Optimization Algorithms: HW0

Lo Chun, Chou
R13922136

March 10, 2025

1

(1)

To show that the optimization problem defining $w^\natural$ is convex, we need to show that both the objective function and the constraint set are convex.

Claim: The objective function $g(w) := \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle x_i, w\rangle)^2$ is convex, and the constraint set $\mathbb{R}^d$ is also convex.

To prove that $g(w)$ is convex, we would use the theorem that:

Theorem. [1] *Assume that a function f is twice differentiable, then f is convex $\Leftrightarrow$ $\mathrm{dom} f$ is convex and its Hessian is positive semidefinite.*

To check that if $g(w)$ is twice differentiable, we first convert the original definition into a matrix-vector form, by letting:

$$X = \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix} \in \mathbb{R}^{n\times d} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n} \quad \text{and} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^{d}$$

Then, we'll get:

$$\begin{aligned} g(w) &= \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^\intercal w)^2 \\ &= \frac{1}{2n}\sum_{i=1}^{n}\left[y_i^2 - 2y_i x_i^\intercal w + (x_i^\intercal w)^2\right] \\ &= \frac{1}{2n}(y^\intercal y - 2y^\intercal Xw + (Xw)^\intercal Xw) \\ &= \frac{1}{2n}(y^\intercal y - 2w^\intercal X^\intercal y + w^\intercal X^\intercal Xw) \end{aligned}$$

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, pp. 71.

Differentiate w.r.t. w:

$$\begin{aligned}\nabla g(w) &= \frac{\partial}{\partial w}\left[\frac{1}{2n}(y^\intercal y - 2w^\intercal X^\intercal y + w^\intercal X^\intercal X w)\right] \\ &= \frac{1}{2n}\left[0 - 2X^\intercal y + 2X^\intercal X w\right] \\ &= -\frac{1}{n}X^\intercal y + \frac{1}{n}X^\intercal X w \end{aligned} \tag{1}$$

Then the second derivative:

$$\begin{aligned}\nabla^2 g(w) &= \frac{\partial}{\partial w}\left[-\frac{1}{n}X^\intercal y + \frac{1}{n}X^\intercal X w\right] \\ &= \frac{1}{n}X^\intercal X \end{aligned} \tag{2}$$

Since $\frac{1}{n}X^\intercal X$ does not depend on w, it is a constant matrix, and therefore the second derivative exists at each point in $\mathrm{dom} f$. We can now check the conditions of the theorem.

The domain of $g(w)$ is $\mathbb{R}^d$, which is convex. [2]

For any $v \in \mathbb{R}^d$, we have:

$$\begin{aligned}v^\intercal \nabla^2 g(w) v &= v^\intercal \frac{1}{n} X^\intercal X v \\ &= \frac{1}{n}(Xv)^\intercal X v \\ &= \frac{1}{n}\|Xv\|_2^2 \geq 0 \end{aligned}$$

Thus, the Hessian of $g(w)$ is positive semidefinite, and $g(w)$ is convex.

Finally, the constraint set $\mathbb{R}^d$ is also convex, as shown above, we can conclude that the optimization problem defining $w^\natural$ is convex. $\square$

(2)

For $t = 1$, we have:

$$w_2 = w_1 - \left(\nabla^2 g(w_1)\right)^{-1} \nabla g(w_1), \qquad \text{where } w_1 = 0 \in \mathbb{R}^d \tag{*}$$

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, pp. 27.

To get w_2, we need to calculate $\nabla g(w_1)$, $\nabla^2 g(w_1)$, from (1) in the previous question, we have:

$$\begin{aligned}\nabla g(w_1) &= -\frac{1}{n}X^\intercal y + \frac{1}{n}X^\intercal X w_1 \\ &= -\frac{1}{n}X^\intercal y\end{aligned}$$

And from (2) in the previous question, we have:

$$\nabla^2 g(w_1) = \frac{1}{n}X^\intercal X$$

Plugging back into $(*)$, we get:

$$\begin{aligned}w_2 &= w_1 - \left(\nabla^2 g(w_1)\right)^{-1} \nabla g(w_1) \\ &= 0 - \left(\frac{1}{n}X^\intercal X\right)^{-1} \left(-\frac{1}{n}X^\intercal y\right) \\ &= 0 + n(X^\intercal X)^{-1}\frac{1}{n}X^\intercal y \\ &= (X^\intercal X)^{-1}X^\intercal y\end{aligned} \tag{1}$$

To show that $w_2 = w^\natural$, observe that $\nabla g(w^\natural) = 0$, using (1) in the previous question, we have:

$$\begin{aligned}\nabla g(w^\natural) &= -\frac{1}{n}X^\intercal y + \frac{1}{n}X^\intercal X w^\natural \\ &= -\frac{1}{n}X^\intercal y + \frac{1}{n}X^\intercal X w^\natural = 0\end{aligned}$$

Reorder and simplify the terms, we get:

$$X^\intercal X w^\natural = X^\intercal y$$

Since $rank(X) = d$, $rank(X^\intercal X) = rank(X) = d$, $X^\intercal X \in \mathbb{R}^{d\times d}$ is of full rank and therefore invertible (and positive definite).
Thus, we can write:

$$w^\natural = (X^\intercal X)^{-1}X^\intercal y$$

which is the same as w_2 in (1). $\square$

2

(1)

Since $y_1, \ldots, y_n$ are random variables that satisfy:

$$\mathsf{P}(y_i = 1) = 1 - \mathsf{P}(y_i = 0) = \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}$$

We knew that the probability of $\mathsf{P}(y_i = 0)$ is:

$$\begin{aligned}
\mathsf{P}(y_i = 0) &= 1 - \mathsf{P}(y_i = 1) \\
&= \frac{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} - \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \\
&= \frac{\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}
\end{aligned}$$

We can write the pmf of given $x_i, \theta^\natural$, observed y_i as:

$$p(y_i | x_i, \theta^\natural) = \left(\frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \right)^{y_i} \left(\frac{\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \right)^{1 - 2y_i}$$

