HW4

October 7, 2017

1 ANLY 561 HW

Name:Yuqi Wang
NetID:yw545

1.1 Problem 1

Prove that

$$\nabla(f \circ g)(x, y) = Dg(x, y)^T \nabla f(g(x, y))$$

using the chain rule

$$(f \circ \gamma)'(t) = \nabla(f(\gamma(t))^T \gamma'(t)$$

PROOF:

$$\nabla(f \circ g)(x, y) = \begin{pmatrix} \frac{\partial(f \circ g)(x,y)}{\partial x} \\ \frac{\partial(f \circ g)(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} \nabla f(g(x, y))^T \frac{\partial g(x,y)}{\partial x} \\ \nabla f(g(x, y))^T \frac{\partial g(x,y)}{\partial y} \end{pmatrix} \text{ (Using the chain rule)}$$

Since

$$\nabla f(g(x, y))^T = \begin{pmatrix} \frac{\partial f(g(x,y))}{\partial g_1(x,y)} & \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \end{pmatrix} \text{ and } \frac{\partial g(x, y)}{\partial x} = \begin{pmatrix} \frac{\partial g_1(x,y)}{\partial x} \\ \frac{\partial g_2(x,y)}{\partial x} \end{pmatrix}, \frac{\partial g(x, y)}{\partial y} = \begin{pmatrix} \frac{\partial g_1(x,y)}{\partial y} \\ \frac{\partial g_2(x,y)}{\partial y} \end{pmatrix}$$

Then,

$$\begin{pmatrix} \nabla f(g(x, y))^T \frac{\partial g(x,y)}{\partial x} \\ \nabla f(g(x, y))^T \frac{\partial g(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \frac{\partial f(g(x,y))}{\partial g_1(x,y)} & \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \end{pmatrix} \begin{pmatrix} \frac{\partial g_1(x,y)}{\partial x} \\ \frac{\partial g_2(x,y)}{\partial x} \end{pmatrix} \\ \begin{pmatrix} \frac{\partial f(g(x,y))}{\partial g_1(x,y)} & \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \end{pmatrix} \begin{pmatrix} \frac{\partial g_1(x,y)}{\partial y} \\ \frac{\partial g_2(x,y)}{\partial y} \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \frac{\partial f(g(x,y))}{\partial g_1(x,y)} \frac{\partial g_1(x,y)}{\partial x} + \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \frac{\partial g_2(x,y)}{\partial x} \\ \frac{\partial f(g(x,y))}{\partial g_1(x,y)} \frac{\partial g_1(x,y)}{\partial y} + \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \frac{\partial g_2(x,y)}{\partial y} \end{pmatrix}$$

For the right side,

$$Dg(x, y)^T \nabla f(g(x, y)) = \begin{pmatrix} \frac{\partial g_1(x,y)}{\partial x} & \frac{\partial g_2(x,y)}{\partial x} \\ \frac{\partial g_1(x,y)}{\partial y} & \frac{\partial g_2(x,y)}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{\partial f(g(x,y))}{\partial g_1(x,y)} \\ \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \end{pmatrix} = \begin{pmatrix} \frac{\partial f(g(x,y))}{\partial g_1(x,y)} \frac{\partial g_1(x,y)}{\partial x} + \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \frac{\partial g_2(x,y)}{\partial x} \\ \frac{\partial f(g(x,y))}{\partial g_1(x,y)} \frac{\partial g_1(x,y)}{\partial y} + \frac{\partial f(g(x,y))}{\partial g_2(x,y)} \frac{\partial g_2(x,y)}{\partial y} \end{pmatrix}$$

Therefore, $\nabla(f \circ g)(x, y) = Dg(x, y)^T \nabla f(g(x, y))$.

