CS 228T Problem Set 1

April 8, 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. *Annealed importance sampling* (18 points, 2 pages). Do exercise 12.25 from the book.

   Depending on the edition of the textbook you have, there may be a small typo in part (b) of the question, so we have reproduced it here. Define

   $$f^*(\mathbf{x}_1, \ldots, \mathbf{x}_k) = f_0(\mathbf{x}_1) \prod_{i=1}^{k-1} \mathcal{T}_i^{-1}(\mathbf{x}_i \to \mathbf{x}_{i+1}),$$

   and define $p^*(\mathbf{x}_1, \ldots, \mathbf{x}_k) \propto f^*(\mathbf{x}_1, \ldots, \mathbf{x}_k)$. Use your answer from part (a) to show that $p^*(\mathbf{x}_1) = p(\mathbf{x}_1)$.

2. *Sampling for the correspondence problem* (18 points, 1 page). Let $G = (U, V, E)$ be an undirected weighted bipartite graph, where U is the first set of nodes, V is the second set of nodes, and E is the set of edges connecting nodes in U to nodes in V. The weight of an edge is given by $w(u, v) \geq 0$. Suppose the graph is fully connected, so $E = U \times V$, and that $|U| = |V| = n$.

   An *assignment* A is a set of n edges where each vertex is incident to exactly one edge. Let $A(u)$ denote the node in V such that $(u, A(u)) \in A$. Let

   $$p(A) \propto \exp\left\{ -\sum_{u \in U} w(u, A(u)) \right\}$$

   be a distribution over assignments A. This question examines proposal distributions $q(A \to A')$ for sampling from p using the Metropolis-Hastings algorithm. The most obvious proposal distribution takes two matched edges (u_1, v_1) and (u_2, v_2) in A at random and swaps the assignments, but this method often mixes extremely slowly, so we consider the alternative method below.

   Let A be some initial assignment. We describe a particular method of traversing edges in G, and will denote this sequence of edges as R.

(a) Pick a random node $u \in U$.

(b) Traverse the edge (u, v) according to the transition probabilities

$$r(u, v) = \frac{\exp(-w(u, v))}{\sum_v \exp(-w(u, v))}.$$

(c) The node v was previously matched to some $u' \in U$. Traverse the previous matched edge (v, u').

(d) Go back to (b) and repeat until a cycle is formed, ending at some u'' visited earlier.

(e) It is possible that u'' is not the same as u, in which case the path R looks like a cycle with a 'tail' hanging off the end; the tail is a path between u'' and the initial node u. If we erase this tail, a cycle C remains. This cycle is *alternating* in the sense that it alternates between edges that are missing and edges that are present in A. Then the proposal is $A' = A \oplus C$, where $\oplus$ denotes symmetric difference.

For example, suppose the sequence of nodes traversed is

$$u_1 \to v_1 \to u_2 \to v_2 \to u_3 \to v_3 \to u_4 \to v_1 \to u_2.$$

In this case, C is $u_2 - v_2 - u_3 - v_3 - u_4 - v_1 - u_2$, and the tail is $u_1 - v_1$.

Note that it is possible for R to be a path through the graph that does not result in the assignment changing at all. This is because, in step (b), the selected v may be the one that was already matched to u. In this case, R is the path of traversed edges, while C is the empty set, so $A' = A \oplus C = A$. In other cases, C may be equal to R.

Show that the acceptance ratio

$$\alpha = \frac{p(A')}{p(A)} \cdot \frac{q(A' \to A)}{q(A \to A')}$$

for this method is always 1.

Hint. Try to express the ratio $p(A')/p(A)$ in terms of $r(u, v)$ to have factors cancel with factors in $q(A' \to A)$ and $q(A \to A')$.

3. *Auxiliary variable methods and log-linear models* (10 points, 1 page). Recall that in auxiliary variable methods, we define a space of auxiliary variables u with conditional distribution $p(u \mid x)$, then sample from the joint distribution $\pi(x, u) = \pi(x)p(u \mid x)$ by alternating the transition

$$T^U((x, u) \to (x, u')) = p(u' \mid x)$$

with some transition $T^X((x, u) \to (x', u))$ that satisfies detailed balance with respect to $\pi(x, u)$. Because $\pi(u, x) = \pi(x)p(u \mid x)$, we can then throw away the u component from the resulting sample.

Suppose X is a pairwise Markov random field with graph $G = (V, E)$ and distribution

$$\pi(X) \propto \pi_0(X) \exp\left\{ \sum_{(i,j) \in E} \sum_{k,l} \beta_{ij}^{kl} f_{ij}^{kl}(X_i, X_j) \right\}$$

where $\pi_0(X) = \prod_i \phi_i(X_i)$ is a product of singleton factors, $\beta_{ij}^{kl} > 0$, and

$$f_{ij}^{kl}(X_i, X_j) = 1[X_i = x_i^k, X_j = x_j^l],$$

where x_i^k denotes the kth outcome for variable X_i.

