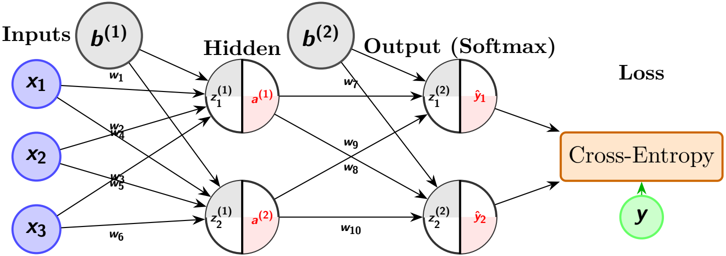
Gradient of Softmax + Cross-Entropy w.r.t Logits

Bindeshwar Singh Kushwaha
PostNetwork Academy

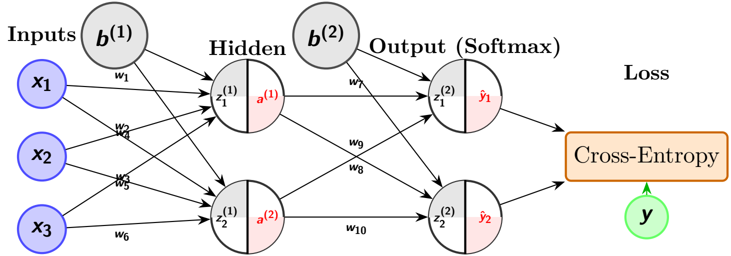
Goal: Compute $\frac{\partial L}{\partial z_j}$. **Notation:**

- Logits: $z = [z_1, z_2, \ldots, z_C]$
- Softmax: $\hat{y}_i = \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}}$
- Cross-Entropy Loss: $L = -\sum_{i=1}^{C} y_i \log \hat{y}_i$, where y_i is one-hot.



$$\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

Explanation: From $\frac{d}{dx}\log x = 1/x$ and negative sign.



Softmax derivative is a Jacobian matrix

Softmax:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}}$$

Observation: The derivative of $\hat{\mathbf{y}}$ w.r.t $\mathbf{z}$ is not a simple vector, but a **matrix** (Jacobian):

$$\frac{\partial \hat{y}}{\partial z} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial z_1} & \frac{\partial \hat{y}_1}{\partial z_2} & \cdots & \frac{\partial \hat{y}_1}{\partial z_C} \\ \frac{\partial \hat{y}_2}{\partial z_1} & \frac{\partial \hat{y}_2}{\partial z_2} & \cdots & \frac{\partial \hat{y}_2}{\partial z_C} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_C}{\partial z_1} & \frac{\partial \hat{y}_C}{\partial z_2} & \cdots & \frac{\partial \hat{y}_C}{\partial z_C} \end{bmatrix}$$

Interpretation:

- Diagonal elements ($i = j$) show how $\hat{y}_i$ changes with its own logit.
- Off-diagonal elements ($i \neq j$) show how $\hat{y}_i$ changes with other logits.

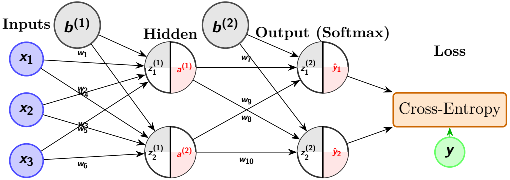
Recall Softmax:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}}$$

Case 1: $i = j$

$$\frac{\partial \hat{y}_i}{\partial z_i} = \frac{(\sum_{k=1}^{C} e^{z_k})e^{z_i} - e^{z_i} \cdot e^{z_i}}{(\sum_{k=1}^{C} e^{z_k})^2}$$

$$= \frac{e^{z_i}}{\sum_k e^{z_k}} \left(1 - \frac{e^{z_i}}{\sum_k e^{z_k}}\right) = \hat{y}_i(1 - \hat{y}_i)$$

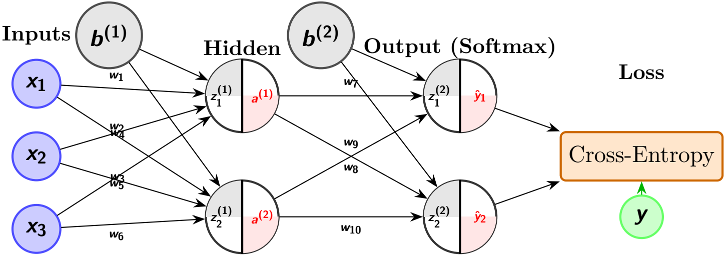


Softmax derivative w.r.t logits (i≠j)

Case 2: $i \neq j$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{0 \cdot (\sum_k e^{z_k}) - e^{z_i} \cdot e^{z_j}}{(\sum_k e^{z_k})^2} = -\frac{e^{z_i}}{\sum_k e^{z_k}} \frac{e^{z_j}}{\sum_k e^{z_k}} = -\hat{y}_i \hat{y}_j$$

Interpretation: Increasing z_j decreases $\hat{y}_i$ for $i \neq j$ because the sum in the denominator increases.

