A Singular Value Decomposition Analysis of Grade Distributions

Stephanie Chung and Caroline Seabrook

School of Mathematics

Georgia Institute of Technology

VIGRE REU

July 2004

Abstract

Students use grade distributions from previous semesters to help them in selecting future classes to take. However, there is more information to be told by these distributions than a simple average GPA or percentage of failures. Using Singular Value Decomposition (SVD), we analyzed this data to discover underlying patterns. We also discovered patterns using probability distributions and entropy. For background, we will discuss error reduction over noisy channels. We also explain Shannon information content and some of its applications, particularly in data compression. We then discuss the SVD and its applications in data compression and noise suppression. For

our research, we used the grade distributions from the math department of Georgia Tech. First we constructed a matrix of the grade distributions from each section of undergraduate classes with more than 10 students. This data was obtained from classes held over the past five years, when Georgia Tech started the semester system. We computed the SVD of this matrix, and then we broke the grades down by class and professor and computed those SVD's. We used truncated SVD's to reconstruct our data and found that the singular values acted differently for each grade. We also computed frequency graphs for our data and observed that A's B's and C's have similar distributions and D's and F's have similar distributions. We were also able to use these frequency graphs to compute the entropy of our data and come to conclusions about a particular professor or class based on its overall entropy.

1 Introduction

One of the main considerations of graduate admissions officers and employers of college graduates is a job candidate's academic transcript, in which, the student's performance in each class is reflected by a single letter. What information is really stored in these letter grades? It is fairly safe to say that grades are a matter of concern to most college students at any university or institution, including Georgia Tech, and any additional information about grades would be valuable.

As the students of Georgia Tech register for classes each semester, many students look up grade distributions with the online Course Critique [1] in order to figure out which classes are easiest or which professors give the most

A's. What other information can be extracted from this data? What else can aid students in their selection of classes? In our project, we attempt to answer these questions.

To begin, we need some background knowledge from information theory, which was developed to study the theoretical limitations and potentials of communication systems given noisy channels [2]. A natural measure of information content of an outcome x is measured in bits by the Shannon information content, which is defined by

$$h(x) = \log_2 \frac{1}{P(x)}, \tag{1}$$

where $P(x)$ is the probability of outcome x. For an ensemble of outcomes, X, the entropy is the average Shannon information content. Entropy, also measured in bits, is defined by

$$H(X) \equiv \sum_{x \in X} P(x) \log_2 \frac{1}{P(x)}. \tag{2}$$

It vanishes when one outcome contains all the probability, that is, if $P(x_i) = 1$ for some x_i, then $H(X) = 1 \cdot \log_2(1) = 0$. Entropy is maximized when the ensemble is uniformly distributed such that $P(x_i) = 1/|X|$ for all x_i.

We also utilize singular value decompositions (SVDs), which are obtained by decomposing an $m \times n$ matrix A into

$$A = U \Sigma V^T, \tag{3}$$

where U and V are orthogonal matrices, and Σ is an $m \times n$ matrix with "singular values" along the main diagonal entries and zero everywhere else. The built-in Matlab SVD function easily computes the vector of singular

values, s, which is subsequently normalized by dividing it by the square of its length. This gives how much data lies in the direction of each singular vector. We subsequently interpret these singular values in the context of the problem we study.

The data used in this research project comes from the Georgia Tech Critique Database, downloadable from the critique website [3]. The database gives the percentage of students who received grades A, B, C, D, and F for each class. These percentages also include students who took the class pass/fail or who withdrew from the class before drop day, so we adjusted the data to obtain percentages only taking students with letter grades into account. We then organized the raw data and grouped it by semester, class, and teacher to be imported into Matlab as matrices where each row corresponds to a class section. The first column represents the percent of A's, the second column represents the percent of B's, and so forth, so that each matrix has five columns. We used this source for convenience despite the fact that the data are rounded to the nearest whole percent. We also obtained the data to two decimal places and found that the effects of rounding are not significant, as indicated by Table 1.

The full database lists all classes from Spring 2000 to present, although the data also is available in other formats for Fall 1999, when the Institute changed to the semester system. Any data from quarters before Fall 1999 are difficult to compare to data from semesters because the material in Georgia Tech's classes were divided differently. For example, a year of calculus is divided into two classes under the semester system, but it is three classes under the quarter system.

Effects of Rounding Data

2 decimals	Rounded	Difference
0.8003	0.8008	0.0005
0.0914	0.0911	-0.0003
0.0632	0.0629	-0.0003
0.0318	0.0319	0.0001
0.0133	0.0133	0.0000

Table 1: Normalized singular values rounded to 2 decimals, rounded to the nearest whole percent, and the differences between these two quantities using data from all math classes in fall semester 1999. The very small differences indicate that using rounded data does not greatly affect results.

Not all professors or classes were analyzed because those professors or classes with only a small number of sections created matrices with too few rows to be analyzed reasonably using SVDs. Larger matrices can be analyzed more reliably; more data reduces the standard error and produces more accurate results. Also, we chose not to include sections containing fewer than ten students or graduate classes because small classes may skew results, and they tend to enter many zeros into the matrix. A class with only two students, for example, may have a grade distributions with those two students both receiving A's (100% A's) and zero percent B's, C's, D's, and F's.

In this paper, we examine various approaches to analyzing and interpreting this data. In section 2 we examine the frequency distributions of each letter grade, the shapes of the distributions. In section 3, we discuss singular value decomposition and its applications. In section 4, we perform

singular value decomposition analysis on the matrices and on matrices of the grades relative to average distribution, and we subsequently reconstruct the matrices with truncated singular value decompositions to see how much information each singular value encodes. In section 5, we calculate the entropy of the data sets. We summarize our results in a concluding section and discuss background research on information theory, entropy, inference, and data compression in an appendix.

