Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

Problem Set 7

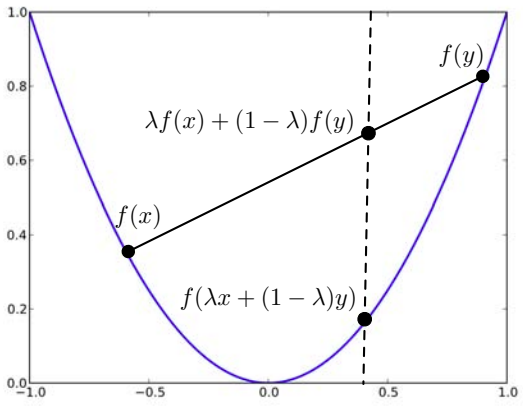
Issued: Thursday, November 6, 2014 **Due:** Tuesday, November 18, 2014

Problem 7.1

A function f over a scalar or vector-valued variable $\mathbf{x}$ is said to be *convex* if it satisfies:

$$f\left(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\right) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

Intuitively, convex functions are bowl-shaped, i.e. they curve upwards. For instance, the scalar function $f(x) = x^2$ is convex, as shown:



Convexity is an important property in optimization, because it implies that any local minimum of the function is also a global minimum. Furthermore, we can find this minimum using a greedy algorithm which moves downhill, such as gradient descent.

Convex functions can also be characterized in terms of the second derivative when it exists. For a scalar function f, if f'' exists and is nonnegative everywhere, then f is convex. (Although this is not required for this problem, this criterion extends to functions of vectors as well. If the Hessian matrix of f exists everywhere and is positive semidefinite, then f is convex.)

(a) Let q_x and p_x be two distributions over the set $\{1, \ldots, k\}$, parameterized by:

$$q_i = q_x(i)$$
$$p_i = p_x(i)$$

Show that the information divergence $D(q_x \| p_x)$ is convex with respect to q_x, holding p_x fixed.

(b) Now suppose we have an undirected graphical model $\mathcal{G}$ where all of the variables $\mathbf{x} = (x_1, \ldots, x_N)$ take values in $\{0, 1\}$. Let $p_{\mathbf{x}}$ denote the joint distribution defined by the network. Let q be an approximating distribution fully factorized across nodes, i.e. $q_{\mathbf{x}}(\mathbf{x}) = \prod_{i \in \mathcal{V}} b_i(x_i)$. Show that the information divergence $D(q_{\mathbf{x}} \| p_{\mathbf{x}})$ is not necessarily convex with respect to $\boldsymbol{b} = (b_1(1), \ldots, b_N(1))^T$ by giving a counterexample.

(c) Briefly explain why parts (a) and (b) are not contradictory.

Problem 7.2

In this problem, you will work with an Ising model on a torroidal grid graph. Recall that an Ising model defined on a graph $(\mathcal{V}, \mathcal{E})$ with binary variables $x_i \in \{-1, +1\}$ is described by the factorization

$$p(\mathbf{x}; \theta) \propto \exp \left\{ \sum_{i \in \mathcal{V}} \theta_i x_i + \sum_{(j,k) \in \mathcal{E}} \theta_{jk} x_j x_k \right\}$$

and also recall that a torroidal grid graph is a grid graph with cyclic boundaries, as shown in Figure 7.1.

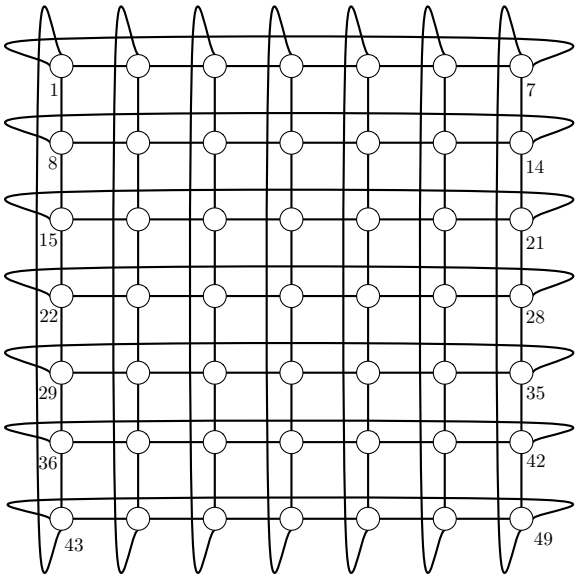


Figure 7.1

(a) In general, the mean field updates for a fully factorized approximation are given

