Assignment 2

1. Maximum likelihood estimator:

a) $p(x, y \mid \theta) = p(x \mid \theta_1) \cdot p(y \mid \theta_2, x)$

we already have: $p(x=0 \mid \theta_1) = 1 - \theta_1$

$p(x=1 \mid \theta_1) = \theta_1$

so the joint probability distribution $p(x, y \mid \theta)$:

	$y=0$	$y=1$
$x=0$	$\theta_2 \cdot (1-\theta_1)$	$(1-\theta_2) \cdot (1-\theta_1)$
$x=1$	$(1-\theta_2) \cdot \theta_1$	$\theta_2 \cdot \theta_1$

b) $p(x, y \mid \theta) = p(x \mid \theta_1) \cdot p(y \mid \theta_2, x)$

$\Rightarrow \log(p(x, y \mid \theta)) = \log(p(x \mid \theta_1)) + \log(p(y \mid \theta_2, x))$

$= \log\left(\prod_{i=1}^{n} p(x_i \mid \theta_1)\right) + \log\left(\prod_{i=1}^{n} p(y_i \mid \theta_2, x_i)\right)$

$= \sum_{i=1}^{n} \log(p(x_i \mid \theta_1)) + \sum_{i=1}^{n} \log(p(y_i \mid \theta_2, x_i))$

We can see that we can find MLE of θ_1 and θ_2 independently

We have:

$\prod_{i=1}^{n} p(x_i \mid \theta_1) = \theta_1^{h} \cdot (1-\theta_1)^{t}$ with h is number of heads

t is number of tails

$h + t = n$ is total times

$\Rightarrow \log\left(\prod_{i=1}^{n} p(x_i \mid \theta_1)\right) = \log(\theta_1^{h} \cdot (1-\theta_1)^{t})$

$= h \cdot \log(\theta_1) + t \cdot \log(1-\theta_1)$

So:

$$\frac{\partial}{\partial \theta_1} p(x, y \mid \theta) = \frac{h}{\theta_1} - \frac{t}{1-\theta_1} = 0$$

$\Leftrightarrow (1-\theta_1) \cdot h = \theta_1 \cdot t = \theta_1 \cdot (n-h)$

$\Rightarrow \theta_1 = \frac{h}{n} \Rightarrow \hat{\theta}_1 = \frac{h}{n}$ (*) $= \frac{4}{7}$

Also:

$\prod_{i=1}^{n} p(y_i | \theta_2, x_i) = \theta_2^{s} \cdot (1-\theta_2)^{d}$ with s is nb of times $x = y$

d is nb of times $x \neq y$

$s + d = n$ is total of times

Because it the same like the previous one

so we have

$$\frac{\partial}{\partial \theta_2} p(x, y | \theta) = \frac{s}{\theta_2} - \frac{d}{1-\theta_2} = 0$$

$$\Leftrightarrow \theta_2 = \frac{s}{n} \Rightarrow \hat{\theta}_2 = \frac{s}{n} \quad (*) = \frac{4}{7}$$

Next:

$$p(D | \hat{\theta}, M_2) = \theta_1^{h} \cdot (1-\theta_1)^{t} \cdot \theta_2^{s} \cdot (1-\theta_2)^{d}$$

$$= \frac{4}{7}^{4} \cdot \frac{3}{7}^{3} \cdot \frac{4}{7}^{4} \cdot \frac{3}{7}^{3}$$

$$= \frac{4}{7}^{8} \cdot \frac{3}{7}^{6} = \frac{4^8 \cdot 3^6}{7^{14}} \approx 0,0000704$$

c) $p(x, y | \theta) = p(x, y | \theta_{0,0}) \cdot p(x, y | \theta_{0,1}) \cdot p(x, y | \theta_{1,0}) \cdot p(x, y | \theta_{1,1})$

$= \theta_{0,0}^{n(x=y=0)} \cdot \theta_{0,1}^{n(x=0,y=1)} \cdot \theta_{1,0}^{n(x=1,y=0)} \cdot \theta_{1,1}^{n(x=y=1)}$

$\Leftrightarrow \log(p(x, y | \theta)) = n_{0,0} \cdot \log(\theta_{0,0}) + n_{0,1} \cdot \log(\theta_{0,1})$

