CPSC 540 Assignment 5 (due April 4 at midnight)

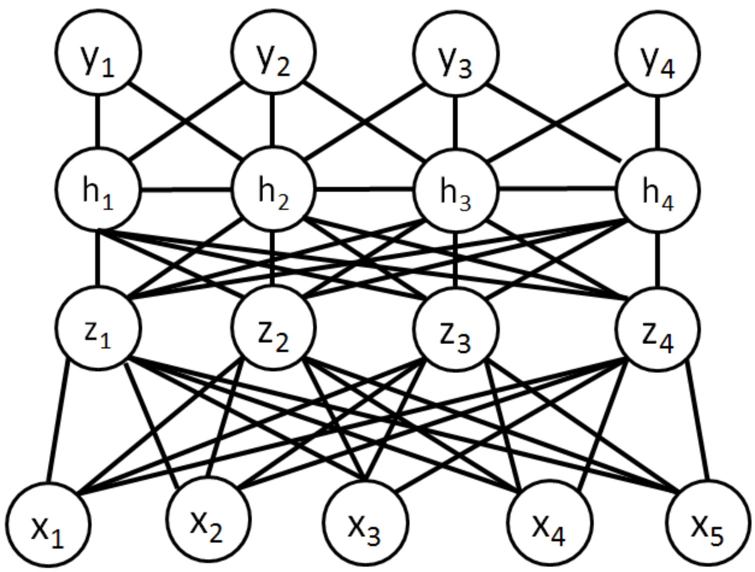
The assignment instructions are the same as for the previous assignment.

1. Name(s):
2. Student ID(s):

1 Undirected Graphical Models

1.1 Conditional UGM

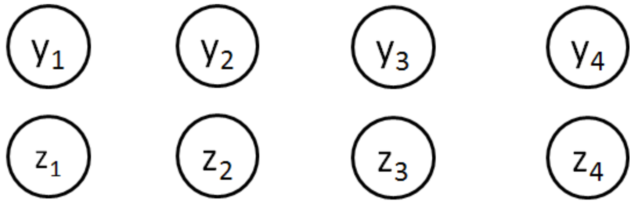
Consider modeling the dependencies between sets of binary variables x_j and y_j with the following UGM which is a variation on a stacked RBM:



Computing univariate marginals in this model will be NP-hard in general, but the graph structure allows efficient block updates by conditioning on suitable subsets of the variables (this could be useful for designing approximate inference methods). For each of the conditioning scenarios below, draw the conditional UGM and informally comment on how expensive it would be to compute univariate marginals (for all variables) in the conditional UGM.

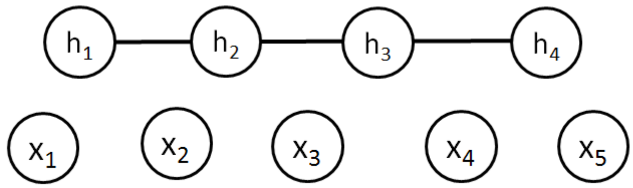
1. Conditioning on all the x and h values.

   Answer: This gives a disconnected graph, so computing marginals is trivial (it only involves univarate probabilities).



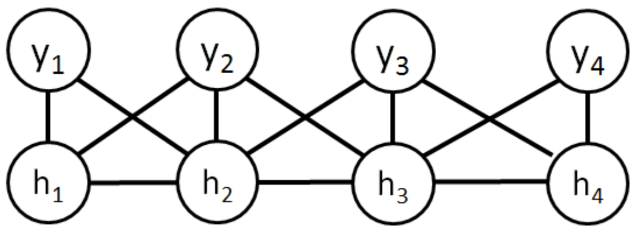
2. Conditioning on all the z and y values.

   Answer: This gives a disconnected graph and a chain, so computing marginals is cheap by using message-passing (the messages will depend on only 1 variable).



3. Conditioning on all the x and z values.

   Answer: This graph is not a tree so message passing will be less cheap (though it's still tractable if you combine each y_i and h_i together into one state and write it a as Markov chain).



