Ho Chi Minh City → Da Nang 27/01/2025

Now we have found expressions for $\frac{\partial}{\partial \theta_j}[C(\underline{\theta})]$ for each $\theta_j \in \underline{\theta}$.

To do next time: ① write iterative equation for gradient descent
② try the same exercise for a 2-hidden-layer network

① Iterative eqn for gradient descent (0-hidden-layer network).

For each $\theta_j \in \underline{\theta} = \{b^{(1)}\} \cup \{w^{(1)}_{i,j} : i \in [0,783] \cap \mathbb{Z},\ j \in [0,9] \cap \mathbb{Z}\}$,

- Initialise $\theta_j^{[0]} :=$ random (between -10 & 10 idk?)
- Step $\theta^{[n+1]} := \theta^{[n]} - \nu \frac{\partial}{\partial \theta_j}[C(\underline{\theta})]$ where $\nu > 0$ is "learning rate".

Can code logic for Step to repeat until we have hit some max number of iterations OR difference between this step vs previous is negligible.

② 2-hidden-layer network

Let L_i denote the number of neurons in layer i. (minus 1, due to 0 indexing) We want this:
- $L_0 = 783$ (since 784 pixels in MNIST images)
- $L_1 = 127$
- $L_2 = 63$
- $L_3 = 9$ (10 digits in final layer)

Otherwise, structure is identical to the 0-hidden-layer network I've been working on. So, in particular:
- Each non-zero layer has a bias parameter $b^{(i)}$ added to the weighted sums evaluated in that layer (i.e. $b^{(1)}, b^{(2)}, b^{(3)}$)
- Each possible pairwise connection between layer $i-1$ and the next layer has a weight parameter associated (i.e. for layer i, $W^{(i)} = \{w^{(i)}_{j,j} : i \in [0, L_i] \cap \mathbb{Z},\ j \in [0, L_{i+1}] \cap \mathbb{Z}\}$. $i \in \{1,2,3\}$
- ReLU activation used: $f(x): \mathbb{R} \to \mathbb{R}$ where $f(x) := \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$
- Cross-entropy loss used for training:

$$C(\underline{\theta}) = \frac{-1}{K} \sum_{k=0}^{K-1} \sum_{j=0}^{9} t_{j,k} \log(\hat{p}_{j,k})$$ as previously defined!

Da Nang → Hanoi 31/01/2025

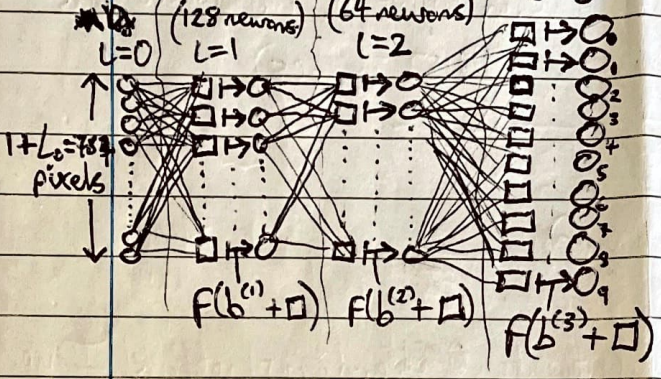
Before differentiating $C(\underline{\theta})$ by all these new parameters:

$$\underline{\theta} = \{b^{(1)}, b^{(2)}, b^{(3)}\} \cup W^{(1)} \cup W^{(2)} \cup W^{(3)}$$

Let us write clearly how neuron values in $C(\underline{\theta})$ are defined...

- $p^{(0)}_{n,k} :=$ n^{th} pixel value in training image k. (valid for $n \in \{0, 1, \dots, 783\}$)
- $p^{(1)}_{n,k} = f\left(b^{(1)} + \sum_{i=0}^{783} w^{(1)}_{i,n}\, p^{(0)}_{i,k}\right)$ for $n \in \{0, 1, \dots, L_1 = 127\}$
- $p^{(2)}_{n,k} = f\left(b^{(2)} + \sum_{i=0}^{127} w^{(2)}_{i,n}\, p^{(1)}_{i,k}\right)$ for $n \in \{0, 1, \dots, L_2 = 63\}$
- $p^{(3)}_{n,k} = f\left(b^{(3)} + \sum_{i=0}^{63} w^{(3)}_{i,n}\, p^{(2)}_{i,k}\right)$ for $n \in \{0, 1, \dots, L_3 = 9\}$
- Then, $\hat{p}_{n,k} = \dfrac{\exp\{p^{(3)}_{n,k}\}}{\sum_{z=0}^{9} \exp\{p^{(3)}_{z,k}\}}$.

