1. (**30 points**) Find the Maximum Likelihood Estimation (MLE) of θ in the following probabilistic density functions. In each case, consider a random sample of size n. Show your calculation:

   Observe the following notation for the random sample X of size n: $X = \{x_i\}_{i=1}^n$

   (a) $f(x|\theta) = \frac{x}{\theta^2} \exp\left\{\frac{-x^2}{2\theta^2}\right\}, x \geq 0$

   Log likelihood:

$$\begin{aligned}
\mathcal{L}(\theta|X) &= \log(\ell(\theta|X)) \\
&= \log\left(\prod_{i=1}^{n} f(x_i|\theta)\right) \\
&= \log\left(\prod_{i=1}^{n} \frac{x_i}{\theta^2} \exp\left\{\frac{-(x_i)^2}{2\theta^2}\right\}\right) \\
&= \sum_{i=1}^{n}\left(\log\left(\frac{x_i}{\theta^2}\right) + \log\left(\exp\left\{\frac{-(x_i)^2}{2\theta^2}\right\}\right)\right) \\
&= \sum_{i=1}^{n}\left(\log(x_i) - \log(\theta^2) + \frac{-(x_i)^2}{2\theta^2}\right)
\end{aligned}$$

   It follows that:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta|X)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left(\sum_{i=1}^{n}\left(\log(x_i) - \log(\theta^2) + \frac{-(x_i)^2}{2\theta^2}\right)\right) \\
&= 0 + \sum_{i=1}^{n}\left(\frac{-2}{\theta} + \frac{(x_i)^2}{\theta^3}\right) \\
&= \frac{-2n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n}\left((x_i)^2\right)
\end{aligned}$$

   Setting equal to zero and solving for θ gives:

$$\begin{aligned}
0 &= \frac{-2n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n}\left((x_i)^2\right) \\
\iff \quad \frac{2n}{\theta} &= \frac{1}{\theta^3}\sum_{i=1}^{n}\left((x_i)^2\right) \\
\iff \quad 2n\theta^2 &= \sum_{i=1}^{n}\left((x_i)^2\right) \\
\iff \quad \theta^2 &= \frac{\sum_{i=1}^{n}\left((x_i)^2\right)}{2n} \\
\iff \quad \theta &= \pm\sqrt{\frac{\sum_{i=1}^{n}\left((x_i)^2\right)}{2n}}
\end{aligned}$$

   Hence, the MLE of θ is:

$$\hat{\theta} = \texttt{argmax}_\theta \mathcal{L}(\theta|X) = \pm\sqrt{\frac{\sum_{i=1}^{n}\left((x_i)^2\right)}{2n}}$$

(b) $f(x|\alpha,\beta,\theta) = \alpha\theta^{-\alpha\beta}x^{\beta}\exp\left\{-\left(\frac{x}{\theta}\right)^{\beta}\right\}, x \geq 0, \alpha > 0, \beta > 0, \theta > 0$

Log likelihood:

$$\begin{aligned}
\mathcal{L}(\alpha,\beta,\theta|X) &= \log(\ell(\alpha,\beta,\theta|X)) \\
&= \log\left(\prod_{i=1}^{n} f(x_i|\alpha,\beta,\theta)\right) \\
&= \log\left(\prod_{i=1}^{n} \alpha\theta^{-\alpha\beta}x_i^{\beta}\exp\left\{-\left(\frac{x_i}{\theta}\right)^{\beta}\right\}\right) \\
&= \sum_{i=1}^{n}\left(\log(\alpha) - \alpha\beta\log(\theta) + \log\left(x_i^{\beta}\right) - \left(\frac{x_i}{\theta}\right)^{\beta}\right) \\
&= n\log(\alpha) - n\alpha\beta\log(\theta) + \sum_{i=1}^{n}\left(\beta\log(x_i) - \frac{x_i^{\beta}}{\theta^{\beta}}\right)
\end{aligned}$$