2 Frequency Distributions of Letter Grades

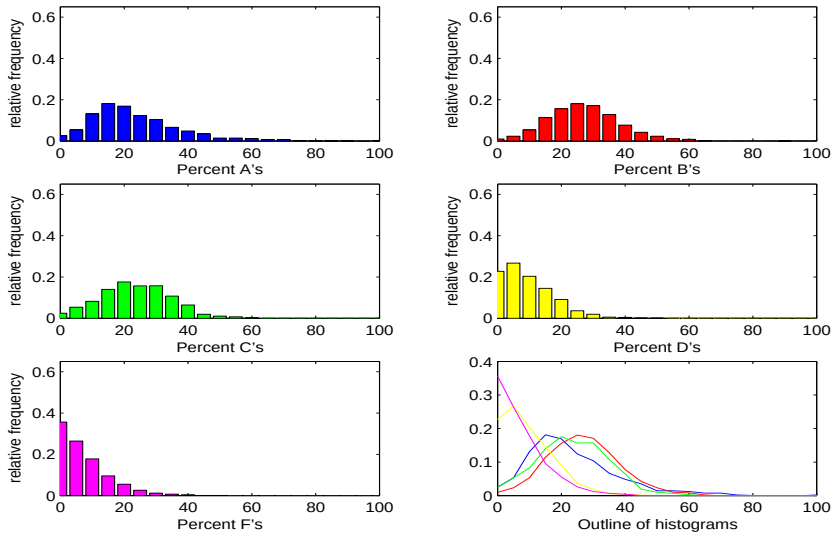
In a typical college course, one might have certain expectations regarding the grade distributions– such as greater percentages of A's, B's, and C's, and fewer D's and F's. Examining means of different sets of the Course Critique data, this is confirmed; grades of A, B, and C make up a greater percentage of the grade distribution than do grades of D and F. Table 2 shows the mean percentage of each letter grade given in the set of all undergraduate classes with at least ten students, all sections of Math1501 (Calculus I), and all classes taught by math professors #1, #22, and #27 (names withheld). We observe similar results with almost all data sets. There are some exceptions such as math professor #27, half of whose students, on average, received D's and F's. Otherwise, it is generally true that D's and F's together comprise approximately one fifth of the recorded grades.

	All	Math1501	MaProf_1	MaProf_{22}	MaProf_{27}
A's	25.47	21.58	22.42	36.10	21.60
B's	29.01	27.23	33.91	38.10	15.50
C's	25.53	27.08	22.84	15.90	17.00
D's	11.10	12.91	10.14	3.90	22.10
F's	8.83	11.15	10.74	5.70	23.80

Table 2: Mean percent of letter grades for all undergraduate classes with at least ten students, all sections of Math1501 (Calculus I), and all classes taught by math professors #1, #22, and #27 (names concealed). In general, A's, B's, and C's make up a greater percentage of the grade distribution than do D's and F's.

Probability Models

To discover unexpected results, it is insightful to examine the distribution of individual letter grades in each data set by constructing frequency distributions (Figure 1). The x-axis of each histogram ranges from 0 to 100 and is divided into twenty intervals (each with length five). Choosing a different interval length will produce different frequency histograms. For future research, it would be interesting to use different interval lengths and observe whether and how the results change. The y-axis represents the relative frequency that the percent of A's, B's, C's, D's, or F's given for the data set falls in each interval. At first glance, one notices that the distribution of A's, B's, and C's have approximately the same shape, and that the distribution of D's and F's have approximately the same shape. This gives two distinct groupings.

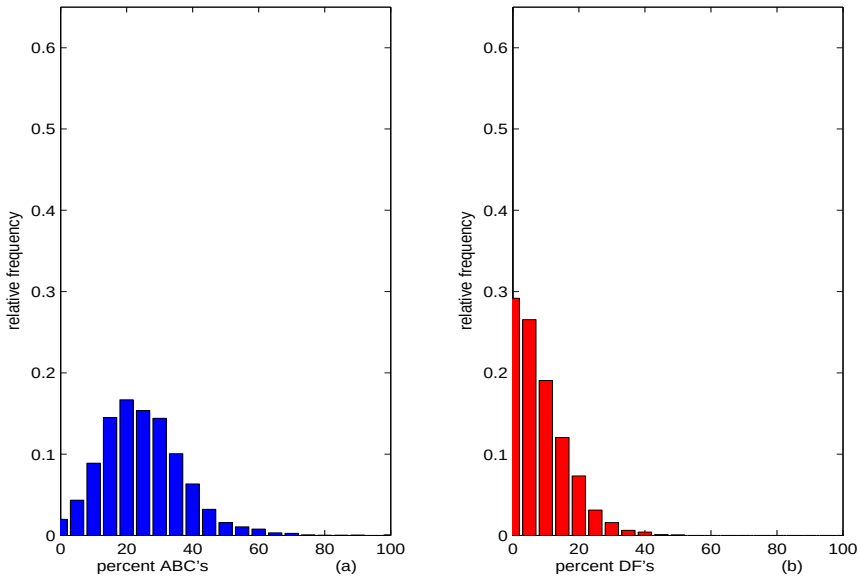


..

Figure 1: Frequency distributions of the percent of A's, B's, C's, D's, and F's in all classes with at least ten students. The plot in the lower right corner shows the outlines of each of the other histograms on the same axis so they may be more easily compared. The distributions of A's, B's, and C's are shaped similarly to each other, and the distributions of D's and F's are shaped similarly to each other.

We depict the data with A's, B's, and C's on the same histogram and D's and F's together on another histogram in Figure 2. We wish to determine what type of probability distribution best describes A's, B's, and C's, and what type best describes D's and F's. We observe that A's, B's, and C's appear to be roughly described by shifted normal distributions, whereas D's and F's appear to be exponentially distributed.

The probability density function of the unshifted normal distribution is



..

Figure 2: Frequency histograms for the set of all classes with at least ten students with A's, B's, and C's combined on the left graph, and D's and F's combined on the right graph.

defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_n)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty, \tag{4}$$

where μ_n is the mean and σ is the standard deviation. The mean and standard deviation are computed for each set of A's, B's, and C's, and the values are inserted into the normal probability density function to obtain a fit for each graph (Figures 3a,b,c). Although the shape seems to match reasonably well, a small shift to the left would improve the accuracy of the approximation. We see this in most of the fittings of the distribution of A's, B's, and C's.

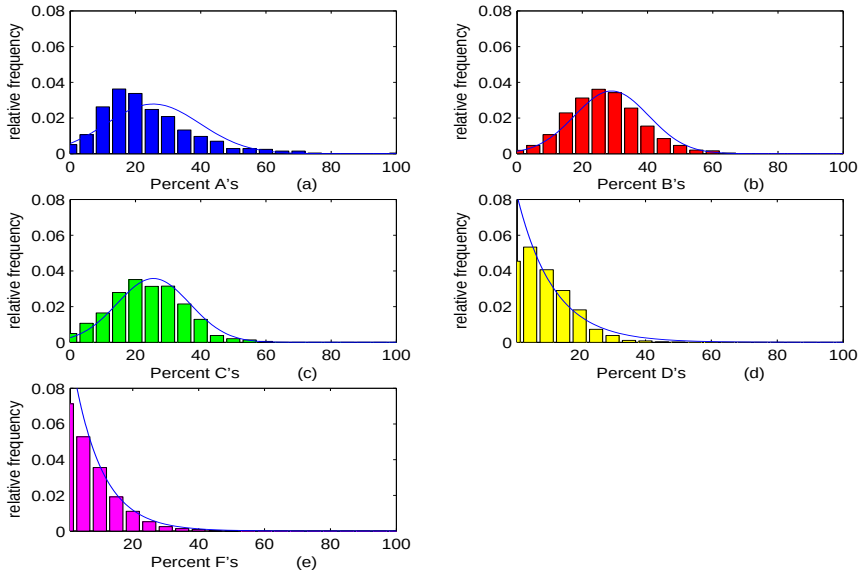


Figure 3: Frequency histograms of percent A's, B's, C's, D's, and F's in the set of all undergraduate classes with at least ten students and the superimposed distribution fittings. Graphs (a), (b), and (c) are approximated by a shifted normal distribution although we fit it to an unshifted normal here. Graphs (d) and (e) are approximated by an exponential distribution.