So, in general we have $p^{(l)}_{n,k} = f\left(b^{(l)} + \sum_{i=0}^{L_{l-1}} w^{(l)}_{i,n}\, p^{(l-1)}_{i,k}\right)$ $\forall l \in \{0,1,2,3\}$



then the digit from 0–9 with the highest neuron-value in this final layer is our network's prediction.

(diagram, where □ represents each weighted sum of neurons)

Now we have everything we need, on the next sheet I'll find $\frac{\partial}{\partial \theta_j}[C(\underline{\theta})]$ for all $\theta_j \in \underline{\theta}$...
(start with biases, then do weights. Work from layer 3 to layer 1).
I will use this fact in each case to kick things off:

$$\frac{\partial}{\partial \theta_j}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{9} t_{m,k}\,\frac{1}{\hat{p}_{m,k}}\,\frac{\partial}{\partial \theta_j}\left[\hat{p}_{m,k}\right].$$

$$\frac{\partial}{\partial b^{(3)}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{j=0}^{q} t_{j,k}\frac{\partial}{\partial b^{(3)}}\left[\log(\hat{p}_{j,k})\right]$$

$b^{(3)}$

$$\frac{\partial}{\partial b^{(3)}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q}\left(\frac{t_{m,k}}{\hat{p}_{m,k}}\frac{\partial}{\partial b^{(3)}}\left[C(\underline{\theta})\right]\frac{\partial}{\partial b^{(3)}}\left[\hat{p}_{m,k}\right]\right)$$

Now, $\frac{\partial}{\partial b^{(3)}}\left[\exp\{p^{(3)}_{m,k}\}\right] =$ chain rule & $\frac{\partial}{\partial x}(e^x) = e^x$

$$= \exp\{p^{(3)}_{m,k}\}\frac{\partial}{\partial b^{(3)}}\left[p^{(3)}_{m,k}\right]$$

$$= \exp\{p^{(3)}_{m,k}\}\frac{\partial}{\partial x}\left[f(x)\right]\frac{\partial}{\partial b^{(3)}}\left[b^{(3)} + \sum_{i=0}^{63} w^{(3)}_{i,m}\rho^{(2)}_{i,k}\right]$$

$$= \exp\{p^{(3)}_{m,k}\}\mathbb{I}\{\Sigma^{(3)}_{m,k} > 0\}.$$

using our old friend defined earlier $\Sigma^{(l)}_{j,k}$!= "sum of l^{th} bias + weighted sum going into j^{th} neuron in layer l"

and the fact $f'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$

So, using the quotient rule in essentially the same way we did for the $b^{(1)}$ partial derivative in my 0-hidden-layer network:

$$\hat{p}_{m,k} = \frac{\exp\{p^{(3)}_{m,k}\}}{\sum_{z=0}^{q}\exp\{p^{(3)}_{z,k}\}} = \frac{u}{v}$$

$$u' = \exp\{p^{(3)}_{m,k}\}\mathbb{I}\{\Sigma^{(3)}_{m,k} > 0\}.$$

$$v' = \sum_{z=0}^{q}\exp\{p^{(3)}_{z,k}\}\mathbb{I}\{\Sigma^{(3)}_{z,k} > 0\}.$$

(derived earlier so copied here!)

$$\Rightarrow \frac{\partial}{\partial b^{(3)}}\left[\hat{p}_{m,k}\right] = \hat{p}_{m,k}\left(\mathbb{I}\{\Sigma^{(3)}_{m,k} > 0\} - \sum_{z=0}^{q}\hat{p}_{z,k}\mathbb{I}\{\Sigma^{(3)}_{z,k} > 0\}\right)$$

$$\Rightarrow \frac{\partial}{\partial b^{(3)}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q} t_{m,k}\left(\mathbb{I}\{\Sigma^{(3)}_{m,k} > 0\} - \sum_{z=0}^{q}\hat{p}_{z,k}\mathbb{I}\{\Sigma^{(3)}_{z,k} > 0\}\right)$$

$b^{(2)}$

$$\frac{\partial}{\partial b^{(2)}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q}\left(\frac{t_{m,k}}{\hat{p}_{m,k}}\frac{\partial}{\partial b^{(2)}}\left[\hat{p}_{m,k}\right]\right).$$

