70007: Computational Optimisation
Coursework 1

Author: Ding Ke (**kd120**)

CATE Disclosure Date: 18:00 22-01-2021
Due on CATE: 19:00 12-02-2021

1 Part 1

1.1

Prove Log-Sum-Exp $\left(\log \sum\limits_{k=1}^{10} \exp\left(B_{j\,k}\right)\right)$ *is convex.*

The strategy for the proof is using the second derivative **sufficient** condition for convex functions: that is if the Hessian is proved to be always positive semi-definite on $\mathbb{R}^n$, then the function is said to be convex.

Let the function of interest be denoted as $f(\boldsymbol{B}) = \log \sum\limits_{k=1}^{10} \exp\left(B_k\right)$, so that now the Hessian can be expressed as:

$$\begin{aligned} H(\boldsymbol{B}) &= \nabla^2 f(\boldsymbol{B}) \\ &= \frac{\partial^2}{\partial B_p B_q} f(\boldsymbol{B}) \end{aligned}$$

Where indices of Hessian matrix are $\begin{cases} p = 1...10 \\ q = 1...10 \end{cases}$, and Hessian matrix $H(\boldsymbol{B}) \in \mathbb{R}^{p \times q}$.

Write the problem as $\frac{\partial}{\partial B_p} \frac{\partial \log \sum\limits_{k=1}^{10} \exp(B_k)}{\partial B_q}$ and deal with the first derivative part

$\frac{\partial \log \sum\limits_{k=1}^{10} \exp(B_k)}{\partial B_q}$ first, using chain rule:

$$f' = \frac{\partial}{\partial B_q} \log \sum_{k=1}^{10} \exp\left(B_k\right) \tag{1.1}$$

$$= \left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)} \right) \cdot \left(\frac{\partial \sum\limits_{k=1}^{10} \exp\left(B_k\right)}{\partial B_q} \right) \tag{1.2}$$

One can observe that $\frac{\partial \sum\limits_{k=1}^{10} \exp(B_k)}{\partial B_q} = \begin{cases} \exp\left(B_q\right) & \text{when k = q} \\ 0 & \textit{otherwise} \end{cases}$, therefore equation (1.2) becomes:

$$\frac{\partial}{\partial B_q} \log \sum_{k=1}^{10} \exp\left(B_k\right) = \left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)} \right) \cdot \exp\left(B_q\right) \tag{1.3}$$

Now elements of the Hessian can be expressed with the index notation:

$$H_{pq} = \frac{\partial}{\partial B_p} \left(\left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)} \right) \cdot \exp\left(B_q\right) \right) \tag{1.4}$$

To continue with the second derivative computation of Hessian, there are two separate cases of consideration: (1) when $p = q$, i.e. all elements along the matrix diagonal; (2) when $p \neq q$, i.e. all the other elements.

For the diagonal scenario($p = q$), equation (1.4) becomes:

$$H_{pp} = \frac{\partial}{\partial B_p} \left(\left(\sum_{k=1}^{10} \exp\left(B_k\right) \right)^{-1} \cdot \exp\left(B_p\right) \right) \tag{1.5}$$

And by product rule:

$$H_{pp} = -\left(\sum_{k=1}^{10} \exp\left(B_k\right) \right)^{-2} \cdot \left(\frac{\partial \sum\limits_{k=1}^{10} \exp\left(B_k\right)}{\partial B_p} \right) \cdot \exp\left(B_p\right) + \left(\sum_{k=1}^{10} \exp\left(B_k\right) \right)^{-1} \cdot \left(\frac{\partial \exp\left(B_p\right)}{\partial B_p} \right) \tag{1.6}$$

Observe that $\frac{\partial \sum\limits_{k=1}^{10} \exp(B_k)}{\partial B_p} = \begin{cases} \exp\left(B_p\right) & \text{when k = p} \\ 0 & otherwise \end{cases}$, and $\frac{\partial \exp(B_p)}{\partial B_p} = \exp\left(B_p\right)$, therefore equation (1.6) becomes:

$$H_{p\,p} = \left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)}\right) \cdot \exp\left(B_p\right) - \left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)}\right)^2 \cdot \exp\left(2B_p\right) \qquad (1.7)$$

Next for the scenario where $p \neq q$, compute from equation (1.4), again using product rule:

$$H_{p\,q} = \left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)}\right) \cdot \left(\frac{\partial \exp\left(B_q\right)}{\partial B_p}\right) + \left(-\left(\sum_{k=1}^{10} \exp\left(B_k\right)\right)^{-2} \cdot \left(\frac{\partial \sum\limits_{k=1}^{10} \exp\left(B_k\right)}{\partial B_p}\right) \cdot \exp\left(B_q\right)\right) \qquad (1.8)$$

Observe that $\frac{\partial \exp(B_q)}{\partial B_p} = 0$, and from previous result of $\frac{\partial \sum\limits_{k=1}^{10} \exp(B_k)}{\partial B_p} = \exp(B_p)$ therefore equation (1.8) becomes:

$$H_{p\,q} = -\left(\frac{1}{\sum\limits_{k=1}^{10} \exp\left(B_k\right)}\right)^2 \cdot \exp\left(B_p\right) \cdot \exp\left(B_q\right) \qquad (1.9)$$