1.2 Fitting a UGM to PINs

The function *example_UGM.jl* loads a dataset X containing samples of PIN numbers, based on the probabilities from the article at this URL: `http://www.datagenetics.com/blog/september32012`.[1]

This function fits a UGM model to the dataset, where all node/edge parameters are untied and the graph is empty. It then performs decoding/inference/sampling in the fitted model. The decoding is reasonable (it's $x = \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}$) and the univariate marginals are reasonable (it says the first number is 1 approximately 40% of the time and the last number is 4 approximately 20% of the time), but because it assumes the variables are independent we can see that this is not a very good model:

1. The sampler doesn't tend to generate the decoding ($x = \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}$) as often as we would expect. Since it happens in more than 1/10 of the training examples, we should be seeing it in more than 1/10 of the samples.

2. Conditioned on the first three numbers being 1 2 3, the probability that the last number is 4 is only around 20%, whereas in the data it's more than 90% in this scenario.

In this question, you'll explore using (non-degenerate UGMs) to try to fix the above issues:

1. Write an equation for $p(x_1, x_2, x_3, x_4)$ in terms of the parameters w being used by the code.

[1] I got the probabilities from the reverse-engineered heatmap here: `http://jemore.free.fr/wordpress/?p=73`.

Answer:
$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\exp(w_{x_1} + w_{x_2} + w_{x_3} + w_{x_4}).$$

2. How would the answer to the previous question change (in terms of w and v) if we use `E = [1 2]`?

   Answer:
   $$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\exp(w_{1,x_1} + w_{2,x_2} + w_{3,x_3} + w_{4,x_4} + v_{x_1,x_2,1}).$$

3. Modify the demo to use chain-structured dependency. Comment on whether this fixes each of the above 2 issues.

   Answer: With a Markov chain, it alleviates these issues but doesn't fix them. It now sometimes generates 1 2 3 4 and the conditional probability of 4 is now over 50%.

4. Modify the demo to use a completely-connected graph. Comment on whether this fixes each of the above 2 issues.

   Answer: With a full graph, it still isn't perfect but it's generating 1 2 3 4 close to 10% of the time and the probability of seeing a 4 after 1 2 3 is now over 70%.

5. What would the effect of higher-order potentials be? What would the disdavantages of higher-order potentials be?

   Answer: With threeway or a fourway potential you could get closer to the true frequencies in the data. However, the risk of overfitting would get much higher.

If you want to further explore UGMs, there are quite a few Matlab demos on the UGM webpage (`https://www.cs.ubc.ca/~schmidtm/Software/UGM.html`) that you can go through which cover all sorts of things like approximate inference and CRFs.

2 Bayesian Inference

2.1 Conjugate Priors

Consider a $y \in \{1, 2, 3\}$ following a multinoulli distribution with parameters $\theta = \{\theta_1, \theta_2, \theta_3\}$,
$$y \mid \theta \sim \text{Mult}(\theta_1, \theta_2, \theta_3).$$

We'll assume that θ follows a Dirichlet distribution (the conjugate prior to the multinoulli) with parameters $\alpha = \{\alpha_1, \alpha_2, \alpha_3\}$,
$$\theta \sim \mathcal{D}(\alpha_1, \alpha_2, \alpha_3).$$

Thus we have
$$p(y \mid \theta, \alpha) = p(y \mid \theta) = \theta_1^{I(y=1)}\theta_2^{I(y=2)}\theta_3^{I(y=3)}, \quad p(\theta \mid \alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\theta_3^{\alpha_3 - 1}.$$

Compute the following quantites:

1. The posterior distribution,
   $$p(\theta \mid y, \alpha).$$

Answer:

$$\begin{aligned}
p(\theta \mid y, \alpha) &\propto p(y \mid \theta, \alpha) p(\theta \mid \alpha) \\
&= p(y \mid \theta) p(\theta \mid \alpha) \\
&\propto \theta_1^{I(y=1)} \theta_2^{I(y=2)} \theta_3^{I(y=3)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \theta_3^{\alpha_3 - 1} \\
&= \theta_1^{I(y=1)+\alpha_1 - 1} \theta_2^{I(y=2)+\alpha_2 - 1} \theta_3^{I(y=3)+\alpha_3 - 1} \\
&= \theta_1^{(I(y=1)+\alpha_1) - 1} \theta_2^{(I(y=2)+\alpha_2) - 1} \theta_3^{(I(y=3)+\alpha_3) - 1}.
\end{aligned}$$

