The University of Sydney
MATH3024 Elementary Cryptography and Protocols

Semester 1 **Exercises and Solutions for Week 4** 2004

Information Theory

In this tutorial we consider the information theory of languages. In order to understand naturally occurring languages, we consider the models for finite languages $\mathcal{L}$ consisting of strings of fixed finite length N together with a probability function P which models the natural language. In what follows, for two strings X and Y we denote their concatenation by XY.

1. Consider the language of 1-character strings over $\{\texttt{A},\texttt{B},\texttt{C},\texttt{D}\}$ with associated probabilities 1/3, 1/12, 1/4, and 1/3. What is its corresponding entropy?

Solution The entropy of the language is

$$\frac{1}{3}\log_2(3)+\frac{1}{12}\log_2(12)+\frac{1}{4}\log_2(4)+\frac{1}{3}\log_2(3)$$

$$=\frac{2}{3}\log_2(3)+\frac{1}{12}(2+\log_2(3))+\frac{1}{2}=\frac{2}{3}+\frac{3}{4}\log_2(3),$$

which is approximately 1.855.

2. Consider the language $\mathcal{L}_2$ of all strings of length 2 in $\{\texttt{A},\texttt{B},\texttt{C},\texttt{D}\}$ defined by the probability function of Exercise 1 and 2-character independence: $P(XY)=P(X)P(Y)$. What is the entropy of this language?

Solution By rearranging the sum

$$\sum_{X\in\mathcal{L}}\sum_{Y\in\mathcal{L}}P(XY)\log_2(P(XY))$$
$$=\sum_{X\in\mathcal{L}}\sum_{Y\in\mathcal{L}}P(X)P(Y)\big(\log_2 P(X)+\log_2 P(Y)\big)$$

one finds the entropy to be double that of Exercise 1, or about 3.711. This is consistent with the interpretation of the entropy as the length of a random element of a language in some theoretically optimal encoding.

3. Let $\mathcal{M}$ be the strings of length 2 over $\{\texttt{A},\texttt{B},\texttt{C},\texttt{D}\}$ with the following frequency distribution:

$P(\texttt{AA})=5/36$	$P(\texttt{BA})=0$	$P(\texttt{CA})=1/12$	$P(\texttt{DA})=1/9$
$P(\texttt{AB})=1/36$	$P(\texttt{BB})=1/144$	$P(\texttt{CB})=1/48$	$P(\texttt{DB})=1/36$
$P(\texttt{AC})=7/72$	$P(\texttt{BC})=1/48$	$P(\texttt{CC})=1/16$	$P(\texttt{DC})=5/72$
$P(\texttt{AD})=5/72$	$P(\texttt{BD})=1/18$	$P(\texttt{CD})=1/12$	$P(\texttt{DD})=1/8$

Show that the 1-character frequencies in this language are the same as for the language in Exercise 2.

Solution The 1-character frequencies can be defined as the average of the character frequencies in the 1st and 2nd positions, but these turn out to be the same for each character, and agree witht the frequencies of Exercise 1.

4. Do you expect the entropy of the language of Exercise 3 to be greater or less than that of Exercise 2? What is the entropy of each language?

Solution The entropy of the language of Exercise 3 is approximately 3.633, compared to an entropy of about 3.711 for that of Exercise 2.

The language of Exercise 2 is the most random space with given 1 character frequences. The lower entropy in Exercise 3 could have been predicted since the probabilities agrees with the 1 character frequencies, while additional structure (less uncertainty) is built into the 2 character probabilities, since in general $P(\mathtt{XY}) \neq P(\mathtt{YX})$.

5. Consider the infinite language of all strings over the alphabet $\{\mathtt{A}\}$, with probability function defined such that $P(\mathtt{A}\dots\mathtt{A}) = 1/2^n$, where n is the length of the string $\mathtt{A}\dots\mathtt{A}$. Show that the entropy of this language is 2.

Solution One must verify the equality

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \log_2(2^n) = \sum_{n=1}^{\infty} \frac{n}{2^n} = 2.$$

We do this by first verifying the equality

$$\sum_{n=0}^{\infty} \frac{1}{2^n} + \sum_{n=1}^{\infty} \frac{n}{2^n} = 2\sum_{n=1}^{\infty} \frac{n}{2^n},$$

together with the standard identity $\sum_{n=0}^{\infty} 1/2^n = 2$.

Frequency Analysis

Consider those ciphertexts from the last tutorial which come from a Vigènere cipher. Use the javascript application for analyzing Vigenère ciphers:

```
http://magma.maths.usyd.edu.au/~kohel/
                        teaching/MATH3024/Javascript/vigenere.html
```

to determine the periods and keys for each of the ciphertext samples.