1. **Probability and independence (10 points)** (Question 2.9 from Koller and Friedman) Prove or disprove (by providing a counterexample) each of the following properties of independence.

   (a) $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z})$ implies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$

   (b) $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ and $(\mathbf{X}, \mathbf{Y} \perp \mathbf{W} \mid \mathbf{Z})$ imply $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z})$

   (c) $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z})$ and $(\mathbf{Y} \perp \mathbf{W} \mid \mathbf{Z})$ imply $(\mathbf{X}, \mathbf{W} \perp \mathbf{Y} \mid \mathbf{Z})$

   (d) $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ and $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W})$ imply $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$

   Solution:

   Note: We use W,X,Y,Z in this solution. They should be w,x,y,z as per notation we use in the class.

   (a) $$\begin{aligned} &\text{p(X,Y,W} \mid \text{Z)} = \text{p(X} \mid \text{Z) p(Y,W} \mid \text{Z) (Given) (1)} \\ &\text{p(X,Y} \mid \text{Z)} \quad = \textstyle\sum_W \text{p(X,Y, W} \mid \text{Z)} \\ &\qquad\qquad = \textstyle\sum_W \text{p(X} \mid \text{Z) p(Y,W} \mid \text{Z) (from 1)} \\ &\qquad\qquad = \text{p(X} \mid \text{Z)} \textstyle\sum_W \text{p(Y,W} \mid \text{Z)} \\ &\qquad\qquad = \text{p(X} \mid \text{Z) p(Y} \mid \text{Z)} \end{aligned}$$

   Thus (X $\perp$ Y | Z).

   (b) This is a simple consequence from (a) (replace (X, Y, W) in (a) by (W, X, Y) and you get (b) by commutativity of conditional independence).

   (c) $$\begin{aligned} &\text{p(X,Y,W} \mid \text{Z)} = \text{p(X} \mid \text{Z) p(Y,W} \mid \text{Z) (Given) (1)} \\ &\text{p(Y,W} \mid \text{Z)} = \text{p(Y} \mid \text{Z) p(W} \mid \text{Z) (Given) (2)} \\ &\text{p(X,Y,W} \mid \text{Z)} \quad = \text{p(X} \mid \text{Z) p(Y,W} \mid \text{Z)} \\ &\qquad\qquad = \text{p(X} \mid \text{Z) p(Y} \mid \text{Z) p(W} \mid \text{Z) (from 2)} \\ &\qquad\qquad = \text{p(Y} \mid \text{Z)} \textstyle\sum_Y \text{p(X} \mid \text{Z) p(Y} \mid \text{Z) p(W} \mid \text{Z)} \\ &\qquad\qquad = \text{p(Y} \mid \text{Z)} \textstyle\sum_Y \text{p(X} \mid \text{Z) p(Y,W} \mid \text{Z) (from 2)} \\ &\qquad\qquad = \text{p(Y} \mid \text{Z)} \textstyle\sum_Y \text{p(X,W,Y} \mid \text{Z) (from 1)} \\ &\qquad\qquad = \text{p(Y} \mid \text{Z) p(X,W} \mid \text{Z)} \end{aligned}$$

   Thus (X,W $\perp$ Y | Z).

   (d) This property is not correct.

   Consider the following counter-example: Z = X $\oplus$ (Y $\oplus$ W) where X, Y, and W are mutually independent Bernoulli(1/2) variables, and $\oplus$ denotes XOR (i.e. X $\oplus$ Y is one iff exactly one of X or Y is one). Now given (Z, W), X and Y become dependent. For example, if Z=1 and W=1, then X should be negation of Y.

   This is a generalization of the counter-example Z = X $\oplus$ Y. Here we can have X independent of Y, Y independent of Z, and Z independent of X. However, given Y, Z is dependent on X. Also, given X, Z is dependent on Y.

   Later in the class, we will see this phenomenon with the v-structures in a Bayes net.