Now, $\frac{\partial}{\partial b^{(2)}}\left[\exp\{p^{(3)}_{m,k}\}\right] = \exp\{p^{(3)}_{m,k}\}\frac{\partial}{\partial b^{(2)}}\left[p^{(3)}_{m,k}\right]$

$$= \exp\{p^{(3)}_{m,k}\}\mathbb{I}\{\Sigma^{(3)}_{m,k} > 0\}\frac{\partial}{\partial b^{(2)}}\left[b^{(3)} + \sum_{i=0}^{63} w^{(3)}_{i,m}\rho^{(2)}_{i,k}\right]$$

the $\mathbb{I}\{\Sigma^{(3)}_{m,k} > 0\}$ term comes from differentiating ReLU.

Now, $\frac{\partial}{\partial b^{(2)}}\left[\rho_{i,k}^{(2)}\right] = \frac{\partial}{\partial b^{(2)}}\left[b^{(2)} + \sum_{j=0}^{127} w_{j,i}^{(2)} \rho_{j,k}^{(1)}\right] \cdot \frac{\partial}{\partial x}\left[f(x)\right]$ $\rightarrow x = \sum_{i,k}^{I(2)}$

$$= 1 \cdot \mathbb{I}_{\left\{\sum_{i,k}^{I(2)} > 0\right\}}$$

So, $\frac{\partial}{\partial b^{(2)}}\left[\exp\left\{\rho_{m,k}^{(3)}\right\}\right] = \exp\left\{\rho_{m,k}^{(3)}\right\} \mathbb{I}_{\left\{\sum_{m,k}^{I(3)} > 0\right\}} \left(\sum_{i=0}^{63} w_{i,m}^{(3)} \mathbb{I}_{\left\{\sum_{i,k}^{I(2)} > 0\right\}}\right)$

So, use quotient rule on the following:

$$\hat{\rho}_{m,k} = \frac{u}{v} = \frac{\exp\left\{\rho_{m,k}^{(3)}\right\}}{\sum_{z=0}^{9} \exp\left\{\rho_{z,k}^{(3)}\right\}}$$

u' is given here ↑

$$v' = \sum_{z=0}^{9} \exp\left\{\rho_{z,k}^{(3)}\right\} \mathbb{I}_{\left\{\sum_{z,k}^{I(3)} > 0\right\}} \left(\sum_{i=0}^{63} w_{i,z}^{(3)} \mathbb{I}_{\left\{\sum_{i,k}^{I(2)} > 0\right\}}\right)$$

So, $\frac{\partial}{\partial b^{(2)}}\left[\hat{\rho}_{m,k}\right] = \frac{vu' - uv'}{v^2}$

$$= \frac{1}{\left(\sum_{z=0}^{9} \exp\left\{\rho_{z,k}^{(3)}\right\}\right)^2} \left(\left(\sum_{z=0}^{9} \exp\left\{\rho_{z,k}^{(3)}\right\}\right) \exp\left\{\rho_{m,k}^{(3)}\right\} \mathbb{I}_{\left\{\sum_{m,k}^{I(3)} > 0\right\}} \left(\sum_{i=0}^{63} w_{i,m}^{(3)} \mathbb{I}_{\left\{\sum_{i,k}^{I(2)} > 0\right\}}\right) \right.$$
$$\left. - \left(\exp\left\{\rho_{m,k}^{(3)}\right\}\right) \sum_{z=0}^{9} \exp\left\{\rho_{z,k}^{(3)}\right\} \mathbb{I}_{\left\{\sum_{z,k}^{I(3)} > 0\right\}} \left(\sum_{i=0}^{63} w_{i,z}^{(3)} \mathbb{I}_{\left\{\sum_{i,k}^{I(2)} > 0\right\}}\right) \right)$$

$$= \hat{\rho}_{m,k} \left(\sum_{i=0}^{63} w_{i,m}^{(3)} \mathbb{I}_{\left\{\sum_{m,k}^{I(3)} > 0, \sum_{i,k}^{I(2)} > 0\right\}} - \sum_{z=0}^{9} \hat{\rho}_{z,k} \sum_{i=0}^{63} w_{i,z}^{(3)} \mathbb{I}_{\left\{\sum_{z,k}^{I(3)} > 0, \sum_{i,k}^{I(2)} > 0\right\}} \right)$$

