Probabilistic Graphical Models Programming Assignment #5: Approximate Inference

This assignment is due at 11:59pm PDT (UDT -7) on 24 April 2012.

1 Introduction

Last week, we focused on implementing exact inference methods. Unfortunately, sometimes performing exact inference is intractable and cannot be done as performing exact inference in general networks is NP-hard. Luckily, there are a number of approximate inference methods and in this programming assignment we will investigate two of them: loopy belief propagation and Markov chain Monte Carlo (MCMC).

As you develop and test your code, you will run tests on a simple pairwise Markov network that we have provided. This Markov net is a 4 x 4 grid network of binary variables, parameterized by a set of singleton factors over each variable and a set of pairwise factors over each edge in the grid. This network is created by the function **ConstructToyNetwork.m**, and in this assignment, you will change some of its parameters and observe the effect this has on different inference techniques.

2 Loopy Belief Propagation

The first approximate inference method that we will implement is loopy belief propagation (LBP). This algorithm takes a cluster graph, a set of factors, and a list of evidence, and outputs the *approximate* posterior marginals for each variable. *The message passing framework you implemented for clique trees in PA4 should directly generalize to the case of cluster graphs,* so you should have to write relatively little further code for LBP. In LBP, we do not have a root and a specific up-down message passing order relative to that root. In particular, we can order messages by an arbitrary criterion. However, we want you to experiment with two different message passing orders.

You will first implement a naive message passing order that blindly iterates through all messages without considering other criteria. You will subsequently experiment with alternative message passing orders and analyze their impact on both the convergence of LBP and the values of the marginals at convergence.

- **NaiveGetNextClusters.m (5 points)** — This function should find a clique that is ready to transmit a message to its neighbor. It should return the indices of the two cliques the message is ready to be passed between. In this naive implementation we simply iterate over the cluster pairs, details on this ordering is given within the code file.

Now we can begin the message passing process to calibrate the cluster, but we'll need some infrastructure to make this work. For example, we need a criterion to tell us when we've converged and we need to create a cluster graph in order to run LBP.

- **CreateClusterGraph.m (5 points)** – Given a list of factors, this function will create a Bethe cluster graph with nodes representing single variable clusters and pairwise clusters. A Bethe cluster graph is perhaps the simplest way to create a cluster graph from a network as

we create a separate cluster for every factor and attach them so as to meet the requirements of a cluster graph. The ClusterGraph data structure is the same as the CliqueTree data structure from PA4, but cliqueList is renamed as clusterList.

- **CheckConvergence.m (2 points)** – Given your current set of messages and the set of messages that immediately preceded each of our current messages, this function will determine if the cluster graph has converged, returning 1 if so and 0 otherwise. In this case, we say that the messages have converged if no messages have changed significantly, where "significantly" means by a value of over 10^{-6} in any entry (note: for theoretic calibration you would actually wait until the difference was 0).

- **ClusterGraphCalibrate.m (10 points)** — This function should perform loopy belief propagation over a given cluster graph. It should return an array of final beliefs (factors) for each cluster.

We are now ready to bring together all the steps described above to compute approximate marginals for the variables in our network. You should call the appropriate functions in the file **ComputeApproxMarginalsBP.m** to run approximate inference.

- **ComputeApproxMarginalsBP.m (5 points)** – This function should take a set of initial factors and vector of evidence and compute the approximate marginal probability distribution for each variable in the network. You should be able to reuse much of the code you wrote for **ComputeExactMarginalsBP.m** for assignment 4 in this function.

2.1 LBP Investigation Questions

Answer these questions in the assignment quizzes section on the course website.