Using the pmf, we can get the likelihood function, which can be written as:

$$l(\theta) = \prod_{i=1}^{n} p(y_i | x_i, \theta)^3$$

We can further derive the log-likelihood:

$$\begin{aligned}
\log l(\theta) &= \log \left[\prod_{i=1}^{n} \left(\frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta \rangle}} \right)^{y_i} \left(\frac{\mathrm{e}^{-\langle x_i, \theta \rangle}}{1 + \mathrm{e}^{-\langle x_i, \theta \rangle}} \right)^{1 - 2y_i} \right] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta \rangle}} \right) + (1 - 2y_i) \log \left(\frac{\mathrm{e}^{-\langle x_i, \theta \rangle}}{1 + \mathrm{e}^{-\langle x_i, \theta \rangle}} \right) \right] \qquad (*)
\end{aligned}$$

The terms in the above equation can be simplified:

$$\begin{aligned}
y_i \log \left(\frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta \rangle}} \right) &= y_i \log \left(1 + \mathrm{e}^{-\langle x_i, \theta \rangle} \right)^{-1} \\
&= -y_i \log \left(1 + \mathrm{e}^{-\langle x_i, \theta \rangle} \right)
\end{aligned}$$

$$\begin{aligned}
(1 - 2y_i) \log \left(\frac{\mathrm{e}^{-\langle x_i, \theta \rangle}}{1 + \mathrm{e}^{-\langle x_i, \theta \rangle}} \right) &= (1 - 2y_i) \log \left(\mathrm{e}^{-\langle x_i, \theta \rangle} \right) - (1 - 2y_i) \log \left(1 + \mathrm{e}^{-\langle x_i, \theta \rangle} \right) \\
&= -\langle x_i, \theta \rangle (1 - 2y_i) - \log \left(1 + \mathrm{e}^{-\langle x_i, \theta \rangle} \right) + y_i \log \left(1 + \mathrm{e}^{-\langle x_i, \theta \rangle} \right)
\end{aligned}$$

[3] Robert V. Hogg, Elliot A. Tanis, Dale Zimmerman, *Probability and Statistical Inference*, 9th ed., Pearson Education, 2015, p. 258-259.

Plugging back into $(*)$, we get:

$$\begin{aligned}\log l(\theta) &= \sum_{i=1}^{n}\left[-y_i \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right) - \langle x_i,\theta\rangle(1-2y_i) - \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right) + y_i\log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right]\\ &= \sum_{i=1}^{n}\left[-\langle x_i,\theta\rangle(1-2y_i) - \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right]\end{aligned}$$

We can find the maximum likelihood estimator $\hat{\theta}_n$ by maximizing the log-likelihood function, which is equivalent to find the minimum of the negative log-likelihood function.

$$\begin{aligned}-\log l(\theta) &= -\sum_{i=1}^{n}\left[-\langle x_i,\theta\rangle(1-2y_i) - \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right]\\ &= \sum_{i=1}^{n}\left[\langle x_i,\theta\rangle(1-2y_i) + \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right]\\ &= \sum_{i=1}^{n}\left[\log \mathrm{e}^{\langle x_i,\theta\rangle(1-2y_i)} + \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right]\\ &= \sum_{i=1}^{n}\left[\log\left(\mathrm{e}^{\langle x_i,\theta\rangle(1-2y_i)}\cdot\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right)\right]\\ &= \sum_{i=1}^{n}\left[\log\left(\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle} + \mathrm{e}^{(-2y_i)}\right)\right]\end{aligned}$$

Check this part! (about the 1)

Since $\mathrm{e}^{(-2y_i)}$ evaluates to 1 when $y_i = 0$ and is a small constant when $y_i = 1$, also, taking the average (multiplying by $\frac{1}{n}$) will not change the result, to minimize $-\log l(\theta)$ will be equivalent to minimize the given expression:

$$\hat{\theta}_n \in \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} L(\theta), \quad L(\theta) := \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}\right) \qquad \square$$

(2)

To show that the optimization problem defining the maximum-likelihood estimator is convex, we need to show that both the objective function and the constraint set are convex.

As in 1.(1), we knew that the constraint set $\mathbb{R}^p$ is convex, therefore, we only need to check the convexity of the objective function.

Claim: The objective function $L(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta\rangle}\right)$ is convex.
Following the same steps in 1.(1), we first differentiate $L(\theta)$ w.r.t. θ:

$$\begin{aligned}\nabla L(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta\rangle}} \cdot -2(y_i - \frac{1}{2})x_i \cdot \mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta\rangle} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i(1-2y_i)\mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta\rangle}}{1 + \mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta\rangle}}\end{aligned}$$

To make the equation more readble, we can represent $z_i = 2(y_i - \frac{1}{2})\langle x_i, \theta\rangle$, so that $\nabla L(\theta)$ is equivalent to:

$$\nabla L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i(1-2y_i)\mathrm{e}^{-z_i}}{1 + \mathrm{e}^{-z_i}}$$

In order to calculate the Hessian, we first calculate some of the terms:

$$\frac{d}{d\theta} z_i = 2(y_i - \frac{1}{2})x_i$$

$$\frac{d}{d\theta} e^{-z_i} = -2(y_i - \frac{1}{2})x_i\mathrm{e}^{-z_i}$$

Then we'll have:

$$\begin{aligned}\frac{d}{d\theta}\left(\frac{(1-2y_i)\mathrm{e}^{-z_i}}{1+\mathrm{e}^{-z_i}}\right) &= \frac{\frac{d}{d\theta}\left((1-2y_i)\mathrm{e}^{-z_i}\right) \cdot (1+\mathrm{e}^{-z_i}) - (1-2y_i)\mathrm{e}^{-z_i} \cdot \frac{d}{d\theta}(1+\mathrm{e}^{-z_i})}{(1+\mathrm{e}^{-z_i})^2} \\ &= \frac{-2(1-2y_i)(y_i - \frac{1}{2})x_i\mathrm{e}^{-z_i}(1+\mathrm{e}^{-z_i}) + 2(1-2y_i)\mathrm{e}^{-z_i}(y_i - \frac{1}{2})x_i\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} \\ &= \frac{2(1-2y_i)(y_i - \frac{1}{2})x_i\mathrm{e}^{-z_i}\left[(-1 - e^{-z_i}) + e^{-z_i}\right]}{(1+\mathrm{e}^{-z_i})^2} \\ &= 2(1-2y_i)(y_i - \frac{1}{2})x_i\mathrm{e}^{-z_i}\frac{-1}{(1+\mathrm{e}^{-z_i})^2} \\ &= \frac{-2(1-2y_i)(y_i - \frac{1}{2})x_i\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2}\end{aligned}$$

