CS 109 Quiz 7 (30 points):

1. [12 points] Definitions (4 points each). Choose **ONLY THREE** out of the following to do.

a. For random variables X, Y, $Cov(X, Y) =$

b. Events A and B are conditionally independent given C: $P(A, B|C) =$

c. If X, Y are continuous, $E[X|Y = y] =$

d. Suppose Y is a discrete rv that somehow influences X. Give a formula for $E[X]$ which uses the law of total expectation conditioning on Y. $E[X] =$

e. If X is discrete and Y is continuous, then using Bayes Theorem (watch your p's and f's), $p_{X|Y}(x|y) =$

f. If X, Y are discrete, then in terms of $p_{X,Y}$, $E[g(X, Y)] =$

2. [18 points] Short Answer. Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)$ are iid samples from the $Geo(\theta)$ distribution with pmf:

$$p_X(x; \theta) = (1 - \theta)^{x-1}\theta$$

a. What is the likelihood, $L(\boldsymbol{x}|\theta) = L(x_1, \ldots, x_n|\theta)$? (Use a product symbol)

b. What is the log-likelihood $LL(\boldsymbol{x}|\theta) = \log L(\boldsymbol{x}|\theta)$? (Use a sum symbol)

c. What is the gradient of the log-likelihood with respect to θ, $\nabla_\theta LL(\boldsymbol{x}|\theta)$? (Here since θ is a single parameter and not a vector, the gradient is simply a one-dimensional derivative)

d. If we were to perform gradient ascent on the log-likelihood with step size α and at time t our guess for θ is θ_t, what is the update rule? $\theta_{t+1} \leftarrow$

e. Find a closed form for the maximum likelihood estimate, $\hat{\theta}$, of θ.

1. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations with common pmf $p_X(x; \theta)$, where θ is an unknown but **fixed** parameter. You may wonder why in MLE, we maximize $L(\boldsymbol{x}|\theta)$, rather than $P(\theta|\boldsymbol{x})$. This is because $P(\theta|\boldsymbol{x})$ doesn't make sense unless θ is a random variable. In MAP estimation, we treat the parameter as a random variable (denoted Θ), and attempt to maximize $\pi_\Theta(\theta|x)$, where π_Θ is either the PMF of Θ (if Θ is discrete), or PDF of Θ (if Θ is continuous). By Bayes Theorem, $\pi_\Theta(\theta|x) = \frac{L(x|\theta)\pi_\Theta(\theta)}{L(x)} \propto L(x|\theta)\pi_\Theta(\theta)$ (since we are optimizing over θ, we don't care about the denominator which is a positive constant). We call $\pi_\Theta(\theta)$ the **prior** for Θ and $\pi_\Theta(\theta|x)$ the **posterior** for Θ. Hence MAP maximizes the product of likelihood and prior, where MLE maximizes just the likelihood. If the prior is uniform, we can see that MAP and MLE give the same answer.

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

b. Suppose we impose $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

d. Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

e. Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = K\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$, where K is some normalizing constant. The mode of a $W \sim Beta(\alpha, \beta)$ random variable is $\frac{\alpha-1}{\alpha+\beta-2}$ (the mode is the value with highest density $\arg\max_{w} f_W(w)$). Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP?

i. Which do you think is "better", MLE or MAP?