1. **(5 points)** – Recall our function **CheckConvergence.m** which uses the difference between a message before and after it is updated (called the "residual") as a criterion for convergence. While running LBP with our naive message ordering on the network created by **ConstructRandNetwork** with on-diag weight .3 and off-diag weight .7, print out and plot the residuals of the message $19 \rightarrow 3, 15 \rightarrow 40$, and $17 \rightarrow 2$, with the iteration number on the x-axis (you may want to change the range of the y-axis). Do these messages converge at the same rate? Which converges fastest? (Note, it will be easiest do this assessment within **ClusterGraphCalibrate.m** and use the helper function **MessageDelta** within that file).

2. **(4 points)** – Can changing the message passing order affect convergence in cluster graph calibration? Can it affect the value of the final marginals? Why or why not?

3. **(5 points)** – Now, consider the toy image network constructed in **ConstructToyNetwork.m**. Change the values of the on- and off-diagonal weights of the pairwise factors in this network to different values (which can be done by changing the values passed to this function). First try making the the weights on the diagonal much larger than the off-diagonal weights (1 and .2 respectively), then try the opposite where the off-diagonal weights are much larger(.2 and 1), and then finally try the case where the weights are roughly equal (.5 and .5). For each such model, run LBP and exact inference (using your code from PA4). How do the results on these varied types of graphs compare? Why?

2.2 Improving Message Passing

Taking the analysis of our naive message passing order that you already performed as inspiration, can you think of ways to improve the message passing order? Perhaps you could use the magnitude of the changes in the marginals before and after message passing to help indicate which messages you should prioritize passing. There exist better message passing schedules that improve the convergence of the algorithm, but that is beyond the scope of this Programming Assignment. Therefore, we won't ask you to implement it. For more information, we refer the interested readers to relevant research papers [Elidan, McGraw, and Koller 06] [1].

3 MCMC

Now that we've looked at belief propagation, we will investigate a second class of inference methods based on Markov Chain Monte Carlo (MCMC) sampling. As a reminder, a Markov chain defines a transition model $T(x \rightarrow x')$ between different states x and x'. The chain is initialized to some initial assignment x_0. At each iteration t, a new state, x_{t+1} is sampled from the transition model. The chain is run for some number of iterations, over which a subset of the samples is collected. The collected samples are then used to estimate statistics such as the marginals of individual variables. In this assignment, you will implement Gibbs and Metropolis-Hastings sampling, both of which are MCMC methods that sample from the posterior of a probabilistic graphical model.

A critical issue in the utility of a Markov chain is the rate at which it mixes to the stationary distribution. For example, if the stationary distribution has two modes that are far apart in the state space and the MCMC transition probability only allows local moves in the state space, it is likely that the state of the Markov chain will get stuck near one of the modes. This affects both the number of samples required before the chain forgets its initial state, and the quality of the estimates – unless many samples are collected, most samples are likely to come from one mode or another. If samples are aggregated before a chain has mixed, then the distribution from which they are drawn will be biased toward the initial state and thus will not be a good approximation to the stationary distribution. In the starter code, we provide you with a visualization function, **VisualizeMCMCMarginals.m**, that allows you to analyze a Markov chain to estimate properties such as mixing time and whether or not it is getting stuck in a local optimum. See the description of the code infrastructure below for more details on this function.

3.1 Gibbs

Recall that the Gibbs chain is a Markov chain where the transition probability $T(x \rightarrow x')$ is defined as follows. We iterate over the variables in some fixed order, say $X_1, ..., X_n$. For the variable X_i, we sample a new value from $P(X_i|x_{-i})$ (which is just $P(X_i|\text{MarkovBlanket}(X_i))$), and update its new value. Note that the terms on the right-hand-side of the conditioning bar use the newly sampled assignment to the variables $X_1, ..., X_{i-1}$. Once we sample a new value for each of $X_1, ..., X_n$, the result is our new sample x'. Our first task for the MCMC task is thus to implement a function that computes and samples from $P(X_i|\text{MarkovBlanket}(X_i))$ as well as a helper function to produce sampling distributions. You will then use this function as a transition probability for MCMC sampling. (Recall that a Markov Blanket of a node X is the set of nodes Y such that X_i is independent of all other nodes given Y, thus it consists of X's parents, children, and children's parents.)