Getting back to the Hessian, we have:

$$
\begin{aligned}
\nabla^2 L(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}\left[x_i\left(\frac{(1-2y_i)\mathrm{e}^{-z_i}}{1+\mathrm{e}^{-z_i}}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \frac{-2(1-2y_i)(y_i-\frac{1}{2})x_i\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\intercal \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} \qquad (1)
\end{aligned}
$$

Since $(1+\mathrm{e}^{-z_i})^2$ is strictly positive, the Hessian exists for all point in $\mathbb{R}^p$. Therefore, $L(\theta)$ is twice differentiable.

Since we knew that the domain of $L(\theta)$ is convex, we only need to check if the Hessian is positive semidefinite to prove that $L(\theta)$ is convex.

By (1), for any $v \in \mathbb{R}^p$, we have:

$$
\begin{aligned}
v^\intercal \nabla^2 L(\theta) v &= \frac{1}{n}\sum_{i=1}^{n} v^\intercal x_i x_i^\intercal \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} v \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{-2(1-2y_i)(y_i-\frac{1}{2})v^\intercal x_i x_i^\intercal v}{(1+\mathrm{e}^{-z_i})^2}
\end{aligned}
$$

For the denominator, $(1+\mathrm{e}^{-z_i})^2 > 0$, and for the coefficient, $-2(1-2y_i)(y_i-\frac{1}{2})$, since $y_i \in \{0,1\}$, we have:

$$
-2(1-2y_i)(y_i-\frac{1}{2}) \geq 0
$$

Last, we have $v^\intercal x_i x_i^\intercal v$, this is equivelent to $(v^\intercal x_i)^2$, which is non-negative. Therefore, we have:

$$
\frac{1}{n}\sum_{i=1}^{n} \frac{-2(1-2y_i)(y_i-\frac{1}{2})v^\intercal x_i x_i^\intercal v}{(1+\mathrm{e}^{-z_i})^2} \geq 0
$$

Thus the Hessian is positive semidefinite, and $L(\theta)$ is convex. $\square$

(3)

Let:

$$
X = \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix} \in \mathbb{R}^{n\times p} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n
$$

By the previous subproblem, we have:

$$\nabla L(\theta^\natural) = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i(1-2y_i)\mathrm{e}^{-z_i}}{1+\mathrm{e}^{-z_i}} \qquad \text{where } z_i = 2(y_i - \frac{1}{2})\langle x_i, \theta^\natural\rangle \tag{*}$$

Thus, to show that $\nabla L(\theta^\natural) = -\frac{1}{n}X^\intercal(y - \mathsf{E}[y])$, it is equivalent to prove:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{x_i(1-2y_i)\mathrm{e}^{-z_i}}{1+\mathrm{e}^{-z_i}} &= -\frac{1}{n}X^\intercal(y - \mathsf{E}[y]) \\ \Rightarrow \sum_{i=1}^{n} \frac{x_i(1-2y_i)\mathrm{e}^{-z_i}}{1+\mathrm{e}^{-z_i}} &= X^\intercal(\mathsf{E}[y] - y) \end{aligned} \tag{1}$$

Right-hand side of (1)

First we knew that the definition of $\mathsf{E}[y]$ is:

$$\mathsf{E}[y] = \begin{bmatrix} \mathsf{E}[y_1] \\ \mathsf{E}[y_2] \\ \vdots \\ \mathsf{E}[y_n] \end{bmatrix}$$

So:

$$\mathsf{E}[y] - y = \begin{bmatrix} \mathsf{E}[y_1] - y_1 \\ \mathsf{E}[y_2] - y_2 \\ \vdots \\ \mathsf{E}[y_n] - y_n \end{bmatrix}$$

By subproblem (1), we have:

$$p(y_i|x_i, \theta^\natural) = \left(\frac{1}{1+\mathrm{e}^{-\langle x_i, \theta^\natural\rangle}}\right)^{y_i} \left(\frac{\mathrm{e}^{-\langle x_i, \theta^\natural\rangle}}{1+\mathrm{e}^{-\langle x_i, \theta^\natural\rangle}}\right)^{1-2y_i}$$

Therefore, its expected value is:

$$\begin{aligned} \mathsf{E}[y_i] &= 1 \times \mathsf{P}(y_i = 1) + 0 \times \mathsf{P}(y_i = 0) \\ &= \mathsf{P}(y_i = 1) \\ &= \frac{1}{1+\mathrm{e}^{-\langle x_i, \theta^\natural\rangle}} \end{aligned}$$

Therefore, we can rewrite the right-hand side of (1):

$$X^\intercal(\mathsf{E}[y_i] - y_i) = \sum_{i=1}^{n} \left(\frac{1}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} - y_i \right) x_i$$

Case. For $y_i = 1$

For the term in the summation, consider the case $y_i = 1$, this would evaluate to:

$$\begin{aligned}
(\mathsf{E}[y_i] - y_i)x_i &= \left(\frac{1}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} - 1 \right) x_i \\
&= \left(\frac{1}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} - \frac{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \right) x_i \\
&= \frac{-x_i \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}
\end{aligned}$$

Case. For $y_i = 0$

$$\begin{aligned}
(\mathsf{E}[y_i] - y_i)x_i &= \left(\frac{1}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} - 0 \right) x_i \\
&= \frac{x_i}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}
\end{aligned}$$

Left-hand side of (1)

