Homework 1

A bonus of 20% of the full score will be awarded for the assignment submitted two days before the due date, November 28, 2023. The total score will not exceed 20 points.

1. (7 points)

   Suppose we have a set of distint data points $X = \{x_1, \dots, x_m\}$ in $\mathbb{R}^n$.

   (a) (2 points) Let x_c be the centroid of X, i.e., the average of all the data points. We want to find the point y in the convex hull H of X that is farthest from x_c.

   Formulate the optimization problem to find y and prove that it is a convex optimization problem.

   Prove that the optimal solution is a vertex of the convex hull H.

   Solution: The optimization problem to find y in the convex hull H of X that is farthest from x_c.

$$\min_{y} \quad \|y - x_c\|_2^2 \tag{1}$$

$$\text{subject to} \quad y \in conv(X) \tag{2}$$

   where $conv(X)$ is the convex hull of X. Maximization of a convex function $f(y) = \|y - x_c\|_2^2$ over a convex set $conv(X)$ is achieved at one of the vertices.

   Let $v_1, \dots, v_K$ be vertices of the convex hull of X, where K is the number of vertices. Denote by v_i the vertex farthest from x_c, i.e.,

$$f(v_i) = \max(f(v_1), \dots, f(v_K))$$

   or

$$v_i = \arg\max_{k=1,\dots,K} \|v_k - x_c\|_2^2.$$

   An arbitrary point, y, in the convex hull $conv(X)$ can be represented as convex combination of vertices v_k

$$y = \sum_{k=1}^{K} \alpha_k v_k \tag{3}$$

   where $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$.

Let $y^\star$ be the optimal point of the above optimization problem, and its convex combination $y^\star = \sum_{k=1}^{K} \alpha_k v_k$.

$$\begin{aligned} f(y^\star) &= f(\sum_{k=1}^{K} \alpha_k v_k) \\ &\leq \sum_{k=1}^{K} \alpha_k f(v_k) \\ &\leq \sum_{k=1}^{K} \alpha_k \max(f(v_k)) \\ &\leq f(v_i) \sum_{k=1}^{K} \alpha_k = f(v_i) \end{aligned}$$

This implies that $y^\star$ must be the vertex v_i.

(b) (1 point) Let u be a vector that is orthogonal to $(y - x_c)$, where y is the optimal solution from (a). We want to find the point x in the convex hull H that maximizes the inner product $(x - x_c)^T u$.

Formulate the optimization problem to find x and prove that the optimal solution is another vertex of the convex hull H.

Solution:

$$\begin{aligned} \max \quad & g(x) = (x - x_c)^T u && (4) \\ \text{s.t.} \quad & x \in conv(X) && (5) \end{aligned}$$

Similar to the task (a), since the objective function is linear and $conv(X)$ is a convex set, the optimal solution $x^\star$ is achieved at one of the vertices of $conv(X)$.

Since x_c is the centroid point of X, x_c is in the interior of the convexhull of X. Note that $g(x_c) = 0$. There must exist a direction d such that $u^T d > 0$, and a sufficiently small $\epsilon > 0$ such that $x = x_c + \epsilon d$ in $conv(X)$.

Since $g(x)$ is convex, by the first order convexity condition, we have

$$g(x) \geq g(x_c) + \nabla_x g(x)^T (x - x_c) = 0 + \epsilon u^T d > 0 \tag{6}$$

Since u is orthogonal to $(y - x_c)$

$$g(y^\star) = (y^\star - x_c)^T u = 0 < g(x_c + \epsilon d) \leq g(x^\star)$$

i.e., the optimal solution $x^\star$ must be another vertex which is different from the vertex $y^\star$.

(c) (2 points) Let $E_x = \{x \in \mathbb{R}^n : (x - c)^T Q (x - c) \leq n\}$ be the ellipsoid with the smallest volume that contains all the points in X.

Formulate the optimization problem to find E_x, and prove that the convex hull H of X is contained in E.