[1] `http://www.robotics.stanford.edu/~koller/Papers/Elidan+al:UAI06.pdf`

- **BlockLogDistribution.m: (5 points)** – This is the function that produces the sampling distribution used in Gibbs sampling and (possibly) versions of Metropolis- Hastings. It takes as input a set of variables $\mathbf{X}_I$ and an assignment $\mathbf{x}$ to all variables in the network, and returns the distribution associated with sampling $\mathbf{X}_I$ as a block (i.e. the variables are constrained to take on the same value) given the joint assignment to all other variables in the network. That is, for each value l, we compute the (unnormalized) probability $\tilde{P}(\mathbf{X}_I = l|\mathbf{x}_{-I})$ where $\mathbf{X}_I = l$ is shorthand for the statement"$X_i = l$ for all $X_i \in \mathbf{X}_I$" and $\mathbf{x}_{-I}$ is the assignment to all other variables in the network. Your solution should only contain one for-loop (because for-loops are slow in Matlab). Note that the distribution will be returned in log-space to avoid underflow issues.

- **GibbsTrans.m (5 points)** – This function defines the transition process in the Gibbs chain as described above (ie. we iteratively resample X_i from $P(X_i|\text{MarkovBlanket}(X_i))$ for each i) . It should call BlockLogDistribution to sample a value for each variable in the network.

3.1.1 Running Gibbs Sampling and Questions

Now that our first transition process has been defined, we need to enact a general framework for running our variants of MCMC.

- **MCMCInference.m PART 1 (3 points)**– This function defines the general framework for conducting MCMC inference. It takes as input a probabilistic graphical model (a redundant but convenient data structure), a set of factors, a list of evidence, the name of the MCMC transition to use, and other MCMC parameters such as the target burn-in time and number of samples to collect. As a first step, you only need to implement the logic that transitions the Markov chain to its next state and records the sample.

With the inference engine, let's try to understand the behavior of Gibbs sampling (answer online):

1. **(5 points)** – Let's run an experiment using our Gibbs sampling method. As before, use the toy image network and set the on-diagonal weight of the pairwise factor (in **ConstructToyNetwork.m**) to be 1.0 and the off-diagonal weight to be 0.1. Now run Gibbs sampling a few times, first initializing the state to be all 1's and then initializing the state to be all 2's. What effect does the initial assignment have on the accuracy of Gibbs sampling? Why does this effect occur?

3.2 Metropolis-Hastings

Metropolis-Hastings is a general framework (within the even more general framework of MCMC) that defines the Markov chain transition in terms of a proposal distribution $Q(x \to x')$ and an acceptance probability $A(x \to x')$. The proposal distribution and acceptance probability must satisfy the detailed balance equation in order to generate the correct stationary distribution. It turns out that a satisfying acceptance probability is given as follows (where π is the stationary distribution):

$$A(x \to x') = \min\left[1, \frac{\pi(x')Q(x' \to x)}{\pi(x)Q(x \to x')}\right]$$

In this section of the assignment, you will implement a general Metropolis-Hastings framework that is capable of utilizing different proposal distributions, specifically the uniform distribution

and the Swendsen-Wang distribution (described later). We will provide you with the implementations of these proposal distributions and you will need to compute the correct acceptance probability. Furthermore, you will study the relative merits of each proposal type. To start, let's implement the uniform proposal distribution:

- **MHUniformTrans.m (5 points)**– This function defines the transition process associated with the uniform proposal distribution in Metropolis-Hastings. You should fill in the code to compute the correct acceptance probability.

