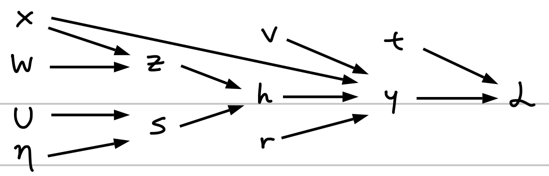
1\. a)



b) $\frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{(1+e^{-x})(0)' - (1)(1+e^{-x})'}{(1+e^{-x})^2} = \frac{-(-e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(\frac{e^{-x}}{1+e^{-x}}\right) = \frac{1}{1+e^{-x}}\left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)$

$= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x))$ (using the derivative quotient rule)

c) Note: $x, z, \eta, s, h, v,$ and r are vectors and U, W are matrices.

$\bar{\mathcal{L}} = 1$

$\bar{y} = \bar{\mathcal{L}}\frac{dL}{dy} = (1)\left(\frac{t}{y} + (-1)\frac{1-t}{1-y}\right) = \frac{t - ty - y + ty}{y(1-y)} = \frac{t-y}{y(1-y)}$

$\bar{h} = \bar{y}\frac{\partial y}{\partial h} = \bar{y}\frac{\partial \sigma(v^T h + r^T x)}{\partial h} = \bar{y}(\sigma(v^T h + r^T x)(1-\sigma(v^T h + r^T x)))\frac{\partial(v^T h + r^T x)}{\partial h} = \bar{y}y(1-y)v = (t-y)v$

$\bar{v} = \bar{y}\frac{\partial y}{\partial v} = \bar{y}\frac{\partial \sigma(v^T h + r^T x)}{\partial v} = \bar{y}(\sigma(v^T h + r^T x)(1-\sigma(v^T h + r^T x)))\frac{\partial(v^T h + r^T x)}{\partial v} = \bar{y}y(1-y)h = (t-y)h$

$\bar{r} = \bar{y}\frac{\partial y}{\partial r} = \bar{y}\frac{\partial \sigma(v^T h + r^T x)}{\partial r} = \bar{y}(\sigma(v^T h + r^T x)(1-\sigma(v^T h + r^T x)))\frac{\partial(v^T h + r^T x)}{\partial r} = \bar{y}y(1-y)x = (t-y)x$

$\bar{z} = \bar{h}\frac{\partial h}{\partial z} = \bar{h}\begin{bmatrix} s_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_n \end{bmatrix} = \bar{h}\,\mathrm{diag}(s) = (t-y)\,\mathrm{diag}(s)v$

$\bar{s} = \bar{h}\frac{\partial h}{\partial s} = \bar{h}\begin{bmatrix} z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & z_n \end{bmatrix} = \bar{h}\,\mathrm{diag}(z) = (t-y)\,\mathrm{diag}(z)v$

$\bar{\eta} = \bar{s}\frac{\partial s}{\partial \eta} = \bar{s}\frac{\partial}{\partial \eta}(U\eta) = U^T\bar{s} = (t-y)U^T\mathrm{diag}(z)v$

$\bar{U} = \bar{s}\frac{\partial s}{\partial U} = \bar{s}\frac{\partial}{\partial U}(U\eta) = \bar{s}\eta^T = (t-y)\,\mathrm{diag}(z)v\eta^T$

$\bar{W} = \bar{z}\frac{\partial z}{\partial W} = \bar{z}\frac{\partial}{\partial W}(Wx) = \bar{z}x^T = (t-y)\,\mathrm{diag}(s)vx^T$

$\bar{x} = \bar{z}\frac{\partial z}{\partial x} + \bar{y}\frac{\partial y}{\partial x} = \bar{z}\frac{\partial}{\partial x}(Wx) + \bar{y}\frac{\partial \sigma(v^T h + r^T x)}{\partial x} = W^T\bar{z} + \bar{y}(\sigma(v^T h + r^T x)(1-\sigma(v^T h + r^T x)))\frac{\partial(v^T h + r^T x)}{\partial x}$

