Homework 12 due April 24

Select five problems from below or make up some of your own.

1. Consider a set consisting of two hundred points in the plane. One hundred of the points are equally spaced on one circle and the other hundred are equally space on a second circle. The two circles have the same center but different radii. Calculate a similarity matrix between pairs of points. The ijth element of the matrix is the similarity of the ith and jth points. We would like to cluster the points. We hope that our clustering algorithm groups the points nto two clusters depending on which circle they are on. One technique is to use a Gaussian kernel. Define the similarity of two points i and j by $e^{-d_{ij}^2/2\sigma^2}$. Find the first two singular vectors of the similarity matrix and plot the points using the coordinates of these vectors. What happens? If the points are not appropriately clustered see if you can modify the problem to make it work.

2. Using the ideas in problem 1 above, write an algorithm to remove duplicates from a mailing list. The problem here is that the names and addresses of duplicates may not be identical. For example, one copy of a persons name may use initials for the first and middle name and the other may write out the first name and use only a middle initial.

3. Consider a coin that comes down heads with probability a. Prove that the expected number of flips before a heads occurs is 1/a.

4. How does one produce a random event with probability $1/2^k$ using only $\log k$ bits.

5. Randomly generate a string $x_1 x_2 \cdots x_n$ of 10^6 0's and 1's with probability ½ of x_i being a 1. Count the number of ones and also estimate the number of ones by the approximate counting algorithm. Repeat the process for p=1/4, 1/8. How close is the approximation?

6. Define a hash function $h:\{1,2,\cdots,m\} \to \{1,2,\cdots,t\}$ that requires only $O(\log t)$ bits.

7. (a) What is the variance of the above method of counting the number of occurrences of a 1 with $\log\log n$ memory?
(b) Can the algorithm be iterated to use only $\log\log\log n$ memory? What happens to the variance?

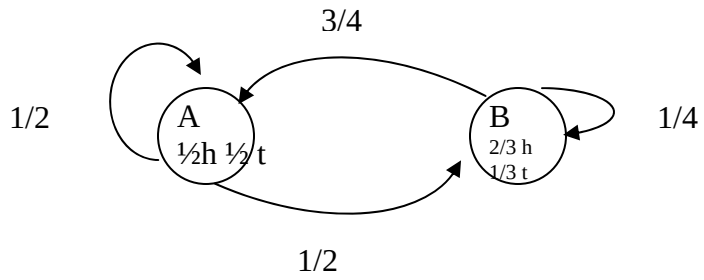
8. Play with the parameters in the algorithms for determining the number of distinct numbers. What if you change the value of t to 2t? What if you require that some a_i hash to 1 and another a_i hash to 2?

9. Program the algorithm for discovering a Hidden Markov Model. Create a small (3 state) Markov model and see if you algorithm can discover it from a long output sequence.

Homework 12 due April 24

10. In order for the HMM algorithm to discover the underlying model, the underlying model must have a strong signature. Play with your algorithm or an algorithm on lthe web to find some three state models with strong signatures.

11. An example of a HMM is the graph with two states A and B illustrated below.



The initial distribution is $\pi(A)=1$ and $\pi(B)=0$. At each step a change of state occurs followed by the output of heads or tails with probability determined by the new state. What is the probability of heads occurring after a sufficiently long sequence of transitions in the above example?