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.

Because this is a midterm, you should not post questions to the maillist. Rather, send requests for clarification to the instructor.

There is a total of 22 points. Because the discretizations are not so clean as the other two take home assignments, we will simply grade it out of 20 for grad students, and out of 14 for undergrads. Different from before, bonus points are available if you choose to hand in more than what is required.

1. (++) Consider a process that generates points which are the result of an edge detector running on an image. We will assume that images are square. The process produces points that are uniform distributed with 50% probability. With 50% probability, the process produces points from one of 5 circles. The circles 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. Similarly circle radii tend to be 10% of the image width, but have some variance.

   Produce a Bayesian model for N observed edge points. For details not specified, make reasonable assumptions. Provide a formulation for the distribution over model parameters based on 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 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 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 at all? Explain.

   (b) Consider the set of all frequency distributions for books. Does this set of distributions have any further structure than L, or can we consider them as simply samples of L?

   (c) Create a graphical model (with a picture) for the words in a bag 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. 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|. 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 one.

   (a) (++) Write a computer program to generate a properly normalized transition matrix and create images that display it. Specifically, you should output images of size B*N by B*N, where B is the size of image blocks that are all the same shade of gray. So, your image has N blocks in each of N rows. Your blocks should be brighter where the transition probability is higher. You may need to fiddle with the mapping from probability to brightness to effectively visualize the matrix. For example, you could consider brightness in proportion to the log of the probability. Provide an image for N=20, and B=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.

   (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. 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 eigen value against N. Comment on what you have found

(8 points total)

4. (++++) 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)

5. (+++) Implement a Metropolis Hastings solution for the Gaussian Mixture Model that we have already done an assignment on. 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.

(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.