Christopher Swingler
EECE 5644
Prof Yeh
Problem Set 2

2.1 $p(x|\theta) = \theta e^{-\theta x}$, $x \geq 0$, $\theta > 0$

a Likelihood function:

$$L(\theta|x) = \prod_{i=1}^{N} \theta e^{-\theta x_i}$$

easier to work with log likelihood function

$$\ln L(\theta|x) = \sum_{i=1}^{N} \ln\theta - \theta x_i$$

derivate log likelihood and set = 0

$$\frac{d}{d\theta} \ln L(\theta|x) = \frac{d}{d\theta} \sum_{i=1}^{N} \ln\theta - \theta x_i$$

$$= \frac{N}{\theta} - \sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow \hat{\theta} = \frac{N}{\sum_{i=1}^{N} x_i} \Leftrightarrow \hat{\theta} = \frac{1}{\bar{x}} \text{ where } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

b Using the above $\hat{\theta}$ and $\bar{x}$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{3}(5+6+4) = 5$$

$$\hat{\theta} = \frac{1}{\bar{x}} = \frac{1}{5}$$

c $p(\theta) = \text{Expon}(\theta|\lambda)$

gamma distribution $\text{Ga}(\theta|a,b) \propto \theta^{a-1} e^{-\theta b}$ → $E[\theta] = \frac{a}{b}$ (mean)

$a = 2 \Rightarrow \frac{1}{3} = \frac{2}{b} \rightarrow b = 6$

$\hat{\lambda} = 6$ to achieve $E[\theta] = 1/3$

d $p(\theta | D, \hat{\lambda})$ $\quad D = \{x_1 = 5, x_2 = 6, x_3 = 4\}$, $\hat{\lambda} = 6$

prior: $p(\theta) = Ga(\theta | a = 2, b = 6) = Expon(\theta | \hat{\lambda} = 6)$

likelihood function: $L(\theta | x) = \prod_{i=1}^{N} \theta e^{-\theta x_i}$

$$p(\theta | D, \hat{\lambda}) \propto L(\theta | x)\, p(\theta)$$
$$\propto [\theta e^{-5\theta} \cdot \theta e^{-6\theta} \cdot \theta e^{-4\theta}] \cdot \theta e^{-6\theta}$$
$$\propto \theta^4 e^{-21\theta}$$

e Yes the exponential prior is conjugate to the exponential likelihood since the resulting was also a gamma distribution. They are of the form $\theta^{a-1} e^{-\theta b}$

F $E[\theta | D, \hat{\lambda}]$ can be found by taking a/b in $\theta^{a-1} e^{-\theta b}$ since it's a gamma distribution.
From above,
$\rightarrow p(\theta | D, \hat{\lambda}) \propto \theta^4 e^{-21\theta}$
$\Rightarrow a/b = 3/21 = 1/7$

g The main reason the MLE (1/5) and posterior mean (1/7) differ is that the MLE does not take into account the prior distribution and instead is solely based on the observed data. The MLE is more reasonable in our situation since the prior data was only estimated by an expert and we should instead just use the observed data here.

2.2 $\hat{y} = w_0 + w_1 x$

$\bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$C_{xx}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, $C_{xy}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$C_{yy}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$

a w_1 is the slope and to estimate the slope,

$$w_1 = \frac{N \Sigma(xy) - \Sigma x \Sigma y}{N \Sigma(x^2) - (\Sigma x)^2}$$ and using the given statistics

$$w_1 = \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}}$$ and so the minimum set of statistics for w_1 is $\{ C_{xx}^{(n)}, C_{xy}^{(n)} \}$

b w_0 is the intercept and to estimate the intercept,

$$w_0 = \frac{\Sigma y - m \Sigma x}{N} = \frac{\Sigma y}{N} - m \frac{\Sigma x}{N}$$ and using given statistics

$$w_0 = \bar{y}^{(n)} - w_1 \bar{x}^{(n)}$$

$$= \bar{y}^{(n)} - \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}} \bar{x}^{(n)}$$ and so the minimum set of statistics for w_0 (assuming we don't have w_1) is $\{ \bar{x}^{(n)}, \bar{y}^{(n)}, C_{xx}^{(n)}, C_{xy}^{(n)} \}$

c $\bar{x}^{(n+1)} \triangleq \frac{1}{n+1} \sum_{i=1}^{n+1} x_i = \frac{1}{n+1} \left(\sum_{i=1}^{n} x_i + x_{n+1} \right)$

we know $\bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} x_i \Rightarrow \sum_{i=1}^{n} x_i = n \bar{x}^{(n)}$

