ISTA 410/510 Final (take home)

For contribution to the final grade, due dates, current late policy, and instructions for handing the assignment in, see the assignment web page.

Please create a PDF document with your answers and/or the results of any programs that you write. You should also hand in your programs. However, the programs will not necessarily be consulted. All information for grading should be in the PDF. Because this final is structured as an assignment, some help can be provided by the instructor in office hours.

There is a total of 30 points. We will grade it out of 20 for grad students, and out of 12 for undergrads. Modest bonus points are available if you choose to hand in more than what is required.

1. (++) Consider a square image which has 5 circles as well as some additional "clutter" which is not of interest. You have the output of an edge detector. With 5% probability, the detector fires at a given pixel due to noise (or clutter) in the image which is uniformly distributed over the image. In addition, if the circle intersects the pixel (think about a graphics program drawing circles), there is a 80% that the detector will fire. You can assume that for a given hypothesis about circle, you have a function that tells you which pixels they occupy. The circles themselves have a tendency to be closer to the center of the image, but they can occur anywhere. More specifically, the probability that a circle center is at an image corner is 10% that of it being in the center, and in-between locations have probabilities in proportion. Notationally, the density is in proportion to

$$p(c_x, c_y) \propto \left\{ 1 - \left(\frac{9}{10} \right) \frac{\sqrt{2}\sqrt{c_x^2 + c_y^2}}{w} \right\}$$, w is image width, image center is (0,0).

However, this formula is meant only to help you imagine the problem setup, and you can just use $p(c_x, c_y)$ in your model. Similarly circle radii tend to be 10% of the image width, but have some variance. The size and position parameters are independent.

Produce a Bayesian model for 5 circles, and N observed edge points, introducing notation as you need it. For details not specified, make reasonable assumptions. Provide a formulation for the joint distribution over model parameters, and the N observed points. Identify priors and likelihoods to demonstrate that you understand these terms.

(3 points total)

2. Consider all the words in a book put into a book word bag, B (i.e., ignore the order). Since we are ignoring order, a lot of information is clearly lost. Consider also a big bag of words, L (for library) that is representative of all books. IE, take the words of all books and put them into one big bag.

(a) Consider reducing L to a frequency distribution that is indicative of probability of word occurrences in written English. Does this have any structure (or information content) at all? Explain.

(b) Consider the set of all frequency distributions for books. Do these distributions have any further structure than L, or can we consider them (i.e., the words in a book bag) as simply samples of L?

(c) Create a graphical model (and provide the graphs as a picture) for the words book bags based on the following idea. Books are on subjects, and subjects imply a distribution over topics. For example, Bayesian modeling books contain the topic "conjugate priors". Words come topics, independent of subject. Are the implications of this model consistent with your answer for (b)?

(3 points total)

3. **A simple sampler**. The code provided on the website, secret.m, is a function that maps any pair of real numbers (but we will assume they are integers) into a probability density. You can assume the real action is in the range of (-25,25)x(-25,25). Consider exploring this function based on the grid points using the basic Metropolis algorithm. A state is a grid location, and the q() function changes to another grid location.

   (a) Propose a q() function that can potentially visit every state. You can assume that secret() has some local structure, so moves that switch to an adjacent grid location might make sense. Make sure that your q() function is symmetric, otherwise the problem is harder.

   (b) Write a computer program to sample the space from a random starting point in the above range. Keep a running count of the number of times the sampler visited each point in the above range. Your sampler should have a non-zero change of leaving that range, and just ignore those samples while you are counting. Visualize the count matrix. One way to do that is with:

```
figure;
colormap('gray');
imagesc(counts);
```

   Provide three different images for each of 1000, 10000, and 100000 iterations (nine figures total). Is the starting point being forgotten?

(6 points total).

4. Consider a set of N states S={1,2,3,...,N} to be visited by a MCMC sampler. Consider the distance between states S1 and S2 to be |S1-S2|. For example, the distance between state 4 and state 7 is 3. Suppose that the probability of transitioning from S1 to S2 is proportional to a Gaussian distribution over x=(S1-S2) with mean zero and variance five.

