Advanced Methods in Machine Learning 6.5.2018

Exercise 3

Nir Raviv 200683548 *Roi Tabach 203022983* *Andrey Leshenko 322026527*

nirraviv@mail.tau.ac.il *roi.tabach@gmail.com* *andrey.leshenko@gmail.com*

Q1

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 $s_i \in \mathbb{R}$ can be both positive or negative. We will show that the LP relaxation we learned in class also solves the MAP problem in this case. The proof follows in the next sections.

a

We will show that the local marginal polytope relaxation (namely $max_{\mu \in \mathcal{M}_L} \mu \cdot \theta$) 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:

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

And the constraints are:

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

For the equivalent local marginal polytope problem we will define the following distribution-like function:

$$\mu_i = \begin{bmatrix}\mu_i(0)\\ \mu_i(1)\end{bmatrix} = \begin{bmatrix}1-\tau_i\\ \tau_i\end{bmatrix} \tag{4}$$

$$\mu_{ij} = \begin{bmatrix}\mu_{ij}(0,0) & \mu_{ij}(0,1)\\ \mu_{ij}(1,0) & \mu_{ij}(1,1)\end{bmatrix} = \begin{bmatrix}(1-\tau_i-\tau_j+\tau_{ij}) & (\tau_j-\tau_{ij})\\ (\tau_i-\tau_{ij}) & (\tau_{ij})\end{bmatrix} \tag{5}$$

First, from the way we defined μ we can see that $f(\tau) = \mu \cdot \theta$ (the θ values are expanded into s values, and many values of μ are multiplied by zeros and cancel out), which is the form of local marginal polytope relaxation. For the local marginal polytope, its function μ must satisfy:

1. All elements are non-negative: $\mu \geq 0$.
2. All distributions sum to one: $\sum_{x_i} \mu_i(x_i) = 1$ and $\sum_{x_i,x_j} \mu_{ij}(x_i,x_j) = 1$.
3. The pairwise distributions agree with the singleton ones:$\sum_{x_i} \mu_{ij}(x_i,x_j) = \mu_j(x_j)$

The first two constraints ensured that all τ_i and τ_{ij} are non-negative, the next two constraints ensure that $\tau_i - \tau_{ij}$ and $\tau_j - \tau_{ij}$ are non-negative (moving all terms to the right), and the final constraint ensures that $1 - \tau_i - \tau_j + \tau_{ij}$ is non-negative (moving all terms to the left). We also need to show that $1 - \tau_i > 0$. If $\tau_i > 1$, then from constraints 4 and 5 we get $\tau_{ij} \geq \tau_i + \tau_j - 1 \geq \tau_i + \tau_{ij} - 1 > 1 + \tau_{ij} - 1$ and we get $\tau_{ij} > \tau_{ij}$ which can never happen. This proves that $\mu > 0$ at all values.

We can see that the distributions sum to 1 by summing the values in the tables that define μ_i and μ_{ij} and seeing that they will always sum to 1.

Finally, we need to show that the pairwise distributions agree with the singleton ones, but this is also easily seen from the tables: summing over the values of x_i is equivalent to summing each column of μ_{ij} which is equal to $[1-\tau_j, \tau_j] = \tau_j$. Similarly, summing over the values of x_j (the rows of μ_{ij}) gives us τ_i. The distributions agree.

This proves that this LP is equivalent to the local marginal polytope relaxation.

b

We would like to show that the LP above has an optimum that has only values 0,1 for the τ variables. Given a solution τ that has fractional elements, we will define a new solution z as:

$$\lambda = \min\left\{\min_{i:\tau_i>0} \tau_i, \min_{ij:\tau_{ij}>0} \tau_{ij}\right\}$$
$$z_i = \tau_i - \lambda \mathcal{I}(0 < \tau_i < 1)$$
$$z_{ij} = \tau_{ij} - \lambda \mathcal{I}(0 < \tau_{ij} < 1)$$

We will also define the solution z' where $\lambda = -\min\left\{\min_{i:\tau_i<1}(1-\tau_i), \min_{ij:\tau_{ij}<1}(1-\tau_{ij})\right\}$.