Consider the left-hand side of (1), expand the expression by plugging in the definition of z_i:

$$\frac{x_i(1-2y_i)\mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{-2(y_i - \frac{1}{2})\langle x_i, \theta^\natural \rangle}}$$

Case. For $y_i = 1$

$$\frac{x_i(1-2)\mathrm{e}^{-2(1 - \frac{1}{2})\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{-2(1 - \frac{1}{2})\langle x_i, \theta^\natural \rangle}} = \frac{-x_i e^{-\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}$$

Case. For $y_i = 0$

$$\frac{x_i(1-0)\mathrm{e}^{-2(0 - \frac{1}{2})\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{-2(0 - \frac{1}{2})\langle x_i, \theta^\natural \rangle}} = \frac{x_i \mathrm{e}^{\langle x_i, \theta^\natural \rangle}}{1+\mathrm{e}^{\langle x_i, \theta^\natural \rangle}}$$

Combine all of the above into equation (1), we found that the equation we need to prove :

$$\sum_{i=1}^{n} \frac{x_i(1-2y_i)\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta^\natural\rangle}}{1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta^\natural\rangle}} = \sum_{i=1}^{n}\left(\frac{1}{1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}} - y_i\right)x_i$$

is actually the same:

$$\sum_{\{i|y_i=1\}} \frac{-x_i\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}}{1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}} + \sum_{\{i|y_i=0\}} \frac{x_i\mathrm{e}^{\langle x_i,\theta^\natural\rangle}}{1+\mathrm{e}^{\langle x_i,\theta^\natural\rangle}} = \sum_{\{i|y_i=1\}} \frac{-x_i\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}}{1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}} + \sum_{\{i|y_i=0\}} \frac{x_i}{1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}} \qquad \square$$

(4)

By equation (1) in 2.(2), we have:

$$\nabla^2 L(\theta^\natural) = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\intercal \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} \qquad \text{where } z_i = 2(y_i-\frac{1}{2})\langle x_i,\theta^\natural\rangle$$

As the approach used in the previous subproblem, we can rewrite the summation part by deviding into the cases that $y_i = 1$ and $y_i = 0$:

$$\begin{aligned}
\nabla^2 L(\theta^\natural) &= \frac{1}{n}\left[\sum_{\{i|y_i=1\}} x_i x_i^\intercal \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} + \sum_{\{i|y_i=0\}} x_i x_i^\intercal \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2}\right] \\
&= \frac{1}{n}\left[\sum_{\{i|y_i=1\}} x_i x_i^\intercal \frac{-2(1-2)(1-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} + \sum_{\{i|y_i=0\}} x_i x_i^\intercal \frac{-2(1-0)(0-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2}\right] \\
&= \frac{1}{n}\left[\sum_{\{i|y_i=1\}} x_i x_i^\intercal \frac{\mathrm{e}^{-\langle x_i,\theta^\natural\rangle}}{(1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle})^2} + \sum_{\{i|y_i=0\}} x_i x_i^\intercal \frac{\mathrm{e}^{\langle x_i,\theta^\natural\rangle}}{(1+\mathrm{e}^{\langle x_i,\theta^\natural\rangle})^2}\right] \qquad (1)
\end{aligned}$$

Need to prove that:

$$\nabla^2 L(\theta^\natural) = X^\intercal D X \qquad \text{where } D = \begin{bmatrix} Var(y_1) & 0 & \cdots & 0 \\ 0 & Var(y_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Var(y_n) \end{bmatrix}$$

To calculate $Var(y_i)$, we knew that:

$$Var(y_i) = \mathsf{E}[y_i^2] - \mathsf{E}[y_i]^2$$

Since $y_i \in \{0, 1\}$, and from the previous subproblem 2.(3), we have:

$$\mathsf{E}[y_i] = \mathsf{P}(y_i = 1) = \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}$$

So we can derive the following:

$$\begin{aligned} \mathsf{E}[y_i^2] &= \mathsf{P}(y_i = 1) \cdot 1^2 + \mathsf{P}(y_i = 0) \cdot 0^2 \\ &= \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \end{aligned}$$

$$\mathsf{E}[y_i]^2 = \left(\frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \right)^2$$

Hence, we have:

$$\begin{aligned} Var(y_i) = \mathsf{E}[y_i^2] - \mathsf{E}[y_i]^2 &= \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} - \left(\frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \right)^2 \\ &= \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \left(1 - \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}} \right) \\ &= \frac{\mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{(1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle})^2} \end{aligned}$$

Calculate $\frac{1}{n} X^\intercal D X$:

$$\begin{aligned} \frac{1}{n} X^\intercal D X &= \frac{1}{n} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \frac{\mathrm{e}^{-\langle x_1, \theta^\natural \rangle}}{(1+\mathrm{e}^{-\langle x_1, \theta^\natural \rangle})^2} & 0 & \cdots & 0 \\ 0 & \frac{\mathrm{e}^{-\langle x_2, \theta^\natural \rangle}}{(1+\mathrm{e}^{-\langle x_2, \theta^\natural \rangle})^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\mathrm{e}^{-\langle x_n, \theta^\natural \rangle}}{(1+\mathrm{e}^{-\langle x_n, \theta^\natural \rangle})^2} \end{bmatrix} \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{x_i x_i^\intercal \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}}{(1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle})^2} \end{aligned}$$

This result can also be split into the cases that $y_i = 1$ and $y_i = 0$:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{x_i x_i^\intercal \mathrm{e}^{-\langle x_i,\theta^\natural\rangle}}{(1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle})^2} = \frac{1}{n}\left[\sum_{\{i|y_i=1\}}\frac{x_i x_i^\intercal \mathrm{e}^{-\langle x_i,\theta^\natural\rangle}}{(1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle})^2} + \sum_{\{i|y_i=0\}}\frac{x_i x_i^\intercal \mathrm{e}^{-\langle x_i,\theta^\natural\rangle}\cdot(\mathrm{e}^{\langle x_i,\theta^\natural\rangle})^2}{(1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle})^2\cdot(\mathrm{e}^{\langle x_i,\theta^\natural\rangle})^2}\right]$$
$$= \frac{1}{n}\left[\sum_{\{i|y_i=1\}}\frac{x_i x_i^\intercal \mathrm{e}^{-\langle x_i,\theta^\natural\rangle}}{(1+\mathrm{e}^{-\langle x_i,\theta^\natural\rangle})^2} + \sum_{\{i|y_i=0\}}\frac{x_i x_i^\intercal \mathrm{e}^{\langle x_i,\theta^\natural\rangle}}{(1+\mathrm{e}^{\langle x_i,\theta^\natural\rangle})^2}\right]$$