We now sample each u_{ij}^{kl} independently uniformly in the interval $[0, \exp(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j))]$, *i.e.*,

$$p(u_{ij}^{kl} \mid x) = \exp(-\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)) \cdot 1[0 \leq u_{ij}^{kl} \leq \exp(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j))].$$

Suppose we think of each indicator as a constraint that is inactive when it evaluates to zero and active when it evaluates to one. The original distribution $\pi(X)$ can then be viewed as a simple distribution π_0 with additional softened constraints mixed in.

(a) How does f_{ij}^{kl} being active or inactive affect the samples u_{ij}^{kl}?

(b) Explicitly characterize the sampling distribution $\pi(X \mid u)$ and state a condition on u_{ij}^{kl} that implies that the sampled x must satisfy the constraint specified by f_{ij}^{kl}. Your expression for $\pi(X \mid u)$ should be in terms of the ϕ_i, β_{ij}^{kl}, and u_{ij}^{kl}. Be sure to specify both the proposal and the acceptance probability.

(c) If the MRF is an Ising model, we can represent it as a log-linear model in two equivalent ways: either by using a single function

$$f_{ij}(X_i, X_j) = 1[X_i = X_j]$$

for each edge, or by using a set of distinct functions

$$f_{ij}^k(X_i, X_j) = 1[X_i = x_i^k, X_j = x_j^k],$$

where all these functions share the same coefficient β_{ij}. If we use this auxiliary variable sampling strategy, the first model gives rise to the Swendsen-Wang algorithm, and the second to the sampling approach you derived in the previous parts. Which is likely to give rise to better mixing in practice?

4. *Approximating the marginal polytope* (18 points, 2 pages). Consider the local consistency polytope defined in §11.3.6.

(a) Show that, for any clique tree, the local consistency polytope is equal to the marginal polytope. (You are allowed to cite any theorems from the textbook; the proof need not be 'from scratch'.)

(b) Show that, for some cluster graph that is *not* a clique tree, the marginal polytope is *strictly* contained in the local consistency polytope. In other words, give an example of a parameterization for a graphical model that satisfies the local consistency constraints but does not correspond to any valid probability distribution.

(c) The marginal polytope is defined by the intersection of a large number of linear constraints. The local consistency polytope is an approximation obtained by taking a small subset of these constraints; geometrically, the local consistency polytope is an 'outer bound' to the marginal polytope. A natural approach to improving this approximation

is to add more constraints that hold for any element of the marginal polytope, thus tightening the outer bound.

One such class of constraints is the set of *cycle inequalities*. Consider a graph $G = (V, E)$ with n vertices corresponding to binary random variables. Given an assignment $x \in \{0,1\}^n$, we say that $(i,j) \in E$ is *cut* if $x_i \neq x_j$.

i. Show that any cycle in G must have an even (possibly zero) number of cut edges.

ii. Using part (i), show that for any cycle C and any $F \subseteq C$ with $|F|$ odd, we have

$$\sum_{(i,j)\in C-F} 1[x_i \neq x_j] + \sum_{(i,j)\in F} 1[x_i = x_j] \geq 1.$$

(Both C and F are sets of *edges* in G, so $|F|$ being odd means that F contains an odd number of edges. Also, note that F need not be a cycle.)

iii. Let $G = (V, E)$ be a pairwise binary Markov random field. Consider a standard Bethe cluster graph for G, so the large clusters are over pairs (X_i, X_j) for edges $(i,j) \in E$. Use the property above to write down a set of constraints on the pseudo-marginals $\beta_{ij}(X_i, X_j)$ that must hold for every distribution Q.

Note. The standard Bethe cluster graph discussed in the book also includes singleton clusters, but you should ignore those for the purposes of this question.

5. *Region graphs and generalized belief propagation* (18 points, 1 page). We will investigate a class of approximations to the exact inference problem that includes Bethe approximation as a special case.

Let Φ be a set of factors for a graphical model over a set of variables X. A *region graph* $R = (V, E)$ is a directed graph where each vertex $r \in V$, also called a *region*, is associated with a set of variables $C_r \subseteq X$ with outcome space $\mathcal{C}_r$, and has a *counting number* $\kappa_r \in \mathbb{R}$. Below, $c_r \in \mathcal{C}_r$ will denote an assignment to C_r. If $s \to r \in E$, then C_r is a subset of C_s. Each factor $\phi \in \Phi$ is associated with a set of regions $\alpha(\phi)$; each $r \in \alpha(\phi)$ must contain the scope of ϕ.

