Softmax Function has a numerical problem: e^{z_i} could be very big

$$\hat{y} = \begin{bmatrix} \frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}} \end{bmatrix}$$

Let's change z_i

$$\boxed{\begin{array}{l} z_i \leftarrow z_i + u \\ \text{for every } i \end{array}}$$

Then, each element of $\hat{y}$ will change or not ?

$$\frac{e^{z_i+u}}{\sum_{k=1}^{K} e^{z_k+u}} = \frac{e^{z_i}\, e^{u}}{\sum_{k=1}^{K} e^{z_k} e^{u}} = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

Thus, each element of $\hat{y}$ is NOT changed after $z_i \leftarrow z_i + u$ for every i

e^{z_i} could be very big e^{100}, it will cause numerical problem (overflow)

Solution: $u = -\max\{z_1, \dots, z_K\}$

e.g. $e^{100} and\ u = -99, then\ e^{100+u} = e^{1}$

Softmax function: a uniqueness problem

$$\hat{y} = \begin{bmatrix} \frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}} \end{bmatrix}$$

$$z_i = w_i^T x + b_i$$

Let's change w_i

$$\boxed{w_i \leftarrow w_i + v}$$

$$\Downarrow$$

$$\boxed{\begin{matrix} z_i \leftarrow z_i + u \\ u = v^T x \end{matrix}}$$

for every i

Then, each element of $\hat{y}$ will change or not ?

$$\frac{e^{z_i+u}}{\sum_{k=1}^{K} e^{z_k+u}} = \frac{e^{z_i}\, e^{u}}{\sum_{k=1}^{K} e^{z_k} e^{u}} = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

Thus, each element of $\hat{y}$ is NOT changed after $w_i \leftarrow w_i + v$ for every i

So, if w_i is an optimal parameter, then $w_i + v$ is also an optimal parameter.

The optimal parameters are not unique…
(the same conclusion for b_i)

Softmax function: a uniqueness problem

The optimal parameters are not unique…

Solution: "remove" z_1, w_1 and b_1

$$\hat{y} = \begin{bmatrix} \frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sum_{k=1}^{K} e^{z_k - z_1}} \\ \vdots \\ \frac{e^{z_i - z_1}}{\sum_{k=1}^{K} e^{z_k - z_1}} \\ \vdots \\ \frac{e^{z_K - z_1}}{\sum_{k=1}^{K} e^{z_k - z_1}} \end{bmatrix}$$

rename $z_k - z_1$ as z_k for $k > 1$