$+ n_{1,0} \cdot \log(\theta_{1,0}) + n_{1,1} \cdot \log(\theta_{1,1})$

We can see that we can also find $\theta_{1,0}; \theta_{0,0}; \theta_{0,1}; \theta_{1,1}$ independently

$\frac{\partial}{\partial \theta}$

c) we can see that this case is similar to the previous case with each $\theta_{x,y}$ can be found independently and through $\prod_{i=1}^{n} p(x_i, y_i | \theta)$ and through the same way as before $\Rightarrow \hat{\theta}_{x,y} = \frac{a_{x,y}}{n}$ with a is the number of time we have x and y

Table of a

	$y=0$	$y=1$
$x=0$	2	1
$x=1$	2	2

Table of $\theta_{x,y}$

	$y=0$	$y=1$
$x=0$	2/7	1/7
$x=1$	2/7	2/7

So: $p(\mathcal{D} | \hat{\theta}, M_4) = \theta_{00}^{a_{0,0}} \cdot \theta_{01}^{a_{0,1}} \cdot \theta_{10}^{a_{1,0}} \cdot \theta_{11}^{a_{1,1}}$

$$= \left(\frac{2}{7}\right)^2 \cdot \left(\frac{1}{7}\right)^1 \cdot \left(\frac{2}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^2$$

$$= \frac{2^6}{7^7} \approx 0,0000777$$

2. Biased Estimator:

$\hat{\mu}$ is the empirical mean $\Rightarrow \hat{\mu} = \frac{\sum_{n=1}^{N} x_n}{N}$

$\Rightarrow N.\hat{\mu} = \sum_{n=1}^{N} x_n$

$$E[\hat{\sigma}^2] = E\left[\frac{1}{N}\cdot\sum_{n=1}^{N}(x_n-\hat{\mu})^2\right]$$

$$= \frac{1}{N}\cdot E\left[\sum_{n=1}^{N}(x_n^2 - 2x_n\cdot\hat{\mu}+\hat{\mu}^2)\right]$$

$$= \frac{1}{N}\cdot E\left[\sum_{n=1}^{N}x_n^2 - 2\cdot\hat{\mu}\cdot\sum_{n=1}^{N}x_n + \sum_{n=1}^{N}\hat{\mu}^2\right]$$

$$= \frac{1}{N}\cdot E\left[\sum_{n=1}^{N}x_n^2 - 2\cdot\hat{\mu}\cdot N\cdot\hat{\mu} + \hat{\mu}^2\cdot N\right]$$

$$= \frac{1}{N}\cdot E\left[\sum_{n=1}^{N}x_n^2 - N\cdot\hat{\mu}^2\right]$$

$$= \frac{1}{N}\cdot E\left[\sum_{n=1}^{N}x_n^2\right] - \frac{1}{N}\cdot N\cdot E[\hat{\mu}^2]$$

$$= \frac{1}{N}\cdot E\left[\sum_{n=1}^{N}x_n^2\right] - E[\hat{\mu}^2]$$

$$= E[x^2] - E[\hat{\mu}^2]$$

we have $\sigma_x^2 = E[x^2] - E[x]^2 \Leftrightarrow E[x^2] = \sigma_x^2 + E[x]^2$

$\sigma_{\hat{\mu}}^2 = E[\hat{\mu}^2] - E[\hat{\mu}]^2 \Leftrightarrow E[\hat{\mu}^2] = \sigma_{\hat{\mu}}^2 + E[\hat{\mu}]^2$

So:

$$E[\hat{\sigma}^2] = \sigma_x^2 + E[x]^2 - \sigma_{\hat{\mu}}^2 - E[\hat{\mu}]^2$$

$$= \sigma_x^2 - \sigma_{\hat{\mu}}^2 \quad \text{cause } E[x] = E[\hat{\mu}]$$

So we can see that

$E[\hat{\sigma}^2] \neq \sigma_x^2 \Rightarrow$ MLE of variance is biased

Futher transformation

$$\sigma_{\hat{\mu}}^2 = \text{var}[\hat{\mu}] = \text{var}\left[\frac{1}{N}\cdot\sum_{n=1}^{N}x_n\right] = \frac{1}{N^2}\cdot\text{var}\left[\sum_{n=1}^{N}x_n\right]$$

