2(b) $\dfrac{\partial(CE(\mathbf{y},\widehat{\mathbf{y}}))}{\partial\boldsymbol{\theta}}$

=	$-\dfrac{\partial}{\partial\boldsymbol{\theta}}\sum_i\left(y_i\log(\widehat{y_i})\right)$	(1)
=	$\begin{bmatrix} -\dfrac{\partial}{\partial\theta_1}\left(\left(y_1\log\widehat{y_1}\right)+\left(y_2\log\widehat{y_2}\right)+\ldots+\left(y_k\log\widehat{y_k}\right)+\ldots\right) \\ -\dfrac{\partial}{\partial\theta_2}\left(\left(y_1\log\widehat{y_1}\right)+\left(y_2\log\widehat{y_2}\right)+\ldots+\left(y_k\log\widehat{y_k}\right)+\ldots\right) \\ \vdots \\ -\dfrac{\partial}{\partial\theta_k}\left(\left(y_1\log\widehat{y_1}\right)+\left(y_2\log\widehat{y_2}\right)+\ldots+\left(y_k\log\widehat{y_k}\right)+\ldots\right) \\ \vdots \end{bmatrix}$	(2)
=	$\begin{bmatrix} -\dfrac{\partial}{\partial\theta_1}\left(\left(y_1\log\left(\dfrac{\exp(\theta_1)}{\sum_m\exp(\theta_m)}\right)\right)+\left(y_2\log\left(\dfrac{\exp(\theta_2)}{\sum_m\exp(\theta_m)}\right)\right)+\ldots+\left(y_k\log\left(\dfrac{\exp(\theta_k)}{\sum_m\exp(\theta_m)}\right)\right)+\ldots\right) \\ -\dfrac{\partial}{\partial\theta_2}\left(\left(y_1\log\left(\dfrac{\exp(\theta_1)}{\sum_m\exp(\theta_m)}\right)\right)+\left(y_2\log\left(\dfrac{\exp(\theta_2)}{\sum_m\exp(\theta_m)}\right)\right)+\ldots+\left(y_k\log\left(\dfrac{\exp(\theta_k)}{\sum_m\exp(\theta_m)}\right)\right)+\ldots\right) \\ \vdots \\ -\dfrac{\partial}{\partial\theta_k}\left(\left(y_1\log\left(\dfrac{\exp(\theta_1)}{\sum_m\exp(\theta_m)}\right)\right)+\left(y_2\log\left(\dfrac{\exp(\theta_2)}{\sum_m\exp(\theta_m)}\right)\right)+\ldots+\left(y_k\log\left(\dfrac{\exp(\theta_k)}{\sum_m\exp(\theta_m)}\right)\right)+\ldots\right) \\ \vdots \end{bmatrix}$	(3)

If we look at the j^{th} row of Equation (3), it is of the form below:

=	$-\dfrac{\partial}{\partial\theta_j}\left(\sum_i y_i\log\left(\dfrac{\exp(\theta_i)}{\sum_m\exp(\theta_m)}\right)\right)$	(4)
=	$-\sum_i\dfrac{\partial}{\partial\theta_j}\left(y_i\log\left(\dfrac{\exp(\theta_i)}{\sum_m\exp(\theta_m)}\right)\right)$	(5)

$$= \quad -\sum_i \frac{\partial}{\partial\theta_j}\left(y_i\left(\log\left(\exp(\theta_i)\right) - \log\sum_m \exp(\theta_m)\right)\right) \tag{6}$$

$$= \quad -\sum_i y_i\left(\frac{\partial\theta_i}{\partial\theta_j} - \frac{\partial}{\partial\theta_j}\left(\log\sum_m \exp(\theta_m)\right)\right) \tag{7}$$

$$= \quad -\sum_i y_i\left(\frac{\partial\theta_i}{\partial\theta_j} - \left(\frac{1}{\sum_m \exp(\theta_m)}\right)\frac{\partial}{\partial\theta_j}\left(\sum_x \exp(\theta_x)\right)\right) \tag{8}$$

$$= \quad -\sum_i y_i\left(\frac{\partial\theta_i}{\partial\theta_j} - \sum_x \frac{\frac{\partial}{\partial\theta_j}\left(\exp(\theta_x)\right)}{\sum_m \exp(\theta_m)}\right) \tag{9}$$

When $i = j$ in Equation (9):

$$= \quad -\left(\ldots + y_j\left(\frac{\partial\theta_j}{\partial\theta_j} - \frac{\frac{\partial}{\partial\theta_j}\left(\exp(\theta_1)+\exp(\theta_2)+\ldots.+\exp(\theta_j)+\ldots\right)}{\sum_m \exp(\theta_m)}\right)+\ldots\right) \tag{10}$$

$$= \quad -\left(\ldots + y_j\left(1 - \frac{\exp(\theta_j)}{\sum_m \exp(\theta_m)}\right)+\ldots\right) \tag{11}$$

$$= \quad -\left(\ldots + y_j\left(1-\hat{y}_j\right)+\ldots\right) \tag{12}$$

$$= \quad -\left(\ldots + y_j - y_j\hat{y}_j + \ldots\right) \tag{13}$$

Now let’s consider all cases where $i \neq j$ in Equation(9):

$$= \quad -\left(\ldots + y_{i\neq j}\left(\frac{\partial\theta_{i\neq j}}{\partial\theta_j} - \frac{\frac{\partial}{\partial\theta_j}\left(\exp(\theta_1)+\exp(\theta_2)+\ldots.+\exp(\theta_j)+\ldots\right)}{\sum_m \exp(\theta_m)}\right)+\ldots\right) \tag{14}$$

$$= \quad -\left(\ldots + y_{i\neq j}\left(0 - \frac{\exp(\theta_j)}{\sum_m \exp(\theta_m)}\right)+\ldots\right) \tag{15}$$

=	$-\left(\ldots+y_{i\neq j}\left(0-\hat{y}_j\right)+\ldots\right)$	(16)
=	$-\left(\ldots+-y_{i\neq j}\hat{y}_j+\ldots\right)$	(17)

Combining Equations (13) and (17):

=	$-\left(y_j-y_j\hat{y}_j-y_{i\neq j}\hat{y}_j+\ldots\right)$	(18)
=	$-y_j+y_j\hat{y}_j+y_{i\neq j}\hat{y}_j+\ldots$	(19)
=	$-y_j+\hat{y}_j\sum_n y_n$	(20)

Now, $\sum_n y_n=1$ because $\mathbf{y}$ is a one-hot label vector whose elements are 0 except for the k^{th} dimension (only $y_k=1$). Substituting in Equation (20):

=	$-y_j+\hat{y}_j(1)$	(21)
=	$\hat{y}_j-y_j$	(22)

Substituting in Equation (3):

$\frac{\partial(CE(\mathbf{y},\hat{\mathbf{y}}))}{\partial\boldsymbol{\theta}}=$	$\begin{bmatrix}\hat{y}_1-y_1\\ \hat{y}_2-y_2\\ \vdots\\ \hat{y}_k-y_k\\ \vdots\end{bmatrix}$	(23)
=	$\hat{\mathbf{y}}-\mathbf{y}$	(24)

Assuming *k* is the correct class (i.e., $y_k=1$):

=	$\begin{Bmatrix}\hat{y}_k-1, & i=k\\ \hat{y}_k, & \text{otherwise}\end{Bmatrix}$	(25)