Advanced Machine Learning: HW-3

Uri Avron [uriavron@gmail.com] [308046994]
Jonathan Somer [jonathan.somer@gmail.com] [307923383]
Matan Harel [matan.harel.mh@gmail.com] [302695721]

May 9, 2018

1. Exact Solution Using The Local Marginal Polytope Approximation

Setting:

- n random variables $X_1, .., X_n$
- Graph E
- MRF defined by:
  1. $\forall ij \in E : \theta_{ij}(x_i, x_j) = \begin{bmatrix} 0 & 0 \\ 0 & s_{ij} \end{bmatrix}$ and $s_{ij} > 0$
  2. $\forall ij \in E : \theta_i(x_i) = \begin{bmatrix} 0 \\ s_i \end{bmatrix}$ and $s_i \neq 0$

(a) Show that $max_{\boldsymbol{\mu} \in M_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta}$ is equivalent to the following LP:

$$\text{Maximize: } f(\boldsymbol{\tau}) = \sum_i s_i \tau_i + \sum_{ij} s_{ij} \tau_{ij}$$

With respect to constraints:

$$\forall ij \in E : \tau_{ij} \geq 0 \tag{1}$$
$$\forall i : \tau_i \geq 0 \tag{2}$$
$$\forall ij \in E : \tau_{ij} \leq \tau_i \tag{3}$$
$$\forall ij \in E : \tau_{ij} \leq \tau_j \tag{4}$$
$$\forall ij \in E : \tau_{ij} \geq \tau_i + \tau_j - 1 \tag{5}$$

We shall start with the local marginal polytope (LMP) relaxation:

$$max_{\boldsymbol{\mu} \in M_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta} = max_{\boldsymbol{\mu}} \sum_{ij} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) + \sum_i \sum_{x_i} \mu_i(x_i)\theta_i(x_i)$$

With constraints:

$$\boldsymbol{\mu} \geq 0 \tag{6}$$

$$\sum_{x_i} \mu_i(x_i) = 1 \tag{7}$$

$$\sum_{x_i, x_j} \mu_{ij}(x_i, x_j) = 1 \tag{8}$$

$$\forall ij \in E, x_j : \sum_{x_i} \mu_{ij}(x_i, x_j) = \mu_j(x_j) \tag{9}$$

$$\forall ij \in E, x_i : \sum_{x_j} \mu_{ij}(x_i, x_j) = \mu_i(x_i) \tag{10}$$

First of all we shall assign $\boldsymbol{\theta}$ its values in our case, noting that its value is 0 in all cases but $(1,1), (1)$

$$max_{\boldsymbol{\mu} \in M_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta} = max_{\boldsymbol{\mu}} \sum_{ij} \mu_{ij}(1,1)\theta_{ij}(1,1) + \sum_i \mu_i(1)\theta_i(1)$$

$$= max_{\boldsymbol{\mu}} \sum_{ij} \mu_{ij}(1,1)s_{ij} + \sum_i \mu_i(1)s_i$$

We now rename $\forall i : \mu_i(1) := \tau_i$ and $\forall ij : \mu_{ij}(1,1) = \tau_{ij}$

$$max_{\boldsymbol{\tau}} \sum_{ij} \tau_{ij}s_{ij} + \sum_i \tau_i s_i$$

This is the exact function we are maximizing in the LP. That is, an optimal solution, over the original constraints, for the LP problem is also an optimal solution for the LMP problem. Thus it is enough to find optimal values for the τ that appear in the new formulation respecting the original constraints on μ. We must now show that:

1. The new LP constraints are not "too tight" so that we are not missing any optimal assignment to τ. We will show this by showing that the new LP constraints can be derived from the LMP constraints.

2. The constraints are not "too loose", that is: maximizing the new target function under the new constraints finds some optimal τ that can satisfy the original constraints on μ. We will prove this by showing that any optimal assignment to τ which satisfies LP can be extended to a valid assignment to μ according to LMP.

We will now show that the 5 constraints in the new problem, denoted by LP,can be derived from the 5 constraints in the LMP relaxation, denoted by LMP. Denote the i^{th} rule in LP by: $LP(i)$, and similarly with LMP. We will also us the notation $LMP(\{1,2\})$ etc. to denote sets of constraints.