Let $U(r)$, $U^*(r)$, $D(r)$, and $D^*(r)$ denote the parents, ancestors, children, and descendants, respectively, of r in R, and let $D^+(r) = \{r\} \cup D^*(r)$. There are some properties a valid region graph must satisfy. The regions with a given variable X_i in scope must form a connected component, and the counting numbers across the regions in the component must sum to 1; the same holds for the set of regions $\alpha(\phi)$ for each $\phi \in \Phi$.

These conditions can be enforced by requiring that

$$\kappa_r = 1 - \sum_{s \in U^*(r)} \kappa_s,$$

which ensures that the sum of the counting numbers of r and its ancestors will be 1. Finally, for each X_i, there must be a unique region r_i such that every other region with X_i in scope is an ancestor of r_i; a similar requirement must hold for each ϕ.

The energy functional associated with a region graph R is

$$\tilde{F}[\tilde{P}, Q] = \sum_{r \in V} \kappa_r \mathbb{E}_{\beta_r}[\log \psi_r] + \sum_{r \in V} \kappa_r H(\beta_r),$$

where ψ_r is the product of the factors assigned to r and β_r is the belief over C_r.

Consider the optimization problem

$$\begin{array}{lll} \text{maximize} & \tilde{F}[\tilde{P}, Q] & \\ \text{subject to} & \sum_{c_r} \beta_r(c_r) = 1 & \forall r \in V \\ & \sum_{c_s \sim c_r} \beta_s(c_s) = \beta_r(c_r) & \forall (s \to r) \in E,\ c_r \in \mathcal{C}_r \\ & \beta_r(c_r) \geq 0 & \forall r \in V,\ c_r \in \mathcal{C}_r, \end{array}$$

with variable $Q = \{\beta_r \mid r \in V\}$. This problem reduces to the usual Bethe approximation for particular choices of regions and counting numbers.

(a) Let G be a pairwise Markov random field in the form of a 3×3 grid. Consider a region graph with 8 regions, where the top layer consists of the four 2×2 grids in G and their children are the two-variable sepsets. Suppose the four top regions have counting number 1 and the four sepsets have counting number -1. Is this a valid region graph for G? If not, modify this construction so the resulting region graph is valid.

(b) Show what fixed point equations a solution must satisfy by forming the Lagrangian and differentiating with respect to $\beta_r(c_r)$.

Note. This question is based on material in §11.3.7 if you want further background.

6. *Exponential families and the marginal polytope* (18 points, 2 pages). In this question, we explore a generalization of the concept of the marginal polytope that holds for any model in the exponential family. This provides an alternative perspective on the marginal polytope and provides some additional geometric intuition.

Let X be a random variable with outcome space $\mathcal{X}$. Let $\mathcal{P}$ be a linear exponential family with sufficient statistics $\tau_\alpha : \mathcal{X} \to \mathbb{R}$, $\alpha = 1, \ldots, K$, and overall sufficient statistics function $\tau : \mathcal{X} \to \mathbb{R}^K$, where $\tau(x) = (\tau_1(x), \ldots, \tau_K(x))$. Members of $\mathcal{P}$ then take the form

$$p_\theta(x) = \exp\{\langle \theta, \tau(x) \rangle - A(\theta)\},$$

where $A(\theta) = \log Z(\theta)$. The auxiliary measure is assumed to be a constant.

Let p be any probability distribution on $\mathcal{X}$, not necessarily in $\mathcal{P}$. The *mean parameter* μ_α associated with a sufficient statistic τ_α is defined as the expectation

$$\mu_\alpha = \mathbb{E}_p[\tau_\alpha(X)]$$

for each $\alpha = 1, \ldots, K$. We can then define the vector of mean parameters $\mu = (\mu_1, \ldots, \mu_K)$ with respect to an arbitrary distribution p. The set of *realizable mean parameters*

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^K \mid \exists p : \mathbb{E}_p[\tau_\alpha(X)] = \mu_\alpha,\ \alpha = 1, \ldots, K\right\}$$

is the set of all expected sufficient statistics that can be obtained for some p.

(a) Show that $\mathcal{M}$ is a convex subset of $\mathbb{R}^K$. Here, X may be a continuous variable.

(b) Suppose $\mathcal{X}$ is finite. Show that $\mathcal{M}$ is the convex hull of $\{\tau(x) \mid x \in \mathcal{X}\}$.

(c) Any discrete Markov random field with graph $G = (V, E)$ can be represented as a linear exponential family, taking the sufficient statistics to be indicator functions of the (finitely many) possible outcomes in each clique. For example, the sufficient statistics for the (pairwise) Ising model would be $1[X_s = \alpha]$ for $s \in V$ and $\alpha \in \{0, 1\}$ and $1[(X_s, X_t) = (\alpha, \beta)]$ for all $(s, t) \in E$ and $\alpha, \beta \in \{0, 1\}$. This is sometimes called the *standard overcomplete representation*.

Explain why $\mathcal{M}$ reduces to the marginal polytope for this class of models.