Q1

a) $E(f(X)) = \sum_{x \in X} f(x)\,p(x) = \underbrace{10 \times \frac{1}{10}}_{\text{for } x=a} + \underbrace{\frac{5}{10} \times \frac{2}{10}}_{\text{for } x=b} + \underbrace{\frac{10}{7} \times \frac{7}{10}}_{\text{for } x=c} = \boxed{3}$

b) $E\left(\frac{1}{p(X)}\right) = \sum_{x \in X} \frac{1}{p(x)} \times p(x) = \frac{1}{p(a)} \times p(a) + \frac{1}{p(b)} \times p(b) + \frac{1}{p(c)} \times p(c) = \boxed{3}$

$\frac{1}{p(a)} = f(a)$, $\frac{1}{p(b)} = f(b)$, $\frac{1}{p(c)} = f(c)$

$\Rightarrow \frac{1}{p(x)} = f(x) \Rightarrow E\left(\frac{1}{p(X)}\right) = E(f(X))$

c) pmf $p \Rightarrow p(a) + p(b) + p(c) + \cdots + p(n) = 1$

$\Rightarrow E\left(\frac{1}{p(X)}\right) = \sum_{\substack{x \in X \\ x=a}}^{n} \frac{1}{p(a)} \times p(a) + \cdots + \frac{1}{p(n)} \times p(n)$

$= |X|$ → number of elements in X or number of elements which x can get

Q2

a)

- $E(X) = E(a_1 X_1 + a_2 X_2 + \cdots + a_m X_m)$

$= E(a_1 X_1) + E(a_2 X_2) + \cdots + E(a_m X_m)$

$= a_1 E(X_1) + a_2 E(X_2) + \cdots + a_m E(X_m)$

$= \sum_{i=1}^{m} a_i E(X_i)$ (I)

- $E(X)$ for guassian distribution is μ. Why?

because: (σ^2 is Σ_i, σ^2 feels better as variance)

$$E(X) = \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} x\,dx \xrightarrow{z = x-\mu}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(z)^2} (z+\mu)\,dz$$

$\hookrightarrow$ no "z"

$$= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \int_{-\infty}^{+\infty} z e^{-\frac{z^2}{2\sigma^2}} dz + \frac{\mu}{(2\pi\sigma^2)^{\frac{1}{2}}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2\sigma^2}} dz$$

odd function $= 0$; it is pdf $= 1$

$\Rightarrow 0 + \mu = \mu \Rightarrow E(X) = \mu$ (II)

- (I) & (II) $\Rightarrow E(X) = \sum_{i=1}^{m} a_i E(X_i) = \boxed{\sum_{i=1}^{m} a_i \mu_i}$

$a_i \rightarrow$ is a constant

$\mu_i \rightarrow$ is $\in \mathbb{R}^d$ $\Rightarrow$ $\boxed{E(X) \text{ has the dimension of } d}$

$\in \mathbb{R}^d$

(2.2) b)

$$Cov(X) = \sum_{i,j=1}^{m} Cov(a_i X_i, a_j X_j)$$

$$= \sum_{i,j=1}^{m} a_i a_j Cov(X_i, X_j)$$

$$= \sum_{i=1}^{m} a_i^2 Var(X_i) + \sum_{i \neq j} a_i a_j Cov(X_i, X_j)$$

$$= \sum_{i=1}^{m} a_i^2 Var(X_i) + 2 \underbrace{\sum_{1 \leq i < j \leq m} a_i a_j Cov(X_i, X_j)}_{0}$$

if random variables $X_1, \dots, X_m$ are independent we will get $Cov(X_i, X_j) = 0$

$$\Rightarrow Cov(X) = \sum_{i=1}^{m} a_i^2 Var(X_i) = \boxed{\sum_{i=1}^{m} a_i^2 \Sigma_i}$$

→ The dimension of $Cov(X)$ will be dxd since $\Sigma_i \in \mathbb{R}^{d \times d}$ and a_i is constant

→ now if $Cov(X_1, X_2) = \Lambda$ we will get:

$$Cov(X) = \sum_{i=1}^{m} a_i^2 \Sigma_i + \underbrace{2 \sum_{2 \leq i < j \leq m} a_i a_j Cov(X_i, X_j)}_{\to 0} + \underbrace{2 \sum_{1 \leq i < j \leq 2} a_i a_j Cov(X_1, X_2)}_{\to \Lambda}$$

$$\Rightarrow \boxed{Cov(X) = \sum_{i=1}^{m} a_i^2 \Sigma_i + 2 a_1 a_2 \Lambda}$$

Q3 a) As I ran the code for this asked question and with the asked configuration, I saw that as the number of samples increased it tends to get closer to 0 which is the μ. Obviously because it should get closer to mean but more samples show the result in the long run.