Which is exactly the same as equation (1), therefore, we have proven that $\nabla^2 L(\theta^\natural) = \frac{1}{n}X^\top D X$ holds. $\square$

(5)

In the last part of the subproblem 2.(2), we have already shown that $0 \le \nabla^2 L(\theta)$. Therefore, we need to show that:

$$\nabla^2 L(\theta) \le \frac{\lambda_{\max}(X^\top X)}{4n} I, \quad \forall\theta \in \mathbb{R}^p$$

Which means that we need to show:

$$\frac{\lambda_{\max}(X^\top X)}{4n} I - \nabla^2 L(\theta), \quad \forall\theta \in \mathbb{R}^p$$

is positive semi-definite.

Since the expression $\in \mathbb{R}^{p\times p}$, given arbitrary $u \in \mathbb{R}^p$, we need to show that:

$$u^\intercal\left(\frac{\lambda_{\max}(X^\top X)}{4n} I - \nabla^2 L(\theta)\right)u \ge 0$$
$$\Rightarrow u^\intercal \frac{\lambda_{\max}(X^\top X)}{4n} I u \ge u^\intercal \nabla^2 L(\theta) u$$

Plugging in the expression of $\nabla^2 L(\theta)$ from equation (1) in subproblem 2.(2), we have:

$$u^\intercal \nabla^2 L(\theta) u = u^\intercal \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\intercal \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} u$$
$$= \frac{1}{n}\sum_{i=1}^{n} u^\intercal x_i x_i^\intercal u \frac{(-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i})}{(1+\mathrm{e}^{-z_i})^2}$$
$$= \frac{1}{n}\sum_{i=1}^{n} u^\intercal x_i x_i^\intercal u \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} \qquad \text{where } z_i = 2(y_i - \frac{1}{2})\langle x_i, \theta\rangle$$

Consider the possible values of the term $\frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2}$:

Case. For $y_i = 1$

$$\frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle})^2} = \frac{-2(1-2)(1-\frac{1}{2})\mathrm{e}^{-2(1-\frac{1}{2})\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{-2(1-\frac{1}{2})\langle x_i,\theta\rangle})^2}$$
$$= \frac{\mathrm{e}^{-\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{-\langle x_i,\theta\rangle})^2}$$

Since we need to check if the maximum possible value of $\nabla^2 L(\theta)$ would exceed $\frac{\lambda_{\max}(X^\intercal X)}{4n}I$, we need to find the maximum possible value of this term.

First, since e^{-z_i} is always positive, $\frac{\mathrm{e}^{-\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{-\langle x_i,\theta\rangle})^2} \geq 0$.
Then, let:

$$f(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2}$$

Differentiate:

$$f'(x) = \frac{(-\mathrm{e}^{-x})(1+\mathrm{e}^{-x})^2 - \mathrm{e}^{-x}\cdot 2(1+\mathrm{e}^{-x})(-\mathrm{e}^{-x})}{(1+\mathrm{e}^{-x})^4}$$
$$= \frac{(-\mathrm{e}^{-x})(1+\mathrm{e}^{-x}) - 2\mathrm{e}^{-x}(-\mathrm{e}^{-x})}{(1+\mathrm{e}^{-x})^3}$$
$$= \frac{(-\mathrm{e}^{-x})(1-\mathrm{e}^{-x})}{(1+\mathrm{e}^{-x})^3}$$

Set $f'(x) = 0$:

$$\frac{(-\mathrm{e}^{-x})(1-\mathrm{e}^{-x})}{(1+\mathrm{e}^{-x})^3} = 0$$

Since $-\mathrm{e}^{-x} \neq 0$, we must have $1 - \mathrm{e}^{-x} = 0$, which means $\mathrm{e}^{-x} = 1$. Therefore, $f(x)$ has critical point at $x = 0$.
Calculate $f(0)$:

$$f(0) = \frac{\mathrm{e}^0}{(1+\mathrm{e}^0)^2} = \frac{1}{4}$$

Check the second derivative:

$$
\begin{aligned}
f''(x) &= \frac{\left[(-\mathrm{e}^{-x})(1-\mathrm{e}^{-x})\right]'(1+\mathrm{e}^{-x})^3 - \left[(1+\mathrm{e}^{-x})^3\right]'\left[(-\mathrm{e}^{-x})(1-\mathrm{e}^{-x})\right]}{(1+\mathrm{e}^{-x})^6} \\
&= \frac{-\mathrm{e}^{-x}(1-2\mathrm{e}^{-x})(1+\mathrm{e}^{-x})^3 - 3(1+\mathrm{e}^{-x})^2\mathrm{e}^{-x}(-1)(-\mathrm{e}^{-x})(1-\mathrm{e}^{-x})}{(1+\mathrm{e}^{-x})^6} \\
&= \frac{\mathrm{e}^{-x}(1+\mathrm{e}^{-x})^2\left[(1-2\mathrm{e}^{-x})(1+\mathrm{e}^{-x}) + 3(-\mathrm{e}^{-x})(1-\mathrm{e}^{-x})\right]}{(1+\mathrm{e}^{-x})^6} \\
&= \frac{\mathrm{e}^{-x}(1+\mathrm{e}^{-x})^2\left[1-2\mathrm{e}^{-x}+\mathrm{e}^{-x}-2\mathrm{e}^{-2x}-3\mathrm{e}^{-x}+3\mathrm{e}^{-2x}\right]}{(1+\mathrm{e}^{-x})^6} \\
&= \frac{\mathrm{e}^{-x}(1+\mathrm{e}^{-x})^2\left[1-4\mathrm{e}^{-x}+\mathrm{e}^{-2x}\right]}{(1+\mathrm{e}^{-x})^6} \\
&= \frac{\mathrm{e}^{-x}\left[1-4\mathrm{e}^{-x}+\mathrm{e}^{-2x}\right]}{(1+\mathrm{e}^{-x})^4}
\end{aligned}
$$

Evaluate at $x = 0$:

$$
f''(0) = \frac{\mathrm{e}^0(1-4\mathrm{e}^0+\mathrm{e}^0)}{(1+\mathrm{e}^0)^4} = \frac{1-4+1}{16} = -\frac{1}{8}
$$

Since $f''(0) < 0$, $f(x)$ has a local maximum at $x = 0$.