   Consider the following example: W = Y $\oplus$ U, and Z = X $\oplus$ U. The Bayes net for this example is as given below:

$$X \to Z \leftarrow U \to W \leftarrow Y$$

(X $\perp$ Y | Z) and (X $\perp$ Y | W). But given both Z and W, X and Y becomes dependent through U (because of the v-structures). Note that by solving for U = Y $\oplus$ W, we get back the original example (where U disappeared).

2. **Bayesian inference and MAP (10 points)**

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim}$ Multinomial$(1, \boldsymbol{\pi})$ on k elements with a similar notation as seen in class: the encoding for a possible value $\boldsymbol{x}_i$ of the random vector $\boldsymbol{X}_i$ is $\boldsymbol{x}_i = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$ with $x_j^{(i)} \in \{0,1\}$ and $\sum_{j'=1}^k x_{j'}^{(i)} = 1$ (that is, we have a j^* where $x_{j^*}^{(i)} = 1$ and for each $j' \neq j^*$, $x_{j'}^{(i)} = 0$). Consider a Dirichlet prior distribution on $\boldsymbol{\pi}$: $\boldsymbol{\pi} \sim$ Dirichlet$(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ and $\alpha_j > 0$ for all j.

(The Dirichlet distribution is a distribution for a *continuous* random vector $\boldsymbol{\pi}$ which lies on the probability simplex Δ_k. Recall $\Delta_k := \{\boldsymbol{\pi} \in \mathbb{R}^k : 0 \leq \pi_j \leq 1 \text{ and } \sum_{j=1}^k \pi_j = 1\}$. Its probability density function[1] is $p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \pi_j^{\alpha_j - 1}$. Note that the beta distribution seen is class is the special case of a Dirichlet distribution for $k = 2$, like the binomial distribution is the special case of a multinomial distribution for $k = 2$.)

(a) Supposing that the data is IID, what are the conditional independence statements that we can state for the joint distribution $p(\boldsymbol{\pi}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$? Write your answer in the form of formal conditional independence statements as in question 1 (a) - (d).

(b) Derive the posterior distribution $p(\boldsymbol{\pi} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

(c) Derive the marginal probability $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ (or equivalently $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid \boldsymbol{\alpha})$.) This quantity is called the *marginal likelihood* and we will see it again when doing model selection later in the course.

(d) Derive the MAP estimate $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ assuming that the hyperparameters for the Dirichlet prior satisfy $\alpha_j > 1$ for all j. Compare this MAP estimator with the MLE estimator for the multinomial distribution seen in class: what can you say when k is extremely large?[2]

Solution:

(a) Given π, any subset of $\{X_1, .., X_n\}$ is independent of any other disjoint subset $\{X_1, .., X_n\}$. Mathematically,

$$X_i \perp X_j \mid \pi$$

X_i and X_j could be any two variables or disjoint subsets of variables.

[1] Formally, this density function is taken with respect to a $(k-1)$-dimensional Lebesgue measure defined on Δ_k. But equivalently, you can also think of the density to be a standard one in dimension $k-1$ defined for the first $k-1$ components $(\pi_1, \ldots, \pi_{k-1})$ which are restricted to the (full) dimensional polytope $T_{k-1} := \{(\pi_1, \ldots, \pi_{k-1}) \in \mathbb{R}^{k-1} : 0 \leq \pi_j \leq 1 \text{ and } \sum_{j=1}^{k-1} \pi_j \leq 1\}$, and then letting $\pi_k := 1 - \sum_{j=1}^{k-1} \pi_j$ in the formula. Note that this bijective transformation from T_{k-1} onto Δ_k has a Jacobian with a determinant of 1, which is why the two Lebesgue measures are equivalent and one does not need to worry about which of the two spaces we are defining the density on.

[2] An example of this is when modeling the appearance of words in a document: here k would be the numbers of words in a vocabulary. The MAP estimator derived above when the prior is a symmetric Dirichlet is called *additive smoothing* or *Laplace smoothing* in statistical NLP.

