Problem 1 (30 marks)

This problem is meant to test your skills in understanding, distilling, and presenting in your own words, complex technical ideas from research papers.

Read the paper "Black Box Variational Inference" by Ranganath et al (2014) and write a summary ($\sim$ 500 words) highlighting the key aspects of this paper. The paper proposes several strategies to reduce the variance of the ELBO's gradients. Your summary must contain a discussion about these strategies. Also discuss any other specific points that you like/dislike about the paper and the methods proposed therein. Feel free to discuss the paper with your classmates but your writeup must be entirely in your own words.

Problem 2 (10 marks)

Consider the Latent Dirichlet Allocation (LDA) model for documents. Suppose that, in addition to the D documents, you are also provided a document-document binary link matrix $\mathbf{A}$ of size $D \times D$ such that if document d is "linked" to document d' (e.g., if one mentions the other) then $A_{dd'} = 1$, and $A_{dd'} = 0$ otherwise (note that $\mathbf{A}$ is symmetric).

Suggest a generative story for each element (a binary observation) of the $\mathbf{A}$ matrix. Specifically, I am looking for a way to generate $\mathbf{A}$ using the latent variable(s) that are part of the original LDA model we saw in the class. In addition to these latent variables, you may introduce additional parameters if you think it would make the model even better. You only need to give the generative model for A and don't need to give the inference algorithm.

Problem 3 (20 marks)

Consider the stochastic blockmodel for a network represented as an $N \times N$ binary adjacency matrix $\mathbf{A}$. Suppose there are a total of K communities and $\boldsymbol{z}_n \in \{1, \ldots, K\}$ denotes the community membership of node n with $p(\boldsymbol{z}_n) = \text{multinoulli}(\pi)$, and $p(\pi) = \text{Dirichlet}(\alpha, \ldots, \alpha)$. Further assume $p(A_{nm}|\boldsymbol{z}_n, \boldsymbol{z}_m, \eta) = \text{Bernoulli}(A_{nm}|\eta_{\boldsymbol{z}_n, \boldsymbol{z}_m})$, where η is a $K \times K$ matrix with each entry $\eta_{k\ell} \in (0, 1)$ being the probability of a link for two nodes belonging in community k and community ℓ, respectively. Assume a beta prior on $\eta_{k\ell}$: $p(\eta_{k\ell}) = \text{Beta}(a, b)$.

Derive a Gibbs sampler for the model. In particular, you need to give the expressions for the conditional posteriors for π, $\eta_{k\ell}$, and $\boldsymbol{z}_n$. You may assume the network to be symmetric (i.e., $A_{nm} = A_{mn}$).

Problem 4 (10 marks)

Using the fact that if $G \sim \text{DP}(\alpha, G_0)$ then the finite dimensional marginal of G is also Dirichlet distributed

$$[G(A_1), G(A_2), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_K))$$

where $A_1, A_2, \ldots, A_K$ is a finite partition of the space Ω over which G_0 is defined, show that given N i.i.d. draws $\theta_1, \theta_2, \ldots, \theta_N$ from the discrete distribution G, the posterior of G is

$$G|\theta_1, \theta_2, \ldots, \theta_N \sim \text{DP}(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\theta_n})$$

.. or equivalently, $G|\theta_1, \theta_2, \ldots, \theta_N \sim \text{DP}(\alpha + N, \frac{\alpha}{\alpha+N} G_0 + \sum_{k=1}^{K} \frac{n_k}{\alpha+N} \delta_{\phi_k})$, where $n_k = \sum_{n=1}^{N} \delta_{\theta_n}(A_k)$. Here $\delta_{\theta_n}(.)$ denotes the Dirac function at θ_n such that $\delta_{\theta_n}(A_k) = 1$ if $\theta_n \in A_k$, and 0 otherwise.

Problem 5 (30 marks)

Just like Problem 1, this problem is meant to test your skills in understanding, distilling, and presenting in your own words, complex technical ideas from research papers.

In particular, your task is to read the paper "Hierarchical Dirichlet Process" (HDP) by Teh et al (2005) upto Section 4 (you aren't required to read Section 5 onwards, which contains details of inference and experiments, to answer this question). As we discussed briefly in the class, HDP is a method to do *joint* mixture modeling of multiple related datasets (as opposed to DP which is used to perform mixture modeling for a single dataset). Joint here means that the clusters can be shared across the multiple datasets.

- Summarize your understanding of HDP in about 200 words, delineating its key differences from Dirichlet Process (DP) which we saw in class (Section 3 of the paper actually contains a fairly good overview of DP). In your summary, focus more on intuition and less on technical aspects. However, you may use some of the relevant equations in your summary.
- Describe the stick-breaking construction for HDP. How is it different from the stick-breaking construction of DP? Again, your answer should focus more on the intuition behind the construction and less on the equations (which are anyway given in the paper), although you may use some of the equations to explain the ideas.
- Describe the Chinese Restaurant Franchise (CRF). How is it differnt from the Chinese Restaurant Process (CRP) metaphor of DP, and why does it make intuitive sense as a prior for jointly doing mixture modeling of multiple related datasets.

You may discuss the paper with your classmates but your writeup must be in your own words.