2. (b)

$$J = -\sum_{i=1}^{V} y_i \ln \hat{y}_i = -\{\boldsymbol{y}\}^{\mathsf{T}} \ln(\operatorname{softmax}(\{\boldsymbol{\theta}\}))$$
$$= -\langle \boldsymbol{y} \rangle \left( \{\boldsymbol{\theta}\} - \ln \sum_{i=1}^{V} e^{\theta_i} \right).$$

$$\begin{split} \frac{\partial J}{\partial \langle \boldsymbol{\theta} \rangle} &= -\langle \boldsymbol{y} \rangle \left( [\boldsymbol{I}] - \frac{1}{\sum e^{\theta_i}} \cdot \sum_{i=1}^{V} \frac{\partial e^{\theta_i}}{\partial \langle \boldsymbol{\theta} \rangle} \right) \\ &= -\langle \boldsymbol{y} \rangle \left( [\boldsymbol{I}] - \frac{e^{\langle \boldsymbol{\theta} \rangle}}{\sum e^{\theta_i}} \right) \\ &= -\langle \boldsymbol{y} \rangle \left( [\boldsymbol{I}] - \langle \hat{\boldsymbol{y}} \rangle \right) \\ &= \langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle \,. \end{split}$$

3. (a)

$$\begin{split} \frac{\partial J}{\partial \left< \boldsymbol{v}_c \right>} &= \frac{\partial J}{\partial \left< \boldsymbol{\theta} \right>} \frac{\partial \left\{ \boldsymbol{\theta} \right\}}{\partial \left< \boldsymbol{v}_c \right>} \\ &= \left( \left< \hat{\boldsymbol{y}} \right> - \left< \boldsymbol{y} \right> \right) \frac{\partial \left\{ \boldsymbol{\theta} \right\}}{\partial \left< \boldsymbol{v}_c \right>} \end{split}$$

Since

$$\left\{oldsymbol{ heta}
ight\} = egin{dcases} \left\langle oldsymbol{u}_1 
ight
angle \left\{ egin{array}{l} \left\langle oldsymbol{u}_2 
ight
angle \left\{ oldsymbol{v}_c 
ight\} \\ dots \\ \left\langle oldsymbol{u}_V 
ight
angle \left\{ oldsymbol{v}_c 
ight\} \\ 
ight. \end{cases} = egin{bmatrix} \left\{ oldsymbol{u}_1 
ight\} & \left\{ oldsymbol{u}_2 
ight\} & \cdots & \left\{ oldsymbol{u}_V 
ight\} \end{bmatrix}^\mathsf{T} \left\{ oldsymbol{v}_c 
ight\}, \end{cases}$$

therefore,

$$\begin{split} &\frac{\partial J}{\partial \left\langle \boldsymbol{v}_{c}\right\rangle }=\left(\left\langle \hat{\boldsymbol{y}}\right\rangle -\left\langle \boldsymbol{y}\right\rangle \right)\left[\boldsymbol{U}\right]^{\mathsf{T}}.\\ &\frac{\partial J}{\partial \left\{\boldsymbol{v}_{c}\right\} }=\left[\boldsymbol{U}\right]\left(\left\{\hat{\boldsymbol{y}}\right\} -\left\{\boldsymbol{y}\right\} \right). \end{split}$$

3. (b)

$$\begin{split} \frac{\partial J}{\partial \left[\boldsymbol{U}\right]} &= \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right) \frac{\partial \left\{\boldsymbol{\theta}\right\}}{\partial \left[\boldsymbol{U}\right]} \\ &= \begin{bmatrix} \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right) & \begin{bmatrix} \frac{\partial \left\{u_{1}\right\}^{\mathsf{T}} \left\{v_{c}\right\}}{\partial \left\langle u_{1}\right\rangle} \\ \vdots \\ \frac{\partial \left\{u_{1}\right\}^{\mathsf{T}} \left\{v_{c}\right\}}{\partial \left\langle u_{1}\right\rangle} \end{bmatrix} \\ &\vdots \\ \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right) & \begin{bmatrix} \frac{\partial \left\{u_{2}\right\}^{\mathsf{T}} \left\{v_{c}\right\}}{\partial \left\langle u_{1}\right\rangle} \\ \vdots \\ \frac{\partial \left\{u_{2}\right\}^{\mathsf{T}} \left\{v_{c}\right\}}{\partial \left\langle u_{1}\right\rangle} \end{bmatrix} \\ &\vdots \\ \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right) & \begin{bmatrix} \partial \left\{u_{V}\right\}^{\mathsf{T}} \left\{v_{c}\right\}}{\partial \left\langle u_{V}\right\rangle} \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right) & \begin{bmatrix} \langle \boldsymbol{v}_{c} \rangle \\ \langle \boldsymbol{0} \rangle \\ \vdots \\ \langle \boldsymbol{0} \rangle \end{bmatrix} \\ &\vdots \\ \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right) & \begin{bmatrix} \langle \boldsymbol{0} \rangle \\ \vdots \\ \langle \boldsymbol{v}_{c} \rangle \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right)_{*1} \cdot \langle \boldsymbol{v}_{c} \rangle \\ \vdots \\ \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right)_{*2} \cdot \langle \boldsymbol{v}_{c} \rangle \\ \vdots \\ \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right)_{*2} \cdot \langle \boldsymbol{v}_{c} \rangle \\ \vdots \\ \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right)_{*2} \cdot \langle \boldsymbol{v}_{c} \rangle \end{bmatrix} \\ &= \left\{\boldsymbol{v}_{c}\right\} \left(\langle \hat{\boldsymbol{y}} \rangle - \langle \boldsymbol{y} \rangle\right)_{*}. \end{split}$$