bkj2111
Bahul Jain

Homework 3

Problem 1.

Let F we a random variable given by –

$$F = \begin{cases} 1 & if\ dice\ is\ fair \\ 0 & if\ dice\ is\ unfair \end{cases}$$

Given: $P(F = 0) = \frac{1}{1000}\ and\ P(F = 1) = \frac{999}{1000}$

Let W be another random variable given by –

$$W = \begin{cases} 1 & if\ player\ wins\ round \\ 0 & if\ player\ loses\ round \end{cases}$$

Now the possible ways of winning are if we get (3,4),(4,3),(2,5),(5,2),(1,6),(6,1) combinations on the two dices of the 36 different combinations possible. If the dices are fair the probability of getting any number between 1-6 is same $= 1/6$. Thus probability of winning given dice is fair is given by…

$$P(W = 1|F = 1) = \frac{6}{36} = \frac{1}{6}$$

For an unfair dice we are given that …

$$P(Dice\ gives\ 3 \mid F = 0)\ = \frac{1}{3}$$

$$P(Dice\ gives\ 4 \mid F = 0) = \frac{1}{3}$$

And rest outcomes are equally likely, which means…

$$P(Dice\ gives\ 1/2/5/6 \mid F = 0) = \frac{\left(1 - \frac{1}{3} - \frac{1}{3}\right)}{4} = \frac{1}{12}$$

Therefore now probability of winning given dice is unfair can be written as…

$$P(W = 1|F = 0) = 2\left(\frac{1}{3}\right)^2 + 4\left(\frac{1}{12}\right)^2 = \frac{1}{4}$$

1. Given n rounds have been played, the probability of winning k rounds considering the dice is unfair can be written as…

   $$P(W = 1\ k\ times \mid F = 0) = \binom{n}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k}$$

   Similarly, given n rounds have been played, the probability of winning k rounds considering the dice is fair can be written as…

   $$P(W = 1\ k\ times \mid F = 1) = \binom{n}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{n-k}$$

   Now we have to find the probability of the dice being unfair given that of n rounds the player won k rounds. This is given by…

   $$P(F = 0 \mid W = 1\ k\ times) = \frac{P(W = 1\ k\ times \mid F = 0).P(F = 0)}{P(W = 1\ k\ times)}$$

To find $P(W = 1\ k\ times)$ in n rounds, we use marginalization property which gives us …

$$P(W\ k\ times) = P(W = 1\ k\ times|F = 1).P(F = 1) + P(W = 1\ k\ times|F = 0).P(F = 0)$$

Since we are given that the dice is most likely unfair the value on the right should be always greater than 0.5. Substituting all the values computed given in the above equation we get…

$$0.5 < \frac{\binom{n}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k}\frac{1}{1000}}{\binom{n}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k}\left(\frac{1}{1000}\right) + \binom{n}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{n-k}\left(\frac{999}{1000}\right)}$$

$$2 > 1 + \frac{\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{n-k}\left(\frac{999}{1000}\right)}{\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k}\frac{1}{1000}}$$

$$1 > \left(\frac{2}{3}\right)^k\left(\frac{10}{9}\right)^{n-k}999$$

$$999 < \left(\frac{3}{2}\right)^k\left(\frac{9}{10}\right)^{n-k}$$

Taking log on both sides we get…

$$\log 999 < k\log 1.5 + (n-k)\log 0.9$$

$$\log 999 - n\log 0.9 < k\log\frac{5}{3}$$

$$k > \frac{\log 999 - n\log 0.9}{\log\frac{5}{3}}$$

Since k is the number of rounds won, k has to be an integer. Also since we have to find the minimum number of rounds won, we take the immediate next value of k. Hence we get the value of k as…

$$\boldsymbol{k = \left\lceil \frac{\log 999 - n\log 0.9}{\log\frac{5}{3}} \right\rceil}$$

