Advanced Methods in ML 2018 - Exercise 3

1. In this exercise we will see a case where the local marginal polytope approximation yields an exact solution, and the graph is not a tree. Consider a set of n binary variables $X_1, \ldots, X_n$, and a graph with edges E that is not necessarily a tree. The MRF will be defined via pairwise functions:

$$\theta_{ij}(x_i, x_j) = \begin{bmatrix} 0 & 0 \\ 0 & s_{ij} \end{bmatrix} \tag{1}$$

for some parameter $s_{ij} > 0$. The singleton functions will be:

$$\theta_i(x_i) = \begin{bmatrix} 0 \\ s_i \end{bmatrix} \tag{2}$$

where $s_i \in \mathbb{R}$ and can be both positive or negative. It turns out that MAP for this problem can be solved with a min-cut algorithm. Here we will show that the LP relaxation we learned in class also solves it. The proof will proceed via the following steps.

   (a) Show that the local marginal polytope relaxation (namely $\max_{\boldsymbol{\mu} \in \mathcal{M}_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta}$) is equivalent to the following LP. The variables of the LP are a scalar τ_{ij} for each edge $ij \in E$, and a scalar τ_i for each variable. The objective is to maximize the function:

$$f(\boldsymbol{\tau}) = \sum_i s_i \tau_i + \sum_{ij} s_{ij} \tau_{ij} \tag{3}$$

   And the constraints are:

$$\begin{aligned} \tau_{ij} &\geq 0 \quad \forall ij \in E \\ \tau_i &\geq 0 \quad \forall i \\ \tau_{ij} &\leq \tau_i \quad \forall ij \in E \\ \tau_{ij} &\leq \tau_j \quad \forall ij \in E \\ \tau_{ij} &\geq \tau_i + \tau_j - 1 \end{aligned}$$

   Hint: the variables τ_{ij} and τ_i correspond to the local marginal polytope variables $\mu_{ij}(1,1)$ and $\mu_i(1)$ respectively.

   (b) We now want to show that the LP above has an optimum that has only values $0, 1$ for the τ variables. To show this, consider a solution τ that has a fractional elements. Define a new solution $\boldsymbol{z}$ as:

$$\begin{aligned} z_i &= \tau_i - \lambda \mathcal{I}(0 < \tau_i < 1) \\ z_{ij} &= \tau_{ij} - \lambda \mathcal{I}(0 < \tau_{ij} < 1) \end{aligned}$$

   where $\lambda = \min_{i:\tau_i > 0} \tau_i$ (note this is a strict inequality, so it is the minimum non-integral value). Show that the new z has less fractional values that τ. Convince yourself (no need to prove explicitly) that defining $\lambda = -\min_{i:\tau_i < 1}(1 - \tau_i)$ will also result in a feasible solution with less fractional values.

(c) Show that using one of the above λ will result in a new solution such that $f(\boldsymbol{\tau}) \leq f(\boldsymbol{z})$.

(d) Conclude that there is always an integral solution to the LP and that this solution is the exact MAP (to make the argument complete, we still need to show that such an integral solution can be found efficiently, but you can assume that's the case).

(e) Now assume the pairwise marginal has four non-zero elements:

$$\theta_{ij}(x_i, x_j) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{4}$$

For simplicity assume they are the same for all $ij \in E$. Assume the singleton terms $\theta_i(x_i)$ can be arbitrary. Show that if $A + D - B - C > 0$ then you can bring this problem to the form with $s_{ij} > 0$ and s_i above, and therefore solve it exactly. Hint: consider transformations on the θ_{ij}, θ_i variables that do not change the maximizer.

2. In this question we will consider importance sampling, and conclude what the optimal proposal distribution q should be. Recall that the importance sampling estimate of $\mathbb{E}_p\left[f(X)\right]$ is the random variable:

$$Z = \frac{1}{T}\sum_{i=1}^{T} \frac{p(X^{(i)})}{q(X^{(i)})} f(X^{(i)}) \tag{5}$$

(a) Show that $\mathbb{E}_{q^n}\left[Z\right] = \mathbb{E}_p\left[f(X)\right]$, where q^n is the distribution sampling n IID samples of $X^{(i)}$ from $q(x)$. In such a case we say that the estimator Z is an unbiased estimate of the true expected value.

(b) Given that Z is correct in expectation, it is interesting to consider its variance. Show that the variance is minimized by the following distribution:

$$q(x) \propto |f(x)|p(x) \tag{6}$$

Hint: Use Jensen's inequality to show that:

- $\mathbb{E}_q\left[f^2(X)\frac{p^2(X)}{q^2(X)}\right] \geq \left(\mathbb{E}_q\left[|f(X)|\frac{p(X)}{q(X)}\right]\right)^2$

Use this to show a lower bound on the variance, and show that the lower bound can be attained with the distribution in Equation 6. Note the interesting conclusion here that p is not the optimal proposal distribution!

3. In this question we will explore the connection between entropy maximization and MRFs.

(a) Given a set of d functions $f_1(x), \ldots, f_d(x)$ and d scalars $a_1, \ldots, a_d$, denote by $q(x)$ a distribution that has maximum entropy among all distributions that satisfy $\mathbb{E}_p\left[f_i(x)\right] = a_i$ for all i. Prove that $q(x)$ is any distribution of the form:

$$q(x) \propto e^{\sum_i \lambda_i f_i(x)} \tag{7}$$

where $\lambda_i \in \mathbb{R}$ are such that $q(x)$ satisfies the expectation constraints. Hint: Use Lagrange multipliers, and ignore the constraint that $p(x)$ is non-negative (explain why that is ok here).

(b) Given a set of marginals $\mu_{ij}(x_i, x_j)$ corresponding to edges $ij \in E$ in a graph, show that among all distributions $p(x_1, \ldots, x_n)$ that have the given pairwise marginals, the one maximizing the entropy is a pairwise MRF. You may use the result from the first part of the question.

4. For a pairwise MRF, show that the log partition-function $\log Z(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$. Hint: show that the Hessian is a PSD matrix by showing that it can be written as $\sum_i z_i z_i^T$ for some vectors z_i (prove that this indeed implies it is PSD).