(b) $\pi \sim$ Dirichlet(α). So,

$$p(\pi;\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha-1}$$

Let us define the normalization factor $D(\alpha)$ for the Dirichlet (we will use it in (c)):

$$D(\alpha) := \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$$

$$p(\pi \mid X_1, ..., X_n) \propto \prod_{i=1}^{n} p(X_i \mid \pi) p(\pi;\alpha)$$

Also,

$$\prod_{i=1}^{n} p(X_i \mid \pi) = \prod_{k=1}^{K} \pi_k^{n^k}$$

where n^k is the number of times k-th element was 1 in n samples. Then we can write,

$$\begin{aligned}
p(\pi \mid X_1, ..., X_n) &\propto \prod_{k=1}^{K} \pi_k^{n^k} \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha-1} \\
&\propto \prod_{k=1}^{K} \pi_k^{n^k} \pi_k^{\alpha-1} \\
&\propto \prod_{k=1}^{K} \pi_k^{n^k+\alpha-1} \\
&= Dir(\pi \mid \alpha + n)
\end{aligned}$$

where $n = \{n^1, ..., n^k\}$. Above, we recognized the shape of the Dirichlet distribution, which is why we did not need to explicitly compute the normalization constant.

(c)

$$\begin{aligned}
p(X_1, ..., X_n \mid \alpha) &= \int_{\pi} \prod_{i=1}^{n} p(X_i \mid \pi) p(\pi \mid \alpha) d\pi \\
&= \int_{\pi} \prod_{k=1}^{K} \pi_k^{n^k} \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha-1} d\pi \\
&= \frac{1}{D(\alpha)} \int_{\pi} \prod_{k=1}^{K} \pi_k^{n^k+\alpha_k-1} d\pi \\
&= \frac{D(n+\alpha)}{D(\alpha)} \int_{\pi} \text{Dir}(\pi|\alpha+n) d\pi \\
&= \frac{D(n+\alpha)}{D(\alpha)}.
\end{aligned}$$

(d)
$$\hat{\pi}^{MAP} = \underset{\pi}{\text{argmax}}\, p(\pi|x_1, ..., x_n, \alpha)$$

We need to take into account of the constraint $\sum_{j=1}^{k} \pi_j = 1$. This can be achieved using a Lagrange multiplier λ and maximizing

$$\sum_{j=1}^{k}(m_j + \alpha_j - 1)\text{log}\pi_j + \lambda(\sum_{j=1}^{k} \pi_j - 1)$$

$$\hat{\pi}^{MAP} = \frac{m_j + \alpha_j - 1}{n + \sum_{j=1}^{k} \alpha_j - k}$$

This can be seen as the smoothed version of MLE estimate which is m_j/n. When k is extremely large, m_j will be very sparse and hence MLE estimate will be mostly zeros. This is undesirable (this is clear overfitting as we can never claim that some class has really *zero* probability unless we have seen an infinite amount of data). On the other hand, MAP estimate will have some small non-zero probability even when m_j is sparse.

3. **Properties of estimators (20 points)**

(a) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Find the maximum likelihood estimator (MLE) and determine its properties: bias, variance, consistency (yes or no). (Recall that the pmf for a Poisson r.v. is $p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$ for $x \in \mathbb{N}$.)

(b) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ where we suppose that $n > 10$. If we take as an estimator of p, $\hat{p} := \frac{1}{10}\sum_{i=1}^{10} X_i$, determine its properties: bias, variance, consistency (yes or no).

(c) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$. Find the MLE and determine its properties: bias, variance, consistency (yes or no).
(Hint: Let $Y = \max\{X_1, \ldots, X_n\}$. For each c, $P(Y < c) = P(X_1 < c, X_2 < c, \ldots, X_n < c) = P(X_1 < c)P(X_2 < c) \cdots P(X_n < c)$).