This is proportional to a Dirichlet distribution,

$$\theta \sim \mathcal{D}(I(y=1)+\alpha_1, I(y=2)+\alpha_2, I(y=3)+\alpha_3),$$

so we have

$$p(\theta \mid y, \alpha) = \frac{\Gamma(\sum_{j=1}^{3}[I(y=j)+\alpha_j])}{\prod_{j=1}^{3}\Gamma(I(y=j)+\alpha_j)} \theta_1^{I(y=1)+\alpha_1 - 1} \theta_2^{I(y=2)+\alpha_2 - 1} \theta_3^{I(y=3)+\alpha_3 - 1}$$

2. The marginal likelihood of y given the hyper-parameters α,

$$p(y \mid \alpha) = \int p(y, \theta \mid \alpha) d\theta,$$

Answer:

$$\begin{aligned}
p(y \mid \alpha) &= \int p(y, \theta \mid \alpha) d\theta \\
&= \int p(y \mid \theta) p(\theta \mid \alpha) d\theta \\
&= \int \prod_{j=1}^{3} \theta_j^{I(y=j)} \frac{1}{D(\alpha)} \prod_{j=1}^{3} \theta_j^{\alpha_j - 1} d\theta \\
&= \frac{1}{D(\alpha)} \int \prod_{j=1}^{3} \theta_j^{I(y=j)+\alpha_j} d\theta \\
&= \frac{D(\alpha^+)}{D(\alpha)}
\end{aligned}$$

This is the normalizing constant of the posterior divided by the normalizing constant of the prior.

3. The posterior mean estimate for θ,

$$\mathbb{E}_{\theta \mid y, \alpha}[\theta_i] = \int \theta_i p(\theta \mid y, \alpha) d\theta,$$

which (after some manipulation) should not involve any Γ functions.

Answer:

$$
\begin{aligned}
\theta_i &= \int \theta_i p(\theta \mid y, \alpha) d\theta \\
&= \int \theta_i \frac{\Gamma(\sum_{j=1}^3 \beta_j)}{\prod_{j=1}^3 \Gamma(\beta_j)} \prod_{j=1}^3 \theta_j^{\beta_j - 1} d\theta \\
&= \frac{\Gamma(\sum_{j=1}^3 \beta_j)}{\prod_{j=1}^3 \Gamma(\beta_j)} \int \theta_i \prod_{j=1}^3 \theta_j^{\beta_j - 1} d\theta \\
&= \frac{\Gamma(\sum_{j=1}^3 \beta_j)}{\prod_{j=1}^3 \Gamma(\beta_j)} \int \prod_{j=1}^3 \theta_j^{I(i=j) + \beta_j - 1} d\theta \\
&= \frac{\Gamma(\sum_{j=1}^3 \beta_j)}{\prod_{j=1}^3 \Gamma(\beta_j)} \frac{\prod_{j=1}^3 \Gamma(I(i=j) + \beta_j)}{\Gamma(\sum_{j=1}^3 I(i=j) + \beta_j)} \\
&= \frac{\Gamma(\sum_{j=1}^3 \beta_j)}{\prod_{j=1}^3 \Gamma(\beta_j)} \frac{\prod_{j=1}^3 \Gamma(I(i=j) + \beta_j)}{\Gamma(1 + \sum_{j=1}^3 \beta_j)} \\
&= \frac{\Gamma(\sum_{j=1}^3 \beta_j)}{\Gamma(1 + \sum_{j=1}^3 \beta_j)} \frac{\prod_{j=1}^3 \Gamma(I(i=j) + \beta_j)}{\prod_{j=1}^3 \Gamma(\beta_j)} \\
&= \frac{\beta_i}{\sum_{j=1}^3 \beta_j} \\
&= \frac{I(y=i) + \alpha_i}{\sum_{j=1}^3 [I(y=i) + \alpha_i]}.
\end{aligned}
$$

(The posterior mean minimizes the ℓ_2-risk when predicting on new data. So if all $\alpha_i = 1$, this gives a theoretical justification for the 'add one' trick that is often used when estimating probabilities. If you were interested in the ℓ_1-risk, you would instead take the posterior median.)

