Accounting: An Information Science

John C. Fellingham
Max M. Fisher College of Business
Ohio State University

2017

Contents

Preface

It is said we live in an *information* age. A number of academic disciplines are referred to, and refer to themselves, as "information sciences." Biology, for example, is heavily motivated by the genetic code, "code" being an important information science word.

So here's a question. Consider a seminar of the best information scientists in the University. Does accounting deserve a seat at the table?

Accounting does come equipped with some unique scientific tools. For example, the double entry system is a linear *code*. (There's that word again.) Indeed, having been in existence for over 5 centuries, the double entry system has a claim to being the oldest, largest, and most famous linear code.

Suppose one starts with the accounting code structure, and develops information theorems. That puts one right in the middle of information issues. One problem, for example, is the achievability of error free transmission through a noisy channel (using a code). This one was resolved theoretically in a surprising way by Claude Shannon, often referred to as the father of the information age.

Another issue is the flip side of error free transmission: that is, how to restrict any information at all from traveling through the channel. That means secret *codes*, and that word appears again. Accountants are routinely charged with safeguarding information. For example, keeping track of the academy award ballots.

This manuscript is intended as a modest case for the academic discipline of accounting as a participant, in good standing, at the information science table. The centerpiece of the case is chapter 8 a bit pretentiously titled "the fundamental theorem of accounting." The theorem establishes (under noted conditions) an equivalency between accounting rate of return and information denominated in Shannon entropy. In other words, the fundamental accounting problems of asset

valuation and income determination can be equivalently framed as an information problem. In addition, any information problem stated in terms of information (entropy reduction) can be equivalently framed as an accounting problem.

Another implication of the theorem in chapter 8 is that paying serious attention to serious accounting has social welfare implications. In that regard the chapter is also meant as a modest proposal for how the accounting profession can maintain the moral high ground as a supplier of an important public good.

The material contained herein was developed while teaching accounting, primarily at Ohio State. Several students provided comments, questions, and general interest for which I am very grateful. A partial list includes J. J. Ericksen, Sam Feller, George Gothot, Claire Guo, Tyler Hankins, Sarah Jadwin, Ray Johnston, Rob Lindner, Jack Parker, Robert Sledge, and Kunjue Wang.

I am also grateful to a number of academic colleagues who read the work and have been very helpful and encouraging with their comments. Here is another partial list: Rick Antle, Haijin Lin, Jonathan Ross, Eric Spires, and Dave Ziebart.

I am particularly grateful to Doug Schroeder who is always willing, and extremely able, to collaborate in figuring out problems relating to the academic discipline of accounting. And to not allow either of us to give up, no matter how difficult things become.

And then there is Joel Demski. The intellectual contributions he has made to accounting and information are unmatched. But the body of work is not as important as the example of what it means to be a true citizen of the University. There is nothing more central than getting the best possible ideas in front of the classroom. And the responsibility for doing so resides solely with faculty scholars.

Columbus, Ohio

1
accounting as an information science

The point of view adopted in this book is that accounting is appropriately treated as an academic discipline, in general, and an information science, in particular. As such, accounting both draws upon, and, importantly, contributes to the intellectual capital of the scholarly community.

Accountants are routinely characterized as suppliers of information in an economic setting. All kinds of problems accountants deal with, and all kinds of jobs accountants do, have information as a core element. At the same time other academic disciplines across the University are becoming, more and more, information sciences. Physics is an example. Pioneering physicist John Wheeler coined the phrase "it from bit" to convey the idea that reality (it) can be viewed and explained using information concepts (bits).

In addition it is not hard to find stated opinions that a large number of future career opportunities will reside in an information industry. Collecting, transforming, and organizing information, using it efficiently, and, pointedly, maintaining control of the information are viewed as important and pervasive job opportunities. The argument is made that a scholarly approach to information is a legitimate, and perhaps optimal, preparation for careers in an ever changing world. Commerce and technology can, and do, change with regularity. Carefully crafted information theorems do not. The objective is to present the best thinking the University has to offer.

It is commonplace to say that we live in an information age: what with gigantic data structures that can only be handled by enormous and speedy computers; as will as uncountable household and personal devices which individually possess more information processing power that the entire Apollo project had available to

land a man on the moon. Accounting, on the other hand, has existed for centuries. Does it belong among modern information sciences?

Consider a metaphorical table where top class scientists from around the University gather to discuss and contemplate information questions. Likely present at the table are physicists, biologists, computer scientists, mathematicians, and many others. What would accountants and the study of accounting bring to the table?

To be a full-fledged contributing participant at the information science table, accounting must come equipped with at least two things. One is a set of tools, and corresponding intellectual experience to effectively power the tools. The tools can be acquired from a serious study of accounting, and, furthermore, might not be so easily acquired in a study of other academic disciplines. The accounting double entry system constitutes one among several examples of a valuable, and distinctly accounting, contribution to information science.

The second prerequisite for full membership at the information science table is a scientific understanding of information. We will attempt to develop a reasonable amount of understanding from the study of accounting topics; information issues arise naturally when accounting is studied rigorously.

Earlier versions of this manuscript were intended, more or less, to cover as many accounting related information theorems as possible. In a way, that is still the case. But as revisions take place, one theorem, in particular, has come to occupy a central and unifying role. The mutual information theorem (which, because of its central importance, we will often refer to as the fundamental theorem of accounting) establishes the equivalence of probability measures of information with optimal growth rates (and dollar values) of general investment activity. The investment activity is general in the sense that it includes activities of economic organizations like firms, as well as investing in already existing securities. In fact, as we shall see from its proofs, the theorem is probably more applicable to non-market (firm) activity.

Accounting is on the dollar side of the theorem's equality; probabilities and information science are on the other. The theorem, then, establishes the operational equivalence of accounting and information science.

Given the mutual information theorem's central role, it is tempting to start our inquiry with the theorem, itself. There are reasons, however, not to do so. One reason is the necessity of building appropriate intellectual foundations for development and appreciation of the theorem. The first two theorems we encounter, then, will be the fundamental theorem of linear algebra and the theorem of the separating hyperplane (known as the fundamental theorem of finance when used in a economic setting). Both of the foundational theorems will be developed in the context of accounting problems and topics. Besides contributing to our intellectual foundations, the accounting problems are interesting in their own right.

After preliminary foundational work the mutual information theorem is developed and proved. We will do exercises to acquire facility with the logic, and, more importantly, gain appreciation for implications. For example, the theorem implies some statements about social welfare.

The mutual information theorem allows access to Claude Shannon's most important theorem: the noisy channel theorem about (error-free) transmission of information through any inherently disruptive channel. Shannon is the founder of information theory. In fact, it is not difficult to find statements comparing Shannon's achievements, both in terms of intellectual elegance and effects on modern life, to more well-known scientific figures like Einstein.[1]

The flip side of error correction is encryption; one activity seeks to eliminate garbling, while the other embraces it.[2] Several theorems are useful in the study of encryption including ones due to Fermat, Euler, and Euclid, as well as the fundamental theorem of arithmetic. Coding and encryption technology lead naturally to quantum information and quantum computation. Quantum encryption is apparently on the frontier of encryption technology, and we aspire to study the frontier. Our journey begins, however, with accounting.

1.1 accounting

Often accounting is studied as a vocational discipline: that is, what is required for an entry level position in the accounting profession. That is not the perspective taken here. Once again, the frame is scientific inquiry.

1.1.1 double entry accounting

Double entry accounting is undeniably an elegant system. Goethe, for example, famously wrote in 1796:

> "It is among the finest inventions of the human mind."

The elegance is intellectually satisfying, and can make the study of accounting a fun enterprise.

Double entry accounting is also durable. It has been in existence for over five centuries. While commerce and technology have changed in dramatic and unpredictable ways, double entry accounting remains recognizable. And it remains central to economic affairs. Perhaps there are reasons for its durability.

The double entry system is linear. Briefly, the first step in the construction of financial statements is the collection of a set of numbers reflecting the transactions and prospects of the entity under consideration. The numbers are collected in journal entries. The process of transforming the journal entry numbers into financial statements is a linear process, and can be represented by a linear operator

---

[1] See, for example, chapter one in *Fortune's Formula* by William Poundstone, or chapter two in *Information Science* by David Luenberger.

[2] The following is attributed to Shannon in Poundstone, p. 25: "A secrecy system is almost identical with a noisy communication system," he claimed. The two lines of inquiry "were so close together you couldn't separate them."

called a matrix. The double entry system gives the matrix a particularly simple and elegant form, which will ease our study of linear systems.

Studying a simple linear system is a good way to learn a number of concepts on linear algebra, itself a foundation of information science. Communication of information from one entity to another is typically accomplished by a coding system, and many codes are linear. Double entry accounting is, itself, a linear code.

Any serious consideration of information in an economic setting can not ignore equilibrium consequences. A simple, but powerful, equilibrium concept is arbitrage free pricing, which is, itself, a *linear* equilibrium concept. Indeed, the linear algebraic formulation of arbitrage equilibria is quite similar to a formulation of double entry accounting. The study of one leads naturally to the other, and double entry examples are used to illuminate arbitrage equilibria.

Finally, it is noted that in numerous linear systems the concept of probability arises naturally. And probabilities are the building blocks of information. For all these reasons a study of the double entry system is a good place to start our journey.

1.1.2 stocks and flows

As systems evolve over time, especially economic systems, the notions of stocks of value and flows of income (that is, change in value) are central. A systematic frame for organizing these concepts is essential, and that is provided by double entry and the tool of T-accounts. At the beginning of the time period under consideration, a T-account contains a beginning balance (stock). As time goes by, the balance is adjusted for various transactions (flows), resulting in an updated ending balance (revised stock). Information is embodied in another stock and flow structure, and it is the notion of probability that holds the structure together. A decision maker begins the time period with probabilities over states of the world - what might happen. These are called prior probabilities and are stocks. During the time period evidence is accumulated and probabilities are revised The evidence is a flow, and the revised probabilities are the updated stock. A consistent way to revise probabilities is the use of Bayes' theorem, and the updating is called Bayesian revision.

It is tempting, and suggestive, to contemplate the possibility of representing the stocks and flows of information updating using the stocks and flow mechanics of the accounting T-account. As it turns out, for illuminating special cases it is possible to do the entire Bayesian updating in a T-account. All that is required is to specify an appropriate amortization rate. Then just do the cash and amortization entries the regular accounting T-account way, and the updated probability parameters pop out naturally.

The particular problem set-up is actually quite complicated (at first glance, anyway). The T-account procedure accomplishes Bayesian revision with a minimum of fuss. That's the sort of contribution an information science is capable of.

Central to the determination of value (stocks) and income (flows) are the related concepts of interest and growth, which are accounting problems. An extremely

important number in the history of mathematics, e, the base of the natural logarithms, first appeared as the answer to an interest problem. Information scientists also find logarithms to be a useful tool. However, they tend to use logarithms to the base 2, as that matches up nicely with bits of information: the switch is on or off, for example. For problems where valuation and growth are central, natural logarithms are more insightful, and that is what we will use in the accounting.

1.1.3 data integrity

Who worries more about the integrity of numbers than accountants? Accountants routinely process a plethora of identification numbers for employees, inventory, and so forth. The data is susceptible to unintentional mistakes caused by everything from typing errors to sun spots. Even more problematic is intentional misuse of the data: from unauthorized use of credit cards and PINs to interception of confidential communications.

As far as unintentional errors are concerned, there are a variety of schemes for error detection, and even error correction. Linear codes like the international standard book number (ISBN) and the universal product code (UPC) are ubiquitous. The basic mechanism is the same as double entry accounting. Clever codes can flag errors as they occur. The error can be corrected manually, or, with linear error correcting codes, the code can correct the error itself without human intervention.

For the problem of *intentional* manipulation of confidential data, encryption is a common response. Many modern encryption schemes are based on number theory results like the fundamental theorem of arithmetic, which shows up, by the way, in linear error detecting codes. Some of the results are centuries old, and only come into play as important parts of encryption methods as computer technology becomes more powerful.

Computer technology continues to get faster and smaller. The apparent limit is computation at the quantum level, some of which is already being done. Encryption follows suit: the topic of quantum processes used in encryption is covered in chapter 11. The topic follows from the problems and techniques of classical coding in the two previous chapters. The math follows, too.

1.1.4 production accounting

A major role of accounting in practice is making sense of the transformation of raw materials into finished goods. Of particular interest is the efficiency of combining production processes, that is, producing multiple types of output using the same factors of production. Indeed, were there not synergies associated with joint production, it is hard to see why production would not be accomplished by market transactions. That is, firms as organizations for production would not arise.

Information is a prime source of synergies for within firm transactions. As information is so central to production accounting, the discussion occurs later in the book after information concepts and theorems are introduced and in a usable form. A perhaps surprising application is the use of quantum processes and quan-

tum information How quantum units interact and transfer information is very mysterious. Einstein, for example, thought things too "spooky" to be entirely true. However, the exchange of information in a production environment seems not quite so mysterious. Questions arise about how synergy works and the role of information therein. The quantum formalism allows addressing the questions. And allows thinking about how production accounting should look to be helpful in an information rich environment. Or, on the other hand, how the accounting can be corrosive. The double entry mechanism turns out to be helpful in keeping track of synergies, both in the quantum world and the production setting.

1.2 information science

1.2.1 concepts of information

The definition of "information" can change according to the context, but we will try to maintain some precision when using the term. Of the different definitions there are two basic ones we will use most often:

- A specification of joint probabilities.
- Entropy.

Joint probabilities specify how likely it is for two random events to both occur: like a prediction of rain and actual rain. Examination of joint probabilities allows us to make statements about how much information about one random variable is available from observation of another.

Entropy is a single number constructed from lots of probability numbers. The entropy number is larger when the state probabilities are all spread out: a situation characterized as "anything can happen." A large entropy number, then, is associated with high uncertainty.

Entropy is cleverly (many would say brilliantly) designed to be additive in nature. So when the probabilities change, and entropy goes down, something else goes up by the same amount. As entropy is uncertainty, the "something else" can be thought of as information. A casual, but often useful, way to think of information, then, is information is the thing that reduces entropy.

Entropy is a central concept in the manuscript, and is the subject of a central (and longest) chapter (8). The concept connects a variety of accounting and information topics which include

- Assigning (determining) state probabilities.
- Rate of communication of information through noisy channels.
- Rates of economic growth.
- Existence of information synergies in production processes.
- Value of information calculations in various economic settings.

1.2.2 information related theorems

Another way to organize the ideas covered in the manuscript is to make a partial list of the theorems encountered. The names are fairly common for most of them; in any event, the names will serve as identifying labels as the discussion progressions.

- Fundamental theorem of linear algebra
- Fundamental theorem of finance
- Fundamental theorem of arithmetic
- Bayes' theorem
- Heisenberg uncertainty principle
- Shannon's entropy theorem
- Shannon's noisy channel theorem
- Mutual information theorem (fundamental theorem of accounting) and Kelly criterion

("Fundamental" in the name of a theorem is attractive, as it implies the potential to build onto the foundation of the theorem.)

Some of the theorems have proofs of varying degrees of rigor. Where a proof, or something like it, is not included, the examples of the theorem's use should give a pretty good idea how it works. The fundamental theorem of linear algebra, for instance, does not have a complete proof included, but the logic is demonstrated for the special case of a double entry matrix. That's one of the advantages of starting with a study of accounting: accounting is an awfully good example of some deeper mathematical relationships.

The theorems are pretty powerful tools, and can be used to solve some problems that look, and indeed are, quite complicated. The powerful tools reduce the complications to fairly simple arithmetic. It is possible, at least sometimes, to solve the problems by remembering a recipe of instructions, and not worrying with how the theorem works and where it came from. But the recipe approach is quite short-sighted. Like any other tool, it can easily be misunderstood and misused. The world changes and new or different uses of theorems might or might not be appropriate. The student is strongly encouraged to seriously ponder how the theorems are developed.

1.3 chapter topics

The discussion begins, of course, with accounting concepts and questions. In chapter 2 are representations of the accounting double entry system; besides the

standard representation of journal entries and T-accounts, double entry is characterized using linear algebra, and visually with directed graphs.

In chapter 3 an accounting question is confronted: How much of the journal entry numbers reach the financial statements? Equivalently, how much of the journal entries can be reconstructed by looking only at the financial statements? The answer to the question is in the form of a non-negative fraction weakly less than one. This anticipates the concept of a probability, as well as being the first example of a noisy communication channel, i. e., the channel from journal entries to financial statements.

In chapter 4 the linear algebra tools from chapter 3 are used to specify a linear arbitrage free equilibrium concept. Based on the fundamental theorem of finance (or, in non-financial settings, known as the theorem of the separating hyperplane), it is an elegant and powerful concept. Some important tools arise which will prove to be useful later, including the concepts of state probabilities and Arrow-Debreu securities.

Chapter 5 is based on an accounting valuation problem: What number should be written down in the financial statements for derivative securities? The criterion used is the financial statement number should be consistent with arbitrage free pricing. Historical cost, and, unfortunately, even market value criteria are generally inconsistent with arbitrage free equilibria. A consistent approach keeps the equilibrium concept and the state probabilities in the frame.

Chapter 6 is a discussion of stocks and flows in accounting. Interest rates and rates of growth are introduced. Of particular interest for later information applications is the development of e, the base of the natural logarithms, and its role in continuously compounded interest. Long-lived assets and liabilities, and contracts relying on interest concepts such as pensions and leases are covered, but the setting remains one of pure certainty.

Uncertainty is introduced into the stocks and flows discussion in chapter 7. The similarity between Bayesian updating and T-account dynamics is emphasized. A case is illustrated wherein Bayesian revision occurs in the T-account: with judicious choice of the amortization rate, the updated mean of the probability distribution naturally shows up as the amortization expense. Making use of a version of Bayes' theorem, the T-account computations, while quite simple, supply the answer to a fairly complicated probability revision exercise.