$= W^T\bar{z} + \bar{y}y(1-y)r = (t-y)W^T\mathrm{diag}(s)v + (t-y)r$

$= (t-y)(W^T\mathrm{diag}(s)v + r)$

2. a) Let $n = \#$ of images and $d = 784$. Then,

$$L(\pi,\theta) = \prod_{i=1}^{n} p(x^{(i)}, c^{(i)} \mid \theta, \pi)$$

$$\ell(\pi,\theta) = \log\left(\prod_{i=1}^{n} p(x^{(i)}, c^{(i)} \mid \theta, \pi)\right) = \sum_{i=1}^{n} \log p(x^{(i)}, c^{(i)} \mid \theta, \pi) = \sum_{i=1}^{n} \log\left[p(c^{(i)} \mid \pi) \prod_{j=1}^{d} p(x_j^{(i)} \mid c^{(i)}, \theta_{jc})\right]$$

$$= \sum_{i=1}^{n}\left[\log p(c^{(i)} \mid \pi) + \sum_{j=1}^{d} \log p(x_j^{(i)} \mid c^{(i)}, \theta_{jc})\right] = \sum_{i=1}^{n} \log p(c^{(i)} \mid \pi) + \sum_{j=1}^{d}\sum_{i=1}^{n} \log p(x_j^{(i)} \mid c^{(i)}, \theta_{jc})$$

To find the MLE for π, we maximize the following:

$$\ell_1(\pi) = \sum_{i=1}^{n} \log p(c^{(i)} \mid \pi) = \sum_{i=1}^{n} \log\left(\prod_{j=0}^{9} \pi_j^{t_j^{(i)}}\right) = \sum_{i=1}^{n}\sum_{j=0}^{9} t_j^{(i)} \log \pi_j = \sum_{i=1}^{n}\left[\sum_{j=0}^{8} t_j^{(i)} \log \pi_j + t_9^{(i)} \log\left(1 - \sum_{j=0}^{8} \pi_j\right)\right]$$

Let $j' \in \{0, \dots, 8\}$. Set the derivative to 0, so

$$\frac{d\ell_1}{d\pi_{j'}} \overset{\text{set}}{=} 0 = \sum_{i=1}^{n}\left[\frac{t_{j'}^{(i)}}{\hat{\pi}_{j'}} + \frac{t_9^{(i)}}{1 - \sum_{j=0}^{8} \hat{\pi}_j}(-1)\right] = \frac{\sum_{i=1}^{n} t_{j'}^{(i)}}{\hat{\pi}_{j'}} - \frac{\sum_{i=1}^{n} t_9^{(i)}}{\hat{\pi}_9} \qquad (*)$$

$$\Rightarrow \frac{\hat{\pi}_{j'}}{\hat{\pi}_9} = \frac{\sum_{i=1}^{n} t_{j'}^{(i)}}{\sum_{i=1}^{n} t_9^{(i)}} \Rightarrow \frac{\hat{\pi}_1}{\hat{\pi}_9} + \dots + \frac{\hat{\pi}_8}{\hat{\pi}_9} = \frac{1 - \hat{\pi}_9}{\hat{\pi}_9} = \frac{\sum_{i=1}^{n} t_1^{(i)}}{\sum_{i=1}^{n} t_9^{(i)}} + \dots + \frac{\sum_{i=1}^{n} t_8^{(i)}}{\sum_{i=1}^{n} t_9^{(i)}} = \frac{n - \sum_{i=1}^{n} t_9^{(i)}}{\sum_{i=1}^{n} t_9^{(i)}}$$

since $\sum_{i=1}^{n} t_1^{(i)} + \dots + \sum_{i=1}^{n} t_9^{(i)} = n$

$$\Rightarrow 1 - \hat{\pi}_9 = \hat{\pi}_9 \frac{n - \sum_{i=1}^{n} t_9^{(i)}}{\sum_{i=1}^{n} t_9^{(i)}} \Rightarrow \hat{\pi}_9\left[\frac{n - \sum_{i=1}^{n} t_9^{(i)}}{\sum_{i=1}^{n} t_9^{(i)}} + 1\right] = \hat{\pi}_9\left[\frac{n}{\sum_{i=1}^{n} t_9^{(i)}}\right] = 1$$