(1) LP can be derived from LMP:

- $LP(1), LP(2)$ directly result from $LMP(1)$ and the way we defined $\boldsymbol{\tau}$
- $LP(3), LP(4)$ result from $LMP(1), LMP(4), LMP(5)$. To show this we will assume by contradiction and w.l.o.g that $LP(3)$ does not hold for some $ij \in E$, that is: $\tau_{ij} > \tau_i$. Note that from the way we defined τ_{ij}, τ_i we have $\mu_{ij}(1,1) > \mu_i(1)$ From $LMP(1)$:

  $$\sum_{x_j} \mu_{ij}(1, x_j) \geq \tau_{ij} > \tau_i = \mu_i(1)$$

  In contradiction to $LMP(5)$.

- We will now show $LP(5)$ results from $LMP(1), LMP(3), LMP(4), LMP(5)$. Let there be some $ij \in E$. By definition:

  $$\tau_i + \tau_j = \mu_i(1) + \mu_j(1)$$

  From $LMP(4), LMP(5)$:

  $$= \sum_{x_j} \mu_{ij}(1, x_j) + \sum_{x_i} \mu_{ij}(x_i, 1)$$

  Add $\mu_{ij}(0,0)$, $LMP(1)$:

  $$\leq \sum_{x_i x_j} \mu_{ij}(x_i, x_j) + \mu_{ij}(1,1)$$

  Definition + $LMP(3)$:

  $$= 1 + \tau_{ij}$$

  Subtracting 1 from both sides of the inequality we arrive at:

  $$\tau_i + \tau_j - 1 \leq \tau_{ij}$$

(2) Any optimal LP-valid assignment to τ can be extended to a LMP-valid assignment to μ:

Let there be some optimal LP-valid assignment to τ.

Extending the optimal solution:

First of all, note that any change to the values of μ who do not correspond to τ (denote by $\mu_{-\tau}$) do not change the value of the target function so we can alter them as we like as long as they do not violate the constraints. Start by assigning $0's$ to all $\mu_{-\tau}$. Note that at this point $LMP(1)$ holds from $LP(1,2)$ and the zero assignment to $\mu_{-\tau}$. From this point on we will only increase values of $\mu_{-\tau}$ (and will not increase to more than 1). So we are done with $LMP(1)$.
Claim: $\forall ij : \tau_{ij} \leq 1$

Assume by contradiction that $\tau_{ij} > 1$ then express τ_{ij} as $1 + \epsilon$, for some $\epsilon > 0$ It follows:

$$1 + \epsilon = \tau_{ij} \geq \tau_i + \tau_j - 1 \geq 1 + \epsilon + 1 + \epsilon - 1$$

Subtract $1 + \epsilon$ from both sides:

$$0 \geq \epsilon$$

Contradiction to the definition of ϵ.
Claim: $\forall i : \tau_i \leq 1$

Assume by contradiction that $\tau_i > 1$. Then:

$$\tau_{ij} \geq \tau_i + \tau_j - 1 = 1 + \epsilon + \tau_j - 1 = \epsilon + \tau_j$$

In contradiction to $LP(4)$

So we now can assume that $\forall ij \in E : 0 \leq \tau_{ij} \leq 1$. and $\forall$i:$0 \leq \tau_i \leq 1$
Let there be some $\tau_{ij}, \tau_i, \tau_j$ in our optimal solution. We will define μ_{ij}, μ_i, μ_j in such a way that all constraints of LMP hold for these values.
First of all, in order to satisfy $LMP(2)$ we must assign:

- $\mu_i(0) = 1 - \mu_i(1) = 1 - \tau_i$
- $\mu_j(0) = 1 - \mu_j(1) = 1 - \tau_j$

We will now satisfy $LMP(\{4,5\})$

- $\mu_j(1) = \tau_j = \sum_{x_i} \mu_{ij}(x_i, 1) = \mu_{ij}(0,1) + \tau_{ij}$. We know that $\tau_{ij} \leq \tau_j$ so we can assign a non-negative value smaller than 1 to $\mu_{ij}(0,1)$ s.t the equality holds.
- $\mu_j(0) = 1 - \tau_j = \sum_{x_i} \mu_{ij}(x_i, 0) = \mu_{ij}(0,0) + \mu_{ij}(1,0)$. We shall assign the value $1 - \tau_j$ to $\mu_{ij}(0,0)$

