NAME: _____ STUDENT NO.: _____

Do ALL questions.

Final exam
ECE1508F, Probabilistic Inference Algorithms and Machine Learning
November 29, 2006

DO ALL QUESTIONS, *BOTH* ON YOUR IN-CLASS PAPER *AND* YOUR TAKE-HOME PAPER

Closed-book, calculators allowed, 2 hours, total of 50 marks

NAME: _____ STUDENT NO.: _____

1. (10 marks) Suppose x and y are binary-valued variables taking values 0 or 1, and that

$$P(x,y) = \begin{cases} 1 & \text{if } x = 0 \text{ and } y = 0 \\ 1 & \text{if } x = 1 \text{ and } y = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose we wish to approximate $P(x,y)$ with a distribution $Q(x,y)$ that assumes x and y are independent, *i.e.*, $Q(x,y) = Q(x)Q(y)$.

1a) (5 marks) Find the distributions $Q(x)$ and $Q(y)$ that minimizes $D(Q,P) = \sum_x \sum_y Q(x)Q(y) \log \frac{Q(x)Q(y)}{P(x,y)}$. If there are multiple solutions that achieve the minimum, give every solution.

1a) (5 marks) Find the distributions $Q(x)$ and $Q(y)$ that minimizes $D(P,Q) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x)Q(y)}$. If there are multiple solutions that achieve the minimum, give every solution.

2. (15 marks) Kalman filtering can be derived by applying the sum-product algorithm to the factor graph in Fig. A. State z_{t+1} is related to the previous state z_t by

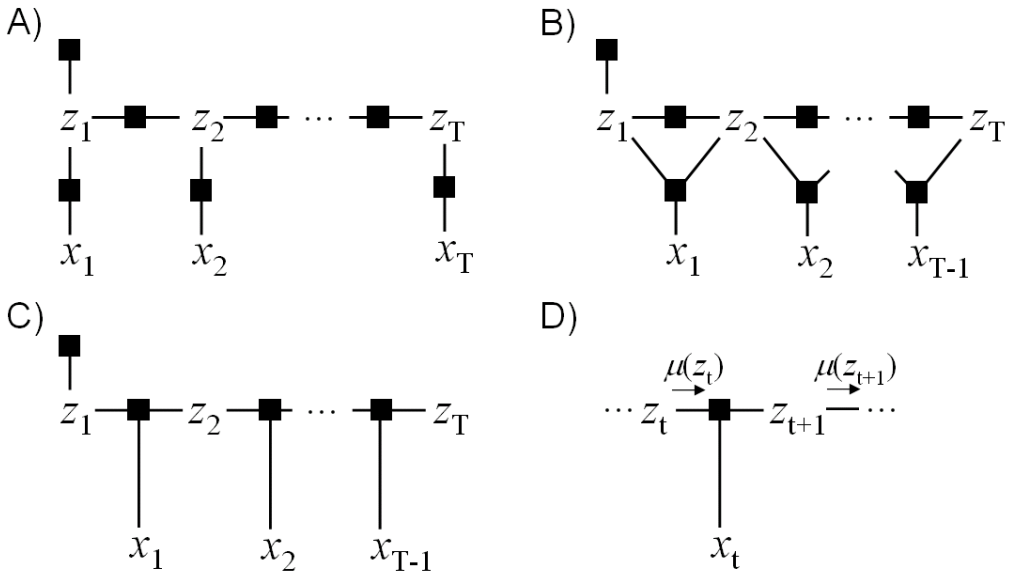
$$p(z_{t+1}|z_t) = \frac{1}{\sqrt{2\pi\alpha^2}} e^{-(z_{t+1}-az_t)^2/2\alpha^2}$$

and the observation likelihood x_t at time t is $p(x_t|z_t) = e^{-(x_t-bz_t)^2/2\beta^2}/\sqrt{2\pi\beta^2}$. The initial state z_1 is Gaussian with mean 0 and variance α^2 – the singleton function node connected to z_1 accounts for this.

