Course Outline

Data Compression COMP 423

(Winter 2008 MWF 12:35:-1:25 LEA 210)

Instructor:	Professor Michael Langer
Office:	McConnell Engineering 329
Tel:	398-3740
Email:	langer@cim.mcgill.ca
Course Web Page:	www.cim.mcgill.ca/~langer/423.html
Office Hours:	by appointment

Teaching Assistant (T.A.)	Robert Kaplow
Email:	rkaplo@cs.mcgill.ca
Office:	ENGMC 111
Hours:	by appointment

Introduction

Data Compression is the computational problem of how to encode a data file (text, image, audio, video) so that the new file has fewer bits the original file. Compression is possible when the original file contains redundancies, such as when certain symbols or sequences of symbols in the file occur more often than others. An example is the redundancy of English text: the letter "e" occurs more frequently than the symbol "z"; the letter "q" is almost always followed by the letter "u", etc.

The basic technique of data compression is to re-encode a data file using shorter codewords (binary strings) for symbols or sequences of symbols that occur more frequently and using long codewords to encode symbols that occur less frequently. This principle is applicable to many types of data including text, still images, video, audio, and speech.

While many theories of data compression apply to all types of data, each type of data contains particular redundancies. Understanding these redundancies on a case-by-case basis is useful for choosing or designing a data compression method that is appropriate for each type of data. As such, this course also covers particular types of data and the specialized methods that have been developed for compressing and decompressing these data.

Data compression is used both for storage and transmission of data. Any storage can become filled with data, and so one can use the storage more efficiently by compressing data prior to storing it. Compression is also important for data transmission, such as downloading files from the world wide web. Since the time of transmission is proportional to the size of the file, one can reduce time by downloading a compressed file rather than a raw data file. Most computer users are familiar with these issues. The specific purpose of the course is to teach advanced undergraduate students the theory and tools of data compression.

The ideas and tools we will cover have applications to other fields, including complexity theory, information theory, statistics, communication, and information retrieval. Thus although we focus on the specific problem of compression, the concepts are relevant to these and other fields.

Prerequisites

The prerequisites for the course are:

- COMP 251 Data Structures and Algorithms
- MATH 323 Probability Theory
- MATH 223 Linear Algebra

These prerequisites are taken seriously. Students who have not passed these courses with a grade of C or better are not allowed to take the course. Moreover, it is strongly recommended that students enrol in this course only if they have achieved a grade of at least B- in these courses.

Topics

- Part 1: Lossless Compression (8 weeks)
  - definitions: codes, average code length, entropy
  - optimal prefix codes, Huffman coding, Shannon code
  - Kraft-McMillan inequality, bounds on average code length, Jensen's inequality
  - codes for positive integers: run length codes (fax), Golumb codes
  - inverted files
  - move-to-front algorithm: best vs. worst case analysis
  - Lempel-Ziv algorithms: worst case analysis
  - marginal, conditional, cumulative probabilities (review)
  - Markov chains and arithmetic coding
- Part 2: Signal Compression (5 weeks)
  - differential coding, transform coding
  - lossless image coding
  - lossless video coding (MPEG)
  - quantization
  - lossy differential coding, linear predictive coding, autocorrelation
  - discrete cosine transform (JPEG)
  - audio (speech, MP3)

Evaluation

Your final grade will be calculated using the following percentage breakdown.

- **Three Assignments (40 %)**
  - A1 to be posted in mid-late January
  - A2 to be posted in mid February
  - A3 to be posted in mid March

  You will be given about two weeks to complete each assignment.
- **Two in-class Quizzes (30 %)**
  - Quiz 1 in early February
  - Quiz 2 in early March *i.e.* after Study Break
- **Final Exam (30 %)** during Examination Period *i.e.* after classes end.

McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures. See **www.mcgill.ca/integrity** *for more information, as well as the Student Guide to Avoid Plagiarism,* **www.mcgill.ca/integrity/studentguide/**.

Lecture Notes

All material covered in the lectures will be made available online on the course web page.

Reference Textbooks

There is no textbook for the course. If you wish to do further background reading, then I would recommend the following which are available on *two hour reserve in the Schulich Library.* Other editions of these books will work fine.

- "Introduction to data compression" by Khalid Sayood.
  San Francisco : Morgan Kaufmann Publishers, 2000.
  Schulich Science & Engineering: TK5102.92 S39 2000
- "Data compression : the complete reference" by David Salomon
  New York : Springer, c2000.
  Schulich Science & Engineering: QA76.9 D33 S25 2000