IE 598 Inference in Graphical Models

Homework 5

Covers lecture slides from

6. Density evolution
8. Variational Inference

Problem 5.1 [Variational approach] In this problem, we are going to compute free energies of simple graphical models and use BP-like fixed point equations to find the stationary points. We shall consider $G_\ell = (V_\ell, E_\ell)$, an $\ell \times \ell$ two-dimensional torus. This has vertex set $V_\ell = [\ell] \times [\ell]$ and, for any two vertices $i, j \in V_\ell$, $i = (i_1, i_2)$, $j = (j_1, j_2)$, $i_1, i_2, j_1, j_2 \in [\ell]$, we let $(i,j) \in E_\ell$ if and only if either $i_1 = j_1$ and $(i_2 - j_2) \in \{+1, -1\}$ *modulo* ℓ, or $i_2 = j_2$ and $(i_1 - j_1) \in \{+1, -1\}$ *modulo* ℓ.

We consider the *homogeneous* Ising model over $x \in \{+1, -1\}^{V_\ell}$

$$\mu(x) = \frac{1}{Z_G} \exp\Big\{\tilde{\epsilon} \sum_{(i,j)\in E_\ell} x_i x_j + \theta_{\mathrm{v}} \sum_{i \in V_\ell} x_i\Big\} ,$$

where $\tilde{\epsilon}, \theta_{\mathrm{v}}$ are parameters.
[It is rare to encounter such a symmetric model in applications. On the other hand, such toy examples are very useful for developing intuition.]

In the following, fix $\ell = 10$, $\theta_{\mathrm{v}} = 0.05$.

(*a*) Consider the *naive mean field approximation*, and write the naive mean field free energy for

$$\mathbb{F}_{\mathrm{MF}}(b) = \mathbb{E}_b[\log \psi_{\mathrm{tot}}(x)] - \sum_i \sum_{x_i} b_i(x_i) \log b_i(x_i) ,$$

where $b = b_1(\cdot) \times \cdots \times b_n(\cdot)$ and $\psi_{\mathrm{tot}}(x) = \prod_{i\in V} \psi_i(x_i) \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)$.

Assume then the further restriction $b_i(x_i) = b_{\mathrm{v}}(x_i)$ for all $i \in V_\ell$ (i.e. the belief is independent of the vertex). Write an expression $\mathbb{F}_{\mathrm{MF}}(b_{\mathrm{v}})$ as a function of b_{v}. Plot the free energy $\mathbb{F}_{\mathrm{MF}}(b_{\mathrm{v}})$ as a function of $a = (b_{\mathrm{v}}(+1) - b_{\mathrm{v}}(-1))$ for $\tilde{\epsilon} \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$.

Maximize $\mathbb{F}_{\mathrm{MF}}(b_{\mathrm{v}})$ with respect to b_{v} and plot the optimal value $b^*_{\mathrm{v}}(+1)$ and $\mathbb{F}_{\mathrm{MF}}(b^*_{\mathrm{v}})$ as a function of $\tilde{\epsilon}$.

(*b*) Repeat the same exercise for the *Bethe free energy*: Write explicitly the Bethe free energy

$$\begin{aligned}\mathbb{F}(b) \quad &= \quad \sum_{(i,j)\in E} \mathbb{E}_{b_{ij}}[\log \psi_{ij}(x_i, x_j)] + \sum_{i \in V} \mathbb{E}_{b_i}[\log \psi_i(x_i)] \\ &\quad - \sum_{(i,j)\in E} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j) - (1 - \deg(i)) \sum_{i\in V} \sum_{x_i} b_i(x_i) \log b_i(x_i) .\end{aligned}$$

Assume the further restriction $b_i(x_i) = b_{\mathrm{v}}(x_i)$ for all $i \in V_\ell$, $b_{ij}(x_i, x_j) = b_{\mathrm{e}}(x_i, x_j)$ (i.e. the belief is independent of the vertex). Write an expression $\mathbb{F}(b_{\mathrm{v}}, b_{\mathrm{e}})$ as a function of $b_{\mathrm{v}}, b_{\mathrm{e}}$.