$$\Rightarrow \frac{\partial}{\partial b^{(2)}}\left[C(\underline{\theta})\right] = \frac{-1}{K} \sum_{k=0}^{K-1} \sum_{m=0}^{9} t_{m,k} \left(\sum_{i=0}^{63} w_{i,m}^{(3)} \mathbb{I}_{\left\{\sum_{m,k}^{I(3)} > 0, \sum_{i,k}^{I(2)} > 0\right\}} - \sum_{z=0}^{9} \hat{\rho}_{z,k} \sum_{i=0}^{63} w_{i,z}^{(3)} \mathbb{I}_{\left\{\sum_{z,k}^{I(3)} > 0, \sum_{i,k}^{I(2)} > 0\right\}} \right)$$

$b^{(1)}$

$$\frac{\partial}{\partial b^{(1)}}\left[C(\underline{\theta})\right] = \frac{-1}{K} \sum_{k=0}^{K-1} \sum_{m=0}^{9} \left(\frac{t_{m,k}}{\hat{\rho}_{m,k}} \frac{\partial}{\partial b^{(1)}}\left[\hat{\rho}_{m,k}\right] \right)$$

Now, $\frac{\partial}{\partial b^{(1)}}\left[\exp\left\{\rho_{m,k}^{(3)}\right\}\right] = \exp\left\{\rho_{m,k}^{(3)}\right\} \frac{\partial}{\partial b^{(1)}}\left[\rho_{m,k}^{(3)}\right]$

$$= \exp\left\{\rho_{m,k}^{(3)}\right\} \mathbb{I}_{\left\{\sum_{m,k}^{I(3)} > 0\right\}} \frac{\partial}{\partial b^{(1)}}\left[b^{(3)} + \sum_{i=0}^{63} w_{i,m}^{(3)} \rho_{i,k}^{(2)}\right]$$

$$= \exp\left\{\rho_{m,k}^{(3)}\right\} \mathbb{I}_{\left\{\sum_{m,k}^{I(3)} > 0\right\}} \left(\sum_{i=0}^{63} w_{i,m}^{(3)} \frac{\partial}{\partial b^{(1)}}\left[\rho_{i,k}^{(2)}\right] \right).$$

Continued next page...

So, $\frac{\partial}{\partial b^{(1)}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q} t_{m,k}\left(\sum_{i=0}^{63}\sum_{j=0}^{127} I^{(3)}_{m,k} w^{(3)}_{i,m} I^{(2)}_{i,k} w^{(2)}_{j,i} I^{(1)}_{j,k} - \sum_{z=0}^{q}\sum_{i=0}^{63}\sum_{j=0}^{127} \hat{p}_{z,k} I^{(3)}_{z,k} w^{(3)}_{i,z} I^{(2)}_{i,k} w^{(2)}_{j,i} I^{(1)}_{j,k}\right)$

$w^{(3)}_{\alpha,\beta}$

$$\frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q} \frac{t_{m,k}}{\hat{p}_{m,k}} \frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[\hat{p}_{m,k}\right]. \quad \text{Now,}$$

$$\frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[\exp\{p^{(3)}_{m,k}\}\right] = \exp\{p^{(3)}_{m,k}\}\frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[p^{(3)}_{m,k}\right]$$

$$= \exp\{p^{(3)}_{m,k}\} I^{(3)}_{m,k} \frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[b^{(3)} + \sum_{i=0}^{63} w^{(3)}_{i,m} p^{(2)}_{i,k}\right]$$

$$= \begin{cases} 0 & \text{if } m \neq \beta \\ \exp\{p^{(3)}_{\beta,k}\} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k} & \text{otherwise} \end{cases}$$

So use quotient rule on the following:

$$\hat{p}_{m,k} = \frac{u}{v} = \frac{\exp\{p^{(3)}_{m,k}\}}{\sum_{z=0}^{q}\exp\{p^{(3)}_{z,k}\}}$$

u' is here

$$u' = \exp\{p^{(3)}_{\beta,k}\} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k} \prod_{\{m=\beta\}}$$

$$v' = \exp\{p^{(3)}_{\beta,k}\} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k}$$