Chapter 8, entitled "the fundamental theorem of accounting," is the central information chapter. Fundamental to the theorem is the concept of entropy. Entropy is a measure of information with attractive and convenient additive properties. These properties ease information computations such as (but not limited to) probability assignment and the valuation of information systems.

In order to access entropy theorems, results from previous chapters are used as foundations:

- linear algebra from chapter 3,
- Arrow-Debreu securities and state probabilities from chapter 4,

- continuous rate of growth and e from chapter 6.

Much of the information discussion relies on the Kelly criterion for evaluating investment opportunities. The criterion is the maximization of expected *long run* growth rate. The objective has some appeal as an individual decision criterion: notably, the long run aspect. It also has some social implications, as maximizing long run growth implies significantly lower leverage (borrowing) than other objectives, such as maximizing expected short run returns. Socially, excessive borrowing has liquidity implications, and "too big to fail" consequences.

But the main advantage of the Kelly criterion for our purposes is that it allows access to the mutual information theorem, which, in turn, isolates and simplifies the effects of information on growth rates and firm (or investment) value. Once again, relatively straightforward computations are sufficient to address some fairly complicated looking information problems.

A sufficient condition for the mutual information theorem to go through is a statement about Arrow-Debreu securities. The chapter has illustrations of problems which are not framed in terms of Arrow-Debreu securities, but can be reframed to meet the conditions of the theorem. Hence, the mutual information theorem is more general than might appear at first glance.

There is more about mutual information in chapter 9, as it is central to Shannon's important noisy channel theorem. The theorem states the maximum error-free rate of transmission through a noisy channel is a (relatively simple) function of mutual information.

Some linear error detecting and correcting codes are presented as examples of mechanisms to deal with noisy channels. Understanding of the linear code structure relies on linearity foundations of the accounting double entry system, itself a linear code.

Secret codes are the subject of chapter 10, wherein the concept of mutual information is turned on its head. Effective secret codes are designed so that an eavesdropper viewing the encrypted message receives *no* information about the plain text message.

Modern encryption techniques are a combination of centuries old mathematical foundations and modern technology. The material in chapter 11 on encryption techniques using quantum physics is a chance to consider the foundations of another information science.

Chapter 12 is about information and synergy. The first part of the chapter uses the separation results of the mutual information theorem to evaluate the synergistic effects of business combinations. Accounting implications, such as goodwill valuation, arise.

The latter part of the chapter uses quantum tools from chapter 11 to explore production based information effects. Nature's production function has powerful information processes, offering an opportunity to explore accounting implications when information is a first order effect. One implication is that using accounting for performance evaluation is a delicate, and potentially corrosive, exercise.

1.4 a little linear algebra

A major idea of the manuscript is that serious consideration of accounting issues is a good way to learn lots of stuff. Indeed, it can be argued that the best way to learn about information is to learn some linear algebra; and the best way to learn linear algebra is to study accounting. All the mathematical theorems will be illustrated by, and often derived from, an accounting problem. In that sense, there is not much in the way of math prerequisites, except, of course, a tolerance for abstraction, and a certain eagerness to learn. Nonetheless, it would not hurt to introduce some linear algebra concepts.

1.4.1 vectors

A vector is an ordered set of things; for us the things will (almost) always be numbers. And we will often use a letter to denote the whole set. Such as

$$a = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

$$b = \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix}$$

$$x = \begin{bmatrix} 7 \\ 6 \\ 1 \\ 3 \\ 9 \end{bmatrix}$$

Vectors are typically thought of as columns where the numbers are stacked one on top of the other, as illustrated above. They can be tipped over using an operation called transpose, and denoted with a superscript T.

$$a^T = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}^T = \begin{bmatrix} 1 & 2 & 4 \end{bmatrix}$$

Two vectors with the same number of elements can be multiplied times each other; that is called vector multiplication. The conventional way to accomplish this is to multiply a row vector times a column vector, so we have to be a little careful which one gets transposed. The answer to the multiplication is one number obtained by multiplying the two vectors term for term, and then adding up the products.

$$a^T b = \begin{bmatrix} 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix} = 1\,(5) + 2\,(1) + 4\,(3) = 19$$

Much of the time it won't matter much to us which is the row and which the column, but we will try to be careful so we can communicate with people who have gotten used to the conventions. The first time we encounter vector multiplication will be the accounting task of posting journal entries to T-accounts.

1.4.2 matrices

A matrix is a set of vectors, and all the matrices we will deal with are rectangular, that is, every row (column) has the same number of elements. Typically matrices are denoted with a capital letter, to contrast with lower case vectors.

$$B = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 1 & 2 \\ 7 & 1 & 3 \end{bmatrix}$$

$$D = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}$$

A matrix can be multiplied times a vector using the regular vector multiplication. Since a matrix can have more than one row, there will be more than one answer, and it will be a vector.

$$Ca = \begin{bmatrix} 1 & 1 & 2 \\ 7 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1(1) + 1(2) + 2(4) \\ 7(1) + 1(2) + 3(4) \end{bmatrix} = \begin{bmatrix} 11 \\ 21 \end{bmatrix}$$

Matrices can be multiplied together using the same vector multiplication operation. Each row in the first matrix is vector multiplied by each column in the second matrix, and the vector product is placed in the appropriate place in a result matrix.

$$BD = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 1(3) + 1(0) & 1(0) + 1(4) \\ 2(3) + 1(0) & 2(0) + 1(4) \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 6 & 4 \end{bmatrix}$$

$$\begin{aligned} BC &= \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 7 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1(1) + 1(7) & 1(1) + 1(1) & 1(2) + 1(3) \\ 2(1) + 1(7) & 2(1) + 1(1) & 2(2) + 1(3) \end{bmatrix} \\ &= \begin{bmatrix} 8 & 2 & 5 \\ 9 & 3 & 7 \end{bmatrix} \end{aligned}$$

To be able to multiply matrices together the number of elements in the rows and columns must match up, so we will have to be careful about that.

There's one more linear algebra concept that will come up after a while. For numbers, the number 1 serves as a multiplicative identity. That is, any number

multiplied times 1 leaves the number unchanged. Also, any number, say x, has an inverse $1/x$, such that x times its inverse, $1/x$, returns the multiplicative identity 1. Correspondingly, square matrices have a multiplicative identity; for the 2 by 2 case the identity matrix is

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

And a square matrix, A, might have an inverse, denoted with a -1 superscript, such that multiplying A times its inverse returns the identity.

$$AA^{-1} = I$$

The concept of an inverse will be quite useful, but we won't actually compute inverses, except for simple cases.

That is plenty to get us started in terms of linear algebra. As far as other prerequisites go, we won't require much, except, of course, a curiousity about things scholarly. Surprisingly, not much in the way of accounting experience is required. The manuscript will contain most (I hope nearly all) of the accounting concepts required to proceed. Since the treatment is atypical, accounting experience ranging from novice to quite experienced will be appropriate for the development herein.

1.5 summary

The main idea is that we will treat the study of accounting as an academic discipline. It is hoped our understanding and appreciation of what is meant by "academic discipline" will be enhanced by the journey. Along the way we will encounter old ideas and new: products of the history of the University. A part of scholarship - a very important part - is the joy of discovery, so, at some level, this should be a fun trip.

1.6 exercises

Exercise 1.1 *Multiply some vectors.*

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}^T \textit{times} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}^T \textit{times} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 \\ 0 \\ 7 \\ 8 \\ 7 \\ 9 \end{bmatrix}^T \textit{times} \begin{bmatrix} 2 \\ 7 \\ 1 \\ 8 \\ 2 \\ 8 \end{bmatrix}$$

Exercise 1.2 *Multiply a matrix times a vector.*

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Exercise 1.3 *Multiply a matrix times a matrix.*

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 5 \\ 2 & 3 \end{bmatrix}$$

$$\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{4} \end{bmatrix}$$

2

alternative representations of the double entry system

The first order of business is to explore the logic of the double entry system; in this chapter we present four renderings of the system. The first two, financial statements and journal entries, are well known and available lots of other places. Accordingly, the discussion here goes awfully fast; nonetheless, everything required for a basic understanding of balance sheet, income statement, and journal entries is included, I think. Most of the chapter is devoted to linear algebra and directed graph representations. These are the tools we will rely heavily upon in later chapters. One additional statement, the statement of cash flow, is introduced in the last section.

2.1 financial statements

Start with the definition of an asset which typically has two components.

Definition 2.1 *An asset possesses two properties*

1. *it has future economic benefit, and*

2. *it is owned by the entity under consideration*

The easiest example is cash whose future economic benefit is self-evident. Inventory is an asset - items held for future sale to customers. Receivables are what someone has promised to pay us in the future, also an asset. Plant and equipment are used to produce inventory and otherwise conduct business.

The first financial statement we encounter is the balance sheet. A balance sheet is composed of two lists: a list of assets and a list of sources of assets. Each item on the list has a number attached to it. The number for cash, for example, is the amount of currency and bank deposits held by the firm. The sums of the two lists are the same. That's the balancing part, hence the name.

There are three basic sources of assets

1. Sale of stock (ownership shares)
2. Borrow assets and promise to repay (liabilities)
3. Generate our own assets by conducting business (called retained earnings)

The third source is so important it has its own financial statement: the income statement. The income statement is composed of two parts: revenues (increases to retained earnings) and expenses (decreases to retained earnings).

Definition 2.2 *Revenues are increases to retained earnings*

Definition 2.3 *Expenses are decreases to retained earnings*

Example 2.1 *It is time for a (very) simple example.*

1. *Firm comes into existence by selling ownership share (capital stock) of $100.*
2. *The firm buys some inventory for $50.*
3. *The firm sells all its inventory for $90.*

Construct the balance sheet after each event.

Balance sheet after event 1

Assets		Equities	
cash	100	capital stock	100
inventory	0	retained earnings	0
total assets	100	total equities	100

Balance sheet after event 2

Assets		Equities	
cash	50	capital stock	100
inventory	50	retained earnings	0
total assets	100	total equities	100

Balance sheet after event 3

Assets		Equities	
cash	140	capital stock	100
inventory	0	retained earnings	40
total assets	140	total equities	140

Since buying and selling inventory is conducting business, an income statement appears.

Income statement	
sales revenue	90
cost of goods sold	50
income	40

Cost of goods sold is the expense account associated with the sale of inventory. The change in retained earnings on the balance sheet (it started at zero) is the amount of income for the period.

2.2 another representation - journal entries

Rather than prepare an entire balance sheet after each transaction, a journal entry captures the effect on the two accounts - there are always two; that's the double entry part. Each journal entry has a debit and a credit. Debits are increases to assets and credits are increases to equities. Furthermore, and this is a particularly clever idea originally written down by Luca Pacioli in 1494, debits can also be decreases to equities and credits can be decreases to assets. Debits are written on the left and credits (indented) on the right.

journal entry one: sale of stock

cash	100	
	capital stock	100

journal entry two: purchase of inventory

inventory	50	
	cash	50

journal entry three: sale - revenue part

cash	90	
	sales revenue	90

journal entry four: sale - expense part

cost of goods sold	50	
	inventory	50

Since revenues are increases in retained earnings, they are increased with a credit. Likewise, since expenses decrease retained earnings, they are increased with a debit.

The effects of each of the four journal entries can be accumulated in T-accounts, one for each account. (Debits on the left; credits on the right.) The journal entry

numbers are noted to the left of the amount, and the ending balance in the account is entered below the horizontal line.

cash	
1) 100	2) 50
3) 90	
140	

inventory	
2) 50	4) 50
0	

capital stock
1) 100
100

sales revenue
3) 90
90

cost of goods sold	
4) 50	
50	

The T-accounts calculate the ending amounts (balance) in each of the financial statement accounts, presented below the bottom horizontal line. Retained earnings is calculated by resetting the expense and revenue accounts to zero (called closing entries).

retained earnings	50	
cost of goods sold		50
sales revenue	90	
retained earnings		90

The resulting retained earnings T-account is presented.

retained earnings	
50	90
	40

2.3 a visual representation - directed graph

A directed graph consists of arrows and nodes. For our purposes, the arrows are journal entries and the nodes are accounts. The convention we will follow is the arrowhead is next to the account debited, and the tail of the arrow is next to the credit account. Figure 2.1 presents the four journal entries in directed graph form.

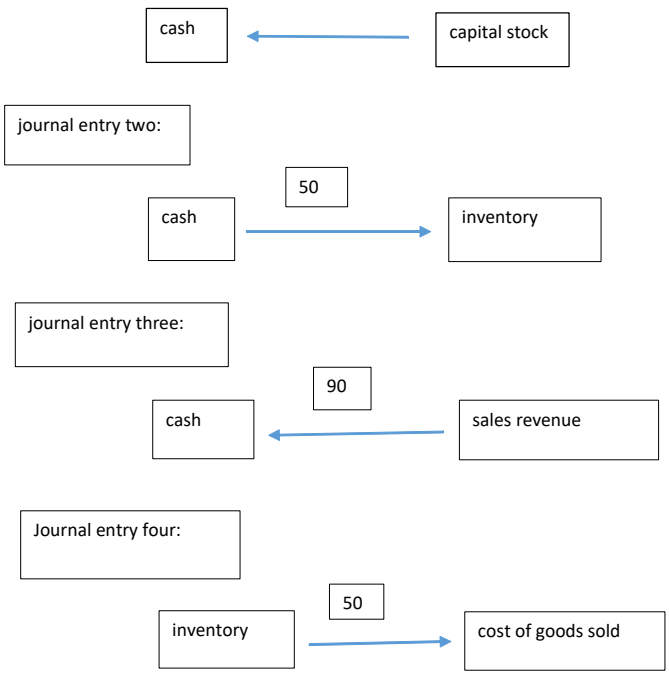


Figure 2.1

Directed graph representation of the four journal entries

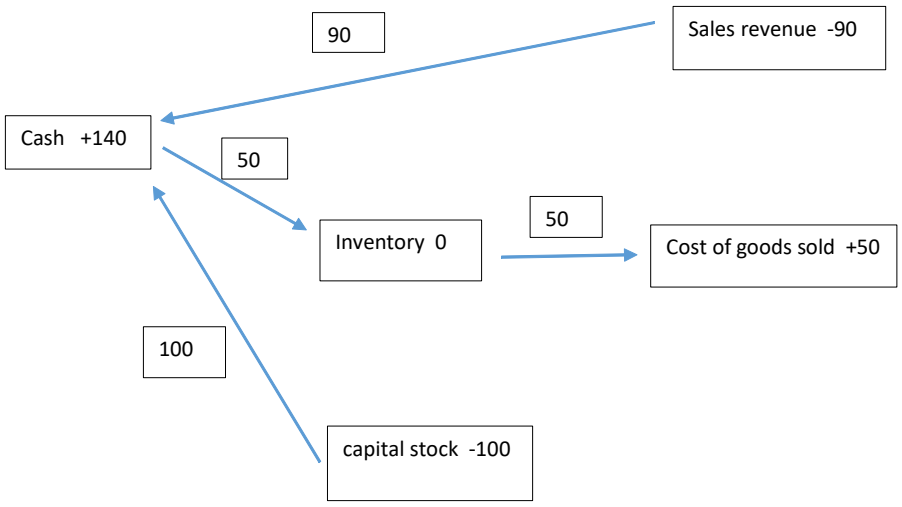


Figure 2.2

Consolidated directed graph with account balances

As the accounts fit together in an organized fashion, it is more instructive to combine the journal entries into one neat package. Figure 2.2 is the same information as in figure 2.1 where each account is only written down once. When more than one journal entry affects a single account, there are arrows for each entry emanating from or entering the account. Also, attached to each account is the ending balance.

To make directed graphs easier to read, the balance sheet accounts are in the middle column, and the income statement accounts are on the right. Further, to leave the income statement accounts in view, they are not closed to retained earnings. Also, notice a credit balance is reported as a negative number. This accomplishes two things. First, it is consistent with the direction of the arrows; an arrow pointing out of the account (credit entry) decreases the balance in the account. Second, negative credit entries allows an easy way to check for the balancing property of double entry: the total balances attached to the nodes sums to zero.

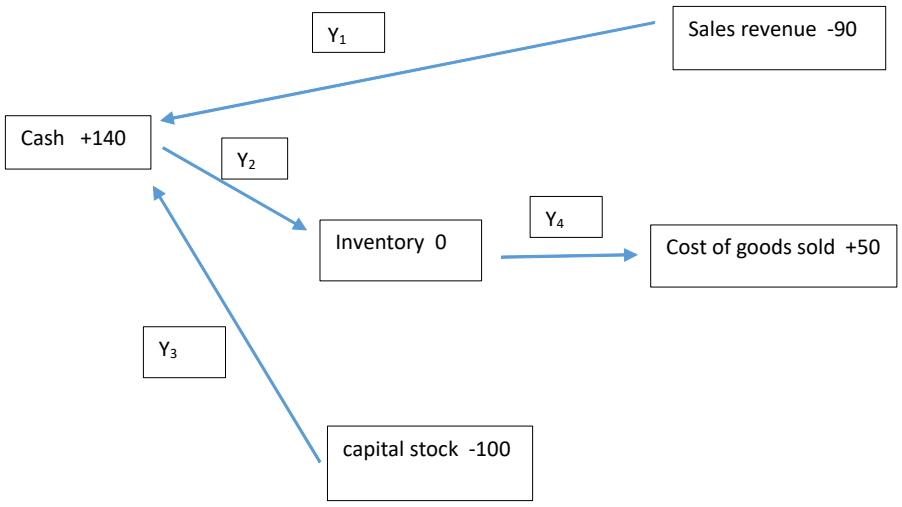


Figure 2.3

Consolidated directed graph with general y_i

2.4 another representation: linear algebra

In order to make things look a little more general, redo the directed graph example with the journal entry amounts in notation where y_i is the amount of journal entry i. See figure 2.3.

