Instructions and Guidelines for Homeworks

- Please answer all of the questions and submit a scanned copy of your written solutions to Gradescope (either hand-written or typed are fine as long as the writing is legible).

- All problems are worth 10 points unless otherwise stated. All homeworks will get equal weight in computation of the final grade for the class.

- The homeworks are intended to help you work through the concepts we discuss in class in more detail. It is important that you try to solve the problems yourself. The homework problems are important to help you better learn and reinforce the material from class. If you don’t do the homeworks you will likely have difficulty in the exams later in the quarter.

- If you can’t solve a problem, you can discuss it verbally with another student. However, please note that before you submit your homework solutions you are not allowed to view (or show to any other student) any written material directly related to the homeworks, including other students’ solutions or drafts of solutions, solutions from previous versions of this class, and so forth. The work you hand in should be your own original work.

- You are allowed to use reference materials in your solutions, such as class notes, textbooks, other reference material (e.g., from the Web), or solutions to other problems in the homework. It is strongly recommended that you first try to solve the problem yourself, without resorting to looking up solutions elsewhere. If you base your solution on material that we did not discuss in class, or is not in the class notes, then you need to clearly provide a reference, e.g., “based on material in Section 2.2 in .....”

- In problems that ask for a proof you should submit a complete mathematical proof (i.e., each line must follow logically from the preceding one, without “hand-waving”). Be as clear as possible in explaining your notation and in stating your reasoning as you go from line to line.

- If you wish to use LaTeX to write up your solutions you may find it useful to use the .tex file for this homework that is posted on the Web page.

Problem 1: Properties of the Beta Density

In class we will discuss the use of a Beta density function as a prior density for a parameter that lies between 0 and 1. The Beta density is defined as:

\[ P(\theta) = Be(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \] (1)

where \( 0 \leq \theta \leq 1 \). The two parameters of this density function are \( \alpha > 0 \) and \( \beta > 0 \) and \( B(\alpha, \beta) \) is a normalization constant to ensure that the density integrates to 1, where \( B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \) and \( \Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx, \ z > 0 \), is the gamma function.

In solving the problems below keep in mind that since \( B(\alpha, \beta) \) is the normalization constant for the density, then by definition \( B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \). Another useful fact that the gamma function has the property that \( \Gamma(z+1) = z\Gamma(z) \).

1. Derive an expression for the mean of a Beta density, as a function of the parameters \( \alpha \) and \( \beta \).

2. Derive an expression for the mode of a Beta density, as a function of the parameters \( \alpha \) and \( \beta \). You can assume in this case that \( \alpha > 1 \) and \( \beta > 1 \).

Problem 2: (Bayesian Estimation for the Multinomial Model)

Consider a data set \( D = \{x_1, \ldots, x_N\} \), with \( x_i \in \{1, \ldots, M\} \) where the \( x_i \)'s are independent draws from a discrete-valued distribution with parameters \( \theta_k = P(x_i = k), 1 \leq k \leq M \), and \( \sum_{k=1}^M \theta_k = 1 \) (i.e., we have a multinomial likelihood for the \( x_i \)'s). Assume that we have a Dirichlet prior for the parameters \( \theta \), where the prior has parameters \( \alpha_1, \ldots, \alpha_M \) and \( \alpha_k > 0 \) and \( 1 \leq k \leq M \).

1. Prove that the posterior distribution on \( \theta_1, \ldots, \theta_K \) is also Dirichlet.

2. Derive an expression for the maximum a posteriori (MAP) estimate for \( \theta_k, 1 \leq k \leq M \). Your solution should be derived from first principles (i.e., using basic calculus to find the mode of the posterior density for \( \theta_k \), working from your solution for \( P(\theta|D) \) from part 1).

Problem 3: (Bayesian Estimation of a Gaussian Model)

For the case of the mean \( \mu \) of a Gaussian model, with known variance \( \sigma^2 \), and with a Gaussian prior on \( \mu \) that has mean \( \mu_0 \) and variance \( s^2 \), we discussed in class the fact that the posterior density for \( \mu \), given \( n \) IID observations \( \{x_1, \ldots, x_n\} \) is also Gaussian with parameters \( \mu_n \) and \( \sigma_n^2 \). Prove that the following expressions for the mean of this posterior and the variance of this posterior are correct, i.e.,

\[ \mu_n = \gamma \hat{\mu}_{ML} + (1 - \gamma)\mu_0 \]
where
\[ \gamma = \frac{ns^2}{n^2 + \sigma^2}. \]
and
\[ \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{s^2}. \]

**Problem 4: (Bayesian Estimation for the Geometric Model)**

Let \( X \) be a geometric random variable taking values \( x \in \{0, 1, 2, \ldots\} \), with probability distribution defined as
\[
P(x) = (1 - \theta)^x \theta, \quad \text{for } x = 0, 1, 2, \ldots
\] (2)
where \( \theta \) is the parameter of the geometric model and \( 0 < \theta < 1 \). The geometric distribution models a problem where we have a sequence of random binary outcomes with “success” probability \( \theta \), where the outcomes are generated independently at each step, and \( x \) is the number of steps before success. For example this could be a model of the number of tails we see in tossing a coin before we see heads, where \( \theta \) is the probability of heads.

