Optimization
Graded Assignment 2

Name: Faran Taimoor Butt

Email: faranbutt@phystech.edu

1. (1 point) Compute the gradients with respect to $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{k\times n}$, $k < n$ of the function

$$J(U,V) = \|UV - Y\|_F^2 + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2),$$

where $\lambda > 0$ is a given number. Also, $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ denotes the Frobenius norm of matrix $A = [a_{ij}]$.

Solution: Given

$$J(U,V) = \|UV - Y\|_F^2 + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)$$

Where $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{k\times n}$, $Y \in \mathbb{R}^{n\times n}$, $\lambda > 0$

As

$$\|A\|_F^2 = \mathrm{Tr}(A^\top A).$$

$\|UV - Y\|_F^2$ can be written as

$$\|UV - Y\|_F^2 = \mathrm{Tr}((UV - Y)^\top (UV - Y))$$

$$= \mathrm{Tr}((UV)^\top UV - 2Y^\top UV + Y^\top Y)$$

$$\|U\|_F^2 = \mathrm{Tr}(U^\top U) \quad \text{and} \quad \|V\|_F^2 = \mathrm{Tr}(V^\top V)$$

then

$$\frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) = \frac{\lambda}{2}\mathrm{Tr}(U^\top U) + \frac{\lambda}{2}\mathrm{Tr}(V^\top V)$$

Gradient with U

$$= \frac{\partial}{\partial U}\|UV - Y\|_F^2$$

$$\frac{\partial}{\partial U}\mathrm{Tr}((UV - Y)^\top (UV - Y)),$$

$$\frac{\partial}{\partial U}\|UV - Y\|_F^2 = \frac{\partial}{\partial U}\mathrm{Tr}((UV)^\top UV - 2Y^\top UV + Y^\top Y)$$

L.H.S

$$\mathrm{Tr}((UV)^\top UV) = \mathrm{Tr}(V^\top U^\top UV)$$

As $\frac{\partial}{\partial B}\text{Tr}(ABC) = CB^\top$

$$\frac{\partial}{\partial U}\text{Tr}(V^\top U^\top UV) = 2UVV^\top.$$

M.H.S

$$\text{Tr}(-2Y^\top UV) = -2\,\text{Tr}(Y^\top UV)$$

$$\frac{\partial}{\partial U}(-2\,\text{Tr}(Y^\top UV)) = -2YV^\top$$

R.H.S

$$\frac{\partial}{\partial U}\text{Tr}(Y^\top Y) = 0$$

So

$$\frac{\partial}{\partial U}\|UV - Y\|_F^2 = 2UVV^\top - 2YV^\top = 2(UV - Y)V^\top$$

Now talking gradient of $\frac{\lambda}{2}\|U\|_F^2$

$$\frac{\partial}{\partial U}\left(\frac{\lambda}{2}\|U\|_F^2\right) = \frac{\partial}{\partial U}\frac{\lambda}{2}\,\text{Tr}(U^\top U)$$

$$= \frac{\lambda}{2}\frac{\partial}{\partial U}\text{Tr}(U^\top U)$$

$$= \frac{\lambda}{2}2U = \lambda U$$

So

$$\frac{\partial J}{\partial U} = 2(UV - Y)V^\top + \lambda U$$

Gradient w.r.t V

$$\frac{\partial}{\partial V}\|UV - Y\|_F^2$$

As

$$\|UV - Y\|_F^2 = \text{Tr}((UV - Y)^\top (UV - Y))$$

$$= \frac{\partial}{\partial V}\text{Tr}((UV)^\top (UV) - 2Y^\top (UV) + Y^\top Y)$$