Case. For $y_i = 0$

$$
\begin{aligned}
\frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle})^2} &= \frac{-2(1-0)(0-\frac{1}{2})\mathrm{e}^{-2(0-\frac{1}{2})\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{-2(0-\frac{1}{2})\langle x_i,\theta\rangle})^2} \\
&= \frac{\mathrm{e}^{\langle x_i,\theta\rangle}}{(1+\mathrm{e}^{\langle x_i,\theta\rangle})^2}
\end{aligned}
$$

Checking the maximum possible value of this term again, let:

$$
g(x) = \frac{\mathrm{e}^x}{(1+\mathrm{e}^x)^2}
$$

Differentiate:

$$
\begin{aligned}
g'(x) &= \frac{\mathrm{e}^x(1+\mathrm{e}^x)^2 - \mathrm{e}^x\cdot 2(1+\mathrm{e}^x)\mathrm{e}^x}{(1+\mathrm{e}^x)^4} \\
&= \frac{\mathrm{e}^x(1+\mathrm{e}^x) - 2\mathrm{e}^x\cdot\mathrm{e}^x}{(1+\mathrm{e}^x)^3} \\
&= \frac{\mathrm{e}^x(1-\mathrm{e}^x)}{(1+\mathrm{e}^x)^3}
\end{aligned}
$$

Set $g'(x) = 0$:

$$\frac{e^x(1-e^x)}{(1+e^x)^3} = 0$$

Since $e^x \neq 0$, we must have $1 - e^x = 0$, which means $e^x = 1$. Therefore, $g(x)$ has critical point at $x = 0$.

Calculate $g(0)$:

$$g(0) = \frac{e^0}{(1+e^0)^2} = \frac{1}{4}$$

Check the second derivative:

$$\begin{aligned}
g''(x) &= \frac{\left[e^x(1-e^x)\right]'(1+e^x)^3 - \left[(1+e^x)^3\right]' e^x(1-e^x)}{(1+e^x)^6} \\
&= \frac{e^x(1-2e^x)(1+e^x)^3 - 3e^x(1+e^x)^2 e^x(1-e^x)}{(1+e^x)^6} \\
&= \frac{e^x(1+e^x)^2\left[(1-2e^x)(1+e^x) - 3e^x(1-e^x)\right]}{(1+e^x)^6} \\
&= \frac{e^x(1+e^x)^2\left[1-2e^x+e^x-2e^{2x}-3e^x+3e^{2x}\right]}{(1+e^x)^6} \\
&= \frac{e^x(1+e^x)^2\left[1-4e^x+e^{2x}\right]}{(1+e^x)^6} \\
&= \frac{e^x\left[1-4e^x+e^{2x}\right]}{(1+e^x)^4}
\end{aligned}$$

Evaluate at $x = 0$:

$$\begin{aligned}
g''(0) &= \frac{e^0(1-4e^0+e^0)}{(1+e^0)^4} \\
&= \frac{1-4+1}{16} \\
&= -\frac{1}{8}
\end{aligned}$$

Since $g''(0) < 0$, $g(x)$ has a local maximum at $x = 0$.

Thus, by the above two cases, we found that the maximum possible value of $\frac{-2(1-2y_i)(y_i-\frac{1}{2})e^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}}{(1+e^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle})^2}$ is bounded by $\frac{1}{4}$, hence we have:

$$
\begin{aligned}
u^\intercal \nabla^2 L(\theta) u &= \frac{1}{n}\sum_{i=1}^{n} u^\intercal x_i x_i^\intercal u \frac{-2(1-2y_i)(y_i-\frac{1}{2})\mathrm{e}^{-z_i}}{(1+\mathrm{e}^{-z_i})^2} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} u^\intercal x_i x_i^\intercal u \frac{1}{4} \\
&= \frac{1}{4n} u^\intercal \sum_{i=1}^{n} (x_i x_i^\intercal) u
\end{aligned}
$$

Which we could observe that since:

$$
X^\intercal X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix} = \sum_{i=1}^{n} x_i x_i^\intercal
$$

We have:

$$
u^\intercal \nabla^2 L(\theta) u \leq \frac{1}{4n} u^\intercal X^\intercal X u
$$

By definition, since $X^\intercal X \in \mathbb{R}^{p\times p}$ is a symmetric matrix, let $Q(u) = u^\intercal X^\intercal X u$ is its quadratic form. [4]

Then, its Rayleigh quotient is to normalize $Q(u)$ by $u^\intercal u$, and its maximum value is the maximum eigenvalue of $X^\intercal X$ (i.e. $\lambda_{\max}(X^\intercal X)$). [5] :

$$
\begin{aligned}
&\max_{u\in\mathbb{R}^p} \frac{u^\intercal X^\intercal X u}{u^\intercal u} = \lambda_{\max}(X^\intercal X) \\
&\Rightarrow u^\intercal X^\intercal X u \leq \lambda_{\max}(X^\intercal X) u^\intercal u \\
&\Rightarrow \frac{1}{4n} u^\intercal X^\intercal X u \leq \frac{1}{4n} \lambda_{\max}(X^\intercal X) u^\intercal u
\end{aligned}
$$

Therefore, we have:

[4] G. Chen, "Lecture 4: Rayleigh Quotient," San Jose State University, p.4. Available at: `https://www.sjsu.edu/faculty/guangliang.chen/Math253S20/lec4RayleighQuotient.pdf`.