$$= \frac{1}{N^2}\cdot\sum_{n=1}^{N}\text{var}[x_n] = \frac{1}{N^2}\cdot\sum_{n=1}^{N}\text{var}[x]$$

$$= \frac{1}{N^2}\cdot N\cdot\text{var}[x] = \frac{1}{N}\cdot\sigma_x^2$$

So $E[\hat{\sigma}^2] = \frac{N-1}{N}\cdot\sigma_x^2$

3. MLE for the Poisson Distribution:

a) $p(x_1, \dots, x_n | \lambda) = \prod_{i=1}^{n} p(x_i | \lambda) = \prod_{i=1}^{n} e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}$

$= e^{-n \cdot \lambda} \cdot \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$

$\ln(p(x_1, \dots x_n | \lambda)) = -n \cdot \lambda + \sum_{i=1}^{n} x_i \cdot \ln(\lambda) - \ln\left(\prod_{i=1}^{n} x_i!\right)$

$\Rightarrow \frac{\partial}{\partial \lambda} p(x_1, \dots x_n | \lambda) = -n + \frac{1}{\lambda} \cdot \sum_{i=1}^{n} x_i = 0$

$\Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$

b) $p(\lambda | D) = \frac{p(D | \lambda) \cdot p(\lambda)}{p(D)}$

~~with: $p(D|\lambda) \sim$ Poi~~

with: $p(D|\lambda) \sim \text{Poi}(\lambda) \Rightarrow p(D|\lambda) \propto e^{-\lambda} \cdot \lambda^{x}$

$p(\lambda) = \text{Ga}(\lambda | a, b) \Rightarrow p(\lambda) \propto \lambda^{a-1} \cdot e^{-\lambda b}$

So we can have: (we can ignore $p(D)$)

$p(\lambda | D) \propto e^{-\lambda} \cdot \lambda^{x} \cdot \lambda^{a-1} \cdot e^{-\lambda b}$

$= e^{-\lambda - \lambda b} \cdot \lambda^{x+a-1}$

with $p(x | \lambda) \sim \text{Poi}(\lambda) \Rightarrow p(x | \lambda) \propto e^{-\lambda} \cdot \lambda^{x}$

$p(D | \lambda) = \prod_{i=1}^{n} p(x_i | \lambda) \propto \prod_{i=1}^{n} e^{-\lambda} \cdot \lambda^{x_i} = e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^{n} x_i}$

$p(\lambda) = \text{Ga}(\lambda | a, b) \Rightarrow p(\lambda) \propto \lambda^{a-1} \cdot e^{-\lambda b}$

So we can have: (We can ignore $p(D)$)

$p(\lambda | D) \propto e^{-n\lambda} \cdot \lambda^{\sum x_i} \cdot \lambda^{a-1} \cdot e^{-\lambda b} = e^{-\lambda(n+b)} \cdot \lambda^{(\sum x_i + a) - 1}$

$\Rightarrow p(\lambda | D) = \text{Ga}\left(\lambda \mid \sum_{i=1}^{n} x_i + a, n + b\right)$

c) mean of $Ga(a, b) = \frac{a}{b}$

$\Rightarrow$ mean of $Ga(\sum_{i=1}^{n} x_i + a, b+n) = \frac{\sum_{i=1}^{n} x_i + a}{b+n}$

When $a \to 0$, $b \to 0$ the mean will be $\frac{\sum_{i=1}^{n} x_i}{n}$

$\Rightarrow$ The posterior mean becomes equal to the sample mean of data

4. Bayesian Analysis of Exponential Distribution:

a) $p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta . e^{-\theta x_i}$

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

$\ln(p(D|\theta)) = n . \ln(\theta) - \theta . \sum_{i=1}^{n} x_i$

$\frac{\partial}{\partial \theta} p(D|\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \theta = \frac{n}{\sum_{i=1}^{n} x_i}$