L.H.S

$$\frac{\partial}{\partial V}\text{Tr}((UV)^\top (UV)) = U(UV)^\top$$

M.H.S

$$\frac{\partial}{\partial V}(-2\text{Tr}(Y^\top (UV))) = -2U^\top Y$$

R.H.S

$$\frac{\partial}{\partial V}\text{Tr}(Y^T Y) = 0$$

So

$$\frac{\partial}{\partial V}\|UV - Y\|_F^2 = 2U^\top(UV - Y)$$

Now talking gradient of $\frac{\lambda}{2}\|V\|_F^2$

$$\frac{\partial}{\partial V}\left(\frac{\lambda}{2}\|V\|_F^2\right) = \frac{\partial}{\partial V}\frac{\lambda}{2}\operatorname{Tr}(V^\top V)$$

$$= \frac{\lambda}{2}\frac{\partial}{\partial V}\operatorname{Tr}(V^\top V)$$

$$= \frac{\lambda}{2}2V = \lambda V$$

Now the equation becomes

$$\frac{\partial J}{\partial V} = 2U^T(UV - Y) + \lambda V$$

The gradients $J(U, V)$ is

$$\nabla U = 2(UV - Y)V^\top + \lambda U$$
$$\nabla V = 2U^T(UV - Y) + \lambda V$$

2. (1 point) Compute the gradient of the following function:

$$f(w) = \sum_{i=1}^{m} \log(1 + e^{-y_i w^\top x_i}) + \frac{1}{2}\|w\|_2^2,$$

Solution: let take derivative of log term

$$\frac{\partial}{\partial w}\log(1 + e^{-y_i w^\top x_i}) = \frac{1}{1 + e^{-y_i w^\top x_i}} \cdot \frac{\partial}{\partial w}(1 + e^{-y_i w^\top x_i})$$

$$\frac{\partial}{\partial w}(1 + e^{-y_i w^\top x_i}) = 0 - y_i x_i e^{-y_i w^\top x_i} = -y_i x_i e^{-y_i w^\top x_i}$$

$$= \frac{1}{1 + e^{-y_i w^\top x_i}} \cdot \left(-y_i x_i e^{-y_i w^\top x_i}\right)$$

$$= \frac{-y_i x_i e^{-y_i w^\top x_i}}{1 + e^{-y_i w^\top x_i}}$$

Sigmoid Function

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

$$\frac{\partial}{\partial w}\log(1 + e^{-y_i w^\top x_i}) = -y_i x_i \sigma(-y_i w^\top x_i).$$

Putting in summation

$$\frac{\partial}{\partial w}\sum_{i=1}^{m}\log(1+e^{-y_i w^\top x_i}) = -\sum_{i=1}^{m} y_i x_i \sigma(-y_i w^\top x_i)$$

L.H.S As

$$\frac{1}{2}\|w\|_2^2 = \frac{1}{2}w^\top w$$

$$\frac{\partial}{\partial w}\frac{1}{2}\|w\|_2^2 = \frac{\partial}{\partial w}\frac{1}{2}w^\top w$$
$$= \frac{1}{2}\frac{\partial}{\partial w}w^\top w$$
$$\partial(< w, w >) =< \partial w, w > + < w, \partial w >$$

$$=< w, \partial w > + < w, \partial w >$$
$$=< 2w, \partial w >$$

$$\nabla w = 2w$$

So equation becomes

$$\frac{1}{2}\cdot 2w = w$$

So final answer is

$$\nabla f(w) = -\sum_{i=1}^{m} y_i x_i \sigma(-y_i w^\top x_i) + w.$$

3. (1 point) Consider the function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$f(w)_j = \frac{e^{w_j}}{\sum_{k=1}^{n} e^{w_k}}.$$

Compute the Jacobian matrix $J_f(w)$ of the function $f(w)$. The Jacobian matrix is defined as

$$J_f(w)_{ij} = \frac{\partial f_i}{\partial w_j}.$$

Also, discuss how to compute this function in a stable manner if one of the elements in w is large.

