IFT6269-A2024	**Hwk 1**	Name:
Prof: Simon Lacoste-Julien	Due date: Sept 26, 2024	Student id:

For each question, provide your derivations and not just the answer.

1. **Probability and independence (10 points)** Prove or disprove 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 $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$

   Hint: If you are convinced a statement is false, come up with a ***concrete and simple*** *counterexample for which the statement is not true.*

2. **Bayesian inference and MAP (10 points)**
   Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\pi})$ on k elements. The encoding for a possible value $\boldsymbol{x}_i$ of the random vector $\boldsymbol{X}_i$ can take is $\boldsymbol{x}_i = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$ with $x_j^{(i)} \in \{0, 1\}$ and $\sum_{j'=1}^k x_{j'}^{(i)} = 1$. In other lingo, $\boldsymbol{X}_i$ is a k-dimensional one-hot vector.

   Consider a Dirichlet prior distribution on $\boldsymbol{\pi}$: $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ and $\alpha_j > 0$ for all j. The Dirichlet distribution describes a *continuous* random vector $\boldsymbol{\pi}$ which lies on the probability simplex $\Delta_k := \{\boldsymbol{\pi} \in \mathbb{R}^k : 0 \leq \pi_j \leq 1 \text{ and } \sum_{j=1}^k \pi_j = 1\}$.

   Its probability density function[1] is $p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \pi_j^{\alpha_j - 1}$. Just like the Binomial distribution is the special case of a Multinomial distribution with $k = 2$, the Beta distribution is the 2-dimensional instantiation of a Dirichlet distribution.

   (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 *one line*, in the form of formal conditional independence statements (like $X_1 \perp X_2 \mid X_3$). Ignore α for this exercise.

   (b) Derive the posterior distribution $p(\boldsymbol{\pi} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The expected answer has the form: *"The posterior is a ____ distribution with parameters ____".*

   (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 in the regime of extremely large k.[2]

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

   (a) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. The pmf for a Poisson r.v. is $p(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ for $x \in \mathbb{N}$. Find the MLE for λ and derive its bias, variance and consistency (Y/N).

   (b) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ and suppose that $n > 10$. Consider $\hat{p} := \frac{1}{10} \sum_{i=1}^{10} X_i$ as an estimator of p. Derive its bias, variance and consistency (Y/N).

(c) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$. Find the MLE for θ and derive its bias, variance and consistency (Y/N).

Hint: For each $c \in \mathbb{R}$, $P(\max\{X_1, \ldots, X_n\} < c) = P(X_1 < c, X_2 < c, \ldots, X_n < c) = P(X_1 < c)P(X_2 < c) \cdots P(X_n < c)$.

(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. Let $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$. Show that the MLE[3] for $\theta := (\mu, \sigma^2)$ is $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Derive the bias, variance and consistency (Y/N) *only* for $\hat{\sigma}^2$.

Hint: Let χ^2_{n-1} *be the chi-squared distribution with* $(n-1)$ *degrees of freedom. When calculating the variance of* $\hat{\sigma}^2$*, you may use the fact that* $Var[\chi^2_{n-1}] = 2(n-1)$*, and that* $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 \overset{d}{=} \chi^2_{n-1}$.

4. **Maximum Likelihood Estimation (10 points)**

Follow the instructions in this Colab notebook: https://colab.research.google.com/drive/1zPg4qksvcOlLhWRfjF0xkXnFoc8WWkiL?usp=sharing

Notes

[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.

[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.