Now, from both equation (1.7) and (1.9), the Hessian can be fully expanded as:

$$H(\boldsymbol{B}) = \begin{bmatrix} \frac{\exp(B_1)}{\sum\limits_{k=1}^{10} \exp(B_k)} - \frac{\exp(2B_1)}{\left(\sum\limits_{k=1}^{10} \exp(B_k)\right)^2} & \cdots & -\frac{\exp(B_1)\exp(B_{10})}{\left(\sum\limits_{k=1}^{10} \exp(B_k)\right)^2} \\ \cdots & \cdots & \cdots \\ -\frac{\exp(B_{10})\exp(B_1)}{\left(\sum\limits_{k=1}^{10} \exp(B_k)\right)^2} & \cdots & \frac{\exp(B_{10})}{\sum\limits_{k=1}^{10} \exp(B_k)} - \frac{\exp(2B_{10})}{\left(\sum\limits_{k=1}^{10} \exp(B_k)\right)^2} \end{bmatrix}$$

Let a vector $\boldsymbol{z} \in \mathbb{R}^{10\times 1}$ be defined as $\boldsymbol{z} = [\exp(B_1), \exp(B_2), \exp(B_3), \cdots, \exp(B_{10})]^T$, thus $1^T\boldsymbol{z} = \sum\limits_{k=1}^{10} \exp\left(B_k\right)$. Observing carefully the above matrix structure, one can quickly notice that the Hessian $H(\boldsymbol{B})$ can be expressed in matrix notation as:

$$H(\boldsymbol{B}) = \frac{1}{1^T\boldsymbol{z}} diag(\boldsymbol{z}) - \frac{1}{(1^T\boldsymbol{z})^2}\boldsymbol{z}\boldsymbol{z}^T \tag{1.10}$$

Where the `diag()` function maps elements of a vector into the diagonal of a matrix(corresponding to the size of the vector).

To see if the Hessian is p.s.d., one need to show that $\boldsymbol{v}^T H(\boldsymbol{B})\boldsymbol{v} \geq 0$, for all $\boldsymbol{v}$:

$$\begin{align} \boldsymbol{v}^T H(\boldsymbol{B})\boldsymbol{v} &= \frac{1}{1^T\boldsymbol{z}}\boldsymbol{v}^T diag(\boldsymbol{z})\boldsymbol{v} - \frac{1}{(1^T\boldsymbol{z})^2}\boldsymbol{v}^T\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{v} \tag{1.11} \\ &= \frac{1^T\boldsymbol{z}}{(1^T\boldsymbol{z})^2}\boldsymbol{v}^T diag(\boldsymbol{z})\boldsymbol{v} - \frac{1}{(1^T\boldsymbol{z})^2}\left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right) \tag{1.12} \\ &= \frac{\left(1^T\boldsymbol{z}\right)\boldsymbol{v}^T diag(\boldsymbol{z})\boldsymbol{v} - \left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right)}{(1^T\boldsymbol{z})^2} \tag{1.13} \end{align}$$

Now focusing on the $\left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right)$ term of equation (1.13), use Cauchy-Schwarz inequality for $\langle\boldsymbol{z}^T\boldsymbol{v}, \boldsymbol{z}^T\boldsymbol{v}\rangle$:

$$\left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right) \leq \|\boldsymbol{z}^T\boldsymbol{v}\|_2\|\boldsymbol{z}^T\boldsymbol{v}\|_2 \tag{1.14}$$

And by the definition of L2 norm, also realising $\boldsymbol{z}^T\boldsymbol{v} = \sum_{k=1}^{10} z_k v_k$ is a real value where $z_k = exp(B_k)$:

$$\begin{align} \left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right) &\leq \sqrt{\sum_{k=1}^{10}(z_k v_k)^2}\sqrt{\sum_{k=1}^{10}(z_k v_k)^2} \tag{1.15} \\ \left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right) &\leq \sum_{k=1}^{10}(z_k v_k)^2 \tag{1.16} \\ \left(\boldsymbol{z}^T\boldsymbol{v}\right)^T\left(\boldsymbol{z}^T\boldsymbol{v}\right) &\leq \sum_{k=1}^{10} z_k \sum_{k=1}^{10} z_k v_k^2 \tag{1.17} \end{align}$$

Writing everything in equation (1.13) and (1.17) in index notation, equation (1.13) becomes:

$$\boldsymbol{v}^T H(\boldsymbol{B})\boldsymbol{v} = \frac{\left(\sum\limits_{k=1}^{10} z_k\right)\left(\sum\limits_{k=1}^{10} z_k v_k^2\right) - \left(\sum\limits_{k=1}^{10} z_k v_k\right)\left(\sum\limits_{k=1}^{10} z_k v_k\right)}{\left(\sum\limits_{k=1}^{10} z_k\right)^2} \tag{1.18}$$

And equation (1.17) in full index notation:

$$\left(\sum_{k=1}^{10} z_k v_k\right)\left(\sum_{k=1}^{10} z_k v_k\right) \leq \left(\sum_{k=1}^{10} z_k\right)\left(\sum_{k=1}^{10} z_k v_k^2\right) \tag{1.19}$$

Substituting equation (1.19) into (1.18) shows that $\boldsymbol{v}^T H(\boldsymbol{B})\boldsymbol{v} \geq 0$ for any $\boldsymbol{v}$, therefore the Hessian $H(\boldsymbol{B})$ is p.s.d. everywhere on $\mathbb{R}^n$, therefore the `Log-Sum-Exp` function is convex.