Note that $LMP(3)$ is satisfied by these assignments:
$\sum_{x_i,x_j} \mu_{ij}(x_i, x_j) = \sum_{x_i} \mu_{ij}(x_i, 0) + \sum_{x_i} \mu_{ij}(x_i, 1) = \tau_j + (1 - \tau_j) = 1$
This extends with no contradicting assignments to all other μ
So... We are done!
□

(b), (c) Exists Optimal $\boldsymbol{\tau}$ With $\{0, 1\}$ Values

Let there be some $\boldsymbol{\tau}$ with fractional values for some variables. We define a new solution $\boldsymbol{z}$ as:

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

Define $\lambda = min_{i:\tau_i > 0} \tau_i$; the minimal non-integeral value.
We shall show in the next two sections that:

1. $\boldsymbol{z}$ has less fractional values
2. $\boldsymbol{z}$ is an optimal solution to the original problem.

Thus, we can apply this method iteratively, assigning $\boldsymbol{\tau}$ to be the $\boldsymbol{z}$ of the previous iteration, until there are no fractional values.

(1) $\boldsymbol{z}$ has less fractional values then $\boldsymbol{\tau}$

Claim: the number of $1's$ and $0's$ does not decrease. So no fractional values are gained.

This results directly from the fact that for any non-fractional value $\mathcal{I}(0 < \tau < 1) = 0$. Thus for any non-fractional value we have: $z = \tau$.
Claim: denote $i = \underset{i:\tau_i > 0}{\text{argmin}} \tau_i$; $z_i = 0$ and τ_i was a fraction. So we will lose at least one fractional value.

We have assumed that $\boldsymbol{\tau}$ had fractional values. Thus i is the index of some fractional value and $\mathcal{I}(0 < \tau_i < 1) = 1$. By definition of λ we have $\lambda = \tau_i$. Thus $z_i = \tau_i - \tau_i = 0$.

Finally, by this definition of $\boldsymbol{z}$ we do not gain fractional values and lose at least one fractional value. So $\boldsymbol{z}$ has less fractional values then $\boldsymbol{\tau}$.

Note: we still need to show that $\boldsymbol{z}$ satisfies the original constraints:

- $LP(\{1, 2\})$ still hold because we subtract the minimal τ_i thus all sizes remain non-negative
- $LP(\{3, 4\})$ still hold, look at 3 cases (considering τ_i, case for τ_j is symmetrical):
  - $\tau_{ij} = 0$. Then the constraints hold from non-negativity of τ_i.
  - $0 < \tau_{ij} < 1$ then the minimal τ will be subtracted from it, but might or might not be subtracted from τ_i and the constraint holds in both cases.
  - $\tau_{ij} = 1$ then the minimal τwill not be subtracted from either τ_{ij} or τ_i and the constraint holds.

- $LP(5)$ if the minimal τ is not subtracted from τ_{ij} the constraint holds. If it is subtracted from τ_{ij} then we must only consider the case where it is not subtracted from either τ_i or τ_j. Assume by contradiction that such a case can occur. In this case τ_{ij} is a fraction and from $LP(\{3,4\})$ and the assumption that we do not subtract form τ_i or τ_j it must be the case that $\tau_i = \tau_j = 1$. Then from $LP(5)$ we have $\tau_{ij} \geq 1$ in contradiction to the assumption that τ_{ij} was a fraction.

(2) $\boldsymbol{z}$ is an optimal solution to the original problem

We will show that:

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

Lets look at $f(\boldsymbol{z})$:

$$f(\boldsymbol{z}) = \sum_i s_i z_i + \sum_{ij} s_{ij} z_{ij}$$

$$\begin{aligned}
\mathrm{f}(\boldsymbol{z}) &= \sum_i s_i z_i + \sum_{ij} s_{ij} z_{ij} \\
&= \sum_i s_i (\tau_i - \lambda \mathcal{I}(0 < \tau_i < 1)) + \sum_{ij} s_{ij} (\tau_{ij} - \lambda \mathcal{I}(0 < \tau_{ij} < 1)) \\
&= \sum_i s_i \tau_i + \sum_{ij} s_{ij} \tau_{ij} - \lambda \left[\sum_{i:0<\tau_i<1} s_i + \sum_{ij:0<\tau_{ij}<1} s_{ij} \right] \\
&= f(\boldsymbol{\tau}) - \lambda \left[\sum_{i:0<\tau_i<1} s_i + \sum_{ij:0<\tau_{ij}<1} s_{ij} \right] \\
&= f(\boldsymbol{\tau}) - \lambda S_{\text{fractions}}
\end{aligned}$$

Note that we can choose a λ that ensures that $-\lambda S_{\text{fractions}} \geq 0$:

- If $S_{\text{fractions}} \geq 0$ we can choose the negative λ from above.
- If $S_{\text{fractions}} \leq 0$ we can choose the positive λ from above.