Each T-account can now be written as a linear equation.

$$\begin{array}{lccccc}
\text{cash} & +y_1 & -y_2 & +y_3 & & = 140 \\
\text{inventory} & & +y_2 & & -y_4 & = 0 \\
\text{capital stock} & -y_1 & & & & = -100 \\
\text{sales revenue} & & & -y_3 & & = -90 \\
\text{cost of goods sold} & & & & +y_4 & = 50
\end{array}$$

It seems like we should be able to do some stuff with these linear equations, and indeed we can. First, let's get the system in a parsimonious form.

Definition 2.4 *a vector is an ordered set of elements.*

The set of account balances can be written as a vector, x. The convention is to write a vector as a column.

$$x = \begin{bmatrix} 140 \\ 0 \\ -100 \\ -90 \\ 50 \end{bmatrix} = \begin{bmatrix} 140 & 0 & -100 & -90 & 50 \end{bmatrix}^T$$

where the superscript T indicates the transpose operation, converting a row vector into a column, or vice-versa.

Definition 2.5 *a matrix is an ordered set of vectors.*

The interesting matrix here is

$$A = \begin{bmatrix} +1 & -1 & +1 & 0 \\ 0 & +1 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & +1 \end{bmatrix}$$

The matrix A captures the positions of the y's in the system of linear equations and the directed graph. Each row is associated with an account (node in the directed graph): the first row is cash, and so forth. Each column represents a journal entry (arc): the plus one is the account debited and the minus one is for the credit account.

Definition 2.6 *an incidence matrix is a matrix in which every column consists of a plus one and a minus one, and all the other elements are zero.*

The incidence structure of A is what establishes the equivalence of the linear algebra representation with double entry accounting. Now specify one more vector and we can write the whole linear system as a matrix equation.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

The linear algebra representation of the double entry system can now be written as one equation.

$$Ay = x \tag{2.1}$$

Ay multiplies each row in A times the vector y. As the first row in A (cash) is $\begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix}$ the first equation in $Ay = x$ is

$$\begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = y_1 - y_2 + y_3 = 140.$$

140 is the first element (cash) in the vector x.

The multiplication of two vectors is accomplished by multiplying term and then adding up, and by convention, the multiplication is row times column. So the second row of A (inventory) times y is

$$\begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = y_2 - y_4 = 0.$$

$Ay = x$, with incidence matrix A, is a parsimonious representation of the double entry process, and one which follows from the rules of linear algebra. It is equivalent to the directed graph representation which offers a more visual interpretation. We will use both of them, a lot.

2.5 another financial statement

Besides the balance sheet and income statement, there is another financial statement documenting the changes in the cash account: the statement of cash flow. There are two ways to prepare the statement, and each of them can be seen by inspection of the directed graph in figure 1.3.

The direct method simply lists the arrows into and out of the cash node, here there are three.

Statement of cash flow - direct method	
sales revenue (y_3)	90
purchase of inventory (y_2)	-50
sale of stock (y_1)	100
change in cash	140

(The organization and order of the line items are specified somewhat, but we're not worried about that now.)

The alternative - indirect method - is to list the changes in all the other (non-cash) balance nodes, and change the sign. Since the nodes must add to zero, this list will also add to the change in cash. Also, the far right column on the directed graph can be combined into one income number.

Statement of cash flow - indirect method	
income	40
change in inventory	0
change in stock	100
change in cash	140

2.6 summary

Linear algebra and the directed graph are the two representations relied on most in the next, and future, chapters. They provide a effective way to illuminate accounting structure, as well as to connect to, and illuminate, results in other academic disciplines. Applied mathematics is a notable example, but other disciplines will appear, as well. We already got to exploit the directed graph representation to illustrate the two distinct methods of preparing the statement of cash flow.

2.7 exercises

Exercise 2.1 *As of around 2008 (subject, of course, to change) the defined benefit plan pension reporting rules don't require reporting both plan assets and a pension liability (called projected benefit obligation or PBO) on the balance sheet, just the net asset or liability. However, supplementary (footnote) disclosure is required, as well as disclosure of the components of pension cost. The components of pension cost are*

- interest accrued on the PBO
- additional employee service which adds to the PBO
- gains or losses on the PBO due to changes in actuarial estimates or contractual changes in the pension benefits
- gains or losses on pension plan assets

Some of these components do not go directly to the income statement as pension expense. Rather, they are temporarily included in equity accounts called "other comprehensive income" (OCI) and then amortized (moved in portions) to pension expense over a number of years.

Here are a set of typical journal entries affecting the pension accounts and the associated directed graph.

- investment in pension plan assets:

plan assets	y_1	
cash		y_1

- recognition of service and interest costs:

pension expense	y_2	
PBO		y_2

- Predicted or expected return on plan assets reduces pension cost.

plan assets	y_3	
pension expense		y_3

- Any unexpected return on plan assets goes to an other comprehensive income (OCI) account.

plan assets	y_4	
unrecognized gain/loss (OCI)		y_4

- As retiree benefits are paid, both plan assets and PBO decline.

PBO	y_5	
plan assets		y_5

- Changes in actuarial estimates which change the pension liability go to OCI initially.

unrecognized gain/loss (OCI)	y_6	
PBO		y_6

- Changes in pension contracts which change the pension liability go to OCI initially.

unrecognized prior service cost (OCI)	y_7	
PBO		y_7

- Finally, the OCI amounts from previous years are apportioned (amortized) to pension expense

pension expense	y_8	
amortization of unrecognized prior service cost		y_8
pension expense	y_9	
amortization of unrecognized gain/loss		y_9

A directed graph portraying the preceding nine journal entries is in figure 2.4.

Required: construct an incidence matrix representing the pension journal entries.

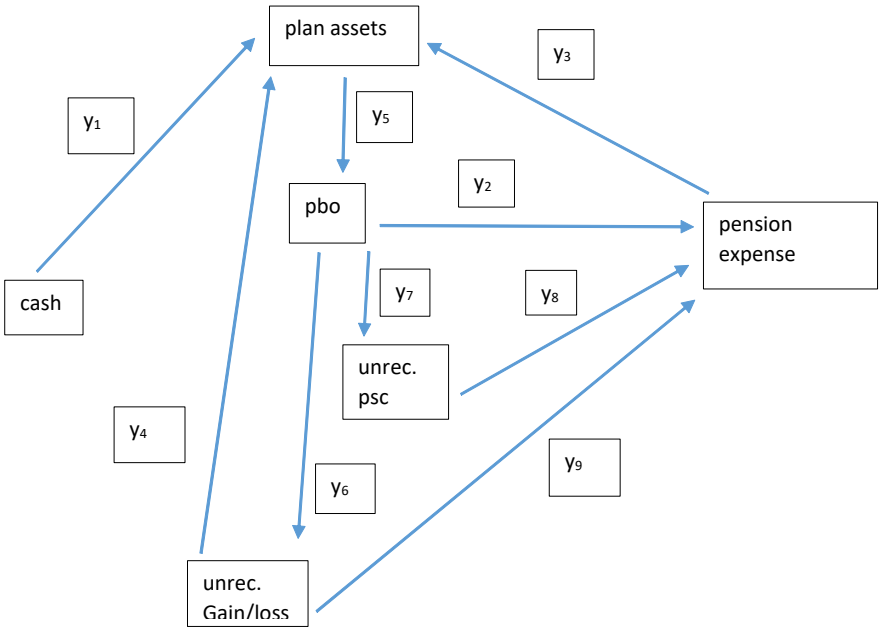


Figure 2.4

Pension activity directed graph

3

accounting as a communication channel

The topic of this chapter is how the double entry system acts as a communication channel. Using the linear algebra representation from chapter 2, that is,

$$Ay = x$$

the central question is easy to state: How much of the information in the vector, y, gets through the double entry matrix, A, to the financial statement vector, x? Once the concept of the row space of a matrix is introduced, the question has a straightforward answer: Only the component of y residing in the row space of A gets through the channel.

Computation of the row component of y can be done a number of ways; the chapter contains five methods. It is not entirely obvious that all methods will always yield the same answer, so some connections are made in that regard.

3.1 the row space of A

The first example will be a simple one: only three journal entries. Cash is paid out for an expense and for an asset. Some of the asset is then amortized (moved to expense).

Example 3.1

expense	4	
cash		4
asset	8	
cash		8

expense	5	
asset		5

If these were the only journal the entries, the resulting financial statements (making use of negative numbers) are

income statement

expense	9

balance sheet

	ending	beginning		ending	beginning
cash	-12	0			
asset	3	0	retained earnings	-9	0
total assets	-9	0	total equities	-9	0

It is useful to access the linear algebra representation for this example.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix}$$

$$A = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

The financial statement vector x is

$$x = \begin{bmatrix} \text{cash} \\ \text{expense} \\ \text{asset} \end{bmatrix} = \begin{bmatrix} -12 \\ 9 \\ 3 \end{bmatrix}$$

And we have the linear algebra representation

$$Ay = x$$

To deal with the question of how much of y gets through A to x, the concept of orthogonality is important.

Definition 3.1 *Two vectors are orthogonal if their vector product is zero.*

Visually, two vectors are orthogonal if they are perpendicular or at right angles. In the (two dimensional) plane, vectors $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$ and $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$ are obviously at right angles, and, just as obviously, the vector product is zero. Another example is $\begin{bmatrix} 1 & 1 \end{bmatrix}^T$ and $\begin{bmatrix} 1 & -1 \end{bmatrix}^T$. The algebraic orthogonality condition for vectors is true in any dimensional space, not just the plane. That's what allows us to think about vectors at right angles even in high dimensions.

Back to the example. Without worrying about where the numbers come from for now, notice y can be decomposed into the following two orthogonal components.

$$y = \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix} = \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix} - 3 \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

It is convenient to have names for the two components. Let them be y_{row} and y_N. The reasons for these particular names will become apparent momentarily.

$$y_{\text{row}} = \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix} \qquad y_N = 3 \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

It is easy to check for orthogonality.

$$y_N^T y_{\text{row}} = 3 \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix} = 3\,(7 - 5 - 2) = 0$$

Vector orthogonality is an important idea, but perhaps more important is the concept of orthogonality of vector spaces. Here we are interested in two vector spaces, and both of them arise from the matrix A: the row space of A, and the null space of A.

Definition 3.2 *The row space of A consists of all the vectors which are linear combinations of the rows of A.*

y_{row} is one vector (among many) in the rows of A, hence the name.

$$y_{\text{row}} = \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix} = -5 \begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

The first vector on the right hand side is the first row of A, and the second is the second row of A. (There are many other ways to form y_{row} from the rows of A.)

Another important space of the matrix A is the null space.

Definition 3.3 *The null space of A consists of all the vectors y_N satisfying $Ay_N = 0$.*

Here we have

$$y_N = 3 \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

So it is easy to check that y_N is in the null space.

$$Ay_N = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} 3 = 3 \begin{bmatrix} -1+1 \\ 1-1 \\ -1+1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

So far we have decomposed y into a row and null component. We have not, as yet, actually computed the components: that is done in succeeding sections of the chapter. But once we have the orthogonal decomposition, we can answer the question about the communication channel. As it turns out, y_{row} goes through the channel, y_N does not.

To see that only y_{row} gets through, suppose all we have is the financial statement vector, x. Then y_{row} is unique, because there is only one solution to

$$\begin{aligned} Ay &= x \\ y^T y_N &= 0 \end{aligned}$$

and it is y_{row}. For our example we have three independent equations and three unknowns. There are two independent T-account equations (the third one is not independent because of double entry), and the orthogonality equation to solve for the three elements of y.

On the other hand, x tells us nothing about y_N. Since $Ay_N = 0$, y_N disappears from $Ay = x$.

$$Ay = A\left(y_{\text{row}} + y_N\right) = Ay_{\text{row}} + Ay_N = Ay_{\text{row}} = x$$

Given x, y_N can be anything in the null space of A.

So the financial statements tell us everything about y_{row} and nothing about y_N. That is the sense in which the row component is all that gets through the channel. And "row component through the channel" will be a recurring theme in later chapters.

One more thing: we can specify *how much* gets through. Because y_{row} and y_N are orthogonal (perpendicular), they obey the theorem of Pythagoras: the sum of the square of the two sides of a right triangle equals the square of the hypotenuse. The "square" of a vector is the vector product of the vector with itself.

$$\begin{aligned} y_{\text{row}}^T y_{\text{row}} &= \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix}^T \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix} = 7^2 + 5^2 + 2^2 = 78 \\ y_N^T y_N &= \begin{bmatrix} -3 \\ 3 \\ 3 \end{bmatrix}^T \begin{bmatrix} -3 \\ 3 \\ 3 \end{bmatrix} = 3^2 + 3^2 + 3^2 = 27 \\ y^T y &= \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix}^T \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix} = 4^2 + 8^2 + 5^2 = 105 \end{aligned}$$

Note the Pythagorean result:

$$\begin{aligned} y_{\text{row}}^T y_{\text{row}} + y_N^T y_N &= y^T y \\ 78 + 27 &= 105 \end{aligned}$$

The fraction getting through, then, is

$$\frac{y_{\text{row}}^T y_{\text{row}}}{y^T y} = \frac{78}{105} \approx .75$$

In this section we derived the main idea of this, and some future, chapters: only the row component gets through. We did not, however, actually compute the row component, y_{row}^T, or the null component, y_N. They were simply stated, and their properties noted and checked. The next several sections of the chapter offers a variety of ways to compute the components. The methods we will use most often are those that solve for the null component first.

3.2 expanded setup

We'll use a slightly expanded example to help demonstrate a number of methods for computing y_{row} from a given set of financial statements.

Example 3.2

balance sheet

	ending	*beginning*		*ending*	*beginning*
cash	*40*	*20*	*accrued liabilities*	*45*	*30*
acc'ts rec.	*35*	*30*	*capital stock*	*40*	*40*
inventory	*50*	*30*	*retained earnings*	*40*	*10*
total assets	*125*	*80*	*total equities*	*125*	*80*

income statement

sales	*120*
cost of goods sold	*60*
gen'l & admin. expenses	*30*
income	*30*

The journal entries (except for closing entries) are presented absent amounts.

accounts receivable	y_1	
sales		y_1
cash	y_2	
accounts receivables		y_2
accrued liabilities	y_3	
cash		y_3
g & a expense	y_4	
accounts receivable		y_4

g & a expense	y_5	
accrued liabilities		y_5
inventory	y_6	
accrued liabilities		y_6
cost of goods sold	y_7	
inventory		y_7

Some of the journal entries might deserve comment.

- y_1: Sales are often made on account, and the cash will be collected later. Accounts receivable is the resulting asset.
- y_2: Cash is collected for accounts receivable.
- y_4: On occasion not all accounts will turn out to be collectible; firms sometimes go bankrupt or otherwise disappear. In that case the receivable asset is reduced, the offsetting debit is often to an income statement account called something like bad debt expense. Here the debit is included in general and administrative expenses.
- y_5 and y_6: Things like labor, raw materials, and supplies are often acquired on account; a liability is thereby created.

Alternative representations for the financial statements are the directed graph and the incidence matrix. The directed graph is in figure 3.1, with cash on the left, balance sheet accounts in the middle column, and income statement accounts on the right. Attached to the account nodes are the (changes in) account balances.

The 6×7 incidence matrix, A, is presented below, along with the account balance vector, x.

x		y_1	y_2	y_3	y_4	y_5	y_6	y_7
20	cash	0	1	-1	0	0	0	0
5	acc'ts rec.	1	-1	0	-1	0	0	0
20	inventory	0	0	0	0	0	1	-1
-15	acc. liab.	0	0	1	0	-1	-1	0
-120	sales	-1	0	0	0	0	0	0
60	cgs	0	0	0	0	0	0	1
30	g&a	0	0	0	1	1	0	0

Our task is to solve for the transaction amounts, y_1 through y_7 which satisfy $Ay = x$. Some of the y values are straightforward. For example, from the directed graph it is immediate that $y_1 = 120$, $y_6 = 80$, and $y_7 = 60$. Those arrow values are the only way to achieve the node balances in sales, inventory, and cgs. The others, however, do not have unique answers.

Definition 3.4 *A loop exists in a directed graph if it is possible to travel from one node, touch others, and return to the original node without traveling on an*

arrow more than once. It is all right to travel backward on the arrow; that is, the debit/credit direction of the arrow does not matter when determining a loop.

Arrows y_2, y_3, y_4, and y_5 constitute a loop in the directed graph. Some possible solutions are

y_2	115	0	105
y_3	95	-20	85
y_4	0	115	10
y_5	30	-85	20

(Check to see these alternate solutions all result in the appropriate node balances.)

We will calculate a unique answer to the problem, $Ay = x$, generating a y vector with certain, perhaps desirable properties. Some of the properties of the answer are

- It is the "shortest" possible y vector which generates the given statements, that is, which solves $Ay = x$.
- It resides entirely within the rows of A, and, indeed, is the only solution to $Ay = x$ which does so. That is, it can be constructed by taking a linear combination of the rows of A. This "residing in the rows" property supplies a convenient name for the vector: y_{row}.
- It is, in the sense discussed previously, the only component of y which gets through the double entry channel. In other words, y_{row} contains all the information available in the financial statements about the transaction amounts.

We cover several methods; the first solves directly for the shortest y vector.

3.3 quadratic programming

The idea is to find the shortest vector satisfying $Ay = x$. Length is defined as the sum of the squared elements of y; squaring eliminates the problem of how negative elements affect the length. It is simple enough to state the problem, and, indeed, it is a simple matter for a computer to solve the problem so stated. So, for the first time through, we will let a computer do the work.

$$\begin{aligned} minimize\ y^T y &= \sum y_i^2 \\ subject\ to\ Ay &= x \end{aligned}$$

$Ay = x$ is a system of seven linear equations - one for each account - with seven unknowns - one for each transaction. Because of the balancing property (or, equivalently, because A is an incidence matrix), there are only six *independent*

equations, and, hence, rather than a unique solution to the system, there are several possible solutions.

