MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science

6.438 ALGORITHMS FOR INFERENCE
Fall 2011

Quiz 2
Monday, December 12, 2011
7:00pm–10:00pm

- This is a closed book exam, but two $8\frac{1}{2}'' \times 11''$ sheets of notes (4 sides total) are allowed.
- Calculators are **not** allowed.
- There are **3** problems of approximately equal value on the exam.
- The problems are not necessarily in order of difficulty. We recommend that you read through all the problems first, then do the problems in whatever order suits you best.
- Record all your solutions in the answer booklet provided. **NOTE: Only the answer booklet is to be handed in—no additional pages will be considered in the grading**. You may want to first work things through on the scratch paper provided and then neatly transfer to the answer sheet the work you would like us to look at. Let us know if you need additional scratch paper.
- A correct answer does not guarantee full credit, and a wrong answer does not guarantee loss of credit. You should clearly but concisely indicate your reasoning and **show all relevant work**. Your grade on each problem will be based on our best assessment of your level of understanding as reflected by what you have written in the answer booklet.
- Please be neat—we can't grade what we can't decipher!

Problem 1

All parts to this problem can be done independently.

We wish to use a variational approach to approximate $p(\mathbf{x}|\mathbf{y})$, where $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)})$ is a collection of L vectors. We will use an approximating distribution q that factors as

$$q(\mathbf{x}) = \prod_{\ell=1}^{L} q_\ell(\mathbf{x}^{(\ell)}) \tag{1}$$

with each q_ℓ summing to 1. Note that each variable in $\mathbf{x}$ is in one and only one q_ℓ.

(a) **(2 Points)** Let q^* be the q that solves our variational optimization problem (i.e. q^* maximizes $\mathbb{E}_q[\log p(\mathbf{x}, \mathbf{y})] + H(q)$ subject to the constraint that q factorizes as above). Justify why q^* can in principle be a better approximation than the mean field approximation where the approximating distribution fully factors $r(\mathbf{x}) = \prod_\ell \prod_i r_i^{(\ell)}(x_i^{(\ell)})$.

(b) **(3 Points)** To search for q^*, we will hold all of the $q_1, \ldots, q_L$ constant except for a specific q_ℓ and maximize with respect to q_ℓ. Show that maximizing $\mathbb{E}_q[\log p(\mathbf{x}, \mathbf{y})] + H(q)$ with respect to q_ℓ keeping the rest of the factors constant is equivalent to minimizing $D(q_\ell \parallel \tilde{p}(\mathbf{x}^{(\ell)}, \mathbf{y}))$ where

$$\tilde{p}(\mathbf{x}^{(\ell)}, \mathbf{y}) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(\mathbf{x}, \mathbf{y})]\right) \tag{2}$$

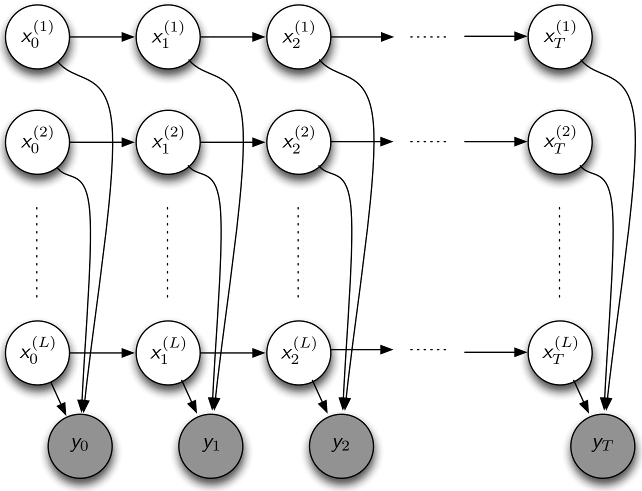
and $\mathbb{E}_{-\ell}[\cdot]$ denotes expectation is over the distribution $\prod_{\substack{i=1 \\ i\neq\ell}}^{L} q_i$, so that $\tilde{p}$ is a distribution over $\mathbf{x}^{(\ell)}$.

