2. (b)

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

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

3. (a)

$$
\begin{aligned}
\frac{\partial J}{\partial \left\langle \boldsymbol{v}_c \right\rangle} &= \frac{\partial J}{\partial \left\langle \boldsymbol{\theta} \right\rangle} \frac{\partial \{\boldsymbol{\theta}\}}{\partial \left\langle \boldsymbol{v}_c \right\rangle} \\
&= \left(\left\langle \hat{\boldsymbol{y}} \right\rangle - \left\langle \boldsymbol{y} \right\rangle \right) \frac{\partial \{\boldsymbol{\theta}\}}{\partial \left\langle \boldsymbol{v}_c \right\rangle}.
\end{aligned}
$$

Since

$$
\{\boldsymbol{\theta}\} = \begin{Bmatrix} \left\langle \boldsymbol{u}_1 \right\rangle \{\boldsymbol{v}_c\} \\ \left\langle \boldsymbol{u}_2 \right\rangle \{\boldsymbol{v}_c\} \\ \vdots \\ \left\langle \boldsymbol{u}_V \right\rangle \{\boldsymbol{v}_c\} \end{Bmatrix} = \begin{bmatrix} \{\boldsymbol{u}_1\} & \{\boldsymbol{u}_2\} & \cdots & \{\boldsymbol{u}_V\} \end{bmatrix}^{\mathsf{T}} \{\boldsymbol{v}_c\},
$$

therefore,

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

3. (b)

$$
\begin{aligned}
\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\{\boldsymbol{u}_1\}^\mathsf{T}\{\boldsymbol{v}_c\}}{\partial\langle\boldsymbol{u}_1\rangle}\\ \vdots \\ \frac{\partial\{\boldsymbol{u}_1\}^\mathsf{T}\{\boldsymbol{v}_c\}}{\partial\langle\boldsymbol{u}_V\rangle}\end{bmatrix}\\
\left(\langle\hat{\boldsymbol{y}}\rangle - \langle\boldsymbol{y}\rangle\right)\begin{bmatrix}\frac{\partial\{\boldsymbol{u}_2\}^\mathsf{T}\{\boldsymbol{v}_c\}}{\partial\langle\boldsymbol{u}_1\rangle}\\ \vdots \\ \frac{\partial\{\boldsymbol{u}_2\}^\mathsf{T}\{\boldsymbol{v}_c\}}{\partial\langle\boldsymbol{u}_V\rangle}\end{bmatrix}\\
\vdots\\
\left(\langle\hat{\boldsymbol{y}}\rangle - \langle\boldsymbol{y}\rangle\right)\begin{bmatrix}\frac{\partial\{\boldsymbol{u}_V\}^\mathsf{T}\{\boldsymbol{v}_c\}}{\partial\langle\boldsymbol{u}_1\rangle}\\ \vdots \\ \frac{\partial\{\boldsymbol{u}_V\}^\mathsf{T}\{\boldsymbol{v}_c\}}{\partial\langle\boldsymbol{u}_V\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}\\
\left(\langle\hat{\boldsymbol{y}}\rangle - \langle\boldsymbol{y}\rangle\right)\begin{bmatrix}\langle\boldsymbol{0}\rangle\\ \langle\boldsymbol{v}_c\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\\
\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)_{*V}\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{aligned}
$$