Problem 1: Exponetial Family

1. Remark poisson distribution $f(k;\lambda) = P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Sufficent statistic carry all the information needed to be inference

$$p(x|\theta) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$$

Consider

$$\text{Possion}(x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda)$$

$$= e^{x\log\lambda} e^{-\lambda} \cdot \frac{1}{x!}$$

let $h(x) = \frac{1}{x!}$, $A(\theta) = \lambda$

then $p(x|\theta) \propto \exp(\theta^t \phi(x))$

$$= \exp(\log\lambda \cdot x)$$

let $\theta^t = \log\lambda$ then $\phi(x) = x$

thus $\phi(x) = [x]$ is the sufficent statistic for possion dist

2. $$p(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

$$= \frac{1}{x!}\exp(x\log\lambda - \lambda)$$

$h(x) = \frac{1}{x!}$

$\theta = \log\lambda$

$\phi(x) = x$

$A(\theta) = \lambda = e^\theta$

by book P214 $p(x|\theta) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$

conjugate prior $p(\theta|\gamma) = h_c(\theta)\exp(\langle\begin{bmatrix}\gamma_1\\ \gamma_2\end{bmatrix}, \begin{bmatrix}\theta\\ -A(\theta)\end{bmatrix}\rangle - A_c(\gamma))$

canonical form $p(\lambda|\alpha,\beta) = h_c(\lambda)\exp(\gamma_1\log\lambda - \gamma_2\lambda - A_c(\gamma))$

(*) $p(\lambda|\alpha,\beta) \propto h_c(\lambda)\, \lambda^{\gamma_1} e^{-\gamma_2\lambda}$

note this close to gamma: $f(x;\alpha,\beta)=\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$

let $\exp(-A_c(\gamma))$ account for $\frac{\beta^\alpha}{\Gamma(\alpha)}$

then let $h_c(x)=1$

$$\begin{bmatrix}\gamma_1\\ \gamma_2\end{bmatrix}=\begin{bmatrix}\alpha-1\\ \beta\end{bmatrix}$$

then $(*) \propto \lambda^{\gamma_1} e^{-\gamma_2\lambda} = \lambda^{\alpha-1} e^{-\beta x}$

conjugate prior is gamma distribution.

Problem 2

consider $L(\theta) = \prod_{i=1}^{N} P(y|x;\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[y_i - (ax_i+b)]^2}{2\sigma^2}}$

$$\log(L(\theta)) = \ell(\theta) = \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \log \prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}(y_i-(ax_i+b))^2}$$

$$= -\frac{1}{2} N \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - (ax_i+b))^2$$

(1)

$$0 \overset{\text{want}}{=} \frac{\partial \ell(\theta)}{\partial a} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(y_i - ax_i - b) \cdot (-x_i)$$

$$= -\frac{1}{\sigma^2}\left[\sum_{i=1}^{N} (y_i - b)x_i - a\sum_{i=1}^{N} x_i^2\right]$$

thus $\sum_{i=1}^{N} x_i^2 \cdot a^{ml} + \sum_{i=1}^{N} x_i \cdot b^{ml} = \sum_{i=1}^{N} x_i y_i \quad (*)$

$$0 \overset{\text{want}}{=} \frac{\partial \ell(\theta)}{\partial b} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(y_i - ax_i - b)$$

thus $\sum_{i=1}^{N} x_i \cdot a^{ml} + N b^{ml} = \sum_{i=1}^{N} y_i \quad (**)$

