Classical Cryptography

Joachim von zur Gathen, Jérémie Detrey

3. Tutorial: Information entropy

(Hand in solutions on Monday, May 26th,
at the beginning of the tutorial)

Exercise 3.1 (Entropy and randomness).

We are given a biased coin whose probability for flipping heads is $p_{\mathrm{H}} = 30\%$.

(i) Compute the information entropy of such a coin toss.

(ii) What is the maximal entropy which can be expected of a coin toss?

(iii) How can you transform this biased coin into a fair(er) coin?

(iv) What would be the value of the entropy, then?

Exercise 3.2 (Entropy and Huffman trees).

We are given an alphabet $\mathbb{A} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$ with the following frequency distribution:

Letter	A	B	C	D	E	F
Frequency	5	18	10	15	45	7

(i) Compute the corresponding entropy.

(ii) Using the same number of bits to encode each letter, how many do we need?

(iii) What is the expected length of an n-letter message with this encoding?

(iv) How can you do better?

Exercise 3.3 (Entropy of the English language). (13 points)

The following table gives the frequency distribution of the letters in English.

Letter	A	B	C	D	E	F	G	H	I
Frequency	8.04	1.54	3.06	3.99	12.51	2.30	1.96	5.49	7.26

Letter	J	K	L	M	N	O	P	Q	R
Frequency	0.16	0.67	4.14	2.53	7.09	7.60	2.00	0.11	6.12

Letter	S	T	U	V	W	X	Y	Z
Frequency	6.54	9.25	2.71	0.99	1.92	0.19	1.73	0.09

$\boxed{1}$ (i) What is the entropy of English?

$\boxed{1}$ (ii) What is the maximal entropy for a 26-letter alphabet?

$\boxed{1}$ (iii) Compute the *redundancy* of English, *i.e.* the entropy distance between English and a uniformly-distributed 26-letter language.

$\boxed{4}$ (iv) Give a Huffman encoding of English according to this frequency distribution.

$\boxed{2}$ (v) Have fun with the Java applet available at this URL:
`http://math.ucsd.edu/~crypto/java/ENTROPY/`
What entropy do you obtain?

$\boxed{2}$ (vi) Why is it lower than the previously computed entropy?

$\boxed{2}$ (vii) Does it mean that we can compress an English text in approximately 1.2 bits per letter? Why aren't such extreme compression techniques not used in practice?

Exercise 3.4 (Entropy of amino acids). (10 points)

In order to build proteins, our living cells use ribosomes to translate strands of *messenger ribonucleic acid* (mRNA) into sequences of amino acids. The mRNA encodes amino acids using *tri-nucleotide codons*, *i.e.* sequences of three nucleotides among Adenine (A), Cytosine (C), Guanine (G) and Uracil (U).

So here, instead of encoding an alphabet of 26 letters (A, ..., Z) using bits (0 or 1), proteins are seen as words over an alphabet of 20 amino acids (Alanine, Cysteine, ...) encoded using nucleotides (A, C, G, or U).

You will find in the following table the actual codon encodings of each amino acid, along with their occurrence in proteins[1]:

Amino acid	Ala	Cys	Asp	Glu	Phe
Codon(s)	GCU, GCC GCA, GCG	UGU, UGC	GAU, GAC	GAA, GAG	UUU, UUC
Frequency	7.8	1.9	5.3	6.3	3.9

Amino acid	Gly	His	Ile	Lys	Leu
Codon(s)	GGU, GGC GGA, GGG	CAU, CAC	AUU, AUC AUA	AAA, AAG	UUA, UUG CUU, CUC CUA, CUG
Frequency	7.2	2.3	5.3	5.9	9.1

Amino acid	Met	Asn	Pro	Gln	Arg
Codon(s)	AUG	AAU, AAC	CCU, CCC CCA, CCG	CAA, CAG	CGU, CGC CGA, CGG AGA, AGG
Frequency	2.3	4.3	5.2	4.2	5.1

Amino acid	Ser	Thr	Val	Trp	Tyr
Codon(s)	UCU, UCC UCA, UCG AGU, AGC	ACU, ACC ACA, ACG	GUU, GUC GUA, GUG	UGG	UAU, UAC
Frequency	6.8	5.9	6.6	1.4	3.2

(i) Compute the entropy of amino acids in proteins in terms of nucleotides. $\boxed{4}$
Warning! The definition of Shannon's entropy as seen in the lecture measures the entropy in terms of bits (*i.e.* encoding the values with only 0's and 1's). But here, we want to measure the entropy in terms of nucleotides (*i.e.* encoding the values with A's, C's, G's and U's). Think carefully and modify the entropy formula accordingly!

(ii) Give a Huffman encoding for amino acids, still using nucleotides. $\boxed{4}$
Warning! Once again, you can encode 4 different values with a single nucleotide. This should give you a hint as to the arity of the Huffman tree.

(iii) Look more closely at the redundancy of the codons. Discuss. $\boxed{2}$

[1] Source: `http://en.wikipedia.org/wiki/List_of_standard_amino_acids`