It follows that:

$$\begin{aligned}
\frac{\partial\mathcal{L}(\alpha,\beta,\theta|X)}{\partial\theta} &= \frac{\partial}{\partial\theta}\left(n\log(\alpha) - n\alpha\beta\log(\theta) + \sum_{i=1}^{n}\left(\beta\log(x_i) - \frac{x_i^{\beta}}{\theta^{\beta}}\right)\right) \\
&= 0 - \frac{n\alpha\beta}{\theta} + \sum_{i=1}^{n}\left(0 + \frac{\beta x_i^{\beta}}{\theta^{\beta+1}}\right) \\
&= \frac{n\alpha\beta}{\theta} + \sum_{i=1}^{n}\left(\frac{\beta x_i^{\beta}}{\theta^{\beta+1}}\right) \\
&= \frac{n\alpha\beta}{\theta} + \frac{n\beta}{\theta^{\beta+1}}\sum_{i=1}^{n}(x_i^{\beta})
\end{aligned}$$

Setting equal to zero and solving for θ gives:

$$\begin{aligned}
0 &= \frac{n\alpha\beta}{\theta} + \frac{n\beta}{\theta^{\beta+1}}\sum_{i=1}^{n}(x_i^{\beta}) \\
\iff \quad 0 &= \alpha\theta^{\beta} + \sum_{i=1}^{n}(x_i^{\beta}) \\
\iff \quad \alpha\theta^{\beta} &= -\sum_{i=1}^{n}(x_i^{\beta}) \\
\iff \quad \theta^{\beta} &= \frac{\sum_{i=1}^{n}(x_i^{\beta})}{-\alpha} \\
\iff \quad \theta &= \sqrt[\beta]{\frac{\sum_{i=1}^{n}(x_i^{\beta})}{-\alpha}}
\end{aligned}$$

Hence, the MLE of θ is:

$$\hat{\theta} = \texttt{argmax}_{\theta}\mathcal{L}(\alpha,\beta,\theta|X) = \sqrt[\beta]{\frac{\sum_{i=1}^{n}(x_i^{\beta})}{-\alpha}}$$

(c) $f(x|\theta) = \frac{1}{\theta}, 0 \leq x \leq \theta, \theta > 0$ (Hint: You can draw the likelihood function)

Realize that the objective is to find the value of θ for which the likelihood function,

$$\ell(\theta|X) = \prod_{i=1}^{n} f(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}$$

is maximized. Clearly, $\frac{1}{\theta^n}$ will be maximized when θ is made as small as possible, while still satisfying the constraint that $\forall x_i \in X, \theta \geq x_i$, and $\theta > 0$. Hence, for the random sample $X = \{x_i\}_{i=1}^{n}$, the MLE of θ is:

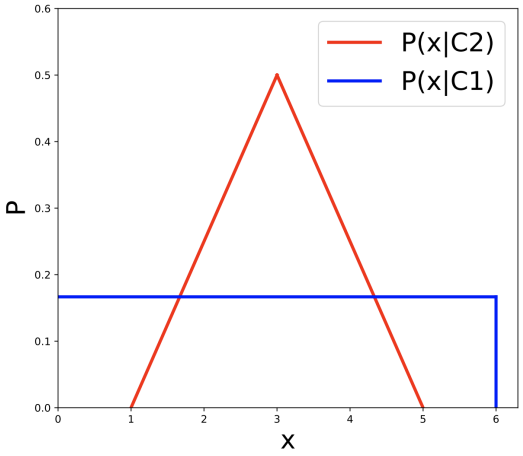
$$\hat{\theta} = \texttt{argmax}_\theta \mathcal{L}(\theta|X) = \max X$$

where $\max X$ is the largest sample point x_i of the random sample X.

2. (**30 points**) We want to build a pattern classifier with continuous attribute using Bayes' Theorem. The object to be classified has one feature, x in the range $0 \leq x < 6$. The conditional probability density functions for each class are listed below:

$$P(x|C_1) = \begin{cases} \frac{1}{6} & if\ 0 \leq x < 6 \\ 0 & otherwise \end{cases}$$

$$P(x|C_2) = \begin{cases} \frac{1}{4}(x-1) & if\ 1 \leq x < 3 \\ \frac{1}{4}(5-x) & if\ 3 \leq x < 5 \\ 0 & otherwise \end{cases}$$



(a) Assuming equal priors, $P(C_1) = P(C_2) = 0.5$, classify an object with the attribute value $x = 2.5$.

