Homework 6

Assigned: 31/10/2011.
Due: 11/11/2011.

Exercise 1. Pseudomarginals and Marginals
Recall the definition of marginal polytope $\mathcal{M}$ and local normalization polytope $\mathcal{M}_{\text{local}}$, and that $\mathcal{M} \subset \mathcal{M}_{\text{local}}$. Make sure you understand that both depend on the graph of a model only. Since LBP operates with *pseudomarginals* $\boldsymbol{\tau} \in \mathcal{M}_{\text{local}}$ only, it can converge to values that are not globally consistent: $\boldsymbol{\tau} \notin \mathcal{M}$, meaning that there is no distribution that exhibits $\boldsymbol{\tau}$ as marginals. In this exercise, you will show that even for very simple graphs (that are not trees), $\mathcal{M}_{\text{local}} \setminus \mathcal{M}$ is not empty.

1. **[1 point]** Consider the three-cycle: three binary nodes with values $\{0, 1\}$, three edges, and some $\boldsymbol{\tau}$ with $\tau_s(x_s) = 1/2$, $s = 1, 2, 3$, moreover $\tau_{st}(x_s, x_t) = \tau_{st} \in [0, 1/2]$ for $x_s = x_t$, $\tau_{st}(x_s, x_t) = 1/2 - \tau_{st}$ otherwise ($x_s \neq x_t$). There are three degrees of freedom τ_{12}, τ_{23}, τ_{13}. Show that for any assignment $\tau_{st} \in [0, 1/2]$: $\boldsymbol{\tau} \in \mathcal{M}_{\text{local}}$.

2. **[3 points]** Set $\tau_{12} = \tau_{23} = 0.4$ and $\tau_{13} = 0.1$. We should be dubious about this assignment: x_1, x_2 and x_2, x_3 are encouraged to be the same, but x_1, x_3 to be different. Prove that $\boldsymbol{\tau}$ is not in $\mathcal{M}$.
Hint: Here is a nice idea, a generalization of which you will learn about in the next lecture. Let $\boldsymbol{y}$ be the vector $(1\, x_1\, x_2\, x_3)^T$, and $\boldsymbol{M} := \mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T]$ for some distribution over $\boldsymbol{x}$. Then, $\boldsymbol{M}$ is positive semidefinite (explain why). Use proof by contradiction. If $\boldsymbol{\tau}$ defined above was in $\mathcal{M}$, there would be a distribution $Q(\boldsymbol{x})$ realizing it. Express $\boldsymbol{M}$ in terms of $\boldsymbol{\tau}$, and prove that it is not positive definite, which is a contradiction (argue why).

Exercise 2. Tree-reweighted Sum-Product
Recall the variational formulation of inference:

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[\boldsymbol{\mu}] \right\}, \quad \mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}.$$

While this is a convex optimization problem, it is not tractable in general, and we have to use relaxations. Some of the most frequently used relaxations, such as structured mean field or LBP, are not convex. In this exercise, you will learn about a convex inference relaxation (more about this in the next lecture). The ingredients are

- A tractable, concave upper bound to the entropy $\mathrm{H}[\boldsymbol{\mu}]$ (which itself is concave, but not tractable)
- A tractable, convex outer bound to the marginal polytope $\mathcal{M}$ (which itself is convex, but not tractable)

Although the method is more general, we'll be concerned here with pairwise MRFs over discrete variables:

$$P(\boldsymbol{x}) = Z^{-1} \exp\left(\sum\nolimits_{s \in V} \theta_s(x_s) + \sum\nolimits_{(s,t) \in E} \theta_{st}(x_s, x_t) \right).$$