$$\Rightarrow \hat{\pi}_9 = \frac{1}{n}\sum_{i=1}^{n} t_9^{(i)}$$

Thus, $(*)$ becomes $\dfrac{\sum_{i=1}^{n} t_{j'}^{(i)}}{\hat{\pi}_{j'}} - \dfrac{\sum_{i=1}^{n} t_9^{(i)}}{\frac{\sum_{i=1}^{n} t_9^{(i)}}{n}} = \dfrac{\sum_{i=1}^{n} t_{j'}^{(i)}}{\hat{\pi}_{j'}} - n = 0 \Rightarrow \hat{\pi}_{j'} = \dfrac{1}{n}\sum_{i=1}^{n} t_{j'}^{(i)}$

Therefore, $\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} t_j^{(i)}$ for $j \in \{0, \dots, 9\}$.

(continued on next page)

To find the MLE for Θ, we maximize the following:

$$\ell_2(\theta) = \sum_{i=1}^{n} \log p(x_j^{(i)} \mid c^{(i)}, \theta_{jc}) = \sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c) \log\left(\theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{(1-x_j^{(i)})}\right)$$

since we only consider terms where $c^{(i)}=c$, by the definition of θ_{jc}

$$= \sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c)\left[x_j^{(i)} \log\theta_{jc} + (1-x_j^{(i)})\log(1-\theta_{jc})\right]$$

Set the derivative to 0, so

$$\frac{d\ell_2}{d\theta_{jc}} \overset{\text{set}}{=} 0 = \sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c)\left[\frac{x_j^{(i)}}{\hat{\theta}_{jc}} - \frac{1-x_j^{(i)}}{1-\hat{\theta}_{jc}}\right] = \sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c)(x_j^{(i)} - \hat{\theta}_{jc})$$

$$\Rightarrow \hat{\theta}_{jc} = \frac{\sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c)\, x_j^{(i)}}{\sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c)} = \frac{\sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c \text{ and } x_j^{(i)}=1)}{\sum_{i=1}^{n} \mathbb{1}(c^{(i)}=c)} \quad \text{as needed.}$$

b) $$\log p(t \mid x, \theta, \pi) = \log \frac{p(x, c \mid \theta, \pi)}{\sum_{c'} p(x, c' \mid \theta, \pi)} = \log(p(c\mid\pi)\, p(x \mid t, \theta, \pi)) - \log\left(\sum_{c'} p(c'\mid\pi)\, p(x\mid t, \theta, \pi)\right)$$

$$= \log\left(p(c\mid\pi) \prod_{j=1}^{784} p(x_j \mid t, \theta_{jc})\right) - \log\left(\sum_{c'} p(c'\mid\pi) \prod_{j=1}^{784} p(x_j \mid t, \theta_{jc'})\right)$$

$$= \log p(c\mid\pi) + \sum_{j=1}^{784} \log\left(\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)}\right) - \log\left(\sum_{c'=0}^{9} p(c'\mid\pi) \prod_{j=1}^{784} \theta_{jc'}^{x_j}(1-\theta_{jc'})^{(1-x_j)}\right)$$

$$= \log \pi_c + \sum_{j=1}^{784}\left[x_j \log\theta_{jc} + (1-x_j)\log(1-\theta_{jc})\right] - \log\left[\sum_{c'=0}^{9} \pi_{c'} \prod_{j=1}^{784} \theta_{jc'}^{x_j}(1-\theta_{jc'})^{(1-x_j)}\right]$$

c) The average log-likelihood using the MLE estimators is NaN since a divide by zero error occurred. This is since some θ_{jc}'s are 0, and $\log\theta_{jc} = \log 0 = -\infty$ for these values, which causes the error.

Part d) on last page.