Solution:

$$f(w)_j = \frac{e^{w_j}}{\sum_{k=1}^{n} e^{w_k}}$$

Let A = $\sum_{k=1}^{n} e^{w_k}$

$$f(w)_j = \frac{e^{w_j}}{A} \tag{1}$$

As we need to take derivative w.r.t to each value in w_j.let i be an iterative value for w_j So there are two cases when

1. $w_i = w_j$
2. $w_i \neq w_j$

For $i = j$:

$$\frac{\partial f(w)_j}{\partial w_j} = \frac{\partial}{\partial w_j}\left(\frac{e^{w_j}}{A}\right).$$
$$= \frac{e^{w_j} \cdot A - e^{w_j} \cdot e^{w_j}}{A^2}$$
$$= \frac{e^{w_j}(A - e^{w_j})}{A^2}$$

As from (1) $e^{w_j} = f(w)_j \cdot A$

$$= \frac{f(w)_j \cdot A \cdot (A - e^{w_j})}{A^2}$$

Also $(A - e^{w_j}) = A - Af(w)_j = A(1 - f(w)_j)$

$$= \frac{f(w)_j \cdot A \cdot (A(1 - f(w)_j))}{A^2}$$

$$= \frac{f(w)_j \cdot A^2 \cdot (1 - f(w)_j))}{A^2}$$

So

$$\frac{\partial f(w)_j}{\partial w_j} = f(w)_j(1 - f(w)_j) \tag{2}$$

For $i \neq j$

$$\frac{\partial f(w)_j}{\partial w_i} = \frac{\partial}{\partial w_i}\left(\frac{e^{w_j}}{A}\right).$$

Using the quotient rule again,

$$\frac{\partial f(w)_j}{\partial w_i} = \frac{0 \cdot A - e^{w_j} \cdot e^{w_i}}{A^2}$$
$$= -\frac{e^{w_j} e^{w_i}}{A^2}$$
$$= -f(w)_j f(w)_i$$
$$J(w_i, w_j) = \begin{cases} f(w)_j(1 - f(w)_j), & \text{if } i = j, \\ -f(w)_i f(w)_j, & \text{if } i \neq j. \end{cases}$$

When components of w are large, computing e^{w_j} can lead to overflow.It means the larger component will have more wightage in overall Jacobian function and the Summation component can become numerically unstable .We can avoid this my subtracting the largest component from each value in w

$$f(w)_j = \frac{e^{w_j - max(w)}}{\sum_{k=1}^{n} e^{w_k}}$$

4. (0.5 points) Compute the gradient of the following functions with respect to matrix X:
   (a) $f(X) = \sum_{i=1}^{n} \lambda_i(X)$
   (b) $f(X) = \prod_{i=1}^{n} \lambda_i(X)$,
   where $\lambda_i(X)$ is the i-th eigenvalue of matrix X.

   1. $f(X) = \sum_{i=1}^{n} \lambda_i(X)$

Solution:
As sum of eigenvalues is known as trace

$$f(X) = \text{tr}(X) = \sum_{i=1}^{n} \lambda_i(X)$$

$$\partial f(X) = \partial \, \text{tr}(X)$$

$$\partial f(X) = \text{tr}(\partial X)$$

$$\partial f(X) \approx \langle \nabla f(X), \partial X \rangle$$
$$\langle A, B \rangle = \text{tr}(A^T B)$$

$$\partial f(X) = \text{tr}(\partial X)$$

$$\partial f(X) = \langle I, \partial X \rangle$$

$$\nabla f(X) = I$$

   2. $f(X) = \prod_{i=1}^{n} \lambda_i(X)$

Solution:
As multiples of eigen value is determinent of a matrix so

$$f(X) = \prod_{i=1}^{n} \lambda_i(X) = \det(X)$$

As determinet of $d(det(x)) = det(x)tr(x^{-1}dx)$

$$\partial f(X) = \det(X) \, \text{tr}(X^{-1} \partial X)$$

$$\partial f(X) \approx \langle \nabla f(X), \partial X \rangle$$

$$= \det(X) \, \text{tr}(X^{-1} \partial X)$$

$$\partial f(X) = \langle \det(X)(X^{-1})^T, \partial X \rangle$$

$$\nabla f(X) = \det(X) \cdot (X^{-1})^T$$