You may use the fact that maximizing $\mathbb{E}_q[\log p(\mathbf{x}, \mathbf{y})] + H(q)$ *is equivalent to minimizing* $D(q \parallel p(\mathbf{x}|\mathbf{y}))$ *the KL divergence between* q *and* $p(\mathbf{x}|\mathbf{y})$.

From this, we conclude that the fixed point equations are given by

$$q_\ell(\mathbf{x}^{(\ell)}) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(\mathbf{x}, \mathbf{y})]\right). \tag{3}$$

Consider the following variation on an HMM where $x_t^{(\ell)}$ is a discrete random variable with K possible values and there are L chains:



The initial distributions $p_{x_0^{(\ell)}}$ are given, the transition distributions $p_{x_t^{(\ell)}|x_{t-1}^{(\ell)}}$ are homogeneous and given, and the emission probabilities $p_{y_t|x_t^{(1)},\dots,x_t^{(L)}}$ are homogeneous and given. In this model, the y_t are observed for $t = 0, \dots, T$. In terms of the notation above, $\mathbf{x}^{(\ell)} = (x_0^{(\ell)}, \dots, x_T^{(\ell)})$ is the vector corresponding to the ℓth chain and $\mathbf{y}$ denotes $(y_0, \dots, y_T)$.

(c) **(3 Points)** For this model, the right hand side of the fixed point equations (3) simplify to

$$q_\ell(\mathbf{x}^{(\ell)}) \propto p(x_0^{(\ell)}) \prod_{t=1}^{T} p(x_t^{(\ell)}|x_{t-1}^{(\ell)}) \prod_{t=0}^{T} \psi_t^{(\ell)}(x_t^{(\ell)}). \tag{4}$$

Determine $\psi_t^{(\ell)}$, expressing your answer in terms of $p_{y_t|x_t^{(1)},\dots,x_t^{(L)}}$ and $q_{\ell'}$'s where $\ell' \neq \ell$.

(d) **(2 Points)** Given a fixed point q^* of (4), draw an undirected graphical model that represents the distribution $q^*(\mathbf{x})$.

Problem 2

All parts to this problem can be done independently.

Suppose we are interested in a distribution of the form

$$p_{\mathsf{x}}(\mathbf{x}) = \frac{1}{Z} R(\mathbf{x}) \tag{5}$$

over random variables $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ which take values in $\{0, 1\}$. We are given R, a strictly positive function we can efficiently evaluate (such as a product of potentials), and we are interested in estimating the partition function $Z = \sum_{\mathbf{x}} R(\mathbf{x})$.

(a) **(1 point)** One approach is based on importance sampling. For a given (strictly positive) proposal distribution q, give a nonnegative function $f(\mathbf{x})$ which can be efficiently evaluated, such that $E_q[f(\mathbf{x})] = Z$.

(b) **(3 points)** Suppose we generate k samples $\mathbf{x}^1, \ldots, \mathbf{x}^k$ from q. Consider the two estimators

$$\alpha(\mathbf{x}^1, \ldots, \mathbf{x}^k) = \frac{1}{k} \sum_{i=1}^{k} f(\mathbf{x}^i) \tag{6}$$

$$\beta(\mathbf{x}^1, \ldots, \mathbf{x}^k) = \exp\left(\frac{1}{k} \sum_{i=1}^{k} \log f(\mathbf{x}^i) \right), \tag{7}$$

where f is the function from part (a). Say whether each of $E[\alpha]$ and $E[\beta]$ is less than, equal to, or greater than Z. Justify your answer. *Hint*: use Jensen's inequality to determine the relationship between α and β for a given (fixed) set of samples $\mathbf{x}^1, \ldots, \mathbf{x}^k$.