Most spreadsheets have an optimizer capable of solving the problem; Excel has a pretty good one called Solver. It doesn't take long to type in the data. If we use the Excel function *sumproduct* to accomplish vector multiplication, we don't have to type in the zeros in the A matrix.

Whatever computer routine we use, the solution appears, called y_{row} .

$$y_{row} = \begin{bmatrix} 120 & 55 & 35 & 60 & -30 & 80 & 60 \end{bmatrix}^T$$

There are a couple of things to notice about the solution.

- For the transactions not in a loop (y_1, y_6, and y_7), the solution is consistent with what is apparent from the directed graph: $y_1 = 120$, $y_6 = 80$, and $y_7 = 60$.
- There is a negative element: $y_5 = -30$. Since we set up the transactions to flow through the accounting system in basically one direction, the existence, in our preferred solution, of a transaction going in the other direction might cause some discomfort. There is no need to stifle the discomfort at this stage.

Example 3.3 *Reconsider the set-up from example 3.1 to compute the y_{row} we already know. Use a computer optimizer like Solver in Excel Recall*

$$x = \begin{bmatrix} -12 \\ 9 \\ 3 \end{bmatrix}$$

$$A = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

It is a fairly simple matter for a computer optimizer like Solver to compute y_{row} using a quadratic program. See exercise 3.9 for a pencil and paper approach.

3.4 regression

3.4.1 computing y_{row} with regression

The first method minimized the sum of the squares of the y vector. Regression also minimizes the sum of squares; this method exploits the connection, and gets us closer to a paper and pencil method to solve the problem. The first step is to find any solution to $Ay = x$; call it y_p for y particular. Then project y_p into the rows of A, that is, run a regression.

Spanning trees are useful for a variety of things, one of which is to find a particular solution, y_p.

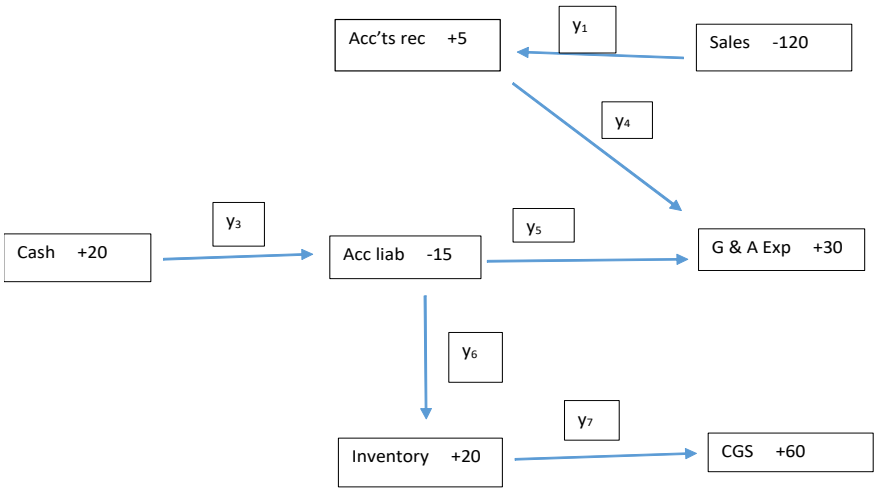


Figure 3.2

Example 3.2 Spanning tree with $y_2 = 0$

Definition 3.5 *A spanning tree is a directed graph with two properties.*

- It spans. That is, every node can be reached from every other node by tracing a path on the arrows. (It's okay to go backward on an arrow.)

- It is a tree. That is, the arrows don't loop. (Most, I think all, trees in nature have this property.)

To form a spanning tree from a directed graph containing loops, erase an arrow from each loop, and set the erased $y_i = 0$. Figure 3.2 contains a spanning tree for example 3.1 with $y_2 = 0$.

It's easy to compute the y vector from a spanning tree: add up the amounts of the node on one end of the arrow (if it's the tail, change the sign of the sum). Since there are no loops, there is no ambiguity about which end of the arrow a node is connected to. For this spanning tree

$$y_p = \begin{bmatrix} 120 & 0 & -20 & 115 & -85 & 80 & 60 \end{bmatrix}^T$$

Now run a regression; Excel is pretty good at this, as well. The dependent variable is y_p and the independent variables (regressors) are the rows of A.

dependent variable	independent variables						omit one column
120	0	1	0	0	-1	0	0
0	1	-1	0	0	0	0	0
-20	-1	0	0	1	0	0	0
115	0	-1	0	0	0	0	1
-85	0	0	0	-1	0	0	1
80	0	0	1	-1	0	0	0
60	0	0	-1	0	0	1	0

The independent variables are recognized as the rows of A in columns: A^T. It is important to eliminate one of the columns before running the regression: Excel complains when the independent variables are not independent of each other. It doesn't matter which column is eliminated; y_{row} is the same, and that should be checked.

Here is (part of) a sample regression output when the last column is eliminated.

coefficients on independent variables	predicted variable	residuals
-5	120	0
-60	55	-55
110	35	-55
30	60	55
-180	-30	-55
170	80	0
	60	0

The predicted variable is seen to be y_{row}. The coefficients are the weights on the rows of A used to generate y_{row}. The first element of y_{row}, for example, is the coefficient vector times the first row of A^T (first column of A omitting the last element).

$$\begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} -5 \\ -60 \\ 110 \\ 30 \\ -180 \\ 170 \end{bmatrix} = -60 - (-180) = 120$$

The other elements of y_{row} are generated in a similar fashion. As y_{row} is constructed as a weighted combination of the rows of A, it is said that y_{row} resides in the rows of A, hence the name y_{row}.

Before leaving the regression output notice the residual vector - this is the difference between the predicted variable, y_{row}, and the dependent variable, y_p. The residual vector actually looks pretty simple, and, indeed, a direct calculation of the residuals is the basis of the next method. But before proceeding to the next method, there is more to say about regressions.

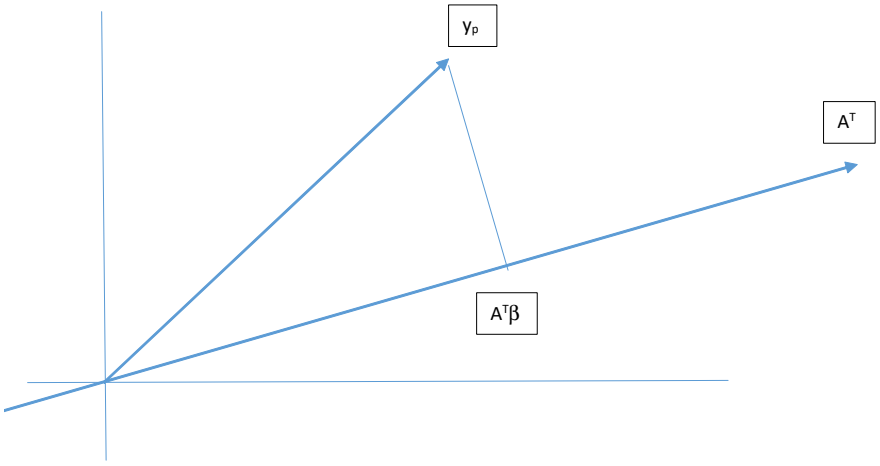


Figure 3.3

Choose β $y_p - A^T\beta$ is minimized

3.4.2 *some more about regressions*

Geometrically, the regression is finding the closest point in the rows of A (in this case, the columns of A^T) to y_p. Roughly speaking, the way to do that is construct a line from y_p to A^T that is perpendicular (orthogonal) to A^T. Geometrically, perpendicular and orthogonal have the same meaning; two lines are perpendicular if they form a right angle. The geometric idea is in figure 3.3.

It's the same idea as the shortest way to get from the interior of a room to a wall is to walk perpendicular to the wall. That simple idea is enough to allow us to write down the basic equation of regression, called the orthogonality conditions. All the regression calculations follow from these conditions.

In Figure 3.3 β is the vector of coefficients on the columns of A^T (rows of A). The regression routine chooses the coefficients so as to make the difference vector, $y_p - A^T\beta$ orthogonal to all the columns of A^T. This requires an algebraic interpretation of orthogonality in terms of vectors: we already know that two vectors are orthogonal if their vector product is zero.

For the regression problem, the vectors we are interested in setting orthogonal to each other are the difference vector $y_p - A^T\beta$ and any vector in A^T, itself. The orthogonality condition is so important, we write it as a theorem.

Theorem 3.1 *The coefficient vector β which minimizes the squared distance from vector y_p to the rows of matrix A solves the orthogonality condition: $A(y_p - A^T\beta) = 0$.*

Once the orthogonality conditions are written down, we can do the regression calculations just like the computer routine does. To solve for β rearrange the orthogonality condition.

$$AA^T\beta = Ay_p$$

This is a parsimonious representation of seven (six independent) equations with seven unknowns.

$$Ay_p = x = \begin{bmatrix} 20 \\ 5 \\ 20 \\ -15 \\ -120 \\ 60 \\ 30 \end{bmatrix}$$

(As Ay_p is equal to x for any y_p, we can see the orthogonality condition is the same for any choice of y_p. In other words, it doesn't matter which particular solution we choose.)

Since

$$A^T = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 2 & -1 & 0 & -1 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 2 & -1 & 0 & -1 & 0 \\ -1 & 0 & -1 & 3 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 2 \end{bmatrix}$$

The matrix equation, $AA^T\beta = Ay_p = x$, can be written as 7 linear equations.

$$\begin{aligned} 2\beta_1 - \beta_2 - \beta_4 &= 20 \\ -\beta_1 + 3\beta_2 - \beta_5 - \beta_7 &= 5 \\ 2\beta_3 - \beta_4 - \beta_6 &= 20 \\ -\beta_1 - \beta_3 + 3\beta_4 - \beta_7 &= -15 \\ -\beta_2 + \beta_5 &= -120 \\ -\beta_3 + \beta_6 &= 60 \\ -\beta_2 - \beta_4 + 2\beta_7 &= 30 \end{aligned}$$

It is not beyond our wit to solve the system, but it is tedious enough so we are grateful for computer aid. But we can verify the regression output solves the system. That is, the row coefficients are

$$\beta = \begin{bmatrix} -5 & -60 & 110 & 30 & -180 & 170 \end{bmatrix}^T$$

with $\beta_7 = 0$ for the omitted row. Our next task is to put the problem in a form that is paper and pencil doable. Before doing so, however, let's do the projection exercise on the simpler problem from example 3.1

Example 3.4 *As there are only two independent rows in A, we can use the first two, and redefine A accordingly.*

$$A = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

And we can use

$$y_p = \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix}$$

So

$$Ay_p = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix} = \begin{bmatrix} -12 \\ 9 \end{bmatrix}$$

(Any y_p satisfying $Ay = x$ will yield the elements of x.)

$$AA^T = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

The orthogonality conditions can be written:

$$\begin{aligned} AA^T\beta &= Ay_p \\ \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\beta &= \begin{bmatrix} -12 \\ 9 \end{bmatrix} \end{aligned}$$

The two orthogonality equations written separately:

$$\begin{aligned} 2\beta_1 - \beta_2 &= -12 \\ -\beta_1 + 2\beta_2 &= 9 \end{aligned}$$

Solving two linear equations in two unknowns is not difficult. Here the solution is

$$\begin{aligned} \beta_1 &= -5 \\ \beta_2 &= 2 \end{aligned}$$

So

$$y_{\text{row}} = -5 \begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix}$$

And, we knew that.

3.5 projection into the nullspace

The method of this section can often be accomplished with a paper and pencil. The idea is to project a particular solution to $Ay = x$ into what is called the nullspace, effectively calculating the residual vector $y_p - y_{row}$ from the previous regression. Notice from the regression output the residual vector looks a bit simpler than y_{row}. What we will be doing is to divide any y_p solution into two orthogonal parts, y_{row} and the null component, denoted y_N.

$$\begin{aligned} y_p &= y_{\text{row}} + y_N \\ \text{where } (y_{\text{row}})^T y_N &= 0 \end{aligned}$$

Recall the nullspace of a matrix A consists of all the vectors that are orthogonal to the rows of A.

Typically, calculating the nullspace of a matrix is a little bit complicated, but for an incidence matrix, there is really nothing to it: simply read off the loops from the associated directed graph. Recall from figure 3.1, the loop consists of y_2, y_3, $-y_5$, and y_4, so the null vector is

$$N = \begin{bmatrix} 0 & 1 & 1 & -1 & 1 & 0 & 0 \end{bmatrix}^T$$

, simply place a one in the position of the arrow in the loop, zero for arrows not in the loop. The direction around the loop is important, so if the arrow is traversed from head to tail, as y_4 is in the example, the sign is negative.

It is a remarkable property of incidence matrices that the nullspace consists of vectors with all positive and negative ones and zeros. Also, it doesn't matter which way the loop is traversed, a counter-clockwise null vector of

$$\begin{bmatrix} 0 & -1 & -1 & 1 & -1 & 0 & 0 \end{bmatrix}^T$$

would work just as well, as can be verified as the procedure unfolds.

It is easy to verify that the loop supplies the null vector, since $AN = 0$.

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & -1 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

When projecting y_p into N, the result is y_N which we shall calculate as N times a regression coefficient.

$$y_N = N\beta$$

The geometric picture is depicted in figure 3.4. The difference vector $y_p - y_N$ is orthogonal to the nullspace, hence it is in the row space, so $y_{row} = y_p - y_N$.

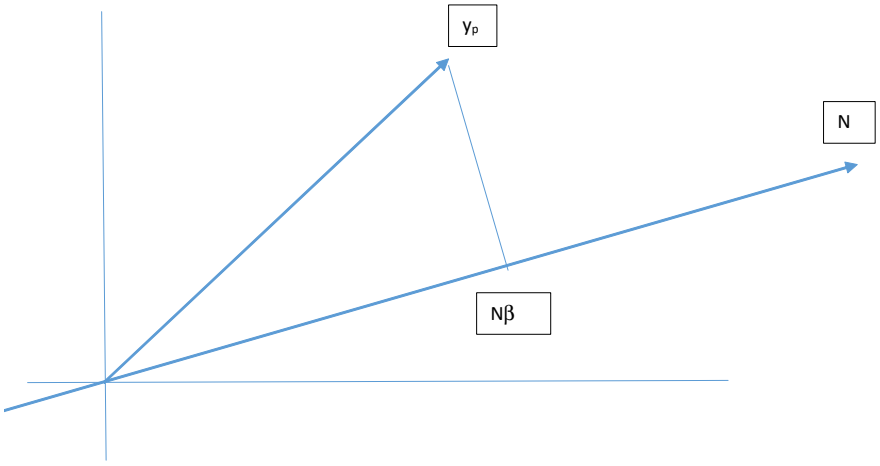


Figure 3.4

Projecting y_p into the nullspace

To calculate the regression coefficient β, use the same y_p as before (any solution to $Ay = x$ will do).

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
y_p	120	0	-20	115	-85	80	60
N^T	0	1	1	-1	1	0	0

The orthogonality condition requires the difference vector, $y_p - y_N$ to be orthogonal to the N vector, so the orthogonality condition is the same as before with A replaced by N as in figure 3.4.

$$\begin{aligned} & N^T (y_p - y_N) \\ = \quad & N^T (y_p - N\beta) = 0 \end{aligned}$$

$$N^T N\beta = N^T y_p$$

$N^T N$ and $N^T y_p$ are computed as vector products.

$$\begin{aligned} N^T N &= 4 \\ N^T y_p &= -20 - 115 - 85 = -220 \end{aligned}$$

Hence,

$$\begin{aligned} 4\beta &= -220 \\ \beta &= -55 \end{aligned}$$

Completing the table.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
y_p	120	0	-20	115	-85	80	60
N	0	1	1	-1	1	0	0
$N\beta$	0	-55	-55	55	-55	0	0
$y_{row} = y_p - N\beta$	120	55	35	60	-30	80	60

y_{row} from this paper and pencil calculation is the same as from the previous methods.

Example 3.5 *For a simpler example reconsider once again the set-up in example 3.1 We have*

$$N = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

which we can derive by following around the loop in the directed graph representation in figure 3.5. We can use

$$y_p = \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix}$$

So we have

$$N^T N = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} = 3$$

$$N^T y_p = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix} = 4 - 8 - 5 = -9$$

So the orthogonality conditions are

$$\begin{aligned} N^T N\beta &= N^T y_p \\ 3\beta &= -9 \\ \beta &= -3 \end{aligned}$$

And

$$\begin{aligned} y_N &= N\beta = -3 \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \\ y_{\text{row}} &= y_p - y_N = ^T \begin{bmatrix} 4 \\ 8 \\ 5 \end{bmatrix} - \begin{bmatrix} -3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix} \end{aligned}$$

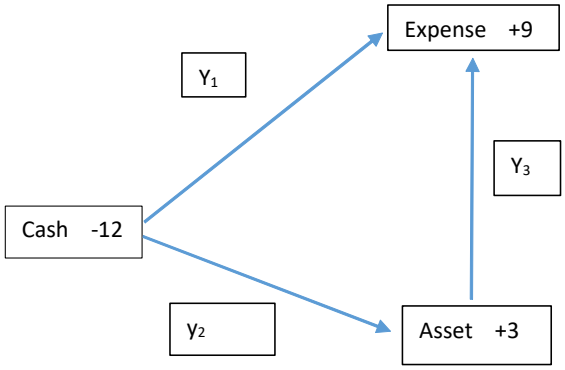


Figure 3.5

Directed graph for example 3.4

3.6 average spanning tree

Calculating the average spanning tree is another method, one which offers a paper and pencil solution for relatively small problems. That is, derive a y_p for every possible spanning tree, add up, and divide by the number of spanning trees. For the ongoing example there are four spanning trees. The spanning tree with $y_2 = 0$, call it spanning tree 1, was presented in Figure 3.2. The other three spanning trees, omitting y_3, y_4, and y_5 in turn, are formed in similar fashion, and the resulting solutions are presented in the following table.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
spanning tree 1	120	0	−20	115	−85	80	60
spanning tree 2	120	20	0	95	−65	80	60
spanning tree 3	120	115	95	0	30	80	60
spanning tree 4	120	85	65	30	0	80	60
sum	480	220	140	240	−120	320	240
$y_{row} = sum/4$	120	55	35	60	−30	80	60

And it is seen that the average spanning tree equals y_{row} from the earlier methods.