The probability density function of the exponential distribution is

$$f(x) = \frac{1}{\mu_e} e^{-x/\mu_e}, \qquad 0 \leq x < \infty, \tag{5}$$

where μ_e is the mean. We compute the mean percent of D's and F's for each set of raw data and insert the value into the exponential probability density function to obtain a fit for each graph (Figures 3d,e). As one can see, the distribution of D's and F's is approximated well by the exponential distribution.

A reasonable conclusion of this analysis is that instructors have a notion that the letter grade D carries the same stigma as an F, whereas A's, B's,

and C's do not. Although technically an F is a failing grade, and D is still a passing grade, the idea is that a D is like a failing grade because many courses require at least a C in order for the class to count towards one's degree. In addition, many majors have GPA cutoffs for good academic standing, so students who do not maintain a minimum GPA are in danger of being dropped from their major. In the math department, for example, the cutoff is 1.70 for freshmen, 1.80 for sophomores, 1.95 for juniors, and 2.00 for seniors. Some departments, such as the School of Industrial and Systems Engineering (ISYE), also have GPA cutoffs to be eligible to transfer into the major. (The ISYE GPA requirements range from 2.4 to 2.8.) Furthermore, the minimum grade point average required to graduate from Georgia Tech with an undergraduate degree is 2.0 (a C average). In other words, a D average is not enough to receive a diploma, despite the fact thata D is technically a passing grade. Thus, as confirmed by our analysis, D's are treated in the same manner as F's when grades are assigned.

Another possible conclusion is that because there are only two types of grades (based on their distribution), having five letter grades is unnecessary. That is, a pass/fail system could accomplish the same goal. This may not be the case for all universities, however, as most do not follow a five-grade system. At a school where each letter grade can also include a plus or minus (where a B^+ has a different grade point value than a B^-), different grades might be treated differently. One might also expect different distributions and different groupings. It is thus worthwhile to repeat our study at other universities to analyze data across different schools.

3 Singular Value Decomposition

Singular value decompositions (SVDs) were originally introduced as an alternative to spectral decomposition. Spectral decomposition decomposes a positive definite matrix A into CLC^T, where L is a diagonal matrix who entries are the eigenvalues of A and C is a matrix of the corresponding eigenvectors of A. However, we cannot use this method of decomposition on $m \times n$ matrices. The SVD theorem states that we can always decompose a real $m \times n$ matrix M into the following form

$$M = U\Sigma V^T \ , \tag{6}$$

where Σ is a diagonal matrix, U is an $m \times m$ matrix and V is an $n \times n$ matrix.[4]

3.1 Computing the SVD

The matrices U and V are constructed from the eigenvectors of MM^T and M^TM, respectively. Because MM^T and M^TM are square, symmetric, real matrices, their eigenvalues are real and positive and their eigenvectors can be made orthonormal. Thus, U and V satisfy

$$UU^T = I_m \quad VV^T = I_n \ .$$

The matrix Σ can be viewed in block form as two matrices If $m > n$,

$$\Sigma = \begin{pmatrix} D \\ 0 \end{pmatrix} \ ,$$

and if $m < n$

$$\Sigma = \begin{pmatrix} D & 0 \end{pmatrix}$$

where D is an $n \times n$ or $m \times m$, respectively, diagonal matrix and 0 is an $|m - n| \times n$ matrix of zeros.

The first k diagonal entries $\sigma_1, \sigma_2, \ldots, \sigma_k$ of the matrix Σ (where M is a rank-k matrix) are the square roots of the eigenvalues of MM^T and M^TM, all of which are positive. The so-called *singular values* σ_i are ordered so that

$$\sigma_1 > \sigma_2 > ... > \sigma_k \ .$$

The remaining $k + 1 \rightarrow n$ entries of D are zero, corresponding to M's nullspace.

Additionally,

$$AA^T u_i = \sigma_i u_i \quad A^T A v_i = \sigma_i v_i, \quad i \leq \min(m, n)$$

where u_i and v_i are the ith columns of U and V, respectively.

We also note the extremely important expansion that follows from equation 6, one obtains the expansion

$$M = \sum_{i=1}^{n} \sigma_i u_i v_i^T \text{ for } n < m \ , \tag{7}$$

which is sometimes written as

$$A = B_1 + B_2 + ... + B_n \ , \tag{8}$$

where the B_i are *modes* of M.

The orthogonality of U and V implies that

$$M = U\Sigma V^T \Rightarrow MV = U\Sigma \ ,$$

which can be written separately for each of the modes,

$$Av_i = \sigma_i u_i \ , \ i = 1, \ 2, \ \ldots \ , \ n \ .$$

3.2 Applications of the SVD

Data represented by matrices often contains large amounts of redundancy. The leading modes in equation 8 represent most of the data. Using SVDs, we can construct a *lossy* compressor. Lossy means that we are guaranteed that we can make the file less than or equal to its current size, but we are not guaranteed to retain all of the information. The use of lossy versus lossless compressors is explained in more detail in the appendix.

Here, we use SVDs for noise filtering. We look at the larger singular values and their associated modes, as they contain most of the data and the smaller singular values are treated as noise. By ignoring the modes associated with the smaller singular values, we can try to reconstruct our data and effectively filter out some of the "noise".

4 SVD of Grade Distributions

At the end of every semester, the Georgia Tech math department records the grade distribution for each class into a database. Large classes are divided into smaller sections, so that all the grade distributions are recorded over

groups of about 10-40 students. Each section is an entry in the database. We can think of each entry as a vector of percentages. For example, the vector

$$g = (25,\ 27,\ 29,\ 9,\ 10)$$

represents a section whose instructor gave 25% A's, 27% B's, 22% C's, 9% D's and 10% F's.

The data we used originally contained the percentages for A's, B's, C's, D's, F's and W's, where W's are the students who withdrew from the class during the semester. The percentages included every student registered for the class, so the row sums did not always equal 100 because data was not included for students who took the class as a pass/fail course. We adjusted the percentages to make the row sums 100.