$\Rightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}$

b) Apply the value to $\hat{\theta}$ we have

$\hat{\theta} = \frac{n}{x_1+x_2+x_3} = \frac{3}{5+6+4} = \frac{1}{5}$

c) we already have $E[\theta] = \frac{1}{\lambda} = \frac{1}{3}$

$\Rightarrow \hat{\lambda} = 3$

d) we have $p(\theta | \mathcal{D}, \hat{\lambda}=3) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta|\hat{\lambda}=3)}{p(\mathcal{D})}$

we can ignore $p(\mathcal{D})$ and :

$p(\mathcal{D}|\theta) = \theta^n \cdot e^{-\theta \cdot \sum x_i}$

$p(\theta | \hat{\lambda}=3) = 3 \cdot e^{-3\theta}$

$\Rightarrow p(\theta | \mathcal{D}, \hat{\lambda}=3) \propto \theta^n \cdot e^{-\theta \cdot \sum x_i} \cdot 3 \cdot e^{-3\theta}$

$\propto \theta^n \cdot e^{-\theta(3+\sum x_i)}$

$\Rightarrow p(\theta | \mathcal{D}, \hat{\lambda}=3) \sim Ga(\theta | n+1, 3+\sum x_i)$

$= Ga(\theta | 3+1, 3+15)$

$= Ga(\theta | 4, 18)$

e) No, the exponential prior conjugate to gamma likelihood

6. Off set Term for Linear Regression :

$E[y|x] = w_0 + w^T \cdot x$

So we have likelihood function in linear regression :

$L(y | x, w_0, w, \sigma^2) = \prod_{i=1}^{n} p(y_i | x_i, w_0, w, \sigma^2)$

$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(y_i - E[y|x])^2}{2\sigma^2}}$

$\Rightarrow \log L(y | x, w_0, w, \sigma^2) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(y_i - E[y|x])^2}{2\sigma^2}} \right)$

$= -\frac{n}{2} \cdot \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (y_i - E[y|x])^2$

$= -\frac{n}{2} \cdot \log(2\pi \cdot \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w_0 - w^T \cdot x_i)^2$

To calculate MLE of w_0 .

$$\frac{\partial L}{\partial w_0} = 0 = -\frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^{n} (y_i - w_0 - w^T \cdot x_{.i})$$

$$= \frac{-1}{\sigma^2} \cdot \sum_{i=1}^{n} (y_i - w_0 - w^T \cdot x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (y_i - w_0 - w^T \cdot x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} w_0 - \sum_{i=1}^{n} w^T \cdot x_i = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (y_i - w^T x_i) = \sum_{i=1}^{n} w_0 = n \cdot w_0$$

$$\Leftrightarrow w_0 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - w^T \cdot x_i)$$

So

$$\hat{w}_0 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - w^T \cdot x_i)$$

5. Linear Regression:

$$\hat{y} = w^T \cdot \phi(x) \Rightarrow p(y \mid x, w) = \prod_{j=1}^{M} \mathcal{N}(y_j \mid w^T \cdot \phi(x), \sigma_j^2)$$

So the log-likelihood will be:

$$\log(p(y \mid x, w)) = \sum_{j=1}^{M} \log(p(y_j \mid x_j, w))$$

$$= -\frac{M}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi - \frac{1}{\sigma^2} \left(\frac{1}{2} \sum_{j=1}^{M} (y_j - w^T \phi(x_j))^2 \right)$$

We can see that MLE of w = LS of w (proven in lecture)

$$\Rightarrow \phi^T \cdot \phi \cdot w = \phi^T \cdot y$$

$$\Rightarrow w = (\phi^T \cdot \phi)^{-1} \cdot \phi^T \cdot y$$

$$\Rightarrow \hat{w} = (\phi^T \cdot \phi)^{-1} \cdot \phi^T \cdot y$$

Apply value of x to $\phi(x)$ we have:

$$\phi = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \Rightarrow \phi^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$\Rightarrow \phi^T . \phi = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} = F$$

We have $F^{-1} = \frac{1}{\det(F)} \cdot \frac{1}{3.3 - 0.0} \cdot \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$

$$= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix}$$

$$\Rightarrow F^{-1} . \phi^T = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}$$

We have

$$y = \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} \Rightarrow F^{-1} . \phi^T . y = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}$$

$$\Rightarrow \hat{W} = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}$$