4. The posterior predictive distribution for a new independent observation $\tilde{y}$ given y,

$$p(\tilde{y} \mid y, \alpha) = \int p(\tilde{y}, \theta \mid y, \alpha) d\theta.$$

Answer:

$$
\begin{aligned}
p(\tilde{y} \mid y, \alpha) &= \int p(\tilde{y}, \theta \mid y, \alpha) d\theta \\
&= \int p(\tilde{y} \mid \theta, y, \alpha) p(\theta \mid y, \alpha) d\theta \\
&= \int p(\tilde{y} \mid \theta) p(\theta \mid \alpha) d\theta \\
&= \int \prod_{j=1}^3 [\theta_j^{I(\tilde{y}=j)}] \frac{1}{D(\alpha^+)} \prod_{j=1}^3 [\theta_j^{I(y=j) + \alpha_j - 1}] d\theta \\
&= \frac{1}{D(\alpha^+)} \int \prod_{j=1}^3 \theta_j^{I(\tilde{y}=j) + I(y=j) + \alpha_j - 1} d\theta \\
&= \frac{D(\alpha^{++})}{D(\alpha^+)},
\end{aligned}
$$

where $D(\alpha^{++})$ is the normalizing constant of the Dirichlet with parameters $(I(\tilde{y} = j) + I(y = j) + \alpha_j)$. Note that this is the marginal likelihood of the new data, if we treat the posterior we got from the old data as our prior.

Hint: You can use $D(\alpha) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1+\alpha_2+\alpha_3)}$ to represent the normalizing constant of the prior and $D(\alpha^+)$ to give the normalizing constant of the posterior. You will also need to use that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. For some calculations you may find it a bit cleaner to parameterize the posterior in terms of $\beta_j = I(y = j) + \alpha_j$, and convert back once you have the final result.

2.2 Empirical Bayes

Consider the model

$$y_i \sim \mathcal{N}(w^T \phi(x^i), \sigma^2), \quad w_j \sim \mathcal{N}(0, \lambda^{-1}),$$

where ϕ is a non-linear transformation of the features x^i (like a polynomial basis or RBFs). By using properties of Gaussians the marginal likelihood (marginalizing over all w_j) has the form

$$p(y \mid X, \sigma, \lambda) = (2\pi)^{-n/2}|C|^{-1/2} \exp\left(-\frac{y^T C^{-1} y}{2}\right),$$

which gives a negative log-marginal likelihood of

$$-\log p(y \mid X, \sigma, \lambda) = \log|C| + y^T C^{-1} y + \text{const.}$$

where

$$C = \sigma^2 I + \frac{1}{\lambda}\Phi(X)\Phi(X)^T,$$

As discussed in class, the marginal likelihood can be used to optimize hyper-parameters like σ, λ, and even the basis ϕ.

The demo *example_basis* loads a dataset and fits a degree-2 polynomial to it. Normally we would use a test set to choose the degree of the polynomial but here we'll use the marginal likelihood of the training set. Write a function, *leastSquaresEmpiricalBaysis*, that uses the marginal likelihood to choose the degree of the polynomial as well as the parameters λ and σ (you can assume that all λ_j are equal, and you can restrict your search for λ and σ to powers of 10). Hand in your code and report the marginally most likely values of the degree, σ, and λ.

Hint: the matrix C can be highly ill-conditioned and thus can cause Julia to do weird things, like saying $y^T C^{-1} y$ is negative even though C is positive-definite by construction. To compute the log-marginal likelihood in a more stable way, you can use the *chol* function (and may need to catch positive-definite errors when the code fails). You can then use the Cholesky factorization to compute $y^T C^{-1} y$ (by solving two linear systems) and $\log|C|$ (as twice the sum of the log of the diagonals).

Answer: This question was a mess, so I think anything should get full marks here. I believe that the reason it doesn't work is the lack of a hyper-prior on λ and σ^2, which seems to lead to some sort of degeneracy. If you click on the link in the lecture notes where I took this example from, it's even followed by a warning not to have a vague prior here. Sorry!

3 Very-Short Answer Questions

Give a short and concise 1-sentence answer to the below questions.

1. In UGMs, why is it easy to evaluate $\tilde{p}(x)$ but not $p(x)$?

   Answer: The reason $p(x)$ is hard to evaluate is due to Z, which isn't need to compute $\tilde{p}(x)$.

2. Why we need to have a "burn in" phase when using a Markov chain Monte Carlo approximation?

   Answer: We may start out in a state that is very unlikely in the stationary distribution, so our initial samples could be really biased.

3. Why do use the parameterization $\phi_j(s) = \exp(w_{j,s})$ in UGMs?

   Answer:

4. The main advantage is that it makes the NLL convex.

5. Suppose you are given the initial probabilities and transition probabilities in a Markov chain. Describe how you could set the parameter of a hidden Markov model so that the x_j follow the given (non-hidden) Markov chain.