1.2

Prove composition of a convex function with an affine mapping $\left(\log \sum\limits_{k=1}^{10} \exp\left(\boldsymbol{x_i}^\top \boldsymbol{\beta_k}\right)\right)$ *is convex.*

The strategy for the proof is using the definition of convex functions.

The composition of `Log-Sum-Exp` function $f(\boldsymbol{B})$ with an affine mapping can be expressed as a new function $g(\boldsymbol{B}) = f(\boldsymbol{x}^T\boldsymbol{B})$, the goal is prove $g(\boldsymbol{B})$ to be convex, given that convexity of `Log-Sum-Exp` function $f(\boldsymbol{B})$ has been proven in the previous section.

First try to create a linear combination structure for $g(\boldsymbol{B})$, let $\boldsymbol{B} = \alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2$, where $\forall\alpha \in (0,1)$. And now by definition of $g(\boldsymbol{B})$:

$$\begin{aligned} g\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) &= f\left(\boldsymbol{x}^T\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right)\right) && (1.20)\\ &= f\left(\alpha\left(\boldsymbol{x}^T\boldsymbol{B}_1\right) + (1-\alpha)\left(\boldsymbol{x}^T\boldsymbol{B}_2\right)\right) && (1.21)\end{aligned}$$

Note that the α term is just a real-valued coefficient, so it can be moved around and outside the brackets.

To prove the convexity of $f(\boldsymbol{x}^T\boldsymbol{B})$ given convexity of $f(\boldsymbol{B})$, first by definition of convexity of `Log-Sum-Exp` function $f(\boldsymbol{B})$:

$$f\left(\alpha\boldsymbol{B}_1+(1-\alpha)\boldsymbol{B}_2\right)\leq\alpha f\left(\boldsymbol{B}_1\right)+(1-\alpha)\,f\left(\boldsymbol{B}_2\right)\tag{1.22}$$

From the above equation, and by definition of the affine mapping $\boldsymbol{x}$:

$$\boldsymbol{x}^T f\left(\alpha\boldsymbol{B}_1+(1-\alpha)\boldsymbol{B}_2\right)\leq\alpha\boldsymbol{x}^T f\left(\boldsymbol{B}_1\right)+(1-\alpha)\,\boldsymbol{x}^T f\left(\boldsymbol{B}_2\right)\tag{1.23}$$

$$f\left(\alpha\left(\boldsymbol{x}^T\boldsymbol{B}_1\right)+(1-\alpha)\left(\boldsymbol{x}^T\boldsymbol{B}_2\right)\right)\leq\alpha f\left(\boldsymbol{x}^T\boldsymbol{B}_1\right)+(1-\alpha)\,f\left(\boldsymbol{x}^T\boldsymbol{B}_2\right)\tag{1.24}$$

Thus the above inequality confirms the convexity of $f(\boldsymbol{x}^T\boldsymbol{B})$, and by definition of function $g(\boldsymbol{B})$ again:

$$g\left(\alpha\boldsymbol{B}_1+(1-\alpha)\boldsymbol{B}_2\right)=f\left(\alpha\left(\boldsymbol{x}^T\boldsymbol{B}_1\right)+(1-\alpha)\left(\boldsymbol{x}^T\boldsymbol{B}_2\right)\right)\leq\alpha f\left(\boldsymbol{x}^T\boldsymbol{B}_1\right)+(1-\alpha)\,f\left(\boldsymbol{x}^T\boldsymbol{B}_2\right)=\alpha g\left(\boldsymbol{B}_1\right)+(1-\alpha)g\left(\boldsymbol{B}_2\right)$$

The above inequality is the exactly the definition of convexity for $g(\boldsymbol{B})$, therefore $g(\boldsymbol{B})$ as an composition of the convex $f(\boldsymbol{B})$ with an affine mapping is also convex.

1.3

Prove affine functions $\left(-\boldsymbol{x_i}^\top\boldsymbol{\beta_{y_i+1}}\right)$ *are convex.*

The strategy for the proof is also using the definition of convex function.

The affine function of interest can be expressed as $f(\boldsymbol{B})=-\boldsymbol{x}^T\boldsymbol{B}$. Now as before, to create a linear combination structure, let $\boldsymbol{B}=\alpha\boldsymbol{B}_1+(1-\alpha)\boldsymbol{B}_2,\forall\alpha\in(0,1)$, and by definition of $f(\boldsymbol{B})$:

$$\begin{aligned}f\left(\alpha\boldsymbol{B}_1+(1-\alpha)\boldsymbol{B}_2\right)&=-\boldsymbol{x}^T\left(\alpha\boldsymbol{B}_1+(1-\alpha)\boldsymbol{B}_2\right)&(1.25)\\&=\alpha\left(-\boldsymbol{x}^T\boldsymbol{B}_1\right)+(1-\alpha)\left(-\boldsymbol{x}^T\boldsymbol{B}_2\right)&(1.26)\\&=\alpha f(\boldsymbol{B}_1)+(1-\alpha)f(\boldsymbol{B}_2)&(1.27)\end{aligned}$$

The above equation shows that affine function $f(\boldsymbol{B})$ is both convex and concave, therefore the affine function is convex.

1.4

Prove that ℓ_1 *Regularisation* $\|\boldsymbol{\beta_k}\|_1$ *is convex.*

The strategy of the proof is using definition of convex function through the triangular inequality.