1.2 Problem 2

1.2.1 Part (a)

$$\ell_{lin}(\beta_0, \beta_1) = \frac{1}{10}\sum_{i=1}^{10}(y_i - \beta_1 x_i - \beta_0)^2$$

We can calculate $\nabla^2 \ell_{lin}$ first, and then varify it's a positive definite matrix.

$$\nabla^2 \ell_{lin} = \begin{pmatrix} \frac{\partial^2 \ell_{lin}}{\partial^2 \beta_0} & \frac{\partial^2 \ell_{lin}}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell_{lin}}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ell_{lin}}{\partial^2 \beta_1} \end{pmatrix}$$

Since

$$\begin{aligned}
\frac{\partial^2 \ell_{lin}}{\partial^2 \beta_0} &= \frac{\partial^2}{\partial^2 \beta_0}\frac{1}{10}\sum_{i=1}^{10}(y_i - \beta_1 x_i - \beta_0)^2 = \frac{1}{10}\sum_{i=1}^{10}\frac{\partial^2}{\partial^2 \beta_0}(y_i - \beta_1 x_i - \beta_0)^2 \\
&= \frac{1}{10}\sum_{i=1}^{10}\frac{\partial}{\partial \beta_0} - 2(y_i - \beta_1 x_i - \beta_0) = \frac{1}{10}\sum_{i=1}^{10} 2 = 2 \\
\frac{\partial^2 \ell_{lin}}{\partial^2 \beta_1} &= \frac{\partial^2}{\partial^2 \beta_1}\frac{1}{10}\sum_{i=1}^{10}(y_i - \beta_1 x_i - \beta_0)^2 = \frac{1}{10}\sum_{i=1}^{10}\frac{\partial^2}{\partial^2 \beta_1}(y_i - \beta_1 x_i - \beta_0)^2 \\
&= \frac{1}{10}\sum_{i=1}^{10}\frac{\partial}{\partial \beta_0} - 2x_i(y_i - \beta_1 x_i - \beta_0) = \frac{1}{10}\sum_{i=1}^{10} 2x_i^2 = \frac{7}{5} \\
\frac{\partial^2 \ell_{lin}}{\partial \beta_0 \partial \beta_1} &= \frac{\partial^2}{\partial \beta_0 \partial \beta_1}\frac{1}{10}\sum_{i=1}^{10}(y_i - \beta_1 x_i - \beta_0)^2 = \frac{1}{10}\sum_{i=1}^{10}\frac{\partial^2}{\partial \beta_0 \partial \beta_1}(y_i - \beta_1 x_i - \beta_0)^2 \\
&= \frac{1}{10}\sum_{i=1}^{10}\frac{\partial}{\partial \beta_1} - 2(y_i - \beta_1 x_i - \beta_0) = \frac{1}{10}\sum_{i=1}^{10} 2x_i = \frac{1}{5}
\end{aligned}$$

Therefore,

$$\nabla^2 \ell_{lin} = \begin{pmatrix} 2 & \frac{1}{5} \\ \frac{1}{5} & \frac{7}{5} \end{pmatrix}$$

Then we calculate the determinant of $\nabla^2 \ell_{lin}$ to prove $\nabla^2 \ell_{lin}$ is a positive definite matrix.

$$\det(\nabla^2 \ell_{lin}) = \det \begin{pmatrix} 2 & \frac{1}{5} \\ \frac{1}{5} & \frac{7}{5} \end{pmatrix} = \frac{14}{5} - \frac{1}{25} = \frac{69}{25} > 0$$

And obviously, $2 > 0$. Therefore, the Sum of Square Errors function, $\ell_{lin}(\beta_0, \beta_1) = \frac{1}{10}\sum_{i=1}^{10}(y_i - \beta_1 x_i - \beta_0)^2$ is strictly convex by the Second Order Conditions for Convexity.