$N \cdot (*) - \sum_{i=1}^{N} x_i \cdot (**) \Rightarrow$

$$a^{ml}\left(N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right) = N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i$$

$$a^{ml} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

by $(**)$ $\quad b^{ml} = \left(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \cdot a^{ml}\right)\frac{1}{N}$

$$= \left(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \cdot \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}\right) \cdot \frac{1}{N}$$

$$0 \overset{\text{want}}{=} \frac{\partial \ell(\theta)}{\partial \sigma} = -\frac{N}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 4\pi\sigma - 2 \cdot \left(-\frac{1}{2}\right) \frac{1}{\sigma^3} \sum_{i=1}^{N} \left(y_i - (ax_i + b)\right)^2$$

$$= -\frac{N}{\sigma} + \frac{1}{\sigma^3} \left(\sum_{i=1}^{N} (y_i - (ax_i + b)\right)^2$$

thus $$(\sigma^{mL})^2 = \frac{1}{N} \left(\sum_{i=1}^{N} (y_i - (a^{mL}x_i + b^{mL})\right)^2$$

Problem 3

$D=\{(\underline{x}_i, y_i)\}_1^N$ $\underline{x}$ d dimensional

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \theta^t \underline{x}_i)^2$$

$\underline{x}_i : d\times 1$

$\theta : d\times 1$

① let $f(\underline{x};\theta) = \theta^t \underline{x}_i$

then want $\hat{\theta} = \arg\min_{\theta} MSE(\theta)$

note $\theta^t \underline{x}_i = \underline{x}_i^t \theta$

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^t\theta)^2$$

$$\overset{let}{=} \frac{1}{N}(y - X\theta)^t(y - X\theta)$$

where $X = \begin{pmatrix} — x_1^t — \\ \vdots \\ — x_N^t — \end{pmatrix}_{N\times d}$ $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}$

consider $\frac{d\,MSE(\theta)}{d\theta} = \frac{1}{N}\frac{d}{d\theta}(y^t y - 2y^t X\theta + \theta^t X^t X\theta)$

$$= \frac{2}{N}(-y^t X + \theta^t X^t X) \overset{want}{=} 0$$

thus $\theta^t(X^t X) = y^t X$

$\theta^t = y^t X (X^t X)^{-1}$

$\theta = (X^t X)^{-1} X^t y$

$(X^t X)\theta = X^t y$

let $A = X^t X$ is $(d\times N)(N\times d) \Rightarrow d\times d$ $X = \begin{pmatrix} — x_1^t — \\ \vdots \\ — x_N^t — \end{pmatrix}$

let $b = X^t y$ is $(d\times N)(N\times 1) \Rightarrow d\times 1$

Assume A be full rank and cannot take inverse

since otherwise A is singular then θ may has

no solution if $b_j \neq 0$ for $\forall j \in \{1,\dots,d\}$

all infinite solutions

when A is full rank

only soln $\hat{\theta} = A^{-1}b$

② $\hat{\theta} = A^{-1}b$ $A^{-1}: d \times d$ $b: d \times 1$ get (A) need $O(Nd^2)$

inverse matrix A need $O(d^3)$ get b need $O(dN)$

multiplication need $O(d \times 1)$

thus time complexity $O(Nd^2 + d^3)$

Note that if A is full rank we need to assume $N > d$ since otherwise let $X_i = \text{ones}(N, d)$ then A will be singular rank(A) not full ✗

thus time complexity is $O(Nd^2)$

thus complexity time is

$O(Nd^2) + O(d^3)$

$\approx O(Nd^2)$ when A full rank

Problem 4

$f(\underline{x};\theta)=\theta^t\underline{x}$

Gaussian regression model: $y|\underline{x}\sim N(\theta^t\underline{x},\sigma^2)$ σ^2 known

independent prior $\theta_j\sim N(\theta;0,s^2)$

① $$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}=\frac{P(Y|X,\theta)P(\theta)}{\int P(Y|X,\theta)P(\theta)d\theta}$$

$$(*)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}(y_i-\theta^t x_i)^2}\cdot\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi s^2}}e^{-\frac{1}{2s^2}(\theta_j)^2}$$

$$P(D)=\int(*)\,d\theta$$

$$P(\theta|D)=\frac{(*)}{P(D)}$$

$$\log P(\theta|D)=-\frac{N}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\theta^t x_i)^2$$
$$-\frac{d}{2}\log(2\pi s^2)-\frac{1}{2s^2}\sum_{j=1}^{d}\theta_j^2-\log P(D)$$

