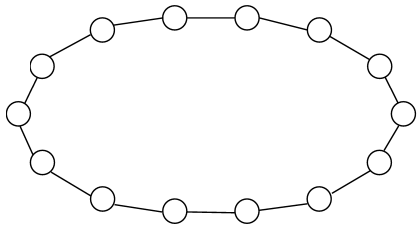
Name: _____ **Student ID:**_____

Final Exam, ECE1508, Fall 2003, Prof. B. J. Frey
3 hours; Total of 30 marks; Closed-book; Calculators allowed

1. (6 marks) You would like to compute the marginals of a set of variables whose distribution is described by a ring-shaped graphical model, *i.e.*, a model consisting of a single cycle as shown below. The variables x_i, $i = 1, \ldots, N$ each have the same alphabet, $x_i \in \{1, \ldots, K\}$. The joint distribution is given by $P(x_1, \ldots, x_N) = \phi_N(x_N, x_1) \prod_{i=1}^{N-1} \phi_i(x_i, x_{i+1})$.



1a) (3 marks) Derive a variational mean field inference algorithm that approximates the distribution over x by $Q(x_1, \ldots, x_N) = Q(x_1)Q(x_2) \cdots Q(x_N)$. Write the mean field free energy, and take the derivatives w.r.t. $Q(x_i)$, $i = 1, \ldots, N$ to find the updates for each $Q(x_i)$. Make sure to include a Lagrange multiplier to ensure that $\sum_{x_i} Q(x_i) = 1$.

1b) (3 marks) To apply the sum-product algorithm in one direction in the ring, you start with a randomly drawn (positive) message, and propagate messages from x_1 to x_2, from x_2 to x_3, and so on, until you propagate a message from x_N to x_1, completing one "iteration". Show that in the limit of an infinite number of iterations, the message arriving at x_1 will converge to the dominant eigenvector of the $K \times K$ matrix $\prod_{i=1}^{N} \mathbf{\Phi}_i$, where $\mathbf{\Phi}_i$ is the $K \times K$ matrix containing the values in the potential $\phi_i()$, *i.e.*, $[\mathbf{\Phi}_i]_{jk} = \phi_i(k, j)$. Hint: Suppose $\mathbf{A}$ is a square matrix with positive real-valued eigenvalues. Then, for most vectors $\mathbf{u}$, $\lim_{m \to \infty} \mathbf{A}^m \mathbf{u} = \mathbf{v}_1$, where $\mathbf{v}_1$ is the dominant eigenvector of $\mathbf{A}$ (the eigenvector with an eigenvalue that is larger than all other eigenvalues).

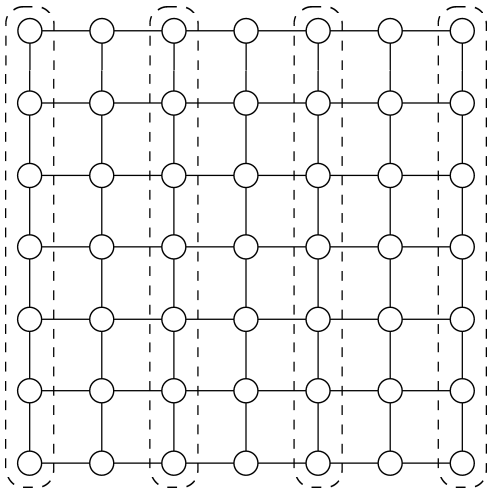
2. (8 marks) A linear dynamic system models a sequence of observed real-valued variables $y_1, y_2, \ldots$ using a sequence of hidden real-valued variables, $x_0, x_1, \ldots$. The joint density function factorizes as follows: $P(x, y) = P(x_0) \prod_{i \geq 1} (P(x_i|x_{i-1}) P(y_i|x_i))$ where $P(x_0) = \mathcal{N}(x_0; 0, 1)$, $P(x_i|x_{i-1}) = \mathcal{N}(x_i; ax_{i-1}, b^2)$ for $i \geq 1$, and $P(y_i|x_i) = \mathcal{N}(y_i; cx_i, d^2)$ for $i \geq 1$.

2a) (1 mark) Draw a cycle-free factor graph for this system.

2b) (2 marks) To infer the state variable x_i given observations $y_1, \ldots, y_i$ the sum-product algorithm can be used. In fact, this is one way to derive the *Kalman filtering* algorithm. Below, you'll show that assuming the message leaving x_{i-1} and heading forward in time is Gaussian with the form $\mathcal{N}(x_{i-1}; \mu_{i-1}, \sigma^2_{i-1})$, then the *normalized* message leaving x_i and heading forward in time is also Gaussian, *i.e.*, has the form $\mathcal{N}(x_i; \mu_i, \sigma_i^2)$. Given this and the properties of the sum-product algorithm in trees, explain why it is that if messages are passed up from the observations and along the chain from x_0 to x_i, then the message leaving x_i, $\mathcal{N}(x_i; \mu_i, \sigma_i^2)$ is equal to $P(x_i|y_1, \ldots, y_i)$.