So, $\frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[\hat{p}_{m,k}\right] = \frac{vu' - uv'}{v^2}$

$$= \frac{1}{\left(\sum_{z=0}^{q}\exp\{p^{(3)}_{z,k}\}\right)^2}\left(\left(\sum_{z=0}^{q}\exp\{p^{(3)}_{z,k}\}\right)\exp\{p^{(3)}_{\beta,k}\} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k} \prod_{\{m=\beta\}} - \exp\{p^{(3)}_{m,k}\}\left(\exp\{p^{(3)}_{\beta,k}\} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k}\right)\right)$$

$$= \hat{p}_{m,k}\left(\prod_{\{m=\beta\}} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k} - \hat{p}_{\beta,k} I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k}\right)$$

can take: $\left(\prod_{\{m=\beta\}} - \hat{p}_{\beta,k}\right) \cdot I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k}$

Therefore we arrive at:

$$\frac{\partial}{\partial w^{(3)}_{\alpha,\beta}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q} t_{m,k}\left(\prod_{\{m=\beta\}} - \hat{p}_{\beta,k}\right) I^{(3)}_{\beta,k} p^{(2)}_{\alpha,k}$$

Still Guangzhou → London ✈

05/02/2025

$W^{(2)}_{\alpha,\beta}$

Start with $\frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[\exp\{\rho^{(3)}_{m,k}\}\right] = \dots$

$$\dots = \exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} \frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[b^{(3)} + \sum_{i=0}^{63} W^{(3)}_{i,m}\rho^{(2)}_{i,k}\right]$$

$$= \sum_{i=0}^{63} W^{(3)}_{i,m} \frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[\rho^{(2)}_{i,k}\right]$$

Now, $\frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[\rho^{(2)}_{i,k}\right] = I^{(2)}_{i,k} \frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[b^{(2)} + \sum_{j=0}^{127} W^{(2)}_{j,i}\rho^{(1)}_{j,k}\right]$

$$= I^{(2)}_{i,k} \amalg_{\{i=\beta\}} \rho^{(1)}_{\alpha,k}$$

$$\left(= I^{(2)}_{\beta,k}\rho^{(1)}_{\alpha,k}\amalg_{\{i=\beta\}}\right)$$

$$\Rightarrow \frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[\exp\{\rho^{(3)}_{m,k}\}\right] = \exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} \sum_{i=0}^{63} W^{(3)}_{i,m} I^{(2)}_{\beta,k}\rho^{(1)}_{\alpha,k}\amalg_{\{i=\beta\}}$$

$$= \exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} W^{(3)}_{\beta,m} I^{(2)}_{\beta,k}\rho^{(1)}_{\alpha,k}$$

Interestingly, there is no index indicator here unlike the $W^{(3)}_{\alpha,\beta}$ derivative. I think this is because the final layer ~~weights~~ do not contribute to ALL 9 OF THE FINAL NEURONS, whereas each 2nd (and 1st) layer weight does contribute to EVERY final neuron!

QUOTIENT RULE

$$\hat{\rho}_{m,k} = \frac{\exp\{\rho^{(3)}_{m,k}\}}{\sum_{z=0}^{9}\exp\{\rho^{(3)}_{z,k}\}} = \frac{u}{v}$$

u' is above.

$$v' = \sum_{z=0}^{9}\exp\{\rho^{(3)}_{z,k}\} I^{(3)}_{z,k} W^{(3)}_{\beta,z} I^{(2)}_{\beta,k}\rho^{(1)}_{\alpha,k}.$$

$$\frac{\partial}{\partial W^{(2)}_{\alpha,\beta}}\left[\hat{\rho}_{m,k}\right] = \frac{vu' - uv'}{v^2} =$$

$$= \frac{1}{\left(\sum_{z=0}^{9}\exp\{\rho^{(3)}_{z,k}\}\right)^2}\left(\sum_{z=0}^{9}\exp\{\rho^{(3)}_{z,k}\}\exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} W^{(3)}_{\beta,m} I^{(2)}_{\beta,k}\rho^{(1)}_{\alpha,k}\right.$$

$$\left. - \exp\{\rho^{(3)}_{m,k}\}\sum_{z=0}^{9}\exp\{\rho^{(3)}_{z,k}\} I^{(3)}_{z,k} W^{(3)}_{\beta,z} I^{(2)}_{\beta,k}\rho^{(1)}_{\alpha,k}\right) = \dots$$

$$= \hat{p}_{m,k}\left(I^{(3)}_{m,k} w^{(3)}_{\beta,m} I^{(2)}_{\beta,k} \rho^{(1)}_{\alpha,k} - \sum_{z=0}^{q} \hat{p}_{z,k} w^{(3)}_{\beta,z} I^{(3)}_{z,k} I^{(2)}_{\beta,k} \rho^{(1)}_{\alpha,k}\right)$$

$$\Rightarrow \frac{\partial}{\partial w^{(2)}_{\alpha,\beta}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q} t_{m,k} I^{(2)}_{\beta,k} \rho^{(1)}_{\alpha,k}\left(I^{(3)}_{m,k} w^{(3)}_{\beta,m} - \sum_{z=0}^{q} \hat{p}_{z,k} I^{(3)}_{z,k} w^{(3)}_{\beta,z}\right)$$