[5] G. Chen, "Lecture 4: Rayleigh Quotient," San Jose State University, p.10-11. Available at: `https://www.sjsu.edu/faculty/guangliang.chen/Math253S20/lec4RayleighQuotient.pdf`.

$$u^\intercal \nabla^2 L(\theta) u \le \frac{1}{4n} \lambda_{\max}(X^\intercal X) u^\intercal u$$

$$\Rightarrow \nabla^2 L(\theta) \le \frac{\lambda_{\max}(X^\intercal X)}{4n} I \quad \square$$

3

(1)

The optimization problem is defined as follows:

$$x_\star \in \underset{x \in \Delta_d}{\operatorname{argmin}} f(x), \quad f(x) := -\sum_{i=1}^{n} w_i \log \langle a_i, x \rangle, \tag{P}$$

To show that P is convex, same theorem and approach used in problem 1.(1) will be used again.

First, we need to show that the objective function f is convex, and the constraint set Δ_d is convex.

Prove that constraint set Δ_d is convex

To prove this by definition, we need to show that for any $x_1, x_2 \in \Delta_d$ and $\lambda \in [0, 1]$,

$$x' = \lambda x_1 + (1 - \lambda) x_2 \in \Delta_d$$

By the definition of Δ_d:

$$\Delta_d := \left\{ x = (x[1], \ldots, x[d]) \in \mathbb{R}^d \,\middle|\, x[i] \ge 0, \sum_{i=1}^{d} x[i] = 1 \right\}$$

this means that x' should satisfy:

$$x'[i] \ge 0, \quad \forall i \in \{1, \ldots, d\}$$

$$\sum_{i=1}^{d} x'[i] = 1$$

Since $x_1, x_2 \in \Delta_d$, we knew that:

$$x[i] \ge 0, \quad y[i] \ge 0 \quad \forall i \in \{1, \ldots, d\}$$

Also, $\lambda \in [0,1]$, so both λ and $1-\lambda$ are nonnegative, so for arbitrary $i \in \{1,\ldots,d\}$, we have:

$$x'[i] = \lambda x_1[i] + (1-\lambda)x_2[i] \geq 0$$

Next, we need to check the sum:

$$\begin{aligned}\sum_{i=1}^{d} x'[i] &= \sum_{i=1}^{d} \lambda x_1[i] + (1-\lambda)x_2[i] \\ &= \lambda \sum_{i=1}^{d} x_1[i] + (1-\lambda)\sum_{i=1}^{d} x_2[i] \\ &= \lambda \cdot 1 + (1-\lambda)\cdot 1 \\ &= 1\end{aligned}$$

Therefore, $x' \in \Delta_d$, and Δ_d is convex.

Prove that objective function f is convex

To use the theorem in 1.(1) , we need to first show that f is twice differentiable, then prove that its Hessian matrix is positive semidefinite.

First, we show that f is twice differentiable.

$$\begin{aligned} f(x) &= -\sum_{i=1}^{n} w_i \log\langle a_i, x\rangle \\ \Rightarrow \nabla f(x) &= -\sum_{i=1}^{n} w_i \frac{a_i}{\langle a_i, x\rangle}\end{aligned}$$

Since $\langle a_i, x\rangle$ is linear in x, and a_i is an entry-wise nonnegative vector $\ni a_i \neq 0$, thus $\langle a_i, x\rangle > 0$. Therefore, for each point x in the domain of f, $\nabla f(x)$ exists, and f is differentiable.

Then we check the second derivative:

$$
\begin{aligned}
\nabla^2 f(x) &= -\sum_{i=1}^{n} w_i \nabla \left(\frac{a_i}{\langle a_i, x \rangle} \right) \\
&= -\sum_{i=1}^{n} w_i \frac{\nabla(a_i)\langle a_i, x\rangle - a_i \nabla(\langle a_i, x\rangle)}{\langle a_i, x\rangle^2} \\
&= -\sum_{i=1}^{n} w_i \frac{0 - a_i a_i^\intercal}{\langle a_i, x\rangle^2} \\
&= \sum_{i=1}^{n} w_i \frac{a_i a_i^\intercal}{\langle a_i, x\rangle^2}
\end{aligned}
$$

It is trivial that $\nabla^2 f(x)$ exists, so we can then check if it is positive semidefinite. For any $x \in \Delta_d$, we have:

$$
\begin{aligned}
x^\intercal \nabla^2 f(x) x &= x^\intercal \sum_{i=1}^{n} w_i \frac{a_i a_i^\intercal}{\langle a_i, x\rangle^2} x \\
&= \sum_{i=1}^{n} w_i \frac{x^\intercal a_i a_i^\intercal x}{\langle a_i, x\rangle^2} \\
&= \sum_{i=1}^{n} w_i \frac{(a_i^\intercal x)^2}{\langle a_i, x\rangle^2} \geq 0
\end{aligned}
$$

We have the above expression ≥ 0 since w_i are given nonnegative. Therefore, $\nabla^2 f(x)$ is positive semidefinite, and f is convex. $\square$

(2)

As stated in the textbook, the ellipsoid method computes $g(x^{(k)})$ via formula (3.2) in the previous cutting plane method. [6] Which is given by:

$$
H^{(k)} = \{x \in \mathbb{R}^n \mid g(x^{(k)})^\intercal (x - x^{(k)}) \leq 0\}
$$

Also, in the remark of Problem 3.(1) [7], it said that we set $g(x) = \nabla f(x)$, and the gurantee for $g(\cdot)$ implies that K lies in the half space:

[6] Y.-T. Wong, *Techniques in optimization and sampling* (Draft),p.30. Available at: `https://github.com/YinTat/optimizationbook`.