2. To obtain the minimum number of rounds played to determine if a dice is unfair, we would have to assume that in every round the player won given the dice is unfair and since the outcome of each rounds are IID's, we can write the probability as…

$$P(W = 1\ n\ times \mid F = 0) = \left(\frac{1}{4}\right)^n$$

Now we are asked to find out the probability of the dice being unfair given that the player has won n rounds consecutively, which can be written as…

$$P(F = 0|W = 1\ n\ times)$$

Using Bayes' Theorem this can be written as…

$$P(F = 0|W = 1\ n\ times) = \frac{P(W = 1\ n\ times \mid F = 0).P(F = 0)}{P(W = 1\ n\ times)}$$

To find $P(W = 1\ n\ times)$ in n rounds, we use marginalization property which gives us …

$$P(W\ n\ times) = P(W = 1\ times \mid F = 1).P(F = 1) + P^n(W = 1 \mid F = 0).P(F = 0)$$

Substituting this value of $P(W)$ in the original equation we now get…

$$P(F = 0 \mid W = 1\ n\ times) = \frac{P(W = 1\ n\ times \mid F = 0).P(F = 0)}{P(W = 1\ n\ times \mid F = 1)P(F = 1) + P(W = 1\ n\ times \mid F = 0)P(F = 0)}$$

Since we are given that this value will reach 0.5 after n rounds we now equate this equation to 0.5 and substitute values for other terms, thus we get…

$$0.5 = \frac{\left(\frac{1}{4}\right)^n \cdot \frac{1}{1000}}{\left(\frac{1}{4}\right)^n \left(\frac{1}{1000}\right) + \left(\frac{1}{6}\right)^n \left(\frac{999}{1000}\right)}$$

Now we simply solve for n…

$$2 = 1 + \frac{\left(\frac{1}{6}\right)^n 999}{\left(\frac{1}{4}\right)^n}$$

$$\frac{1}{999} = \left(\frac{2}{3}\right)^n$$

Taking log on both sides we get…

$$\log 999 = n.\log(1.5)$$

$$n = \frac{\log 999}{\log 1.5}$$

$$n = 17.0345$$

Since we have to find the minimum number of rounds played which will always be an integer, we take the immediate next integer which gives us **n=18**. Hence after winning 18 consecutive round the probability of the fact that dice is unfair has crossed 0.5, which means the dice is most likely unfair.

Problem 2.

A) Uniform probability distribution means that $P_A(\mu) = 1\ \forall\ \mu\ \epsilon\ [0,1]$

B) This given probability distribution can be written as $P_B(\mu) = -6\mu^2 + 6\mu\ \forall\ \mu\ \epsilon\ [0,1]$

D_1) $\{H, T\}$

D_2) $\{T, T, T\}$

$P(H) = \mu$

$P(T) = 1 - \mu$

1) $P(H) = \mu$, thus we simply find the expected value of μ given the 2 probability distributions.

A) $E[\mu] = \int_0^1 P_A(\mu).\mu.d\mu = \int_0^1 \mu.1.d\mu = \left[\frac{\mu^2}{2}\right]_0^1 = \frac{1}{2}$ Hence $\boldsymbol{P(H)} = \mathbf{\frac{1}{2}}$

B) $E[\mu] = \int_0^1 P_B(\mu).\mu.d\mu = \int_0^1 \mu(-6\mu^2 + 6\mu)d\mu = \int_0^1 (-6\mu^3 + 6\mu^2)d\mu = \left[-\frac{6\mu^4}{4} + \frac{6\mu^3}{3}\right]_0^1 = 2 - \frac{3}{2} = \frac{1}{2}$