Solution: The optimization problem to find the ellipsoid E_x with the smallest volume that contains all the points in X

$$\begin{aligned} \min_{Q,c} \quad & -\log\det(Q) && (7)\\ \text{s.t.} \quad & (x_k - c)^T Q(x_k - c) \le n, \quad k = 1, \dots, m && (8)\\ & Q \in S^n_{++} && (9) \end{aligned}$$

An arbitrary point $x \in H$ can be represented as convex combination

$$x = \sum_{i=1}^{m} \alpha_i x_i$$

where $\alpha_i \ge 0$ and $\sum_i \alpha_i = 1$.

Since $Q \in S^n_{++}$, $f(x) = (x-c)^T Q(x-c)$ is convex, and $f(x_i) = (x_i - c)^T Q(x_i - c) \le n$.

$$\begin{aligned} f(x) = f(\sum_i \alpha_i x_i) &\le \sum_i \alpha_i f(x_i) && (10)\\ &\le \sum_i \alpha_i n = n && (11) \end{aligned}$$

i.e., x is also in the ellipsoid E_x.

(d) (1 point) Define $z_k = [x_k, 1]^T$, and $E_z = \{z \in \mathbb{R}^{n+1} : z^T P z \le n + 1\}$ the ellipsoid with smallest volume which contains all z_k.

Prove that

$$P = \begin{bmatrix} Q & -Qc \\ -(Qc)^T & 1 + c^T Qc \end{bmatrix}$$

Solution: For x in the ellipsoid E_x, we have

$$\begin{aligned} n + 1 &\ge (x-c)^T Q(x-c) + 1 \\ &= \begin{bmatrix} x \\ 1 \end{bmatrix}^T \begin{bmatrix} I & 0 \\ -c^T & 1 \end{bmatrix} \begin{bmatrix} Q & \\ & 1 \end{bmatrix} \begin{bmatrix} I & -c^T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} x \\ 1 \end{bmatrix}^T \begin{bmatrix} Q & -Qc \\ -c^T Q & 1 + c^T Qc \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} && (12) \end{aligned}$$

(e) (1 point) Show an example to illustrate finding the ellipsoid E_x from E_z.

2. Derive optimality conditions and analytical solution (if it exists) of the following optimization problems, solve them. Demonstrate the problem and compare the results with those obtained by CVX or CVXpy.

   $\boldsymbol{b} \in \mathbb{R}^n$ is vectorization of an image corrupted by Gaussian noise, $\mathbf{A}$ is of size $n \times m$, $m > n$, $\boldsymbol{x} \in \mathbb{R}^m$. $\mathbf{A}$ can be concatenation of discrete cosine transform matrix and wavelets transform operators.

(a) (2 points)

$$\min_{\boldsymbol{x}} \quad \|\boldsymbol{b}-\boldsymbol{x}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{x} \geq 0,\ \mathbf{1}^T\boldsymbol{x} = 1$$

Solution:

The Lagrangian function of the problem is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \nu) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{b}\|_2^2 - \boldsymbol{\lambda}^T\boldsymbol{x} + \nu(1-\mathbf{1}^T\boldsymbol{x})$$

where $\boldsymbol{\lambda} \geq 0$. The optimal solution of the problem must satisfy the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{x}} &= \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\lambda} - \mathbf{1}\nu = 0 && \text{(stationarity)}\\
\boldsymbol{\lambda} &\geq 0 && \text{(dual feasibility)}\\
\boldsymbol{x} &\geq 0, \quad \mathbf{1}^T\boldsymbol{x} = 1 && \text{(primal feasibility)}\\
\lambda_i x_i &= 0 \quad \forall i = 1, \ldots, n && \text{(complementary slackness)}
\end{aligned}$$