Now as we changed σ from 1 to 10 it means variance of our samples can be larger, so we saw the same number of samples as $\sigma = 1$, for $\sigma = 10$ have more space to spread. Therefore, for small number of samples in $\sigma = 10$ we had more variety of means, in contrast of $\sigma = 1$. However, again when we increased number of samples, samples had less and less space to spread. If we increase σ from 10 to 100 even 1000 samples will find more variety of means.

# of samples → 10

σ ↓ 1 → -0.211, 0.09, -0.232
0.337

σ ↓ 10 → 1.26, -6.451, -2.198
-1.129

more variety of numbers because of increase in variance

# of samples →	100	1000
$\sigma = 1$	0.096, -0.055 0.131, 0.212	-0.013, 0.03, 0.03 0.02
$\sigma = 10$	1.017, 0.620 0.203; -0.47	0.16, 0.25, 0.77 0.024

b) It is basically the covariance matrix of each individual variable:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} Cov(X,X)=Var(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Cov(Y,Y)=Var(Y) & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Cov(Z,Z)=Var(Z) \end{bmatrix}$$

This identity matrix shows X, Y, and Z are independent from each other. Covariance basically shows how two variable, vary with each other. All of them are 0. So they don't vary together and they are indipendent.

c) By this change we will have:

$Cov(X,X)=Var(X)=1$ $Cov(X,Z)=1$ $Cov(Y,Y)=Var(Y)=1$

$Cov(Z,X)=1$ $Cov(Z,Z)=Var(Z)=1$

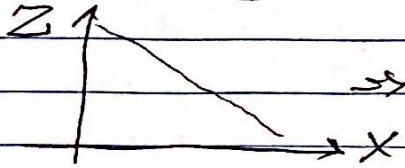
This indicates: $Corr(X,Z) = \frac{Cov(X,Z)}{\sqrt{Var(X)}\sqrt{Var(Z)}} = 1$

⇒ It means X and Z will increase together

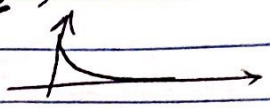
⇒ $Cov(X,Z)=+1$

$Cov(Z,X)$ will do the same as $Cov(X,Z)$, so the result won't change.

If this $Cov(X,Z)=-1$, the exact opposite of this diagram would happen:



⇒ $Cov(X,Z)=-1$

(24) a) based on our initial feelings, since it is an exponential distribution with $\theta = \frac{1}{2}$, we can assume $p(\lambda) = \frac{1}{2} e^{-\frac{1}{2}\lambda}$. It looks like 

which easily has the highest point at $\lambda = 0$.
Now, even with maximum, $\frac{d f(\lambda)}{d\lambda} = -\frac{1}{4} e^{-\frac{\lambda}{2}} = 0 \Rightarrow \lambda = 0$
$\Rightarrow f(0) = \frac{1}{2}$

b) $\lambda_{MLE} = \arg\max \{ P(D|X) \}$

$\rightarrow P(D|X) = P(x_1, x_2, x_3, \dots, x_n | \lambda)$

$= \prod_{i=1}^{n} P(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$

$\Rightarrow \ln(P(D|\lambda)) = \ln \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} = \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$

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

$\Rightarrow \lambda = \underbrace{\left(\sum_{i=1}^{n} x_i\right)}_{79} / n \rightarrow 9 \Rightarrow \boxed{\lambda = \frac{79}{9}}$

c) $\lambda_{MAP} = \arg\max \{ P(D|\lambda)\, P(\lambda) \}$

we have showed that (I) $P(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda n}}{\prod_{i=1}^{n} x_i!}$

$P(\lambda) = \theta e^{-\theta\lambda} \xrightarrow{\theta = \frac{1}{2}} \frac{1}{2} e^{-\frac{1}{2}\lambda}$ (II)

$P(D|\lambda)\, P(\lambda)$ (I)x(II) $\frac{1}{2} \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda/2 - \lambda n}}{\prod_{i=1}^{n} x_i!}$

$\ln(P(D|\lambda)\, P(\lambda)) \Rightarrow \frac{1}{2}\left(\ln\lambda \sum_{i=1}^{n} x_i - \lambda/2 - \lambda n - \sum_{i=0}^{n} \ln(x_i!)\right)$

$\Rightarrow \frac{\partial(\ln(P(D|\lambda)P(\lambda))}{\partial\lambda} = \frac{1}{2}\left(\frac{1}{\lambda}\sum_{i=1}^{n} x_i - \frac{1}{2} - n\right) = 0$

