Ex. 1
Verify that the function $H(p_1, \ldots, p_n) = -\sum_k p_k \log_2 p_k$ satisfies all 8 axioms on H.

Ex. 2
(Not to be handed in). List as many of the 8 axioms as you can, (without just looking at the notes).

Ex. 3
Let X be a random variable, taking the values a_1 and a_2 with probability p_1 and p_2, respectively. Let Y be a random variable, taking the values b_1, b_2 and b_3 with probability q_1, q_2 and q_3, respectively.
Prove that $H(X,Y) \leq H(X) + H(Y)$, with equality if and only if X and Y are independent. (You should work through the proof here, not just say that this is a special case of a more general theorem stated in the lecture).

Ex. 4
Two fair dice are thrown. X denotes the value obtained by the first, and Y the value obtained by the second, and let $Z = X + Y$ be the corresponding random variable.
Evaluate $H(X), H(Y), H(X,Y), H(Z), H(X|Y)$.
From this, verify that $H(X,Y) = H(X) + H(Y)$ and show that $H(Z) < H(X,Y)$.

Ex. 5
Show that, for any random variable $H(X, X^2) = H(X)$.
Show that $H(X^2|X) = 0$ but that $H(X|X^2)$ is not necessarily zero.

Ex. 6
The random variable X takes the values $1, 2, \ldots, 2N$ with equal probability. The random variable Y is defined by

$$Y = \begin{cases} 0 & \text{if } X \text{ is even,} \\ 1 & \text{if } X \text{ is odd.} \end{cases}$$

Evaluate $H(X|Y)$, and show that $H(Y|X) = 0$.

Ex. 7
(Think about this one for a while). Suppose you want to compress, encode and encrypt a message. (Compression for shortening the data if possible, encoding for securing against noise on the channel, encryption to keep the message secret). Does it matter in which order you do this, and if yes, in which order should you do it?

Ex. 8
Make yourself familiar with the restricted bookshelf, there are copies of the recommended books, (Welsh, MCKay, Jones and Jones, Hill). (The restricted bookshelf is behind the counter on the left hand side, then inside at the very end close to the copy machine, 001.5*** .. library coding).

Ex. 9
If X and Y are discrete random variables taking only a finite number of values, show that

$$H(X+Y \mid X) = H(Y \mid X).$$

Also show that

$$H(g(X,Y) \mid X) = H(Y \mid X)$$

does not hold generally for $g : \mathbb{R}^2 \to \mathbb{R}$.

Ex. 10
A statistical survey of married couples shows that 70% of men have dark hair, that 25% of girls are blonde, and that 80% of blonde girls marry dark-haired men. How much information about the colour of a man's hair is conveyed by the colour of his wife's hair?

Ex. 11
A random variable X has the binomial distribution with parameters n and p. That is, for $0 \leq k \leq n$

$$P(X = k) = \binom{n}{k} p^k q^{n-k},$$

where $0 < p < 1$ and $q = 1 - p$. Show that

$$H(X) \leq -n(p \log p + q \log q).$$

Remark:
At the end of the year you will have to write a Thesis. Experience from last years shows that students learn the mathematical word processing language LaTeX just in the final steps of their Thesis.
You are encouraged to learn this during the year. good exercise would be to type some of your weekly homework, and to learn LaTeX in small chunks. Essentially one can learn LaTeX by looking at a brief manual, looking just at some examples. LaTeX is installed on the computers of the Maths department. It's free software. So you can also dowload the package from the web and use on your computer.
A very brief guide:
ftp://ftp.ams.org/pub/tex/doc/amsmath/short-math-guide.pdf
http://www.isg.rhul.ac.uk/∼cjm/Chris_Mitchell.htm#LaTeX
http://www.ma.rhul.ac.uk/current/latex-howto/

Ex. 12
Show that, for any positive integer m there exists an instantaneous code over $\Sigma = \{0, 1\}$ that has words of all lengths in the set $\{1, \ldots, m\}$.

Ex. 13
What is the maximum number of words in a binary instantaneous code in which the maximum word length is 7?

Ex. 14 (Not to be handed in.)
Work through the detailed example in section 1.6.

