CS698X: Topics in Prob. Modeling and Inference *Your Name & Roll no:*

End-sem Exam (2017-2018, Sem-II)

Instructions:

- This question paper is worth a total of 100 marks.
- Total duration is 3 hours.
- Please write your name and roll number at the top of this sheet as well as on the provided answer copy.
- The **true-false** questions have to be answered on this question sheet itself, in the space provided. Please also submit the question sheet.
- To answer the other parts of the question paper, please use the provided answer copy.
- For rough work, please use pages at the back of answer copy. Extra sheets can be provided if needed.

Some distributions and their properties:

- For $x \in (0,1)$, $\text{Beta}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and Γ denotes the gamma function s.t. $\Gamma(x) = (x-1)!$ for a positive integer x. Expectation of a Beta r.v.: $\mathbb{E}[x] = \frac{a}{a+b}$.
- For $x \in \{0,1,2,\ldots\}$ (non-negative integers), $\text{Poisson}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ where λ is the rate parameter.
- For $x \in \mathbb{R}_+$, $\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ (shape and rate parameterization)
- For $x \in \mathbb{R}$, Univariate Gaussian: $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$
- For $x \in \mathbb{R}^D$, D-dimensional Gaussian: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$.
  Trace-based representation: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\Big\}$, $\mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top$.
- For $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$, $\text{Dirichlet}(\boldsymbol{\pi}|\alpha_1,\ldots,\alpha_K) = \frac{1}{B(\alpha_1,\ldots,\alpha_K)} \prod_{k=1}^K \pi_k^{\alpha_k-1}$ where $B(\alpha_1,\ldots,\alpha_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$, and $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}$
- For $x_k \in \{0,N\}$ and $\sum_{k=1}^K x_k = N$, $\text{multinomial}(x_1,\ldots,x_K|N,\boldsymbol{\pi}) = \frac{N!}{\boldsymbol{x}_1!\ldots,x_K!}\pi_1^{x_1}\ldots\pi_K^{x_K}$, where $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$. The multinoulli is the same as multinomial with $N = 1$.

Some other useful results:

- If $\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon$, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$, $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\mathbf{L}^{-1})$ then $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z}+\boldsymbol{b},\mathbf{L}^{-1})$, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu}+\boldsymbol{b},\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1})$, and $p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top\mathbf{L}(\boldsymbol{x}-\boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\},\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}$.
- Marginal and conditional distributions for Gaussians: $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_{aa})$, $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b},\boldsymbol{\Sigma}_{a|b})$ where $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$, $\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\right\} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$, where symbols have their usual meaning. :)
- $\frac{\partial}{\partial\boldsymbol{\mu}}[\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}] = [\mathbf{A}+\mathbf{A}^\top]\boldsymbol{\mu}$, $\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = \mathbf{A}^{-\top}$, $\frac{\partial}{\partial\mathbf{A}}\text{trace}[\mathbf{A}\mathbf{B}] = \mathbf{B}^\top$
- For a random variable vector $\boldsymbol{x}$, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top + \text{cov}[\boldsymbol{x}]$

Questions begin from the next page

Section 1 - true/false (10 questions, 10 marks)

1. [**F**] The EM algorithm is prone to overfitting since it does not provide a way to incorporate any regularization for the parameters being estimated.
2. [**T**] Unlike Bayesian linear regression with fixed hyperparams, Bayesian linear regression with unknown hyperparams must use an iterative method for doing inference.
3. [**F**] For any model, the posterior predictive distribution is the weighted average of the point predictive distributions computed at infinite many possible values of the parameters. (If the parameter is discrete-valued then we will be averaging the point predictives computed at its finite many values).
4. [**F**] For GP regression, although the posterior predictive for a single test point is available in closed form (Gaussian), the joint posterior predictive for multiple test points doesn't have a closed form.
5. [**T**] Storing the VB approximation of a posterior distribution requires less space as compared to its MCMC approximation.
6. [**F**] Metropolis-Hastings sampling with a Gaussian proposal is equivalent to Gibbs sampling.
7. [**T**] The number of parameters to be learned when doing inference in a variational auto-encoder (VAE) model is independent of the total number of observations.
8. [**F**] A hidden Markov model (HMM) can only model discrete-valued observations
9. [**T**] The standard SGLD algorithm cannot be applied to infer all the parameters of a GMM. (GMM has some parameters that are constrained, e.g., mixing proportions. SGLD can't be used for these)
10. [**F**] The computational-cost per iteration of the stochastic variational inference (SVI) algorithm scales in the total number of observations.