First we need to make sure that z is still a valid solution to the LP problem. From the way we selected λ, all non-negativity constraints still hold. Regarding constraints 3 and

4, if τ_i was decreased, then either τ_{ij} is also fractional and will be decreased by the same amount, or it is and stays zero, preserving the constraint. In constraint 5, if both τ_i and τ_j are integral (4 possible cases) the constraint will hold for z; if one or both of τ_i, τ_j is fractional, then the right hand-side of the equation we will get smaller by λ or 2λ (while the left-hand side gets smaller by 0 or λ) and the constraint still holds.

We will now show that z has less fractional values than τ. First notice that the only values that change are the ones which have fractional parts, the integral values of τ move to z unchanged, ensuring that the number of fractional elements in z is not larger than in τ. Out of the fractional values that do change, one will be the minimum and set the value of λ. After λ is subtracted from all fractional elements this element will become zero in z. This shows that the number of fractional elements in z is strictly less than in τ. The solution z' can be analysed in a similar way.

c

We will show that for one of the above λ we have $f(\tau) \le f(z)$. When creating z, we change all fractional elements by a small amount. Reorganizing the objective function we get:

$$f(z) = f(\tau) - \lambda \cdot \left(\sum_{i, 0<\tau_i<1} s_i + \sum_{ij, 0<\tau_{ij}<1} s_{ij} \right)$$

Out of the λ we defined in the previous section, one is always positive while the other is always negative. This means that no matter what the sign of the term inside the brackets is, for one of the above λ we have $f(\tau) \le f(z)$.

d

We have seen that the LP is equivalent to the local marginal polytope, and that some integral solution will be its maximum. From how we defined the function μ, each integral solution corresponds to a binary assignment of the vector x. We can now conclude that there is always an integral solution to the LP which is the exact MAP.

e

We now assume the pairwise marginal has four non-zero elements:

$$\theta_{ij}(x_i, x_j) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{6}$$

where $A + D - B - C > 0$ and the singleton terms $\theta_i(x_i)$ can be arbitrary. We will show that this problem can be brought to the form with $s_{ij} > 0$ and s_i as above, and therefore solved exactly.

Look at the terms that relate to some edge of the graph:

$$\theta_i(x_i) = \begin{bmatrix}\theta_i(0)\\ \theta_i(1)\end{bmatrix}, \theta_{ij}(x_i, x_j) = \begin{bmatrix}A & B\\ C & D\end{bmatrix}, \theta_j(x_j) = \begin{bmatrix}\theta_j(0) & \theta_j(1)\end{bmatrix}^T \tag{7}$$

The possible transformations which not change the maximizer will be:

1. Add the same scalar to all elements of θ_i, θ_{ij}, or θ_j.
2. Add the same scalar to a row of θ_{ij}, and subtract it from the same row of θ_i.
3. Add the same scalar to a column of θ_{ij}, and subtract it from the same column of θ_j.

The first transformation is legal because it increases or decreases the function $f(x;\theta)$ by a constant value regardless of x, therefore it doesn't influence the maximizer. The second and third transformations don't change $f(x;\theta)$, but they shift values from the pairwise terms to the singletons (and vice versa). If $x_i = 1$, then in the sum of $f(x;\theta)$ we will have $\theta_i(1)$ and $\theta_{ij}(1, x_j)$ for some value of x_j which corresponds to the bottom row of θ_{ij}. We can increase $\theta_i(1)$ and decrease the relevant part of θ_{ij} without changing $f(x;\theta)$.

Armed with these three transformations, we can now devise an algorithm that changes the marginals into the wanted form. For **each** edge, our algorithm carries out the following steps:

Beginning from Eq. **??**, subtract B from the second column of θ_{ij}:

$$\begin{bmatrix}\theta_i(0)\\ \theta_i(1)\end{bmatrix}, \begin{bmatrix}A & 0\\ C & D-B\end{bmatrix}, \begin{bmatrix}\theta_j(0) & \theta_j(1)+B\end{bmatrix}^T \tag{8}$$

Subtract C from the second row of θ_{ij}:

$$\begin{bmatrix}\theta_i(0)\\ \theta_i(1)+C\end{bmatrix}, \begin{bmatrix}A & 0\\ 0 & D-B-C\end{bmatrix}, \begin{bmatrix}\theta_j(0) & \theta_j(1)+B\end{bmatrix}^T \tag{9}$$

Add A to all elements of θ_{ij}:

$$\begin{bmatrix}\theta_i(0)\\ \theta_i(1)+C\end{bmatrix}, \begin{bmatrix}2A & A\\ A & A+D-B-C\end{bmatrix}, \begin{bmatrix}\theta_j(0) & \theta_j(1)+B\end{bmatrix}^T \tag{10}$$