$\Rightarrow \lambda = \frac{2\sum_{i=1}^{n} x_i}{2n+1} \xrightarrow{n=9} \frac{158}{19}$

d) by using ML and MAP we get $\arg\max \lambda_{ML}(K)$ and $\arg\max \lambda_{MAP}(K)$. As these are comming from Poisson distribution, the arg max will be the mode of these distributions which is $\lfloor \lambda \rfloor$. both of MAP and ML are predicting 8 incidences (~~$\frac{158}{19} \approx 8$ and $\frac{79}{9} \approx 8$~~) for tomorrow. ML and MAP are not always the same, but this time both are predicting 8 accidents. (see next page for more precise way) (continue next page)

We put the numbers into the distribution for λ of ML and MAP. to find the most probability of accident:

$\lambda_{ML} = \frac{79}{9} \approx 8.7$ $\quad x \to 8, 9$

$x=8 \Rightarrow \left(\frac{79}{9}\right)^8 \times e^{-\frac{79}{9}} / 8! = \boxed{0.1347}$ → more probable

$x=9 \Rightarrow \left(\frac{79}{9}\right)^9 \times e^{-\frac{79}{9}} / 9! = 0.1313$

$\overset{ML}{\Rightarrow}$ number of accidents: $\boxed{8}$

$\lambda_{MAP} = \frac{158}{19}$

$x=8 \Rightarrow \left(\frac{158}{19}\right)^8 \times e^{-\left(\frac{158}{19}\right)} / 8! = \boxed{0.1387}$

$x=9 \Rightarrow \left(\frac{158}{19}\right)^9 \times e^{-\left(\frac{158}{19}\right)} / 9! = 0.1281$

$\overset{MAP}{\Rightarrow}$ number of accidents: $\boxed{8}$

e) We mainly use prior knowledge to enhance our predictions estimation.
In this example, we used this information that λ is exponential distribution.
The prior information will be less and less effective when our samples increase.
In fact ML is as MAP with uniform distribution for λ, an information which is not always true.
If we dont have enough data, these information helps us to increase our accuracy of estimation to avoid overfitting.

In MAP the prior lets us to estimate the most likely parameter given the data under posterior distribution

$P(\lambda|D) = \frac{P(D|\lambda) \times P(\lambda)}{P(D)}$

f) As our goal is to decrease the number of accidents, easily we have to increase θ in the prior. There are two ways to prove it.
First is, as we know, increasing θ means decreasing the mean of the distribution
$\Rightarrow$ estimated λ will decrease as well
Second, you can see by $\frac{(\sum k_i)}{n+\theta}$ in MAP that increasing θ means decreasing the statement(λ) and vice versa.

(25) a)

We have 2 random variables. Lets call "is sunny" as X and "is table free" as Y. For writing the ML for maximize $P(D|\lambda)$ we need to come up with a distribution. Obviously for "D" we are dealing with Bayes rule. So the problem will be:

$$\lambda_{ML} = \arg\max p(D|\lambda)$$

$$\hookrightarrow p(x_i, y_i|\lambda) = P(y_i|x_i,\lambda)P(x_i)$$

For distribution we can use Bernoulli distribution for being sunny and not sunny as win and lose respectively. The same is applicable for vacancy of table.

We also use marginal pmf for "being sunny" was related to "not empty table" and find another distribution related to that But we don't have those information to compute exactly right now.

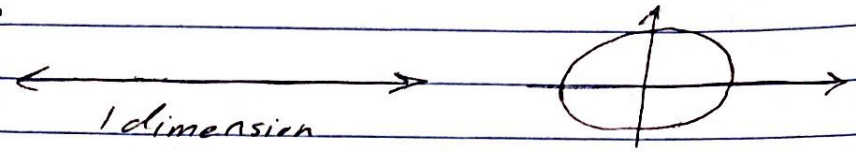
b) Now that we have Maximum likelihood for some number of data (10 days), we can use "x" as sunny parameter. It means we can put "x=1" in our distribution and find its value (mode) from ML technique. It means we have to put our problem equal to zero after derivation of y (emptiness of a table). Now, there is an approach to estimate y since y can only be 0 and 1. We can use a threshold and say if the value that we take is more than the threshold or not. This threshold can be guessed by using mean value if it is not biased it would be 0.5%. In any case with any threshold, we can easily say if it is more than threshold, y is equal to 1 which means table is available, otherwise it is not available.

c) Now we have to add another variable with name z. We can use three different variables for different times of day, but it is not suggested since it cannot be "morning", "afternoon", or "evening" at the same time. So we use a Z with values of 0 ~~for~~ morning, 1 ~~for~~ afternoon, and 2 ~~for~~ evening. Now it means based on z and x we want to estimate y (vacancy of table). Again we have to write down the formula of Bayes which will be used for $P(D|\lambda)$:

$$P(D|\lambda) = P(x_i, y_i, z_i | \lambda) = P(x_i, y_i | z_i, \lambda) P(z_i)$$
$$= P(y_i | x_i, z_i, \lambda) P(x_i) P(z_i)$$

Again like part "a" and b we have to solve ML problem by using derivative and equalizing with 0.