2c) (5 marks) Use the sum-product (or, rather, "integrate-product") algorithm to derive expressions for μ_i and σ_i^2 in terms of the parameters of the previous message μ_{i-1} and σ_{i-1}^2, the new observation y_i, and the model parameters a, b, c, d.

3. (6 marks) You propose an MCMC technique for sampling from variables arranged in a grid MRF, as shown below. Starting from a random configuration, for every column of circled variables, the technique obtains an *exact joint sample* of the variables in the column, given the column's neighboring variables (which are held fixed). These joint samples are drawn for all circled columns, *in parallel*. Then, the columns that were held fixed in the previous sampling step are sampled in a similar fashion, holding the circled variables fixed. This process is repeated, forming an MCMC sequence.



3a) (2 marks) Describe an efficient way to sample an entire column of variables, given the neighboring columns of variables. For a column with n binary variables, your method should take order n time, not order 2^n time.

3b) (2 marks) Show that this MCMC technique can be viewed as Gibbs sampling. Note that you will need to explain why parallel updates are acceptable here, since in general parallel updates will not satisfy detailed balance.

3c) (2 marks) This question is about the Gibbs sampling method in general. Show that an infinite ensemble of Gibbs samplers monotonically decreases the free energy at each step of sampling, where at each step, one of the hidden variables, h_i, is updated in parallel in all Gibbs samplers. (Hint: The free energy at step i is $F^{(i)} = \sum_h Q^{(i)}(h) \ln(Q^{(i)}(h)/P(h,v))$, where $Q^{(i)}(h)$ is the distribution for the ensemble at step i.)

4. (10 marks) You observe a noisy version $\mathbf{y}$ of an N-dimensional vector, $\mathbf{x}$, which *must* lie in a K-dimensional linear subspace defined by $\mathbf{Ax} = \mathbf{0}$, where $K < N$. $\mathbf{A}$ is an $(N-K) \times N$ matrix whose rows span the null space. The noise added to x_i to get y_i is independent zero-mean Gaussian noise with variance σ^2. Assuming $\mathbf{x}$ is uniformly distributed in the linear subspace, the joint distribution on $\mathbf{x}$ and $\mathbf{y}$ can be written

$$P(\mathbf{x}, \mathbf{y}) = \Big(\prod_{j=1}^{N-K} \delta(\mathbf{a}_j \mathbf{x}, 0) \Big) \Big(\prod_{i=1}^{N} \mathcal{N}(y_i; x_i, \sigma^2)) \Big), \tag{1}$$

where $\mathbf{a}_j$ is the jth row of $\mathbf{A}$, and $\delta(c, d) = 1$ if $c = d$ and $\delta(c, d) = 0$ if $c \neq d$. Note that $P(x, y) = 0$ for any vector x that does not satisfy $\mathbf{a}_j \mathbf{x} = 0$ for all $j = 1, \ldots, N - K$.

4a) (1 mark) Draw a factor graph on the variables $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ corresponding to the product of $(N - K) + N$ functions in (1).

4b) (3 marks) To infer the hidden vector, $\mathbf{x}$, you consider 4 techniques: ICM, updating one x_i at a time; Gibbs sampling, updating one x_i at a time; fully-factorized variational inference (mean field), updating one $Q(x_i)$ at a time; and the sum-product algorithm, ignoring the cycles. In fact, all but one of these techniques are so greedy that they will get stuck in a local minimum in the first iteration, making them useless for inference. Identify the 3 overly-greedy techniques, and in each case explain why the technique gets stuck in a local minimum in the first iteration. Also, explain why the remaining technique will not immediately get stuck in a local minimum. Hint: The free energy is infinite if the approximate posterior distribution is non-zero for one or more values of $\mathbf{x}$ where $\mathbf{Ax} \neq \mathbf{0}$.

4c) (1 mark) An alternative formulation of the above problem follows from the fact that $\mathbf{x}$ satisfies $\mathbf{Ax} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{Bz}$, where $\mathbf{B}$ is an $N \times K$ matrix that maps a K-dimensional vector $\mathbf{z}$ to the N-dimensional space of valid $\mathbf{x}$'s, *i.e.*, the columns of $\mathbf{B}$ span the K-dimensional subspace. In this case, $y_i = \mathbf{b}_i\mathbf{z} + noise$, where $\mathbf{b}_i$ is the ith row of $\mathbf{B}$. Assuming $\mathbf{z}$ is uniformly distributed, the joint distribution over $\mathbf{z}$ and $\mathbf{y}$ can be written

$$P(\mathbf{z}, \mathbf{y}) = \prod_{i=1}^{N} \mathcal{N}(y_i; \mathbf{b}_i\mathbf{z}, \sigma^2). \tag{2}$$

Draw the factor graph on the variables $y_1, \ldots, y_N$ and $z_1, \ldots, z_K$ corresponding to the product of N functions in (2).

4d) (3 marks) The 4 techniques described in 4(b) can also be applied in this alternative model, where now the techniques are applied to the z_k's instead of the x_i's. It turns out that in this formulation, none of the 4 techniques will get stuck in a local minimum in the first iteration. Explain why.

4e) (2 marks) Briefly contrast the two formulations in terms of their time efficiency and quality of solution in solving the problem. Give advantages *and* disadvantages for *both* formulations.