[7] Y.-T. Wong, *Techniques in optimization and sampling* (Draft),p.29. Available at: `https://github.com/YinTat/optimizationbook`.

$$H^{(k)} = \{y \mid g(x^{(k)})^\intercal (y - x^{(k)}) \leq 0\}$$

Which means that the mapping $g(\cdot)$ is chosen by calculating the gradient of $f(x)$. And this has already been done in the previous subproblem 3.(1).

Therefore, we have:

$$g(x) = \nabla f(x) = -\sum_{i=1}^{n} w_i \frac{a_i}{\langle a_i, x\rangle} \quad \square$$

(3)

From the original definition of the ellipsoid, we have:

$$E^{(k)} = \{y \in \mathbb{R}^n \mid (y - x^{(k)})^\intercal (A^{(k)})^{-1} (y - x^{(k)}) \leq 1\} \tag{1}$$

in the proof of Lemma 3.3, it said that the affine transformation would transform it into the following by setting $A^{(k)} = I, \quad x^{(k)} = 0, \quad v(x^{(k)}) = e_1$:

$$E^{(k)} = \{y \in \mathbb{R}^n \mid y^\intercal y \leq 1\} \tag{2}$$

And would therfore make the halfspace be transformed from its orignal definition (which we mentioned in the previous subproblem 3.(2)):

$$H^{(k)} = \{x \in \mathbb{R}^n \mid g(x^{(k)})^\intercal (x - x^{(k)}) \leq 0\}$$

to:

$$H^{(k)} = \{x \mid x_1 \leq 0\}$$

So, since affine transformation would not change set containment, suppose we have $y_1 \in E^{(k)}$, then it will satisfy:

$$(y_1 - x^{(k)})^\intercal (A^{(k)})^{-1} (y_1 - x^{(k)}) \leq 1 \tag{*}$$

based on the definition in (1), and after some affine transformation:

$$y_1' = Ay_1 + b \in E^{(k)} \tag{3}$$

and y_1' satisfies:

$$(y_1')^\intercal y_1' \leq 1 \tag{4}$$

based on the definition in (2).

Thus we can derive the following by plugging (3) into (4):

$$\begin{aligned}
&(Ay_1 + b)^\intercal (Ay_1 + b) \leq 1 \\
\Rightarrow\ &(y_1^\intercal A^\intercal + b^\intercal)(Ay_1 + b) \leq 1 \\
\Rightarrow\ &y_1^\intercal A^\intercal A y_1 + y_1^\intercal A^\intercal b + b^\intercal A y_1 + b^\intercal b \leq 1
\end{aligned}$$

Expand $(*)$:

$$\begin{aligned}
&(y_1 - x^{(k)})^\intercal (A^{(k)})^{-1}(y_1 - x^{(k)}) \leq 1 \\
\Rightarrow\ &(y_1^\intercal - (x^{(k)})^\intercal)(A^{(k)})^{-1}(y_1 - x^{(k)}) \leq 1 \\
\Rightarrow\ &\left[y_1^\intercal (A^{(k)})^{-1} - (x^{(k)})^\intercal (A^{(k)})^{-1}\right](y_1 - x^{(k)}) \leq 1 \\
\Rightarrow\ &y_1^\intercal (A^{(k)})^{-1} y_1 - y_1^\intercal (A^{(k)})^{-1} x^{(k)} - x^{(k)\intercal}(A^{(k)})^{-1} y_1 + x^{(k)\intercal}(A^{(k)})^{-1} x^{(k)} \leq 1
\end{aligned}$$

Compare the two expressions, and we can observe that:

$$\begin{aligned}
&y_1^\intercal A^\intercal A y_1 = y_1^\intercal (A^{(k)})^{-1} y_1 \\
\Rightarrow\ &A^\intercal A = (A^{(k)})^{-1} \\
\Rightarrow\ &-y_1^\intercal (A^{(k)})^{-1} x^{(k)} = -y_1^\intercal A^\intercal A x^{(k)} = y_1^\intercal A^\intercal b \\
\Rightarrow\ &b = -Ax^{(k)}
\end{aligned}$$

Verify the result:

$$\begin{aligned}
-x^{(k)\intercal}(A^{(k)})^{-1} y_1 &= -x^{(k)\intercal} A^\intercal A y_1 = b^\intercal A y_1 \\
x^{(k)\intercal}(A^{(k)})^{-1} x^{(k)} &= (-Ax^{(k)})^\intercal (-Ax^{(k)}) = b^\intercal b
\end{aligned}$$

Therefore, we have:

$$\begin{cases} A^\intercal A = (A^{(k)})^{-1} \\ b = -Ax^{(k)} \end{cases}$$

To further derive A, we knew that A is invertible, and $A^\intercal A = (A^\intercal A)^\intercal$, so $A^\intercal A$ is Hermitian. Also, consider any $u \in \mathbb{R}^n, \quad u \neq 0$, we have:

$$u^\intercal (A^\intercal A) u = (Au)^\intercal (Au) = \|Au\|^2 > 0$$

Note that we have $>$ since A is invertible and $u \neq 0$. Thus, $A^\intercal A$ is positive definite.

Using the above result ($A^\intercal A$: positive definite, Hermitian), we knew that $A^\intercal A$ can be written as a product of its square root matrix. [8]

$$(A^\intercal A)^{\frac{1}{2}} = ((A^{(k)})^{-1})^{\frac{1}{2}}$$

Therefore, we have:

$$A = ((A^{(k)})^{-1})^{\frac{1}{2}}$$

[9]

And we conclude that the affine transformation is given by:

$$Ay_1 + b \quad \text{where } A = ((A^{(k)})^{-1})^{\frac{1}{2}}, \quad b = -Ax^{(k)} \quad \square$$

(4)

[8] Wikipedia contributors, "Cholesky decomposition," *Wikipedia, The Free Encyclopedia*, Available at: `https://en.wikipedia.org/wiki/Cholesky_decomposition`. Accessed: Mar. 10, 2025.

[9] Wikipedia contributors, "Square root of a matrix," *Wikipedia, The Free Encyclopedia*, Available at: `https://en.wikipedia.org/wiki/Square_root_of_a_matrix`. Accessed: Mar. 10, 2025.