First prove the triangular inequality holds for L1 norm like $\|\boldsymbol{B}\|_1$, taking square of L1 norm, assuming $\langle , \rangle$ is the inner product between two vectors and it is symmetric:

$$\begin{aligned}
\|\boldsymbol{A}+\boldsymbol{B}\|_1^2 &= \langle \boldsymbol{A}+\boldsymbol{B}, \boldsymbol{A}+\boldsymbol{B}\rangle && (1.28)\\
&= \langle \boldsymbol{A}, \boldsymbol{A}\rangle + \langle \boldsymbol{B}, \boldsymbol{B}\rangle + 2\langle \boldsymbol{A}, \boldsymbol{B}\rangle && (1.29)\\
&= \|\boldsymbol{A}\|_1^2 + \|\boldsymbol{B}\|_1^2 + 2\langle \boldsymbol{A}, \boldsymbol{B}\rangle && (1.30)
\end{aligned}$$

Now use Cauchy-Schwarz inequality on $\langle \boldsymbol{A}, \boldsymbol{B}\rangle$:

$$\begin{aligned}
\|\boldsymbol{A}+\boldsymbol{B}\|_1^2 = \|\boldsymbol{A}\|_1^2 + \|\boldsymbol{B}\|_1^2 + 2\langle \boldsymbol{A}, \boldsymbol{B}\rangle \leq \|\boldsymbol{A}\|_1^2 + \|\boldsymbol{B}\|_1^2 + 2\|\boldsymbol{A}\|_1\|\boldsymbol{B}\|_1 && (1.31)\\
\|\boldsymbol{A}+\boldsymbol{B}\|_1^2 \leq (\|\boldsymbol{A}\|_1 + \|\boldsymbol{B}\|_1)^2 && (1.32)\\
\|\boldsymbol{A}+\boldsymbol{B}\|_1 \leq \|\boldsymbol{A}\|_1 + \|\boldsymbol{B}\|_1 && (1.33)
\end{aligned}$$

Now, creating a linear combination structure $(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2))$ in the L1-norm and one can exploit the above proven triangular inequality, let $\boldsymbol{A} = \alpha\boldsymbol{B}_1$ and $\boldsymbol{B} = (1-\alpha)\boldsymbol{B}_2$, $\forall\alpha \in (0,1)$:

$$\begin{aligned}
\|\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\|_1 \leq \|\alpha\boldsymbol{B}_1\|_1 + \|(1-\alpha)\boldsymbol{B}_2\|_1 && (1.34)\\
\|\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\|_1 \leq \alpha\|\boldsymbol{B}_1\|_1 + (1-\alpha)\|\boldsymbol{B}_2\|_1 && (1.35)
\end{aligned}$$

The above inequality is the exact definition of convexity for L1-norm $\|\cdot\|_1$, therefore the ℓ_1 Regularisation is convex.

1.5

Prove that the entire optimisation problem is convex.

The strategy of proof is decomposing the problem into sub-parts and using the definition of convex function where necessary.

The problem $\min\limits_{\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_{10}}} g(\boldsymbol{B}) + h(\boldsymbol{B}) = \min\limits_{\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_{10}}} \sum\limits_{i=1}^{m}\left(\log\sum\limits_{k=1}^{10}\exp\left(\boldsymbol{x_i}^\top\boldsymbol{\beta_k}\right) - \boldsymbol{x_i}^\top\boldsymbol{\beta_{y_i+1}}\right) + \lambda\sum\limits_{k=1}^{10}\|\boldsymbol{\beta_k}\|_1$ can be broken up into the following parts: (1) let $g_1(\boldsymbol{B}) = \left(\log\sum\limits_{k=1}^{10}\exp\left(\boldsymbol{x_i}^\top\boldsymbol{\beta_k}\right) - \boldsymbol{x_i}^\top\boldsymbol{\beta_{y_i+1}}\right)$; (2) let $h_1(\boldsymbol{B}) = \sum\limits_{k=1}^{10}\|\boldsymbol{\beta_k}\|_1$. We first deal with the $g_1(\boldsymbol{B})$ part first, recognising immediately that $g_1(\boldsymbol{B}) = g_2(\boldsymbol{B}) + g_3(\boldsymbol{B})$ where $g_2(\boldsymbol{B})$ is the `log-sum-exp` whose convexity is proven in section 1.1 and 1.2; the affine function $g_3(\boldsymbol{B}) = -\boldsymbol{x_i}^T \boldsymbol{\beta_{y_i+1}}$ is proven to be convex in section 1.3.

Now what is needed is a proof saying that the unweighted sum of convex functions is still convex. First by definition of g_1 the sum of two convex functions:

$$g_1\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) = g_2\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) + g_3\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) \tag{1.36}$$

By definition of convexity of both g_2 and g_3:

$$g_2\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) \leq \alpha g_2(\boldsymbol{B}_1) + (1-\alpha)g_2(\boldsymbol{B}_2) \tag{1.37}$$

$$g_3\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) \leq \alpha g_3(\boldsymbol{B}_1) + (1-\alpha)g_3(\boldsymbol{B}_2) \tag{1.38}$$

Now add up the above two inequalities, and given that $\forall\alpha \in (0,1)$:

$$g_2\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) + g_3\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) \leq \alpha\left(g_2(\boldsymbol{B}_1) + g_3(\boldsymbol{B}_1)\right) + (1-\alpha)\left(g_2(\boldsymbol{B}_2) + g_3(\boldsymbol{B}_2)\right) \tag{1.39}$$

$$g_1\left(\alpha\boldsymbol{B}_1 + (1-\alpha)\boldsymbol{B}_2\right) \leq \alpha\left(g_1(\boldsymbol{B}_1)\right) + (1-\alpha)\left(g_1(\boldsymbol{B}_2)\right) \tag{1.40}$$

The above inequality shows exactly the convexity of g_1, therefore the unweighted sum of convex functions is convex. Thus $g_1(\boldsymbol{B}) = g_2(\boldsymbol{B}) + g_3(\boldsymbol{B})$ is convex.