Now, consider $\tilde{\epsilon} = 1.0$, and we want to show that $\mathbb{F}(b_{\mathrm{v}}, b_{\mathrm{e}})$ has more than one stationary point. The objective function is $\mathbb{F}(b_v, b_e)$, and the constraint is that $\sum_{x_i} b_e(x_i, x_j) = b_v(x_j)$ and $\sum_{x_j} b_e(x_i, x_j) = b_v(x_i)$. The Lagrangian can be written as

$$L(b_v, b_e, \lambda_1, \lambda_2) = \mathbb{F}(b_v, b_e) + \sum_{x_i} \lambda_1(x_i)(\sum_{x_i} b_e(x_i, x_j) - b_v(x_j)) + \sum_{x_j} \lambda_2(x_j)(\sum_{x_j} b_e(x_i, x_j) - b_v(x_i)) .$$

The derivative gives

$$\begin{aligned}\frac{\partial L}{\partial b_v(x_i)} &= \frac{\partial \mathbb{F}(b_v, b_e)}{\partial b_v(x_i)} - \lambda_1(x_i) - \lambda_2(x_i) + C \\ \frac{\partial L}{\partial b_e(x_i, x_j)} &= \frac{\partial \mathbb{F}(b_v, b_e)}{\partial b_v(x_i)} + \lambda_1(x_i) + \lambda_2(x_j) + C' \ ,\end{aligned}$$

where C and C' are constants (that may differ for each x_i, x_j) that we ignore because we do not care about normalization at this point. Write the explicit derivative of the Lagrangian in terms of l, θ_v, θ_e, $b_v(x_i)$, $b_e(x_i, x_j)$, and Lagrangian multipliers $\lambda_1(x_i)$ and $\lambda_2(x_j)$ which correspond to the constraints $\sum_{x_j} b_e(x_i, x_j) = b_v(x_i)$ and $\sum_{x_i} b_e(x_i, x_j) = b_v(x_j)$.

By symmetry, λ_1 and λ_2 are the same. So we define $\lambda(x_i) = (1/2l^2)\lambda_1(x_i) = (1/2l^2)\lambda_2(x_i)$. Show that $b_v(x_i)$ and $b_e(x_i, x_j)$ at the stationary point satisfy the below equations, by setting the above derivative to zero.

$$\begin{aligned} b_v(x_i) &\propto e^{-(1/3)\theta_v x_i} e^{(4/3)\lambda(x_i)} \\ b_e(x_i, x_j) &\propto e^{\theta_e x_i x_j} e^{(\lambda(x_i)+\lambda(x_j))} \ ,\end{aligned}$$

By the condition that $\sum_{x_i} b_e(x_i, x_j) = b_v(x_j)$, this gives

$$e^{\theta_e x_i + \lambda(+)} + e^{-\theta_e x_i + \lambda(-)} \quad \propto \quad e^{-(1/3)\theta_v x_i + (1/3)\lambda(x_i)} \ ,$$

for $x_i \in \{+1, -1\}$. substituting $x_i = +1$ in the above equation, then dividing by the same function evaluated at $x_i = -1$, we get

$$\frac{e^{\theta_e + \lambda(+)} + e^{-\theta_e + \lambda(-)}}{e^{-\theta_e + \lambda(+)} + e^{+\theta_e + \lambda(-)}} \quad = \quad e^{-(2/3)\theta_v + (1/3)(\lambda(+) - \lambda(-))} \ ,$$

Let $\ell = (1/2)(\lambda(+) - \lambda(-))$ and change variables to get

$$\frac{e^{\theta_e + \ell} + e^{-\theta_e - \ell}}{e^{-\theta_e + \ell} + e^{+\theta_e - \ell}} \quad = \quad e^{-(2/3)\theta_v + (2/3)\ell} \ ,$$

Using the equality that $\operatorname{atanh}(\tanh(a)\tanh(b)) = (1/2)\log\left(\frac{e^{a+b}+e^{-a-b}}{e^{a-b}+e^{-a+b}}\right)$, show that

$$\tanh(\theta_e)\tanh(\ell) \quad = \quad \tanh\left(\frac{1}{3}(\ell - \theta_v)\right) \ . \tag{1}$$

Plot the left-hand side and the right-hand side of the above equations to finish the proof that there are multiple stationary points of Bethe free energy when $\theta_v = 0.05$ and $\theta_e = 1.0$.

(*c*) We want to maximize $\mathbb{F}(p_1, p_2)$ for each value of $\theta_e \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. Using the above fixed point equations in (1), find all the fixed points of ℓ (numerically and/or approximately). For each fixed point ℓ, find the corresponding value of $b_v(\cdot)$, $b_e(\cdot)$, and $\mathbb{F}(b_v, b_e)$. Plot the optimal (i.e., maximum) value $p_1 = b^*_{\mathrm{v}}(+1)$ and the free energy $\mathbb{F}(p^*_1, p^*_2)$ as a function of $\tilde{\epsilon}$.