Section 2 (10 questions, 30 marks)

1. Suppose we have N observations drawn i.i.d. as $x_n \sim \mathcal{N}(0, \sigma^2)$, $n = 1, \ldots, N$, and an inverse-gamma prior $\text{IG}(\sigma^2|a, b)$ on the variance σ^2. The inverse-gamma distribution over some random variable z has the form $p(z|a, b) = \text{IG}(z|a, b) = \frac{b^a}{\Gamma(a)} z^{-a-1} \exp\left(-\frac{\beta}{z}\right)$ where $a, b > 0$ are the shape and scale parameters, respectively. Compute the posterior of σ^2.

   Answer: It is a simple model with conjugacy. The posterior will be $\text{IG}(\sigma^2|a + \frac{N}{2}, b + \frac{1}{2}\sum_{n=1}^{N} x_n^2)$
2. Suppose you have a model with unknowns θ that in turn depend on some "hyperparameters" λ. Briefly describe how would you use EM to find the point estimates of the hyperparameters? Can you use also EM to find the posterior distribution of the hyperparameters? If yes, why? If no, why not?

   Answer: If we want point estimate of the hyperparameters λ, we can do so in the M step by maximizing the expected CLL w.r.t. the hyperparams (and will estimate the posterior of θ in the E step). If we want the posterior of the hyperparameters, we can do it in the E step (and will find the point estimate of θ in the M step).
3. In a probabilistic model with both local and global variables, why might it be okay to only compute point estimates of the global variables, as opposed to their full posterior?

   Answer: Since we usually have plenty of data to estimate the global parameters, there is very little uncertainty in our estimate of the global parameters. Therefore a point estimate is usually okay for global parameters.
4. Briefly describe what collapsing means in the context of Bayesian inference. Why is collapsing useful?

   Answer: Collapsing refers to integrating out some of the unknowns from the model when deriving the inference update equations, so they are no longer required to be estimated during inference. We may

want to do so if we don't care about some of the unknowns (even if they are desirable/necessary while defining the model). Collapsing is often possible if we have a conjugate model. Collapsing is useful since it reduces the number of parameters to be estimated.

5. Consider Bayesian linear regression with a Laplace prior on each component of the weight vector, i.e., $p(w_d) = \frac{\gamma}{2}\exp(-\gamma|w_d|)$. The Laplace prior can also be written as a Gaussian scale mixture prior, i.e., $p(w_d) = \int \mathcal{N}(w_d|0,\tau_d^2)\text{Gamma}(\tau_d^2|1,\frac{\gamma^2}{2})d\tau_d^2$. What is/are the advantage(s) of the latter representation?

   Answer: The Laplace distribution is not conjugate to a Gaussian, so inference would not be straightforward. However, expressing the Laplace as$p(w_d) = \int \mathcal{N}(w_d|0,\tau_d^2)\text{Gamma}(\tau_d^2|1,\frac{\gamma^2}{2})d\tau_d^2$ enables us to use an iterative algorithm such as EM to infer w_d and τ_d in an alternating fashion. In fact, if we can also do Gibbs sampling since the conditional posteriors of w_d given τ_d and vice-versa are available in closed form due to Gaussian-Gaussian and Gaussian-gamma conjugacy.

6. What is the difference between variational EM and standard EM? When would you need to use variational EM as opposed to standard EM?

   Answer: Variational EM is basically EM but the E step uses variational inference. Variational EM is needed instead of standard EM if the posterior over the latent variables to be estimated in the E step is not tractable and requires approximation. So we approximate it using a variational distribution (e.g., mean-field). Both algorithms are identical as far as the M step is concerned.

7. Suppose you are given N paired observations of the form $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ and you would like to learn a mixture model for $\{y_1,\ldots,y_N\}$ while also incorporating the information from $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$. Suggest a model to do this. How does such a model differ from a conventional mixture model that is learned only using $\{y_1,\ldots,y_N\}$? Is learning this model easier or more difficult as compared to learning the conventional mixture model? Briefly explain your answer.