In both cases we will end up with $f(\boldsymbol{\tau}) \leq f(\boldsymbol{z})$
□

(d) Conclusion:

We have seen that for any optimal solution $\boldsymbol{\tau}$ we can iteratively decrease the number of fractions in the solution. As we start with a finite number of fractional

values and decrease by at least one at each stage we must end up with an integeral solution. The solution is exact as we have simply solved an equivalent problem as the original.

The MAP is: $\boldsymbol{x}_{(i)} = \begin{cases} 1 & \tau_i = 1 \\ 0 & else \end{cases}$

It is easy to see that the method for extending our solution for τ(shown in part a) results in a LMP valid μ which behaves as a consistent indicator for the values of $\boldsymbol{x}$.

(e) Generalizing:

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

Assume $A + D - B - C > 0$; and $A, B, C, D \neq 0$.

We will show that we can bring this problem to the form from before. Note that in our case θ_{ij} is the same for all ij

We start with the original LMP form:

$$max_{\boldsymbol{\mu} \in M_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta} = max_{\boldsymbol{\mu}} \sum_{ij} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) + \sum_{i} \sum_{x_i} \mu_i(x_i)\theta_i(x_i)$$

We will wish to alter θ_{ij}, θ_j so that we arrive at:

$$\forall ij : \theta_{ij}(x_i, x_j) = \begin{bmatrix} 0 & 0 \\ 0 & A + D - B - C \end{bmatrix}$$

$$\theta_i = \begin{bmatrix} 0 \\ s_i \end{bmatrix}$$

Which operations can we perform without changing the optimal $\boldsymbol{\mu}$?

1. Subtracting c from a row or column of θ_{ij} (say: $\theta_{ij}(0,0)$ and $\theta_{ij}(1,0)$) and adding c to the respective $\theta_j(x_j)$ (for the case before: $\theta_j(0)$), does not change the optimal $\boldsymbol{\mu}$(other rows/cols are symmetrical) The reason we can perform this is:
   (a) The preference between the two: $\theta_{ij}(0,0)$, $\theta_{ij}(1,0)$ does not change as we have subtracted the same size from both.
   (b) The preference for the two: $\theta_{ij}(0,0)$, $\theta_{ij}(1,0)$ is decreased from the subtraction. From $LMP(4)$ decreasing μ for one of the previous two will immediately demand a decrease of the same size in $\theta_j(0)$. By adding c to $\theta_j(0)$ we de-insentivize the change in μ such that altering it will have no effect and thus the optimal μremains the same.
2. Subtracting/adding the same c from $\theta_i(1)$ and $\theta_i(0)$. The preference between the two remains the same.

Finally, perform the following on all θ_{ij}:

Using rule 1:

- Subtract B from right column $\Rightarrow \begin{bmatrix} A & 0 \\ C & D-B \end{bmatrix}$, and add to the relevant singleton (won't right this in following sections)
- Subtract C from bottom row $\Rightarrow \begin{bmatrix} A & 0 \\ 0 & D-B-C \end{bmatrix}$
- Subtract $A/2$ from top row and left column $\Rightarrow \begin{bmatrix} 0 & -A/2 \\ -A/2 & D-B-C \end{bmatrix}$
- Add $A/2$ to bottom row and right column $\Rightarrow \begin{bmatrix} 0 & 0 \\ 0 & A+D-B-C \end{bmatrix}$

Finall, using rule 2: subtract $c = \theta_i(0)$ from $\theta_i(0), \theta_j(1) \Rightarrow \theta_i = \begin{bmatrix} 0 \\ s_i \end{bmatrix}$ (hopefully $s_i \neq 0$, there is no way to tell as the instructions tell us that they are defined arbitrarily)

Done!