The unique minimizer

$\ell_{lin}(\beta_0, \beta_1)$ is strictly convex, so there is a unique minimizer (β_0^*, β_1^*) must satisfy the condition

$$\nabla \ell_{lin}(\beta_0^*, \beta_1^*) = \begin{pmatrix} \partial_1 \ell_{lin}(\beta_0^*, \beta_1^*) \\ \partial_2 \ell_{lin}(\beta_0^*, \beta_1^*) \end{pmatrix} = \mathbf{0}$$

Then,

$$\partial_1 \ell_{lin}(\beta_0^*, \beta_1^*) = \frac{1}{10}\sum_{i=1}^{10} -2(y_i - \beta_1^* x_i - \beta_0^*) = 0$$

$$\partial_2 \ell_{lin}(\beta_0^*, \beta_1^*) = \frac{1}{10}\sum_{i=1}^{10} -2x_i(y_i - \beta_1^* x_i - \beta_0^*) = 0$$

We get

$$\beta_1^* + 10\beta_0^* = 0$$

$$5 - 7\beta_1^* - \beta_0^* = 0$$

Therefore,

$$\beta_0^* = -\frac{5}{69}$$

$$\beta_1^* = \frac{50}{69}$$

1.2.2 Part (b)

$$\ell_{log}(\beta_0, \beta_1) = \frac{1}{10}\sum_{i=1}^{10} log(1 + e^{-y_i(\beta_1 x_i + \beta_0)})$$

Since

$$\frac{\partial^2 \ell_{log}}{\partial^2 \beta_0} = \frac{1}{10}\sum_{i=1}^{10} \frac{\partial^2}{\partial^2 \beta_0} log(1 + e^{-y_i(\beta_1 x_i + \beta_0)})$$

$$= \frac{1}{10}\sum_{i=1}^{10} \frac{\partial}{\partial \beta_0} - y_i + \frac{y_i}{1 + e^{-y_i(\beta_1 x_i + \beta_0)}} = \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 e^{-y_i(\beta_1 x_i + \beta_0)}}{(1 + e^{-y_i(\beta_1 x_i + \beta_0)})^2}$$

$$\frac{\partial^2 \ell_{log}}{\partial^2 \beta_1} = \frac{1}{10}\sum_{i=1}^{10} \frac{\partial^2}{\partial^2 \beta_1} log(1 + e^{-y_i(\beta_1 x_i + \beta_0)})$$

$$= \frac{1}{10}\sum_{i=1}^{10} \frac{\partial}{\partial \beta_1} - y_i x_i + \frac{y_i x_i}{1 + e^{-y_i(\beta_1 x_i + \beta_0)}} = \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i + \beta_0)}}{(1 + e^{-y_i(\beta_1 x_i + \beta_0)})^2}$$

$$\frac{\partial^2 \ell_{log}}{\partial \beta_0 \partial \beta_1} = \frac{1}{10}\sum_{i=1}^{10} \frac{\partial^2}{\partial \beta_0 \partial \beta_1} log(1 + e^{-y_i(\beta_1 x_i + \beta_0)})$$

$$= \frac{1}{10}\sum_{i=1}^{10} \frac{\partial}{\partial \beta_1} - y_i + \frac{y_i}{1 + e^{-y_i(\beta_1 x_i + \beta_0)}} = \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i e^{-y_i(\beta_1 x_i + \beta_0)}}{(1 + e^{-y_i(\beta_1 x_i + \beta_0)})^2}$$

Let's denote that

$$\nabla^2 \ell_{log} = \begin{pmatrix} \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 e^{-y_i(\beta_1 x_i + \beta_0)}}{(1+e^{-y_i(\beta_1 x_i + \beta_0)})^2} & \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i e^{-y_i(\beta_1 x_i + \beta_0)}}{(1+e^{-y_i(\beta_1 x_i + \beta_0)})^2} \\ \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i e^{-y_i(\beta_1 x_i + \beta_0)}}{(1+e^{-y_i(\beta_1 x_i + \beta_0)})^2} & \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i + \beta_0)}}{(1+e^{-y_i(\beta_1 x_i + \beta_0)})^2} \end{pmatrix} = \begin{pmatrix} a_{(1,1)} & a_{(1,2)} \\ a_{(1,2)} & a_{(2,2)} \end{pmatrix}$$

Firstly, it is obvious that $a_{(1,1)} > 0$ since $y_i^2 = 1 > 0$ for all i and the range of exponential function is $(0, \infty)$.