First, we made a matrix consisting of all such vectors from undergraduate classes (classes numbered 1XXX, 2XXX, 3XXX or 4XXX). We only included courses with at least 10 students.

Computing the SVD of the undergraduate matrix M gives us the matrices U and V and the vector σ, which contains the five singular values $\{\sigma_i\}$ (the diagonal entries of the Σ matrix). For the undergraduate matrix,

$$\sigma = (1860,\ 643.9,\ 469.6,\ 358.1,\ 272.1)^T\ .$$

We then normalize σ by defining

$$\widetilde{\sigma_i} = \frac{\sigma_i}{\sum_{i=1}^{5} \sigma_i^2}\ .$$

Dropping the $\tilde{}$ for convenience, we obtain a vector of percentages

$$\sigma = (0.805,\ 0.0965,\ 0.0513,\ 0.0298,\ 0.0172)^T \ .$$

For this matrix, the first singular value encodes about 80% of our original information. Using the second singular value as well, we can retain about 90%. But what does this mean?

Digital Signal Processing Theory

We motivate the answer to our question with digital signal processing theory. In digital signal processing, one has a matrix M corresponding to a noisy signal. We compute the SVD of the matrix M and discard the smaller singular values, which represent noise [5].

Neglecting the smaller singular values, we can reconstruct a new projection matrix of rank k (where k is the number of remaining singular values), given by

$$M_k = \sum_{i=1}^{k} u_i \sigma_i v_i^T \ .$$

Recall from section 3 that this is the truncated matrix. We have projected the original data onto a basis of the first k modes. Our "noise" comes from abnormal data. Extreme situations, such as an entire section receiving A's would skew our data and thus viewed as noise. and no other grades, which would skew our data, are seen as noise are made to look more like the leading modes.

To analyze these projection matrices, we look at Figures 4 and 5. The

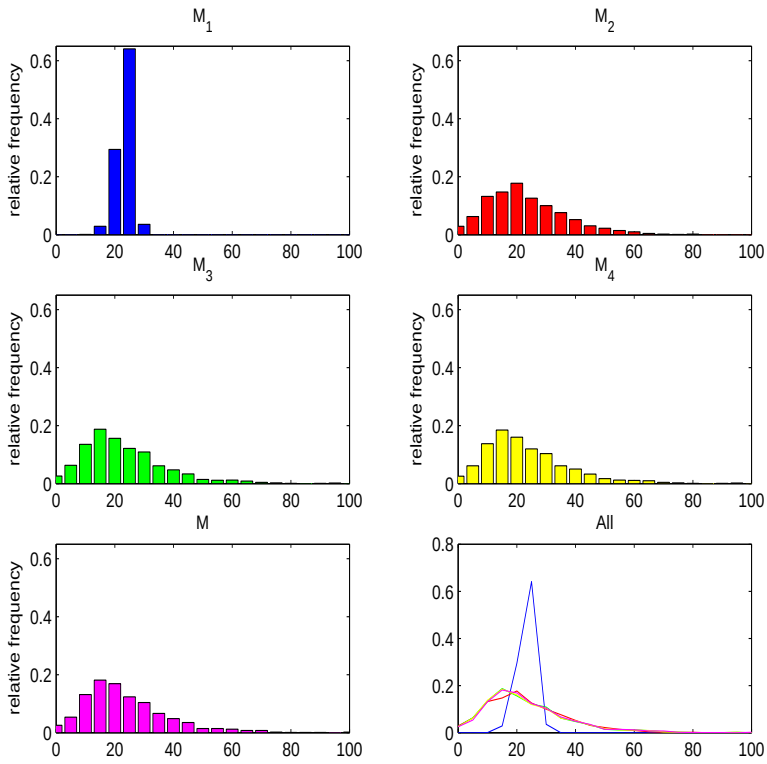


Figure 4: Frequency distribution of the percentage of A's given by instructors in undergraduate classes. We also show the frequency distributions for M_2, M_3, M_4, and the original data from M. The last plot shows all five graphs on the same axis. As one can see, M_1 has a steep peak, whereas the rest of the graphs are nearly identical. Hence, a two-dimensional truncation suffices to explain the data

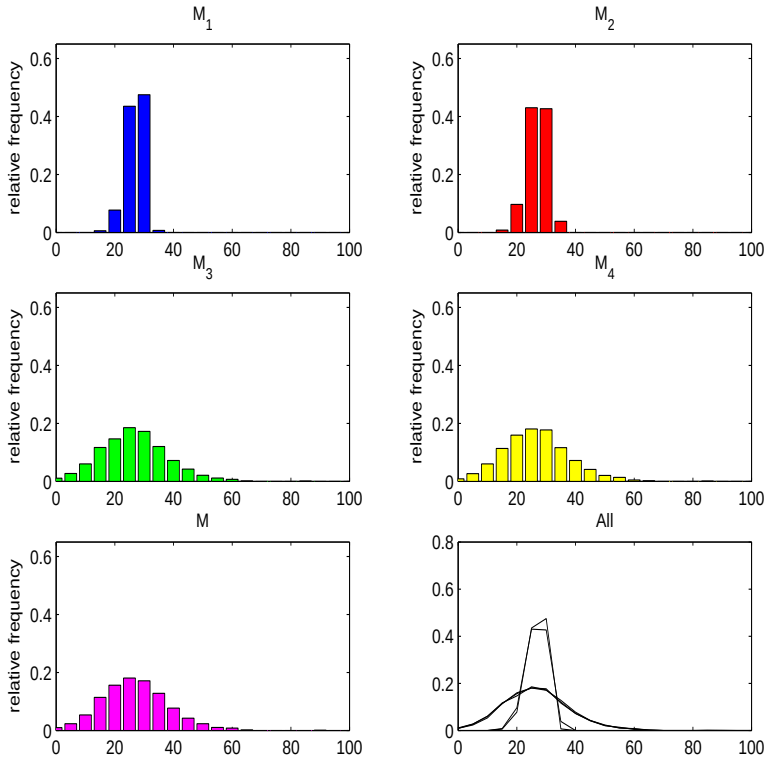


Figure 5: Frequency distribution of the percentage of B's given by instructors in undergraduate classes. The first graph is for M_1, the frequency distribution obtained when just keeping one singular value. We also show the frequency distributions for M_2, M_3 and M_4 and the original data from M. The last plot shows all five graphs on the same axis. As one can see, M_1 and M_2 are nearly identical for the distribution of B's, whereas M_3 and M_4 are nearly identical to each other and the original data. A three-dimensional truncation suffices to explain this data.

former shows the distributions of the percentage of A's given out by instructors in different sections. Keeping just one singular value gives the original mean, but the effects of the extreme cases have been largely ignored. Keeping two singular values produces nearly the same distribution as the original distribution in M. Hence, even when we keep 80% of the information, one still misses vital information from the distribution of A's, but when we keep 90% we have virtually everything we need.