Next we deal with the $h_1(\boldsymbol{B}) = \sum\limits_{k=1}^{10} \|\boldsymbol{\beta_k}\|_1$ part. As the L1 norm $\|\boldsymbol{\beta_k}\|_1$ is proven to convex in section 1.4, and we have just shown from above that sum of convex functions is convex, therefore $h_1(\boldsymbol{B})$ being the sum of 10 L1 norms is convex.

Similarly, $g(\boldsymbol{B}) = \sum\limits_{i=1}^{m} g_1(\boldsymbol{B})$ is the sum of m convex functions $g_1(\boldsymbol{B})$, thus $g(\boldsymbol{B})$ is also convex.

Now, this problem has been reduced into $\min\limits_{\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_{10}}} g(\boldsymbol{B}) + h(\boldsymbol{B}) = \min\limits_{\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_{10}}} g(\boldsymbol{B}) + \lambda h_1(\boldsymbol{B})$. For this is a positively weighted sum of convex functions $g(\boldsymbol{B})$ and $h_1(\boldsymbol{B})$, where $\lambda \in \mathbb{R}, \lambda > 0$, the proof of convexity follows the same procedure as the unweighted sum one, except for the adding of the convex definition inequality of $g(\boldsymbol{B})$ and λ times the convex definition inequality of $h_1(\boldsymbol{B})$. Now as the weight λ is a real-valued positive number, the sum of these inequalities still preserves the inequality, therefore the sum $g(\boldsymbol{B}) + h(\boldsymbol{B}) = g(\boldsymbol{B}) + \lambda h_1(\boldsymbol{B})$ is convex. And thus the whole problem is convex.

2 Part 2

2.1

Prove that $h(\boldsymbol{B}) = \lambda \sum_{k=1}^{10} \|\boldsymbol{\beta_k}\|_1 \notin C^1$.

The strategy of proof is decomposing the function into the smallest part which is a single absolute function and prove that it is not continuously differentiable

The ℓ_1 Regularisation $\|\boldsymbol{\beta_k}\|_1$ can be decomposed as the sum of the absoulte values of the β_k recall the definition of L1 norm.

$$h(\boldsymbol{B}) = \lambda \sum_{k=1}^{10} (|\beta_{k1}| + |\beta_{k2}| + \cdots + |\beta_{kn}|) \tag{2.1}$$

Let $f(\beta) = |\beta|$. Taking partial derivatives will end in adding up several derivatives of single absolute functions

$$\nabla h(\boldsymbol{B}) = \lambda \sum_{k=1}^{10} ([f'(\beta_{k1}) \quad f'(\beta_{k2}) \quad \cdots \quad f'(\beta_{kn})]) \tag{2.2}$$

Now the question becomes proving whether the derivative of the absolute function $f'(\beta)$ is continuously differentiable. Using the basic definition of the derivative, it can be found that at the point $\beta = 0$, the derivative obtained from the left is different from the derivative obtained from the right. In other words, the function is not differentiable at the point $\beta = 0$:

$$\underset{0^- \ \to \ 0}{lim} \frac{[f(0+\Delta\beta) - f(0)]}{\Delta\beta} = \frac{[0 - \Delta\beta - 0]}{\Delta\beta} = \frac{-\Delta\beta}{\Delta\beta} = -1 \tag{2.3}$$

$$\underset{0^+ \ \to \ 0}{lim} \frac{[f(0+\Delta\beta) - f(0)]}{\Delta\beta} = \frac{[0 + \Delta\beta - 0]}{\Delta\beta} = \frac{\Delta\beta}{\Delta\beta} = 1 \tag{2.4}$$

Since some part of the function is not differentiable, clearly the original function $h(\boldsymbol{B})$ is not once continuously differentiable.

2.2

Prove that the new optimisation problem(using ℓ_2 regularisation instead) is convex and once continuously differentiable.

The strategy of proof is similar to section 1.5. This conclusion can be proved by decomposing the problem into sub-parts and using the Hessian matrix to prove the convexity when needed.

The problem $\min_{\boldsymbol{\beta_1},\dots,\boldsymbol{\beta_{10}}} f(\boldsymbol{B}) = \min_{\boldsymbol{\beta_1},\dots,\boldsymbol{\beta_{10}}} \sum_{i=1}^{m} \left(\log \sum_{k=1}^{10} \exp\left(\boldsymbol{x_i}^\top \boldsymbol{\beta_k}\right) - \boldsymbol{x_i}^\top \boldsymbol{\beta_{y_i+1}} \right) + \lambda \sum_{k=1}^{10} \|\boldsymbol{\beta_k}\|_2^2$ again can be broken up into the following two parts: (1) let $g(\boldsymbol{B}) = \left(\log \sum_{k=1}^{10} \exp\left(\boldsymbol{x_i}^\top \boldsymbol{\beta_k}\right) - \boldsymbol{x_i}^\top \boldsymbol{\beta_{y_i+1}} \right)$; (2) let $h(\boldsymbol{B}) = \sum_{k=1}^{10} \|\boldsymbol{\beta_k}\|_2^2$.