$w^{(1)}_{\alpha,\beta}$

Start with $\frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[\exp\{\rho^{(3)}_{m,k}\}\right] = \ldots$

$$= \exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} \sum_{i=0}^{63} w^{(3)}_{i,m} \frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[\rho^{(2)}_{i,k}\right]$$

Now $\frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[\rho^{(2)}_{i,k}\right] = I^{(2)}_{i,k} \frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[b^{(2)} + \sum_{j=0}^{127} w^{(2)}_{j,i} \rho^{(1)}_{j,k}\right]$

$$= I^{(2)}_{i,k} \sum_{j=0}^{127} w^{(2)}_{j,i} \frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[\rho^{(1)}_{j,k}\right]$$

$$= I^{(2)}_{i,k} \sum_{j=0}^{127} w^{(2)}_{j,i} I^{(1)}_{j,k} \frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[b^{(1)} + \sum_{n=0}^{783} w^{(1)}_{n,j} \rho^{(0)}_{n,k}\right]$$

$$= I^{(2)}_{i,k} \sum_{j=0}^{127} w^{(2)}_{j,i} I^{(1)}_{j,k} \prod_{\{j=\beta\}} \rho^{(0)}_{\alpha,k}$$

$$= I^{(2)}_{i,k} w^{(2)}_{\beta,i} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k}$$ (since summing over j zeroes all except $j=\beta$)

So, subbing back in:

$$\frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[\exp\{\rho^{(3)}_{m,k}\}\right] = \exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} \sum_{i=0}^{63} w^{(3)}_{i,m} I^{(2)}_{i,k} w^{(2)}_{\beta,i} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k}$$

(QUOTIENT RULE) $\hat{p}_{m,k} = \frac{u}{v} = \frac{\exp\{\rho^{(3)}_{m,k}\}}{\sum_{z=0}^{q}\exp\{\rho^{(3)}_{z,k}\}}$. $u' = \exp\{\rho^{(3)}_{m,k}\} I^{(3)}_{m,k} \sum_{i=0}^{63} w^{(3)}_{i,m} I^{(2)}_{i,k} w^{(2)}_{\beta,i} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k}$. $v' = \sum_{z=0}^{q} \exp\{\rho^{(3)}_{z,k}\} I^{(3)}_{z,k} \sum_{i=0}^{63} w^{(3)}_{i,z} I^{(2)}_{i,k} w^{(2)}_{\beta,i} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k}$.

$$\Rightarrow \frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[\hat{p}_{m,k}\right] = \frac{vu' - uv'}{v^2} = \hat{p}_{m,k}\left(I^{(3)}_{m,k} \sum_{i=0}^{63} w^{(3)}_{i,m} I^{(2)}_{i,k} w^{(2)}_{\beta,i} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k} - \sum_{z=0}^{q} \hat{p}_{z,k} I^{(3)}_{z,k} \sum_{i=0}^{63} w^{(3)}_{i,z} I^{(2)}_{i,k} w^{(2)}_{\beta,i} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k}\right)$$

$$\Rightarrow \frac{\partial}{\partial w^{(1)}_{\alpha,\beta}}\left[C(\underline{\theta})\right] = \frac{-1}{K}\sum_{k=0}^{K-1}\sum_{m=0}^{q} t_{m,k} I^{(1)}_{\beta,k} \rho^{(0)}_{\alpha,k}\left(I^{(3)}_{m,k} \sum_{i=0}^{63} w^{(3)}_{i,m} I^{(2)}_{i,k} w^{(2)}_{\beta,i} - \sum_{z=0}^{q} \hat{p}_{z,k} I^{(3)}_{z,k} \sum_{i=0}^{63} w^{(3)}_{i,z} I^{(2)}_{i,k} w^{(2)}_{\beta,i}\right).$$