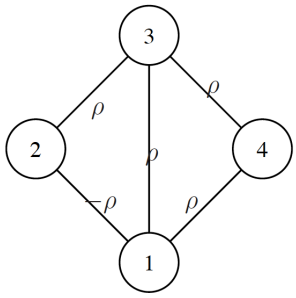
HW3 (4 problems, 1 of which is optional)

Jamie Morgenstern *CSE515*

1. *[2 points]* **(Loopy belief propagation)** Consider the Gaussian graphical model depicted below. More precisely, if we let x denote the 4-dimensional vector of variables at the 4 nodes (ordered according to the node numbering given), then $x \sim \mathcal{N}^{-1}(h, J)$, where J has diagonal values all equal to 1 and non-zero off-diagonal entries as indicated in the figure (e.g., $J_{12} = -\rho$).



   (a) Confirm (e.g., by checking Sylvester's criterion to see if the determinants of all principal minors are positive or by checking the smallest eigen value is positive) that J is a valid information matrix–i.e., it is positive definite–if $\rho = .39$ or $\rho = .4$. Compute the variances for each of the components (i.e., the diagonal elements of $\Lambda = J^{-1}$)–you can use any software to do this if you'd like.

   (b) We now want to examine Loopy BP for this model, focusing on the recursions for the information matrix parameters. Write out these recursions in detail for this model. Implement these **recursions** and try for $\rho = .39$ and $\rho = .4$ with 5,000 iterations. Plot the computed marginals (just the J_i's and not h_i's) at the 4 nodes over the iterations. Describe the behavior that you observe.

   (c) Construct the computation tree for this model with node three as the root node. Note that the effective "J" – parameters for this model are copies of the corresponding ones for the original model (so that every time the edge $(1, 2)$ appears in the computation tree, the corresponding J-component is $-\rho$). Use Matlab, Python or any other numerical tools to check the positive-definiteness of these implied models on computation trees for different depths (use iterations 1,2, and 3) and for two different values of ρ ($\rho \in \{0.3, 0.46\}$). What do you observe that would explain the result in part (b)?

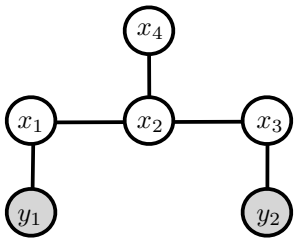
2. *[2 points]* **(Gaussian graphical model and Gaussian BP)**
   Let $x \sim \mathcal{N}^{-1}(h_x, J_x)$, and $y = Cx + v$, where $v \sim \mathcal{N}(0, R)$.

   (*a*) Find the potential vector $h_{y|x}$ and the information matrix $J_{y|x}$ of $p(y|x)$.

   (*b*) Find the potential vector $h_{x,y}$ and the information matrix $J_{x,y}$ of $p(x, y)$.

   (*c*) Find the potential vector $h_{x|y}$ and the information matrix $J_{x|y}$ of $p(x|y)$.

   (*d*) Consider the following Gaussian graphical model. We will use the type of BP where the messages are distributions over the source, i.e. the message $\mathcal{N}^{-1}(h_{i\to j}, J_{i\to j})$ represents a distribution over x_i (as opposed to x_j).



Let $y_1 = x_1 + v_1$, $y_2 = x_3 + v_2$, and $R = I$ is the identity matrix. Find C. Represent messages $h_{x_3 \to x_2}$ and $J_{x_3 \to x_2}$ in terms of y_2 and the elements of h_x and J_x. [y_1 and y_2 are measurements, which should be treated as given and deterministically known.]

(*e*) Now assume that we have an additional measurement $y_3 = x_3 + v_3$, where v_3 is a zero-mean Gaussian variable with variance 1 and is independent from all other variables. Find the new C. Represent messages $h_{x_3 \to x_2}$ and $J_{x_3 \to x_2}$ in terms of y_2, y_3 and the elements of h_x and J_x. [again y_2 should be considered as a measurement which is given, and deterministically known.]

(*f*) The BP message from x_3 to x_2 define a Gaussian distribution with mean $m_{x_3 \to x_2} = J^{-1}_{x_3 \to x_2} h_{x_3 \to x_2}$ and variance $\sigma_{x_3 \to x_2} = J^{-1}_{x_3 \to x_2}$. Compare the difference in the variance of this message when computed using a single observation y_2 versus when computed using multiple observations (y_2, y_3). What is the mean and variance of the BP message when the number of observations grows to infinity?

3. *[2 points]* **(Free energy)**
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 $x \in \{+1, -1\}^{V_\ell}$

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

where $\theta_e, \theta_{\rm v}$ 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 $\ell = 10$, $\theta_{\rm v} = 0.05$.

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

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

where $b = b_1(\cdot) \times \cdots \times b_n(\cdot)$ and $\psi_{\rm tot}(x) = \prod_{i \in V} \psi_i(x_i) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j)$.
Assume then the further restriction $b_i(x_i) = b_{\rm v}(x_i)$ for all $i \in V_\ell$ (i.e. the belief is independent of the vertex). Write an expression $\mathbb{F}_{\rm MF}(b_{\rm v})$ as a function of $b_{\rm v} \in \mathbb{R}^2$. This is the objective function to be maximized. Plot the free energy $\mathbb{F}_{\rm MF}(b_{\rm v})$ as a function of a scalar variable $a = (b_{\rm v}(+1) -$

$b_{\rm v}(-1)) \in \mathbb{R}$ for $\theta_e \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. This is equivalent to setting $b_{\rm v}(+1) = (1+a)/2$ and $b_{\rm v}(-1) = (1-a)/2$.