   Answer: Set the probabilities over the z variables to be equal to the probabilities over the x variables, then make the probabilities over the x variables be an identity function.

6. What is the key advantage of the graph structure in restricted Boltzmann machines?

   Answer: It allows efficient blocks Gibbs sampling (visibles are independent given hidden, and vice versa).

7. What is the difference between a generative model and a discriminative model?

   Answer: Generative models $p(y, x)$, discriminative only models $p(y \mid x)$.

8. Why can fully-convolutional networks segment images of different sizes?

   Answer: The parameters are tied across space in the output layer.

9. What is the key feature of a "sequence to sequence" RNN?

   Answer: Both the inputs and outputs are sequences.

10. What are two advantages of the Bayesian approach to learning?

    Answer: Possible answers include optimal decisions, estimates of variance/confidence, optimal model selection/averaging criterion, relaxing the IID assumption, allowing infinite/unknown number of parameters.

11. What is the difference between the posterior distribution and the posterior predictive distribution?

    Answer: The posterior is over parameters, the posterior predictive is over new data.

12. What is the key property of a conjugate prior?

    Answer: The posterior comes from the same family.

4 Literature Survey

Reading academic papers is a skill that takes practice. When you first start out reading papers, you may find that you need to re-read things several times before you understand them, or that details will still be very fuzzy even after you've put a great amount of effort into trying to understand a paper. Don't panic, this is normal.

Even if you are used to reading papers from your particular sub-area, it can be challenging to read papers about a completely different topic. Usually, people in different areas use different language/notation and focus on very different issues. Nevertheless, many of the most-successful people in academia and industry are those that are able to understand/adapt ideas from different areas. (There are a ton of smart people in the world working on all sorts of amazing things, it's good to know how to communicate with as many of them as possible.)

A common technique when trying to understand a new topic (or reading scientific papers for the first time) is to read and write notes on 10 papers on the topic. When you read the first paper, you'll often find that it's hard to follow. This can make reading take a long time and might still leave you feeling that many things don't make sense; keep reading and trying to take notes. When you get to the second paper, it might still be very hard to follow. But when you start getting to the 8th or 9th paper, things often start making more sense. You'll start to form an impression of what the influential works in the area are, you'll start getting to used to the language and jargon, you'll start to understand what the main issues that people who work on the topic care about, and you'll probably notice some important references that weren't on your initial list of 10 papers. Ideally, you'll also start to notice how the topic has changed over time and you may get ideas of future work that you could do on the topic.

To help you make progress on your project or to give you an excuse to learn about a new topic, for this part you should write a literature survey of at least 10 academic papers on a particular topic. While your personal notes on the papers may be longer, the survey should be at most 4 pages of text (excluding references/tables/figures) in a format similar to the one for this document. Some logical components of a literature survey might be:

- A description of the overall topic, and the key themes/trends across the papers.
- A short high-level description of what was explored in each paper. For example, describe the problem being addressed, the key components of the proposed solution, and how it was evaluated. In addition, it is important to comment on the *why* questions: why is this problem important and why would this particular solution method make progress on it? It's also useful to comment on the strengths and weaknesses of the various works, and it's particularly nice if you can show how some works address the weaknesses of prior works (or introduce new weaknesses).
- One or more logical "groupings" of the papers. This could be in terms of the variant of the topic that they address, in terms of the solution techniques used, or in chronological terms.

Some advice on choosing the topic:

- The most logical/easy topic for your literature survey is a topic related to your course project, given that your final report will need a (shorter) literature survey included.
- If you are an undergrad, or a masters student without a research project yet, you may alternately want to choose a general area (like variance-reduced stochastic gradient, non-Gaussian graphical models, recurrent neural networks, matrix factorization, neural artistic style transfer, Bayesian optimization, etc.) as your topic.
- If you are a masters student that already has a thesis project, it could make sense to do a survey on a topic where ML intersects with your thesis (or where ML *could* intersect your thesis).
- If you are a PhD student, I would recommend using this an excuse to learn about a *completely different* topic than what you normally work on. Choose something hard that you would like to learn about, but previously haven't been to justify exploring carefully. This can be invaluable to your future research, because during your PhD it's often hard to allocate time to learn completely new topics.