Make sure to understand this notation. V is the set of variables (nodes), E the set of edges (cliques). The feature mapping (called $\boldsymbol{f}(\boldsymbol{x})$ in the lecture) consists of indicators $\boldsymbol{x} \mapsto \mathrm{I}_{\{x_s=\tilde{x}\}}$ for all $\tilde{x}$, $s \in V$, and $\boldsymbol{x} \mapsto \mathrm{I}_{\{x_s=\tilde{x},x_t=\tilde{x}'\}}$ for all $\tilde{x}$, $\tilde{x}'$, $(s,t) \in E$, and in $\theta_s(x_s)$, $\theta_{st}(x_s,x_t)$, x_s and x_t are indices into the overall vector $\boldsymbol{\theta}$. Therefore,

$$\boldsymbol{\theta}^T \boldsymbol{f}(\boldsymbol{x}) = \sum\nolimits_{s\in V} \theta_s(x_s) + \sum\nolimits_{(s,t)\in E} \theta_{st}(x_s,x_t),$$
$$\boldsymbol{\theta}^T \boldsymbol{\mu} = \sum\nolimits_{s\in V} \sum\nolimits_{x_s} \theta_s(x_s)\mu_s(x_s) + \sum\nolimits_{(s,t)\in E} \sum\nolimits_{x_s,x_t} \theta_{st}(x_s,x_t)\mu_{st}(x_s,x_t).$$

The graph is given by $G = (V,E)$. Assume that it is connected, but not a tree (otherwise we do not have to approximate inference). We need the following result, which you do not have to prove. Let $\tilde{G}$ be a subgraph of G, in the sense that $\tilde{G} = (V,\tilde{E})$, $\tilde{E} \subset E$, and let $\mathcal{M}_G$ and $\mathcal{M}_{\tilde{G}}$ be the marginal polytopes w.r.t. either graph. If $\boldsymbol{\mu} \in \mathcal{M}_G$, define $\boldsymbol{\mu}_{\tilde{G}}$ as projection onto $\tilde{G}$'s structure, meaning that $\boldsymbol{\mu}_{\tilde{G}}$ consists of $\mu_s(\cdot)$, $s \in V$, and $\mu_{st}(\cdot,\cdot)$, $(s,t) \in \tilde{E}$, then $\boldsymbol{\mu}_{\tilde{G}} \in \mathcal{M}_{\tilde{G}}$ (this is obvious). The real statement is this: for $\boldsymbol{\mu} \in \mathcal{M}_G$,

$$\mathrm{H}[\boldsymbol{\mu}] \leq \mathrm{H}[\boldsymbol{\mu}_{\tilde{G}}]. \tag{1}$$

Make sure to understand what $\mathrm{H}[\boldsymbol{\mu}_{\tilde{G}}]$ means (from the lecture): there is a unique distribution on the graph $\tilde{G}$ (not on G!) whose moments are $\boldsymbol{\mu}_{\tilde{G}}$, and $\mathrm{H}[\boldsymbol{\mu}_{\tilde{G}}]$ is its entropy. So, (1) is really about two distributions, one on G (with moments $\boldsymbol{\mu}$) and one on $\tilde{G}$ (with moments $\boldsymbol{\mu}_{\tilde{G}}$). You know a special case of this: if $\tilde{E} = \emptyset$, then $\mathrm{H}[\boldsymbol{\mu}_{\tilde{G}}] = \sum_{s\in V} \mathrm{H}[\mu_s] \geq \mathrm{H}[\boldsymbol{\mu}]$.