In your application, the observations depend on the *difference* $z_{t+1} - z_t$ in the states from one time step to the next, so you construct the factor graph shown in Fig. B, where

$$p(x_t|z_t, z_{t+1}) = \frac{1}{\sqrt{2\pi\beta^2}} e^{-(x_t-b(z_{t+1}-z_t))^2/2\beta^2}$$

and $p(z_{t+1}|z_t)$ is the same as before. Here, you'll derive a new Kalman filter for this kind of application. (10 marks)



a) (2 marks) The factor graph in Fig. B is not a tree, but it can be converted to the tree shown in Fig. C. Write down an expression for the function $f(z_t, z_{t+1}, x_t)$ in this factor graph.

b) (8 marks) Fig. D shows a fragment of the factor graph from Fig. C. Suppose the message $\mu(z_t)$ is Gaussian with mean m_t and variance σ_t^2. Find expressions for m_{t+1} and σ_{t+1}^2 in terms of x_t, m_t, σ_t^2, a, b, α^2 and β^2. (You can make use properties of Gaussians that we have previously worked out in class or class assignments.)

c) (5 marks) Filtering consists of computing $p(z_i|x_1, \ldots, x_i)$, whereas prediction consists of computing $p(x_i|x_1, \ldots, x_{i-1})$. To compute these, we first set $m_1 = 0$ and $\sigma_1^2 = \alpha^2$ and then apply the recursions you were asked to derive in (b) until we have computed m_i and σ_i^2, which are the mean and the variance of the message $\mu(z_i)$. Give expressions for the mean and variance of z_i given $x_1, \ldots, x_i$ (the filtering task) *and* x_i given $x_1, \ldots, x_{i-1}$ (the prediction task). (Note that each of these distributions is conditioned on a different set of variables!)

3. (25 marks) You would like to apply 1-dimensional factor analysis (*i.e.*, there is one real-valued hidden variable z) to analyze some vectors, but you notice that occasionally, a sensor value (*not* the entire vector) is an outlier. So, you propose to extend factor analysis to include one binary variable s_k for each element x_k in the data vector $\mathbf{x} = (x_1, \ldots, x_K)^{\mathrm{T}}$. $s_k = 1$ indicates that x_k is sporadic noise, while $s_k = 0$ indicates that x_k should fit the factor analysis model. You assume the s's are i.i.d. with $P(s_k = 1) = \rho$. Assume that if $s_k = 1$, the variance of x_k is ψ_{1k} and if $s_k = 0$, the variance of x_k is ψ_{0k}. For a single training case, the joint distribution of the latent real variable z, the binary noise indicator vector $\mathbf{s}$ and the visible variables $\mathbf{x}$ is

$$P(\mathbf{x}, \mathbf{s}, z) = \Big(\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\Big)\Big(\prod_{k=1}^{K} \rho^{s_k}(1-\rho)^{1-s_k} \frac{1}{\sqrt{2\pi\psi_{s_k k}}} e^{-(x_k - \lambda_k z)^2/2\psi_{s_k k}}\Big),$$

where $\psi_{s_k k} = \psi_{0k}$ if $s_k = 0$ and $\psi_{s_k k} = \psi_{1k}$ if $s_k = 1$.

To avoid computing the exact posterior distribution $P(z, \mathbf{s}|\mathbf{x})$, which takes an amount of time that is exponential in K, you use a variational approximation: $P(z, \mathbf{s}|\mathbf{x}) \approx Q(z, \mathbf{s})$, where

$$Q(z, \mathbf{s}) = Q(z) \prod_{k=1}^{K} Q(s_k),$$
$$Q(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\mu)^2/2\sigma^2},$$
$$Q(s_k = 1) = \gamma_k, \quad Q(s_k = 0) = 1 - \gamma_k$$

a) (10 marks) The free energy is $F = \sum_{\mathbf{s}} \int_z Q(z, \mathbf{s}) \log Q(z, \mathbf{s})/P(\mathbf{x}, \mathbf{s}, z)$. Substitute the above expressions for Q and P into the free energy and simplify it so that F depends only on the data vector $\mathbf{x}$, the model parameters ρ, λ, ψ and the variational parameters μ, σ^2 and γ (with appropriate indices). *Make sure that your expression for the free energy can be computed in time that is linear in K.*

3a) continued

b) (15 marks) Take the derivatives of F wrt the variational parameters μ, σ^2 and γ and specify an implementable procedure for minimizing F wrt these variational parameters.

3b) continued