Hence $\boldsymbol{P(H)} = \mathbf{\frac{1}{2}}$

2) We know that…

$$P(\mu \mid D) = P(D \mid \mu).\frac{P(\mu)}{P(D)}$$

For D_1 and (A)…

$$P(\mu \mid D_1) = P(D_1 \mid \mu).\frac{P_A(\mu)}{P(D_1)}$$

$$P_A(\mu \mid D_1) = (\mu)(1-\mu).\frac{1}{\int_0^1 P(D_1 \mid \mu).P_A(\mu)d\mu}$$

$$P_A(\mu \mid D_1) = \frac{\mu(1-\mu)}{\int_0^1 \mu - \mu^2 d\mu}$$

$$P_A(\mu \mid D_1) = \frac{\mu(1-\mu)}{\left[\frac{\mu^2}{2} - \frac{\mu^3}{3}\right]_0^1} = \frac{\mu(1-\mu)}{\frac{1}{6}} = 6\mu(1-\mu)$$

$$\boldsymbol{P_A(\mu \mid D_1) = 6\mu(1-\mu)}$$

For D_2 and (A)…

$$P_A(\mu \mid D_2) = P(D_2 \mid \mu).\frac{P_A(\mu)}{P(D_2)}$$

$$P_A(\mu \mid D_2) = (1-\mu)^3.\frac{1}{\int_0^1 P(D_2 \mid \mu).P_A(\mu)d\mu}$$

$$P_A(\mu \mid D_2) = \frac{(1-\mu)^3}{\int_0^1 (1-\mu)^3 d\mu}$$

$$P_A(\mu \mid D_2) = \frac{(1-\mu)^3}{\left[\frac{(1-\mu)^4}{-4}\right]_0^1} = \frac{(1-\mu)^3}{0+\frac{1}{4}} = 4(1-\mu)^3$$

$$\boldsymbol{P_A(\mu \mid D_2) = 4(1-\mu)^3}$$

For D_1 and (B)...

$$P_B(\mu \mid D_1) = P(D_1 \mid \mu).\frac{P_B(\mu)}{P(D_1)}$$

$$P_B(\mu \mid D_1) = (\mu)(1-\mu).\frac{(-6\mu^2+6\mu)}{\int_0^1 P(D_1 \mid \mu).P_B(\mu)d\mu}$$

$$P_B(\mu \mid D_1) = \frac{6\mu^2(1-\mu)^2}{\int_0^1 6\mu^2(1-\mu)^2 d\mu}$$

$$P_B(\mu \mid D_1) = \frac{6\mu^2(1-\mu)^2}{\frac{1}{5}} = 30\mu^2(1-\mu)^2$$

$$\boldsymbol{P_B(\mu \mid D_1) = 30\mu^2(1-\mu)^2}$$

For D_2 and (B)…

$$P_B(\mu \mid D_2) = P(D_2 \mid \mu).\frac{P_B(\mu)}{P(D_2)}$$

$$P_B(\mu \mid D_2) = (1-\mu)^3.\frac{(-6\mu^2+6\mu)}{\int_0^1 P(D_2 \mid \mu).P_B(\mu)d\mu}$$

$$P_B(\mu \mid D_2) = \frac{6\mu(1-\mu)^4}{\int_0^1 6\mu(1-\mu)^4 d\mu}$$

$$P_B(\mu \mid D_2) = \frac{6\mu(1-\mu)^4}{0+\frac{1}{5}} = 30\mu(1-\mu)^4$$

$$\boldsymbol{P_B(\mu \mid D_2) = 30\mu(1-\mu)^4}$$

3) For the given two distributions μ_{ML} can be found out by differentiating the below equation and setting the value to 0…

$$L(\mu) = \prod P(D \mid \mu)$$

For D_1…

$$L(\mu) = \mu(1-\mu)$$
$$\frac{d}{d\mu}L(\mu) = 1 - 2\mu = 0$$
$$\boldsymbol{\mu_{ML} = \frac{1}{2}}$$

For D_2…

$$L(\mu) = (1-\mu)^3$$
$$\frac{d}{d\mu}L(\mu) = -3(1-\mu)^2 = 0$$
$$\boldsymbol{\mu_{ML} = 1}$$