Now we consider a more accurate procedure called annealed importance sampling. We are given functions $R^{(t)}$ which define a series of $T + 1$ distributions $p_{\mathsf{x}}^{(0)}, \ldots, p_{\mathsf{x}}^{(T)}$:

$$p_{\mathsf{x}}^{(t)}(\mathbf{x}) = \frac{1}{Z^{(t)}} R^{(t)}(\mathbf{x}) \tag{8}$$

such that $p_{\mathsf{x}}^{(T)} = p_{\mathsf{x}}$ and $Z^{(T)} = Z$. The initial distribution $p_{\mathsf{x}}^{(0)}$ is one for which we can easily generate samples and calculate the partition function. (For instance, a uniform distribution.) Consider the following procedure:

```
Sample $\mathbf{x}^{(0)}$ from the distribution $p_{\mathsf{x}}^{(0)}$, and set $w^{(0)} := Z^{(0)}$.
For $t = 0, \ldots, T - 1$,
    Set $w^{(t+1)} := w^{(t)} \frac{R^{(t+1)}(\mathbf{x}^{(t)})}{R^{(t)}(\mathbf{x}^{(t)})}$
    Let $\mathbf{x}^{(t+1)}$ be the result of applying one step of Gibbs sampling to $\mathbf{x}^{(t)}$ with respect to the distribution $p_{\mathsf{x}}^{(t+1)}$.
```

(Recall that, in Gibbs sampling, we randomly choose a variable x_i and set x_i to a sample from the conditional distribution $p^{(t+1)}_{\mathsf{x}_i|\boldsymbol{\mathsf{x}}_{-i}}(\cdot|\mathbf{x}_{-i})$, where $\boldsymbol{\mathsf{x}}_{-i}$ denotes the variables other than x_i.) Note that, while this is a sequential sampling procedure like particle filtering, there is only one sample $\mathbf{x}^{(t)}$ and one weight $w^{(t)}$ in each step.

Let $q_{\boldsymbol{\mathsf{x}}^{(t)},\mathsf{w}^{(t)}}$ denote the joint distribution over the state and weight at a given time step, i.e. the randomness is with respect to the choices made in the above algorithm.

We will prove inductively that in each iteration, $E_{q_{\boldsymbol{\mathsf{x}}^{(t)},\mathsf{w}^{(t)}}}[w^{(t)}f(\mathbf{x}^{(t)})] = Z^{(t)}E_{p^{(t)}}[f(\mathbf{x})]$ for all functions f. This yields the partition function when we plug in $t = T$ and $f(\mathbf{x}) = 1$. The base case is clear from the definition. Parts (c) and (d) give the inductive step. Note that the two parts can be completed independently of each other.

(c) **(3 points)** Suppose that

$$E_{q_{\boldsymbol{\mathsf{x}}^{(t)},\mathsf{w}^{(t)}}}[w^{(t)}f(\mathbf{x}^{(t)})] = Z^{(t)}E_{p^{(t)}_{\boldsymbol{\mathsf{x}}}}[f(\mathbf{x})] \tag{9}$$

for all functions f. Show that

$$E_{q_{\boldsymbol{\mathsf{x}}^{(t)},\mathsf{w}^{(t+1)}}}[w^{(t+1)}g(\mathbf{x}^{(t)})] = Z^{(t+1)}E_{p^{(t+1)}_{\boldsymbol{\mathsf{x}}}}[g(\mathbf{x})] \tag{10}$$

for all functions g. *Hint:* In evaluating the left hand side of (10), use (9) with a suitable choice of f.

(d) **(3 points)** Suppose that

$$E_{q_{\boldsymbol{\mathsf{x}}^{(t)},\mathsf{w}^{(t+1)}}}[w^{(t+1)}f(\mathbf{x}^{(t)})] = Z^{(t+1)}E_{p^{(t+1)}_{\boldsymbol{\mathsf{x}}}}[f(\mathbf{x})] \tag{11}$$

for all functions f. Show that

$$E_{q_{\boldsymbol{\mathsf{x}}^{(t+1)},\mathsf{w}^{(t+1)}}}[w^{(t+1)}g(\mathbf{x}^{(t+1)})] = Z^{(t+1)}E_{p^{(t+1)}_{\boldsymbol{\mathsf{x}}}}[g(\mathbf{x})] \tag{12}$$

for all functions g.

