HW 2 Solution

1 Problem 1 (10 points)

a) 5 points

$$\begin{aligned}\frac{\partial(U^T AV)}{\partial X} &= \frac{\partial V}{\partial X} A^T U + \frac{\partial(A^T U)}{\partial X} V \\ &= \frac{\partial V}{\partial X} A^T U + \frac{\partial U}{\partial X} AV\end{aligned}$$

The first equality is the result of applying the chain rule and the second one holds due to the fact that A is not a function of X.
3 points for applying the chain rule, and **2 points** for simplification in the second step.

b) 5 points

$$\begin{aligned}\frac{\partial(U^T AV)}{\partial X} &= \underbrace{\frac{\partial U}{\partial X}}_{=0} AV + \frac{\partial(AV)}{\partial X} U \\ &= \underbrace{\frac{\partial V}{\partial X}}_{=0} A^T U + \left[\left[\left(\frac{\partial A}{\partial X}\right)^{M\times N\times K} V^{K\times 1}\right]^{M\times N} U^{N\times 1}\right]^{M\times 1} \\ &= \sum_i \sum_j U_i V_j \frac{\partial A_{ij}}{\partial X}\end{aligned}$$

2 points for applying the chain rule, **1 point** for recognizing the zero terms, and **2 points** for finding the final answer with the correct dimension.

2 Problem 2 (10 points)

a) 5 points

The sigmoid function $\sigma(z)$ can be written as:

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \tag{1}$$

Taking derivative of the sigmoid function with respect to z, we can get:

$$\begin{aligned}&\frac{\partial \sigma(z)}{\partial z} = \frac{\exp(-z)}{(1 + \exp(-z))^2} = \frac{\exp(-z)}{1 + \exp(-z)} * \frac{1}{1 + \exp(-z)} \\ &= \sigma(z)(1 - \sigma(z))\end{aligned} \tag{2}$$

4 points for calculating the derivative correctly, and **1 point** for writing the final answer in terms of sigmoid functions.

b) 5 points

The multiclass sigmoid function can be expressed as:

$$\sigma_M(z_k) = \frac{\exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)} \tag{3}$$

Taking derivative of the multiclass sigmoid function with respect to z_k, we can get:

$$\begin{aligned}&\frac{\partial \sigma_M(z_k)}{\partial z_k} = \frac{\exp(z_k)(\sum_{i=1}^{K} \exp(z_i)) - \exp(z_k)\exp(z_k)}{(\sum_{i=1}^{K} \exp(z_i))^2} \\ &= \frac{\exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)} - (\frac{\exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)})^2 \\ &= \sigma_M(z_k)(1 - \sigma_M(z_k))\end{aligned} \tag{4}$$

4 points for calculating the derivative correctly, **2 points** for each of the numerator and denominator, and **1 points** for writing the final answer in terms of sigmoid functions.

3 Problem 3 (10 points)

Assume there are m training samples $\mathcal{D}_{m=1}^{M} = \{\boldsymbol{x}[m], \boldsymbol{y}[m]\}$. Construct the data matrix $\boldsymbol{A}$:

$$\boldsymbol{A}^{M\times(N+1)} = \begin{bmatrix} A[1] \\ A[2] \\ \vdots \\ A[M] \end{bmatrix} \tag{5}$$

with mth row composed of: $A[m] = \begin{bmatrix} \boldsymbol{x}[m] & 1 \end{bmatrix}^{1\times(N+1)}$. And the output data matrix $\boldsymbol{y}$:

$$\boldsymbol{y}^{M\times 2} = \begin{bmatrix} y_1[1] & y_2[1] \\ y_1[2] & y_2[2] \\ \vdots & \\ y_1[M] & y_2[M] \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 \end{bmatrix} \tag{6}$$

So the predicted output matrix $\hat{\boldsymbol{y}}$ can be expressed as:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{A}^{M\times(N+1)} W^{(N+1)\times 2} \\ &= \boldsymbol{A}^{M\times(N+1)} \begin{bmatrix} W_1^{(N+1)\times 1} & W_2^{(N+1)\times 1} \end{bmatrix} \end{aligned} \tag{7}$$

The mean squared errors are:

$$L_{MSE}(D : W) = \frac{1}{M}[(\boldsymbol{A}W_1 - \boldsymbol{y}_1)^t(\boldsymbol{A}W_1 - \boldsymbol{y}_1) + (\boldsymbol{A}W_2 - \boldsymbol{y}_2)^t(\boldsymbol{A}W_2 - \boldsymbol{y}_2)] \tag{8}$$

Taking derivative of the mean squared error with respect to W_1, we can get:

$$\begin{aligned} \nabla_{W_1} L_{MSE} &= \frac{\partial L_{MSE}}{\partial W_1} \\ &= \frac{\partial \frac{1}{M}[(\boldsymbol{A}W_1 - \boldsymbol{y}_1)^t(\boldsymbol{A}W_1 - \boldsymbol{y}_1) + (\boldsymbol{A}W_2 - \boldsymbol{y}_2)^t(\boldsymbol{A}W_2 - \boldsymbol{y}_2)]}{\partial W_1} \\ &= \frac{2}{M}\boldsymbol{A}^t(\boldsymbol{A}W_1 - \boldsymbol{y}_1) \end{aligned} \tag{9}$$

Following the same procedure, we can get:

$$\nabla_{W_2} L_{MSE} = \frac{2}{M}\boldsymbol{A}^t(\boldsymbol{A}W_2 - \boldsymbol{y}_2) \tag{10}$$

Setting the gradient equal to zero, we can get:

$$W = \begin{bmatrix} W_1 & W_2 \end{bmatrix} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}\boldsymbol{y} \tag{11}$$

while $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 \end{bmatrix}$.