   Answer: We can use a mixture of experts (MoE) model. The key difference is that the component distributions and the mixing proportions depend on (i.e., *conditioned* on) $\boldsymbol{x}_n$. Learning the MoE is usually more difficult since the conditioning on $\boldsymbol{x}_n$ introduces additional parameters, estimating which requires additional effort (e.g., the gating network which defines the mixing proportions in MoE usually requires learning a softmax model).

8. Consider a mixture model with K clusters with each cluster id $\boldsymbol{z}_n$ having a K dimensional multinomial prior. Suppose we would like to learn another model with equal expressive power but using only $\log_2 K$ clusters. Suggest such a model along with the appropriate the prior on $\boldsymbol{z}_n$.

   Answer: We can learn a model that defines $\boldsymbol{z}_n$ as a binary vector $\{0,1\}^K$ instead of a one-hot vector. In the binary vector representation, a $\log_2 K$ length binary vector can model K clusters (each binary vector representation would correspond to one of the K clusters). We can assume a Bernoulli prior on each component of $\boldsymbol{z}_n$, i.e., $p(z_{nk}) = \text{Bernoulli}(z_{nk}|\pi_k)$ with $\pi_k \in (0,1)$. However, note that here $\sum_{k=1}^K \pi_k \neq 1$. Also note that this is the same prior that we have seen in the "subset sum" problems.

9. Both GP as well as Bayesian neural network can be used for learning a nonlinear regression model. Give 2 reasons each as to why you would prefer one over the other for nonlinear regression.

   Answer: GP regression has closed form inference and posterior predictive. It is also nonparametric (model complexity can grow with more training examples). Bayesian NN is faster during training and prediction (unlike GP for which training/test cost scales in the number of training examples). It is also good at extracting good features automatically (unlike GP for which we need to choose a kernel or learn a kernel which may be doable but can be somewhat expensive computationally).

10. For Bayesian Optimization, why is relying *solely* on the variance of the posterior of the function being optimized not a good idea to select the next query point?

    Answer: Because in BO, our goal is not only to learn the underlying function but also locate its optima. Therefore the criteria to quantify the utility of a potential query point must not only take into account the variance of the function at that point but also the extent/probability of improvement the point would give as compared to the previous best point (where the improvement refers to whether the new point is better than the previous best in being the optimum).

Section 3 (6 questions, 60 marks)

1. Consider the following hierarchical model: $p(x|\lambda) = \text{Poisson}(x|\lambda)$, $p(\lambda|r,\theta) = \text{Gamma}\left(\lambda|r, \frac{\theta}{1-\theta}\right)$. Assume r to be known, $p(\theta|a,b) = \text{Beta}(\theta|a,b)$, and a, b to be known. Derive the expression for the posterior $p(\theta|x)$. Is this posterior available in closed form?

   Answer: From Bayes rule, $p(\theta|x) \propto p(\theta)p(x|\theta)$. We are given that the prior $p(\theta) = \text{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$ and to compute the likelihood $p(x|\theta)$, we need to first integrate out λ as $p(x|\theta) = \int p(x|\lambda)p(\lambda|r,\theta)d\lambda$. You actually did it in one of the homeworks and $p(x|\theta)$ turns out to be negative-Binomial distribution, i.e., $p(x|\theta) = \text{NB}(x|r,\theta) \propto (1-\theta)^r\theta^x$. Easy to see that negative-Binomial and beta are conjugate, and the posterior is $\text{Beta}(\theta|a+x, b+r)$, and therefore available in closed form.

2. Consider a K component mixture model with prior on its mixing proportions $p(\pi) = \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$. Given the cluster ids $\mathbf{Z} = \{\boldsymbol{z}_n\}_{n=1}^N$ of the N observations $\{\boldsymbol{x}_n\}_{n=1}^N$ with each $\boldsymbol{z}_n$ assumed drawn i.i.d. from multinoulli(π), derive the predictive distribution $p(\boldsymbol{z}_{N+1}|\mathbf{Z})$ for the cluster id of a new observation $\boldsymbol{x}_{n+1}$. Specifically, write down the name of this distribution and the expression for its parameters.

   Now suppose we allow $K \to \infty$ (i.e., unbounded) for this model. What will be the distribution $p(\boldsymbol{z}_{N+1}|\mathbf{Z})$ in this case and what will be its parameters?