From the stationarity condition, we can obtain the optimal primal solution as

$$\boldsymbol{x}^\star = \boldsymbol{b} + \boldsymbol{\lambda} + \mathbf{1}\nu$$

Denote by $\mathcal{I} = \{i : x_i^\star > 0\}$ the index set of the positive components of $\boldsymbol{x}^\star$. By the complementarity condition, we have

$$\lambda_i = 0, \quad \forall i \in \mathcal{I}$$

Hence, we can simplify the expression of $\boldsymbol{x}^\star$ as

$$\boldsymbol{x}_{\mathcal{I}}^\star = \boldsymbol{b}_{\mathcal{I}} + \mathbf{1}\nu$$

where $\boldsymbol{b}_{\mathcal{I}} = [b_{i\in\mathcal{I}}]$ is the subvector of $\boldsymbol{b}$ corresponding to the indices in $\mathcal{I}$. Using the equality constraint, we can derive the optimal dual variable ν as

$$\nu = \frac{1-\mathbf{1}^T\boldsymbol{b}_{\mathcal{I}}}{|\mathcal{I}|}$$

where $|\mathcal{I}|$ is the cardinality of $\mathcal{I}$. In addition, since $\boldsymbol{x}^\star > 0$, we must have $b_i > -\nu$ for all $i \in \mathcal{I}$.

For $j \notin \mathcal{I}$, we have $x_j^\star = 0$. From the stationarity condition, we get

$$0 = x_j^\star = b_j + \lambda_j + \nu \geq b_j + \nu$$

which implies that $b_j \leq -\nu$ for all $j \notin \mathcal{I}$.

This means that there is a gap between the values of $\boldsymbol{b}$ that correspond to the positive and zero components of $\boldsymbol{x}^\star$, i.e.,

$$\min_{i\in\mathcal{I}}(b_i) > -\nu \geq \max_{j\notin\mathcal{I}}(b_j)$$

If we sort the elements of $\boldsymbol{b}$ in a descending order, we have:

$$b_1 \geq \cdots \geq b_I > -\nu \geq b_{I+1} \geq \cdots \geq b_n$$

(b) (2 points)

$$\min_{\boldsymbol{x}} \quad \|\boldsymbol{b}-\boldsymbol{x}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{x} \geq 0,\ \|\boldsymbol{x}\|_\infty \leq 1$$

Solution: Since $\boldsymbol{x} \geq 0$, the optimization problem is rewritten as

$$\min_{\boldsymbol{x}} \quad \|\boldsymbol{b}-\boldsymbol{x}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{x} \geq 0, \boldsymbol{x} \leq 1$$

The Lagrangian function of the problem is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \nu) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{b}\|_2^2 - \boldsymbol{\lambda}^T\boldsymbol{x} + \boldsymbol{\nu}^T(\boldsymbol{x}-1)$$

where $\boldsymbol{\lambda} \geq 0$, $\boldsymbol{\nu} \geq 0$. The optimal solution of the problem must satisfy the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{x}} &= \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\lambda} + \boldsymbol{\nu} = 0 && \text{(stationarity)}\\
\boldsymbol{\lambda} &\geq 0, \quad \boldsymbol{\nu} \geq 0 && \text{(dual feasibility)}\\
0 \leq x_i &\leq 1 && \text{(primal feasibility)}\\
\lambda_i x_i &= 0 \quad \forall i = 1, \ldots, n && \text{(complementary slackness)}\\
\nu_i(x_i - 1) &= 0 \quad \forall i = 1, \ldots, n &&
\end{aligned}$$

From the stationarity condition, we can obtain the optimal primal solution as

$$\boldsymbol{x}^\star = \boldsymbol{b} + \boldsymbol{\lambda} - \boldsymbol{\nu}$$

Denote by $\mathcal{I} = \{i : 0 < x_i^\star < 1\}$ the index set of $\boldsymbol{x}^\star$ that are strictly between 0 and 1. By the complementarity conditions, we have

$$\lambda_i = 0, \quad \nu_i = 0, \quad i \in \mathcal{I}$$

the expression of $\boldsymbol{x}^\star_{\mathcal{I}}$

$$\boldsymbol{x}^\star_{\mathcal{I}} = \boldsymbol{b}_{\mathcal{I}}\,.$$

This implies that $0 < b_i < 1$ for all $i \in \mathcal{I}$.
For $j \notin \mathcal{I}$, we have either $x_j^\star = 0$ or $x_j^\star = 1$. From the stationarity condition, we get:

$$x_j^\star = b_j + \lambda_j - \nu_j$$

If $x_j^\star = 0$, we must have $\nu_j = 0$ and $x_j^\star = b_j + \lambda_j = 0$, which implies that $b_j = -\lambda_j \leq 0$.