② $$\log P(\theta|D)=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\theta^t x_i)^2-\frac{1}{2s^2}\sum_{j=1}^{d}\theta_j^2+\text{something not depends on }\theta$$

Note $\theta^t x_i=x_i^t\theta$

$$=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-x_i^t\theta)^2-\frac{1}{2s^2}\sum_{j=1}^{d}\theta_j^2+\text{else}$$

$\underline{X}_{n\times d}$ as

$$=-\frac{1}{2\sigma^2}(\underline{y}-\underline{X}\theta)^t(\underline{y}-\underline{X}\theta)-\frac{1}{2s^2}\theta^t\theta+\text{else}$$

where $\underline{y}=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}$, $\underline{X}=\begin{pmatrix}-x_1^t-\\ -x_n^t-\end{pmatrix}_{n\times d}$ $\theta_{1\times d}^t$

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{1}{2\sigma^2}\frac{d}{d\theta}\left(\underline{y}^t\underline{y} - 2\underline{y}^t\underline{X}\theta + \theta^t\underline{X}^t\underline{X}\theta\right) - \frac{1}{2s^2}\frac{d}{d\theta}(\theta^t\theta)$$

$$= -\frac{1}{2\sigma^2}\left(-2\underline{y}^t\underline{X} + 2\theta^t(\underline{X}^t\underline{X})\right) - \frac{1}{2s^2}\,2\theta^t$$

$$= -\frac{1}{\sigma^2}\left(-\underline{y}^t\underline{X} + \theta^t(\underline{X}^t\underline{X})\right) - \frac{\theta^t}{s^2}$$

$$= \frac{\underline{y}^t\underline{X}}{\sigma^2} - \theta^t\left(\frac{\underline{X}^t\underline{X}}{\sigma^2} + \frac{1}{s^2}\right)$$

③ $\underline{y}$: $n\times 1$ $\underline{X}$: $n\times d$ $\theta^t_{1\times d}$ $O(\underline{X}^t\underline{X}) = O(d\times n\times d) = O(nd^2)$

$$O\left(\frac{\underline{y}^t\underline{X}}{\sigma^2}\right) = O(nd) + O(d)$$

$$O\left(\frac{\theta^t\underline{X}^t\underline{X}}{\sigma^2}\right) = O(d^2n) + O(nd) + O(d)$$

$$O\left(\theta^t/s^2\right) = O(d)$$

O(gradient descent) $= O(MNd^2)$ proof next page

O(normal method) $= O(Nd^2) + O(d^3)$

normal method has larger time complexity if $d > MN$

else gradient decent will larger time complexity

double check for gradient method

$$\sum_{i=1}^{N}(y_i-\theta^T x_i)^2+\sum_{j=1}^{d}\theta_j^2$$

$$(*)=\sum_{i=1}^{N}\Big(y_i-\sum_{j=1}^{d}\theta_j x_{ij}\Big)^2+\sum_{j=1}^{d}\theta_j^2$$

$$=\sum_{i=1}^{N}\Big[y_i^2-2y_i\sum_{j=1}^{d}\theta_j x_{ij}+\underbrace{\sum_{k=1}^{d}\theta_k x_{ik}\sum_{j=1}^{d}\theta_j x_{ij}}_{\sum_{k=1}^{d}\sum_{j=1}^{d}\theta_k\theta_j x_{ik}x_{ij}}\Big]+\sum_{j=1}^{d}\theta_j^2$$

$$\frac{\partial(*)}{\partial\theta_h}=\sum_{i=1}^{N}\Big[\underbrace{-2y_i\theta_h x_{ih}}_{O(1)}+\underbrace{\sum_{j=1}^{d}\theta_j x_{ih}x_{ij}}_{O(d)}+\underbrace{\sum_{k=1}^{d}\theta_k x_{ik}x_{ih}}_{O(d)}\Big]+\underbrace{2\sum_{j=1}^{d}\theta_j}_{O(d)}$$