Subtract A from the first column and first row of θ_{ij}:

$$\begin{bmatrix}\theta_i(0)+A\\ \theta_i(1)+C\end{bmatrix}, \begin{bmatrix}0 & 0\\ 0 & A+D-B-C\end{bmatrix}, \begin{bmatrix}\theta_j(0)+A & \theta_j(1)+B\end{bmatrix}^T \tag{11}$$

Subtract $\theta_i(0)+A$ from all values of θ_i and subtract $\theta_j(0)+A$ from all values of θ_j:

$$\begin{bmatrix}0\\ \theta_i(1)+C-\theta_i(0)-A\end{bmatrix}, \begin{bmatrix}0 & 0\\ 0 & A+D-B-C\end{bmatrix}, \begin{bmatrix}0 & \theta_j(1)+B-\theta_j(0)-A\end{bmatrix}^T \tag{12}$$

Given that $s_{ij} = A+D-B-C > 0$, we successfully simplified the problem without changing its minimizer (after doing these steps for each edge in the graph). QED.

Q2

Let us recall that the importance sampling estimate of $\mathbb{E}_p[f(X)]$ is the random variable:

$$Z = \frac{1}{T}\sum_{i=1}^{T} \frac{p(X^{(i)})}{q(X^{(i)})} f(X^{(i)}) \tag{13}$$

a

q^n is a distribution sampling n IID samples of $X(i)$ from $q(x)$. We need to show that

$$\mathbb{E}_{q^n}[Z] = \mathbb{E}_p[f(X)] \tag{14}$$

Because $X^{(i)}$ are IID then $Z^{(i)} = \frac{p(X^{(i)})}{q(X^{(i)})} f(X^{(i)})$ are IID as well, so it is sufficient to check that $\mathbb{E}\left[Z^{(1)}\right] = \mathbb{E}_p[f(X)]$. We will prove that in the continuous case (but it is similar in the discrete case).

$$\mathbb{E}_{q^n}[Z^{(1)}] = \int \frac{p(y)}{q(y)} q(y)dy = \int f(y)q(y)dy = \mathbb{E}_p[f(X)]$$

QED

b

We need to show that the variance is minimized by the following distribution:

$$q^*(x) = \frac{|f(x)|p(x)}{\mathbb{E}_p[|f(X)|]} \propto |f(x)|p(x) \tag{15}$$

For this section we will use Jensen's inequality. if X is a random variable and ϕ is convex function, then

$$\phi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\phi(X)\right] \tag{16}$$

We will prove again for the continuous case.
First, we will calculate the variance of Z and since $Z^{(i)}$ are IID, $Var(Z) = \frac{Var(Z^{(1)})}{T}$

$$\mathbb{E}[(Z^{(i)})^2] = \int f^2(y)\frac{p^2(y)}{q^2(y)} q(y)dy = \int f^2(y)\frac{p(y)}{q(y)} p(y) = \mathbb{E}_p\left[f^2(X)\frac{p(X)}{q(X)}\right] \tag{17}$$

Thus,

$$\sigma_{i,q}^2 = Var\left(Z^{(i)}\right) = \mathbb{E}\left[(Z^{(i)})^2\right] - \left(\mathbb{E}[(Z^{(i)})]\right)^2 = \mathbb{E}_p\left[f^2(X)\frac{p(X)}{q(X)}\right] - \left(\mathbb{E}_p[f(X)]\right)^2$$

Secondly, we will use the hint to find the lower bound for the first term (in the last equality)

$$\left(\mathbb{E}_p[|f(X)|]\right)^2 \overset{(??)}{=} \left(\mathbb{E}_q|f(X)|\frac{p(X)}{q(X)}\right)^2 \overset{(??)}{\leq} \mathbb{E}_q\left[f^2(X)\frac{p^2(X)}{q^2(X)}\right] \overset{(??)}{=} \mathbb{E}_p\left[f^2(X)\frac{p(X)}{q(X)}\right]$$

Finally, we will show that the equality takes place for $q^*(x)$

$$\mathbb{E}_p\left[f^2(X)\frac{p(X)}{q^*(X)}\right] = \int f^2(y)\frac{p^2(y)}{q^*(y)}dy \overset{(??)}{=} \int |f(y)|\mathbb{E}_p[|f(X)|]p(y)dy = (\mathbb{E}_p[|f(X)|])^2$$

Thus, $\sigma^2_{i,q^*} \leq \sigma^2_{i,q}$
QED

Q3

a

We want to solve the following maximization problem:

$$max_p - \sum_x p(x)log(p(x))$$

where

$$\forall i \in [1 \cdots d] \sum_x p(x)f_i(x) = a_i$$

and

$$\sum_x p(x) = 1$$

we should have added $\forall x p(x) \geq 0$, but instead we'll notice that the p we're about to find is already non-negative, as an exponent, even without adding this constraint.