If $x_j^\star = 1$, we must have $\lambda_j = 0$, $x_j^\star = b_j - \nu_j = 1$, which implies that $b_j = 1 + \nu_j \geq 1$. Therefore, we can write the optimal solution as:

$$\boldsymbol{x}^\star = \max(0, \min(1, \boldsymbol{b}))$$

where the max and min operations are applied element-wise.

(c) (2 points)

$$\min_{\boldsymbol{x}} \quad \|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2^2 \leq 1$$

Solution: See Slides for Quadratically constrained least squares.

The Lagrangian function of the problem is

$$L(\boldsymbol{x}, \lambda) = \frac{1}{2}\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \frac{1}{2}\lambda(\|\boldsymbol{x}\|_2^2 - 1)$$

where $\lambda \geq 0$. The optimal solution of the problem must satisfy the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned}
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{A}^T\mathbf{A}\boldsymbol{x} - \mathbf{A}^T\boldsymbol{b} + \lambda\boldsymbol{x} = 0 \qquad & \text{(stationarity)} \\
\lambda \geq 0 \qquad & \text{(dual feasibility)} \\
|\boldsymbol{x}|_2^2 \leq 1 \qquad & \text{(primal feasibility)} \\
\lambda(|\boldsymbol{x}|_2^2 - 1) = 0 \qquad & \text{(complementary slackness)}
\end{aligned}$$

From the stationarity condition, we can obtain the optimal primal solution as:

$$\boldsymbol{x}^\star = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{b}$$

where $\mathbf{I}$ is the identity matrix. The dual function is then given by

$$\begin{aligned}
g(\lambda) &= \inf_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) \\
&= \frac{1}{2}\|\mathbf{A}(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{b} - \boldsymbol{b}\|_2^2 + \frac{\lambda}{2}\|((\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{b})\|_2^2 - \frac{\lambda}{2}
\end{aligned}$$

Denote SVD of $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})$, and $\boldsymbol{z} = \mathbf{U}^T\boldsymbol{b}$. The dual function is simplified into

$$g(\lambda) = \frac{\lambda}{2}\sum_i \frac{z_i^2}{\sigma_i^2 + \lambda} - \frac{\lambda}{2}$$

and the optimal $\lambda \geq 0$ is root of the derivative or a secular function

$$\sum_{i=1}^n \frac{z_i^2\sigma_i^2}{(\sigma_i^2 + \lambda)^2} - 1 = 0$$

This equation can be solved numerically using bisection or Newton's method.

3. (4 points) **Clustering**

Consider images in the MNIST dataset for three digits 0, 1, and 2. Denote by $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K]$ a matrix of vectorization of the images.

(a) Solve the following optimization problem to find an orthogonal matrix $\mathbf{U}$ and a feature matrix, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K]$, of size $R \times K$

$$\begin{aligned} \min \quad & \|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 \\ \text{s.t.} \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I}_R \\ & \mathbf{X} \geq 0, \quad \mathbf{X}^T\mathbf{1}_R = \mathbf{1}_K \end{aligned}$$

(b) On the basis of extracted features, $\boldsymbol{x}_k$, apply the K-means algorithm to predict categorical labels of images.

Solution:

(a) Assume $\mathbf{Y}$ is of size $N \times K$. Given $\mathbf{X}$, define SVD of $\mathbf{Y}\mathbf{X}^T = \mathbf{A}\,\text{diag}(\boldsymbol{\sigma})\mathbf{B}^T$.
Update U

$$\begin{aligned} \mathbf{U}^\star &= \arg\min_{\mathbf{U}} \quad \|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 \\ &= \arg\min_{\mathbf{U}} \quad \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 - 2\,\text{tr}(\mathbf{U}\mathbf{X}\mathbf{Y}^T) \\ &= \arg\max_{\mathbf{U}} \quad \text{tr}((\mathbf{A}^T\mathbf{U}\mathbf{B})\,\text{diag}(\boldsymbol{\sigma})) \\ &= \arg\max_{\mathbf{U}} \quad \text{diag}(\mathbf{A}^T\mathbf{U}\mathbf{B})^T\boldsymbol{\sigma} \end{aligned}$$