Problem 5.2 [Crowdsourcing] From the lecture, we studied a message passing algorithm (developed as a belief propagation for Haldane prior):

- initialize: $y^{(0)}_{j\to i}$'s as independent and identically distributed Gaussian random variable with mean one and variance one (this is one choice of initialization and any reasonable choice works as well)

- update messages:

$$x_{i\to j}^{(\tau+1)} = \sum_{k\in\partial i\setminus j} y_{j\to i}^{(\tau)} A_{ik}$$
$$y_{j\to i}^{(\tau+1)} = \sum_{k\in\partial j\setminus i} x_{k\to j}^{(\tau+1)} A_{kj}$$

- after enough number (e.g. T) of iterations estimate each task label by

$$\hat{t}_i = \text{sign}\Big(\sum_{k\in\partial i} y_{k\to i}^{(T)} A_{ik}\Big)$$

We will implement this algorithm for the following setting:

- the number of tasks $n = 100$
- the number of workers $m = 100$
- the (average) degree of a task node is ℓ
- the (average) degree of a worker node is also ℓ
- generate random graph as follows: for each task-worker pair (i, j), connect the two nodes with an edge with probability ℓ/m and otherwise do not connect with an edge: for example you can use the following Matlab script to generate such a graph with adjacency matrix E

```
E = zeros(n,m);
E = ceil( rand(n,m)-1+(l/m) );
```

- generate random n task labels i.i.d, such that $t_i = +1$ with probability $1/2$ and -1 with probability $1/2$

```
t = sign( rand(n,1)-0.5 );
```

- generate random m worker reliabilities i.i.d., such that p_j is drawn from the uniform distribution over the interval $[a, b]$ for some $0 < a < b < 1$

```
p = a+(b-a)*rand(m,1);
```

We will fix $a = 0.3$ and $b = 0.95$. For each value of $\ell \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$, we will generate 100 instances of {random graph, task labels, worker reliabilities}, and for each instance of the problem, generate the responses of the workers on those tasks assigned to the workers according to the **Dawid-Skene model**, i.e.

$$A_{ij} = \begin{cases} t_i & \text{with probability } p_j \\ -t_i & \text{with probability } 1 - p_j \end{cases}$$

for all $(i, j) \in E$.

For each instance of the problem, use the proposed algorithm to find the estimates $\{\hat{t}_i\}_{i\in n}$, and compute the error probability:

$$P_e(\ell) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(t_i \neq \hat{t}_i)$$

We will compare it to majority voting error rate:

$$P_{\text{MV}}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\Big(t_i \neq \text{sign}\big(\sum_{j\in\partial i} A_{ij}\big)\Big)$$

For each value of ℓ plot the $P_e(\ell)$ and $P_{\rm MV}(\ell)$ averaged over the 100 random instances of the problem, as a function of $\ell \in \{2,3,4,5,6,7,8,9,10\}$.

Problem 5.3 In this problem, we explore the connections between minimum cut of a graph and pairwise Markov random fields in binary alphabets. Consider a graphical model defined on an undirected graph $G(V,E)$,

$$\mu(x) \;=\; \frac{1}{Z}\exp\{-\sum_{i\in V}\phi_i(x_i) - \sum_{(i,j)\in E}\phi_{ij}(x_i,x_j)\}\;,$$

for $x=[x_1,\dots,x_n]\in\{0,1\}^n$. We further assume for now that $\phi_{ij}(0,0)=\phi_{ij}(1,1)=0$ for all $(i,j)\in E$ (meaning they are zero-diagonal when we consider the functions as 2×2 matrices) such that

$$\phi_i(\cdot) \;=\; \begin{bmatrix}\phi_i(0)\\ \phi_i(1)\end{bmatrix}\;,\quad\text{and}\quad \phi_{ij}(\cdot,\cdot) \;=\; \begin{bmatrix}0 & \phi_{ij}(0,1)\\ \phi_{ij}(1,0) & 0\end{bmatrix}\;.$$

Our goal is to find the maximum likelihood estimate, the one that maximizes the above joint distribution. In order to find the maximizer, we pose this question as a problem of finding the minimum cut of a graph.

Given a pairwise MRF on $G(V,E)$ and the compatibility functions $\phi_{ij}(\cdot,\cdot)$'s, we first create a new **directed** and **weighted** graph as follows.