2 Importance Sampling

(a) Let us inspect $\mathbb{E}_{q^n}[Z]$:

$$\begin{aligned}
\mathbb{E}_{q^n}[Z] &= \mathbb{E}_{q^n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{p(X^{(i)})}{q(X^{(i)})}f(X^{(i)})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_q\left[\frac{p(X^{(i)})}{q(X^{(i)})}f(X^{(i)})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_q\left[\frac{p(X)}{q(X)}f(X)\right] \\
&= \mathbb{E}_q\left[\frac{p(X)}{q(X)}f(X)\right] \\
&= \sum_x q(x)\frac{p(x)}{q(x)}f(x) \\
&= \sum_x p(x)f(x) \\
&= \mathbb{E}_p[f(X)]
\end{aligned}$$

when the first two steps uses the linearity of expectation and the fact that we sample n IID samples. □

(b) By Jensen's inequality, for $\varphi(Y) = Y^2$, $Y = |f(X)|\frac{p(X)}{q(X)}$, it holds that $\mathbb{E}_p\left[f^2(X)\frac{p^2(X)}{q^2(X)}\right] \geq \left(\mathbb{E}_p\left[|f(X)|\frac{p(X)}{q(X)}\right]\right)^2$. Using this, we get a lower bound on the variance:

$$\begin{aligned}
\mathbb{V}_{q^n}[Z] &= \mathbb{E}_{q^n}\left[Z^2\right] - \mathbb{E}^2_{q^n}[Z] \\
&= \mathbb{E}_p\left[f^2(X)\frac{p^2(X)}{q^2(X)}\right] - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&\geq \left(\mathbb{E}_p\left[|f(X)|\frac{p(X)}{q(X)}\right]\right)^2 - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2
\end{aligned}$$

Let $q(x) \propto |f(x)|p(x)$ with constant c. For that q:

$$\begin{aligned}
\mathbb{V}_{q^n}[Z] &= \mathbb{E}_p\left[f^2(X)\frac{p^2(X)}{q^2(X)}\right] - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&= \mathbb{E}_p\left[f^2(X)\frac{p^2(X)}{\left(c|f(X)|p(X)\right)^2}\right] - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&= \mathbb{E}_p\left[\frac{1}{c^2}\right] - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&= \left(\mathbb{E}_p\left[\frac{1}{c}\right]\right)^2 - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&= \left(\mathbb{E}_p\left[\frac{1}{c}\frac{|f(X)|}{|f(X)|}\frac{p(X)}{p(X)}\right]\right)^2 - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&= \left(\mathbb{E}_p\left[|f(X)|\frac{p(X)}{c|f(X)|p(X)}\right]\right)^2 - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2 \\
&= \left(\mathbb{E}_p\left[|f(X)|\frac{p(X)}{q(X)}\right]\right)^2 - \left(\mathbb{E}_p\left[f(X)\frac{p(X)}{q(X)}\right]\right)^2
\end{aligned}$$

Hence, the lower bound is attained with q, which means q minimizes the variance. □

3. Entropy Maximization and MRFs

(a)

Assume $f_1(x), ..., f_d(x)$ to be d functions, and $a_1, ..., a_d$ to be d scalars. We're looking for a distribution $q(x)$ which maximize the entropy and satisfies that $\mathbb{E}[\mathrm{f_i(x)}] = a_i$ for all i.

Let's take a look on the following maximization problem:

$$\max_p -\sum_x p(x) \times \log(p(x))$$
$$\forall_i \mathbb{E}[f_i(x)] = a_i$$
$$\sum_x p(x) = 1$$

We will use Lagrange multiplier to find a solution for this problem. $\forall_{i\in[1,d]}\ \lambda_i$ is the Lagrange multiplier for the constrain $\mathbb{E}[f_i(x)] = a_i$, and c be the Lagrange multiplier to the constrain $\sum_x p(x) = 1$.
In addition by definition $E[f_i(x)] = \sum_x p(x) \times f_i(x)$. Then we get:

$$\mathcal{L}(p, \lambda, c) = -\sum_x p(x) \times \log(p(x)) + \sum_i \lambda_i (\sum_x p(x) \times f_i(x) - a_i) + c(\sum_x p(x) - 1)$$

Let's derivative with respect to p(x) and equal to zero in order to find q(x):

$$\mathcal{L}(p, \lambda, c)' = -1 - \log(p(x)) + \sum_i \lambda_i f_i(x) + c$$
$$log(q(x)) = \sum_i \lambda_i f_i(x) + c - 1$$
$$q(x) = e^{\sum_i \lambda_i f_i(x) + c - 1} = e^{c-1} \times e^{\sum_i \lambda_i f_i(x)}$$