4) For the given two distributions μ_{MAP} can be found out by differentiating the below equation and setting the value to 0…

$$L(\mu) = \left(\prod P(D \mid \mu)\right) P(\mu)$$

For D_1 and (A)…

$$L(\mu) = \left(\prod P(D_1 \mid \mu)\right) P_A(\mu)$$
$$L(\mu) = \mu(1-\mu)$$
$$\frac{d}{d\mu}L(\mu) = 1 - 2\mu = 0$$
$$\boldsymbol{\mu_{MAP} = \frac{1}{2}}$$

For D_2 and (A)…

$$L(\mu) = \left(\prod P(D_2 \mid \mu)\right) P_A(\mu)$$
$$L(\mu) = (1-\mu)^3$$
$$\frac{d}{d\mu}L(\mu) = -3(1-\mu)^2 = 0$$
$$\boldsymbol{\mu_{MAP} = 1}$$

For D_1 and (B)…

$$L(\mu) = \left(\prod P(D_1 \mid \mu)\right) P_B(\mu)$$
$$L(\mu) = 6\mu^2(1-\mu)^2$$
$$\frac{d}{d\mu}L(\mu) = 12\mu(1-\mu)^2 - 12\mu^2(1-\mu) = 0$$
$$\mu = 1 - \mu$$

At this point μ can have the values 1, 0 and 1/2. These values could give us maximum likelihood as well as minimum likelihood. So now we will find the second derivative and only consider those values that give negative result for second derivative.

$$\frac{d^2}{d\mu^2}L(\mu) = 12 - 24\mu + 12\mu^2 - 24\mu + 24\mu^2 - 24\mu + 36\mu^2 = 12 - 72\mu + 72\mu^2$$

From the above equation we see that second derivative is only negative for $\mu = \frac{1}{2}$, hence we get…

$$\boldsymbol{\mu_{MAP} = \frac{1}{2}}$$

For D_2 and (B)…

$$L(\mu) = \left(\prod\prod P(D_2 \mid \mu)\right) P_B(\mu)$$

$$L(\mu) = 6\mu(1-\mu)^4$$

$$\frac{d}{d\mu}L(\mu) = 6(1-\mu)^4 - 24\mu(1-\mu)^3 = 0$$

$$(1-\mu) = 4\mu$$

$$5\mu = 1 \rightarrow \mu = \frac{1}{5}$$

Again we find the second derivative to determine if we are getting the maxima or minima….

$$\frac{d^2}{d\mu^2}L(\mu) = -24(1-\mu)^3 - 24(1-\mu)^3 + 72\mu(1-\mu)^2$$

For the above equation the only value of μ that is giving a negative values is $\mu = \frac{1}{5}$, hence we get…

$$\boldsymbol{\mu_{MAP} = \frac{1}{5}}$$

5) To find $P(H \mid D)$ we use the Bayesian inference equation given by…

$$P(H \mid D) = \int_0^1 P(H, \mu \mid D)d\mu$$

$$P(H \mid D) = \int_0^1 P(H \mid \mu)P(\mu \mid D)d\mu$$

Now for D_1 and (A)

$$P_A(H \mid D_1) = \int_0^1 \mu . P_A(\mu \mid D_1)d\mu$$

From part 2 we use the value of $P_A(\mu \mid D_1)$

$$P_A(H \mid D_1) = \int_0^1 (6\mu^2 - 6\mu^3)d\mu$$

$$P_A(H \mid D_1) = \left[\frac{6\mu^3}{3} - \frac{6\mu^4}{4}\right]_0^1 = \frac{1}{2}$$

$$\boldsymbol{P_A(H \mid D_1) = \frac{1}{2}}$$

Now for D_2 and (A)