$$\frac{\partial(*)}{\partial\theta_h}=O(Nd)+O(d) \qquad \theta\ d\times 1$$

$$\frac{\partial(*)}{\partial\theta}=O(Nd^2)$$

Need M steps $O(MNd^2)$

Problem 5

$$MSE(\theta) = \sum_{i=1}^{N} (y_i - f(\underline{x}_i;\theta))^2$$

$$D = \{(\underline{x}_i, y_i)\}$$

$$r(\theta) = \sum_{j=1}^{p} |\theta_j|$$

minimize $MSE(\theta) + \lambda r(\theta)$

$$\log p(\theta|X,y) = \log p(y|X,\theta) + \log p(\theta) + \text{const}$$

new loss function $E(\theta) = \sum_{i=1}^{N} (y_i - f(\underline{x}_i;\theta))^2 + \lambda \sum_{j=1}^{p} |\theta_j|$

WLOG, say $\underline{x}_i$ is $d \times 1$

let $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} -x_1^t- \\ -x_N^t- \end{pmatrix}_{n \times d}$ $\theta^t_{1\times d}$ $\theta_{d\times 1}$

then $E(\theta) = (y - X\theta)^t (y - X\theta) + \lambda \sum_{j=1}^{p} \theta_j$

want $\arg\min_{\theta} E(\theta)$

equal to $\arg\max_{\theta} -\frac{1}{2\sigma^2} E(\theta)$

Consider from ML of Least-Squares Regression of Gaussian in problem 2

$$L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i^t\theta)^2}{2\sigma^2}}$$ can be consider as likelihood

since $\ell(\theta) \propto -\frac{1}{2\sigma^2} E(\theta)$

similarly $\prod_{i=1}^{p} \frac{\lambda_i(\sigma^2)}{2} e^{-\lambda_i(\sigma^2)|\theta|}$ can be considered as prior

called double exponential distribution (Laplace dist) which is symmetric a exponetial dist by y axis and rescale sum to 1

since $\ell(\theta) \propto -\lambda\sigma^2 \sum_{i=1}^{p} |\theta|$

consider rescale $e^{\frac{-(y_i - x_i^t\theta)^2}{2\sigma^2}}$

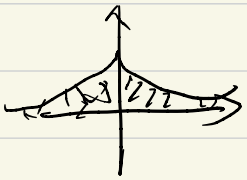
let $\bar{y}_i = y_i/\sqrt{\sigma^2}$, $\bar{\theta} = \theta/\sqrt{\sigma^2}$

then $\Rightarrow e^{\frac{-(\bar{y}_i - x_i^t\bar{\theta})^2}{2}}$ kernel with mean $x_i^t\bar{\theta}$ and variance 1

use same parameters rescaled

$$\lambda^*\cdot(\sigma^2)\, e^{-\lambda_*(\sigma^2)|\theta|} \Rightarrow \frac{\lambda^{**}}{2\sqrt{\sigma^2}} e^{-\lambda^{**}|\bar{\theta}|/\sqrt{\sigma^2}}$$

force sum to 1



$$\Rightarrow \frac{\lambda^{**}}{2\sqrt{\sigma^2}} e^{-\lambda^{**}|\bar{\theta}|/\sqrt{\sigma^2}}$$

thus consider prior as $p(\bar{\theta}|\sigma^2) = \pi(\sigma^2)\prod_{j=1}^{p}\frac{\lambda^{**}}{2\sqrt{\sigma^2}} e^{-\lambda^{**}|\bar{\theta}|/\sqrt{\sigma^2}}$

where $\pi(\sigma^2)$ is hyper prior

check for unimodal

$$p(\bar{\theta}|Y,X,\sigma^2) \propto \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - x_i^t\theta)^2}{2\sigma^2}} \prod_{i=1}^{p}\frac{\lambda^{**}}{2\sqrt{\sigma^2}} e^{-\lambda^{**}|\theta|/\sqrt{\sigma^2}} \cdot \pi(\sigma^2)$$

change of variables let $\bar{\theta} = \theta/\sqrt{\sigma^2}$, $\bar{y} = y/\sqrt{\sigma^2}$

$$\ell(p(\cdot)) = \log(\pi(\sigma^2)) - \frac{n+p+1}{2}\ln(\sigma^2) - \frac{1}{2}\|\bar{y} - X\bar{\theta}\|_2^2 - \lambda^{**}|\bar{\theta}|$$