by:

$$b_i(x_i) = K \exp\left\{ \sum_{\substack{C \in \mathcal{C} \text{ s.t. } \\ i \in C}} \sum_{x_C \setminus \{x_i\}} \log \psi_C(x_C) \prod_{\substack{j \in C \\ j \neq i}} b_j(x_j) \right\} \tag{1}$$

Use the general result to find the mean field updates for our Ising model assuming an approximating model that is fully factorized across nodes, i.e. $q(\mathbf{x}) = \prod_{i \in \mathcal{V}} b_i(x_i)$.

(b) It is a useful exercise to derive the mean field updates for this special case. Derive the mean field updates of part (a) for $b_1(x_1)$ by writing out the appropriate information divergence and differentiating with respect to $b_1(x_1)$.

(c) Now we want to implement the algorithm in MATLAB® on a particular model instance. Set $\theta_{jk} = 0.25$ for all edges and $\theta_i = (-1)^i$ for all $i \in \mathcal{V} = \{1, \ldots, 49\}$, where nodes are numbered as in Figure 7.1. Calculate the approximated mean for each node, τ_i for each $i \in \mathcal{V}$.

Problem 7.3

In this problem, we consider a variational approximation to a multivariate Gaussian distribution using scalar Gaussian factors. Consider a Gaussian distribution $p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, J^{-1})$ for which we identify the two components of the vector $\mathbf{x} \in \mathbb{R}^2$:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}$$

where $J_{12} = J_{21}$ and each element of J is a scalar. The distribution is also assumed non-degenerate, i.e. J is positive definite.

We want to approximate the distribution p with a factorized distribution $q_{\mathbf{x}}(\mathbf{x}) = b_1(x_1)b_2(x_2)$, for some b_1 and b_2 which we do not a priori constrain to be of a particular class of distributions.

Analogously to the discrete case, if we drop reference to any graphical structure in the distribution (i.e. consider the graph to be fully connected) and consider the two-factor case, we have

$$b_1(x_1) = K' \exp\left\{ E_{b_2}\left[\ln p_{\mathbf{x}}(\mathbf{x}) \right] \right\} \tag{2}$$

where E_{b_2} denotes the expectation with respect to the variational distribution over node 2, i.e.

$$E_{b_2}\left[f(x_2) \right] = \int f(x_2) b_2(x_2) \, dx_2$$

Note that we have replaced the unnormalized potential ψ with the joint probability distribution $p_{\mathbf{x}}$, which only affects the normalization constant K'. The expression for

the optimal update to b_2 is analogous to Eq.(2) with the factor b_1 switched in to the expression.

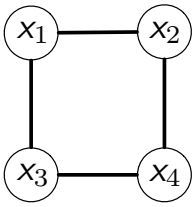
(a) Using the formula in Eq.(2), derive the optimal mean field updates for factors b_1 and b_2. You should show that the updated factors are themselves Gaussian densities, and use that observation and the technique of completing the square to find the normalization constant K' by comparison to the standard Gaussian form.

(b) Show that the update equations are satisfied when the variational mean is equal to the true mean. Discuss why the optimal variational approximation makes intuitive sense.

Problem 7.4 (Practice)

(a) In our discussion of variational inference, we minimized $D(b_{\mathbf{x}} \| p_{\mathbf{x}})$ with respect to $b_{\mathbf{x}}$. In this part, we consider the other direction. Suppose that calculating the marginals of $p_{\mathbf{x}}(\mathbf{x})$ is computationally intractable. Could it be tractable to find a fully factorized approximating distribution $b(\mathbf{x}) = \prod_{i=1}^{N} b_i(x_i)$ by minimizing $D(p_{\mathbf{x}} \| b_{\mathbf{x}})$ with respect to $b_{\mathbf{x}}(\mathbf{x})$?