$$P_A(H \mid D_2) = \int_0^1 \mu . P_A(\mu \mid D_2)d\mu$$

From part 2 we use the value of $P_A(\mu \mid D_2)$

$$P_A(H \mid D_2) = \int_0^1 4\mu(1-\mu)^3 d\mu$$

$$P_A(H \mid D_2) = 4\int_0^1 (\mu - \mu^4 - 3\mu^2 + 3\mu^3)d\mu = 4\left[\frac{\mu^2}{2} - \frac{\mu^5}{5} - \frac{3\mu^3}{3} + \frac{3\mu^4}{4}\right]_0^1 = \frac{1}{5}$$

$$\boldsymbol{P_A(H \mid D_2) = \frac{1}{5}}$$

Now for D_1 and (B)

$$P_B(H \mid D_1) = \int_0^1 \mu . P_B(\mu \mid D_1) d\mu$$

From part 2 we use the value of $P_B(\mu \mid D_1)$

$$P_B(H \mid D_1) = \int_0^1 36\mu^3(1-\mu)^2 d\mu$$

$$P_B(H \mid D_1) = 36 \int_0^1 (\mu^3 - 2.\mu^4 + \mu^5)\, d\mu = 30 \left[\frac{\mu^4}{4} - \frac{2\mu^5}{5} + \frac{\mu^6}{6}\right]_0^1 = \frac{1}{2}$$

$$\boldsymbol{P_B(H \mid D_1) = \frac{1}{2}}$$

Now for D_2 and (B)

$$P_B(H \mid D_2) = \int_0^1 \mu . P_B(\mu \mid D_2) d\mu$$

From part 2 we use the value of $P_B(\mu \mid D_2)$

$$P_B(H \mid D_2) = \int_0^1 30\mu(1-\mu)^4 d\mu$$

$$\boldsymbol{P_B(H \mid D_2) = \frac{2}{7}}$$

6) Variance for the posterior distribution $P(\mu \mid D)$ is given by…

$$Var\{\mu \mid D\} = E\{\mu^2 \mid D\} - E^2\{\mu \mid D\}$$

$$E\{\mu \mid D\} = \int_0^1 \mu . P(\mu \mid D) d\mu$$

This we computed in part 5…

$$E\{\mu^2 \mid D\} = \int_0^1 \mu^2 . P(\mu \mid D) d\mu$$

Now for D_1 and (A)…

$$Var\{\mu \mid D_1\} = \int_0^1 \mu^2 . P_A(\mu \mid D_1) d\mu - \left(\int_0^1 \mu . P_A(\mu \mid D_1) d\mu\right)^2$$

$$Var\{\mu \mid D_1\} = \int_0^1 6\mu^3(1-\mu) d\mu - \frac{1}{4} = 6\left[\frac{\mu^4}{4} - \frac{\mu^5}{5}\right]_0^1 - \frac{1}{4} = \boldsymbol{\frac{1}{20}}$$

Now for D_2 and (A)…

$$Var\{\mu \mid D_2\} = \int_0^1 \mu^2 . P_A(\mu \mid D_2) d\mu - \left(\int_0^1 \mu . P_A(\mu \mid D_2) d\mu\right)^2$$

$$Var\{\mu \mid D_2\} = \int_0^1 4\mu^2(1-\mu)^3 d\mu - \frac{1}{25} = \frac{1}{15} - \frac{1}{25} = \boldsymbol{\frac{2}{75}}$$

Now for D_1 and (B)…

$$Var\{\mu \mid D_1\} = \int_0^1 \mu^2 . P_B(\mu \mid D_1) d\mu - \left(\int_0^1 \mu . P_B(\mu \mid D_1) d\mu\right)^2$$

$$Var\{\mu \mid D_1\} = \int_0^1 30\mu^4(1-\mu)^2 d\mu - \frac{1}{4} = 30\left[\frac{\mu^5}{5} - \frac{2.\mu^6}{6} + \frac{\mu^7}{7}\right]_0^1 - \frac{1}{25} = \boldsymbol{\frac{1}{28}}$$