Then we calculate the determinant of $\nabla^2 \ell_{lin}$ to prove $\nabla^2 \ell_{lin}$ is a positive definite matrix.

$$\det(\nabla^2 \ell_{lin}) = \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 e^{-y_i(\beta_1 x_i + \beta_0)}}{(1 + e^{-y_i(\beta_1 x_i + \beta_0)})^2} \frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i + \beta_0)}}{(1 + e^{-y_i(\beta_1 x_i + \beta_0)})^2} - \left(\frac{1}{10}\sum_{i=1}^{10} \frac{y_i^2 x_i e^{-y_i(\beta_1 x_i + \beta_0)}}{(1 + e^{-y_i(\beta_1 x_i + \beta_0)})^2}\right)^2$$

$$= \frac{1}{100}\sum_{i=1}^{10}\frac{y_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2}\sum_{i=1}^{10}\frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2} - \frac{1}{100}\left(\sum_{i=1}^{10}\frac{y_i^2 x_i e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2}\right)^2$$

By the Cauchy–Schwarz inequality,

$$\text{LEFT} = \left(\sum_{i=1}^{10}\frac{y_i^2 x_i e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2}\right)^2 = \left(\sum_{i=1}^{10}\sqrt{\frac{y_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2}}\sqrt{\frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2}}\right)^2$$

$$< \sum_{i=1}^{10}\frac{y_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2}\sum_{i=1}^{10}\frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{(1+e^{-y_i(\beta_1 x_i+\beta_0)})^2} = \text{RIGHT}$$

The LEFT = RIGHT equality doesn't hold because

$$\sqrt{\frac{y_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{\left(1+e^{-y_i(\beta_1 x_i+\beta_0)}\right)^2}} \neq \sqrt{\frac{y_i^2 x_i^2 e^{-y_i(\beta_1 x_i+\beta_0)}}{\left(1+e^{-y_i(\beta_1 x_i+\beta_0)}\right)^2}} \text{ for all } i.$$

Hence

$$\det(\nabla^2 \ell_{lin}) > 0$$

Therefore, $\ell_{log}(\beta_0, \beta_1)$ is strictly convex by the Second Order Conditions for Convexity.

The necessary and sufficient conditions for optimality

The necessary condition for optimality of (β_0^*, β_1^*) is that if (β_0^*, β_1^*) is a minimiser of $\ell_{log}(\beta_0, \beta_1)$, then $\nabla \ell_{log}(\beta_0^*, \beta_1^*) = \mathbf{0}$:

$$\frac{1}{10}\sum_{i=1}^{10} -y_i + \frac{y_i}{1+e^{-y_i(\beta_1 x_i+\beta_0)}} = 0 \text{ and } \frac{1}{10}\sum_{i=1}^{10} -y_i x_i + \frac{y_i x_i}{1+e^{-y_i(\beta_1 x_i+\beta_0)}} = 0$$

The sufficient condition for optimality of (β_0^*, β_1^*) is that if ℓ_{log} is strictly convex, and $\nabla \ell_{log}(\beta_0^*, \beta_1^*) = \mathbf{0}$:

$$\frac{1}{10}\sum_{i=1}^{10} -y_i + \frac{y_i}{1+e^{-y_i(\beta_1 x_i+\beta_0)}} = 0 \text{ and } \frac{1}{10}\sum_{i=1}^{10} -y_i x_i + \frac{y_i x_i}{1+e^{-y_i(\beta_1 x_i+\beta_0)}} = 0$$

Then (β_0^*, β_1^*) is the unique minimiser of $\ell_{log}(\beta_0, \beta_1)$.

