CS 530: Geometric and Probabilistic Methods in Computer Science Homework 2 (Fall '13)

1. You are given a deck of ordinary playing cards from which all cards except the jacks, queens, kings, and aces have been removed. Because I hate the color red, I also have removed the queens and kings of $\diamondsuit$ and $\heartsuit$. The discrete r.v. X has outcomes $\{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}$ and the discrete r.v. Y has outcomes $\{\text{jack}, \text{queen}, \text{king}, \text{ace}\}$.

   (a) Compute the marginal p.m.f.'s, p_X and p_Y, and the joint p.m.f., p_{XY}.

   (b) Are X and Y statistically independent?

   (c) Compute the conditional p.m.f.'s, $p_{X|Y}$ and $p_{Y|X}$.

   (d) Compute the entropies H_X, H_Y, and H_{XY}.

   (e) Compute the conditional entropies $H_{X|Y}$ and $H_{Y|X}$.

   (f) Compute the mutual information, I_{XY}.

   (g) I draw a card. How much information do you receive if you are told the card is black? red?

2. A fair coin is flipped until the first head appears. Let X denote the number of flips required. Find the entropy, H_X, in bits. The following expressions may be useful: $\sum_{n=1}^{\infty} r^n = r/(1-r)$ and $\sum_{n=1}^{\infty} nr^n = r/(1-r)^2$.

3. Two discrete random variables, X and Y, have outcomes, $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$, which occur with probabilities, $p_X(x_i) = \{1/2, 1/4, 1/4\}$ and $p_Y(y_j) = \{2/3, 1/6, 1/6\}$. Compute the maximum entropy joint distribution, P_{XY}, with these marginals. Prove your result is correct.

4. The p.m.f. for the 12367 most frequently used words in English is approximately:

$$p(n) = \begin{cases} \frac{0.1}{n} & \text{for } 1 \leq n \leq 12367 \\ 0 & n > 12367. \end{cases}$$

   This remarkable fact is known as Zipf's law, and applies to many languages (Zipf, 1949). If we assume that English is generated by picking

words at random according to this distribution, what is the entropy of English (per word)?

5. JPEG[1] is by far the most widely used compressed image format. However, unlike the GIF format, which uses a *lossless* compression method, JPEG compression decreases image quality, *i.e.*, it is a *lossy* method. In this exercise, JPEG will be viewed as an *information channel*. The grey values of the pixels of an image before and after JPEG compression will be considered to be samples of two non-independent discrete r.v.'s X and Y. Note that a pixel of an uncompressed image with 256 grey levels can contain at most 8 bits of information. The *lena* image on the class homepage is stored in a PGM format. This format does no compression. The *lena-jpeg* image is also stored in the PGM format. However, the *lena-jpeg* image has already undergone JPEG compression. Using pixels of the *lena* and *lena-jpeg* images, compute the following:

   - H_X - The entropy of the *lena* image.
   - H_Y - The entropy of the *lena-jpeg* image.
   - $H_{Y|X}$ - The channel noise.
   - $H_{X|Y}$ - The channel loss.
   - I_{XY} - The mutual information, *i.e.*, the amount of information which actually passes through the JPEG information channel.

6. Write a MATLAB function, *cdf*, which given an image with grey values, k, in the range 0 to 255, returns a vector of length 256 representing the discrete c.d.f., $F_K(k) = \sum_{i=0}^{k} f_K(k)$, where $f_K(k)$ is the image's histogram. The output of *cdf* should be normalized so that its minimum value is 0 and its maximum value is 255. Test your function on the Ganymede image. Plot your result.

7. Write a MATLAB function, *icdf*, which given an image with grey values in the range 0 to 255, returns a vector of length 256 representing the inverse function of the discrete c.d.f., F_K, defined as above. The output of *icdf* should be normalized so that its minimum value is 0 and its maximum value is 255. Test your function on the Callisto image. Plot your result.

[1] Joint Photograph Experts Group.

8. Using *cdf* and *icdf*, write a MATLAB function, *match*, which given two images, I_1 and I_2, returns a new image, I_3. The value of every pixel, (i, j), in I_3 is computed as follows:

$$I_3(i, j) = F_2^{-1}(F_1(I_1(i, j)))$$

where F_1 is the c.d.f. of I_1 and F_2^{-1} is the inverse c.d.f. of I_2. Compute I_3 where I_1 is Ganymede and I_2 is Callisto. Display I_3. Plot the histograms of I_2 and I_3.