Now for D_2 and (B)…

$$Var\{\mu \mid D_2\} = \int_0^1 \mu^2 . P_B(\mu \mid D_2) d\mu - \left(\int_0^1 \mu . P_B(\mu \mid D_2) d\mu\right)^2$$

$$Var\{\mu \mid D_2\} = \int_0^1 30\mu^3(1-\mu)^4 d\mu - \frac{4}{49} = \frac{3}{28} - \frac{4}{49} = \boldsymbol{\frac{5}{196}}$$

Final Result of all parts

	(A) and D1	(A) and D2	(B) and D1	(B) and D2
$\boldsymbol{P(H)}$	½	½	½	½
$\boldsymbol{P(\mu \mid D)}$	$6\mu(1-\mu)$	$4(1-\mu)^3$	$30\mu^2(1-\mu)^2$	$30\mu(1-\mu)^4$
$\boldsymbol{\mu_{ML}}$	½	1	½	1
$\boldsymbol{\mu_{MAP}}$	½	1	½	1/5
$\boldsymbol{P(H \mid D)}$	½	1/5	½	2/7
$\boldsymbol{Var\{\mu \mid D\}}$	1/20	2/75	1/28	5/196

The maximum likelihood estimate is not reliable in this context because for D_2, we get the value of μ to be 1, which means that the model gives the probability of heads $P(H) = 1$. This is not a valid estimator because $D_2 = \{T, T, T\}$, would actually never be obtained using μ_{ML}.

Problem 3.

1. $a \perp\!\!\!\perp b|c \rightarrow a \perp\!\!\!\perp b$ – This essentially means that given an event c, events a *and* b are independent of each other.
   Let's take event c: 'father has diabetes'
   Event a: 'sibling 1 has diabetes'
   Event b: 'sibling 2 has diabetes'
   Now given that only event a has occurred, there arises a probability that sibling 1 could have inherited the disease from their parents which increases the probability of sibling 2 also having diabetes. This means that the two events are dependent. Now given that event c has occurred, i.e. father has diabetes, the fact that sibling 1 has diabetes will provide us no information that will change the probability of sibling 2 having diabetes (since event c has already incorporated the possibility of inheriting the disease from parent which we were considering as the 'dependent factor' between the two events in the previous case). Hence they are independent now. To sum this up we can say that event a and event b are independent given event c has occurred, but when nothing is given, these events depend on each. The above conclusion disproves the claim made above and hence it is **false**.

   Let's do this using values now …

   When a and b are conditionally independent given c $\rightarrow P(a, b \mid c) = P(a \mid c)P(b \mid c)$. When a and b are independent $\rightarrow P(a, b) = P(a)P(b)$. For the above claim to hold true $P(a, b \mid c) = P(a, b)$ should also hold true. Let us take counter examples to disprove the claim…

$$P(a = 1) = 0.3, P(a = 0) = 0.7$$
$$P(b = 1) = 0.4, P(b = 0) = 0.6$$
$$P(a = 1 \mid c) = 0.5, P(a = 0 \mid c) = 0.5$$
$$P(b = 1 \mid c) = 0.8, P(b = 0 \mid c) = 0.2$$

   Substituting these values in the above equations we get…

$$P(a, b \mid c) = P(a \mid c)P(b \mid c) = 0.5 \times 0.8 = 0.4$$
$$P(a, b) = P(a)P(b) = 0.3 \times 0.4 = 0.12$$

   Since these two values are not equal we have disproved the claim.

2. $a \perp\!\!\!\perp b \rightarrow a \perp\!\!\!\perp b|c$ - According to this claim if a *and* b are independent of each then they will always be independent of each other given that event c has happened.
   Let event a: Getting heads in a coin flip – $P(a) = 0.5$
   Let event b: Getting heads in a coin flip – $P(b) = 0.5$
   We can clearly say that a & b are independent of each other, i.e.
   Probability that event a and b both happen $P(a, b) = P(a)P(b) = 0.25$
   Now let's take another event c such that…
   Event c: Getting same outcome in both flips
   So now if I know that in event a I get a heads with a probability $P(a) = 0.5$ and given that outcomes of both flips will be the same then I know for sure that in event b I will definitely get a head with probability P(b)=1.
   Now…
   Probability that event a and b both happen $P(a, b|c) = P(a|c)P(b|c) = 0.5$
   This disproves the claim given. Hence it's false.