1. **[4 points]** G is not a tree, but since it is connected, it contains spanning trees (subgraphs over all nodes V that are trees). Most graphs with many cycles have very many embedded spanning trees. Let $\mathcal{F} = \{F\}$ be a set of spanning trees F embedded in G (need not be all of them, but $\mathcal{F} \neq \emptyset$), and $\rho(F)$ a distribution over $\mathcal{F}$: $\rho(F) \geq 0$ and $\sum_{F\in\mathcal{F}} \rho(F) = 1$. Show that for any $\boldsymbol{\mu} \in \mathcal{M}$: $\mathrm{H}[\boldsymbol{\mu}] \leq \sum_{F\in\mathcal{F}} \rho(F)\mathrm{H}[\boldsymbol{\mu}_F]$, and that $\boldsymbol{\mu} \mapsto \sum_{F\in\mathcal{F}} \rho(F)\mathrm{H}[\boldsymbol{\mu}_F]$ is concave.
   Hint: Recall that if $\boldsymbol{\mu} \in \mathcal{M}_G$, then $\boldsymbol{\mu}_F \in \mathcal{M}_F$. Use (1). Be careful when arguing why the upper bound is concave. Use the following statements (but prove that they hold before using them): (a) if $\boldsymbol{\mu}_F \mapsto f(\boldsymbol{\mu}_F)$ is concave, then $\boldsymbol{\mu} \mapsto f(\boldsymbol{\mu}_F)$ is concave as well. (b) if $\mathrm{H}[\boldsymbol{\mu}_F]$ are concave for all $F \in \mathcal{F}$ and $\rho(F)$ is a distribution, then $\sum_{F\in\mathcal{F}} \rho(F)\mathrm{H}[\boldsymbol{\mu}_F]$ is concave as well. (c) $\boldsymbol{\mu}_F \mapsto \mathrm{H}[\boldsymbol{\mu}_F]$ is concave (what graph does this situation correspond to?).

2. **[5 points]** You know how $\mathrm{H}[\boldsymbol{\mu}_F]$ looks like, if F is a tree (embedded in G). Write the upper bound $\sum_{F\in\mathcal{F}} \rho(F)\mathrm{H}[\boldsymbol{\mu}_F]$ in terms of single node entropies $\mathrm{H}[\mu_s]$, double node entropies $\mathrm{H}[\mu_{st}]$, and so-called edge appearance probabilities $\rho_{st} = \Pr_{F\sim\rho}[(s,t) \in E_F]$, where $F = (V,E_F)$ (so E_F is the edge set of F). Draw a graph with cycles, state some distribution $\rho(F)$ over embedded spanning trees (draw them as well), and work out these numbers ρ_{st}.
   Make sure you understand that this cannot be done in general with $\mathrm{H}[\boldsymbol{\mu}]$, and that it is the fact that we can write the upper bound like this (a simple function of the local marginals) that makes it tractable.
   Hint: Pull the expectation over $\rho(F)$ inside. It is simpler if, for every tree F, you group together expressions $\mathrm{H}[\mu_{st}] - \mathrm{H}[\mu_s] - \mathrm{H}[\mu_t]$ (you may recognize them, from the lecture, as local negative mutual information terms), this leaves you with an expression that does not depend on counting numbers n_s.

3. **[5 points]** At this point, we have a relaxation which, if we could solve it, would give us an upper bound to $\log Z$. But we are still stuck with $\mathcal{M} = \mathcal{M}_G$. Show that if $\boldsymbol{\mu} \in \mathcal{M}_{\mathrm{local}}$ (the local marginalization polytope from the lecture), then for every tree $F \in \mathcal{F}$: $\boldsymbol{\mu}_F \in \mathcal{M}_F$. Argue why this implies that the entropy upper bound derived above is defined on $\mathcal{M}_{\mathrm{local}}$ (while $\mathrm{H}[\boldsymbol{\mu}]$ itself is defined on $\mathcal{M}$ only). Therefore, if we relax $\boldsymbol{\mu} \in \mathcal{M}$ to $\boldsymbol{\mu} \in \mathcal{M}_{\mathrm{local}}$, we have a tractable convex relaxation. Show that solving