Hint: Start by arguing that

$$E_{q_{\boldsymbol{\mathsf{x}}^{(t+1)},\mathsf{w}^{(t+1)}}}[w^{(t+1)}g(\mathbf{x}^{(t+1)})] = E_{q_{\boldsymbol{\mathsf{x}}^{(t)},\mathsf{w}^{(t+1)}}}[w^{(t+1)}E_{q_{\boldsymbol{\mathsf{x}}^{(t+1)}|\boldsymbol{\mathsf{x}}^{(t)}}}[g(x^{(t+1)})|\mathbf{x}^{(t)}]]. \tag{13}$$

To evaluate the right hand side of (13), use (11) with a suitable choice of f. Then use the fact that $p^{(t+1)}_{\boldsymbol{\mathsf{x}}}$ is a fixed point of the Gibbs sampling transition operator.

Problem 3

All parts to this problem can be done independently.
Suppose we have a zero-mean Gaussian random vector $\mathsf{x} = (\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_3) \in \mathbb{R}^3$ with *unknown* information matrix $\mathbf{J}$, i.e.,

$$\mathsf{x} = \begin{pmatrix} \mathsf{x}_1 \\ \mathsf{x}_2 \\ \mathsf{x}_3 \end{pmatrix} \sim \mathtt{N}^{-1}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \right). \tag{14}$$

Our goal is to determine the graph structure over random variables $\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_3$ given observations $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}$, sampled i.i.d. from $\mathtt{N}^{-1}(\mathbf{0}, \mathbf{J})$. Letting $\boldsymbol{\mu} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}^{(m)}$, the problem with taking the sample covariance matrix $\frac{1}{M}\sum_{m=1}^{M}(\mathbf{x}^{(m)} - \boldsymbol{\mu})(\mathbf{x}^{(m)} - \boldsymbol{\mu})^{\mathrm{T}}$ and inverting it to estimate the information matrix is that inverting a sample covariance matrix will almost surely give a matrix that has all non-zero entries, corresponding to a fully connected graph, which isn't very interesting.

(a) **(2 points)** We know that $\mathsf{x}_1 | (\mathsf{x}_2, \mathsf{x}_3) \sim \mathtt{N}^{-1}(\hat{h}_1, J_{11})$ for some $\hat{h}_1$.

Show that $\widehat{h}_1 = f(J_{12})\mathsf{x}_2 + g(J_{13})\mathsf{x}_3$ where $f(J_{12}) = 0$ if and only if $J_{12} = 0$ and $g(J_{13}) = 0$ if and only if $J_{13} = 0$.

(b) **(2 points)** For our model, we have $\mathbb{E}[\mathsf{x}_1 \mid \mathsf{x}_2, \mathsf{x}_3] = \alpha \mathsf{x}_2 + \beta \mathsf{x}_3$. Provide expressions for α and β in terms of $f(J_{12})$, $g(J_{13})$, and J_{11}. You should find that $\alpha = 0$ if and only if $f(J_{12}) = 0$, and $\beta = 0$ if and only if $g(J_{13}) = 0$.

(c) **(1 point)** Using parts (a) and (b), justify why $\alpha = 0$ if and only if there is no edge between x_1 and x_2, and similarly why $\beta = 0$ if and only if there is no edge between x_1 and x_3.