For even slightly larger problems, it is not always easy to confidently list all the spanning trees. There is a quite remarkable theorem, called the matrix tree theorem, which provides a simple and useful expression for the number of spanning trees.

Theorem 3.2 *For any directed graph the number of spanning trees is given by the determinant of $N^T N$, where N is the matrix with the nullspace vectors in the columns.*[1]

For the example the single nullspace vector is $\begin{bmatrix} 0 & 1 & 1 & -1 & 1 & 0 & 0 \end{bmatrix}^T$ so $N^T N$ is 4, which is also the determinant. The determinant of a two by two matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $ad - bc$. The formula for the determinant of larger matrices gets a little bit complicated; in any event, spreadsheets can calculate them numerically.

Example 3.6 *Referring to figure 3.5, the three spanning trees for the simpler example are*

$$y = \begin{bmatrix} 0 \\ 12 \\ 9 \end{bmatrix}, \begin{bmatrix} 12 \\ 0 \\ -3 \end{bmatrix}, \text{ and } \begin{bmatrix} 9 \\ 3 \\ 0 \end{bmatrix}.$$

The average spanning tree amounts are

$$\frac{1}{3} \begin{matrix} 0+12+9 \\ 12+0+3 \\ 9-3+0 \end{matrix} = \begin{bmatrix} 7 \\ 5 \\ 3 \end{bmatrix}$$

3.7 augmented A matrix

Finally, another method is to add enough equations to $Ay = x$ so the system has a solution, and, further, the solution is y_{row}. Here we can use the extra equation(s) implied by the nullspace relationships. We have seen that y_{row} is orthogonal to the nullspace. In the example the nullspace consisted of only one vector $\begin{bmatrix} 0 & 1 & 1 & -1 & 1 & 0 & 0 \end{bmatrix}^T$ The extra equation is the vector product of the nullspace vector with y is equal to zero which is the last row of augmented A and the last element of augmented x.

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & -1 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 20 \\ 5 \\ 20 \\ -15 \\ -120 \\ 30 \\ 60 \\ 0 \end{bmatrix}$$

[1] The matrix tree theorem is typically stated and proved in terms of the incidence matrix which eliminates confusion when there are no loops. See, for example, Harris, Hirst, and Mossinghoff, page 29.

There are now seven independent linear equations which can be used to solve for the seven elements of y_{row}, as the null row adds another independent equation. (Although there are 8 equations, only 7 are independent, as any particular T-account can be constructed from the others.) Solving seven linear equations is a little bit tedious, but it is easy to verify (or use a spreadsheet to find) that the system is solved by

$$y_{\text{row}} = \begin{bmatrix} 120 & 55 & 35 & 60 & -30 & 80 & 60 \end{bmatrix}^T$$

And we have five different methods to find y_{row}, but there are still some things to learn and connections to make, one of which is the fundamental theorem of linear algebra.

Example 3.7 *For the simpler example the augmented equations are*

$$\begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -12 \\ 9 \\ 3 \\ 0 \end{bmatrix}$$

The unique solution to all 4 equations is

$$y_{\text{row}} = \begin{bmatrix} 7 \\ 5 \\ 2 \end{bmatrix}$$

3.8 the fundamental theorem of linear algebra

As far as the ongoing numerical example is concerned, all the foregoing methods for finding y_{row} arrive at the same answer. It is, indeed, true for all examples. However, the example approach does not demonstrate the equivalence of all the methods for all possible problems. In this section we show the equivalence for two of the methods: quadratic programming and regression. The demonstration uses the fundamental theorem of linear algebra.

For the example we have been doing, the row space of A (where A has m rows and n columns) is not "complete" in the sense that not all vectors with n elements (where n is the number of journal entries) can be formed using weighted combinations of the rows. For example, the chosen y_p's could not be so formed, because a regression had a residual vector $y_p - A^T\beta$. So, in cases like this, it is said the row space does not *span* the set, or *space*, of n element vectors.

According to the fundamental theorem of linear algebra the nullspace completes the row space, and they are said to be *complements*. That is, the combination of the two spaces spans the space of n element vectors. Furthermore, as all the vectors in the nullspace are orthogonal to all the vectors in the row space, they are said to be orthogonal complements. All matrices have the same property as stated in the theorem.

Theorem 3.3 *The row space and nullspace of any matrix are orthogonal complements.*[2]

The theorem allows any vector composed of n elements to be decomposed into orthogonal components, and that's often an instructive thing to do. Any solution to $Ay = x$ can be written as $y = y_{\text{row}} + Nk$, where k is a vector of weights on the nullvector(s) N. The decomposition is always possible by the fundamental theorem. Now we can revisit the first method to see that y_{row} is always a solution to the quadratic program.

$$\begin{aligned} &\min\ y^T y \\ &s.t.\ Ay = x \end{aligned}$$

Substitute in the objective function.

$$\begin{aligned} y^T y &= (y_{\text{row}} + Nk)^T (y_{\text{row}} + Nk) \\ &= y_{\text{row}}^T y_{\text{row}} + 2 (Nk)^T y_{\text{row}} + (Nk)^T (Nk) \end{aligned}$$

The crossproduct term $(Nk)^T y_{\text{row}}$ is zero by orthogonality. Hence, finding the minimum $y^T y$ is equivalent to finding the k vector that minimizes

$$y_{\text{row}}^T y_{\text{row}} + (Nk)^T (Nk)$$

Choosing k equal to zero obviously does the trick. Hence, we have shown y_{row}, derived as the solution to the projection problem, is also the solution to the original (method 1) quadratic programming problem.

A relative of the fundamental theorem, called Euler's theorem, is useful when looking for the loops in a directed graph. The number of loops is equivalent to the number of independent vectors in the null space. When the graph is extensive it is not always easy to find the independent loops; Euler's theorem tells us how many loops we are looking for.

Theorem 3.4 *Euler's theorem says the number of independent loops in a directed graph is equal to the number of arcs (journal entries) minus the number of nodes (accounts) plus one.*

From the fundamental theorem the total number of independent vectors in the row space of A plus the nullspace is n, where n is the number of columns (and journal entries).[3] The number of independent vectors in the row space is one

[2] See Strang, page 138. The theorem is also true for the transpose of a matrix; the orthogonal complements are called the column space and the left nullspace.

[3] The number of independent vectors in the row space is called the "rank" of the matrix. It is kind of remarkable that the rank is the same whether looking at the matrix or its transpose. That is, the number of independent vectors in the row space is equal to the number of independent vectors in the column space.

less than the number of rows, m, as double entry implies one T-account must be determinable from all the others. Since the sum of the loops plus the sum of independent row vectors is n, we have

$$\begin{aligned} \text{\# of loops } + (m-1) &= n \\ \text{\# of loops} &= n - m + 1 \\ &= \text{\# of journal entries - \# of accounts +1} \end{aligned}$$

3.9 multiple loops

When there is more than one loop, calculation of y_{row} proceeds in the same way. To project into the nullspace, the orthogonality conditions are used. However, instead of just one equation, when there are two loops, there are two equations to solve for two regression coefficients.

Here is an example, the form of which will reappear in later sections. The (partial) financial statements have three assets and two expense accounts. The assets could be various prepayments or equipment, and the two expenses could be general expenses and cost of goods sold. Cash is another asset, so the example has 6 total accounts.

Example 3.8 *Here are partial financial statements.*

partial balance sheet

	ending balance	*beginning balance*
asset 1	*10*	*10*
asset 2	*15*	*15*
asset 3	*8*	*5*

partial income statement

expense 1	*6*
expense 2	*3*

Here are the only journal entries affecting the above accounts.

asset 1	y_1	
cash		y_1
asset 2	y_2	
cash		y_2
asset 3	y_3	
cash		y_3
expense 1	y_4	
asset 1		y_4

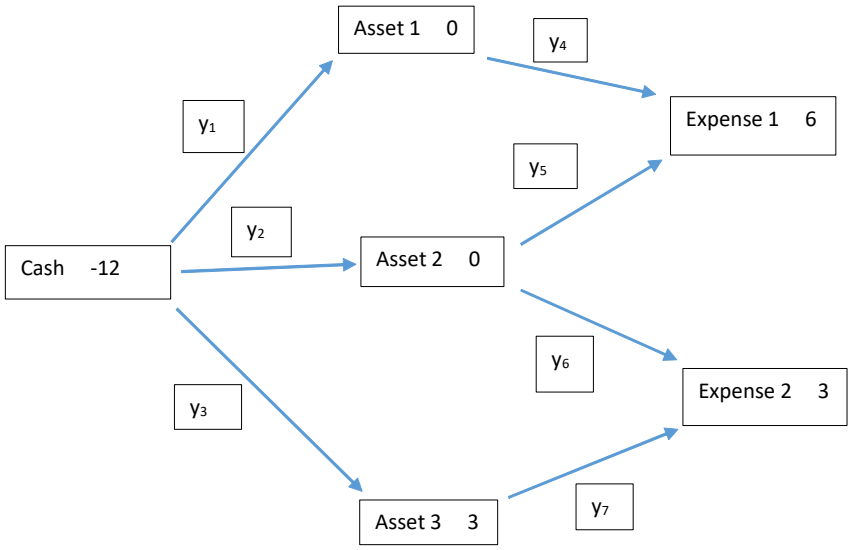


Figure 3.6

Example 3.8 in directed graph format

expense 1 y_5
asset 2 y_5

expense 2 y_6
asset 2 y_6

expense 2 y_7
asset 3 y_7

Compute y_{row}.

To identify the loops (the nullspace vectors) construct the graph in figure 3.6.

It is clear from inspection (and can also be verified by Euler's theorem) that there are two independent loops.[4] An algebraic characterization of the two loops,

[4]The number of loops is the number of journal entries (7) minus the number of accounts (6) plus one.

along with a sample y_p solution, is presented in the table.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
y_p	0	9	3	0	6	3	0
N^T	1	−1	0	1	−1	0	0
	0	1	−1	0	0	1	−1

The orthogonality condition requires the difference vector $y_p - N\beta$ be orthogonal to both columns of N.

$$N^T(y_p - N\beta) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The two zeros on the right are for the two vector products. Restating

$$N^T N\beta = N^T y_p$$

represents two linear equations with two unknowns, the regression coefficients $\beta = \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}^T$. Computing the vector products

$$\begin{aligned} N^T N &= \begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \\ 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} N^T y_p &= \begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 9 \\ 3 \\ 0 \\ 6 \\ 3 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -9-6 \\ 9-3+3 \end{bmatrix} = \begin{bmatrix} -15 \\ 9 \end{bmatrix} \end{aligned}$$

Substituting into the orthogonality conditions

$$\begin{aligned} N^T N\beta &= N^T y_p \\ \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= \begin{bmatrix} -15 \\ 9 \end{bmatrix} \end{aligned}$$

The two linear equations written separately are

$$\begin{aligned} 4\beta_1 - \beta_2 &= -15 \\ -\beta_1 + 4\beta_2 &= 9 \end{aligned}$$

There are many ways to solve two linear equations in two unknowns. One way that does the trick is to multiply the first equation times 4, and then add the two equations together, thereby eliminating β_2. The solution

$$\begin{aligned}\beta_1 &= -3.4\\ \beta_2 &= 1.4\end{aligned}$$

The nullspace vector y_N is computed as $N\beta$..

$$\begin{aligned} y_N &= N\beta \\ \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \\ 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -3.4 \\ 1.4 \end{bmatrix} &= \begin{bmatrix} -3.4 \\ 3.4+1.4 \\ -1.4 \\ -3.4 \\ 3.4 \\ 1.4 \\ -1.4 \end{bmatrix} = \begin{bmatrix} -3.4 \\ 4.8 \\ -1.4 \\ -3.4 \\ 3.4 \\ 1.4 \\ -1.4 \end{bmatrix} \end{aligned}$$

And y_{row} is computed from

$$\begin{aligned} y_p &= y\,\text{row} + y_N \\ y_{\text{row}} &= y_p - y_N \\ &= \begin{bmatrix} 0 \\ 9 \\ 3 \\ 0 \\ 6 \\ 3 \\ 0 \end{bmatrix} - \begin{bmatrix} -3.4 \\ 4.8 \\ -1.4 \\ -3.4 \\ 3.4 \\ 1.4 \\ -1.4 \end{bmatrix} = \begin{bmatrix} 3.4 \\ 4.2 \\ 4.4 \\ 3.4 \\ 2.6 \\ 1.6 \\ 1.4 \end{bmatrix} \end{aligned}$$

The calculations are summarized in the table.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
y_p	0	9	3	0	6	3	0
N	1	−1	0	1	−1	0	0
	0	1	−1	0	0	1	−1
$N\beta$	−3.4	4.8	−1.4	−3.4	3.4	1.4	−1.4
$y_{row} = y_p - N\beta$	3.4	4.2	4.4	3.4	2.6	1.6	1.4

y_{row} could, and should, be checked by verifying

$$\begin{aligned} Ay_{\text{row}} &= x \\ N^T y_{\text{row}} &= 0 \end{aligned}$$

One way to verify both conditions is to check the directed graph in figure 3.7. The orthogonality condition can be verified by computing the sum of the directed amounts around each of the loops. The first loop, for example, is

$$3.4 + 3.4 - 2.6 - 4.2 = 0$$

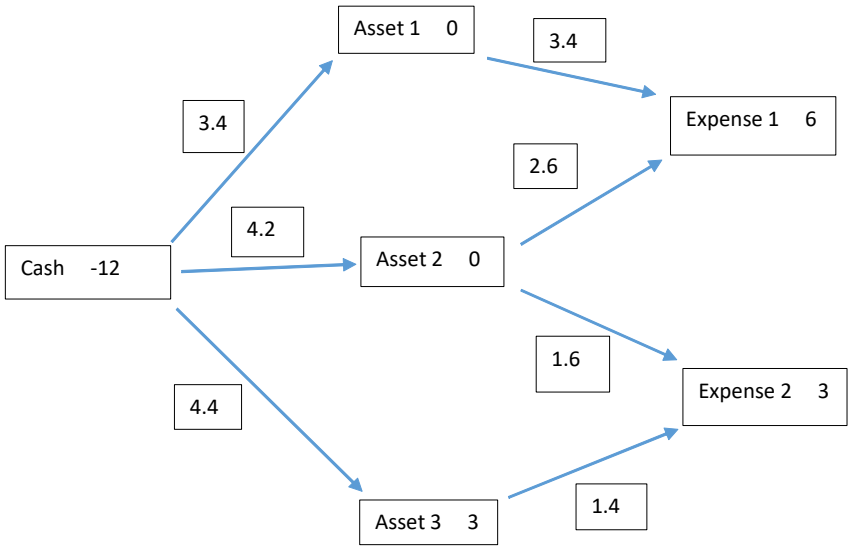


Figure 3.7

y_{row} for example 3.6

3.10 summary

The basic process in the chapter was to start with a journal entry vector, y, operate with a double entry accounting matrix, A, and generate a financial statement vector, x. The question is how much of the information in y reaches the vector x. And the answer is simple to state: the information in the row component in y, called y_{row}, gets through the channel to x, and it is all that gets to x.

The central part of the problem is the computation of y_{row}. We studied five solution techniques; it is a little bit remarkable that all five techniques yield the same solution. The fundamental theorem of linear algebra is instructive on this point. Each solution technique, in turn, adds more interpretation to the solution. As accounting is an information science, an information interpretation is useful: the unique solution, y_{row}, is all the information available in the financial statement vector, x, about the underlying journal entry amounts. Along the way the

five solution techniques illustrate and invoke some theorems, notably fundamental theorems about the two basic activities of applied mathematics: optimization, estimation, and their interrelationships.

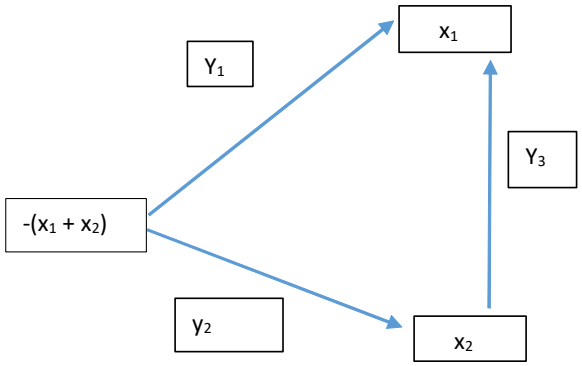


Figure 3.8

Exercise 3.1 directed graph

3.11 reference

Strang, Gilbert, *Linear Algebra and Its Applications*. Harcourt Brace Jovanovich, 1986.

3.12 exercises

Exercise 3.1 *Figure 3.8 presents a double entry example in directed graph form. It is equivalent to a cash payment, some of which goes to an asset, some to expense, and some of the asset is amortized over time.*

a. Suppose $x_1 = x_2 = 5$, and $y = \begin{bmatrix} 2 & 8 & 3 \end{bmatrix}^T$. Compute y_{row}. What fraction of y gets through to x?

b. Suppose $x_1 = 4$ and $x_2 = 6$, and $y = \begin{bmatrix} 2 & 8 & 2 \end{bmatrix}^T$. Compute y_{row}. What fraction of y gets through to x?

c. Use general x_1 and x_2. Compute y_{row} in terms of x.