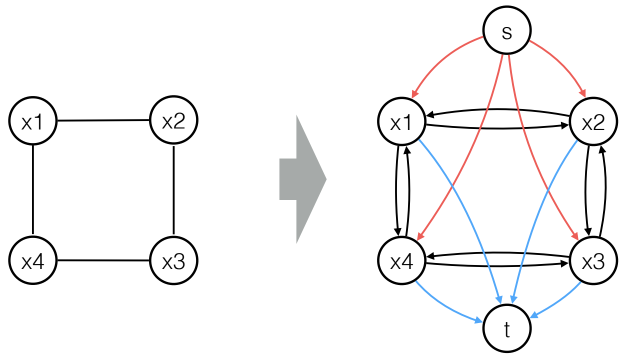
- Add one node for the source s and one node for the sink t.
- Add an edge from source s to all nodes in V (red edges in the figure below).
- Add an edge from all nodes in V to the sink t (blue edges in the figure below).
- make all edges in E reciprocal (by taking the undirected edge E and making them in to two edges in opposite directions; black edges in the figure below).

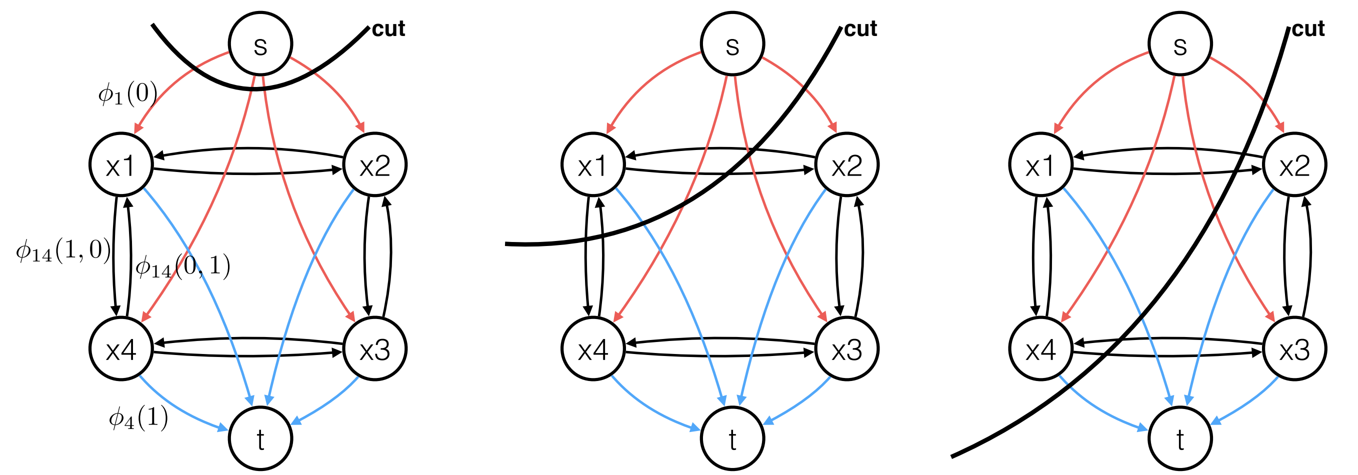
An example of a 2×2 grid G, that is transformed is shown below. The colors do not have particular meanings, it is there to help you understand the creation of the new graph. We will find the minimum cut in this transformed graph, after putting appropriate non-negative weights on the edges. A **cut** in a graph is partition of the nodes into two disjoint sets, one containing the source and the other containing the sink. **The value of a cut** is the total weight of the edges that start from a node in the same partition as the source and end in a node in the sink side of the partition, i.e. those that go from the source side of the partition to the other. Note that in the minimum cut, for each node in V, EITHER the edge connecting to the sink will be cut, OR the edge connecting from the source will be cut, but NOT BOTH (since the source and the sink are constrained to be on different sides of the cut). Once we find the minimum cut in this graph, we will assign ZERO to the sink side of the cut and ONE to the source side. This defines a one-to-one mapping between an assignment of binary values in the MRF and a cut in the transformed graph $H(V\cup\{s,t\},D)$.

Our goal is to minimize $E(x)\triangleq\sum_{i\in V}\phi_i(x_i)+\sum_{(i,j)\in E}\phi_{ij}(x_i,x_j)$ (which is equivalent as finding the most likely assignment). The following costs on the edges (also called capacities in max-flow min-cut context) ensures that the min-cut of the transformed graph H corresponds to the minimizer of $E(x)$.

