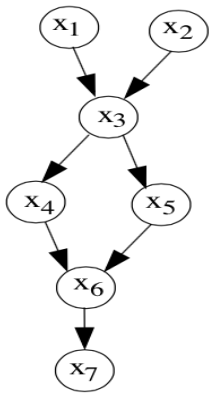
CSC 412
Probabilistic Graphical Models
Assignment 2
Out: Feb. 9
Due: Mar. 1

1 Graphical model distributions [20 points]

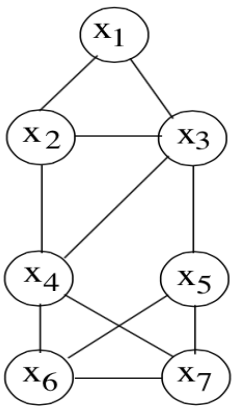
The figure below shows a directed graphical model (D) and an undirected one (U), each representing a distribution over six variables.

For each of the following statements and each graphical model, prove whether the statement *must* be true of the distribution represented by the model, *could* be true but we do not know, or *cannot* be true.

1. $\mathbf{x}_4$ is conditionally independent of $\mathbf{x}_5$ given $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$
2. $\mathbf{x}_3$ is conditionally dependent on $\mathbf{x}_7$ given $\mathbf{x}_4, \mathbf{x}_5$
3. $\mathbf{x}_1$ is marginally independent of $\mathbf{x}_2$
4. $\mathbf{x}_4$ is conditionally independent of $\mathbf{x}_5$ given $\mathbf{x}_3, \mathbf{x}_6$



D



U

2 Complete graphs [20 points]

Draw a directed graphical mode on five variables which (a) can capture any joint distribution and (b) is acyclic.

Can any edges be added to or removed from your graph and still preserve both the properties (a) and (b) above? If so show the addition or removal; if not say why not.

Draw an undirected graphical model on four variables which can capture any joint distribution. List all the maximal cliques.

Can any edges be added to or removed from your graph and still preserve the above property? If so show the addition or removal; if not say why not.

[Bonus] Prove or disprove the following statement: All (un)directed graphical models on k variables which can capture any joint distribution have exactly $k^2/2 - k/2$ edges.

3 Learning undirected models [60 points] (Sudderth)

We now develop algorithms for learning, from complete observations, undirected graphical models of N binary variables $x_s \in \{0, 1\}$. We focus on models with *pairwise* dependencies, and use a minimal parameterization which allocates one parameter θ_s for each node $s \in \mathcal{V}$, and one parameter θ_{st} for each edge $(s, t) \in \mathcal{E}$. The overall joint distribution is then:

$$p(x|\theta) = \exp\left\{\sum_{s\in\mathcal{V}} \theta_s x_s + \sum_{(s,t)\in\mathcal{E}} \theta_{st} x_s x_t - \Phi(\theta)\right\} \tag{1}$$

$$\Phi(\theta) = \log\left(\sum_x \exp\left\{\sum_{s\in\mathcal{V}} \theta_s x_s + \sum_{(s,t)\in\mathcal{E}} \theta_{st} x_s x_t\right\}\right) \tag{2}$$

The questions below use voting records from senators during the 112th United States Congress (January 3, 2011 until January 3, 2013), as collected by `voteview.org`. We provide data for $N = 13$ senators, and all $L = 486$ bills, in the binary matrix `senatorVotes`. If x_{sl} is the vote of senator s on measure l, we let $x_{sl} = 1$ if their vote was "Yea", and $x_{sl} = 0$ otherwise (a vote of "Nay", or more rarely a failure to vote). We also provide the last names, home states, and party affliations (1 for Democrat, 2 for Republican) of each Senator. For learning, we interpret the bills as L independent samples from some joint distribution on Senate votes. For some of the questions below, you need to solve L_1-regularized optimization problems. We recommend using Schmidt's `L1General` Matlab package, which is available at: `http://www.di.ens.fr/~mschmidt/Software/L1General.html`. Also be sure to read the INSTRUCTIONS file in the package.

a) Consider a pairwise binary MRF as in Eqn. (1). Suppose that the graph is fully disconnected ($\mathcal{E} = \phi$). Derive a closed form expression for the maximum likelihood (ML) estimates of the node parameters θ_s. Compute these ML estimates using the full vote dataset, and plot the estimated θ_s values for each senator.

b) Now allow the pairwise binary MRF of Eqn. (1) to have some arbitrary, fixed set of edges. Consider the joint log-likelihood of a dataset with L independent observations. Derive an expression for the gradient of this log-likelihood with respect to some vectorized ordering of the parameters θ. Simplify your answer to be as explicit as possible. Write a function that, given some training dataset, computes the log-likelihood objective value and gradient corresponding to any candidate model parameters θ.

c) Using the gradient objective from part (b), and an optimization package such as `L1General`, write code which computes the ML estimate of the model parameters θ. Assume a fully connected pairwise graphical model, for which $\mathcal{E}$ contains an edge linking every pair of nodes. Apply this code to the full Senate voting record, and plot the log-likelihood of the estimated model after each optimization iteration. Initialize the node parameters θ_s to the ML estimates from part (a), and the edge parameters $\theta_{st} = 0$.

d) Can the fully connected, pairwise graphical model estimated in part (c) represent an arbitrary joint distribution on N binary variables? If so, explain why. If not, discuss which statistics of the data it does capture.

e) Consider two different models learned from the voting data: the factorized model from part (a), and the fully connected model from part (c). For each model, compute the binary entropy of the corresponding joint distribution:

$$H(\theta) = -\sum_x p(x|\theta) \log_2 p(x|\theta)$$

What do these numbers suggest about the voting patterns of US senators?

f) Suppose that we place a factorized Laplacian prior on our model parameters:

$$p(\theta|\lambda) = \prod_{s \in \mathcal{V}} Lap(\theta_s|\lambda_s) \prod_{(s,t) \in \mathcal{E}} Lap(\theta_{st}|\lambda_{st})$$

$$Lap(\theta|\lambda) = \frac{\lambda}{2} \exp\{-\lambda|\theta|\}$$

Derive an objective whose minimum, given a dataset with L training examples and fixed hyperparameters λ, gives the maximum a posteriori (MAP) estimate of θ. Adapt the code from part (c) to numerically compute MAP estimates.

g) Let $\lambda_{st} = \bar{\lambda}$ for all pairs of nodes, and $\lambda_s \to 0$ (the limit as the variance of the node parameter priors approaches infinity). Use the votes for the first 400 bills as a training set, and the remaining 86 bills as a validation set. For a range of possible $\bar{\lambda}$, and MAP parameter estimates. For each learned model, evaluate the log-probability of the validation data, and plot these probabilities as a function of $\bar{\lambda}$. Use a logarithmic scale when choosing candidate $\bar{\lambda}$, and when making this plot.

h) Suggest a way of associating graphical structures to the models learned in part (g). Plot the graph corresponding to the $\bar{\lambda}$ which produced the highest validation log-likelihood, as well as the graph with the smallest number of edges that was nevertheless connected. In both cases, label the nodes of the graphs with the names of the corresponding senators.