Exercise 3.2 *Figure 3.9 presents a double entry example in directed graph form. It is generally equivalent to cash outlays to three cost pools which are then converted into two output products.*

a. Let $x = 5$ and $y = \begin{bmatrix} 2 & 5 & 3 & 2 & 3 & 2 & 3 \end{bmatrix}^T$. Compute y_{row} and the fraction of y reaching x (R^2).

b. Use general x. Compute y_{row} in terms of x.

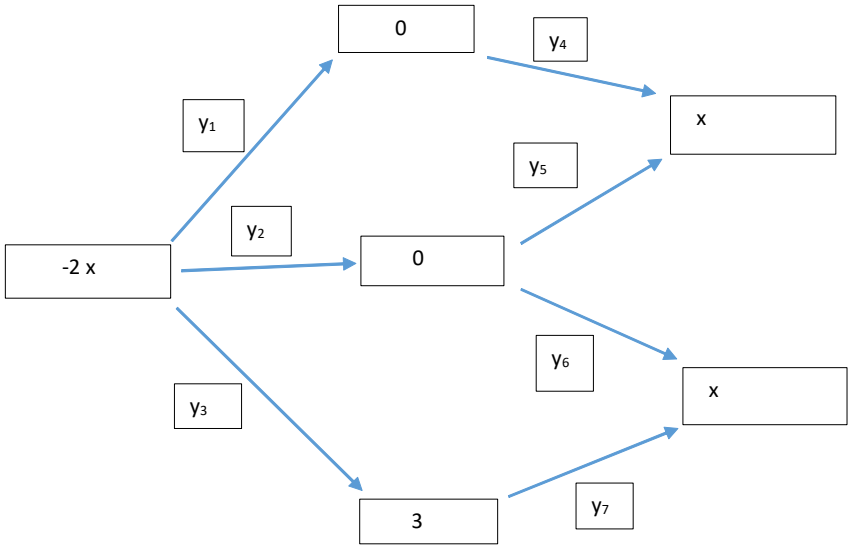


Figure 3.9

Exeercise 3.2 in directed graph format

Exercise 3.3 *Here are partial financial statements.*

partial balance sheet

	ending balance	*beginning balance*
asset 1	*100*	*100*
asset 2	*150*	*150*
asset 3	*115*	*70*

partial income statement

expense 1	*75*
expense 2	*105*

Here are the only journal entries affecting the above accounts.

asset 1	y_1	
cash		y_1
asset 2	y_2	
cash		y_2
asset 3	y_3	
cash		y_3
expense 1	y_4	
asset 1		y_4
expense 1	y_5	
asset 2		y_5
expense 2	y_6	
asset 2		y_6
expense 2	y_7	
asset 3		y_7

Compute y_{row}.

Exercise 3.4 *Here are financial statements.*

balance sheet

	ending	*beginning*		*ending*	*beginning*
cash	*4*	*2*	*payables*	*11*	*8*
receivables	*8*	*4*	*capital stock*	*10*	*10*
inventory	*8*	*9*	*retained earnings*	*9*	*4*
equipment	*10*	*7*			
total assets	*30*	*22*	*total equities*	*30*	*22*

income statement

sales	*10*
cost of goods sold	*2*
gen'l & admin. expenses	*3*
income	*5*

Here are the only journal entries affecting the above accounts.

receivables	y_1	
sales		y_1
cash	y_2	
receivables		y_2
payables	y_3	
cash		y_3
equipment	y_4	
cash		y_4
g&a expense	y_5	
payables		y_5
CGS	y_6	
inventory		y_6
inventory	y_7	
equipment		y_7
inventory	y_8	
payables		y_8

Compute y_{row}.

Exercise 3.5 *Here are partial financial statements.*

partial balance sheet

	ending balance	*beginning balance*
asset 1	*90*	*100*
asset 2	*180*	*150*

partial income statement

expense	*80*

Here are the only journal entries affecting the above accounts.

asset 1	y_1	
cash		y_1
asset 2	y_2	
cash		y_2
expense	y_3	
cash		y_3

expense	y_4	
asset 1		y_4
expense	y_5	
asset 2		y_5

Compute y_{row}.

The following three exercises are about pension accounting. The idea is that directed graphs are a useful tool for unraveling pension accounting, in particular, computing numbers like the pension investment made during the period, and pension payments made to pensioners.

Exercise 3.6 *Revisit exercise 2.1, particularly the directed graph depicting pension activity. Here is an example of the supplementary disclosure accompanying financial statements.*

	2007	*2006*
projected obligation	*$289.500*	*$265.000*
plan assets	*190.600*	*159.600*
prepaid/(accrued) pension cost	*$(98.900)*	*$(105.400)*
comprehensive income adjustments:		
unrecognized (gain)/loss	*23.728*	*29.940*
unrecognized prior service cost	*14.400*	*32.000*

Notice the supplementary disclosure allows computing the x vector. For example, the PBO node is -24.5 (increase in a credit balance). Also available is the pension cost: 44.312.

Construct a directed graph with all the accounts affected by the pension journal entries. Attach the change in the account balances (x) wherever possible.

Exercise 3.7 *This is a continuation of the previous problem. Additional supplementary disclosure includes the components of pension cost. This allows specifying some of the elements of the y vector.*

components of pension cost:	*2007*
service cost	*$16.000*
interest	*26.500*
(return)	*(22.000)*
unexpected gain/loss	*6.040*
amort. of prior service cost	*17.600*
amort. of unrecog. (gain)/loss	*0.172*
net pension cost	*$44.312*

Which y vector generates the pension disclosure amounts? Is there more than one solution to $Ay = x$?

Exercise 3.8 *Here is another example of supplementary financial disclosure for pensions.*

	2007	*2006*
projected obligation	*$60*	*$50*
plan assets	*38*	*30*
prepaid/(accrued) pension cost	*$(22)*	*$(20)*
comprehensive income adjustments:		
unrecognized (gain)/loss	*10*	*6*
unrecognized prior service cost	*8*	*12*

components of pension cost:	2007
service cost	20
interest	10
(return)	(15)
unexpected gain/loss	3
amort. of prior service cost	4
amort. of unrecog. (gain)/loss	2
net pension cost	$24

Compute a solution to $Ay = x$. Is there more than one solution?

Exercise 3.9 *Reconsider example 3.3. The quadratic program is (after eliminating a redundant constraint)*

$$\begin{aligned} & & Min\ \ y_1^2 + y_2^2 + y_3^2 \\ s.t. y_1 + y_2 &= 12 \\ y_1 + y_3 &= 9 \end{aligned}$$

The method of Lagrange combines the objective with the left-hand side of the constraints into one expression.

$$\mathcal{L} = y_1^2 + y_2^2 + y_3^2 + \lambda_1 (y_1 + y_2) + \lambda_2 (y_1 + y_3)$$

The λ's are Lagrange multipliers (shadow prices on the constraints). The first order conditions for a local optimum are the partial derivatives of $\mathcal{L}$ are equal to zero.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial y_1} &= 2y_1 + \lambda_1 + \lambda_2 = 0 \\ \frac{\partial \mathcal{L}}{\partial y_2} &= 2y_2 + \lambda_1 = 0 \\ \frac{\partial \mathcal{L}}{\partial y_1} &= 2y_3 + \lambda_2 = 0 \end{aligned}$$

With the two original constraints we have five linear equations in five unknowns. In this problem the multipliers are easily substituted out.

$$\begin{aligned} \lambda_1 &= -2y_2 \\ \lambda_2 &= -2y_3 \end{aligned}$$

And we are left with three relatively simple linear equations in three unknowns to get y_{row}.

$$\begin{aligned} 2y_1 - 2y_2 - 2y_3 &= 0 \\ y_1 + y_2 &= 12 \\ y_1 + y_3 &= 9 \end{aligned}$$

Exercise 3.10 *Decompose the vector* $y^T = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ *into 2 orthogonal components, one of which is a scalar multiple of* $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$. *What is the* R^2*?*

Exercise 3.11 *Decompose the vector* $y^T = \begin{bmatrix} 0 & 5 & 10 \end{bmatrix}$ *into 2 orthogonal components, one of which is in the row space of*

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$

Exercise 3.12 *Redo the previous exercise with* $y^T = \begin{bmatrix} 5 & 0 & 10 \end{bmatrix}$.

4

the theorem of the separating hyperplane

One purpose of these chapters is to proceed through a series of activities.

1. Use double entry accounting to acquire and understand some intellectual tools. Often the tools are in the form of theorems, but, more generally, the idea is to build a solid academic foundation which will support careful study of various topics.

2. Use the tools so acquired to illuminate other (non-accounting) problems and endeavors.

3. Use the understanding of other problems to illuminate accounting problems.

In this chapter, and the next, the plan is to proceed through the three step sequence using an important theorem, an important result from another (non-accounting) intellectual topic, and an important problem in accounting.

1. Use double entry accounting to illustrate, and hopefully understand, the theorem of the separating hyperplane.

2. Use the theorem of the separating hyperplane to illustrate (and understand) the concept of equilibrium prices in a setting without arbitrage opportunities. The important result in this area is often called the fundamental theorem of finance which is a restatement of the theorem of the separating hyperplane.

3. Use the concept of equilibrium prices to illuminate an accounting problem. In particular, phrases like "fair value" and "market value" have often been

used to prescribe how to value assets and liabilities on the balance sheet. The issue is how to think carefully about these phrases in an equilibrium setting.

Before starting the example, state the theorem.

Theorem 4.1 *For any matrix A and appropriate size vector x, either $Ay = x$ has a solution $y \geq 0$, or there exists a λ such that $A^T\lambda \geq 0$ and $\lambda^T x < 0$. One, or the other, statement is true, but not both. See Strang, p. 420 and Gale, p. 44.*

The theorem is true for any matrix, not just incidence matrices. Because of the accounting connections, and because the examples, are pretty crisp, we will use incidence matrices to illustrate and explore the theorem.

4.1 accounting illustration of the theorem

Reconsider the simple example from the previous chapter.

$$A = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} -12 \\ 9 \\ 3 \end{bmatrix}$$

We already have $y = \begin{bmatrix} 7 & 5 & 2 \end{bmatrix}^T$, and numerous other $y \geq 0$, that satisfy $Ay = x$. The theorem states, then, that there is not a λ vector satisfying

$$A^T\lambda \geq 0 \quad \text{and} \quad \lambda^T x < 0$$

Rather than spend time looking fruitlessly for a λ, change the example. Now let

$$x = \begin{bmatrix} -12 \\ -1 \\ 13 \end{bmatrix}$$

The directed graph representation of the revised example is in figure 4.1

It is obvious from the directed graph that there is no positive y vector satisfying $Ay = x$. As the expense account is negative, and both the arrows touching it, y_1 and y_3, are pointing into the account, at least one of the arrows *must* be negative in order to decrease expense by 1. Choice of a λ vector exploits those two properties:

x	λ		$A:$	
-12	0	-1	-1	0
-1	1	1	0	1
13	0	0	1	-1
$\lambda^T x = -1$	$A^T\lambda$	1	0	1

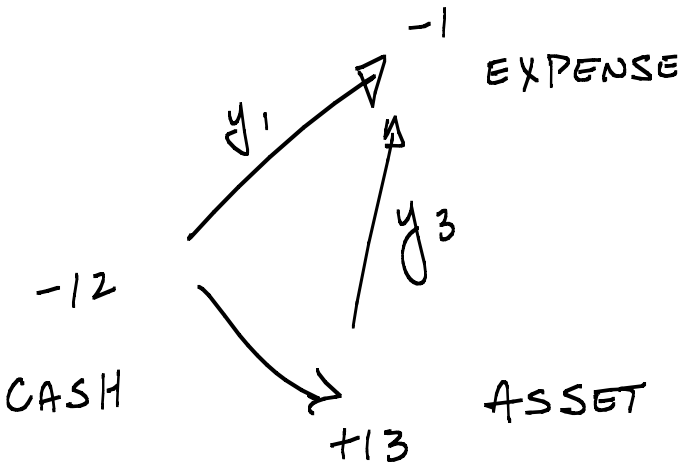


FIGURE 4.1
SIMPLE EXAMPLE
THEOREM OF THE SEPARATING
HYPERPLANE

The λ vector has a 1 in the position corresponding to the expense account, and zeros elsewhere. $A^T\lambda$ is the vector products of λ with the columns of A. Since an arrow into expense means a $+1$ in the corresponding column of A (the debit is the arrowhead), the vector products will be 1 for y_1 and y_3, and zero for y_2. Likewise, the vector product $\lambda^T x$ is the balance in the expense account, -1. So the two properties of the graph which make it impossible for there to be a non-negative solution to $Ay = x$, are the same two properties exploited by the choice of the λ vector:

- all arrows into the account ($A^T\lambda \geq 0$),
- negative balance in the account ($\lambda^T x < 0$).

4.2 another accounting illustration of the theorem: set-up

Here's another, slightly more extensive, example of the theorem of the separating hyperplane, utilizing the double entry structure, and especially the directed graph representation.

Example 4.1

balance sheet

	ending	*beginning*		*ending*	*beginning*
cash	*6*	*2*	*payables*	*12*	*8*
receivables	*5*	*4*	*capital stock*	*10*	*10*
inventory	*8*	*9*	*retained earnings*	*7*	*4*
plant	*10*	*7*			
total assets	*29*	*22*	*total equities*	*29*	*22*

income statement

sales	*7*
cost of goods sold	*2*
gen'l & admin. expenses	*2*
income	*3*

The allowable journal entries are denoted in the directed graph in figure 4.2. There are a couple of new journal entries.

gen'l & admin. expenses y_8
plant y_8

inventory y_{10}
plant y_{10}

Long-lived assets are used up in the production process. Sometimes the using up results in a new asset, inventory, and sometimes no new asset is produced; rather,

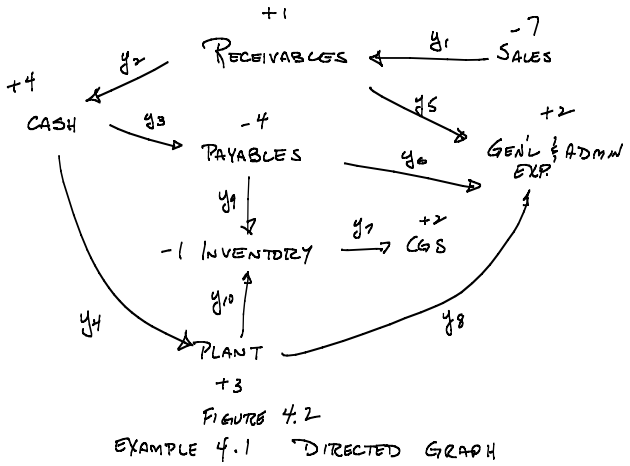


FIGURE 4.2
EXAMPLE 4.1 DIRECTED GRAPH

an expense is recognized. We note in passing that this reconstitution of assets makes it more difficult than one might think to cleanly distinguish one asset from another.

Construct the A matrix and the x vector.

x		y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
4	cash	0	1	-1	-1	0	0	0	0	0	0
1	rec.	1	-1	0	0	-1	0	0	0	0	0
-1	inv.	0	0	0	0	0	0	-1	0	1	1
3	plant	0	0	0	1	0	0	0	-1	0	-1
-4	pbles	0	0	1	0	0	-1	0	0	-1	0
-7	sales	-1	0	0	0	0	0	0	0	0	0
2	cgs	0	0	0	0	0	0	1	0	0	0
2	g&a	0	0	0	0	1	1	0	1	0	0

4.3 another accounting illustration of the theorem

Start the inquiry with the standard question: what is y_{row}? There are, of course, several methods, all of which yield the following.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
y_{row}	7	$\frac{10}{3}$	$-\frac{3}{2}$	$\frac{5}{6}$	$\frac{8}{3}$	$\frac{5}{6}$	2	$-\frac{3}{2}$	$\frac{5}{3}$	$-\frac{2}{3}$

While it might be a little bit difficult to find y_{row} with a problem of this size,[1] it is easily verified that the proposed y_{row} satisfies $Ay = x$ and is orthogonal to

[1]The paper and pencil methods are a little tedious, but by no means impossible. There are 36 spanning trees, for example, and projection into the nullspace effectively requires solving 3 linear equations for 3 unknowns.