Note that $\mathbf{U}\mathbf{B}$ is an orthogonal matrix $(\mathbf{U}\mathbf{B})^T(\mathbf{U}\mathbf{B}) = \mathbf{B}^T\mathbf{U}^T\mathbf{U}\mathbf{B} = \mathbf{I}$ and $|\boldsymbol{a}_i^T\boldsymbol{u}_i|^2 \leq \|\boldsymbol{a}_i\|_2^2\,\|\boldsymbol{u}_i\|_2^2 = 1$ for $i = 1, 2, \ldots$. Hence $(\mathbf{A}^T\mathbf{U}\mathbf{B})(i,i) \leq 1$. The above problem achieves a maximum when $\mathbf{A}^T\mathbf{U}\mathbf{B}$ is an identity matrix, or $\mathbf{U} = \mathbf{A}\mathbf{B}^T$.
Update $\boldsymbol{x}_k$ as projection of $\mathbf{U}^T\boldsymbol{y}_k$ onto the probability simplex $\Delta(R)$

$$\begin{aligned} \boldsymbol{x}_k^\star &= \arg\min_{\boldsymbol{x}\in\Delta(R)} \quad \|\boldsymbol{y}_k - \mathbf{U}\boldsymbol{x}_k\|_2^2 \\ &= \arg\min_{\boldsymbol{x}\in\Delta(R)} \quad \|\mathbf{U}^T\boldsymbol{y}_k - \boldsymbol{x}_k\|_2^2 \\ &= \text{proj}_{\Delta(R)}(\mathbf{U}^T\boldsymbol{y}_k) \end{aligned}$$

4. (2 points) Solve the following optimization problem by applying optimality condition

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i \ln x_i \\ \text{s.t.} \quad & \sum_{i=1}^n x_i = 1, \quad \boldsymbol{x} \geq 0 \end{aligned}$$

where $\alpha_i < 0$.

Solution: The objective function $f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i \ln x_i$ is convex since the Hessian $\nabla^2 f(x) = -\operatorname{diag}(\dots, \frac{\alpha_i}{x_i^2}, \dots) > 0$.

Next consider a normalized objective function which is scaled by $-\sum_i \alpha_i (> 0)$. Define $\tilde{\alpha}_i = \frac{-\alpha_i}{\sum_i \alpha_i}$, then $\sum_i \tilde{\alpha}_i = -1$.

We solve an alternative optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & g(\boldsymbol{x}) = \sum_{i=1}^n \tilde{\alpha}_i \ln x_i \\ \text{s.t.} \quad & \sum_{i=1}^n x_i = 1, \quad \boldsymbol{x} \geq 0 \end{aligned}$$

Applying the optimality condition, for all $\boldsymbol{y}$ in the n- simplex Δ_n, we have

$$0 \leq (\nabla g(\boldsymbol{x}^\star))^T (\boldsymbol{y} - \boldsymbol{x}^\star) = \sum_i \frac{\tilde{\alpha}_i}{x_i^\star}(y_i - x_i^\star)$$

or

$$h(\boldsymbol{y}) = \sum_i \frac{-\tilde{\alpha}_i}{x_i^\star} y_i \leq \sum_i (-\tilde{\alpha}_i) = 1 \quad \forall \boldsymbol{y} \in \Delta_n. \tag{13}$$

It is clear that $\boldsymbol{x}^\star = -\tilde{\boldsymbol{\alpha}}$ is the optimal solution since it holds the optimality condition.

Another way is to prove $h(\boldsymbol{y}) \leq 1$ for all $\boldsymbol{y}$ in the simplex Δ_n only when $\boldsymbol{x}^\star = -\tilde{\boldsymbol{\alpha}}$.