Figure 5 shows the distributions of the percentage of B's. Keeping one singular value gives us the original mean, but keeping only one or two singular values insufficiently approximates the original grade distribution. We can conclude that there is more noise in the distribution of B's than that of A's.

We considered some possibilities for this difference in the distribution of the B's as compared to that for A's. In Georgia, a significant number of students that go to college get the HOPE scholarship for living in Georgia and having above a 3.0 grade average [6]. Because receiving a B at Georgia Tech earns the student a 3.0 towards their grade point average, this grade is typically considered very pivotal by students and professors. One might also question the reliability of a 3.0 grade average, because of the significant amount of noise in this range of grades. Hence, this SVD analysis highlights a significant concern in the grades handed out at Georgia Tech.

Let's now consider C's, D's, and F's. We find that C's behave much like A's in that their distribution can be approximated with just two singular values when including all undergraduate classes. However, for D's and F's, it is necessary to keep four singular values to get a good approximation for the original data. That is, D's and F's contain the most. We can compare

this finding with that found earlier in the frequency distributions. The D's and F's have the most extreme tails, foreshadowing this significant noise.

While this information is interesting, it is particularly instructive to apply SVDs to individual classes and instructors. Figure 6 is similar to Figure 4, except we only use information from sections of Math 1501 (Calculus I). Nearly every student at Georgia Tech must take this course, so it is one of the most significant courses at the school. For a student trying to predict grade distributions for a future class of Math 1501, he or she would get the most accurate prediction by reducing the impact that noise has on the data. In Figure 6, we see that using M_2 allows us to more accurately predict a future grade distribution. In this particular example, sections in which more than 50% of the students received A's are not present in M_2 and there are fewer section with no A's. Both these cases are unlikely in a section of Math 1501, thus we consider them "noise."

Figure 7 compares the grade distribution of the M matrix to that of M_2, the 2-dimensional truncation of M.

5 Entropy of Grade Distributions

We can also use entropy to quantitatively measure the average information content of grades. First, we connect the idea of relative frequency of an outcome in a data set to the probability that the outcome will occur, assuming that grades follow the same distribution and that it is valid to make inferences based on our data. We obtain an ensemble X for each letter grade from the relative frequency histogram. The set of possible outcomes, $\mathcal{A}_X$, consists

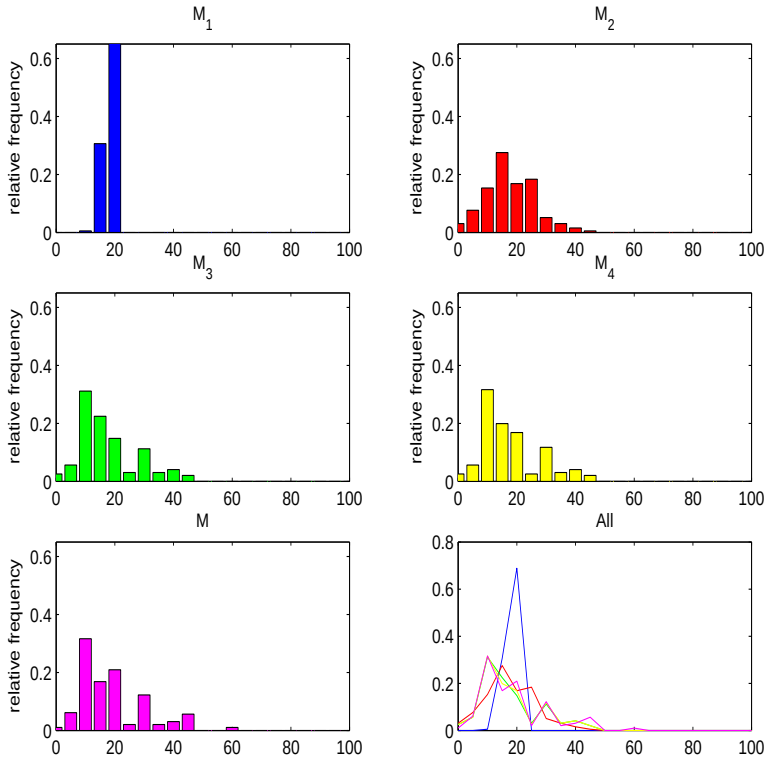


Figure 6: Frequency distribution of the percentage of A's given in Math 1501. The first graph depicts M_1, the frequency distribution obtained in the one-mode projection. We also show the frequency distributions for M_2, M_3, and M_4 and the original data from M. The last plot shows all five graphs on the same axis. Though M_1 preserves the mean, we have lost some important information. In M_2 the mean and most of the important information is preserved, while data at the tails is discarded in favor of the leading modes.

