 \end{bmatrix} \rightarrow z_1 - z_2 \rightarrow \boxed{\text{LOGISTIC}} \rightarrow y = p(a_1|x)$$