Optimization Algorithms: HW0

Lo Chun, Chou
R13922136

February 25, 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 $\Leftrightarrow$ $\operatorname{dom} f$ 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:

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

Then the second derivative:

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

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:

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

[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:

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

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 we can only show that $X^\intercal X$ is positive semi-definite, we cannot guarantee that the inverse $(X^\intercal X)^{-1}$ exists, so we should take the Moore-Penrose pseudo-inverse, which is uniquely defined for any matrix:

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

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:

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

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

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

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:

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

The terms in the above equation can be simplified:

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

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

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

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

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

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.

Thus, we should compute the gradient of the negative log-likelihood w.r.t. θ, set to 0 and solve for θ [4] :

$$
\begin{aligned}
\nabla\left(-\log l(\theta)\right) &= \nabla\left(-\sum_{i=1}^{n}\left[-\langle x_i,\theta\rangle(1-y_i) - \log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right]\right) \\
&= \nabla\left(\sum_{i=1}^{n}\langle x_i,\theta\rangle(1-y_i)\right) + \nabla\left(\sum_{i=1}^{n}\log\left(1+\mathrm{e}^{-\langle x_i,\theta\rangle}\right)\right) \\
&= \sum_{i=1}^{n}\left[x_i(1-y_i) - x_i\frac{\mathrm{e}^{-\langle x_i,\theta\rangle}}{1+\mathrm{e}^{-\langle x_i,\theta\rangle}}\right] = 0
\end{aligned}
$$

If $\theta^\natural$ is given by:

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

Then $\hat{\theta}_n$ should also satisfy $\nabla L(\theta) = 0$, so we have:
In order to calculate it, we can use the fact that the term inside the summation is of the form:

$$
\log(1+e^z)
$$

So taking the derivative w.r.t. z gives:

$$
\frac{d}{dz}\log(1+e^z) = \frac{e^z}{1+e^z}
$$

Therefore we have $z = -2(y_i - \frac{1}{2})\langle x_i,\theta\rangle$, and combined using the chain rule will give:

[4] Deisenroth, Marc Peter, Faisal, A. Aldo, Ong, Cheng Soon, *Mathematics for Machine Learning*, Cambridge University Press, 2020, pp. 351.

$$
\begin{aligned}
\nabla L(\theta) &= \nabla\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}\right)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\nabla\left(\log\left(1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}\right)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}}\cdot\nabla\left(-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}}\cdot -2(y_i-\frac{1}{2})x_i = 0
\end{aligned}
$$

Simplify:

$$
\sum_{i=1}^{n}\frac{(1-y_i)x_i}{1+\mathrm{e}^{-2(y_i-\frac{1}{2})\langle x_i,\theta\rangle}} = 0 \tag{1}
$$

Observe that when $y_i = 1$, the term in the summation will become 0, and if $y_i = 0$, the term will become:

$$
\frac{x_i}{1+e^{\langle x_i,\theta\rangle}}
$$

Thus, we can rewrite (1) as:

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

(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:

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

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:

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

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

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

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

Then we'll have:

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

Getting back to the Hessian, we have:

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

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:

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

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

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

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:

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

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:

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

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

$$
\sum_{i=1}^{n} \frac{x_i(1-y_i)\mathrm{e}^{-z_i}}{1+\mathrm{e}^{-z_i}} = X^\intercal(y-\mathsf{E}[y])
$$

=== could be wrong ===

And since $\theta^\natural$ is the true parameter, this implies that it would minimize the error function $L(\theta)$, which is equivalent to satisfy:

$$\nabla L(\theta^\natural) = 0$$