

# Optimization Algorithms: HW0

Lo Chun, Chou  
R13922136

March 10, 2025

-0 1

(1)

To show that the optimization problem defining  $w^*$  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.**<sup>1</sup> Assume that a function  $f$  is twice differentiable, then  $f$  is convex  $\Leftrightarrow \text{dom} f$  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:

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

Then, we'll get:

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

---

<sup>1</sup>S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, pp. 71.

Differentiate w.r.t.  $w$ :

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

Then the second derivative:

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

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

The domain of  $g(w)$  is  $\mathbb{R}^d$ , which is convex.<sup>2</sup>

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

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

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:

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

---

<sup>2</sup>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:

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

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

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

Plugging back into (\*), we get:

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

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

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

Reorder and simplify the terms, we get:

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

Since  $\text{rank}(X) = d$ ,  $\text{rank}(X^\top X) = \text{rank}(X) = d$ ,  $X^\top X \in \mathbb{R}^{d \times d}$  is of full rank and therefore invertible (and positive definite).

Thus, we can write:

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

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

## 2



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

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

We knew that the probability of  $P(y_i = 0)$  is:

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

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

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

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:

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

The terms in the above equation can be simplified:

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

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

---

<sup>3</sup>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:

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

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.

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

Since  $e^{(-2y_i)}$  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:

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



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  $L(\theta) := \frac{1}{n} \sum_{i=1}^n \log \left( 1 + e^{-2(y_i - \frac{1}{2})\langle x_i, \theta \rangle} \right)$  is convex.

Following the same steps in 1.(1), we first differentiate  $L(\theta)$  w.r.t.  $\theta$ :

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

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

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

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

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

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

Then we'll have:

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

Getting back to the Hessian, we have:

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

Since  $(1+e^{-z_i})^2$  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:

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

For the denominator,  $(1+e^{-z_i})^2 > 0$ , 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  $v^\top x_i x_i^\top v$ , this is equivalent to  $(v^\top x_i)^2$ , which is non-negative. Therefore, we have:

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

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

(2)  
 Let:

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

By the previous subproblem, we have:

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

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

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

**Right-hand side of (1)**

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

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

So:

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

By subproblem (1), we have:

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

Therefore, its expected value is:

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



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

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

**Case.** For  $y_i = 1$

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

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

**Case.** For  $y_i = 0$

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

### Left-hand side of (1)

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

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

**Case.** For  $y_i = 1$

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

**Case.** For  $y_i = 0$

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

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

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

is actually the same:

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



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

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

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

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

Need to prove that:

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

To calculate  $\text{Var}(y_i)$ , we knew that:

$$Var(y_i) = E[y_i^2] - E[y_i]^2$$

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

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

So we can derive the following:

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

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

Hence, we have:

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

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

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

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

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

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.  $\square$



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

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

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:

$$\begin{aligned}
u^\top \left( \frac{\lambda_{\max}(X^\top X)}{4n} I - \nabla^2 L(\theta) \right) u &\geq 0 \\
\Rightarrow u^\top \frac{\lambda_{\max}(X^\top X)}{4n} I u &\geq u^\top \nabla^2 L(\theta) u
\end{aligned}$$

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

$$\begin{aligned}
u^\top \nabla^2 L(\theta) u &= u^\top \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \frac{-2(1-2y_i)(y_i - \frac{1}{2})e^{-z_i}}{(1 + e^{-z_i})^2} u \\
&= \frac{1}{n} \sum_{i=1}^n u^\top x_i x_i^\top u \frac{(-2(1-2y_i)(y_i - \frac{1}{2})e^{-z_i})}{(1 + e^{-z_i})^2} \\
&= \frac{1}{n} \sum_{i=1}^n u^\top x_i x_i^\top u \frac{-2(1-2y_i)(y_i - \frac{1}{2})e^{-z_i}}{(1 + e^{-z_i})^2} \quad \text{where } z_i = 2(y_i - \frac{1}{2})\langle x_i, \theta \rangle
\end{aligned}$$

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

**Case.** For  $y_i = 1$

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

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

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

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

Differentiate:

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

Set  $f'(x) = 0$ :

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

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

Calculate  $f(0)$ :

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

Check the second derivative:

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

Evaluate at  $x = 0$ :

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

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

**Case.** For  $y_i = 0$

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

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

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

Differentiate:

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

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:

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

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:

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

Which we could observe that since:

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

We have:

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

By definition, since  $X^\top X \in \mathbb{R}^{p \times p}$  is a symmetric matrix, let  $Q(u) = u^\top X^\top X u$  is its quadratic form.<sup>4</sup>

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

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

Therefore, we have:

---

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

<sup>5</sup>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>.



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

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

### 3

(1) 

The optimization problem is defined as follows:

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

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  $\Delta_d$  is convex.