   Answer: Due to Dirichlet-multinoulli conjugacy, it is easy to see that the posterior $p(\pi|\mathbf{Z}) = \text{Dirichlet}(\frac{\alpha}{K} + n_1, \ldots, \frac{\alpha}{K} + n_K)$, where n_k denotes the number of points with $\boldsymbol{z}_n = k$. The predictive distribution can be computed as $p(\boldsymbol{z}_{N+1} = k|\mathbf{Z}) = \int p(\boldsymbol{z}_{N+1}|\pi)p(\pi|\mathbf{Z})d\pi = \int \pi_k p(\pi|\mathbf{Z})d\pi$, which is the expectation of a Dirichlet. Therefore $p(\boldsymbol{z}_{N+1} = k|\mathbf{Z}) = \frac{\alpha/K+n_k}{\sum_{k=1}^K \alpha/K+n_k} = \frac{\alpha/K+n_k}{\alpha+N}$. Therefore $p(\boldsymbol{z}_{N+1}|\mathbf{Z})$ is a K dimensional multinoulli with K dimensional parameter vector $\left(\frac{\alpha/K+n_1}{\alpha+N}, \ldots, \frac{\alpha/K+n_K}{\alpha+N}\right)$

   When $K \to \infty$, we can easily see that $p(\boldsymbol{z}_{N+1} = k|\mathbf{Z}) = \frac{n_k}{\alpha+N}$. Note that this doesn't sum to 1 when summed over k. The "left-over" probability $\frac{\alpha}{\alpha+N}$ is the probability that the new point will be assigned to a new cluster (call it cluster number $K+1$). This is basically the CRP mixture model. Therefore in the case with $K \to \infty$, $p(\boldsymbol{z}_{N+1}|\mathbf{Z})$ will be a $K+1$ dimensional multinoulli with $K+1$ dimensional parameter vector $\left(\frac{\alpha/K+n_1}{\alpha+N}, \ldots, \frac{\alpha/K+n_K}{\alpha+N}, \frac{\alpha}{\alpha+N}\right)$.

3. Consider a linear regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ with $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$ is the $N \times 1$ response vector, $\mathbf{X}$ is the $N \times D$ feature matrix, and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$ is the $N \times 1$ vector of i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$. Let us assume the following prior on each entry of the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$p(w_d|\sigma, \gamma_d) = \begin{cases} \mathcal{N}(0, \sigma^2 v_0), & \text{if } \gamma_d = 0 \\ \mathcal{N}(0, \sigma^2 v_1), & \text{if } \gamma_d = 1 \end{cases}$$

   where $v_1 \gg v_0 > 0$. Further assume the priors $p(\gamma_d) = \text{Bernoulli}(\theta)$, $d = 1, \ldots, D$, $p(\theta) = \text{Beta}(a_0, b_0)$, and $p(\sigma^2) = \text{IG}(\nu/2, \nu\lambda/2)$, where IG denotes the inverse-gamma prior in its shape-scale parameriza-tion. Let us denote $\boldsymbol{\gamma}$ to be the vector of indicator variables $\gamma_1, \ldots, \gamma_D$, and assume σ^2 to be known.

   Derive the expressions for the conditional posteriors for each of the unknowns $\boldsymbol{w}, \boldsymbol{\gamma}$ and θ.

   Answer: We looked at this model in one of the homework problems where you derived an EM algorithm. Here you are required to derive the conditional posteriors. Due to conjugacy, each of these are available in closed form. The Gaussian prior on $\boldsymbol{w}$ and the Gaussian likelihood of linear regression leads to a Gaussian conditional posterior on $\boldsymbol{w}$ (which you derived in the E step of the homework question). The posterior over each component of γ can be computed as $p(\gamma_d = 1|w_d) = p_d$ where $p_d = a_d/(a_d + b_d)$ with $a_d = p(\gamma_d = 1)p(w_d|\sigma, \gamma_d = 1) = \theta\mathcal{N}(0, \sigma^2 v_1)$ and $b_d = p(\gamma_d = 1)p(w_d|\sigma, \gamma_d = 0) = (1-\theta)\mathcal{N}(0, \sigma^2 v_0)$. Finally, the conditional posterior over θ is Beta due to the Bernoulli-Beta conjugacy and is given by $p(\theta|\gamma) = \text{Beta}(a_0 + \sum_{d=1}^D \gamma_d, b_0 + D - \sum_{d=1}^D \gamma_d)$.