Q5 d) as number of dimensions increase, obviously the mean of distance also increases. It is because of the fact that with the same variance and no covariance between each of the two dimensions, we have way more space to spread our point. for example:



for 1 dimension which is a line, we have to spread 1000 points. Now for a circle (2d) we have more space to spread 1000 points. for a sphere (3d) we have again more space to spread the points. As we can see, from each point to center of our shape we have same variance and more space causes more chance for points to spread far away from center. (also $\sqrt{x^2} < \sqrt{x^2+y^2} < \sqrt{x^2+y^2+z^2} < \dots$)
Consequently, for K-means, if number of classes stay the same we have to increase its range because for example for 1D we have 0.78 as mean and for 256D we have 16.008 as mean. We cannot cover all numbers (if the number of classes and their range stays the same) for 256D as it will be covered for 1D. So, either the number of classes should increase or the range of the classes that will cover the points should increase or both. It is happening while we have same number of points. If now we increase number of points as we increase the dimensions, still we have to do the same because even if we increase the samples and afterward the ~~number of~~ mean decreases. for higher dimension to the exact mean of lower dimensions, the biggest number of each sample has been increased as dimension increased. So increasing range of classes or number of classes for higher dimensions is the only way. (PICTURE)

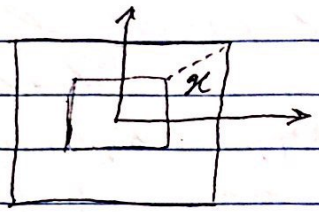
now based on K-means definition, we have an specified number of clusters each with a center which is the mid point. When we introduce another dimension it will increase the distance of each point to the center of the cluster because obviously $\sqrt{x^2+y^2} < \sqrt{x^2+y^2+z^2}$ which means distance will increase as D increases. Later it will make the points to find the nearest center of cluster and then the distance of the center of cluster to origin point will increase. It causes, at the end, the distance of center of clusters to the origin point to increase, the range of distances that a cluster supports to increase, and some points change their clusters as the number of dimensions increase.

Q5 e) This question means we have two d-dimensional hypercubes inside of each other. One with 1 side length and one with $(1-\epsilon)$ side length which means the smaller one is inside of the larger one. The volume of the bigger hypercube is 1^d and its " for the smaller one is $(1-\epsilon)^d \Rightarrow$

$\Rightarrow \frac{1^d}{(1-\epsilon)^d}$ is the ratio of the first hypercube to the volume of the second one or in the inverse order it will be $\boxed{\frac{(1-\epsilon)^d}{1^d}}$.

As d gets larger 1^d will stay 1 and $(1-\epsilon)^d$ will be smaller since $0 < 1-\epsilon < 1$. So the volume of the " hypercube decreases. It means that distance of each point on the corner of bigger hyper cube to the corresponding corner of the smaller hypercube will increase. It causes some points in the smaller hypercube falls into the bigger hypercube. If the points are on the surface of the hyper cubes the distance between each two of them from bigger hypercube to the smaller one will increase. (if we assume each hypercube is our samples) Now, if the points should spread inside of the hypercube as the volume decreases distance of each angle to the center will decrease (for the smaller hypercube). But for the larger it will stay the same. At the same time by adding a dimension we are adding a new variable to our distance (e.g. $\sqrt{x^2} \leq \sqrt{x^2+y^2} \leq \sqrt{x^2+y^2+z^2} \leq \dots$) so distance will increase. It will make the density to increase. ~~how we cannot say much about its distance to origin point~~ We can see it mathematically in the next page.

for example for 2D we have:



$$x = \sqrt{\underbrace{\left(\frac{1-(1-\epsilon)}{2}\right)^2}_{\alpha} + \left(\frac{1-(1-\epsilon)}{2}\right)^2}$$

$$x = \sqrt{2\alpha^2} = \sqrt{2}\,\alpha$$

for 3D:

$$x = \sqrt{3\alpha^2} = \sqrt{3}\,\alpha$$

for d dimensions:

$$x = \sqrt{d\,\alpha^2} = \sqrt{d}\,\sqrt{\alpha^2} = \sqrt{d}\,\alpha$$

→ ~~we are~~ as you can see, as d increases x gets larger and the volume between larger hypercube and the smaller one increases so it proves that as dimensions increase, each two points' ~~bet~~ distance will also increase.