For the convexity part of the problem, recall some conclusions from section 1.5, the first part $g(\boldsymbol{B})$ is convex and the positively weighted sum of two convex functions is convex. Thus, the problem becomes proving the convexity of the second part$h(\boldsymbol{B})$.

$$h(\boldsymbol{B}) = \sum_{k=1}^{10} \|\boldsymbol{\beta_k}\|_2^2 = \sum_{k=1}^{10} \left(\sqrt{\beta_{k1}^2 + \beta_{k2}^2 + \cdots + \beta_{kn}^2} \right)^2 = \sum_{k=1}^{10} \sum_{i=1}^{n} \beta_{ki}^2 \tag{2.5}$$

Let function $f_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} \beta_i^2$ and the function $h(\boldsymbol{B})$ now becomes $h(\boldsymbol{B}) = \sum_{k=1}^{10} f_1(\boldsymbol{\beta_k})$. The convexity of function $f_1(\boldsymbol{\beta})$ can be proved by calculating its Hessian matrix:

$$H(\boldsymbol{\beta}) = \nabla^2 f(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \beta_i \beta_j} f_1(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_i} \frac{\partial \sum_{i=1}^{n} \beta^2}{\partial \beta_i} \tag{2.6}$$

Where the indices of the Hessian matrix i, j $\in \{1, 2, \cdots, n\}$, and thus the Hessian matrix $H(\boldsymbol{\beta}) \in \mathbb{R}^{n \times n}$.

For the case when $i = j$, the diagonal values of the Hessian matrix can be calculated as below:

$$H_{ii} = \frac{\partial}{\partial \beta_i} 2\beta_i = 2 \tag{2.7}$$

For all the other normal cases when $i \neq j$, the values of the Hessian matrix are:

$$H_{ij} \frac{\partial}{\partial \beta_i} 2\beta_j = 0 \tag{2.8}$$

Therefore, the whole Hessian matrix of the function $f_1(\boldsymbol{\beta})$ can be fully expanded as below:

$$H(\boldsymbol{\beta}) = \begin{bmatrix} 2 & 0 & 0 & \cdots & 0 \\ 0 & 2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 2 & 0 \\ 0 & \cdots & 0 & 0 & 2 \end{bmatrix} = 2\mathbf{I}$$

Clearly that the Hessian matrix is positive definite $\forall \beta_{ki} \in R$. Therefore the function $f_1(\boldsymbol{\beta})$ is a convex function. It has been proven in section 1.5 that the unweighted sum of convex functions is convex. Applying this, the convexity of the second part function $h(\boldsymbol{B})$ can be proved. And thus the whole optimisation problem is convex.

For the continuously differentiability part, the strategy is to prove that the two parts are both continuously differentiable. There is no doubt that the sum of two continuously differentiable functions is continuously differentiable as well according to the sum rule of the derivative.

Firstly, we solve for the first part. From the giving information, the gradient of the function is given as:

$$\nabla_{\boldsymbol{B}} g(\boldsymbol{B}) = \boldsymbol{X}^\top (\boldsymbol{Z} \exp(\boldsymbol{X}\boldsymbol{B}) - \boldsymbol{Y}) \tag{2.9}$$

$$\boldsymbol{Z}_{ii} = \frac{1}{\sum\limits_{k=1}^{10} \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta_k})} \quad \forall i \in \{1, \cdots, m\} \tag{2.10}$$

According to the definition, the function is continuously differentiable if and only if there exists a real value derivative for all feasible values of the variable. Therefore, we need to prove that this gradient function is meaningful for all possible values of $\boldsymbol{B}$.

Checking the existence of the gradient is divided into two steps. Step 1, checking the dimensions of the matrixes. $\boldsymbol{Z} \in \mathbb{R}^{m\times m}, \boldsymbol{X} \in \mathbb{R}^{m\times n}\ and\ \boldsymbol{B} \in \mathbb{R}^{n\times 10}$. Multiplying these three matrices will result in a dimension of $\mathbb{R}^{m\times 10}$ which matches the dimension of $\boldsymbol{Y}$. The dimension of $\boldsymbol{X}^\top$ matches the matrixes in the bracket as well and this will give a final dimension of $\mathbb{R}^{n\times 10}$ which fits the size of $\boldsymbol{B}$. Step 2, checking whether these matrixes exist. According to the given information, the matrix $\boldsymbol{X}, \boldsymbol{B}\ and\ \boldsymbol{Y}$ will always exist and will have real values. As for the matrix $\boldsymbol{Z}$, since the function on the denominator is exponential and is larger than 0, the matrix $\boldsymbol{Z}$ will always exist and have real values as well. In conclusion, the gradient of function $g(\boldsymbol{B})$ will always be meaningful and thus this function is once continuously differentiable.