3. $(a \perp\!\!\!\perp b) \wedge (b \perp\!\!\!\perp c) \rightarrow (a \perp\!\!\!\perp c)$ - Now it's said that if events a and b are independent and events b and c are independent then events a and c are also independent.
   Let event a: Picking an apple from a fruit basket (Basket 1) having 1 apple and 3 oranges - $P(a) = 0.25$
   Let event b: Picking a banana from a fruit basket (Basket 2) having 1 apple and 1 banana - $P(b) = 0.5$
   And let event c: Picking an orange from a fruit basket (Basket 1) - $P(b) = 0.75$
   Now we can clearly see that if event b occurs it will definitely not change $P(a)$ or $P(c)$, i.e. events a and c are independent of event b. But given that event a occurs, i.e. I pick an apple from fruit basket 1, the probability of event c happening will change since there are only oranges left in fruit basket and so P(c) will become 1.
   This disproves the claim given. Hence it's false.

Problem 4.

Given…

$$P(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}$$

The likelihood can be found out using the equation given below…

$$L(\lambda) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_i^{\alpha-1} e^{-\lambda x_i}$$

$$\log(L(\lambda)) = \prod_{i=1}^{n} \log\left(\frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_i^{\alpha-1} e^{-\lambda x_i}\right)$$

To find the maximum likelihood we differentiate the above equation and set it to zero…

$$\frac{\partial}{\partial \lambda}\left(\log\left(L(\lambda)\right)\right) = \sum_{i=1}^{n} \left(\log\left(\frac{1}{\Gamma(\alpha)}\right) + \alpha \log(\lambda) + (\alpha - 1) \log x_i - \lambda x_i\right)$$

$$= \sum_{i=1}^{n} \left(0 + \frac{\alpha}{\lambda} + 0 - x_i\right)$$

$$\frac{\partial}{\partial \lambda}\left(\log\left(L(\lambda)\right)\right) = \sum_{i=1}^{n} \left(\frac{\alpha}{\lambda} - x_i\right)$$

$$\sum_{i=1}^{n} \frac{\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} \frac{\alpha}{\lambda} = \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i = \frac{n\alpha}{\lambda}$$

$$\lambda = \frac{n\alpha}{\sum_{i=1}^{n} x_i}$$

To prove this lambda corresponds to the maximum likelihood, we take the double derivative and show that it is negative.

$$\frac{\partial^2}{\partial \lambda^2}\left(\log\left(L(\lambda)\right)\right) = \frac{\partial}{\partial \lambda}\left(\frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i\right) = -\frac{\alpha n}{\lambda^2}$$

Since all n, alpha and lambda are positive values the above term will always be negative. Hence we have proved that the above value of lambda corresponds to the maxima.

Problem 5.

Given the cost function…

$$J(w) = -\frac{1}{N}\sum_{i=1}^{n} log\left(\sigma(y_i w^T k_i)\right) + \lambda w^T w$$

We will first find the derivative of the cost function to find the optimum model…

$$\frac{d}{dw}J(w) = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sigma(y_i w^T k_i)}.\sigma(y_i w^T k_i)\left(1 - \sigma(y_i w^T k_i)\right).y_i k_i + 2\lambda w$$

$$\nabla = \frac{d}{dw}J(w) = -\frac{1}{N}\sum_{i=1}^{N} (y_i k_i - \sigma(y_i w^T k_i).y_i k_i) + 2\lambda w$$

Setting this equal to zero and solving for w will be quite a tedious task, hence we use gradient descent approach to obtain optimum model using the above value as the change in the model. This is given by….

$$w_{t+1} = w_t - \eta.\nabla$$

For stochastic gradient descent the following equations are used as updates to the model…