(d) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ (where $\mu \in \mathbb{R}$) for $n \geq 2$ to simplify. Show that the MLE[3] for $\theta := (\mu, \sigma^2)$ is $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Also determine the properties only for $\hat{\sigma}^2$: its bias, the variance and whether it is consistent.
(Hint: for the variance of $\hat{\sigma}^2$ calculation, you may use the fact that $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \overset{\text{d}}{=} \chi^2_{n-1}$, where χ^2_{n-1} is the chi-squared distribution with $(n-1)$ degrees of freedom, and that $\text{Var}[\chi^2_{n-1}] = 2(n-1)$.)

Solution:

(a) $X_1, ..., X_n \sim \text{Poisson}(\lambda)$.

[3] Note that formally we should use the notation $\hat{\sigma^2}$ (which looks ugly!) as we are estimating the variance σ^2 of a Gaussian rather than its standard deviation σ. But as the MLE is invariant to a re-parameterization of the full parameter space (from σ^2 to σ e.g.), then we simply have $\hat{\sigma^2} = \hat{\sigma}^2$ and the distinction is irrelevant.

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}[\ln(p(X_1, X_2, ...X_n \mid \lambda))] &= 0 \\
\frac{\partial}{\partial\lambda}[\ln(p(X_1 \mid \lambda)p(X_2 \mid \lambda)...p(X_n \mid \lambda))] &= 0 \\
\frac{\partial}{\partial\lambda}[\sum_{i=1}^{n} \ln(p(X_i \mid \lambda)] &= 0 \\
\frac{\partial}{\partial\lambda}[\sum_{i=1}^{n} \ln(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!})] &= 0 \\
\frac{\partial}{\partial\lambda}[\sum_{i=1}^{n} -\lambda + x_i\ln\lambda - \ln x_i!] &= 0 \\
\frac{\partial}{\partial\lambda}[-n\lambda + \ln\lambda\sum_i X_i - \ln\prod_i x_i!] &= 0 \\
-n + \frac{\sum X_i}{\lambda} &= 0 \\
\hat{\lambda} &= \frac{\sum X_i}{n}
\end{aligned}
$$

is the MLE.
Bias :

$$
\begin{aligned}
E[\hat{\lambda}] &= E[\frac{\sum X_i}{n}] \\
&= \frac{1}{n}\sum E[X_i] \\
&= \frac{1}{n}n\lambda \\
&= \lambda
\end{aligned}
$$

There is no bias.
Variance:

$$
\begin{aligned}
Var[\hat{\lambda}] &= Var[\frac{\sum X_i}{n}] \\
&= \frac{1}{n^2}Var[\sum X_i] \\
&= \frac{1}{n^2}\sum Var[X_i] \\
&= \frac{1}{n^2}n\lambda \\
&= \frac{\lambda}{n}
\end{aligned}
$$

Consistency: Yes this is consistent. Because, bias is zero and $\sqrt{\frac{\lambda}{n}} \to$ as $n \to \infty$.

(b) $X_1, ..., X_n \sim$ Bernoulli(p).
$\hat{p} = \frac{1}{10}\sum_{i=1}^{10} X_i$
Bias :

$$\begin{aligned} E[\hat{p}] &= \frac{1}{10}\sum_{i=1}^{10} E[X_i] \\ &= \frac{1}{10} 10p \\ &= p \end{aligned}$$

There is no bias.
Variance:

$$\begin{aligned} Var[\hat{p}] &= Var[\frac{1}{10}\sum_{i=1}^{10} X_i] \\ &= \frac{1}{100}\sum_{i=1}^{10} Var[X_i]] \\ &= \frac{1}{100} 10p(1-p) \\ &= \frac{1}{10} p(1-p) \end{aligned}$$

Consistency: This is not a consistent estimator. Because, bias is zero but variance will never go to zero as $n \to \infty$.

(c) $X_1, ..., X_n \sim$ Uniform$(0, \theta)$.

$p(X_1, ..., X_n; \theta) = \prod p(X_i; \theta) = \frac{1}{\theta^n} \quad if \ 0 \leq X_i \leq \theta$
p is 0 if any of the X_i is greater than θ. To get MLE for θ, we know that $\theta \geq X_i$ and it should maximize $1/\theta^n$. As θ increases $1/\theta^n$ decreases. Thus

$$\hat{\theta} = max(X_1, ..., X_n)$$

is the MLE.