Secondly, we move on to the ℓ_2 regularisation function $h(\boldsymbol{B})$. Taking the derivative of the function gives:

$$h(\boldsymbol{B}) = \sum_{k=1}^{10} \left(\sqrt{\beta_{k1}^2 + \beta_{k2}^2 + \cdots + \beta_{kn}^2} \right)^2 = \sum_{k=1}^{10} \boldsymbol{\beta_k}^\top \boldsymbol{\beta_k} \tag{2.11}$$

$$\nabla_{\boldsymbol{B}} h(\boldsymbol{B}) = 2 \sum_{k=1}^{10} \boldsymbol{\beta_k}^\top \tag{2.12}$$

Considering that the vector $\boldsymbol{\beta_k}$ always exists, the gradient will then be meaningful for all the feasible values of $\boldsymbol{B}$

Finally, as stated above, the derivative of a complex function is equal to the sum of the derivatives of each individual part of the function according to the sum rule of the derivative. Since the general optimisation problem is the weighted sum of two parts, its derivative will simply be the weighted sum of the individual derivatives. Thus, the optimisation problem will be once continuously differentiable. In other words, $f(\boldsymbol{B}) \in C^1$

2.3

Show the three fixed lines of the code.

line 79: convgsd(i) = norm(beta_grad);
line 84: lenXsd(i) = norm(beta_guess_iter(i+1,:) - beta_guess_iter(i,:));
line 89: diffFsd(i) = abs(fcn_val_iter(i+1) - fcn_val_iter(i));

2.4

How do the tolerances in Part 2.3 correspond to the FONC, SONC, and SOSC? Show how each of Lines 79, 84, & 89 in SolveMNIST_Gradient now correspond to an optimality condition and state the relevant condition.

2.4.1

The first tolerance check $\|\nabla f(\boldsymbol{B}^{(j)})\|_2 < \epsilon$ corresponds to the optimality condition of FONC. Here is the proof.

As the termination tolerance is set $\epsilon = 1 \times 10^{-4}$ which is a number very close to 0. And the tolerance check $\|\nabla f(\boldsymbol{B}^{(j)})\|_2 < \epsilon$ essentially tries to terminate the program when a point $\boldsymbol{B}^*$ is reached such that the $\|\nabla f(\boldsymbol{B}^*)\|_2$ is sufficiently close to 0. Thus the intention of the first check can be formulated as:

$$\|\nabla f(\boldsymbol{B}^*)\|_2 = 0 \tag{2.13}$$

Assuming the Jacobian of $\boldsymbol{B}$ collapses into a vector so that only vector norms are under consideration here, and by definition of vector L2-norm:

$$\sqrt{\sum_{i=1}^{n} \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_i} \right)^2} = 0 \tag{2.14}$$

$$\sum_{i=1}^{n} \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_i} \right)^2 = 0 \tag{2.15}$$

$$\left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_1} \right)^2 + \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_2} \right)^2 + \cdots + \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_n} \right)^2 = 0 \tag{2.16}$$

As each of the element $\left(\frac{\partial f(\boldsymbol{B}^*)}{\partial \boldsymbol{B}_i} \right)^2 \geq 0$:

$$\left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_1} \right)^2 = \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_2} \right)^2 = \cdots = \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_n} \right)^2 = 0 \tag{2.17}$$

$$\left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_1} \right) = \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_2} \right) = \cdots = \left(\frac{\partial f\left(\boldsymbol{B}^*\right)}{\partial \boldsymbol{B}_n} \right) = 0 \tag{2.18}$$

With each element of the Jacobian being 0, therefore:

$$\nabla_{\boldsymbol{B}} f(\boldsymbol{B}^*) = 0 \tag{2.19}$$

Now from the above equation we have $\boldsymbol{d}^T \nabla f(\boldsymbol{B}^*) = 0$ at the local minimiser $\boldsymbol{B}^*$, for any direction vector $\boldsymbol{d}$. This is exactly the definition of FONC, thus the first tolerance check is indeed effectively the FONC.

2.4.2

Since the first tolerance check corresponds to the FONC, we assume that the second tolerance check $\|\boldsymbol{B}^{(j+1)} - \boldsymbol{B}^{(j)}\|_2 < \epsilon$ corresponds to the SONC or SOSC. In order to prove this, it is necessary to check whether the Hessian matrix is positive definite.

Once again, we break the optimisation problem $f(\boldsymbol{B})$ into two parts:(1) let $g(\boldsymbol{B}) = \left(\log \sum_{k=1}^{10} \exp\left(\boldsymbol{x_i}^\top \boldsymbol{\beta_k}\right) - \boldsymbol{x_i}^\top \boldsymbol{\beta_{y_i+1}}\right)$; (2) let $h(\boldsymbol{B}) = \sum_{k=1}^{10} \|\boldsymbol{\beta_k}\|_2^2$.

The gradient of function $g(\boldsymbol{B})$ is given as function 2.9 above. Using a slight different approach from pervious section, the gradient of the function $h(\boldsymbol{B})$ in a format of matrix can be obtained:

$$h(\boldsymbol{B}) = \sum_{k=1}^{10} \|\boldsymbol{\beta_k}\|_2^2 = \sum_{k=1}^{10} \left(\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}\right)^2 = \sum_{k=1}^{10} \boldsymbol{\beta}^\top \boldsymbol{\beta} = \boldsymbol{B}^\top \boldsymbol{B} \tag{2.20}$$

$$\nabla_{\boldsymbol{B}} h(\boldsymbol{B}) = \frac{\partial \boldsymbol{B}^\top \boldsymbol{B}}{\partial \boldsymbol{B}} = 2\boldsymbol{B} \tag{2.21}$$

