1. What is a parameter? What is a random variable?

2. What is a prior distribution? What is a posterior distribution? What does it mean for a prior distribution to be a conjugate prior?

3. Suppose you have a coin. Let $X$ be the random variable indicating whether a flip of the coin is heads (1) or tails (0). You flip the coin a number of times and obtain the following outcomes: $D = (H, T, H, H, T, H, T, T, H)$. If we assume that the coin is modeled using a Bernoulli distribution

$$p(X = k) = \begin{cases} \theta & \text{if } k = 1 \\ 1 - \theta & \text{otherwise,} \end{cases}$$

what is the maximum likelihood estimate for $\theta$?

4. Suppose we impose a Beta distribution

$$p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

(with $\alpha, \beta > 1$) as a prior distribution over the parameter $\theta$. What is the posterior distribution?

5. We want to obtain an estimate for $\theta$ consistent with this prior. Specify two different mathematical approaches for obtaining an estimate for $\theta$. Do they both give you the same results?

6. PLSA has a log-likelihood function

$$\log p(W \mid \theta, \phi) = \sum_{d=1}^{M} \sum_{i=1}^{[wd]} \log \left( \sum_{k=1}^{K} p(z_{d,i} = k \mid \theta_d)p(w_{d,i} \mid \phi_k) \right).$$

What are the parameters of the model? What are the random variables? How did you know?
7. Briefly explain why PLSA is not considered a “fully generative” model. Why does this matter, and how could we fix it?

8. Draw your modified model proposal using plate notation.

9. Suppose you have a document $w_d$. What is the posterior distribution of the latent variables for document $w_d$? Briefly describe why computing this distribution is intractable.

10. Variational inference uses a simpler, factorized distribution as an approximation to the posterior distribution of interest. In mean field variational inference, that distribution is chosen to be fully factorized. Draw a fully factorized variational distribution for LDA using plate notation and write its mathematical form as the distribution $q$. What are its parameters? What are its latent variables?

11. We want to find the setting of the parameters for this simpler distribution $q$ that best approximate $p$. One way of quantifying the quality of this approximation is by using the KL-divergence. Write down the formula for $D(q \parallel p)$, the KL-divergence from $p$ to $q$. 


12. Write down the updating equations for the coordinate ascent algorithm to minimize the above KL-divergence equation.

13. Describe how to update the parameters $\phi_k$ after running the optimization algorithm for each document in the corpus.

14. *(shifting gears.)* The Dirichlet distribution is defined as

$$p(\theta \mid \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta^{\alpha_i - 1}$$

where $B(\alpha)$ is the multivariate beta function. Show that the Dirichlet distribution is the conjugate prior of the multinomial.

15. Prove the following:

$$\int \prod_{i=1}^{K} \theta^{n_i + \alpha_i - 1} d\theta = B(\alpha + n)$$
16. Using the fact that the Dirichlet is conjugate to the multinomial, perform the following integration to “collapse” out the prior over $\Theta$ in LDA:

$$ P(Z | \alpha) = \int P(\Theta | \alpha)P(Z | \Theta)d\Theta $$

17. Now, collapse the prior over $\Phi$ in a similar way:

$$ P(W | Z, \beta) = \int P(\Phi | \beta)P(W | Z, \Phi)d\Phi $$

18. Draw the collapsed model using plate notation, and write down its joint distribution as an equation.