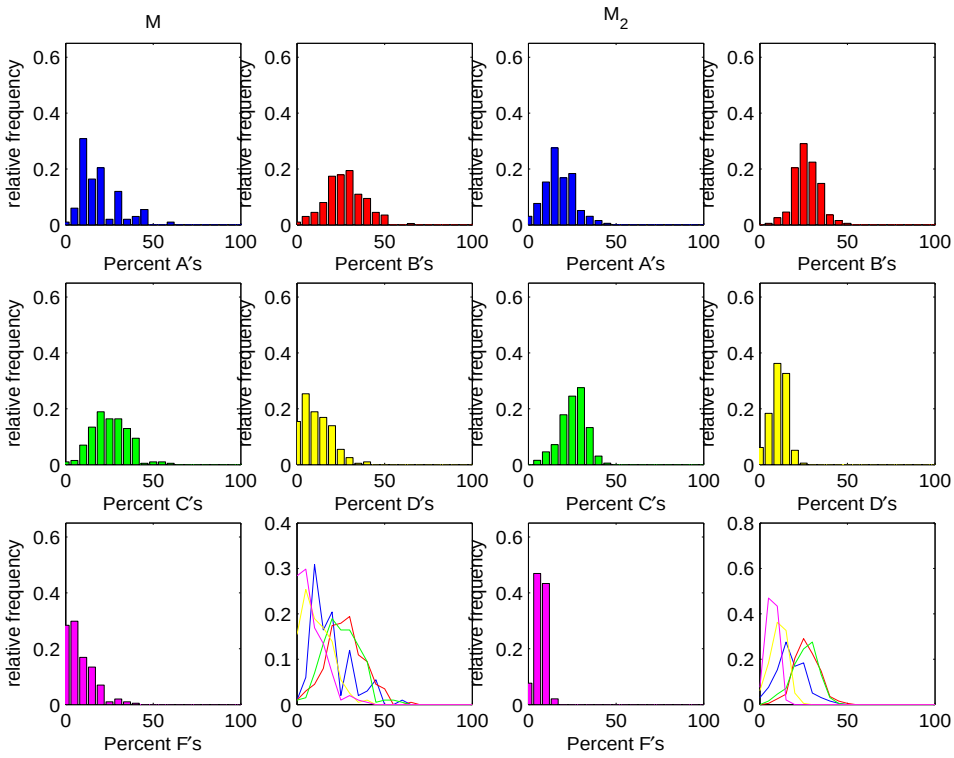


Figure 7: Comparison between the original matrix M and the truncated matrix M_2. The means remain the same, but the data is more evenly distributed around the mean in M_2. The leading modes take over the data and the noise at the tails has been reduced.

of the twenty intervals that we use for constructing the frequency histograms, and outcomes have probabilities equal to the relative frequencies:

$$\begin{aligned} \mathcal{A}_X &= \{[0,5),[5,10),\ldots,[95,100]\}, & |\mathcal{A}_X| = 20 \\ \mathcal{P}_X &= \{p_1,p_2,\cdots,p_{20}\}, & \sum_{i=1}^{20} p_i = 1. \end{aligned} \tag{9}$$

Now, using equation (2), we calculate the entropy for the letter grades A, B, C, D, and F for each data set. The maximum value of entropy in this case is

$$\max H(X) = \log_2 |\mathcal{A}_X| = \log_2 |20| \approx 4.3219. \tag{10}$$

This maximum value would change if we chose a different interval length because then the number of intervals (or outcomes) $|\mathcal{A}_X|$ would change. The maximum value occurs when the probability distribution is uniform, so an entropy close to 4.3219 bits implies that the distribution is close to uniform. We have already found that the distribution of letter grades can be closely modelled with a shifted normal or exponential distribution, rather than the uniform distribution, so we expect that the entropies will not be close to the maximum value. Table 3 shows the entropies of the ensembles corresponding to each letter grade for the set of all undergraduate classes with at least ten students and all class sections taught by math professors #15 and #30.

We interpret these entropies as average information content, or the number of bits of information an average piece of data supplies. However, we do not see any clear pattern with these entropies across data sets. In some cases, one acquires more information from an outcome of A, and in others, there is more information from an outcome of F. A high entropy implies that it is almost equally probable for an outcome to fall in any of the twenty intervals

	All	MaProf_{15}	MaProf_{30}
$H(X_A)$	3.3959	1.9212	3.4031
$H(X_B)$	3.2661	2.2028	3.9038
$H(X_C)$	3.2628	2.1800	3.7919
$H(X_D)$	2.7128	2.3826	3.6880
$H(X_F)$	2.5034	2.4581	3.3710

Table 3: Entropies of ensembles for A's, B's, C's, D's, and F's using the data set of all undergraduate classes with at least ten students, and all class sections taught by math professors #15 and #30. Math professor #15's relatively low entropies reflect a tendency to give grades in a certain interval. Math professor #30's relatively high entropies reflect more uniform (random) distributions.

of percents; that is, the instructor does not favor a particular percentage range for the letter grade. A low entropy implies the opposite; there may be some intervals in which outcomes are more probable, and the instructor may consistently give grades in those intervals.

We computed entropies of the ensembles for each letter grade for thirty-seven instructors; of these, the average of the five entropies was lowest for math professor #15 and highest for math professor #30 (Table 3). We plot the rows of the matrices formed by these data sets to examine how grade distributions of single class sections vary within each data set (Figure 8). The distributions for math professor #15 follow a clearer pattern than do those of math professor #30. Thus, math professor #15 favors a particular distribution of grades, whereas math professor #30 does not; this is reflected

by the entropies.

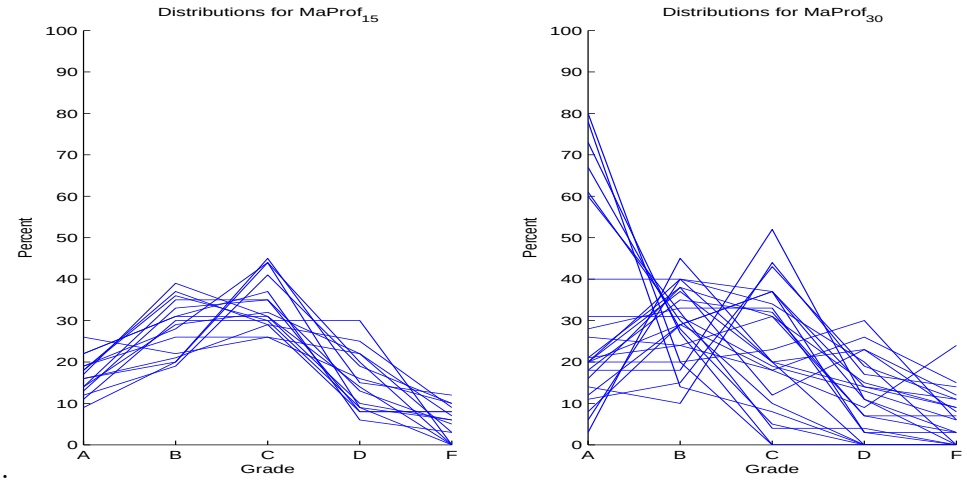


Figure 8: Line plots of the rows of the matrices for math professors #15 and #30. Math professor #15 favors a particular distribution of grades, whereas math professor #30 does not.

Instead of considering each distribution of letter grades as a separate ensemble, we can examine the overall entropy of a data set with a different ensemble, Y, defined so that the outcomes are the letter grades. We assign each letter grade a probability equal to the expected value (i.e., mean) of the letter grade. For the ensemble Y,

$$\begin{aligned} \mathcal{A}_Y &= \{A, B, C, D, F\}, \qquad\qquad |\mathcal{A}_Y| = 5, \\ \mathcal{P}_Y &= \{\mu_A, \mu_B, \mu_C, \mu_D, \mu_F\}. \end{aligned} \tag{11}$$

The maximum entropy is now

$$\max H(Y) = \log_2 |\mathcal{A}_Y| = \log_2 |5| \approx 2.3219. \tag{12}$$

An entropy close to 2.3219 now means that grades are nearly uniformly distributed across the five letter grades. We saw earlier that this is usually not

the case, so we do not expect that the entropies will be very close to the maximum value.