   (a) (++) Write a computer program to generate a properly normalized transition matrix and create and image to display it with blocks of darker or brighter shades indicating lower or higher probability. In Matlab, the following might do the trick (stretch to get bigger blocks),

```
figure;
colormap('gray');
imagesc(T);
```

   but you may need to further scale the transitions non-linearly to view it properly.Provide an image for N=20.

   (b) For your N=20 matrix, find a stationary probability vector, and produce Matlab (or other computer program output) that (1) shows how you found that vector, and (2) verifies that it is in fact a stationary probability vector. You can cut and paste the Matlab output into your PDF.

   (c) (++) Recall that we considered that we can start with any probability vector, and successive applications of the transition function will lead to a stationary vector. Use a uniform vector as your initial probability vector. Measure the convergence by the differences between the vectors of successive iterations by $R=\|V1-V2\| / (\|V1\| +\|V2\|)$. Define convergence time, T, by the number of iterations to get R below a certain threshold (try 1e-08). For some reasonable definition of R, plot T versus N for some reasonable values of N. You should have at least 10 different values of N on your plot. What defines "reasonable"? You want to either establish that there is a systematic effect on changing N that is exposed by plotting, or establish that there is probably no such effect. R must be small enough to capture the difference, and your values of N (which need not be a linear, it could be exponential) should make an interesting plot, or you need to explore a big enough range of N to suggest that the plot is not likely to get interesting.

   (d) Redo (c) but now plot the magnitude of the second eigenvalue against N. Comment on what you have found

   (8 points total)

5. (+++) Consider the theorem from Neal 93 and its proof which are reproduced in the following two pages. The proof is given, but providing more details helps us understand it. Provide the following details.

   (i) More detailed justification (sub-proof) of the claim that $v \leq 1$.

   (ii) A brief justification of steps 3.17 through 3.22. In most cases, referring to a simple fact (e.g., "commutativity of addition" will suffice, but in some cases you may want to add some intermediate steps.

   (iii) Confirm by doing it that it is "easy to show" $\sum_x r_{\bar{n}+1}(x) = 1$.

   (iv) A brief justification of steps 3.23 through 3.30. In most cases, referring to a simple fact (e.g., "commutativity of addition" will suffice, but in some cases you may want to add some intermediate steps.

(4 points total)

FUNDAMENTAL THEOREM. *If a homogeneous Markov chain on a finite state space with transition probabilities $T(x, x')$ has π as an invariant distribution and*

$$\nu \quad = \quad \min_x \min_{x':\pi(x')>0} T(x, x')/\pi(x') \quad > \quad 0 \tag{3.12}$$

then the Markov chain is ergodic, i.e., regardless of the initial probabilities, $p_0(x)$

$$\lim_{n\to\infty} p_n(x) \quad = \quad \pi(x) \tag{3.13}$$

for all x. A bound on the rate of convergence is given by

$$|\pi(x) - p_n(x)| \quad \leq \quad (1-\nu)^n \tag{3.14}$$

Furthermore, if $a(x)$ is any real-valued function of the state, then the expectation of a with respect to the distribution p_n, written $E_n[a]$, converges to its expectation with respect to π, written $\langle a \rangle$, with

$$|\langle a \rangle - E_n[a]| \quad \leq \quad (1-\nu)^n \max_{x,x'} |a(x) - a(x')| \tag{3.15}$$