Ex. 15 (Not to be handed in.)
Welsh's argument on McMillan's inequality is a bit short. Read it carefully and try to understand the details. Below is the argument (almost a word by word quotation).

Suppose we have a uniquely decipherable code C with word lengths $l_1, \ldots, l_N$. If $l = \max_i l_i$, then, for any positive integer r, we have

$$(D^{-l_1} + \ldots + D^{-l_N})^r = \sum_{i=1}^{rl} b_i D^{-i},$$

where b_i is a non-negative integer.
Here b_i counts the number of ways in which a string of length i of symbols from the alphabet Σ can be made up by stringing together r words of lengths chosen from the set $\{l_1, \ldots, l_N\}$. (This is a partition problem, also studied in the theory of generating functions. Construct some small examples to see what happens here!)
But if the code C is uniquely decipherable, it must be the case that any string of length i formed from codewords must correspond to at most one sequence of code words. Hence we must have

$$b_i \leq D^i, \quad (1 \leq i \leq rl).$$

Hence we obtain

$$(D^{-l_1} + \ldots + D^{-l_N})^r \leq rl.$$

Therefore

$$\sum_{k=1}^{N} D^{-l_k} \leq l^{\frac{1}{r}} r^{\frac{1}{r}},$$

for all positive integers r. We can let $r \to \infty$, so that $l^{\frac{1}{r}} r^{\frac{1}{r}} \to 1$, and McMillan's inequality follows.

Ex. 16
Use Huffman's encoding for the source given in example 2.4.2 ($p_1 = 0.6, p_2 = 0.13, p_3 = 0.12, p_4 = 0.1, p_5 = 0.05$). Compare the average word length with the Shannon-Fano encoding.
Also calculate the entropy $H = -\sum p_i \log_2 p_i$ and compare.

Ex. 17
Use Shannon-Fano encoding for the source given in example 2.5.4 ($p_1 = 0.5, p_2 = 0.2, p_3 = 0.15, p_4 = 0.1, p_5 = 0.05$). Compare the average word length with the Hufmann coding. Also calculate the entropy H and compare.

Ex. 18
Find a simple necessary condition so that in Kraft's and McMillan's inequality equality $= 1$ can hold. In the noiseless coding therem, when can the lower bound be attained?

Ex. 19
Compare the noiseles coding theorem with the length of a compact encoding of $2^k - 1$ binary words with equal probability $p = \frac{1}{2^k - 1}$.

Ex. 20
You are allowed six questions that will truthfully be answered by Yes or NO. Describe briefly a strategy how one can determine one square of a chessboard (64 squares). How many questions does one need to specify one square on and $n \times n$ board?

Ex. 21
Examine whether the following three codes are uniquley decipherable, prefix codes, and/or instantaneous codes.

$$C_1 = \{0, 010, 01, 10\}, C_2 = \{10, 00, 11, 110\}, C_3 = \{0, 10, 110, 111\}.$$

(Prove your statements).

Ex. 22
You are given a balance and nine apparently identical coins. One coin is different from the rest. Devise a strategy of three weigings to find the coin and whether it is heavier or lighter. Try to generalise.

Ex. 23
Find the compact code over $\{0, 1\}$ for a source that emits words $w_1, \ldots w_6$ with

$$P(w_1) = \frac{1}{3}, P(w_2) = \frac{1}{4}, P(w_3) = \frac{1}{6}, P(w_4) = P(w_5) = P(w_6) = \frac{1}{12}$$

and compare its average length with the upper and lower bounds given by the noiseless coding theorem.

Ex. 24 (easy, with simple result from probability theory.)
A message consisting of N binary digits is transmitted through a binary symmetric channel having error probability p. Show that the expected number of errors is Np.

Ex. 25
A code consists of 4 codewords $c_1 = 1000, c_2 = 0110, c_3 = 0001, c_4 = 1111$. Assume that the probabilities that these words occur are

$$P(c_1) = P(c_2) = \frac{1}{3}, P(c_3) = P(c_4) = \frac{1}{6}.$$

You use a binary symmetric channel with error probability $p = \frac{1}{10}$. You receive 1001. How should you decode (using the maximum likelihood method)? What is the error probability?