So the lagrangian function is this:

$$\boldsymbol{L}(p, \vec{\lambda}, \nu) = -\sum_x p(x)log(p(x)) + \sum_i \lambda_i \left(\sum_x p(x)f_i(x) - a_i\right) + \nu \cdot \left(\sum_x p(x) - 1\right)$$

We derive w.r.t. p, to get

$$0 = \frac{\delta \boldsymbol{L}}{\delta p} = -\sum_x (1 + log(p(x)) + \sum_i \lambda_i \left(\sum_x f_i(x)\right) + \nu \cdot \sum_x 1$$

$$\sum_x log(p(x)) = \sum_x \left(\nu - 1 + \sum_i \lambda_i f_i(x)\right)$$

and so

$$p(x) = e^{\nu - 1} \cdot e^{\sum_i \lambda_i f_i(x)}$$

As requested.

b

We now have the marginals given for $ij \in E$ for some graph. We want to create the $\{f_i, a_i\}_{i=1\cdots d}$ so that the probability p that has $E_p\left(f_i(x)\right) = a_i$ that has the maximal entropy is the Pairwise MRF with those marginals. Recall that we can write the pairwise MRF as:

$$p_{p-MRF} = e^{\sum_{ij \in E} log(\mu_{ij}(x_i,x_j))}$$

We denote $\theta_{ij} := log(\mu_{ij})$.

$$= e^{\sum_{ij}\left(\theta_{ij}(x_i,x_j)\cdot \mathbb{1}_{ij\in E}\right)}$$

Which can be written similar to the first part of the question, with $\lambda_{ij} = \mathbb{1}_{ij\in E}$ and $f_{ij} = \theta_{ij}\left(x_i, x_j\right)$. Note that in the first part we had $i = 1\cdots d$ and d wasn't bound, here we have ij as the notation but we can still use the results from part 1.

So we need to see which a_is to choose.

$$\mathbb{E}_{p_{p-MRF}}(f_{ij}(x)) = \sum_x (p_{p-MRF}(x) f_{ij}(x)) = \sum_{x_{ij}} \left(\sum_{x_{[n]\setminus\{i,j\}}} p_{p-MRF}(x) f_{ij}(x) \right)$$

$$= \sum_{x_{ij}} f_{ij}(x) \left(\sum_{x_{[n]\setminus\{i,j\}}} p_{p-MRF}(x) \right)$$

now we will start calling p_{p-MRF} simply p for convenience. Notice that for every distribution, the bracketed part is simply the marginal distribution! So we will call it p_{ij} to get:

$$= \sum_{x_{ij}} f_{ij}(x)\left(p_{ij}(x)\right) = \sum_{x_{ij}} f_{ij}(x)\left(\mu_{ij}(x_i,x_j)\right)$$

$$= \sum_{x_{ij}} log(\mu_{ij}(x_i,x_j))\left(\mu_{ij}(x_i,x_j)\right) = -H\left[\mu_{ij}\right] := a_{ij}$$

And this concludes the question.

Q4

We need to show that for a pairwise MRF on graph, G, the log partition function, $logZ(\theta)$ is a convex function of θ. To prove that sufficient to show that $logZ(\theta)$ is PSD. First, we will show that the log partition function is a covariance matrix and then we will prove the covariance matrix is PSD.