Denote by m the index such that $\frac{-\tilde{\alpha}_m}{x_m^\star} = \max\left(\dots, \frac{-\tilde{\alpha}_i}{x_i^\star}, \dots\right)$. Note that $\frac{-\tilde{\alpha}_m}{x_m^\star} \geq 1$, otherwise $1 = \sum_i -\tilde{\alpha}_i < \sum_i x_i^\star = 1$. Applying the Holder's inequality

$$h(\boldsymbol{y}) = \sum_i \frac{-\tilde{\alpha}_i}{x_i^\star} y_i \leq \| \frac{-\tilde{\boldsymbol{\alpha}}}{\boldsymbol{x}^\star} \|_\infty \, \|\boldsymbol{y}\|_1 = \frac{-\tilde{\alpha}_m}{x_m^\star}$$

The equality holds when $\boldsymbol{y} = \boldsymbol{e}_m$ is the m-th unit vector. From (13),

$$1 \geq h(\boldsymbol{y} = \boldsymbol{e}_m) = \sum_i \frac{-\tilde{\alpha}_i}{x_i^\star} y_i = \frac{-\tilde{\alpha}_m}{x_m^\star} \geq 1$$

Implying that $\frac{-\tilde{\alpha}_m}{x_m^\star} = 1 \geq \frac{-\tilde{\alpha}_i}{x_i^\star}$ for $i \neq m$. If there exists $\frac{-\tilde{\alpha}_j}{x_j^\star} < 1$, then $\sum_i -\tilde{\alpha}_i < \sum_i x_i^\star = 1$. Contradiction. Finally, $\frac{-\tilde{\alpha}_i}{x_i^\star} = 1$ for all $i = 1, \dots, n$.

5. (2 points) Prove that the function $f(x) = \log(\sum_k \exp(x_k))$
   - is convex
   - but not strictly convex

Solution

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) = \begin{cases} \frac{-e^{x_i}e^{x_j}}{(\sum_j e^{x_j})^2}, & i \neq j \\ \frac{-e^{x_i}e^{x_j}}{(\sum_j e^{x_j})^2} + \frac{e^{x_j}}{\sum_j e^{x_j}}, & i = j \end{cases}$$

The Hessian matrix is then written as

$$\nabla^2 f(\boldsymbol{x}) = \text{diag}(\mathbf{z}) - \mathbf{z}\,\mathbf{z}^T, \tag{14}$$

where $\mathbf{z} = [z_i]$, $z_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$ and $\boldsymbol{z}^T\mathbf{1} = 1$

To show that f is convex, we need to show that $\boldsymbol{v}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{v} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^n$.

$$\boldsymbol{v}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{v} = \boldsymbol{v}^T\text{diag}(\boldsymbol{z})\boldsymbol{v} - \boldsymbol{v}^T\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{v} = \sum_{i=1}^n z_i v_i^2 - \left(\sum_{i=1}^n z_i v_i\right)^2$$

Using the Cauchy-Schwarz inequality we have

$$\left(\sum_{i=1}^n z_i v_i\right)^2 \leq \left(\sum_{i=1}^n z_i^2\right)\left(\sum_{i=1}^n v_i^2\right) \leq \left(\sum_{i=1}^n z_i\right)\left(\sum_{i=1}^n z_i v_i^2\right) = \sum_{i=1}^n z_i v_i^2$$

where the last inequality follows from the fact that $\sum_{i=1}^n z_i = 1$ and $0 \leq z_i \leq 1$ for all i. Therefore, we have

$$\boldsymbol{v}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{v} \geq \sum_{i=1}^n z_i v_i^2 - \left(\sum_{i=1}^n z_i v_i\right)^2 \geq 0$$

which proves that f is convex.

To show that f is not strictly convex, we need to find an $\boldsymbol{x} \in \mathbb{R}^n$ and a nonzero $\boldsymbol{v} \in \mathbb{R}^n \setminus \mathbf{0}$ such that $\boldsymbol{v}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{v} = 0$. A simple choice is to take $\boldsymbol{v} = \mathbf{1}$, where $\mathbf{1}$ is the vector of all ones. Then, we have:

$$\boldsymbol{v}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{v} = \mathbf{1}^T\text{diag}(\boldsymbol{z})\mathbf{1} - \mathbf{1}^T\boldsymbol{z}\boldsymbol{z}^T\mathbf{1} = \sum_{i=1}^n z_i - \left(\sum_{i=1}^n z_i\right)^2 = 1 - 1 = 0$$

This proves that f is not strictly convex.