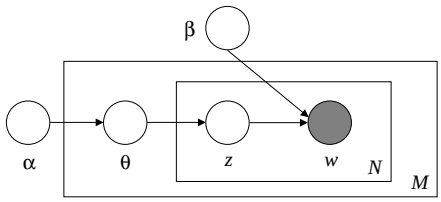
CS 281A/Stat 241A Homework Assignment 5 (due November 29)

1. *Relationship between Gibbs sampling and mean-field.*

   Suppose we have a probability model $p(x_1, \ldots, x_n)$. Let $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots)$. Recall that in Gibbs sampling, we draw $x_i \sim p(x_i \mid \mathbf{x}_{-i})$. Consider a naive mean-field approximation where $q(\mathbf{x}) = \prod_{i=1}^{n} q_i(x_i)$. Recall that in mean-field variational inference, we try to minimize $\mathrm{KL}(q(\mathbf{x})||p(\mathbf{x}))$. Derive a general form for the mean-field update for $q_i(x_i)$ in terms of $p(x_i \mid \mathbf{x}_{-i})$ and $q_j(x_j)$, $j \neq i$. (Hint: use the fact that KL-divergence is minimized when the two arguments are equal).

2. *Latent Dirichlet allocation.*

   Latent Dirichlet allocation (LDA) is a model for discovering topics in sets of documents. Simplifying slightly, the model is as follows:



   (a) For each document, $m = 1, \ldots, M$
      i. Draw topic probabilities $\theta_m \sim p(\theta|\alpha)$
      ii. For each of the N words:
         A. Draw a topic $z_{mn} \sim p(z|\theta_m)$
         B. Draw a word $w_{mn} \sim p(w|z_{mn}, \beta)$,

   where $p(\theta|\alpha)$ is a Dirichlet distribution, and where $p(z|\theta_m)$ and $p(w|z_{mn}, \beta)$ are multinomial distributions. Treat α and β as fixed hyperparameters. Note that β is a matrix, with one column per topic, and the multinomial variable z_{mn} selects one of the columns of β to yield multinomial probabilities for w_{mn}. (See the paper "Latent Dirichlet allocation" on the course website for more details if needed).

   (a) Write down a Gibbs sampler for the LDA model. (I.e., write down the set of conditional probabilities for the sampler).

   (b) Write down a collapsed Gibbs sampler for the LDA model, where you integrate out the topic probabilities θ_m.

3. *Rasch Model and Metropolis within Gibbs.*

   Consider the following model:

$$
\begin{aligned}
\tau_\theta^2 &\sim Gamma(\alpha_\theta, \eta_\theta) \\
\tau_\beta^2 &\sim Gamma(\alpha_\beta, \eta_\beta) \\
\mu|\tau_\beta^2 &\sim N(0, 1/c\tau_\beta^2) \\
\theta_i|\tau_\theta^2 &\sim N(0, 1/\tau_\theta^2) \\
\beta_j|\tau_\beta^2, \mu &\sim N(\mu, 1/\tau_\beta^2) \\
z_{ij}|\theta_i, \beta_j &\sim Bernoulli\left(\frac{1}{1+\exp(\theta_i - \beta_j)}\right)
\end{aligned}
$$

This is a hierarchical Bayesian version of the *Rasch model*. The Rasch model is an important model in educational testing where z_{ij} is binary, and $z_{ij} = 1$ if student j answers question i correctly; θ_i represents question difficulty; and β_j represents student ability.

We wish to derive an MCMC sampler for the posterior distribution of the parameters. Let $\Theta^{(t)}$ denote the current value of the parameters.

(a) For most of the parameters, the hierarchical nature of the model and the choice of priors make a Gibbs sampler a natural choice. In particular, what is the conditional distribution of τ_θ given everything else? I.e., what is the distribution of $\tau_\theta | \vec{z}, \Theta^{(t)} \backslash \{\tau_\theta^{(t)}\}$? (Give the family it belongs to and specify the parameters). Likewise τ_β and μ have simple conditionals given everything else. State what family each belongs to. (No need to perform calculations or specify parameter values).

(b) The β's and θ's do not have as nice a form. Specify, up to a normalizing constant, the conditional density of θ_j given everything else. I.e., find g_j such that

$$p(\theta_j | \vec{z}, \Theta^{(t)} \backslash \{\theta_j^{(t)}\}) \propto g_j(\theta_j).$$

Derive a quadratic approximation to $\log g_j$. Taking the exponential of this quadratic approximation gives us something proportional to the density of some normal density f_j. (Hint: Note that the Rasch model is a hierarchical logistic model. Recall that IRLS was based on a quadratic approximation to the log-likelihood).

(c) Instead of sampling exactly from the full conditional for θ_j given everything else, we may treat the normal density f_j as an approximation to the full conditional and use it as a proposal density for Metropolis-Hastings. (This is sometimes called "Metropolis within Gibbs.") Show that we need not express the acceptance probability ρ in terms of the full posterior density and f_j, but that it simplifies to an expression in terms of g_j and f_j.

(d) One may also perform the Gibbs step for θ_j using slice sampling. Give a decomposition $g_j(\theta_j) = \prod_i g_{ij}(\theta_j)$ where each g_{ij} is unimodal. (Hint: assign a small part of the prior $p(\theta|\tau_\theta)$ to each g_{ij}).

Let $A_i(w) = \{y : g_{ij}(y) \geq w\}$. Use the inequality $\log(1 + e^x) > x$ to derive a region $B_i(w)$ such that $A_i(w) \subset B_i(w)$. Combine this with rejection sampling to obtain a slice sampler for θ_j.