Table 4 shows the entropies for all thirty-seven math professors for whom we had sufficient data. The highest entropy is that of math professor #27, whose entropy 2.3032 is close to the maximum, 2.3219, corresponding to a nearly uniform distribution among the letter grades. The lowest entropy belongs to math professor #22, at 1.0910, whose distribution of letter grades is more greatly skewed. Recall from Table 2 that math professor #27 assigned a large number of D's and F's, whereas math professor #22 assigned, on average, very few D's and F's (less than 10% combined) and had relatively high means for the percent of A's and B's given.

A low entropy does not indicate the outcome(s) toward which the probabilities are skewed; it may be that there are very few A's, B's, and C's, but many D's and F's. However, based on the assumption that instructors generally give more A's, B's, and C's, we may conclude that a low entropy is likely to be indicative of a very low percentage of D's and F's in the data set. A high entropy indicates that the chances of getting an A is approximately the same as the chances of getting any other grade. We might compare this situation to an instructor who has a fair five-sided die and assigns grades by rolling the die.

Given these results, a student would likely prefer professors with a low overall entropy. Assembling these results allows students to easily compare instructors by comparing a single value instead of looking at all the data for each class section. It would also be useful to compute entropies not only for different instructors, but also for different classes or semesters in order

MaProf	entropy	MaProf	entropy	MaProf	entropy	MaProf	entropy
1	2.1798	11	2.2116	21	2.1945	31	2.0314
2	2.0894	12	2.2649	22	1.9010	32	2.2205
3	2.2422	13	2.1870	23	2.0553	33	2.1913
4	2.0975	14	2.2778	24	2.1766	34	2.2303
5	1.9419	15	2.1197	25	2.2637	35	2.0889
6	2.2556	16	2.2788	26	2.0901	36	2.0421
7	2.1597	17	2.2329	27	2.3032	37	2.2104
8	2.1957	18	2.0740	28	2.2353	All	2.1846
9	2.1649	19	2.1064	29	2.0124		
10	1.9379	20	2.0816	30	2.1127		

Table 4: Entropies for math professors #1 through #37 and for all undergraduate classes with at least ten students. Math professor #27, whose distribution of letter grades is close to uniform, has the highest entropy. Math professor #22, whose distribution of letter grades is skewed, has the lowest entropy.

to aid students in selecting which classes to take and when to take them. For example, most freshmen who test out of Calculus I take Calculus II in the fall, and these students are usually very bright and were motivated to take college-level calculus during high school. Therefore, taking Calculus II in the fall may be different from taking it in the spring because the student make-up of the classes are often different.

6 Conclusions

By considering a probability model for grade distributions, we have unearthed a few interesting results. Firstly, when taken over many sections, the frequency distribution of A's, B's and C's given in a course can be approximated by a normal distribution, while the D's and F's are approximated by an exponential distribution. This implies that when distributing grades, instructors treat D's like F's. This can be attributed to the fact that being in good academic standing at Georgia Tech requires a student to have above a C grade average.

When we consider the grade distribution database as a noisy system and compute the SVD, we discover that a projection onto the first mode will retain the original mean, however the distribution is largely skewed. For A's and C's, a 2-dimensional truncation accurately approximates the data, however for B's a 3-dimensional truncation is needed. A significant factor in this difference comes from students struggling to maintain at least a B average in order to keep scholarships. However, the fact that the B's need a 3-dimensional truncation for a good approximation implies that the variance

of the B's is higher than that of A's or C's, and so the grade of a B is less reliable as a judge of a student's performance in a class. D's and F's required a projection onto 4 modes to accurately approximate the data. Thus, D's and F's have the largest amount of noise data.

Once again considering the probability model for grade distributions, we can use the frequency distributions to compute entropies for different professors. While interpreting the exact meaning of the entropy of an ensemble is difficult without knowing any more information about the ensemble, we know that a high entropy implies that the instructor associated with the ensemble does not favor a particular grade distribution or letter grade. A low entropy implies that certain grade distributions are more probable than others and the instructor consistently gives this pattern of grades. While we cannot exactly judge the grade distribution that an instructor with a low entropy gives out, based on our knowledge that A's, B's, and C's have a higher frequency, we can conclude a low entropy is most likely associated with a low percentage of D's and F's in the data set. A high entropy implies the chance of getting an A is the same as that of an F, or any other grade.

For our future research, we intend to analyze the entropy of different courses during different semesters and comparing our results. In addition, doing the same computations for other departments of Georgia Tech and other colleges will be important to assure that the math department of Georgia Tech is not a special case. We also would like to discover the effect of grading systems that make use of +/- instead of only 5 letter grades.

Appendix: Background Research

Before beginning our research, we studied David MacKay's book, *Information Theory, Inference, and Learning Algorithms*[2]. We focused on the chapters about information theory. The following subsections will summarize what we covered.

A.1 Introduction to Information Theory

One of the most important questions raised in information theory is how to communicate perfectly over an imperfect, noisy communication channel. For instance, on a computer one may want to write some information, in the form of a binary string, from the memory to the disk drive. On the disk drive we, distinguish a 0 from a 1 by aligning a magnetic strip in one direction for a 0 and another for a 1. There is a chance that these bits may spontaneously flip, resulting in noise.

There are two types of ways to deal with such noise problems. One is improving the physical system, perhaps by using better materials to construct the disk drive. The other uses information and coding theory. More specifically, one can add redundancy through an encoder and decoder to a message to reduce the risk of bit error. Two of the simpler examples of encoding are *Repetition Codes* and *Block Codes*.

Repetitions codes repeat every bit in a message N times. If the original probability of bit error is f, the repetition code has a new probability of bit error

$$p_b = \sum_{n=(N+1)/2}^{N} \binom{N}{n} f^n (1-f)^{N-n}$$

However, there is a trade-off, as our new rate of communication is $1/N$.

On the other hand, block codes provide a way of converting a source message $\mathbf{s}$ of length K into a transmitted message $\mathbf{t}$ of length N. The first K bits of $\mathbf{t}$ are the original message $\mathbf{s}$ and the last $N-K$ bits are parity checks on the original message. In the $(7,4)$ *Hamming Code*, a type of block code, an original message of length four is sent as a message of length 7. The first four bits are the original message and the last 3 bits are parity checks. The probability of bit error for the $(7,4)$ Hamming Code is

$$p_b = \sum_{r=2}^{7} \binom{7}{r} f^r (1-f)^{7-r}$$

The new rate of communication here is $4/7$.

Although there seems to be a trade-off between bit error and communication rate, Claude Shannon proved that the maximum possible rate of communication does not vanish as the probability of bit error goes to zero.

A.2 Shannon Information and Entropy

The most important results of this section concern Shannon information content and Shannon entropy.