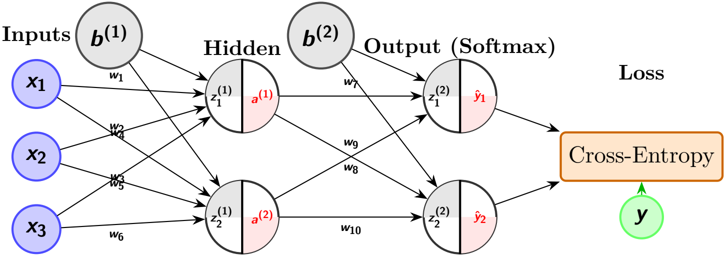


Compact formula:

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1-\hat{y}_i), & i = j \\ -\hat{y}_i\hat{y}_j, & i \neq j \end{cases}$$

Summary:

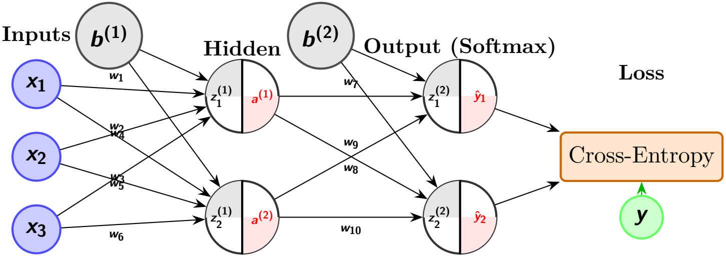
- The diagonal terms ($i = j$) are positive, representing self-influence.
- The off-diagonal terms ($i \neq j$) are negative, representing the competition between classes.
- This is the Jacobian of the softmax function.



Chain rule

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{C} \frac{\partial L}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_j} = \sum_{i=1}^{C} \left(-\frac{y_i}{\hat{y}_i}\right) \frac{\partial \hat{y}_i}{\partial z_j}$$

$$i = j : -y_j(1 - \hat{y}_j), \quad i \neq j : \sum_{i \neq j} y_i \hat{y}_j$$



Gradient of the Loss with respect to z_j

We start from:

$$\frac{\partial L}{\partial z_j} = -y_j(1 - \hat{y}_j) + \sum_{i \neq j} y_i \hat{y}_j$$

Step 1 (True class j):

$$-y_j(1 - \hat{y}_j)$$

Step 2 (Other classes $i \neq j$):

$$\sum_{i \neq j} y_i \hat{y}_j$$

Since only one $y_i = 1$ (true class), all others are $\mathbf{0}$.

Simplification of the Gradient

Step 3: Replace the sum. Since $\sum_{i \neq j} y_i = 1 - y_j$, we rewrite:

$$\frac{\partial L}{\partial z_j} = -y_j(1 - \hat{y}_j) + (1 - y_j)\hat{y}_j$$

Step 4: Expand the terms.

$$= -y_j + y_j\hat{y}_j + \hat{y}_j - y_j\hat{y}_j$$

Step 5: Cancel common terms.

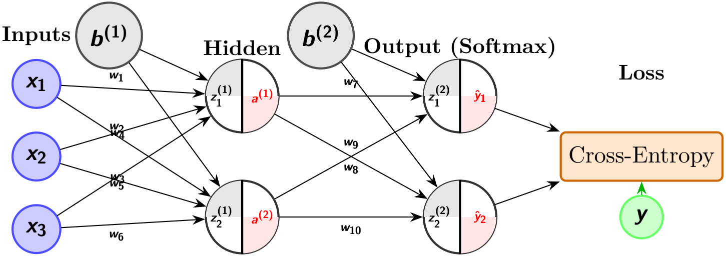
$$= \hat{y}_j - y_j$$

Interpretation: The gradient is just the difference between predicted probability $\hat{y}_j$ and true label y_j.

Combine terms

$$\frac{\partial L}{\partial z_j} = -y_j(1 - \hat{y}_j) + \sum_{i \neq j} y_i \hat{y}_j$$

$$\implies \frac{\partial L}{\partial z_j} = \hat{y}_j - y_j$$

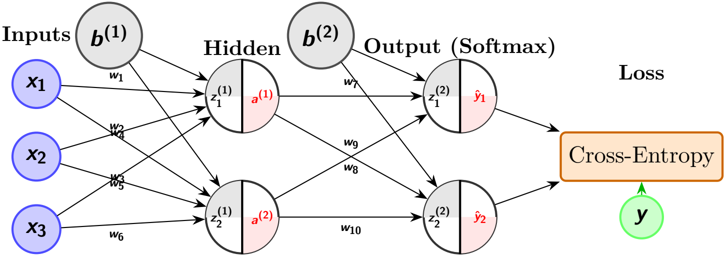


Conclusion

$$\frac{\partial L}{\partial z} = \hat{y} - y$$

Summary:

- Gradient is predicted minus target.
- No need for full Jacobian.
- Efficient for classification tasks.



Reach PostNetwork Academy

Website

www.postnetwork.co

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

GitHub Repositories

www.github.com/postnetworkacademy

Thank You!