(d) **(4 points)** For this part only, we'll assume that there is no edge between x_1 and x_3, so by the results of parts (b) and (c), we have $\mathbb{E}[\mathsf{x}_1 \mid \mathsf{x}_2] = \mathbb{E}[\mathsf{x}_1 \mid \mathsf{x}_2, \mathsf{x}_3] = \alpha \mathsf{x}_2$. We want to see whether there is an edge between x_1 and x_2. We write $\mathbb{E}[\mathsf{x}_1 \mid \mathsf{x}_2]$ in variational form:

$$\mathbb{E}[\mathsf{x}_1 \mid \mathsf{x}_2] = \operatorname*{argmin}_{h \text{ s.t. } h(\mathsf{x}_2) = a\mathsf{x}_2 \text{ for some } a \in \mathbb{R}} \mathbb{E}[(\mathsf{x}_1 - h(\mathsf{x}_2))^2]. \tag{15}$$

We can't compute the expectation on the right-hand side since we don't know $\mathbf{J}$, so we replace the expectation with an empirical mean and instead solve:

$$\widehat{\alpha} = \operatorname*{argmin}_{a \in \mathbb{R}} \frac{1}{M} \sum_{m=1}^{M} (x_1^{(m)} - a x_2^{(m)})^2 = \frac{\sum_{m=1}^{M} x_1^{(m)} x_2^{(m)}}{\sum_{m=1}^{M} (x_2^{(m)})^2}, \tag{16}$$

where the second equality follows from setting the derivative of the objective function to 0. The resulting $\widehat{\alpha}$ is our estimate of α in the previous parts; thus, we use $\widehat{\alpha}$ to infer whether there is an edge between x_1 and x_2. Unfortunately, the right-hand side is nonzero with probability 1, so we always have an edge between x_1 and x_2.

Consider instead if we replaced optimization problem (16) with the following new optimization problem:

$$\alpha^* = \operatorname*{argmin}_{a \in \mathbb{R}} \left\{ \frac{1}{M} \sum_{m=1}^{M} (x_1^{(m)} - a x_2^{(m)})^2 + \lambda |a| \right\}, \tag{17}$$

where $\lambda > 0$ is a fixed constant. The new term added, $\lambda|a|$, encourages sparsity.

Provide an expression for α^* *under the assumption that we are only optimizing over* $a \in [0, \infty)$, and determine constant L in terms of λ and M such that $\alpha^* = 0$ whenever $\sum_{m=1}^{M} x_1^{(m)} x_2^{(m)} \leq L$.

Remark: If we optimize over $a \in \mathbb{R}$ rather than just $a \in [0, \infty)$, then $\alpha^* = 0$ whenever $-L \leq \sum_{m=1}^{M} x_1^{(m)} x_2^{(m)} \leq L$, but you don't need to show this.

(e) **(1 point)** Optimization problem (17) can be extended to work when x_1 has more than just one neighbor. In fact, if $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ is instead an N-dimensional Gaussian, then we can solve:

$$\boldsymbol{\alpha}^* = \operatorname*{argmin}_{\mathbf{a}=(a_2,a_3,\ldots,a_N) \in \mathbb{R}^{N-1}} \left\{ \frac{1}{M} \sum_{m=1}^{M} \left(x_1^{(m)} - \sum_{j=2}^{N} a_j x_j^{(m)} \right)^2 + \lambda \sum_{j=2}^{N} |a_j| \right\}, \tag{18}$$

for some fixed, pre-specified $\lambda > 0$. Sparsity in $\boldsymbol{\alpha}^*$ tells us which neighbors x_1 has, i.e., $\boldsymbol{\alpha}_j^* = 0$ implies that x_1 and x_j are not neighbors.

Suppose $M \ll N$. In this setting, the sample covariance matrix lacks an inverse as it is not full rank. However, we can still solve optimization problem (18). Assume we have an algorithm that solves (18) in time $O(NM^2)$. Describe an $O(N^2M^2)$ procedure that estimates the graph structure across all N variables — not just which nodes are neighbors of x_1. Feel free to resolve inconsistencies regarding whether two nodes are neighbors arbitrarily.

Your answer should be brief.

6.438 Algorithms for Inference
Fall 2014