CS 228T Problem Set 2

April 22, 2011

Instructions. The lengths listed for each problem are suggested *maximum* lengths for typed solutions, not minimum; solving the problems fully in less space is possible. Some questions may be related to published research papers, so do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you work in groups, indicate in your solutions who you worked with.

1. *Inference with the convex-concave procedure* (13 points, 2 pages). A main difficulty with using loopy belief propagation to solve the Bethe approximation to the exact inference problem is that it often has trouble converging. In this question, we explore the use of a different algorithm that is guaranteed to converge.

   Sequential convex programming (SCP) refers to a class of algorithms for nonconvex optimization that involves solving a sequence of convex approximations to the nonconvex problem. We consider a particular SCP algorithm for problems in the form

   $$\begin{array}{ll} \text{minimize} & f(x) - g(x) \\ \text{subject to} & x \in \mathcal{C}, \end{array}$$

   where f and g are convex functions and $\mathcal{C}$ is a convex set. Problems of this form, sometimes called 'difference of convex' optimization problems, need not be convex because the objective function involves the sum of a convex and a concave function.

   The *convex-concave procedure* (CCCP) approaches such problems by replacing the objective with the following convex upper bound at iteration k:

   $$\hat{f}_k(x) = f(x) - g(x^k) - \nabla g(x^k)^T (x - x^k),$$

   i.e., by linearizing the concave term $-g$ at x^k. The algorithm can then be written

   $$x^{k+1} := \operatorname*{argmin}_{x \in \mathcal{C}} \hat{f}_k(x) = \operatorname*{argmin}_{x \in \mathcal{C}} \left(f(x) - \nabla g(x^k)^T x \right).$$

   (a) Let $G = (V, E)$ be a cluster graph for a graphical model with factors Φ. Derive a CCCP-based algorithm for solving the approximate inference problem

   $$\begin{array}{ll} \text{maximize} & \sum_{i \in V} \mathbb{E}_{\beta_i}[\log \psi_i] + \sum_{i \in V} H(\beta_i) - \sum_{(i,j) \in E} H(\mu_{ij}) \\ \text{subject to} & q \in \mathbb{L}(G), \end{array}$$

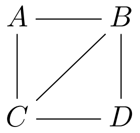
   where $q = \{\beta_i\}_{i \in V} \cup \{\mu_{ij}\}_{(i,j) \in E}$ and $\mathbb{L}(G)$ is the local consistency polytope for G.

(b) It is possible to apply CCCP in a more refined way than simply linearizing an entire summation in the problem above. For example, suppose G is the Bethe cluster graph for factors Φ and variables $X_1, \ldots, X_n$. Let β_ϕ denote the belief for the cluster corresponding to factor ϕ, and let β_i be the belief for variable X_i. In this case, the problem above can be rewritten

$$\begin{array}{ll} \text{maximize} & \sum_{\phi \in \Phi} \mathbb{E}_{\beta_\phi}[\log \phi] + \sum_{\phi \in \Phi} H(\beta_\phi) - \sum_{i=1}^{n} (d_i - 1) H(\beta_i) \\ \text{subject to} & q \in \mathbb{L}(G), \end{array}$$

where d_i is the number of factors with X_i in scope. In other words, this is an alternate way of expressing the Bethe approximation.

Consider the particular instance of this problem corresponding to the Bethe cluster graph for the pairwise Markov random field given by



Explain how we can use CCCP to solve this problem instance while only approximating a single term. (For example, linearize only $-H(X)$ for some single variable X.)

2. *Structured variational methods* (18 points, 2 pages). Consider the structured variational approximation of equation 11.61. As discussed, to execute the update, we need to collect the expectation of several factors, and each of these requires that we compare expectations given different assignments to the factor of interest.

Specifically, consider the case of the chain-structured variational approximation for the $n \times n$ grid, as illustrated in Figure 11.17a.

Show how we can use a dynamic programming algorithm to reuse computation so as to evaluate these *asynchronous* updates more efficiently. (Here, 'asynchronous' refers to the fact that we update a single ψ_j at a time, then use its updated values when we move on to the next step, which involves updating some ψ_k for $k \neq j$.)

(This problem is based on exercise 11.27.)

3. *Cluster variational methods* (18 points, 2 pages). Do exercise 11.29.

4. *Nonconvexity of mean field* (10 points, 1 page). Consider the Ising model over the undirected graph $G = (V, E)$.

(a) Let $\mathcal{Q}$ be the feasible set for the naïve mean field problem, *i.e.*, the set of fully factored distributions. Using the convex hull characterization of the marginal polytope from question 6 on the first problem set, show that $\mathcal{Q}$ is a subset of the marginal polytope containing all its extreme points.

(b) Use part (a) to show that the mean field problem is nonconvex.

5. *Graph cuts for MAP inference* (16 points, 3 pages).

(a) Do exercise 13.14.

(b) Do exercise 13.15.

6. *Suboptimality bounds for α-expansion* (13 points, 2 pages). Let $\mathcal{X}$ be a pairwise metric Markov random field over a graph $G = (V, E)$. Suppose that the variables are nonbinary and that the node potentials are nonnegative. Let $\mathcal{A}$ denote the set of labels for each $X \in \mathcal{X}$. Though it is not possible to (tractably) find the globally optimal assignment $x^\star$ in general, the α-expansion algorithm provides a method for finding assignments $\hat{x}$ that are locally optimal with respect to a large set of transformations, *i.e.*, the possible α-expansion moves.

   Despite the fact that α-expansion only produces a locally optimal MAP assignment, it is possible to prove that the energy of this assignment is within a known factor of the energy of the globally optimal solution $x^\star$. In fact, this is a special case of a more general principle that applies to a wide variety of algorithms, including max-product belief propagation and more general move-making algorithms: If one can prove that the solutions obtained by the algorithm are 'strong local minima', *i.e.*, local minima with respect to a large set of potential moves, then it is possible to derive bounds on the (global) suboptimality of these solutions, and the quality of the bounds will depend on the nature of the moves considered. (There is a precise definition of 'large set of moves'.)

   Consider the following approach to proving the suboptimality bound for α-expansion.

   (a) Let $\hat{x}$ be a local minimum with respect to expansion moves. For each $\alpha \in \mathcal{A}$, let

   $$V^\alpha = \{s \in V \mid x_s^\star = \alpha\},$$

   i.e., the set of nodes labelled α in the global minimum. Let x' be an assignment that is equal to $x^\star$ on V^α and equal to $\hat{x}$ elsewhere; this is an α-expansion of $\hat{x}$. Verify that

   $$E(x^\star) \leq E(\hat{x}) \leq E(x').$$

   (b) Building on the previous part, show that

   $$E(\hat{x}) \leq 2cE(x^\star),$$

   where

   $$c = \max_{(s,t) \in E} \left(\frac{\max_{\alpha \neq \beta} \varepsilon_{st}(\alpha, \beta)}{\min_{\alpha \neq \beta} \varepsilon_{st}(\alpha, \beta)} \right)$$

   and E denotes the energy of an assignment.

   Hint. Think about where x' agrees with $\hat{x}$ and where it agrees with $x^\star$.

7. *Global optimality and max-product belief propagation* (12 points, 2 pages). Do exercise 13.11.