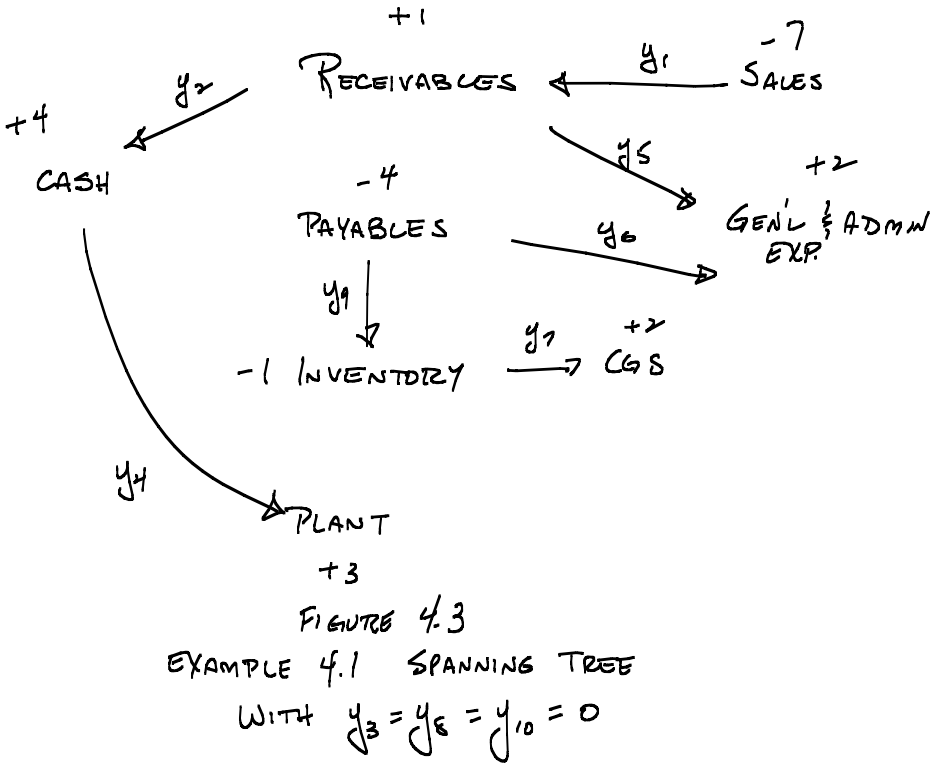


FIGURE 4.3
EXAMPLE 4.1 SPANNING TREE
WITH $y_3 = y_8 = y_{10} = 0$

the nullspace of A. The nullspace can be constructed, as before, by isolating the loops in figure 4.2.

nullspace of A

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
0	1	1	0	−1	1	0	0	0	0
0	0	1	−1	0	0	0	0	1	−1
0	0	0	0	0	1	0	−1	−1	1

Now consider y_{row}. We note that y_{row} has some negative elements: y_3, y_8, and y_{10}. The question here, and the question the theorem of the separating hyperplane can answer, is whether there exists a non-negative solution to $Ay = x$.

When we have encountered negative elements in y_{row} in previous examples, we found non-negative solutions in the set of spanning trees. As we have three loops and three negative elements in y_{row}, construct the spanning tree by setting the troublesome elements, y_3, y_8, and y_{10}, equal to zero. The resulting spanning tree is in figure 4.3.

We note that the spanning tree doesn't solve the non-negativity problem. The entire y vector is

$$y = \begin{bmatrix} 7 & 7 & 0 & 3 & -1 & 3 & 2 & 0 & 1 & 0 \end{bmatrix}$$

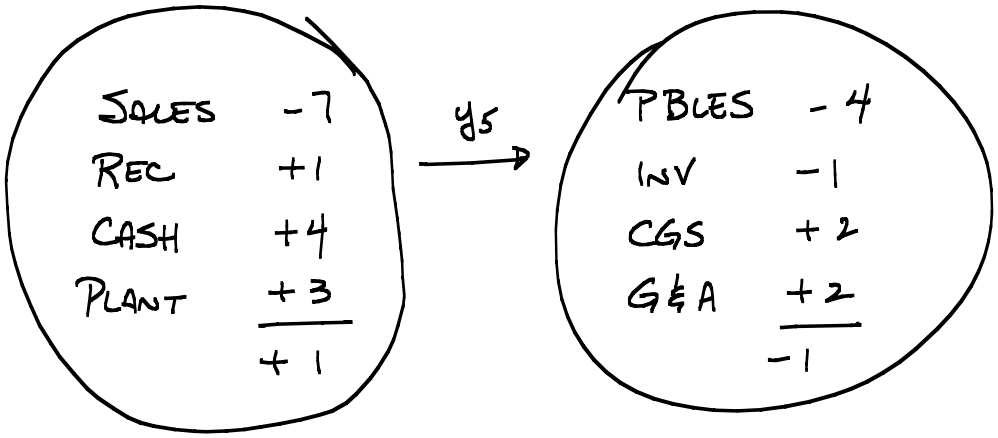


FIGURE 4.4

EXAMPLE 4.1 - CIRCLE PICTURE WITH y_5 & COMPOSITE ACCOUNTS

There is still one negative element: $y_5 = -1$. This, however, suggests another representation: combine all the accounts on either side of y_5 into two composite accounts as depicted in the circle picture in figure 4.4.

Now re-enter y_3, y_8, and y_{10} into the circle picture. y_3 is a debit to payables and a credit to cash, so it connects the two composite accounts with an arrow from left to right, as do y_8 and y_{10}. The augmented circle picture is in figure 4.5.

The augmented circle picture makes it clear: a non-negative solution to $Ay = x$ can not exist. In order for the right hand circle to have a balance of minus one, one dollar must have been moved from the right to the left circle. But all the arrows are pointing from the left to the right. Hence, at least one of the four journal entries (y_3, y_5, y_8, or y_{10}) must be strictly negative. That is, one of the journal entries must move backwards on an arrow.

So the circle picture answers the existence question, and, since that is the same question resolved by the theorem, it seems there is a way to connect, and indeed there is. To connect to the theorem fill in the λ vector: set $\lambda_i = 1$ for all accounts in the right circle, and $\lambda_i = 0$ for all the accounts in the left. Add an extra column,

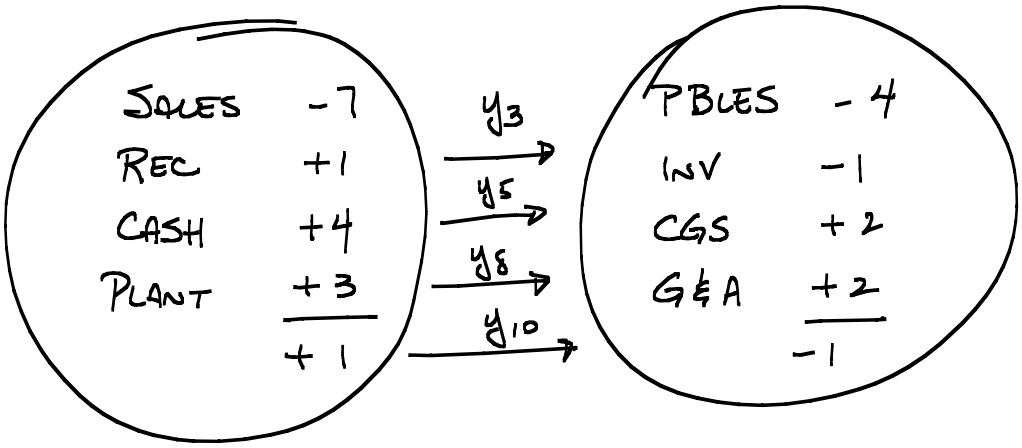


FIGURE 4.5

EXAMPLE 4.1 - CIRCLE PICTURE WITH $y_3, y_5, y_8, \& y_{10}$

λ, to the A matrix and x vector, as well as a row for $A^T\lambda$ and a cell for $\lambda^T x$.

λ	x		y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
0	4	cash	0	1	-1	-1	0	0	0	0	0	0
0	1	rec.	1	-1	0	0	-1	0	0	0	0	0
1	-1	inv.	0	0	0	0	0	0	-1	0	1	1
0	3	plant	0	0	0	1	0	0	0	-1	0	-1
1	-4	pbles	0	0	1	0	0	-1	0	0	-1	0
0	-7	sales	-1	0	0	0	0	0	0	0	0	0
1	2	cgs	0	0	0	0	0	0	1	0	0	0
1	2	g&a	0	0	0	0	1	1	0	1	0	0
$\lambda^T x = -1$		$A^T\lambda$	0	0	1	0	1	0	0	1	0	1

$A^T\lambda$ is calculated as the vector product of λ with each column (journal entry) in A. The only non-zero vector products are for y_3, y_5, y_8, and y_{10}, as they are the only ones crossing the gap between the circles, and, because of the arrow direction, they are $+1$. $\lambda^T x$ is the total balance in the right circle.

The theorem tells us the same thing as the circle picture. Either there is a non-negative solution to $Ay = x$, or there exists a λ with $\lambda^T x < 0$ and $A^T\lambda \geq 0$. Here $\lambda^T x = -1$ and all the elements of $A^T\lambda$ are either zero or one. So, for the example, there exists no non-negative y vector, and further searching would be a waste of time. But the real reason for the exercise was to gain some understanding of the theorem. Now we are sufficiently armed to confront another hard problem in accounting: what to do with derivatives and other assets in incomplete markets.

4.4 arbitrage free equilibrium pricing

The theorem of the separating hyperplane is the foundation for no arbitrage equilibrium pricing, of which options pricing is a special case. As usual we start with a numerical example; this one is adapted from Cox, Ross, and Rubinstein. The example is presented in state-act-outcome format. Next period's (dollar) outcome is jointly determined by an action chosen by a decision maker, and next period's state of nature, the realization of which is beyond the control of the decision maker.

Example 4.2 *There are two possible states of nature, denoted θ_1 and θ_2, and three possible actions this period: purchase a bond, a share of stock, or a call option. Next period's outcomes for each security in each state are presented in the table.*

price	*security*	θ_1	θ_2
.8	*bond*	*1*	*1*
50	*stock*	*25*	*100*
c(?)	*call*	*0*	*50*

The prices of the different securities are included in the left column; the (equilibrium) price of the call, c, is unknown. The return on the bond is the same in both states (it is a risk free security), while the returns on the stock and call depend on which state is realized. θ_2 is a high return state for both securities.

Perhaps the states could be thought of as next period's weather: say, rain for θ_1 and sunshine for θ_2. Consistent with the weather interpretation, the stock could be ownership of an ice cream manufacturer. If the sun shines next period, an investment in ice cream is worth more.

The first question of interest is the equilibrium price of the call. One way to proceed is to construct a portfolio of the bond and stock, such that the state returns on the call are reproduced exactly by the portfolio. For the portfolio to return 0 in θ_1 and 50 in θ_2, just like the call, the following equations must hold. Here the λ vector represents the portfolio weights: that is, λ_1 is the number of bonds in the portfolio.

$$\begin{aligned} 1\lambda_1 + 25\lambda_2 &= 0 \\ 1\lambda_1 + 100\lambda_2 &= 50 \end{aligned}$$

Solving the two independent linear equations:

$$\begin{aligned} \lambda_1 &= -\frac{50}{3} \\ \lambda_2 &= \frac{2}{3} \end{aligned}$$

The portfolio we form consists of *selling* 50/3 bonds (as $\lambda_1 = -50/3$), and *buying* 2/3 of a share of stock. Then in the next period, returns will be 0 in θ_1 and 50 in θ_2. Importantly, we know how much the portfolio will cost:

$$-\frac{50}{3}(.8) + \frac{2}{3}(50) = 20$$

As the portfolio which reproduces the returns of the call costs 20, it seems reasonable that the call, itself, should also cost 20. Indeed, if the price of the call is something other than 20, we can exploit the situation. That's where arbitrage comes in.

Suppose the call price is 30. What we can do is buy the stock-bond portfolio for 20, sell the call for 30, and come out 10 to the good. (If we don't actually possess a call to sell, we can do what is called "sell short," that is, promise someone 50 if θ_2 occurs, and 0 in θ_1, and charge them 30 for the promise.) Whatever happens in the next period, we are covered. If θ_2 occurs the stocks and bonds will generate 50, exactly enough to cover the cost of the call promise. If θ_1 occurs, no money changes hands. That's what is called arbitrage.

Definition 4.1 *An arbitrage opportunity exists if it is possible, given the prices, to construct a portfolio of securities which generates a non-negative return in all states in the next period, and the portfolio has a strictly negative price.*

The negative price means the market will pay us to hold the portfolio; that's the net 10 we receive for buying the bond-stock portfolio for 20 and selling the call for 30. Let's see if we can represent the arbitrage portfolio in notation.

In terms of portfolio weights, selling the call means $\lambda_3 = -1$, so the λ vector is

$$\lambda = \begin{bmatrix} -\frac{50}{3} \\ \frac{2}{3} \\ -1 \end{bmatrix}$$

The A matrix consists of the returns on each of the securities.

$$A = \begin{bmatrix} 1 & 1 \\ 25 & 100 \\ 0 & 50 \end{bmatrix}$$

And x is the vector of security prices, with the call priced at 30 for the example.

$$x = \begin{bmatrix} .8 \\ 50 \\ 30 \end{bmatrix}$$

The fact that we are covered no matter what happens next period is in the vector product $A^T\lambda$.

$$\begin{aligned} A^T\lambda &= \begin{bmatrix} 1 & 1 \\ 25 & 100 \\ 0 & 50 \end{bmatrix}^T \begin{bmatrix} -\frac{50}{3} \\ \frac{2}{3} \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 25 & 0 \\ 1 & 100 & 50 \end{bmatrix} \begin{bmatrix} -\frac{50}{3} \\ \frac{2}{3} \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned}$$

The (negative) price of the portfolio is $\lambda^T x$.

$$\lambda^T x = \begin{bmatrix} -\frac{50}{3} \\ \frac{2}{3} \\ -1 \end{bmatrix}^T \begin{bmatrix} .8 \\ 50 \\ 30 \end{bmatrix} = -10 < 0$$

So the two parts of the definition of an arbitrage opportunity portfolio have linear algebra representations.

- generates a non-negative return whatever happens next period:

$$A^T \lambda \geq 0;$$

- has a strictly negative price:

$$\lambda^T x < 0$$

It is important to notice that, if the price of the call is 20 (as implied by the price of the equivalent bond-stock portfolio), the arbitrage opportunity would disappear. In this case, $\lambda^T x$ would be zero. The equilibrium idea is that, if arbitrage opportunities exist, the forces of supply and demand will adjust the prices. That is, an attractive portfolio would be bid up in price, and the arbitrage opportunity would disappear. In the case of the example, since everyone would want to sell the call for 30 to earn arbitrage, the price of the call would fall.

Definition 4.2 *An arbitrage free equilibrium is a set of prices such that no arbitrage opportunities exist.*

For the example, the set of arbitrage free prices is

$$x = \begin{bmatrix} .8 \\ 50 \\ 20 \end{bmatrix}$$

So far we haven't used the theorem of the separating hyperplane, but now is the time. In an arbitrage free equilibrium, there does not exist a λ vector that satisfies

$$\begin{aligned} A^T \lambda &\geq 0 \\ \lambda^T x &< 0 \end{aligned}$$

The theorem tells us that can not happen if there is a $y \geq 0$ that satisfies

$$Ay = x$$

It is usually much (much) easier to solve for y than it is to search for arbitrage portfolio weights. When the theorem is used in this context, it is often referred to as the fundamental theorem of finance.

To solve for y in the example:

$$\begin{aligned} Ay &= x \\ \begin{bmatrix} 1 & 1 \\ 25 & 100 \\ 0 & 50 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} .8 \\ 50 \\ c \end{bmatrix} \end{aligned}$$

The first two equations are often referred to as the state price equations.

$$\begin{aligned} 1y_1 + 1y_2 &= .8 \\ 25y_1 + 100y_2 &= 50 \end{aligned}$$

The solution is $y_1 = y_2 = .4$. Substituting into the third equation of $Ay = x$ yields

$$\begin{aligned} 0y_1 + 50y_2 &= c \\ 20 &= c \end{aligned}$$

And we get the equilibrium call price directly from the equations. We don't have to search for portfolio weights. The theorem tells us the weights will not exist that allow us (or anyone) to generate arbitrage returns. There is no λ if we have a non-negative y.

It remains to give y a name; they are typically referred to as state prices.

Definition 4.3 *The state price for state i is the amount paid in order to generate an outcome of \$1 in state i.*

We know from the state price equations that an investment of .8 (in a bond) will generate 1 dollar in state 1 (θ_1), as well as 1 dollar in state 2 (θ_2). Likewise, an investment of 50 will generate 25 dollars in θ_1 and 100 in θ_2. It is implied, therefore, that it takes .4 to generate 1 dollar in θ_1 as well as in θ_2.

In the future we will often find it useful to think in terms of Arrow-Debreu securities (named after famous economists Kenneth Arrow and Gerard Debreu).

Definition 4.4 *An Arrow-Debreu security returns one dollar in some state i and zero in all other states.*

The price of an Arrow-Debreu security that returns one dollar in state i is the state price for state i, that is y_i. We can specify the characteristics of any security by treating it as combination of Arrow-Debreu securities. Solving for the state price vector y, and invoking the fundamental theorem of finance, enables us to specify the arbitrage free equilibrium price for all kinds of securities. In the next section we consider settings where the state prices are not unique.

It is important to note that state probabilities were absent as we determined arbitrage free prices. Indeed, a strength of arbitrage pricing is that we need not specify the probabilities of states to proceed. Any time we have an arbitrage free equilibrium, the equilibrium conditions hold for any probabilities attached to state

realization. We can, however, go in the other direction. That is, given equilibrium prices, we can infer the nature of the state probabilities from observation of equilibrium prices.

State probabilities are closely related to the state prices; indeed, the state prices, themselves, are already very close to probabilities. We will exploit this similarity to help with the probability assignment task in chapter 8. For now a brief introduction will suffice.

Recall from the example we have state prices

$$y_1 = y_2 = .4$$

Probability numbers, p_i, lie between zero and one, and sum to one.

$$\begin{aligned} 0 &\leq p_i \leq 1 \\ \sum_i p_i &= 1 \end{aligned}$$

The $y's$ satisfy the first property but not the second. One way to achieve probabilities is to simply rescale the state prices, that is, divide by the sum of the $y's$.

$$\begin{aligned} p_1 &= \frac{y_1}{\sum y_i} = .\frac{.4}{.8} = .5 \\ p_2 &= \frac{y_2}{\sum y_i} = .\frac{.4}{.8} = .5 \end{aligned}$$

And these are one version of state probabilities which will reappear in chapter 8. (Sometimes they are referred to as risk neutral probabilities.)