2 points for forming the augmented input matrix and the output matrix, **2 points** for writing the total cost functions in matrix format, **3 points** for correctly calculating the gradient, and **3 points** for finding the closed-from for the final solution.

4 Problem 4 (10 points)

By using the same notations as in the class, we have vector

$$y[m] = \big[y[m][1], y[m][2], y[m][3]\big]^T$$

as the output for data point m, where $y[m][i] = 1$ when $\boldsymbol{y}[m] = i$, and define $X[m] = [x[m], 1]^T$ as the augmented input, where $x[m]$ is the feature vector for data point m. We know that when we use softmax, the conditional likelihood of the output vector is

$$p(y[m] \big| X[m], \boldsymbol{\theta}) = \prod_{k=1}^{3} \Big(\sigma_3(\boldsymbol{\theta}_k^T X[m])\Big)^{y[m][k]} .$$

Therefore, the overall cost function, which is the negative log conditional likelihood added by the L-1 norm, is

$$\begin{aligned}
L(\boldsymbol{D} : \boldsymbol{\theta}) &= -\sum_{m=1}^{N} \log p(y[m] \big| X[m], \boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta}) \\
&= -\sum_{m=1}^{N} \log \prod_{k=1}^{3} \Big(\sigma_3(\boldsymbol{\theta}_k^T X[m])\Big)^{y[m][k]} + \lambda R(\boldsymbol{\theta}) \\
&= -\sum_{m=1}^{N} \sum_{k=1}^{3} y[m][k] \log \Big(\sigma_3(\boldsymbol{\theta}_k^T X[m])\Big) + \lambda R(\boldsymbol{\theta}) \\
&= -\sum_{m=1}^{N} \sum_{k=1}^{3} y[m][k] \log \frac{\exp\{\boldsymbol{\theta}_k^T X[m]\}}{\sum_{\ell=1}^{3} \exp\{\boldsymbol{\theta}_\ell^T X[m]\}} + \lambda R(\boldsymbol{\theta}) \\
&= -\sum_{m=1}^{N} \sum_{k=1}^{3} y[m][k] \Big[\boldsymbol{\theta}_k^T X[m] - \log \sum_{\ell=1}^{3} \exp\{\boldsymbol{\theta}_\ell^T X[m]\}\Big] + \lambda R(\boldsymbol{\theta})
\end{aligned}$$

Now, the gradient equation for any $\boldsymbol{\theta}_k,\ k \in \{1, 2, 3\}$ can be obtained from the fact that only *one* of the elements of $y[m]$ is 1 for any m and the rests are 0, and also the derivative of the L-1 norm is the sign function, where

here we assume that it is applied element-wise on any vector

$$
\begin{aligned}
\frac{\partial L(\boldsymbol{D}:\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} &= -\sum_{m=1}^{N}\sum_{j=1}^{3}\frac{\partial}{\partial \boldsymbol{\theta}_k}\left[y[m][j]\left[\boldsymbol{\theta}_j^T X[m] - \log\sum_{\ell=1}^{3}\exp\{\boldsymbol{\theta}_\ell^T X[m]\}\right]\right] + \lambda\frac{\partial R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} \\
&= -\sum_{m=1}^{N}\left[y[m][k]X[m] - \frac{\partial}{\partial \boldsymbol{\theta}_k}\underbrace{\sum_{j=1}^{3}y[m][j]}_{=1}\log\sum_{\ell=1}^{3}\exp\{\boldsymbol{\theta}_\ell^T X[m]\}\right] + \lambda\mathrm{sign}(\boldsymbol{\theta}_k) \\
&= -\sum_{m=1}^{N}\left[y[m][k]X[m] - \frac{\partial}{\partial \boldsymbol{\theta}_k}\log\sum_{\ell=1}^{3}\exp\{\boldsymbol{\theta}_\ell^T X[m]\}\right] + \lambda\mathrm{sign}(\boldsymbol{\theta}_k) \\
&= -\sum_{m=1}^{N}\left[y[m][k]X[m] - \frac{X[m]\exp\{\boldsymbol{\theta}_k^T X[m]\}}{\sum_{\ell=1}^{3}\exp\{\boldsymbol{\theta}_\ell^T X[m]\}}\right] + \lambda\mathrm{sign}(\boldsymbol{\theta}_k) \\
&= -\sum_{m=1}^{N}X[m]\left[y[m][k] - \frac{\exp\{\boldsymbol{\theta}_k^T X[m]\}}{\sum_{\ell=1}^{3}\exp\{\boldsymbol{\theta}_\ell^T X[m]\}}\right] + \lambda\mathrm{sign}(\boldsymbol{\theta}_k) \\
&= -\sum_{m=1}^{N}X[m]\Big[y[m][k] - \sigma_3(\boldsymbol{\theta}_k^T X[m])\Big] + \lambda\mathrm{sign}(\boldsymbol{\theta}_k)
\end{aligned}
$$

2 points for forming the negative log conditional likelihood, **2 points** for calculating the derivative of the L-1 norm correctly, **3 points** for correctly identifying the relevant term for taking the derivatives, and **3 points** for finding the closed-from for the final solution.