2.3 We can write the gradient $\nabla_\theta L(\theta)$ as

$$\nabla_\theta L(\theta) = \left[\frac{dL}{d\theta_0}, \frac{dL}{d\theta_1}, \cdots \frac{dL}{d\theta_{M-1}} \right]$$

where

$$\frac{dL}{d\theta_k} = \sum_{n=0}^{N-1} \left[-x_{n,k} + \sum_{i=0}^{M-1} \left[x_{n,i} \frac{e^{\theta_k}}{\sum_{j=0}^{M-1} e^{\theta_j}} \right] \right] \qquad k \in [0, M-1]$$

$$\frac{dL}{d\theta_k} = \sum_{n=0}^{N-1} \sum_{i=0}^{M-1} \left[-x_{n,i} \left[\delta(k-i) - \frac{e^{\theta_k}}{\sum_{j=0}^{M-1} e^{\theta_j}} \right] \right]$$

$$\frac{dL}{d\theta_k} = \sum_{i=0}^{M-1} \left[\underbrace{\sum_{n=0}^{N-1} -x_{n,i}}_{\text{From } 2.1} \left[\delta(k-i) - \frac{e^{\theta_k}}{\sum_{j=0}^{M-1} e^{\theta_j}} \right] \right]$$

$$\frac{dL}{d\theta_k} = \sum_{i=0}^{M-1} \left[-N_i \left[\delta(k-i) - \frac{e^{\theta_k}}{\sum_{j=0}^{M-1} e^{\theta_j}} \right] \right] \qquad \left(\sum_t g(t)\,\delta(t-t_0) = g(t_0) \right)$$

$$\frac{dL}{d\theta_k} = -N_k + \left[\sum_{i=0}^{M-1} N_i \right] \frac{e^{\theta_k}}{\sum_{j=0}^{M-1} e^{\theta_j}}$$

- To obtain θ_k^* that minimizes Loss

$$\frac{dL}{d\theta_k} = 0$$

$$-N_k + \left[\sum_{i=0}^{M-1} N_i \right] \frac{e^{\theta_k^*}}{\sum_{j=0}^{M-1} e^{\theta_j}} = 0$$

$$e^{\theta_k^*} = \underbrace{\frac{\sum_{j=0}^{M-1} e^{\theta_j}}{\sum_{i=0}^{M-1} N_i}}_{\alpha} \cdot N_k$$

$$\Rightarrow e^{\theta_k^*} = \alpha N_k$$

$$\theta_k^* = \beta + \log N_k \qquad (\beta = \log \alpha)$$

β: For any constant β