$\ell(p(\cdot))$ is convex since $\log$, $\|\cdot\|$ are convex operator on $\bar{\theta}$ and σ^2

choose $\pi(\sigma^2)$ to be convex also

then $p(\bar{\theta}|Y,X,\sigma^2)$ is unimodel

also $\ell(p(\cdot)) \propto -\frac{1}{2}\|\bar{y} - X\bar{\theta}\|_2^2 - \lambda^{**}|\bar{\theta}|$

choose $\lambda = \frac{1}{2}\lambda^{**}$

then $-\ell(p(\bar{\theta}|Y, X, \sigma^2)) \propto \|\bar{y} - X\bar{\theta}\|_2^2 + \lambda|\bar{\theta}|$

thus lasso gaussian prior is

$$\pi(\sigma^2) \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\theta|/\sqrt{\sigma^2}}$$

is the double exponential distribution (laplace)

where $\pi(\sigma^2)$ is the hyperparameter distribution for unimodal of posterior

Note that if we define MSE as

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i;\theta))^2$$

then let $\lambda = \frac{\lambda}{N}$

the lasso gaussian prior will be same as above

Problem 6

① $y_i | \underline{x}_i, \theta \sim \text{Poss}(\lambda(\underline{x})) = \text{Poss}(\theta^t \underline{x}) = \text{Poss}(\underline{x}^t \theta)$

$$P(y_i|\lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

recall exponetial family from problem 1

$$p(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$
$$= \frac{1}{x!} \exp(x \log\lambda - \lambda)$$

$h(x) = \frac{1}{x!}$

$\theta = \log\lambda$

$\phi(x) = x$

$A(\theta) = \lambda = e^\theta$

$$p(y|\lambda) = \frac{1}{y!} \exp(y \log\lambda - \lambda)$$

$h(y) = \frac{1}{y!}$

$A(\theta) = \lambda = e^{\theta^t \underline{x}}$

thus $p(y_i|\underline{x}_i, \theta) = \frac{1}{y!} \exp(y_i(\theta^t \underline{x}_i) - e^{\theta^t \underline{x}_i})$

$$P(D|, \theta) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{y_i(\theta^t \underline{x}_i) - e^{\theta^t \underline{x}_i}}$$

$$\ell(\theta) = \sum_{i=1}^{n} (y_i(\theta^t \underline{x}_i) - e^{\theta^t \underline{x}_i} - \log(y_i!))$$

$\theta^t \underline{x}_i$: $1 \times d \times d \times 1$

$\underline{x}_i^t \theta$: $1 \times d \times d \times 1$

$$= \sum_{i=1}^{n} (y_i(\underline{x}_i^t \theta) - e^{\underline{x}_i^t \theta} - \log(y_i!))$$

② $\frac{\partial l(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left(y_i \frac{\partial (x_i^t \theta)}{\partial \theta} \right) - \frac{\partial e^{x_i^t \theta}}{\partial \theta} - 0$

