

CS 530: Geometric and Probabilistic Methods in Computer Science Homework 2 (Fall '08)

1. The joint probability distribution of two discrete random variables X and Y is given below:

$$X \begin{matrix} & \begin{matrix} Y \\ 0 \\ 1/9 \\ 1/4 \end{matrix} \\ \begin{matrix} 0 \\ 1/12 \\ 1/36 \end{matrix} & \begin{bmatrix} 5/36 & 1/3 \\ 1/9 & 1/18 \\ 1/4 & 0 \end{bmatrix} \end{matrix}.$$

Compute:

- (a) Marginal information, H_X .
 - (b) Marginal information, H_Y .
 - (c) Conditional information, $H_{X|Y}$.
 - (d) Conditional information, $H_{Y|X}$.
 - (e) Joint information, H_{XY} .
 - (f) Mutual information, I_{XY} .
2. A roulette wheel is subdivided into 38 numbered compartments of various colors. The distribution of the compartments according to color is 2 green, 18 red, and 18 black. The experiment consists of throwing a small ball onto the rotating roulette wheel.
- How much information do you receive if you are told the color?
 - How much information do you receive if you are told the color and the number?
 - How much additional information do you receive if you are told the number, given that you already know the color?
3. Of a group of students, 25% are not suitable for university. However, 25% of these unsuitable students' applications are accepted. 50% of all students' applications are accepted.
- How much information is received by a student when he hears whether or not he has been accepted?

- How much information is received by a student, who knows he is not suitable for university, when he hears whether or not he has been accepted?
 - Answer the same question if applications are accepted or rejected by flipping a coin.
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. JPEG¹ 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.

¹Joint Photograph Experts Group.

6. Write a MATLAB function, *cdf*, which given a vector p representing a discrete p.m.f., returns a vector of the same length representing the discrete c.d.f., $P_K(k) = \sum_{i=0}^k p_K(k)$. Test your function on the normalized histogram of the *mars* image. Plot your result.
7. Write a MATLAB function, *icdf*, which given a vector P representing a discrete c.d.f. and an integer N , returns a vector of length N representing the inverse function of the discrete c.d.f. sampled at N equally spaced points. Test your function on the c.d.f. of the *mars* image with N equal to 1000. Plot your result.
8. The MATLAB function, *rand*, takes an integer argument, N and returns an $N \times N$ matrix of random numbers uniformly distributed in the interval, $[0, 1]$. Write a function, *rand2*, which, given a vector, p , representing a discrete p.m.f., and an integer, N , returns an $N \times N$ matrix of random numbers with the distribution, p . Test your function on the following p :

```
p = exp(-(0:255).^2/1000); p = p/sum(p);
```

Plot p and the histogram of *rand2*(p , 64).