Let $Y = \max(X_1, ..., X_n)$.

$$\begin{aligned} \text{cdf}(y) = F(y) &= p(Y \leq y) \\ &= p(max(X_i) \leq y) \\ &= p(X_1 \leq y, ..., X_n \leq y,) \\ &= p(X_1 \leq y) ... P(X_n \leq y,) \\ &= p^n(x \leq y) \\ &= (\frac{y}{\theta})^n \end{aligned}$$

Now we can derive the pdf p as follows:

$$\begin{aligned} p(y) &= F'(y) \\ &= n(\frac{y}{\theta})^{n-1}\frac{1}{\theta} \\ &= n(\frac{1}{\theta})^{n}y^{n-1} \quad \text{for } 0 \leq y \leq \theta \end{aligned}$$

Bias:

$$\begin{aligned} E[\hat{\theta}] &= \int_{-\infty}^{\infty} yp(y)dy \\ &= \int_{0}^{\theta} yn(\frac{1}{\theta})^{n}y^{n-1} \\ &= n(\frac{1}{\theta})^{n}\int_{0}^{\theta} yy^{n-1} \\ &= n(\frac{1}{\theta})^{n}\int_{0}^{\theta} y^{n} \\ &= n(\frac{1}{\theta})^{n}\frac{\theta^{n+1}}{n+1} \\ &= \frac{n}{n+1}\theta \end{aligned}$$

Thus this estimator is biased.

$$\begin{aligned} E[\hat{\theta}^2] &= \int_{-\infty}^{\infty} y^2p(y)dy \\ &= \int_{0}^{\theta} y^2n(\frac{1}{\theta})^{n}y^{n-1} \\ &= n(\frac{1}{\theta})^{n}\int_{0}^{\theta} y^2y^{n-1} \\ &= n(\frac{1}{\theta})^{n}\int_{0}^{\theta} y^{n+1} \\ &= n(\frac{1}{\theta})^{n}\frac{\theta^{n+2}}{n+2} \\ &= \frac{n}{n+2}\theta^2 \end{aligned}$$

Variance:

$$\begin{aligned}
var(\hat{\theta}) &= E[\hat{\theta}^2] - E[\hat{\theta}]^2 \\
&= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 \\
&= n\theta^2[\frac{1}{n+2} - \frac{n}{(n+1)^2}]
\end{aligned}$$

This estimator is consistent since both bias and variance becomes 0 as $n \to \infty$

(d) (solution reproduced with permission from Dong-Hyun Lee)The likelihood is

$$\begin{aligned}
P(X_1, ..., X_n|\mu, \sigma^2) &= \prod_{i=1}^{n} P(X_i|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(X_i - \mu)^2\right] \\
\log P(X_1, ..., X_n|\mu, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2
\end{aligned} \tag{1}$$

The derivative of log likelihood w.r.t μ, σ^{-2} zero for MLE,

$$\begin{aligned}
-\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu - X_i) &= 0 \\
\frac{n}{2}\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2 &= 0
\end{aligned} \tag{2}$$

From this, we can get the MLE for μ, σ^{-2}.

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.
\end{aligned} \tag{3}$$

Simon's addition: to be rigorous, we need to check that this is really a maximum (and not a minimum or a saddle point).[4] Taking the derivative of (2), we get that the Hessian with respect to (μ, σ^{-2}) evaluated at $(\hat{\mu}, \hat{\sigma}^{-2})$ is diagonal, with only negative entries, and thus negative definite, i.e., we have a strict local maximum. Even if this the only stationary point of the log-likelihood (which is continuously differentiable on the interior of its domain), we still need to check that the boundary of the domain cannot get bigger values.[5] In our case, the objective goes to $-\infty$ as $\|\mu\| \to \infty$ for any σ^2; and it also goes to $-\infty$ as $\sigma^2 \to 0$ or $\sigma^2 \to \infty$ for any μ as long as we assume that there are at least two X_i's which have different values (i.e. the empirical variance is non-zero). So under the assumption that $\hat{\sigma}^2 > 0$ in (3), we have that the stationary point of (3) is the global maximum. If all the X_i's are the same, i.e. $\hat{\sigma}^2 = 0$, then (3) gives the *limiting* global

[4] As I mentioned in class, sometimes you can be unlucky and what you computed is actually a local minimum, and the maximum is actually achieved at the boundary of the domain! See the next footnote.