The Hessian matrix for the can be calculated by taking the partial derivative of the gradient function:

$$\nabla^2_{\boldsymbol{B}} f(\boldsymbol{B}) = \nabla^2_{\boldsymbol{B}} g(\boldsymbol{B}) + \nabla^2_{\boldsymbol{B}} h(\boldsymbol{B}) = \frac{\partial \boldsymbol{X}^\top (\boldsymbol{Z} \exp(\boldsymbol{X}\boldsymbol{B}) - \boldsymbol{Y})}{\partial \boldsymbol{B}} + \frac{\partial 2\boldsymbol{B}}{\partial \boldsymbol{B}} \tag{2.22}$$

Now that the equation for the Hessian matrix is derived, it is necessary to find the relationship between the tolerance check and the Hessian matrix. As the tolerance is small, similar to section 2.4.1, the intention of the second check can also be formulated as:

$$\|\boldsymbol{B}^{(j+1)} - \boldsymbol{B}^{(j)}\|_2 = \|\Delta \boldsymbol{B}\|_2 = 0 \tag{2.23}$$

Following a similar procedure as the section 2.4.1, it can be proven that $\Delta \boldsymbol{B} = 0$. From the code and the giving information, the weight matrix $\boldsymbol{B}$ is trained using a gradient-based method. Thus the difference of the weight matrix between each training epoch can be generalised as below:

$$\boldsymbol{B^{(j+1)}} = \boldsymbol{B^{(j)}} - k_t \nabla f(\boldsymbol{B}) \tag{2.24}$$

$$\Delta \boldsymbol{B} = -k_t \nabla f(\boldsymbol{B}) \tag{2.25}$$

where k_t is user-defined learning rate. Combining the equation 2.23 and 2.25, it can be found that the second tolerance check will eventually lead to the same result, that is the gradient of the loss function becomes zero.

Combining all of the discussions above, it is found that the second tolerance check results in the same optimality condition. The second tolerance check is FONC as well.

2.4.3

Finally, it comes to the third tolerance check $|f(\boldsymbol{B^{(j+1)}}) - f(\boldsymbol{B^{(j)}}) < \epsilon|$. Since the Hessian matrix has already be calculated in section 2.4.2, we only need to find the relationship between the tolerance check and the optimality conditions.

The range of the difference of the loss when the terminate state reaches can be simply derived by removing the absolute function in the third check:

$$-\epsilon < f(\boldsymbol{B^{(j+1)}}) - f(\boldsymbol{B^{(j)}}) < \epsilon \tag{2.26}$$

Considering that the termination tolerance is set very close to 0, the intention of the third check can be then formulated as:

$$f(\boldsymbol{B^{(j+1)}}) - f(\boldsymbol{B^{(j)}}) = 0 \tag{2.27}$$

Let $\Delta\boldsymbol{B} = \boldsymbol{B^{(j+1)}}) - f(\boldsymbol{B^{(j)}}$, from the definition of the derivative:

$$\frac{df(\boldsymbol{B})}{d\boldsymbol{B}} = \underset{\Delta\boldsymbol{B} \to 0}{lim} \frac{f(\boldsymbol{B^{(j+1)}}) - f(\boldsymbol{B^{(j)}})}{\Delta\boldsymbol{B}} \tag{2.28}$$

It is clear that when the numerator $f(\boldsymbol{B}^{(j+1)}) - f(\boldsymbol{B}^{(j)})$ is zero, all the partial derivative values in the Jacobian will be zero, and thus $\nabla f(\boldsymbol{B}) = 0$. Now let $\boldsymbol{B}^j$ be the local minimiser such that $\boldsymbol{d}^T\nabla f(\boldsymbol{B}^{(j)}) = 0$, and $\boldsymbol{B}^{j+1} = \boldsymbol{B}^j + \alpha\boldsymbol{d}$. Use Taylor theorem to expand $f(\boldsymbol{B}^{j+1})$ as:

$$f(\boldsymbol{B}^{j+1}) = f(\boldsymbol{B}^j) + \alpha\nabla\boldsymbol{d}^T f(\boldsymbol{B}^j) + \frac{1}{2}(\alpha)^2\boldsymbol{d}^T\nabla^2 f(\boldsymbol{B}^j)\boldsymbol{d} \tag{2.29}$$

$$f(\boldsymbol{B}^{j+1}) = f(\boldsymbol{B}^j) + \frac{1}{2}(\alpha)^2\boldsymbol{d}^T\nabla^2 f(\boldsymbol{B}^j)\boldsymbol{d} \tag{2.30}$$

$$f(\boldsymbol{B}^{j+1}) - f(\boldsymbol{B}^j) = \frac{1}{2}(\alpha)^2\boldsymbol{d}^T\nabla^2 f(\boldsymbol{B}^j)\boldsymbol{d} \tag{2.31}$$

By the above equation and from the given tolerance check $|f(\boldsymbol{B^{(j+1)}}) - f(\boldsymbol{B^{(j)}}) < \epsilon|$, where ϵ is very close to 0, one can obtain the relation of $|\frac{1}{2}(\alpha)^2\boldsymbol{d}^T\nabla^2 f(\boldsymbol{B}^j)\boldsymbol{d}| = 0$, therefore $\boldsymbol{d}^T\nabla^2 f(\boldsymbol{B}^j)\boldsymbol{d} = 0$. And thus the SONC is proved.

From all of the discussions above, a conclusion can be drawn that the third tolerance check corresponds to the SONC.