- Assign $\phi_i(0)$ to the edge from the source (s,i).
- Assign $\phi_i(1)$ to the edge to the sink (i,t).
- Assign $\phi(1,0)$ to the edge (i,j) and $\phi_{ij}(0,1)$ to the edge (j,i).

An example below shows that this assignment ensures that the value of the cut corresponds to the energy $E(x)$ of the corresponding assignment. In general, cut values are equal to the energy $E(x)$ pf the corresponding assignment x.





assignment: $x = [0,0,0,0]$
cut value: $\phi_1(0) + \phi_2(0) + \phi_3(0) + \phi_4(0)$

$x = [1,0,0,0]$
$\phi_1(1) + \phi_{14}(1,0) + \phi_{12}(1,0)$
$+\phi_2(0) + \phi_3(0) + \phi_4(0)$

$x = [1,0,0,1]$
$\phi_1(1) + \phi_{12}(1,0)$
$+\phi_2(0) + \phi_3(0) + \phi_4(1) + \phi_{34}(01)$

It is known that when the cost on the edges are non-negative, the minimum cut can be found efficiently. Hence, when all $\phi_{ij}(0,0) = \phi_{ij}(1,1) = 0$ and $\phi_i(x_i)$'s, $\phi_{ij}(0,1)$'s and $\phi_{ij}(1,0)$'s are all non-negative, then the costs on the edges are all non-negative and the minimizer of $E(x)$ can be found efficiently by running the off-the-shelf min-cut solvers on H.

(a) Suppose $\phi_1(0) < 0$, and the rest of the compatibility functions are all non-negative, and $\phi_{ij}(0,0) = \phi_{ij}(1,1) = 0$ for all $(i,j) \in E$. Find a new $\phi'_1(x_1)$ such that

- $\phi'_1(0)$ and $\phi'_1(1)$ are non-negative; and
- the minimizer of $E'(x) = \phi'_1(x_1) + \sum_{i \in V \setminus \{1\}} \phi_i(x_i) + \sum_{(i,j)\in E} \phi_{ij}(x_i, x_j)$ is the minimizer of $E(x)$.

Then, the corresponding transformed graph H with the new costs from $\phi'_1(x_1)$ can be solved for min-cut, since all costs are non-negative.

(b) Now, consider a general case when $\phi_{ij}(0,0)$'s and $\phi_{ij}(1,1)$'s are not necessarily zero. Explain how to assign costs to the directed edges of H (not just for the example given above, but for general $H(V \cup \{s,t\}, D)$ defined from general $G(V,E)$), such that **the value of a cut in this new H is equal to the energy $E(x) = \sum_{i\in V} \phi_i(x_i) + \sum_{(i,j)\in E} \phi_{ij}(x_i,x_j)$ for the corresponding assignment x.** Note that we do not worry about computational complexity of finding the minimum-cut in this part, and focus in

posing the problem as a min-cut problem.
[hint: consider changing $\phi_i(x_i)$'s and $\phi_{ij}(x_i, x_j)$'s in order to get new $\phi'_{ij}(x_i, x_j)$'s such that the diagonals are zero.]

(c) Suppose $\phi_i(x_i)$'s are all non-negative and $\phi_{ij}(x_i, x_j)$'s are also all non-negative. Assigning costs to the edges of H as per the solution of part (b), it is possible that some edges are assigned negative costs. This is problematic, since min-cut cannot be efficiently solved. However, when all pairwise compatibility functions are **sub-modular**, then the minimizer of $E(x)$ can be found efficiently. We will prove that this is possible, by constructing a new graph H with non-negative costs under sub-modularity assumption.

A function $f(\cdot)$ over two binary variables is said to be sub-modular if and only if

$$f(0,0) + f(1,1) \quad \leq \quad f(0,1) + f(1,0) \; .$$

Suppose $\phi_i(x_i)$'s are non-negative and $\phi_{ij}(x_i, x_j)$'s are non-negative and sub-modular. Explain how to assign costs to the directed edges of H (not just for the example given above, but for general $H(V \cup \{s,t\}, D)$ defined from general $G(V, E)$), such that

- the value of a cut in this new H is equal to the energy $E(x) = \sum_{i \in V} \phi_i(x_i) + \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j)$ for the corresponding assignment x; and
- all costs are non-negative.

[hint: consider changing $\phi_i(x_i)$'s and $\phi_{ij}(x_i, x_j)$'s in order to get new $\phi'_{ij}(x_i, x_j)$'s such that the diagonals are zero and the off-diagonals are non-negative.]