We will use the constrain$\sum_x p(x) = 1$ to get

$$\sum_x q(x) = \sum_x e^{c-1} \times e^{\sum_i \lambda_i f_i(x)} = 1$$
$$e^{c-1} = \frac{1}{\sum_x e^{\sum_i \lambda_i f_i(x)}}$$

So e^{c-1} is the normalization factor. Now we'll prove that it's an maximum by derivative once again:

$$\mathcal{L}(p, \lambda, c)'' = \frac{-1}{q(x)} < 0$$

We can ignore the constrains $\forall_x p(x) \geq 0$ because the solution we got is $q(x) = e^{\sum_i \lambda_i f_i(x) + c - 1} = e^{c-1} \times e^{\sum_i \lambda_i f_i(x)} > 0$ for all x.
Therefore $q(x) \propto e^{\sum_i \lambda_i f_i(x)}$

(b)

For each k, l we'll define an indicator function $f_{k,l}(x_i, x_j)$ that is 1 only when $k = i$ and $l = j$ otherwise 0.
Now we'll require that $\mathbb{E}_p[f_{i,j}(x_i, x_j)] = \mu_{i,j}(x_i, x_j)$ and by that we get:

$$\mathbb{E}_p[f_{i,j}(x_i, x_j)] = \sum_{x_k, x_l} p(x_k, x_l) \times f_{i,j}(x_k, x_l) = p(x_i, x_j) = \mu_{i,j}(x_i, , x_j)$$

Let's define $\theta_{i,j}(x_i, x_j) = \lambda_{i,j}$ where $\lambda_{i,j}$ is the Lagrange multiplier that maximizes the maximization problem, then we get:

$$p(x) \propto e^{\sum_{i,j} \lambda_{i,j} \times f_{i,j}(x_i, x_j)} = e^{\sum_{i,j} \theta_{i,j}(x_i, x_j)}$$

and that is exactly a pairwise MRF and it maximizes entorpy.

4 Log Partition Function is Convex

Consider the quadruplets $Q = \{(i, j, x_i, x_j) | i, j \in [n], x_i \in X_i, x_j \in X_j\}$ and let us define some reasonable order $\text{ord}(i, j, x_i, x_j)$ on Q. We define vector $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}_{\text{ord}(i,j,x_i,x_j)} = \theta_{ij}(x_i, x_j)$. Moreover, we define function f such that for $k = \text{ord}(i, j, x_i, x_j)$:

$$(f(x'))_k = I\left[x'_i = x_i \wedge x'_j = x_j \wedge ij \in E\right]$$

$$f_k(x') \coloneqq (f(x'))_k$$

It follows that

$$\log Z(\boldsymbol{\theta}) = \sum_{x \in X} \exp\left(\boldsymbol{\theta} \cdot f(x)\right)$$

Let H be the hessian of $\log Z(\boldsymbol{\theta})$. Therefore (using the chain rule):

$$\begin{aligned}
H_{ij} &= \frac{\partial^2 \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \\
&= \frac{\partial \sum_x \exp\left(\boldsymbol{\theta} \cdot f(x)\right) f_i(x)}{\partial \boldsymbol{\theta}_j} \\
&= \sum_x \exp\left(\boldsymbol{\theta} \cdot f(x)\right) f_i(x) f_j(x) \\
&= \sum_x \exp\left(\frac{1}{2}\boldsymbol{\theta} \cdot f(x)\right) \exp\left(\frac{1}{2}\boldsymbol{\theta} \cdot f(x)\right) f_i(x) f_j(x) \\
&= \sum_x \left(f_i(x) \exp\left(\frac{1}{2}\boldsymbol{\theta} \cdot f(x)\right)\right) \cdot \left(f_j(x) \exp\left(\frac{1}{2}\boldsymbol{\theta} \cdot f(x)\right)\right)
\end{aligned}$$

Hence, by defining vector z such that$z_k = f(x)_k \exp\left(\frac{1}{2}\boldsymbol{\theta} \cdot f(x)\right)$ we get $H = \sum_k z_k z_k^T$. Now, for every vector v, it holds that

$$v^T H v = v^T \left(\sum_k z_k z_k^T\right) v = \sum_k v^T z_k z_k^T v = \sum_k (v^T z_k)^T (v^T z_k) = \sum_k (v^T z_k)^2 \geq 0$$

Thus, by definition, H is a PSD and hence $\log Z(\boldsymbol{\theta})$ is convex. $\square$