IFT6269-A2016	**Hwk 1**	Name:
Prof: Simon Lacoste-Julien	Due date: Sept 27, 2016	Student id:

[Updated version of Sept 16th, 2016].
For each question, give your derivations (not just the answer).

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

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 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 $\boldsymbol{\pi}$: $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, 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 properties 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]

[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. **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 $\text{Var}[\chi^2_{n-1}] = 2(n-1)$.)

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!)

[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.