$\Rightarrow \frac{1}{n+1} \left(n \bar{x}^{(n)} + x_{n+1} \right) = \frac{n \bar{x}^{(n)}}{n+1} + \frac{x_{n+1}}{n+1}$

we can rewrite $\frac{n \bar{x}^{(n)}}{n+1} = \frac{(n+1)\bar{x}^{(n)}}{n+1} - \frac{\bar{x}^{(n)}}{n+1} = \bar{x}^{(n)} - \frac{\bar{x}^{(n)}}{n+1}$

$$\bar{x}^{(n)} - \frac{\bar{x}^{(n)}}{n+1} + \frac{x_{n+1}}{n+1} = \bar{x}^{(n)} + \frac{1}{n+1} \left(x_{n+1} - \bar{x}^{(n)} \right)$$

(c) $\bar{y}^{(n+1)} \triangleq \frac{1}{n+1}\sum_{i=1}^{n+1} y_i = \frac{1}{n+1}\left(\sum_{i=1}^{n} y_i + y_{n+1}\right)$

from $\bar{y}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} y_i \Rightarrow \sum_{i=1}^{n} y_i = n\bar{y}^{(n)}$

$\Rightarrow \frac{1}{n+1}\left(n\bar{y}^{(n)} + y_{n+1}\right) = \frac{n\bar{y}^{(n)}}{n+1} + \frac{y_{n+1}}{n+1}$

rewrite $\frac{n\bar{y}^{(n)}}{n+1} = \frac{(n+1)\bar{y}^{(n)}}{n+1} - \frac{\bar{y}^{(n)}}{n+1} = \bar{y}^{(n)} - \frac{\bar{y}^{(n)}}{n+1}$

$\Rightarrow \bar{y}^{(n)} - \frac{\bar{y}^{(n)}}{n+1} + \frac{y_{n+1}}{n+1} = \bar{y}^{(n)} + \frac{1}{n+1}\left(y_{n+1} - \bar{y}^{(n)}\right)$

d Let's start by rewriting $C_{xy}^{(n)}$

$C_{xy}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}^{(n)})(y_i - \bar{y}^{(n)})$

$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}^{(n)} - \sum_{i=1}^{n} \bar{x}^{(n)} y_i + \sum_{i=1}^{n} \bar{x}^{(n)}\bar{y}^{(n)}\right]$

we know $\sum_{i=1}^{n} x_i = n\bar{x}^{(n)}$ and $\sum_{i=1}^{n} y_i = n\bar{y}^{(n)}$

and we know $\sum_{i=1}^{n} \bar{x}^{(n)}\bar{y}^{(n)} = n\bar{x}^{(n)}\bar{y}^{(n)}$

$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}^{(n)}\bar{y}^{(n)} - n\bar{x}^{(n)}\bar{y}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)}\right]$

$C_{xy}^{(n)} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}^{(n)}\bar{y}^{(n)}\right]$

$\Rightarrow \sum_{i=1}^{n} x_i y_i = nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)}$

Now we can look at $C_{xy}^{(n+1)}$

$C_{xy}^{(n+1)} = \frac{1}{n+1}\sum_{i=1}^{n+1}(x_i - \bar{x}^{(n+1)})(y_i - \bar{y}^{(n+1)})$

$= \frac{1}{n+1}\left[\sum_{i=1}^{n+1} x_i y_i - \sum_{i=1}^{n+1} x_i \bar{y}^{(n+1)} - \sum_{i=1}^{n+1} \bar{x}^{(n+1)} y_i + \sum_{i=1}^{n+1} \bar{x}^{(n+1)}\bar{y}^{(n+1)}\right]$

Following a similar simplification above ...

$= \frac{1}{n+1}\left[\sum_{i=1}^{n+1} x_i y_i - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)} + (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\right]$

$= \frac{1}{n+1}\left[\sum_{i=1}^{n+1} x_i y_i - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\right]$

↖ let's separate this term

(d) $= \frac{1}{n+1}\left[x_{n+1}y_{n+1} + \sum_{i=1}^{n} x_i y_i - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\right]$

↖ let's substitute from above

$C_{xy}^{(n+1)} = \frac{1}{n+1}\left[x_{n+1}y_{n+1} + nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\right]$