Now that we have that baseline, let's move on to Swendsen-Wang. Swendsen-Wang was designed to propose more global moves in the context of MCMC for pairwise Markov networks of the type used for image segmentation or Ising models, where adjacent variables like to take the same value. At its core, it is a graph node clustering algorithm. Given a pairwise Markov network and a current joint assignment x to all variables, it generates clusters as follows: first it eliminates all edges in the Markov network between variables that have different values in x. Then, for each remaining edge $\{i, j\}$, it activates the edge with some probability $q_{i,j}$ (which can depend on the variables i and j but not on their values in x). It then computes the connected components of the graph over the activated edges. Finally, it selects one connected component, $\mathbf{Y}$, uniformly at random from all connected components. Note that all nodes in $\mathbf{Y}$ will have the same label l. We then (randomly) choose a new value l' that will be taken by all nodes in this connected component. These variables are then updated in the joint assignment to produce the new assignment x'. In other words, the new assignment x' is the same as x, except that the variables in $\mathbf{Y}$ are all labeled l' instead of l. Note that this proposed move flips a large number of variables at the same time, and thus it takes much larger steps in the space than a local Gibbs or Metropolis-Hastings sampler for this Markov network.

Let $q(Y|x)$ be the probability that a set $\mathbf{Y}$ is selected to be updated using this procedure. It is possible to show that

$$\frac{q(\mathbf{Y}|x')}{q(\mathbf{Y}|x)} = \frac{\prod_{(i,j)\in\mathcal{E}(\mathbf{Y},(\mathbf{X}'_{\mathbf{l}'}-\mathbf{Y}))}(1-q_{i,j})}{\prod_{(i,j)\in\mathcal{E}(\mathbf{Y},(\mathbf{X}_{\mathbf{l}}-\mathbf{Y}))}(1-q_{i,j})} \tag{1}$$

where: $\mathbf{X_l}$ is the set of vertices with label l in x, $\mathbf{X}'_{\mathbf{l}'}$ is the set of vertices with label l' in $\mathbf{x}'$; and where $\mathcal{E}(\mathbf{Y}, \mathbf{Z})$ (between two disjoint sets $\mathbf{Y}, \mathbf{Z}$) is the set of edges connecting nodes in $\mathbf{Y}$ to nodes in $\mathbf{Z}$. (NOTE: The log of the quotient in equation 1 is called log_QY_ratio in the code.) Then we have that

$$\frac{\mathcal{T}^Q(x' \to x)}{\mathcal{T}^Q(x \to x')} = \frac{q(\mathbf{Y}|x')}{q(\mathbf{Y}|x)} \frac{R(\mathbf{Y} = l|x'_{-\mathbf{Y}})}{R(\mathbf{Y} = l'|x_{-\mathbf{Y}})} \tag{2}$$

where: $R(\mathbf{Y} = l'|x_{-\mathbf{Y}})$ is a distribution specified by you for choosing the label l' for $\mathbf{Y}$ given $x_{-\mathbf{Y}}$ (the assignment to all variables outside of $\mathbf{Y}$). Note that $x_{-\mathbf{Y}} = x'_{-\mathbf{Y}}$. In this assignment, the code for generating a Swendsen-Wang proposal is given to you, but you will have to compute the acceptance probability and use that to define the sampling process for the Markov chain. You will implement 2 variants that experiment with different parameters for the proposal distribution. In particular, you will change the value of the $q_{i,j}$'s and $R(\mathbf{Y} = l|x_{-\mathbf{Y}})$. The two variants are as follows:

1. Set the $q_{i,j}$'s to be uniformly 0.5, and set the distribution R to be uniform.
2. Set R to be the block-sampling distribution (as defined in BlockLogDistribution.m) for sampling a new label and make $q_{i,j}$ dependent on the pairwise factor $F_{i,j}$ between i and j.