We defined the *Shannon information content* for an outcome x in section **??** as

$$h(x) = \log_2 \frac{1}{P(x)}$$

The function $h(a_i)$, a natural measure of the information content of an event $x = a_i$, is measured in bits.

The *entropy* of an ensemble X is defined as the average information content in the set of outcomes, $\mathcal{A}_X$,

$$H(X) \equiv \sum_{x \in A_X} P(x) \log \frac{1}{P(x)} \ .$$

The entropy is also measured in bits.

The entropy function satisfies the following properties:

1. $H(x) \geq 0$ with equality if and only if (iff) $p_i = 1$ for only one i.

2. Entropy is maximized if the probabilities of the outcomes in the set $\mathcal{A}_X$ are uniform. That is

$$H(X) \leq \log(|\mathcal{A}_X|) \quad \text{with equality iff} \quad p_i = 1/|X| \quad \text{for all } i$$

The *relative entropy* or *Kullback-Leibler divergence* between two probability distributions (defined over the same outcome set $\mathcal{A}_X$) is

$$D_K L(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \ .$$

The relative entropy satisfies *Gibbs' Inequality*

$$D_K L(P||Q) \geq 0 \ ,$$

where equality equality if and only if $P = Q$.

A.3 Inference

In *inference*, probabilities are interpreted not as frequencies or proportions, but rather as degrees of belief. For an example of an inference problem,

consider a bent coin. We toss it F times and observe a sequence of heads and tails. We want to know the bias of the coin and want to predict the probability that the next toss will be a head.

To do this, we must first make an assumption about the prior distribution. Inference is always conditional on assumptions. When doing inference problems, it is important to be explicit about one's assumptions so that they can be easily noted and (if necessary) modified. One's computations also must be easily reproducedible. Popular applications of inference include the analysis of evidence in legal cases.

A.4 Lossy versus Lossless Compression

Consider a guessing game in which a player attempts to identify an outcome by asking as few yes/no questions as possible. Assuming all outcomes are equally probable, the answer is yes or no with probability 0.5 each, and the questions are independent. Each question has Shannon information content $\log_2 \frac{1}{1/2} = 1$ bit. If the answer to each question is encoded with 1 for yes and 0 for no, then each outcome is encoded by a binary file with length $\log_2 |\mathcal{A}_X|$, where $\mathcal{A}_X$ is the set of all possible outcomes. This length, defined to be the raw bit content of X, is the minimum number of questions necessary. The total information content equals the total length of the binary file.

We gain the most information from an outcome if the probability distribution over the outcomes is uniform. Also, less probable outcomes provide more information than more probable outcomes. For example, the probability that an English word begins with 'xyl-' is very small, whereas there is a much greater probability that an English word begins with 'pro-'. That the

first three letters of a word are 'xyl-' conveys a lot information, as it is easy to guess the remaining letters in the word. Conversely, it is unlikely that we would correctly guess the word beginning with 'pro-'.

There are two types of data compression: lossy and lossless. A lossy compressor compresses some files, but others may have the same encoding and may be confused with each other. If the probability that a lossy compressor will fail, δ, is small, then it may still be useful. The smallest δ-sufficient subset, S_δ is formed by adding elements in order from most probable to least probable until the total probability satisfies

$$P(x \in S_\delta) \geq 1 - \delta. \tag{13}$$

The essential bit content of X,

$$H_\delta(X) = \log_2 |S_\delta|, \tag{14}$$

is then the length of the binary encoding of each element in the smallest sufficient subset.

Extended ensembles, X^N consist of a string of N independent, identically distributed (i.i.d.) random variables from X. The essential bit content of X^N approaches $NH(X)$, where $H(X)$ is the entropy of a single random variable. This leads to Shannon's source coding theorem:

Theorem 1 *Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,*

$$|\tfrac{1}{N} H_\delta(X^N) - H| < \epsilon,$$

where H is the entropy of the ensemble X.

This means that N i.i.d. random variables compressed into more than $NH(X)$ bits has negligible risk of information loss, whereas it is almost

certain that information will be lost if they are compressed into fewer than $NH(X)$ bits.

A lossless compressor encodes all files with distinct encodings. A well designed lossless compressor has high probability of shortening a file and low probability of lengthening one, although it is necessary that both shortening and lengthening occur. Binary symbol codes map each outcome of an ensemble to a 'codeword', or binary string, $c(x)$ with length $l(x)$. A string of outcomes is encoded by concatenation of the corresponding codewords, and any encoded string must be uniquely and easily decodeable. Prefix codes, which can be represented by binary trees, are uniquely decodeable symbol codes in which no codeword is a prefix of another. Furthermore, any set of codewords satisfying the Kraft inequality,

$$\sum_{i=1}^{|\mathcal{A}_X|} 2^{-l(x_i)} \leq 1, \tag{15}$$

always has a prefix code with the given codeword lengths.

Symbol codes must also achieve the most efficient compression possible; that is, the expected length of a code should be small. The expected length is minimized for code lengths $l(x_i)$ equal to the Shannon information content $\log_2(1/p_i)$, so the minimum expected length equals the entropy. This occurs when equation (15) is an equality and leads to the following source coding theorem for symbol codes:

Theorem 2 *For an ensemble X, there exists a prefix code C with expected length $L(C, X)$ satisfying*

$$H(X) \leq L(C, X) < H(X) + 1.$$

To find the optimal prefix code, we employ the Huffman coding algorithm, which builds a binary tree from the bottom up. The algorithm combines the two symbols with the smallest probabilities into one symbol and iterates this procedure until all symbols have been used. This method insures that the least probable symbols have the longest codewords, minimizing the expected length.

Acknowledgements

We would like to thank the following people for their assistance in the research project: Shui-Nee Chow, Mason Porter, Alexander Grigo, and Ying Wang served as our advisors. Tom Trotter, Rena Brakebill, and Rhonda Mozingo assisted in data acquisition, and Hao Min Zhou provided useful discussion about image processing. This project was funded by the Georgia Tech Math Department's REU program, funded by an NSF VIGRE grant.

References

[1] Georgia Tech Course Critique. Website, June 2004. www.sga.gatech.edu/critique/.

[2] David MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[3] Georgia Tech Course Critique Database. Website, June 2004. www.sga.gatech.edu/critique/Database.php.

[4] G. Golub and C. van Loan. *Matrix computations*. The Johns Hopkins University Press, London, 1996.

[5] Per Christian Hansen. The truncated svd as a method for regularzation. *BIT*, 27:534–553, 1987.

[6] Georgia's HOPE Scholarshp Program. Website, September 2004. www.gsfc.org/HOPE/Index.cfm.