Use the time to revise old handouts, home work, Welsh's book, or to experiment with LaTeX.

Ex. 26 (not to be handed in)
Calculate the capacity of the binary symmetric channel with error probability ε. (Go through the details of the proof given in the lecture, or read about it in Welsh's book).

Ex. 27
Calculate the capacity of the binary erasure channel with error probability ε.

Ex. 28
Plot the real function $C : [0,1] \to \mathbb{R}$ with $C(p) = 1 + p\log_2 p + (1-p)\log_2(1-p)$. Evaluate it for $p = 0, p = 0.01, p = 0.02, p = 0.05, p = 0.2, p = 0.5, p = 0.99, p = 1$ and explain what it has to do with channels.

Ex. 29 (not so relevant for the course)
For a code with code word length n

- Given any codeword $\vec{c}$. Count how many vectors $\vec{x} \in V_n = \{0,1\}^n$ have Hamming distance $d(\vec{x},\vec{c}) = t$. Do the same for $d(\vec{x},\vec{c}) \leq t$.
- Using maximum likelihood decoding, which is the same as nearest neighbour decoding, i.e. received any word, decode as the nearest codeword. At most how many codewords can there be, if any two codewords have distance at least 3? (generally, if the minimum distance is d, where d is odd)?

  Let $n = 7, d = 3$. How many codewords can there be at most?
- Use a computer. Let n be an integer. Generate M random binary codewords of length n.

Ex. 30
A source emits words with probabilities

$$p_1 = \frac{1}{3}, p_2 = \frac{1}{3}, p_3 = \frac{1}{4}, p_4 = \frac{1}{12}.$$

Study the possible code word lengths in the Huffman code(s) and compare with the word lengths of Shannon encoding.

Ex. 31
A binary symmetric channel with symbol error probability $p = 0.05$ can transmit 800 binary digits per second. How many bits can it transmit accurately per second? Hint: use the noisy coding theorem.

Ex. 32
A binary symmetric channel which can can physically transmit 800 binary digits per second, can transmit 500 digits per second with arbitrarily small error. What does this tell about the error probability of the channel? Hint: use the noisy coding theorem.

Ex. 33
State in your own words the noisy coding theorem and write a short essay about its meaning (200-300 words).

Ex. 34
Give an example to show that equality can hold in Fano's inequality.

Ex. 35
Two binary vectors are chosen at random. What is the probability that their Hamming distance is at least k?

Ex. 36 (For programming)
Let V_n denote the set of binary words of length n. Choose k codewords $w_1, \ldots, w_k$ at random. Let D_k denote their minimum distance, i.e. the minimum distance between any pairs of two distinct codewords. Estimate the expected value by a computer simulation.
(Example: for $k = 3, 4$, and $1 \leq n \leq 20$ do a few thousands random tests.)

Ex. 37
Given a channel with matrix

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

Show that the capacity is given by $C = \log 2^{\frac{5}{3}} - \log 3$.

Ex. 38
A memoryless source of entropy 15 bits per sourceword is connected to a binary symmetric channel of symbol error probability $p = 0.1$, which can pass 1000 binary digits per second through the receiver. At which rate can the source words be emitted if they are to be communicated accurately through the channel?

Ex. 39
A source S is such that, for each n, the probability distribution of X_n depends on X_{n-1} in the following way:

$$\begin{aligned} P(X_n \mid X_{n-1}, X_{n-2}, , \ldots, X_1) &= P(X_n \mid X_{n-1}) \\ P(X_n = 0 \mid X_{n-1} = 0) &= P(X_n = 1 \mid X_{n-1} = 1) = p \\ P(X_n = 1 \mid X_{n-1} = 0) &= P(X_n = 0 \mid X_{n-1} = 1) = q, \end{aligned}$$

where $0 \leq p \leq 1$ and $q = 1 - p$. Show that $H(S)$ exists and find it.

Ex. 40
A source behaves as follows: it emits, each with probability $\frac{1}{2}$, either an infinite string of zeros or a purely random string of zeros and ones. Does S have an entropy?