Let \( D = \{x_1, \ldots, x_N\} \) be an observed data set where we assume the samples were generated in an IID manner from \( P(x) \).

1. Define the likelihood function for this problem.

2. Prove that the Beta density \( \text{Be}(\alpha, \beta) \) is a conjugate prior for the Geometric likelihood and derive an equation for the posterior density for \( \theta \). (A prior is said to be *conjugate* to a particular type of likelihood function whenever the posterior has the same functional form (the same type of density) as the prior).

**Problem 5: Bayesian Estimation for the Poisson Model**

Consider a random variable \( X \) taking values \( x \in \{0, 1, 2, \ldots\} \), with a Poisson distribution with parameter \( \lambda > 0 \), i.e., \( P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \). Let \( P(\lambda|\alpha, \beta) \) be a Gamma prior for \( \lambda \) defined as
\[
P(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta \lambda}, \quad \alpha, \beta > 0
\]
where \( \Gamma(\alpha) \) is the gamma function we discussed in Problem 1.

1. Consider a data set consisting of a single observation \( x \). Show clearly that the posterior distribution for \( \lambda \), \( P(\lambda|x, \alpha, \beta) \), is also a Gamma distribution and express the parameters of this posterior Gamma distribution as a function of \( \alpha, \beta, \) and \( x \).
2. Now consider the case of a data set $D = \{x_1, \ldots, x_N\}, N \geq 2$. Show that the posterior distribution for $\lambda, P(\lambda|D, \alpha, \beta)$ is also a Gamma distribution and express the parameters of this posterior Gamma distribution as a function of $\alpha, \beta$, and the data $D$.

3. What are the posterior mean and the posterior mode for part 2 above? (you don’t need to derive them: you can look up what the properties of a Gamma distribution are to answer this).

Problem 6: Different Types of Point Estimates of Parameters

In class we discussed the “fully Bayesian” approach to parameter estimation, where we report the full posterior density $p(\theta|D)$. Assume that $D$ is a data set with $N$ IID observations. We also discussed reporting a point estimate instead of the full posterior density, both the expected or mean value of the posterior density $E_{p(\theta|D)}[\theta]$ (also known as $\theta_{MPE}$), and the maximum a posterior value $\theta_{MAP} = \arg \max p(\theta|D)$.

To answer the questions below you can use specific examples of functional forms for $p(\theta)$ and $p(D|\theta)$, either in equation form or with plots (hand-sketched is fine):

1. Describe a situation where $\theta_{MPE}$ would be a better summary in some respect than $\theta_{MAP}$.
2. Describe a situation where $\theta_{MAP}$ would be a better summary in some respect than $\theta_{MPE}$.
3. Describe a situation where neither is really adequate.

Problem 7: (Computer Simulations for the Dirichlet-Multinomial Model)

1. Assume we have a random variable $X$ taking values $x \in \{1, \ldots, K\}$ with parameters $\theta_k = p(x = k)$ with $\sum_{k=1}^{K} \theta_k = 1$. Say we want to simulate a data set $D = \{x_1, \ldots, x_N\}$ with $N$ IID samples from $X$. Assume you have access to a pseudorandom number generator (e.g., in your favorite programming language) called $\text{rand()}$ and when $\text{rand()}$ is called it returns a pseudorandom number from the uniform distribution $U[0, 1]$. Explain clearly how you can use this $\text{rand()}$ function to simulate a data set with $N$ IID samples of $X$ given multinomial parameters $\theta_1, \ldots, \theta_K$. In addition, provide a few lines of pseudocode explaining how you would write code to do this, e.g., part of it can consist of a for-loop over simulated data points from $i = 1$ to $i = N$.

2. Write a function (in whatever language you prefer, e.g., Matlab, Python, R, etc) that takes as input the parameters of a Dirichlet distribution with $K = 3$, i.e., $\text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ and generates a contour plot of the Dirichlet density function (with these parameter values) over the 2-dimensional simplex where the simplex has $\theta_1$ as the x-axis and $\theta_2$ as the y-axis. You do not need to write your own function for generating contours: you should be able to find a plotting function to do this in whatever language you are using. Using your code generate plots that show:

(a) A prior defined as $\text{Dir}(4, 4, 4)$
(b) The posterior density with $Dir(4, 4, 4)$ as the prior and with $N = 20$ data points simulated from a multinomial with $\theta_1 = 0.1, \theta_2 = 0.5, \theta_3 = 0.4$.

(c) The posterior density with the same prior and data from the same multinomial but now with $N = 200$ data points.

Use your solution from part 1 to write code to simulate the data points. Submit the 3 plots above. No need to submit your code.

3. Now repeat the previous part of this question but with a prior defined as $Dir(50, 10, 40)$. Submit the 3 plots again and write a few sentences comparing the two sets of 3 plots (with the two different priors).