Maximize $\mathbb{F}_{\rm MF}(b_{\rm v})$ (using numerical methods) with respect to $b_{\rm v}$ and plot the optimal value $b^*_{\rm v}(+1)$ and $\mathbb{F}_{\rm MF}(b^*_{\rm v})$ as a function of θ_e.

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

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

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

Now, consider $\theta_e = 1.0$, and we want to show that $\mathbb{F}(b_{\rm v}, b_{\rm e})$ 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

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

Notice that we omitted the constraints $\sum_{x_i} b_v(x_i) = 1$ condition to save some notations and space. It is without loss of generality as we will only specify the values up to a scaling in the following derivations. In the end, we will uniquely identify the scaling such that $\sum_{x_i} b_v(x_i) = 1$ is satisfied.

The derivative gives

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

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

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

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

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

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

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

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

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

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 $\theta_e = 1.0$. This means that plot two curves, $y = \tanh(\theta_e)\tanh(x)$ and $y = \tanh(\frac{1}{3}(x - \theta_v))$, and inspect how many places they meet.

(*c*) We want to maximize $\mathbb{F}(p_1, p_2)$ for each value of $\theta_e \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, where $p_1 = b_v(+1)$ and $p_2 = b_e$. Using the above fixed point equations in (1), find all the fixed points of w (numerically and/or approximately). For each fixed point w, 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 $p_1^* = b_v^*(+1)$ and the free energy $\mathbb{F}(p_1^*, p_2^*)$ as a function of θ_e.

4. *OPTIONAL, [2 points]* **(Application of minimum cut)**
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 $G(V, E)$,

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

for $x = [x_1, \ldots, x_n] \in \{0, 1\}^n$. 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

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

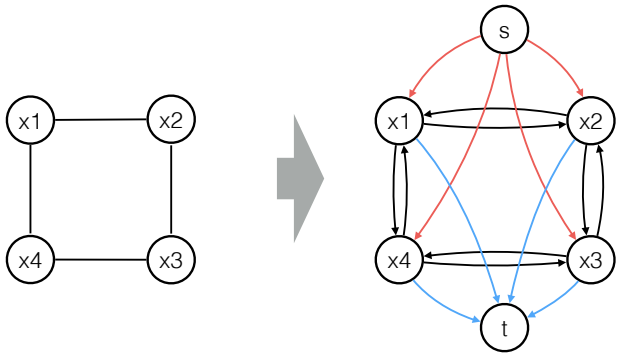
Our goal is to find the maximum likelihood estimate, the one that maximizes the above joint distribution. In order to find the maximizer, we pose this question as a problem of finding the minimum cut of a graph.

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).

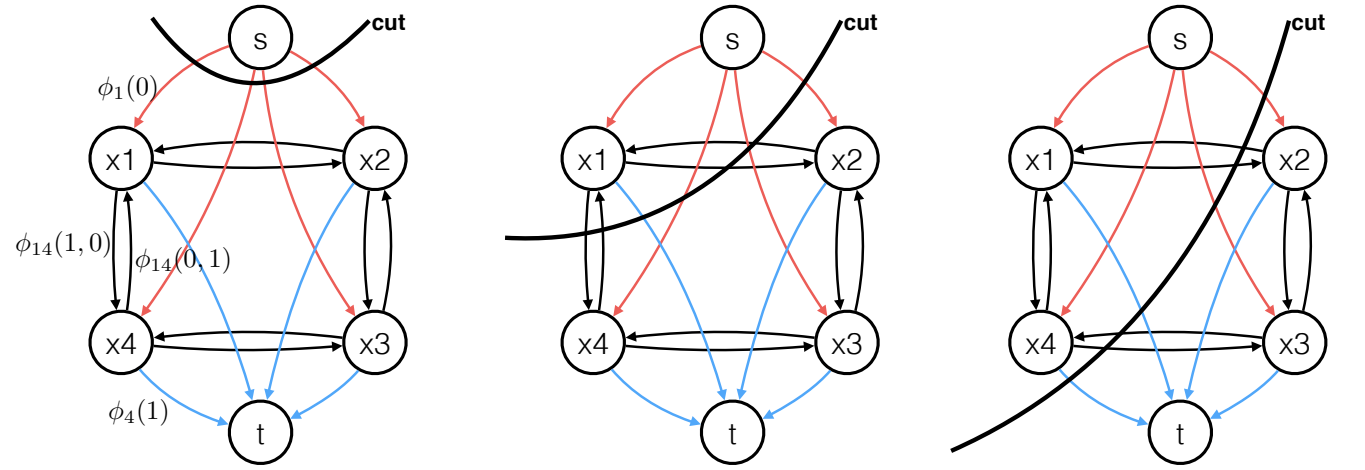
An example below shows that this assignment ensures that the value of the cut corresponds to the energy $E(x)$ of the corresponding assignment. In general, cut values are equal to the energy $E(x)$ pf the corresponding assignment x.

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



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

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

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

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.]