Ex. 41
Consider a source S whose output $(X_n : n = 1, 2, \ldots)$ is as follows: $X_{2k} = 1$ for all k and $(X_1, X_3, X_5, \ldots)$ is a purely random source. Show that $H(S)$ exists, but that $\lim_{n\to\infty} H(X_n \mid X_1, \ldots, X_{n-1})$ does not.

Ex. 42
Consider the sources of the three problems above. Are they stationary or not?

Ex. 43
A memoryless source emits only vowels, each with the following probabilities:

$$P(A) = 0.2, P(E) = 0.3, P(I) = P(O) = 0.2, P(U) = 0.1.$$

Estimate the number of typical outputs of length n. (Describe what you are doing: define what you mean by typical.)

Ex. 44
A memoryless source over the 26-letter alphabet has a vocabulary of about 10^n sequences of length n, for suifficiently large n. Estimate the entropy of the source. (Hint: The answer is simple and short.)

Ex. 45 (nice problem but not exam-relevant)
Consider the infinite square lattice consisting of all integer-coordinated points of the plane and with nearest neighbours in the direction of the coordinate axes joined by an edge. A self avoiding walk of length n is a sequence of n edges starting from the origin, each pair of consecutive edges having a common point, and at no stage revisiting a point already visited. If $f(n)$ denotes the number of self-avoiding walks of length n, then $f(1) = 4, f(2) = 12$ and so on. Prove that

$$f(m + n) \leq f(m)f(n),$$

and hence deduce that

$$\lim_{n\to\infty} (f(n))^{\frac{1}{n}} = \inf_{n\geq 1} (f(n))^{\frac{1}{n}} = \theta$$

exists. Determine $f(3)$. Draw the situation for $n = 1, 2, 3$. Prove that $2 \leq \theta \leq 3$.

Ex. 46
With probability $\frac{1}{3}$, a source $\mathcal{S}$ emits a random string of zeros and ones; with probability $\frac{2}{3}$, it emits a random string of ones and twos. Show that the source is not ergodic.

Ex. 47 (Study the examples from the lectures:)
Find the entropy of the Markov source whose transition matrix is given by

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The same for

$$\begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

Ex. 48
Which of the Markov sources having transition matrices as shown are irreducible?

$$M_1 = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$M_3 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}, \quad M_4 = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

This is a relatively easy problem sheet, reflecting that the chapter is rather on observations and empirical studies. Use also the time to catch up with previous problem sheets. Good further problems for exam preparation purposes can be found on the course's webpage and at Prof. Cover's webpage. http://www.stanford.edu/class/ee376a/
Ignore problems on material that is not covered in this course...)

For exam preparation:
1. Revise the problem sheets.
2. Revise the sketched solutions.
3. Iterate the above, modify the questions by taking different numerical values.
4. Go through the lecture notes and/or Welsh's book.
5. Last years exam paper is available online (library, you need to login with your account).

Ex. 49
What is the average length of a word in the first order approximation to English if the probability of a space is $p = 0.18$?

Ex. 50
Taking the entropy of English as 1.5 bits, estimate the number of meaningful strings of N symbols in English.

Ex. 51
If you assume $H_E = 1.2$ bits show, assuming the noiseless coding theorem, that 100 letters of ordinary text can be encoded in ~ 25.2 characters of recorded text without loss of information.

Ex. 52
Prove that the average word length in the first-order and second-order approximation to English are the same.

Ex. 53
Two languages obey Zipf's law exactly, but the first has twice as many words as the second. Show that, if q_1 and p_1 are the probabilities of the most frequent words in the two languages, then

$$p_1 \approx \frac{q_1}{1 - q_1 \ln 2}.$$

Good further problems for exam preparation purposes can be found on the course's webpage and at Prof. Cover's webpage. http://www.stanford.edu/class/ee376a/ Ignore problems on material that is not covered in this course...

Ex. 54
Read the description of the Lempel Ziv algorithm in McKay's book. Describe, in your own words the method, and briefly discuss its performance, advantages/disadvantages. How good would it be for large examples? Can the basic method be improved?

Ex. 55
Encode the string using the Lempel Ziv method.

$$000000000000100000000000.$$

Ex. 56
Decode the string

$$001010111011001001000110101010000011,$$

which was encoded by the basic Lempel Ziv encoding in McKay's notation.