

Date Assigned: 20 October 2004.

Date Due: 27 October 2004 in class.

Suggested Reading: Section 8.3 on properties of the optimal input distribution and Section 8.11 on Hamming codes. Chapter 9 and Sections 2.3 and 2.4 of your text book.

1. RAID is an acronym for Redundant Arrays of Independent Disks and is an information storage system consisting of up to ten disk drives. The simplest RAID system involves 7 disk drives which store data at rate 4/7 using a (7,4) Hamming code: each successive four information bits are encoded with the code and the resulting seven codeword bits are written one to each disk. Suppose that either a disk is fully correct or it is entirely dead. Further, assume that we know when a disk is dead.
 - (a) Model this data storage system as communicating the information bits over an appropriate noisy channel.
 - (b) Argue that if no more than two disk drives are lost, then the others can recover the information.
 - (c) Give an example of three disk drives that, if lost, lead to a failure to recover the information from the others. On the other hand, give an example of three disk drives that, if lost, one can indeed recover the information from the remaining.

RAID is a commercially very successful data storage system. It was proposed by UC Berkeley researchers (the same team that also proposed the RISC processor design).

2. Most books have an ISBN (International Standard Book Number), a sequence of 10 digits. For instance, your text book has the ISBN 0-471-06259-6. The first digit indicates the language, the next two digits indicate the publisher, the next six digits the book number assigned by the publisher and the last digit is a “check sum” that is defined as:

$$\text{digit}_{10} = \sum_{i=1}^9 i \cdot \text{digit}_i \quad \text{modulo } 11. \quad (1)$$

In arithmetic modulo 11, one needs a symbol for 10 – this symbol is X .¹ You might want to verify that the last digit of the ISBN of your text book (6) satisfies this equation.

- (a) Verify that the ISBN can be used to detect single errors.

¹Not very many books seem to have X as its last digit (the book *Error Control Coding: Fundamentals and Applications*, by Lin and Costello, has the ISBN 0-13-283796- X). If the first 9 digits were i.i.d. uniformly chosen between 0 and 9, what do you think the probability of the last digit of the corresponding ISBN being X is?

- (b) Verify that the ISBN can be used to detect transpositions of two symbols, a common mistake made in transcribing numbers.
3. Problem 9.1 of your text book.
 4. Problem 9.4(a) and 9.4(b) of your text book.
 5. Consider a DMC $p(y|x)$. Prove that no output y is unused by an optimal input distribution unless it is unreachable, that is, has $p(y|x) = 0$ for every x in the input alphabet.
 6. The input distribution that maximizes $I(X;Y)$ through a DMC may not be unique. However, show that all the optimal input distributions induce the *same* output probability distribution.
 7. *Bonus:*² Let m be the smallest number of inputs that can be used with nonzero probability to achieve the capacity of a DMC and let $A \subset \mathcal{X}$ be such a set of input alphabets. Then show that $m \leq |\mathcal{Y}|$ and the input probability assignment on A to achieve capacity using only inputs in A is *unique*.

²Again, this matters to only those students who expect more challenges from this course.