Part A (40 points) Approximate Inference

1. (7 points):

   a) All messages are initialized to 1.0 and we can begine anywhere. We stop when the changes seem to be small (or zero) or when the graph is calibrated. Convergence is not assured (2 p)

   b) between any two nodes with a common variable there exists exactly one path along with the varible is in all sepsets (2p)

   c) This says that the marginalization of the cluster beliefs of two nodes connected by an edge is equal if the marginalization include all variables except the sepset of the edge. (1p)

   d) $\frac{\prod \beta_i(C_i)}{\prod \mu_{ij}(S_{ij})}$ (2p)

2. Variational methods (14 points):

   a) $-E_q[ln(q(\mathbf{z}))]$. (2p)

   b) $\frac{1}{2} + \frac{1}{2}ln(2\pi\sigma^2)$ (2p)

   c) $E_q[ln(\frac{q}{p})]$ (2p)

   d) I-Projection $argmin_q D_{KL}(q \mid\mid p)$) M-Projection $argmin_q D_{KL}(p \mid\mid q)$) (3p)

   e) $ELBO = ln(p(x)) - D_{KL}(q(z) \mid\mid p(z \mid x))$ (2p)

   f) $q(\mathbf{z}) = \prod q_i(z_i)$ (2p)

   g) $q_2(z_2)^* \propto exp(E_{-2}[ln(P(\mathbf{z}, \mathbf{x}))])$ where the -2 means expectation uses all the q_i except i=2. (1p)

3. Sampling based Approximation (9 points):

   a) $E_p[f(\mathbf{x})] \approx \frac{\sum_i^M f(\mathbf{x}_i)}{M}$ where $\mathbf{x}_i$ are samples from the distribution p (2p)

   b) $\{(0,1),(1,1),(1,0),(0,0)\}$ (3p)

   c) With (0,1) $P(z = 0 \mid 0, 1) = .9 > .5$ so here we select z=0 which disagrees with evidence so reject the sample (0,1,0). Similarly using .4 with (1,1) we pick z=0 and reject the sample.
   Final set of samples is $\{(1,0,1),(0,0,1)\}$ (2p)

   d) $w_i \propto 0, .2$ so sample 2 gets normalized weights of 1 while sample 1 has 0.(2p)

4. Markov Chain Monte Carlo Methods, MCMC (10 points)

   a) A Markov chain is a sequence where each element depends on the previous element. A Monte Carlo method is one in which sampling is used to estimate a distribution. In MCMC we gnerate the samples of our target distribution by sampling a markov chain that is chosen to converge to samples from our target. (2p)

   b) $P(\mathbf{x}) = \sum_{x'} T(\mathbf{x} \mid \mathbf{x}')P(\mathbf{x}')$ and $T(\mathbf{x}' \mid \mathbf{x})P(\mathbf{x}) = T(\mathbf{x} \mid \mathbf{x}')P(\mathbf{x}')$ (3p)

   c) Sample with Q and accept the sample with probability $max(1, \frac{P(\mathbf{x})Q(\mathbf{x}'|\mathbf{x})}{P(\mathbf{x}')Q(\mathbf{x}|\mathbf{x}')})$ (2p)

   d) for $i = 1...n$ sample x_i from $p(x_i \mid x_1...x_{i-1}, x_{i+1}...x_n)$ (2p)

   e) Mixing is the property of the kernel being able to generate samples from all the modes of our distribution from any starting point. It is important to prevent the MC sequence from becoming stuck in one mode. (1p)

Part B (20 points) Learning

5. Maximum Likelihood Estimation (8 points)

   a) $\sum_i^m ln(P(\mathbf{x_i} \mid \theta))$ (2p)

   b) $argmax_\theta P(\mathbf{x_i} \mid \theta))$ (2p)

   c) θ_i would be three multinomial rates, one for each value that would add up to 1. The sufficient stats are the number of times each value occurs in the data (or the frequencies) The MLE rates are then just those frequencies $\theta_i = N_i/N$ (2p)

   d) The sufficient stats are $\sum_i x_i$ and $\sum_i x_i^2$. The MLE of the mean is then $\mu = \frac{\sum_i x_i}{m}$ and the variance is $\sigma^2 = \frac{\sum_i x_i^2}{m} - (\frac{\sum_i x_i}{m})^2$ (2p)

6. Bayesian Parameter Estimation (8 points)

   a)$p(\mathbf{x} \mid \theta) = \int p(\mathbf{x} \mid \theta)p(\theta \mid \{\mathbf{x_i}\})d\theta$ where $p(\theta \mid \{\mathbf{x_i}\}) = \frac{p(\theta)\sum_i p(\mathbf{x_i}|\theta)}{p(\{\mathbf{x_i}\})}$ (2p)

   b) A Conjugate prior is a choice for $p(\theta)$ such that when multiplied by $p(x \mid \theta)$ it has the same form with different parameters (2p)

   c) Beta for a Binomial, Dirichlet for the multinmial, (or Gaussian for a Gaussian) (4p)

7. Partially Observed Data (4 points)

   a) MCAR has the observation process independant of the data. MAR has the observation of a value independent of that value given the other observations. (3p)

   b) $\theta \neq \theta' \rightarrow P(x \mid \theta) \neq P(x \mid \theta')$ for some data set x.(1p)