[5] For example, consider minimizing the function $(x^2 + \delta)\exp(-x)$, with δ a small number. It has a strict local minimum at $x \approx \delta/2$ and a strict local maximum at $x \approx 2$, but it reaches its global minimum at the boundary of the domain ($x \to \infty$).

maximum of the log-likelihood (i.e. is achieved at the boundary of the domain). Also note that the log-likelihood function here is not concave in σ^2, which is why we needed to consider the boundary. For a concave function, stationarity is sufficient for global optimality, so things are much simpler. We will also see later that using the exponential family perspective simplifies a lot the above computations for the Gaussian.

Next, $\hat{\sigma}^2$ is biased because

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2\sum_{i=1}^{n}X_i\bar{X} + \sum_{i=1}^{n}\bar{X}^2\right]\\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right]\\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\bar{X}^2\right] = \mathrm{Var}\left[X\right] + \mathbb{E}\left[X\right]^2 - \mathrm{Var}\left[\bar{X}\right] - \mathbb{E}\left[\bar{X}\right]^2 = \mathrm{Var}\left[X\right] - \mathrm{Var}\left[\bar{X}\right]\\ &= \sigma^2 - \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2\end{aligned} \tag{4}$$

The bias is $-\sigma^2/n$. The variance of the estimator is

$$\begin{aligned}\mathrm{Var}\left[\hat{\sigma}^2\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{\sigma^4}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right]\\ &= \frac{\sigma^4}{n^2}\mathrm{Var}\left[\chi^2_{n-1}\right] = \frac{\sigma^4}{n^2}2(n-1) = \frac{2(n-1)}{n^2}\sigma^4\end{aligned} \tag{5}$$

The estimator is consistent in that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[\hat{\sigma^2}\right] &= \lim_{n\to\infty}\frac{n-1}{n}\sigma^2 = \sigma^2\\ \lim_{n\to\infty}\mathrm{Var}\left[\hat{\sigma^2}\right] &= \lim_{n\to\infty}\frac{2(n-1)}{n^2}\sigma^4 = 0\end{aligned} \tag{6}$$

so $\hat{\sigma}^2 \to \sigma^2$ as $n \to \infty$, as both the bias and variance go to zero as $n \to \infty$.

4. **Empirical experimentation (simple programming assignment) (10 points)**
In this question, we are going to numerically explore the MLE (maximum likelihood estimator) of the variance parameter of the Gaussian, with the formula that was given in Question 3(d) above.

   (a) Draw $n = 5$ samples from the standard Gaussian distribution, $\mathcal{N}(0, 1)$.

   (b) Using the samples as data, compute the ML estimate $\hat{\mu}$ for the mean and $\hat{\sigma}^2$ for the variance of the Gaussian, as given in Question 3(d) above.

   (c) Repeat steps (a) and (b) 10,000 times. Plot a histogram of the 10,000 estimates of the Gaussian variance parameter to show its empirical distribution. Do you recognize its shape?

(d) Use these 10,000 repeated trials to numerically estimate the (frequentist) bias and variance of the ML estimate $\hat{\sigma}^2$ of the Gaussian variance parameter.

(e) Compare the results of (d) with the theoretical (frequentist) bias and variance that you can compute from the formula you derived in Question 3(d). (Hint: if your numerical estimates are very far from the theoretical formula, you made a mistake somewhere!)

Solution:

See `q4.py` – you can run it using Anaconda Python for example, a very popular distribution of python packages which contains everything to do scientific python. You can download it here.