#### Prove that constraint set $\Delta_d$ 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  $\Delta_d$ :

$$\Delta_d := \left\{ x = (x[1], \dots, x[d]) \in \mathbb{R}^d \mid x[i] \geq 0, \sum_{i=1}^d x[i] = 1 \right\}$$

this means that  $x'$  should satisfy:

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

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

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

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

Also,  $\lambda \in [0, 1]$ , so both  $\lambda$  and  $1 - \lambda$  are nonnegative, so for arbitrary  $i \in \{1, \dots, d\}$ , 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  $\Delta_d$  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:

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

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:

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

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

(2)



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

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

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

<sup>6</sup>Y.-T. Wong, \*Techniques in optimization and sampling\* (Draft),p.30. Available at: <https://github.com/YinTat/optimizationbook>.

<sup>7</sup>Y.-T. Wong, \*Techniques in optimization and sampling\* (Draft),p.29. Available at: <https://github.com/YinTat/optimizationbook>.

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

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:

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

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

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

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

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

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:

$$(y_1 - x^{(k)})^\top (A^{(k)})^{-1} (y_1 - x^{(k)}) \leq 1 \quad (*)$$

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

$$y'_1 = Ay_1 + b \in E^{(k)} \quad (3)$$

and  $y'_1$  satisfies:

$$(y'_1)^\top y'_1 \leq 1 \quad (4)$$

based on the definition in (2).

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

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

Expand (\*):

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

Compare the two expressions, and we can observe that:

$$\begin{aligned} & y_1^\top A^\top Ay_1 = y_1^\top (A^{(k)})^{-1} y_1 \\ \Rightarrow & A^\top A = (A^{(k)})^{-1} \\ \Rightarrow & -y_1^\top (A^{(k)})^{-1} x^{(k)} = -y_1^\top A^\top Ax^{(k)} = y_1^\top A^\top b \\ \Rightarrow & b = -Ax^{(k)} \end{aligned}$$

Verify the result:

$$\begin{aligned} & -x^{(k)\top} (A^{(k)})^{-1} y_1 = -x^{(k)\top} A^\top Ay_1 = b^\top Ay_1 \\ & x^{(k)\top} (A^{(k)})^{-1} x^{(k)} = (-Ax^{(k)})^\top (-Ax^{(k)}) = b^\top b \end{aligned}$$

Therefore, we have:

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

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

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

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

Using the above result ( $A^\top A$ : positive definite, Hermitian), we knew that  $A^\top A$  can be written as a product of its square root matrix.<sup>8</sup>

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

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)

(5)

In algorithm 5, the initial ellipsoid  $E^{(0)}$  is given by:

$$E^{(0)} = \{y \in \mathbb{R}^n \mid (y - x^{(0)})^\top (A^{(0)})^{-1} (y - x^{(0)}) \leq 1\}$$

and in theorem 3.7, it said that  $\Omega$  is any set that is a subset of  $E^{(0)}$ , therefore, if we choose some point  $y_0 \in \Omega \subset E^{(0)}$ , where  $y_0 \notin \Delta_d$ , for example, there are negative values in some dimension of  $y_0$  (i.e.  $y_0[k_i] < 0$  for some  $k_i \in \{1, \dots, d\}$ ,  $i = 1, \dots, n$ ), such that:

$$\langle a_i, y_0 \rangle < 0, \quad \text{for some } i \in \{1, \dots, n\}$$

then  $f(x)$  is undefined at  $x = y_0$ , since the definition of  $f(x)$  ( $f(x) := -\sum_{i=1}^n w_i \log \langle a_i, x \rangle$ ) contains logarithm of  $\langle a_i, x \rangle$ .

<sup>8</sup>Wikipedia contributors, "Cholesky decomposition," *Wikipedia, The Free Encyclopedia*, Available at: [https://en.wikipedia.org/wiki/Cholesky\\_decomposition](https://en.wikipedia.org/wiki/Cholesky_decomposition). Accessed: Mar. 10, 2025.

<sup>9</sup>Wikipedia contributors, "Square root of a matrix," *Wikipedia, The Free Encyclopedia*, Available at: [https://en.wikipedia.org/wiki/Square\\_root\\_of\\_a\\_matrix](https://en.wikipedia.org/wiki/Square_root_of_a_matrix). Accessed: Mar. 10, 2025.

Therefore, we can not imply the inequality given in the proof:

$$f(x^{(i)}) \leq f(y) \leq (1 - \alpha)f(x^*) + \alpha f(z)$$

by the convexity of  $f$ , since  $f$  is not well-defined everywhere in  $E^{(0)}$ .

Also, the feasible set of  $P$  (the probability simplex) is given by:

$$\Delta_d := \left\{ x = (x[1], \dots, x[d]) \in \mathbb{R}^d \mid x[i] \geq 0, \sum_{i=1}^d x[i] = 1 \right\}$$

the set  $\Delta_d$  is in fact not of dimension  $d$ , but  $d - 1$ , since the last dimension is determined by  $1 - \sum_{i=1}^{d-1} x[i]$ .

Therefore the mapping  $V$  can not be properly defined, since  $V$  maps from subsets of  $\mathbb{R}^n$  and  $\Delta_d \not\subseteq \mathbb{R}^n$ .<sup>10</sup>  $\square$

---

<sup>10</sup>I asked ChatGPT for explaining and giving possible reasons but formulated the above proof by myself without copying anything.