1.3 Problem 3

if A is a symmetric 2 by 2 matrix, then any solution to

$$\max_{x\in\mathbb{R}^2} \frac{1}{2}\mathbf{x}^T A\mathbf{x} \text{ subject to } ||\mathbf{x}||^2 = 1$$

is an eigenvector of A corresponding to the largest eigenvalue of A.

Firstly, we know that

$$\max_{x\in\mathbb{R}^2} \frac{1}{2}\mathbf{x}^T A\mathbf{x} = \min_{x\in\mathbb{R}^2} -\frac{1}{2}\mathbf{x}^T A\mathbf{x}$$

Let $f(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T A\mathbf{x}$, $g(\mathbf{x}) = ||\mathbf{x}||^2 - 1$, and $A = \begin{pmatrix} a_{(1,1)} & a_{(1,2)} \\ a_{(1,2)} & a_{(2,2)} \end{pmatrix}$, then $f(x_1, x_2) = -\frac{1}{2}(a_{(1,1)}x_1^2 + 2a_{(1,2)}x_1x_2 + a_{(2,2)}x_2^2)$.

Suppose $\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix}$ is a minimizer of $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x}^*) = ||\mathbf{x}^*||^2 - 1 = 0$.
So $\mathbf{x}^* \neq \mathbf{0}$ and

$$\nabla g(\mathbf{x}^*) = \begin{pmatrix} 2x_1^* \\ 2x_2^* \end{pmatrix} = 2\mathbf{x}^* \neq \mathbf{0}$$

By the Theorem (Lagrange Multipliers), there exists a $\lambda \in \mathbb{R}$ such that $\nabla f(\mathbf{x}^*) = \lambda \nabla g(\mathbf{x}^*)$.

$$f(\mathbf{x}) = f(x_1, x_2) = \begin{pmatrix} -a_{(1,1)}x_1 - a_{(1,2)}x_2 \\ -a_{(1,2)}x_1 - a_{(2,2)}x_2 \end{pmatrix} = -A\mathbf{x}$$

Then

$$\nabla f(\mathbf{x}^*) = -A\mathbf{x}^* = \lambda \nabla g(\mathbf{x}^*) = 2\lambda \mathbf{x}^* - A\mathbf{x}^* = 2\lambda \mathbf{x}^* A\mathbf{x}^* = -2\lambda \mathbf{x}^*$$

So -2λ is an eigenvalue of A and $\mathbf{x}^*$ is its corresponding eigenvector.
Suppose that -2λ is not the largest eigenvalue of A, then there exist a $\lambda' > -2\lambda$ and $\mathbf{x}'$ such that $A\mathbf{x}' = \lambda'\mathbf{x}'$ and $||\mathbf{x}'||^2 = 1$.
Then

$$\frac{1}{2}\mathbf{x}^{*T} A\mathbf{x}^* = \frac{1}{2}\mathbf{x}^{*T}(-2\lambda \mathbf{x}^*) = -\lambda ||\mathbf{x}^*||^2 = -\lambda \frac{1}{2}\mathbf{x}'^T A\mathbf{x}' = \frac{1}{2}\mathbf{x}'^T(\lambda'\mathbf{x}') = \frac{\lambda'}{2}||\mathbf{x}'||^2 = \frac{\lambda'}{2}$$

Since $\lambda' > -2\lambda$, then $\frac{1}{2}\mathbf{x}^{*T} A\mathbf{x}^* < \frac{1}{2}\mathbf{x}'^T A\mathbf{x}'$.
However, we know that $\mathbf{x}^*$ is a solution of $\max_{x \in \mathbb{R}^2} \frac{1}{2}\mathbf{x}^T A\mathbf{x}$, so we get a contradiction.
Thus $\mathbf{x}^*$ is an eigenvector of A corresponding to the largest eigenvalue of A, -2λ.