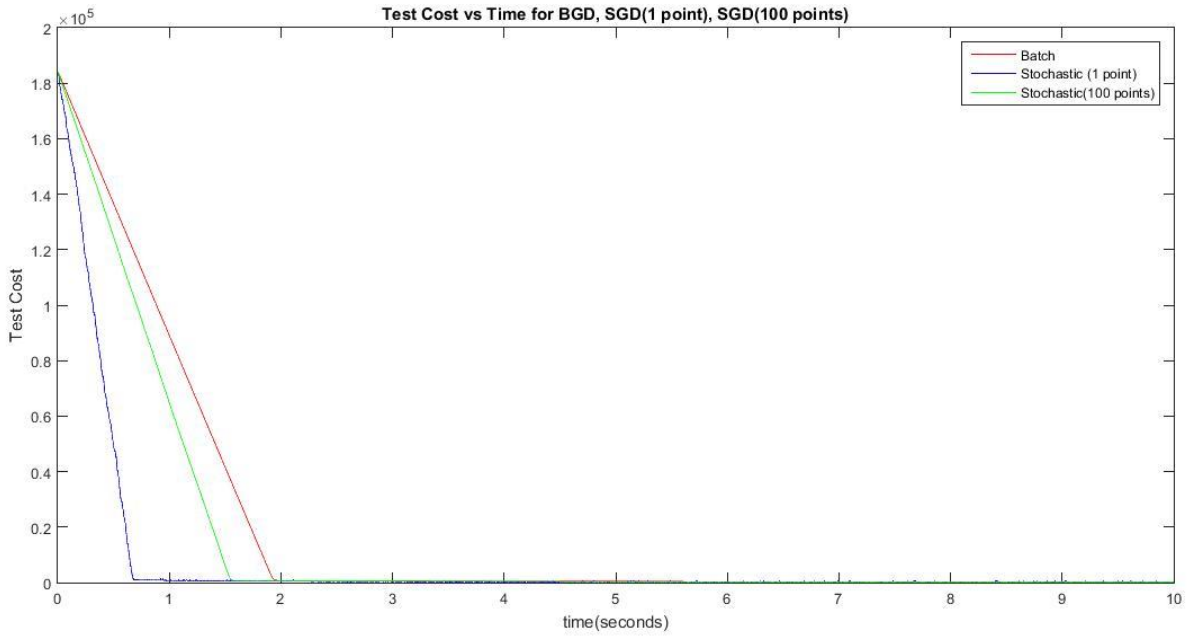
For 1 random point….

$$\nabla = -(y_i k_i - \sigma(y_i w^T k_i).y_i k_i) + 2\lambda w$$

For 100 random points…

$$\nabla = -\frac{1}{100}\sum_{i=1}^{100} (y_i k_i - \sigma(y_i w^T k_i).y_i k_i) + 2\lambda w$$

After the first 10 seconds of running each of the three learning methods we obtained the graph as follows…



From the above graph we can say that the test cost is reducing the fastest for stochastic gradient descent with only 1 random point, followed by stochastic gradient descent using 100 random points and finally batch gradient descent.

Now we let batch gradient descent converge till the gradient becomes very small (i.e. it falls below the tolerance value). The time to converge is recorded and the other two learning methods are let to run for at least this recorded time. After all the learning methods have been allowed to run for comparable time we compare the results from each method. The results have been shown below…

	BGD	SGD (1 random point)	SGD (100 random points)
η (step size)	$1e-2$	$1e-2$	$1e-2$
λ	$1e-3$	$1e-3$	$1e-3$
ϵ (tolerance)	$1e-2$	NA	NA
# Iterations	15,844	50,170	30,737
Training Accuracy (%)	91.4%	92.0%	91.4%
Test Accuracy (%)	92.0%	92.2%	92.3%
Test Cost	241.98	224.08	232.00
Time Elapsed (seconds)	61	62	62

End.