Exercise: Symmetric version of ℓ_2 regularized multinomial logistic regression

(Source: Ex 18.3 of (**?**).)

Multiclass logistic regression has the form

$$p(y = c|\mathbf{x}, \mathbf{W}) = \frac{\exp(w_{c0} + \mathbf{w}_c^T \mathbf{x})}{\sum_{k=1}^{C} \exp(w_{k0} + \mathbf{w}_k^T \mathbf{x})} \tag{1}$$

where $\mathbf{W}$ is a $(D+1) \times C$ weight matrix. We can arbitrarily define $\mathbf{w}_c = \mathbf{0}$ for one of the classes, say $c = C$, since $p(y = C|\mathbf{x}, \mathbf{W}) = 1 - \sum_{c=1}^{C-1} p(y = c|\mathbf{x}, \mathbf{w})$. In this case, the model has the form

$$p(y = c|\mathbf{x}, \mathbf{W}) = \frac{\exp(w_{c0} + \mathbf{w}_c^T \mathbf{x})}{1 + \sum_{k=1}^{C-1} \exp(w_{k0} + \mathbf{w}_k^T \mathbf{x})} \tag{2}$$

If we don't "clamp" one of the vectors to some constant value, the parameters will be unidentifiable. However, suppose we don't clamp $\mathbf{w}_c = \mathbf{0}$, so we are using Equation 1, but we add ℓ_2 regularization by optimizing

$$\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \mathbf{W}) - \lambda \sum_{c=1}^{C} ||\mathbf{w}_c||_2^2 \tag{3}$$

Show that at the optimum we have $\sum_{c=1}^{C} \hat{w}_{cj} = 0$ for $j = 1 : D$. (For the unregularized $\hat{w}_{c0}$ terms, we still need to enforce that $w_{0C} = 0$ to ensure identifiability of the offset.)