Specifically, we will see that the distribution at time n can be written as

$$p_n(x) \;=\; [1-(1-\nu)^n]\,\pi(x) \;+\; (1-\nu)^n\, r_n(x) \tag{3.16}$$

with r_n being a valid probability distribution. Note that $\nu \leq 1$, since we cannot have $\pi(x') < T(x,x')$ for all x'. The above formula can be satisfied for $n = 0$ — just set $r_0(x) = p_0(x)$. If it holds for $n = \bar{n}$, then

$$p_{\bar{n}+1}(x) \;=\; \sum_{\tilde{x}} p_{\bar{n}}(\tilde{x})\, T(\tilde{x},x) \tag{3.17}$$

$$\;=\; [1-(1-\nu)^{\bar{n}}] \sum_{\tilde{x}} \pi(\tilde{x})\, T(\tilde{x},x) \;+\; (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\, T(\tilde{x},x) \tag{3.18}$$

$$\;=\; [1-(1-\nu)^{\bar{n}}]\,\pi(x) \;+\; (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x}) \left[T(\tilde{x},x) - \nu\pi(x) + \nu\pi(x)\right] \tag{3.19}$$

$$\;=\; [1-(1-\nu)^{\bar{n}}]\,\pi(x) \;+\; (1-\nu)^{\bar{n}}\,\nu\,\pi(x) + (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x}) \left[T(\tilde{x},x) - \nu\pi(x)\right] \tag{3.20}$$

$$\;=\; [1-(1-\nu)^{\bar{n}+1}]\,\pi(x) \;+\; (1-\nu)^{\bar{n}+1} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\, \frac{T(\tilde{x},x) - \nu\pi(x)}{1-\nu} \tag{3.21}$$

$$\;=\; [1-(1-\nu)^{\bar{n}+1}]\,\pi(x) \;+\; (1-\nu)^{\bar{n}+1}\, r_{\bar{n}+1}(x) \tag{3.22}$$

where $r_{\bar{n}+1}(x) = \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu\pi(x)]/(1-\nu)$. From (3.12), we find that $r_{\bar{n}+1}(x) \geq 0$ for all x. One can also easily show that $\sum_x r_{\bar{n}+1}(x) = 1$. The $r_{\bar{n}+1}(x)$ therefore define a probability distribution, establishing (3.16) for $n = \bar{n}+1$, and, by induction, for all n.

Using (3.16), we can now show that (3.14) holds:

$$|\pi(x) - p_n(x)| \;=\; |\,\pi(x) \;-\; [1-(1-\nu)^n]\,\pi(x) \;-\; (1-\nu)^n\, r_n(x)\,| \tag{3.23}$$

$$\;=\; |\,(1-\nu)^n\,\pi(x) \;-\; (1-\nu)^n\, r_n(x)\,| \tag{3.24}$$

$$\;=\; (1-\nu)^n\,|\pi(x) - r_n(x)| \tag{3.25}$$

$$\;\leq\; (1-\nu)^n \tag{3.26}$$

We can show (3.15) similarly:

$$|\langle a\rangle - E_n[a]| \;=\; \left|\,\textstyle\sum_{\tilde{x}} a(\tilde{x})\,\pi(\tilde{x}) \;-\; \sum_{\tilde{x}} a(\tilde{x})\, p_n(\tilde{x})\,\right| \tag{3.27}$$

$$\;=\; \left|\,\textstyle\sum_{\tilde{x}} a(\tilde{x})\,[(1-\nu)^n\pi(\tilde{x}) \;-\; (1-\nu)^n r_n(\tilde{x})]\,\right| \tag{3.28}$$

$$\;=\; (1-\nu)^n \left|\,\textstyle\sum_{\tilde{x}} a(\tilde{x})\,\pi(\tilde{x}) \;-\; \sum_{\tilde{x}} a(\tilde{x})\, r_n(\tilde{x})\,\right| \tag{3.29}$$

$$\;\leq\; (1-\nu)^n \max_{x,\,x'} |a(x) - a(x')| \tag{3.30}$$

This completes the proof.

6. (+++++) Implement a Metropolis Hastings solution for the Gaussian Mixture Model that we have already done using EM in a previous assignment. By using your EM code, time spent on coding should not be overly high because much of the problem infrastructure is in place. Redo experiments from question 1 of assignment 5, and report whether you are able to find a better solution.

   Alternatively, if you worked with the vision libraries in assignment 5, you could consider doing the above in the context of the vision library.

   Alternatively, if you want to explore a Metropolis Hastings implementation in some other context, this can be considered. Contact Kobus with a proposal.

(6 points total)