For the remainder of this question, we return to the usual variational objective function $D(b_{\mathbf{x}} \| p_{\mathbf{x}})$. We restrict our attention to the following undirected graphical model:



Associated with each node in the graph is a random variable x_i, and the joint distribution for these variables is of the form

$$p_{x_1,x_2,x_3,x_4}(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \psi_{i,j}(x_i, x_j) \tag{3}$$

where Z is the partition function, $\psi_{i,j}(x_i, x_j)$ is a clique potential for edge (i, j), and the set of edges is $\mathcal{E} = \{(1, 2), (1, 3), (2, 4), (3, 4)\}$.

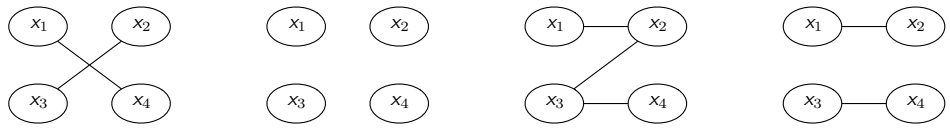
(b) Determine whether the following statement is TRUE or FALSE. If you answer TRUE, be sure to provide a proof; if you answer FALSE, be sure to provide a counterexample.

STATEMENT: Consider the family of approximating distributions

$$b_{\mathbf{x}}(\mathbf{x}) = \prod_{(i,j)\in\mathcal{E}_b} b_{i,j}(x_i, x_j)$$

for some alternative edge set $\mathcal{E}_b$. Then the optimal variational approximation in this family has the property that $b_{i,j}(x_i, x_j) \propto \psi_{i,j}(x_i, x_j)$ for all $(i,j) \in \mathcal{E}_b \cap \mathcal{E}$.

(c) Consider families of approximating distributions $b_{\mathbf{x}}(\mathbf{x})$ described by the four graphical models shown below.



I II III IV

Suppose that we have a black-box procedure that can obtain the best variational approximations within each of these families. Rank the graphs in descending order of the quality of their resulting approximations as obtained from the black-box procedure, and explicitly specify when there is a tie. Be sure to fully justify your reasoning.

Hint: If graph $\mathcal{G}_1$ is contained in $\mathcal{G}_2$, try deriving the variational inference update rule for the graph $\mathcal{G}_2$. Argue that if the updates are contained in the set of approximating distributions given by $\mathcal{G}_1$, then there is a tie.

Problem 7.5 (Practice)

Consider the distribution over binary variables $x_i \in \{0, 1\}$ given by

$$p(\mathbf{x}) \propto 1 - \prod_{i=1}^{n} x_i \tag{4}$$

(a) What undirected graphical model on n nodes is a minimal I-Map for this distribution?

(b) To develop some intuition for approximations based on minimizing information divergence, consider approximating the distribution p in Eq.(4) with a fully factorized distribution with only one parameter, i.e. $q(\mathbf{x}) = \prod_{i=1}^{n} q(x_i)$ (noting that there is only one $q(\cdot)$ function). Show that there is only one unique minimizer for the expression $\mathrm{D}(q||p)$ and describe its form.

(c) Now consider approximating the distribution p from Eq.(4) with a fully factorized distribution in which there are different parameters at each node, i.e. $q(\mathbf{x}) = \prod_{i=1}^{n} q_i(x_i)$. Show that there will be n symmetric minimizers of the $\mathrm{D}(q||p)$ expression and describe them.

(d) Note that the class of approximating distributions in both parts (b) and (c) correspond to fully disconnected graphs, i.e. graphs with empty edge sets. Suppose now that we allow our approximating distribution q to be represented as a product over potentials on maximal cliques for some (not fully-connected) graph. Describe the form of the clique potentials on the graph that would minimize the quantity $\mathrm{D}(q||p)$.

6.438 Algorithms for Inference

Fall 2014