We'll now derive a similar expression for $C_{xx}^{(n+1)}$

to start, we rewrite $C_{xx}^{(n)}$

$$C_{xx}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}^{(n)})^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x}^{(n)} + \sum_{i=1}^{n}\bar{x}^{(n)2}\right]$$

we rewrite $\sum_{i=1}^{n} x_i$ as $n\bar{x}^{(n)}$

and this as $n\bar{x}^{(n)2}$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^{(n)2} + n\bar{x}^{(n)2}\right]$$

$$C_{xx}^{(n)} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^{(n)2}\right]$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 = nC_{xx}^{(n)} + n\bar{x}^{(n)2}$$

so now for $C_{xx}^{(n+1)}$

$$C_{xx}^{(n+1)} = \frac{1}{n+1}\sum_{i=1}^{n+1}(x_i - \bar{x}^{(n+1)})^2$$

$$= \frac{1}{n+1}\left[\sum_{i=1}^{n+1} x_i^2 - 2\sum_{i=1}^{n+1} x_i\bar{x}^{(n+1)} + \sum_{i=1}^{n+1}\bar{x}^{(n+1)2}\right]$$

similarly substituting...

$$= \frac{1}{n+1}\left[\sum_{i=1}^{n+1} x_i^2 - 2(n+1)\bar{x}^{(n+1)2} + (n+1)\bar{x}^{(n+1)2}\right]$$

↖ separate ↖ combine

$$= \frac{1}{n+1}\left[x_{n+1}^2 + \sum_{i=1}^{n} x_i^2 - (n+1)\bar{x}^{(n+1)2}\right]$$

↑ and substituting from above

$$C_{xx}^{(n+1)} = \frac{1}{n+1}\left[x_{n+1}^2 + nC_{xx}^{(n)} + n\bar{x}^{(n)2} - (n+1)\bar{x}^{(n+1)2}\right]$$

2.3 add l2 regularization by optimizing

$$\sum_{i=1}^{N} \log_2(y_i | x_i, W) - \lambda \sum_{c=1}^{C} \|w_c\|_2^2$$

Intuitively, l2 regularization encourages smaller weights and so by optimizing, we minimize loss while keeping the weights small. If we were to have found the optimum, we'd expect the sum of the weights $\sum_{c=1}^{C} \hat{w}_{cj} = 0$. Let's consider the weights matrix. For a column c, the derivative of the weights takes into account if a particular $y_i = c$. When we have the optimum, this gradient vector is 0 and so if this is the case, the regularization term $\lambda \sum w_c$ is 0 and therefore $\sum_{c=1}^{C} \hat{w}_{cj} = 0$ for $j = 1:D$

2.4 $K(x,z) = e^{\frac{-\|x-z\|^2}{\tau^2}}$

a decision function learned by SVM

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b$$

training data $\{(x^{(1)}, y^{(1)}) \dots (x^{(m)}, y^{(m)})\}$
separated by at least ϵ: $\|x^{(i)} - x^{(j)}\| \geq \epsilon$ for any $i \neq j$

let's say $\alpha_i = 1$ for all i and $b = 0 \Rightarrow f(x) = \sum_{i=1}^{m} y^{(i)} K(x^{(i)}, x)$
for $y \in \{-1, 1\}$, the prediction $x^{(i)}$ is correct if

$$|f(x^{(i)}) - y^{(i)}| < 1$$

↓ this $x^{(i)}$ comes from plugging in for $f(x^{(i)})$

$$\left| \sum_{i=1}^{m} y^{(i)} K(x^{(i)}, x^{(i)}) - y^{(i)} \right| < 1$$

$$\left| \sum_{i=1}^{m} y^{(i)} \exp(-\|x^{(i)} - x^{(i)}\|^2 / \tau^2) - y^{(i)} \right| < 1$$

↳ this expression can be simplified if we don't have duplicate points so let's simplify the summation to have no duplicate points

pull out from summation / use j to avoid confusion

$$\left| y^{(i)} + \sum_{j \neq i} y^{(j)} \exp(-\|x^{(j)} - x^{(i)}\|^2 / \tau^2) - y^{(i)} \right| < 1$$

$$\left| \sum_{j \neq i} y^{(j)} \exp(-\|x^{(j)} - x^{(i)}\|^2 / \tau^2) \right| < 1$$

↳ we can now simplify since $j \neq i$

$$\left| \sum_{j \neq i} \exp(-\epsilon^2 / \tau^2) \right| < 1$$

↳ this summation happens $m-1$ times since we exclude $j = i$

$$(m-1)\exp(-\epsilon^2/\tau^2) < 1$$

$$-\epsilon^2/\tau^2 < \ln\left(\frac{1}{m-1}\right)$$

$$-\epsilon^2/\tau^2 < -\ln(m-1)$$

$$\tau^2 < \frac{\epsilon^2}{\ln(m-1)}$$

we can pick $\tau = \frac{\epsilon}{\ln(m)}$ with $\alpha_i = 1 \ \forall i$ and $b = 0$

b Slack variables are useful when data isn't linearly separable or if outliers exist. They aim to maximize the margin and minimize misclassifications. They allow some misclassifications and margin violations. If the bandwidth parameter γ can achieve zero training error, introducing slack variables will also obtain zero training error.