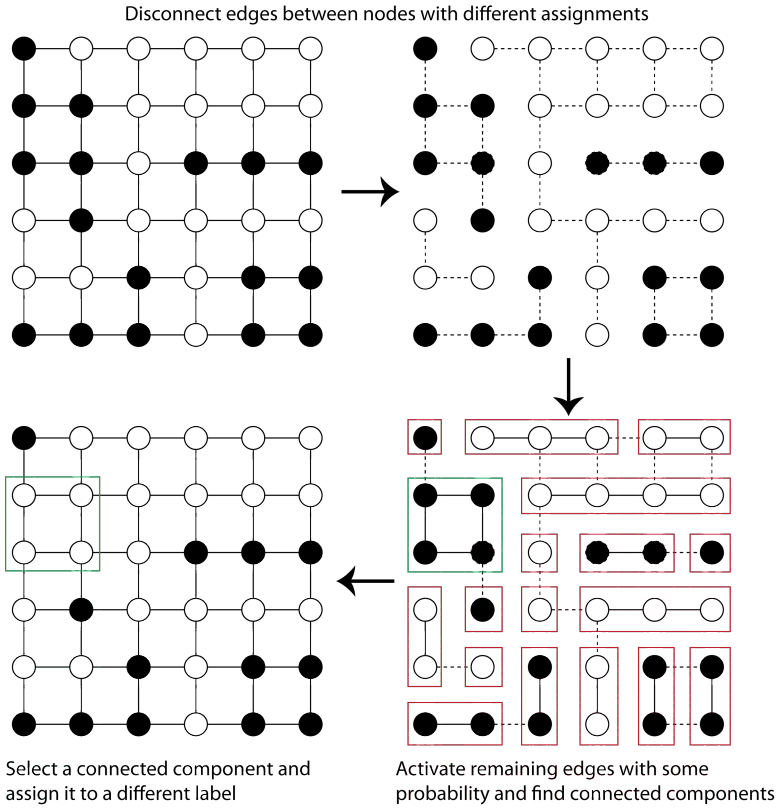

Figure 1: Visualization of Swendsen-Wang Procedure

In particular, set

$$q_{i,j} := \frac{\sum_u F_{i,j}(u,u)}{\sum_{u,v} F_{i,j}(u,v)}$$

- **MHSWTrans.m (Variant 1) (3 points)** – This function defines the transition process associated with the Swendsen-Wang proposal distribution in Metropolis-Hastings. You should fill in the code to compute the proposal distribution values and then use these to compute the acceptance probability. Implement the first variant for this test.
- **MHSWTrans.m (Variant 2) (3 points)** – Now implement the second variant of SW. Note: the first variant should still function after the second variant has been implemented.

With Swendsen-Wang, we will need to compute the values of our $q_{i,j}$'s, so we must update our inference engine:

- **MCMCInference.m PART 2 (4 points)**– Flesh this function out to run our Swendsen-Wang variants in addition to Gibbs. Your task here is to implement the calculations of the $q_{i,j}$'s for both variants of Swendsen-Wang in this function. (The reason that this is done here and not in MHSWTrans.m is to improve efficiency.)

Now that we've finished implementing all of these functions, let's compare these inference algorithms.

1. **(10 points)** – For this question, repeat the experiment for LBP where we ran our Toy Network while changing the on and off-diagonal network. Again, we will consider the cases where the on-diagonal weights are much larger, much smaller, and about equal to the off-diagonal weights. While running this, use **VisualizeMCMCMarginals.m** to visualize the distribution of the Markov chain for **multiple** runs of MCMC (see **TestToy.m** for how to do this). We will examine how the mixing behavior and final marginals for each chain change in response to the change in the pairwise factor.
   (a) **(5 points)** – Set the on-diagonal weight of our toy image network to 1 and off-diagonal weight to .2. Now visualize **multiple** runs with each of Gibbs, MHUniform, Swendsen-Wang variant 1, and Swendsen-Wang variant 2 using **VisualizeMCMC-Marginals.m** (see **TestToy.m** for how to do this). How do the mixing times of these chains compare? How do the final marginals compare to the exact marginals? Why?
   (b) **(5 points)** – Set the on-diagonal weight of our toy image network to .5 and off-diagonal weight to .5. Now visualize **multiple** runs with each of Gibbs, MHUniform, Swendsen-Wang variant 1, and Swendsen-Wang variant 2 using **VisualizeMCMC-Marginals.m** (see **TestToy.m** for how to do this). How do the mixing times of these chains compare? How do the final marginals compare to the exact marginals? Why?
2. **(3 points)** When creating our proposal distribution for Swendsen-Wang, if you set all the $q_{i,j}$'s to zero, what does Swendsen-Wang reduce to?