$= \sum_{i=1}^{n} y_i x_i^t - x_i^t e^{\theta^t x_i}$

$= \sum_{i=1}^{n} x_i^t (y_i - e^{\theta^t x_i})$

x_i $d \times 1$

x_i^t $1 \times d$

$e^{x_i^t \theta}$ $1 \times d \times d \times 1$

$y_i x_i^t$: $1 \times 1 \times 1 \times d$

$x_i^t e^{x_i \theta}$: $1 \times d \times$

Problem 7

① $f(x_i;\theta) = \frac{1}{1+e^{(-\theta^t x_i)}}$

$$CE(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \underbrace{\log(f(x_i;\theta))}_{\text{part ①}} + (1-y_i)\underbrace{\log(1-f(x_i;\theta))}_{\text{part ②}}$$

$$\frac{d-\log(f(x_i;\theta))}{d\theta} = \frac{d\log(1+e^{-\theta^t x_i})}{d\theta} = \frac{1}{1+e^{-\theta^t x}}\cdot(-e^{-\theta^t x})\cdot(-x_i)$$

$$= \frac{x e^{-\theta^t x}}{1+e^{-\theta^t x}} = x(1-f(x_i;\theta))$$

$$\frac{d-\log(1-f(x_i;\theta))}{d\theta} = \frac{d\log\left(\frac{e^{-\theta^t x_i}}{1+e^{-\theta^t}}\right)^{-1}}{d\theta} = \frac{d\log(1+e^{-\theta^t x_i}) - \log(e^{-\theta^t x_i})}{d\theta}$$

$= x_i(1-f(x_i;\theta)) + x_i$ first part same as above

thus $\frac{dCE(\theta)}{d\theta} = \frac{1}{N}\sum_{i=1}^{N} y_i x_i (1-f(x_i;\theta)) + (1-y_i)[x_i(1-f(x_i;\theta)) + x_i]$

②

Theorem 2.6.1: *If the function f has a second derivative which is non-negative (positive) everywhere, then the function is convex (strictly convex).*

WTS hessian matrix is SPD (semi-positive definite)

hessian for part ①

$$\frac{\partial\, x_i(1-f(x_i;\theta))}{\partial\theta} = \frac{\partial\, -x_i f(x_i;\theta)}{\partial\theta}$$

$$= \frac{\partial\, \frac{-x_i}{1+e^{-\theta^t x}}}{\partial\theta}$$

$$= \frac{0 + x_i^2 e^{-\theta^t x_i}}{(1+e^{-\theta^t x_i})^2} = f(x_i;\theta)(1-f(x_i;\theta))x_i x_i^t$$

let $z \in \mathbb{R}_{d\times 1}$

then $(*): z^t f(x_i;\theta)(1-f(x_i;\theta)) x_i x_i^t z$

since $f(x_i;\theta) \in (0,1]$

then $f(x_i;\theta)(1-f(x_i;\theta)) \geq 0$

thus $(*)$ = nonnegative number $\cdot (x_i^t z)^2 \geq 0$ SPD

hessian for part ②

$$\frac{\partial\, x_i(1-f(x_i;\theta)) + x_i}{\partial\theta} \overset{\text{same as above}}{=} f(x_i;\theta)(1-f(x_i;\theta)) x_i x_i^t$$

Same as $(*)$ part ② hessain is SPD

Since $0 \leq y^i \in \{0,1\}$ $0 \leq 1-y^i \in \{0,1\}$

$$CE(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(f(x_i;\theta)) + (1-y_i)\log(1-f(x_i;\theta))$$

since if f convex func g convex func $\lambda_1 \geq 0$ $\lambda_2 \geq 0$

then $(\lambda_1 f + \lambda_2 g)(\alpha x + (1-\alpha)y)$

$$\underset{\text{convex}}{=} \lambda_1 f(\alpha x + (1-\alpha)y) + \lambda_2 g(\alpha x + (1-\alpha)y)$$

$$\leq \lambda_1 f\alpha x + \lambda_1 f(1-\alpha)y + \lambda_2 g\alpha x + \lambda_2 g(1-\alpha)y$$

$$= \alpha(\lambda_1 f_1 + \lambda g)x + (1-\alpha)(\lambda_1 f + \lambda_2 g)y$$

$\lambda_1 f + \lambda_2 g$ also convex

thus since $-\log f(x_i;\theta)$ and $-\log(1-f(x_i;\theta))$ convex

$y_i, 1-y_i \geq 0$

thus $CE(\theta)$ is convex