1. What is the posterior distribution over the latent variables for a document $w_d$? Briefly describe why computing this distribution is intractable.

2. 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?

3. 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 $q$.

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

5. Describe how to update the parameters $\phi_k$ after running the optimization algorithm for each document in the corpus.
6. Briefly describe how you could use parallel processing to speed up the inference algorithm.

7. (shifting gears.) What are Monte Carlo methods? Briefly describe a Monte Carlo method for computing $\pi$.

8. What is a Markov chain? What does it mean for a Markov chain to have a stationary distribution?

9. What is a Markov chain Monte Carlo (MCMC) algorithm? Name two different MCMC algorithms, and briefly describe the difference between them.

10. 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.

11. Prove the following:

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

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

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

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

14. Draw the collapsed model using plate notation, and write down its joint distribution as an equation.
15. Explain why the definition of the joint distribution above is sufficient for computing the sampling probability $P(z_{d,i} = k \mid Z_{-d,i}, W, \alpha, \beta)$ needed at each step of the Gibbs sampler.

16. Write down the sampling probability $P(z_{d,i} = k \mid Z_{-d,i}, W, \alpha, \beta)$ used in each step of the collapsed Gibbs sampler. Can you intuitively describe the two parts of the equation?

17. What must be stored in memory to run the collapsed Gibbs sampler?

18. Describe how to infer the marginalized variables $\Theta$ and $\Phi$ from the current state of the Markov chain.