Assuming that the priors are equal, we can predict that the object with the attribute value $x = 2.5$ belongs to class C_2. Observe,

$$P(x = 2.5 \mid C_1) = \frac{1}{6} \quad \texttt{and} \quad P(x = 2.5 \mid C_2) = \frac{1}{4}(2.5 - 1) = \frac{3}{8}$$

It follows that:

$$\begin{aligned} & P(x = 2.5 \mid C_2) > P(x = 2.5 \mid C_1) \\ \iff \quad & \frac{P(C_1)}{P(x = 2.5)} P(x = 2.5 \mid C_2) > \frac{P(C_2)}{P(x = 2.5)} P(x = 2.5 \mid C_1) \quad \texttt{// by assumption that priors are equal} \\ \iff \quad & P(C_2 \mid x = 2.5) > P(C_1 \mid x = 2.5) \quad \texttt{// Bayes' Rule} \end{aligned}$$

Hence, the posterior probability for class C_2 is greater than that of class C_1, and, as such, class C_2 should be the predicted classification.

(b) Assuming unequal priors, $P(C_1) = 0.7, P(C_2) = 0.3$, classify an object with the attribute value $x = 4$.

Using Bayes' rule, we can make a class prediction by determining the posterior probabilities for each class as follows:

$$P(C_1 \mid x = 4) = \frac{P(C_1)P(x = 4 \mid C_1)}{P(x = 4)} = \frac{(0.7)(\frac{1}{6})}{P(x = 4)} = \frac{\frac{7}{60}}{P(x = 4)}$$

$$P(C_2 \mid x = 4) = \frac{P(C_2)P(x = 4 \mid C_2)}{P(x = 4)} = \frac{(0.3)(\frac{1}{4}(5 - 4))}{P(x = 4)} = \frac{\frac{3}{40}}{P(x = 4)}$$

Hence, with priors of $P(C_1) = 0.7$ and $P(C_2) = 0.3$, the predicted class of the object with the attribute value $x = 4$ is C_1 since $\frac{7}{60} > \frac{3}{40}$ which, in turn, means that $P(C_1 \mid x = 4) > P(C_2 \mid x = 4)$.

Continued on next page ⟶

(c) Consider a decision function $\phi(x)$ of the form $\phi(x) = (|x-3|) - \alpha$ with one free parameter α in the range $0 \leq \alpha \leq 2$. You classify a given input x as class 2 if and only if $\phi(x) < 0$, or equivalently $3-\alpha < x < 3+\alpha$, otherwise you choose x as class 1. Assume equal priors, $P(C_1) = P(C_2) = 0.5$, what is the optimal decision boundary - that is, what is the value of α which minimizes the probability of misclassification? What is the resulting probability of misclassification with this optimal value for α? (Hint: take advantage of the symmetry around $x = 3$.)

First, let us find the values of x for which the two conditional PDF functions intersect:

$$\begin{aligned} & P(x|C_1) = P(x|C_2) \\ \iff \quad & \frac{1}{6} = \frac{1}{4}(x-1) \quad \texttt{or} \quad \frac{1}{6} = \frac{1}{4}(5-x) \\ \iff \quad & x = \frac{4}{6} + 1 = \frac{5}{3} \quad \texttt{or} \quad x = 5 - \frac{4}{6} = \frac{13}{3} \end{aligned}$$

As a result of the equal priors, in order to mimimize the probability of misclassification we want the decision function to evaluate to a value less than zero for values of $x \in (\frac{5}{3}, \frac{13}{3})$:

$$\begin{aligned} |\frac{5}{3} - 3| - \alpha = |\frac{13}{3} - 3| - \alpha = \frac{4}{3} - \alpha = 0 \\ \iff \alpha = \frac{4}{3} \end{aligned}$$

Notice that with this optimal value for α, there is a 0% chance of misclassification for $x \in [0,1) \cup (5,6)$. For $x \in [1,5]$, then, we can find the probability of misclassification (p_{miss}) by realizing that the classifier will make the wrong classification for $x \in [3, \frac{13}{3}]$ and $x \in [\frac{13}{3}, 5]$ when x is of class C_1 and class C_2 respectively. Exploiting the symmetry about $x = 3$, then gives rise to the following calculation:

$$\begin{aligned} p_{miss} &= 2\left(\int_3^{\frac{13}{3}} P(C_1|x)P(x)dx + \int_{\frac{13}{3}}^5 P(C_2|x)P(x)dx\right) \\ &= 2\left(\int_3^{\frac{13}{3}} \frac{P(C_1)P(x|C_1)}{P(x)}P(x)dx + \int_{\frac{13}{3}}^5 \frac{P(C_2)P(x|C_2)}{P(x)}P(x)dx\right) \\ &= 2\left(\int_3^{\frac{13}{3}} P(C_1)P(x|C_1)dx + \int_{\frac{13}{3}}^5 P(C_2)P(x|C_2)dx\right) \\ &= 2\left(\int_3^{\frac{13}{3}} (0.5)P(x|C_1)dx + \int_{\frac{13}{3}}^5 (0.5)P(x|C_2)dx\right) \\ &= \int_3^{\frac{13}{3}} P(x|C_1)dx + \int_{\frac{13}{3}}^5 P(x|C_2)dx \\ &= \int_3^{\frac{13}{3}} \frac{1}{6}dx + \int_{\frac{13}{3}}^5 \frac{1}{4}(5-x)dx \\ &= \frac{1}{6}(\frac{13}{3} - 3) + \frac{5}{4}(5 - \frac{13}{3}) - \frac{1}{8}(5^2 - \left(\frac{13}{3}\right)^2) \\ &= \frac{2}{9} + \frac{1}{18} = \frac{5}{18} = 0.2\overline{7} \end{aligned}$$

Hence, there is a $27.\overline{7}\%$ chance of misclassification with the optimal value of α.

3. (**40 points**) In this programming exercise you will implement three multivariate Gaussian classifiers, with different assumptions as follows:

   - Assume S_1 and S_2 are learned independently (learned from the data from each class).
   - Assume $S_1 = S_2$ (learned from the data from both classes).
   - Assume $S_1 = S_2$ (learned from the data from both classes), and the covariance is a diagonal matrix.

   What is the discriminant function in each case? Show in your report and briefly explain.

   In each case, the discriminant function is nothing more than an expansion of the following with any constants dropped:

   $$g_i(x) = \log\left(p(x|C_i) + \log P(C_i)\right)$$

   It follows that in the case of the first assumption that S_1 and S_2 are learned independently, the following discrimant function was used:

   $$g_i(x^t) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x^t - \mu_i)^T\Sigma_i^{-1}(x^t - \mu_i) + \log P(C_i)$$

   In the case of assumptions two and three, the following discrimant function was used:

   $$g_i(x^t) = -\frac{1}{2}(x^t - \mu_i)^T\Sigma^{-1}(x^t - \mu_i) + \log P(C_i)$$

   Note that the $-\frac{1}{2}\log|\Sigma_i|$ term is dropped as this becomes a constant when $S_1 = S_2$. The fact that the covariance matrix is diagonal under assumption three means that the computation of Σ^{-1} is made trivial as all that is required is to take the reciprocal of each diagonal entry.

   Report the confusion matrix on the test set for each assumption. Briefly explain the results.

```
[anthony@ <<22:58:21>> ~/.../hw1/hw1_programming]$ python3 hw1.py
Confusion Matrix for Gaussian Discriminant with class-dependent covariance
[[14  5]
 [16 65]]
Confusion Matrix for Gaussian Discriminant with class-independent covariance
[[24  5]
 [ 6 65]]
Confusion Matrix for Gaussian Discriminant with diagonal covariance
[[26 11]
 [ 4 59]]
```

Figure 1: Confusion matrices over test set for each assumption

Each 2-by-2 confusion matrix displays the number of correct classifications over the test set on the diagonal entries, and the incorrect classifications on the off diagonal entries. For example, in the first cofusion matrix in figure 1, of the 30 total samples that are labeled as belonging to class 1, 14 were correctly classified by the class-dependent covariance classifier, while 16 were incorrectly classified. Similarly, of the 70 total samples that are labeled as belonging to class 2, 5 were incorrectly classified, while 65 were correctly classified. It follows that both the class-dependent covariance, and class-independent covariance classifiers performed equally well in classifying samples belonging to class 2, beating the performance of the diagonal covariance classifier. The diagonal covariance classifier outperformed the other two in its classification of samples belonging to class 1–though the class-independent covariance classifier was not far behind. These results seem to suggest that the sample features are class-independent (as evidenced by the improved performance of the class-independent classifier over the class-dependent classifier), and mostly independent of each other (as evidenced by the similar performance of the diagonal and class-independent covariance classifiers). To further elicidate these results, more training data could be included as the current training set is comprised of only 50 samples.