CS 573: Algorithms, Fall 2009

Homework 6, due Wednesday, December 9, 23:59:59, 2009

Version 1.01

Name:	
Net ID:	Alias:

#	Score	Grader
1.		
2.		
3.		
4.		
5.		
Total		

Neatly print your name(s), NetID(s), and the alias(es) you used for Homework 0 in the boxes above. Staple this sheet to the top of your homework. If you are on campus, submit the homework by submitting it in SC 3306 (or sliding it under the door).

> What do you know? You know just what you perceive.
> What can you show? Nothin' of what you believe,
> And as you grow, each thread of life that you leave
> Will spin around your deeds and dictate your needs
> As you sell your soul and you sow your seeds,
> And you wound yourself and your loved one bleeds,
> And your habits grow, and your conscience feeds
> On all that you thought you should be –
> I never thought this could happen to Meeeeeeeee
> – Dreidel, Don McLean At other times you seemed to me either pitiable or contemptible,

Required Problems

1. Naive compression.
   [20 Points]

   We wish to compress a sequence of independent, identically distributed random variables $X_1, X_2, \ldots$. Each X_j takes on one of n values. The ith value occurs with probability p_i, where $p_1 \geq p_2 \geq \ldots \geq p_n$. The result is compressed as follows. Set

$$T_i = \sum_{j=1}^{i-1} p_j,$$

and let the ith codeword be the first $\lceil \lg(1/p_i) \rceil$ bits (in the binary representation) of T_i. Start with an empty string, and consider X_j in order. If X_j takes on the ith value, append the ith codeword to the end of the string.

(A) Show that no codeword is the prefix of any other codeword.

(B) Let Z be the average number of bits appended for each random variable X_j. Show that

$$\mathbb{H}(X_j) \leq z \leq \mathbb{H}(X_j) + 1.$$

2. Code this.
[20 Points]

Arithmetic coding is a standard compression method. In the case when the string to be compressed is a sequence of biased coin flips, it can be described as follows. Suppose that we have a sequence of bits $X = (X_1, X_2, \ldots, X_n)$, where each X_i is independently 0 with probability p and 1 with probability $1 - p$. The sequences can be ordered lexicographically, so for $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, we say that $x < y$ if $x_i = 0$ and $y_i = 1$ in the first coordinate i such that $x_i \neq y_i$. If $z(x)$ is the number of zeroes in the string x, then define $p(x) = p^{z(x)}(1-p)^{n-z(x)}$ and

$$q(x) = \sum_{y<x} p(y).$$

(A) Suppose we are given $X = (X_1, X_2, \ldots, X_n)$. Explain how to compute $q(X)$ in time $O(n)$ (assume that any reasonable operation on real numbers takes constant time).

(B) Argue that the intervals $\big[q(x), q(x) + p(x)\big)$ are disjoint subintervals of $[0, 1)$.

(C) Given (A) and (B), the sequence X can be represented by any point in the interval $I(X) = \big[q(X), q(X) + p(X)\big)$. Show that we can choose a codeword in $I(X)$ with $\lceil \lg(1/p(X)) \rceil + 1$ binary digits to represent X in such a way that no codeword is the prefix of any other codeword.

(D) Given a codeword chosen as in (C), explain how to decompress it to determine the corresponding sequence $(X_1, X_2, \ldots, X_n)$.

(E) (Extra credit.) Using the Chernoff inequality, argue that $\lg(1/p(X))$ is close to $n\mathbb{H}(p)$ with high probability. Thus, this approach yields an effective compression scheme.

3. Maximizing Entropy
[20 Points] Consider an n-sided die, where the ith face comes up with probability p_i. Show that the entropy of a die roll is maximized when each face comes up with equal probability $1/n$.

4. Extraction to the limit,
[20 Points] We have shown that we can extract, on average, at least $\lfloor \lg m \rfloor - 1$ independent, unbiased bits from a number chosen uniformly at random from $\{0, \ldots, m-1\}$. It follows that if we have k numbers chosen independently and uniformly at random from $\{0, \ldots, m-1\}$ then we can extract, on average, at least $k \lfloor \lg m \rfloor - k$ independent, unbiased bits from them. Give a better procedure that extracts, on average, at least $k \lfloor \lg m \rfloor - 1$ independent, unbiased bits from these numbers.

5. Easy inequality.
[20 Points]

Assume you have a (valid) prefix code with n codewords, where the ith codeword is made out of ℓ_i bits. Prove that

$$\sum_{i=1}^{n} \frac{1}{2^{l_i}} \leq 1.$$

1 Practice problems

1. Computing entropy.
[20 Points]

   (a) Let $S = \sum_{i=1}^{10} 1/i^2$. Consider a random variable X such that $\mathbf{Pr}[X = i] = 1/(Si^2)$, for $i = 1, \ldots, 10$. Compute $\mathbb{H}(X)$.

   (b) Let $S = \sum_{i=1}^{10} 1/i^3$. Consider a random variable X such that $\mathbf{Pr}[X = i] = 1/(Si^3)$, for $i = 1, \ldots, 10$. Compute $\mathbb{H}(X)$.

   (c) Let $S(\alpha) = \sum_{i=1}^{10} 1/i^\alpha$, for $\alpha > 1$. Consider a random variable X such that $\mathbf{Pr}[X = i] = 1/(S(\alpha)i^\alpha)$, for $i = 1, \ldots, 10$. Prove that $\mathbb{H}(X)$ is either increasing or decreasing as a function of α (you can assume that α is an integer).

2. Conditional Entropy
The ***conditional entropy*** $\mathbb{H}(Y|X)$ is defined by

$$\mathbb{H}(Y|X) = \sum_{x,y} \mathbf{Pr}[(X = x) \cap (Y = y)] \lg \frac{1}{\mathbf{Pr}[Y = y | X = x]}.$$

If $Z = (X, Y)$, prove that

$$\mathbb{H}(Z) = \mathbb{H}(X) + \mathbb{H}(Y|X).$$