4. Consider the standard Latent Dirichlet Allocation (LDA) topic model with K topics. Now suppose each document d also has an observed categorical label $y_d \in \{1, \ldots, C\}$ where C denotes the total number of classes. Suggest TWO ways to extend the standard LDA model to incorporate the document label

information. Write down all the steps in the generative stories for both of your proposed models (akin to the original LDA generative story), clearly specifying the priors on the unknowns of the model (expect for the fixed hyperparameters). Draw the plate notation as well. You do not need to give the inference algorithm for the proposed models.

Answer: One way would be to regress from the topic proportion vectors θ_d to the document label y_d (or from empirical topic proportions $\frac{1}{N_d}\sum_{n=1}^{N_d} \boldsymbol{z}_{n,d}$ to the document label y_d). It can be done via a softmax regression model. Another way would be to draw the topic proportion vector θ_d of each document from a mixture model rather than from the same Dirichlet$(\alpha, \ldots, \alpha)$ distribution. So instead of a single Dirichlet$(\alpha, \ldots, \alpha)$ distribution, we can have C Dirichlet distributions and for each document, we can choose the one depending on the document's label y_d. I am skipping the full generative model and plate notation diagrams.

5. 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., $A_{nm} = A_{mn}$).

   Answer: The conditional posterior for π is straightforward and due to Dirichlet-multinoulli conjugacy will be a Dirichlet (similar to a mixture model). Denoting $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, $p(\pi|\mathbf{Z}) = \text{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_K)$, where n_k denotes the number of nodes with $\boldsymbol{z}_n = k$.

   The conditional posterior of $\eta_{k\ell}$ will be $p(\eta_{k\ell}|\mathbf{A}) \propto p(\eta_{k\ell}) \prod p(A_{nm}|\boldsymbol{z}_n = k, \boldsymbol{z}_m = \ell)$. Note that the prior is Beta and the likelihood is a product of Bernoulli distributions each of which with parameter $\eta_{k\ell}$. However, for the likelihood part, the product only involves those entries in $\mathbf{A}$ for which node n belongs to community k and node m belongs to community ℓ. The resulting posterior will be Beta$(a + N_1, b + N_0)$ where $N_1 = \sum_{\boldsymbol{z}_n=k, \boldsymbol{z}_m=\ell} \mathbb{I}[A_{nm} = 1]$ and $N_0 = \sum_{\boldsymbol{z}_n=k, \boldsymbol{z}_m=\ell} \mathbb{I}[A_{nm} = 0]$.

   Finally, the conditional posterior of $\boldsymbol{z}_n$ can be computed as

$$p(\boldsymbol{z}_n = k|\mathbf{A}, \pi, \eta, \mathbf{Z}_{-n}) \propto p(\boldsymbol{z}_n = k|\pi) \prod_{m<n} p(A_{nm}|\boldsymbol{z}_n = k, \boldsymbol{z}_m, \eta) \quad \text{where } p(\boldsymbol{z}_n = k|\pi) = \pi_k$$

   Each term in the product above is a Bernoulli and only involves those entries of $\mathbf{A}$ for which the first node is n and the index of the second node is $m < n$ (since $\mathbf{A}$ is assumed to be symmetric). We can compute $p(\boldsymbol{z}_n = k|\mathbf{A}, \pi, \eta, \mathbf{Z}_{-n})$ for $k = 1, \ldots, K$ and normalize these probabilities. This will give us a probability vector of size K which will be the parameter vector of the multinoulli posterior over $\boldsymbol{z}_n$.

6. Consider a regression problem where we are modeling count-valued responses using a Poisson distribution $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \text{Poisson}(y_n|\exp(\boldsymbol{w}^\top \boldsymbol{x}_n))$, $n = 1, \ldots, N$. Assume a Gaussian prior on the weights, i.e., $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I})$ with λ known. Suggest a sampling based inference algorithm (also justify your choice) for the posterior over $\boldsymbol{w}$ and give the necessary update equations.

   Answer: Note that the model does not have conjugacy (since Poisson and Gaussian are not conjugate to each other). One of the easiest sampling based inference algorithms for this model would be SLGD. This would require writing down the log likelihood (log of Poisson$(y_n|\exp(\boldsymbol{w}^\top \boldsymbol{x}_n))$ this this case) and log prior (log of $\mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I})$ in this case), and taking derivatives of the resulting expression. Once we have the derivatives, you can plug these into the SLGD update equations. I am skipping the detailed equations since this is mostly straightforward.