this relaxation leads to an upper bound to $\log Z$.
Hint: Please do this carefully, for your own benefit of understanding. What are the contraints of $\mathcal{M}_F$ for a tree F, how do they relate to $\mathcal{M}_{\text{local}}$? For the upper bound on $\log Z$, go step by step, and explain each inequality: first, upper bound on $\mathrm{H}[\boldsymbol{\mu}]$, for $\boldsymbol{\mu} \in \mathcal{M}$; second, outer bound $\mathcal{M}_{\text{local}}$.
Note: For the meticulous. You may be worried that if $\mathrm{H}[\boldsymbol{\mu}]$ is not defined on $\mathcal{M}_{\text{local}}$, whether we can really do this second step. Couldn't it be that $\mathrm{H}[\boldsymbol{\tau}] > \sum_{F \in \mathcal{F}} \rho(F) \mathrm{H}[\boldsymbol{\tau}_F]$ for some $\boldsymbol{\tau} \in \mathcal{M}_{\text{local}} \setminus \mathcal{M}$, whatever $\mathrm{H}[\boldsymbol{\tau}]$ may be defined as? In fact, Wainwright and Jordan show that $\mathrm{H}[\boldsymbol{\tau}] = -\infty$ for $\boldsymbol{\tau} \notin \mathcal{M}$, so the upper bound is valid everywhere. But this argument is not even necessary, because the relaxation works in two steps. In the first, $\mathrm{H}[\boldsymbol{\mu}]$ is eliminated, in the second, $\mathcal{M}$ is extended. Both are relaxations (lead to inequalities), *however* $\sum_{F \in \mathcal{F}} \rho(F) \mathrm{H}[\boldsymbol{\tau}_F]$ is defined outside of $\mathcal{M}$. Of course, we would like to have it being concave also on $\mathcal{M}_{\text{local}}$ (which it is), but this is just to obtain a *convex* relaxation.

4. **[2 points]** Compare the resulting convex variational problem with the Bethe variational problem. For which values ρ_{st} are they the same? Argue that this can happen only if G is a tree in the first place.

5. **[optional; 5 bonus points]** Prove (1).
Hint: $\boldsymbol{\mu}_{\tilde{G}}$ is a valid moment vector, so is realized by some model on $\tilde{G}$ with potentials $e^{\theta_{st}(x_s, x_t)}$ on $(st) \in \tilde{E}$. But this distribution is also consistent with G: just add further potentials $e^{\theta_{st}(x_s, x_t)}$, $(st) \in E \setminus \tilde{E}$, with $\theta_{st} \equiv 0$, giving rise to a model on G with potential parameters $\boldsymbol{\theta}$. Write down the variational problem for this model, renaming the optimization variables to $\boldsymbol{\tau}$ (to distinguish them from the fixed value $\boldsymbol{\mu}$). Compare $\boldsymbol{\theta}^T \boldsymbol{\tau}$ for $\boldsymbol{\tau} = \boldsymbol{\mu}$ and the maximizer $\boldsymbol{\tau} = \boldsymbol{\tau}_*$. Relate $\boldsymbol{\tau}_*$ to $\boldsymbol{\mu}_{\tilde{G}}$.

This convex problems differs from the Bethe problem "only" by reweighting of entropy terms. Since LBP corresponds to the Bethe problem (at least as far as fixed points are concerned), there is an equivalent *reweighted* sum-product algorithm for solving the relaxation here. Its equations can be deduced just like LBP from the Bethe approximation (recall the handout). Interestingly, all this does not really depend on the set $\mathcal{F}$ or the distribution $\rho(F)$, but only on the vector $\boldsymbol{\rho} = (\rho_{st})$. What then are all valid assignments to this vector? You can find details in the original paper, or the Wainwright and Jordan monograph. The set of all valid $\boldsymbol{\rho}$ vectors is called the *spanning tree polytope*, and has been analyzed in combinatorics.
Are these the only reweightings of LBP that lead to convex relaxations? No, there are many others. We will learn about others in the next lecture. Just as well, there are other convex relaxations of variational inference that do not lead to reweighted LBP algorithms. And the question when to use which relaxation for best results, is pretty much an open problem. This is an ongoing, active part of Bayesian machine learning.