4 Conclusion

Congratulations! You've now implemented a full suite of inference engines for exact inference, approximate inference, and sampling. These methods are useful for making predictions and gaining understanding of the world around us, a process we'll explore more in PA6. Of course, one underlying assumption has been that we already know the basic facts of which variables influence one another – an assumption that is certainly not always the case! So stay tuned, we'll look into how to eliminate more of our assumptions later in the course.

5 Infrastructure Reference

A few methods you may find useful:

1. **exampleIOPA5.mat:** Mat-file containing example input and output corresponding to the 13 preliminary tests for PA5. For argument j of the function call in part i, you should use **exampleINPUT.t**$\#_i$**a**$\#_j$ (replacing the $\#_i$ with i). If there are multiple function calls in one test (for example, we iterate over multiple inputs) then for iteration k you should reference **exampleINPUT.t**$\#_i$**a**$\#_j$.$\{\#_k\}$. For output, look at **exampleOUTPUT.t**$\#_i$ for the output to part i. If there are multiple outputs or iterations, the functionality is the same as for the input example.
2. **ConstructToyNetwork.m:** Function that constructs a toy pairwise Markov Network that you will use in some of the questions. This function accepts two arguments, an "on-diagonal" weight and an "off-diagonal" weight. These refer to the weights in our image network's pairwise factors, where on-diagonal refers to the weight associated with adjacent nodes agreeing and off-diagonal corresponds to having different assignments. The output network will be a 4 x 4 grid.

3. **ConstructRandNetwork.m:** Function that constructs a randomized pairwise Markov Network that you will use in some of the questions. The functionality is essentially the same as **ConstructToyNetwork.m**.

4. **VisualizeMCMCMarginals.m:** This displays two things. First, it displays a plot of the log-likelihood of each sample over time. Recall that in quickly mixing chains, this value should increase until it roughly converges to some constant log-likelihood, where it should remain (with some occasional jumps down). This function also visualizes the estimate of the marginal distributions of specified variables in the network as estimated by a string of samples obtained from MCMC over time. In particular, it takes a fixed-window subset of the samples around a given iteration t and uses these to compute a sliding-window average of the estimated marginal(s). It then plots the sliding-window average of each value in the marginal as its estimate progresses over time. The function also can accept samples from more than one MCMC run, in which case the marginal values that correspond to one another are plotted in the same color, allowing you to determine whether the different MCMC runs are converging to the same result. This is particularly helpful if you are trying to identify whether the chain is susceptible to local optima (in which case, different runs will converge to different marginals) or whether the chain has mixed by a given iteration.

5. **TestToy.m:** This function constructs a toy image network where each variable is a binary pixel that can take on a value of 1 or 2. This network is a pairwise Markov net structured as a 4 x 4 grid. The parameterization for this network can be found in TestToy.m and you will tune the parameters of the pairwise factors to study the corresponding behavior of different inference techniques. You can visualize the marginal strengths of this toy image by calling the function VisualizeToyImageMarginals.m, which will display the marginals as a gray-scale image, where the intensity of each pixel represents the probability that that pixel has the label 2.

6. **VisualizeToyImageMarginals.m:** Visualizes the marginals of the variables in the toy network on a 4x4 grid. We have provided a lot of the infrastructure code for you so that you can concentrate on the details of the inference algorithms.