We have seen in class that the log partition function can be represented as follows,

$$f(\theta) = logZ(\theta) = log\left(\sum_{x_1,\ldots,x_n} e^{\sum_{ij}\theta_{ij}(x_i,x_j)+\sum_i \theta_i(x_i)}\right) = log\left(\sum_{\boldsymbol{x}} e^{\boldsymbol{\mu(x)}^T\boldsymbol{\theta}}\right)$$

Where $\boldsymbol{\mu}$ comes from some distribution. Thus it should hold that all elements are non-negative, all distributions sum to one and pairwise distributions agree with the singleton ones.
We now want to show that the Hessian is a PSD matrix.

$$\frac{\partial f(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} log\left(\sum_{\boldsymbol{x}} exp(\mu^T\theta)\right) = \frac{1}{\sum_{\boldsymbol{x}} exp(\mu^T\theta)} \sum_{\boldsymbol{x}} \frac{\partial}{\partial \theta_i} exp(\mu^T\theta)$$

$$= \frac{1}{\sum_{\boldsymbol{x}} exp(\mu^T\theta)} \sum_{\boldsymbol{x}} \mu_i(\boldsymbol{x}) exp(\mu^T\theta) = \sum_{\boldsymbol{x}} \frac{exp(\mu^T\theta)}{\sum_{\tilde{\boldsymbol{x}}} exp(\mu^T\theta)} \mu_i(\boldsymbol{x})$$

$$= \sum_{\boldsymbol{x}} p(\boldsymbol{x})\mu_i(\boldsymbol{x}) = \mathbb{E}[\mu_i(\boldsymbol{x})]$$

$$\frac{\partial^2 f(\theta)}{\partial \theta_j \partial \theta_i} = \frac{\partial}{\partial \theta_j} \frac{\sum_{\boldsymbol{x}} \mu_i(\boldsymbol{x}) exp(\mu^T\theta)}{\sum_{\tilde{\boldsymbol{x}}} exp(\mu^T\theta)} = \frac{\sum_{\boldsymbol{x}} \mu_i(\boldsymbol{x})\mu_j(\boldsymbol{x}) e^{\mu^T\theta} \sum_{\tilde{\boldsymbol{x}}} e^{\mu^T\theta}}{(\sum_{\tilde{\boldsymbol{x}}} e^{\mu^T\theta})^2} -$$

$$\frac{\sum_{\boldsymbol{x}} \mu_i(\boldsymbol{x}) e^{\mu^T\theta} \sum_{\tilde{\boldsymbol{x}}} \mu_j(\tilde{\boldsymbol{x}}) e^{\mu^T\theta}}{(\sum_{\tilde{\boldsymbol{x}}} e^{\mu^T\theta})^2} = \sum_{\boldsymbol{x}} \frac{\mu_i(\boldsymbol{x})\mu_j(\boldsymbol{x}) e^{\mu^T\theta}}{\sum_{\tilde{\boldsymbol{x}}} e^{\mu^T\theta}} -$$

$$\sum_{\boldsymbol{x}} \frac{exp(\mu^T\theta)}{\sum_{\tilde{\boldsymbol{x}}} exp(\mu^T\theta)} \mu_i(\boldsymbol{x}) \sum_{\boldsymbol{x}} \frac{exp(\mu^T\theta)}{\sum_{\tilde{\boldsymbol{x}}} exp(\mu^T\theta)} \mu_j(\boldsymbol{x})$$

$$= \sum_{\boldsymbol{x}} p(\boldsymbol{x})\mu_i(\boldsymbol{x})\mu_j(\boldsymbol{x}) - \sum_{\boldsymbol{x}} p(\boldsymbol{x})\mu_i(\boldsymbol{x}) \sum_{\boldsymbol{x}} p(\boldsymbol{x})\mu_j(\boldsymbol{x})$$

$$= \mathbb{E}[\mu_i(\boldsymbol{x})\mu_j(\boldsymbol{x})] - \mathbb{E}[\mu_i(\boldsymbol{x})]\mathbb{E}[\mu_j(\boldsymbol{x})] = cov(\mu_i(\boldsymbol{x}), \mu_j(\boldsymbol{x}))$$

Thus, the Hessian matrix is $cov(\boldsymbol{\mu}(\boldsymbol{x}))$
Finally, we need to prove that the covariance matrix is PSD for any random vector x with expectation $\bar{x}$.

$$v^T C_x v = v^T \mathbb{E}[(x-\bar{x})(x-\bar{x})^T]v = \mathbb{E}[v^T(x-\bar{x})(x-\bar{x})^T v] \overset{u=v^T(x-\bar{x})}{=} \mathbb{E}[uu^T] = \sigma_u^2 \overset{(*)}{\geq} 0$$

(*) variance of a zero-mean random variable is non-negative.
QED