4.5 multiple equilibrium prices

Expand the ongoing example by adding another state.

price	security	θ_1	θ_2	θ_3
.8	bond	1	1	1
50	stock	25	50	100
c(?)	call	0	0	50

The initial question is the same as before: what is the equilibrium price of the call? This time through, use the theorem directly.

$$\begin{aligned} A &= \begin{bmatrix} 1 & 1 & 1 \\ 25 & 50 & 100 \\ 0 & 0 & 50 \end{bmatrix} \\ x &= \begin{bmatrix} .8 \\ 50 \\ c \end{bmatrix} \end{aligned}$$

According to the theorem, there are no arbitrage opportunities as long as $Ay = x$ has a non-negative solution. When we find $y \geq 0$, simply calculate $c = 50y_3$ to find the equilibrium call price.

Finding a non-negative y vector is not so difficult. To start, set $y_1 = 0$ and solve the remaining two independent linear equations for the other two unknowns.

$$\begin{aligned} y_2 + y_3 &= .8 \\ 50y_2 + 100y_3 &= 50 \end{aligned}$$

$y_2 = .6$ and $y_3 = .2$ solves the system, so $y = \begin{bmatrix} 0 & .6 & .2 \end{bmatrix}^T$, and $c = 10$, is a non-negative solution to $Ay = x$, thereby specifying a no arbitrage equilibrium. But for this problem, there are other solutions. Setting $y_2 = 0$ implies the following equations.

$$\begin{aligned} y_1 + y_3 &= .8 \\ 25y_1 + 100y_3 &= 50 \end{aligned}$$

And we have another solution, $y = \begin{bmatrix} .4 & 0 & .4 \end{bmatrix}^T$, and $c = 20,$ Setting $y_3 = 0$ does not yield another non-negative solution, as it implies

$$y_1 = -.4 < 0$$

The two acceptable solutions introduce us to the possibility of multiple equilibrium prices.

$$\begin{aligned} \text{For any } 0 &\leq \alpha \leq 1 \\ y &= \alpha \begin{bmatrix} .4 \\ 0 \\ .4 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 0 \\ .6 \\ .2 \end{bmatrix} \end{aligned}$$

is a non-negative solution of $Ay = x$. As far as the state price is concerned, y_3 can be anything between .2 and .4. This, in turn, implies the equilibrium price of the call can be anything between 10 and 20. For this simple example the price is not unique.

To check our thinking about the fundamental theorem of finance, consider what happens when a call price outside the bounds occurs. Suppose, for example, we observe $c = 5$. Can we find portfolio weights λ satisfying the two arbitrage conditions: $A^T\lambda \geq 0$ and $\lambda^T x < 0$? The theorem says we can.

Arbitrage portfolio weights are included:

λ	x		θ_1	θ_2	θ_3
50	.8	$A:$	1	1	1
-1	50		25	50	100
1	5		0	0	50
$\lambda^T x = -5 < 0$		$A^T\lambda$	25	0	0

As the call is undervalued relative to an arbitrage free equilibrium, it is rational to buy as much as possible. Buying one unit is enough to illustrate the arbitrage possibilities; that is, set $\lambda_3 = 1$. The stock-bond portion of the portfolio can then be chosen so as to guarantee an outcome of at least zero in every state. The closest equilibrium price to $c = 5$ is $c = 10$ when $y_1 = 0$. So it seems safe to ignore the state 1 conditions. (After all, $y_1 = 0$ is as if θ_1 does not exist.) Try, then, the portfolio weights, which satisfy the following equations, one for state 2 and one for state 3. For state 3 the stock-bond portion can pay 50, since the call will generate 50.

$$\begin{aligned} 1\lambda_1 + 50\lambda_2 &= 0 \\ 1\lambda_1 + 100\lambda_2 &= -50 \end{aligned}$$

The solution is $\lambda_1 = 50$ and $\lambda_2 = -1$, and arbitrage is easily verified. Also note $\lambda^T x = -5$, which means the trading profit is 5. That is, the market is paying us 5 to hold the portfolio. This seems sensible, as the price of the derivative is 5 less than we think it should be: we pay 5 and we think it is worth at least 10.

4.6 multiple equilibria and accounting

We have a tendency in accounting to think, and talk, as if every asset has a unique equilibrium value associated with it. The concept of fair value is offered as an answer to accounting conundrums when there exist alternative accounting methods. A relatively recent pronouncement of the Financial Accounting Standards Board (FAS 157) is an example of the fair value prescription.

If multiple equilibrium values, as was illustrated in the previous example, tend to exist in economic and accounting contexts, then at least a little more thought is required before accepting a fair value prescription. At a minimum the question arises: which equilibrium value should be used? Or, at another level, is it possible for accounting to report (no) arbitrage equilibrium values?

The multiple equilibrium situation arose in the example because there were three possible states and only two priced securities (from which the price of the third security was derived). Roughly speaking, there were states for which it was not possible to make a meaningful trade. That is, market transactions were not available even though participants might wish to transfer outcomes from different states. We note that firms, especially those involved in production, typically do not rely exclusively on market transactions. We do not observe, for example, firms participating in the spot market for labor.

At a slightly deeper level, it is often argued that information problems foreclose the possibility of market transactions in some circumstances. A privately informed seller of a used car, for example, might have difficulty finding a buyer, the buyer reasoning that, if the person familiar with the car wants to sell it, there must be something wrong. Akerlof (1970) discusses the effect of the used car problem on market transactions; Demski (2008, chapter 10) also develops the im-

plications of private information for a restriction in mutually beneficial trading opportunites.

In any event, we will move forward on the premise that multiple equilibria are sufficiently pervasive and important phenomena to warrant further study from an accounting point of view. In particular, the next chapter will consider the possibility that accounting is able to represent, in a meaningful way, (no) arbitrage equilibrium prices.

4.7 summary

As the title implies, this chapter is about the theorem of the separating hyperplane. It was first applied to an accounting problem - the existence of non-negative journal entries. Then the theorem was used to value derivative securities in an equilibrium without arbitrage opportunities. In that context the theorem is known as the fundamental theorem of finance.

The logic of the fundamental theorem of finance follows from our understanding of linear operations (honed in our study of double entry accounting). Besides an equilibrium concept, other ideas which will prove to be useful to us include Arrow-Debreu securities, state prices and related probability concepts .

For accounting purposes, an intriguing result is that the equilibrium value of an asset might not be unique. The result arises in incomplete market settings, or market settings with information problems, where accounting information might be salient.

The next chapter considers possible accounting responses to the existence of multiple equilibrium prices.

4.8 references

Akerlof, George, "The Market for Lemons: Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics*, 84(3): 488-500, 1970.

Cox, John C., Stephen A. Ross, and Mark Rubinstein, "Option Pricing: A Simplified Approach." *Journal of Financial Economics* 7: 229-63, 1979.

Demski, Joel, *Managerial Uses of Accounting Information*. Springer, 2008.

Gale, David, *The Theory of Linear Economic Models*. The University of Chicago Press, 1960.

Strang, Gilbert, *Linear Algebra and Its Applications*. Harcourt Brace Jovanovich, 1986.

4.9 exercises

Exercise 4.1 *Ralph faces an uncertain future captured by the following state-act-outcome matrix with two states and three securities.*

		outcomes	
security	*price*	θ_1	θ_2
bond	*1*	*1*	*1*
stock	*20*	*10*	*40*
call	*c?*	*0*	*10*

Using arbitrage pricing theory and the fundamental theorem of finance, find all the equilibrium values of the call. Also calculate all the equilibrium values of the two Arrow-Debreu securities (y_1 and y_2 using the notation in the chapter).

Suppose a call price of 4 is observed. Is it possible, trading the call at that price, to form an arbitrage portfolio? If so, specify the portfolio weights, setting the weight on the call to be plus or minus one.

Exercise 4.2 *Relative to the preceding problem Ralph's world becomes more uncertain and a third outcome state enters the picture.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	*1*	*1*	*1*	*1*
stock	*20*	*10*	*40*	*40*
call	*c?*	*0*	*0*	*10*

What are all the equilibrium values of the call? What are all the equilibrium values of the three Arrow-Debreu securities?

Exercise 4.3 *Here is a three security, three state setting in state-act-outcome format. The price of the call, c, is unspecified.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	*1*	*1*	*1*	*1*
stock	*70*	*30*	*65*	*90*
call	*c?*	*0*	*0*	*40*

Compute the no arbitrage equilibrium price of the call, as well as the equilibrium state prices.

Suppose a call price of 4 is observed. Is it possible, trading the call at that price, to form an arbitrage portfolio? If so, specify the portfolio weights, setting the weight on the call to be plus or minus one.

Exercise 4.4 *Here is a three security, three state setting in state-act-outcome format. The price of the generic derivative security is unspecified.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	*1*	*1*	*1*	*1*
stock	*1*	*0*	*1*	*2*
derivative	*?*	*2*	*1*	*0*

Compute the no arbitrage equilibrium price of the derivative, as well as the equilibrium state prices.

Suppose a derivative price of 4 is observed. Is it possible, trading at that price, to form an arbitrage portfolio? If so, specify the portfolio weights, setting the weight on the call to be plus or minus one.

Exercise 4.5 *Consider a horse race with three horses, conveniently named 1, 2, and 3. There are different bets (securities) available. A "win" bet pays off when the named horse wins. There are also "exacta" bets that pay off if the first two finishers are named in order. The cost of all bets is 2 dollars. Here are the payoffs: w1 is a bet that horse 1 wins; x12 is a bet the first two finishers are 1 and 2 in that order.*

bet	*payoff*
w1	$20/3$
w2	3
w3	10
x12	20
x13	10
x21	$20/3$
x23	10
x31	20
x33	20

So far the betting scheme is similar to what is available at racing venues, but here is a difference. In order to compare with securities trading, it is also possible to go short. That is, the bet (security) can be sold as well as purchased. We can act like the track or a bookie (at least in the old movies) by accepting 2 dollars and paying off when the specified outcome occurs.

Here is a state-act-outcome representation of the same bet payoffs. There are six states for the possible orders of finish.

order of finish	123	132	213	231	312	321
state	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
w1	$20/3$	$20/3$	0	0	0	0
w2	0	0	3	3	0	0
w3	0	0	0	0	10	10
x12	20	0	0	0	0	0
x13	0	10	0	0	0	0
x21	0	0	$20/3$	0	0	0
x23	0	0	0	10	0	0
x31	0	0	0	0	20	0
x32	0	0	0	0	0	20

Using the fundamental theorem of finance, determine if there are arbitrage opportunities available. If there are arbitrage opportunities, specify a vector of portfolio weights, λ, for each bet. Recall, short selling is allowed and all bets are priced at 2 dollars.

Exercise 4.6 *For this exercise there is no exacta betting, but "place" bets are allowed. A place bet pays off if the specified horse finishes either first or second. For example, a place bet on horse 1 pays off in state θ_1, θ_2, θ_3, and θ_5. "p2", for horse 2, pays off in state θ_1, θ_3, θ_4, and θ_6.*

order of finish	*123*	*132*	*213*	*231*	*312*	*321*
state	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
w1	$20/3$	$20/3$	0	0	0	0
w2	0	0	4	4	0	0
w3	0	0	0	0	10	10
p1	$20/7$	$20/7$	$20/7$	0	$20/7$	0
p2	2	0	2	2	0	2
p3	0	$10/3$	0	$10/3$	$10/3$	$10/3$

Arbitrage opportunities available? If so, specify a portfolio weight vector. (For this problem it might be useful to have a computer at hand to run a regression.) Short selling remains.

Exercise 4.7 *Same situation, and question, as the previous exercise with some slight changes in some of the payoff amounts.*

order of finish	*123*	*132*	*213*	*231*	*312*	*321*
state	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
w1	$20/3$	$20/3$	0	0	0	0
w2	0	0	4	4	0	0
w3	0	0	0	0	10	10
p1	$20/7$	$20/7$	$20/7$	0	$20/7$	0
p2	$20/7$	0	$20/7$	$20/7$	0	$20/7$
p3	0	$10/3$	0	$10/3$	$10/3$	$10/3$

Exercise 4.8 *Race tracks often use a system called parimutuel betting where a fraction of the money bet is retained by the track before bets are paid off. For this exercise the effect is simulated by reducing the win payoffs from the previous exercise by 10%. No place bets are included in the payoffs, but, for this exercise, a risk free security (bet) is available. The bettor (or the track) can invest (keep) the 2 dollar bet price so the payoff is 2 dollars in all states.*

order of finish	*123*	*132*	*213*	*231*	*312*	*321*
state	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
w1	6	6	0	0	0	0
w2	0	0	3.6	3.6	0	0
w3	0	0	0	0	9	9
risk free	2	2	2	2	2	2

Arbitrage opportunities? (As in the previous exercises, short selling remains.)

Exercise 4.9 *Compare and contrast horse race betting and trading in credit default swaps. Is the same mathematics descriptive and/or illuminating? What about social welfare implications?*

Exercise 4.10 *Here is a three security, three state setting in state-act-outcome format. The price of the call, c, is unspecified.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	*1*	*1.2*	*1.2*	*1.2*
stock	*70*	*36*	*78*	*108*
call	*c?*	*0*	*0*	*48*

Compute the no arbitrage equilibrium price of the call, as well as the equilibrium state prices and the risk neutral state probabilities.

Suppose a call price of 4 is observed. Is it possible, trading the call at that price, to form an arbitrage portfolio? If so, specify the portfolio weights, setting the weight on the call to be plus or minus one. (Hint: compare with exercise 4.3.)

Exercise 4.11 *Here are partial financial statements.*

partial balance sheet

	ending	*beginning*
asset 1	*16*	*5*
asset 2	*10*	*20*

partial income statement

expense	*9*

And here are the journal entries.

asset 1	y_1	
cash		y_1
asset 2	y_2	
cash		y_2
expense	y_3	
cash		y_3
expense	y_4	
asset 1		y_4
expense	y_5	
asset 2		y_5

Let A be the incidence matrix describing the journal entries, and x the vector of changes in account balances. According to the theorem of the separating hyperplane, there exists either a λ or a non-negative y. Which is it? And compute an example of the vector that does exist.

5

accounting and equilibrium: valuation in the row space

The previous chapter identified a problem with a fair value approach to accounting disclosure. That is, information problems tend to create multiple equilibrium values for the same asset; choosing only one of those values ignores the complexity and subtlety of the setting, and can hardly enhance the information environment. It doesn't seem fair to register a complaint about accounting of a unique fair value without offering an alternative. In this chapter we explore some of the accounting complications, and offer some modest suggestions.

The first part of the chapter uses some standard accounting methods to value the derivative portfolio: historical cost and mark to market. These methods can cause problems. In particular, there is no assurance the reported value is consistent with arbitrage free pricing. In fact, it is likely not, as no equilibrium concept is in sight while doing the accounting. Indeed, if the portfolio were actually available to be bought or sold at the reported price, arbitrage profits would be available.

In the latter part of the chapter reporting is done consistent with arbitrage free pricing. Using the logic of the fundamental theorem of linear algebra, the derivative portfolio returns vector can be decomposed into two orthogonal components. One component can be constructed in the rows of the outcome matrix; that is, the row component consists of bond and stock returns. The row component price is unambiguous, consisting of priced stocks and bonds, and often is consistent with arbitrage free pricing.

The row component can be computed in two different ways. The derivative security returns can be projected into the stock and bond rows. Or any set of state prices can similarly be projected into the same rows to yield y_{row}. Then y_{row} is used to value the derivative portfolio. Both methods yield the same portfolio value, and, since accounting is all about checking, it is nice to have a check result.

The first method allows the derivative portfolio to be stated in risk-free (bond) and risky (stock) components. The second method verifies equilibrium, since the state prices are arbitrage free as long as they are non-negative, according to the fundamental theorem of finance. If y_{row} is *not* non-negative, then a little extra work is required to price the orthogonal residual component.

5.1 historical cost

Start with an example based on the two security-three state setting of the previous chapter.

		state outcomes		
price	security	θ_1	θ_2	θ_3
.8	bond	1	1	1
50	stock	25	50	100

The multiple equilibria state prices were computed in the previous chapter.

$$y = \begin{bmatrix} .4 \\ 0 \\ .4 \end{bmatrix} \text{ or } y = \begin{bmatrix} 0 \\ .6 \\ .2 \end{bmatrix}$$

or any linear, non-negative combination.

Example 5.1 *Ralph runs a trading operation in Arrow-Debreu securities: denote them a_1, a_2, a_3 for the three possible states. Ralph's trading activity is summarized in the tables.*

Purchases	:	
400 a_1 @ .3	=	120
650 a_2 @ .5	=	325
1200 a_3 @ .3	=	360

Sales	:	
50 a_1 @ .4	=	20
50 a_2 @ .6	=	30
100 a_3 @ .4	=	40

All the sales occurred on the last day of the trading period; the purchases were made before that. Notice all the trading prices are within the no arbitrage equilibrium bounds.

The question is the standard accounting issue: what is the appropriate balance sheet valuation for the Arrow-Debreu portfolio? Once that is determined, of course, income measurement is simultaneously decided. The data admit of three distinct valuation methods.

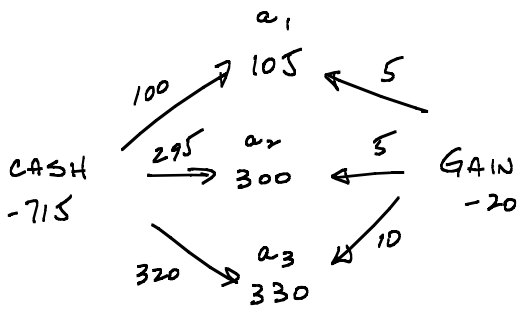


FIGURE 5.1
HISTORICAL COST ACCOUNTING
FOR ARROW-DEBREU SECURITIES

A historical cost balance sheet is straightforward: record each asset at its acquisition cost.

$$\begin{aligned} 350\ a_1\ @\ .3 &= 105 \\ 600\ a_2\ @\ .5 &= 300 \\ 1100\ a_3\ @\ .3 &= 330 \\ \text{total assets} &: 735 \end{aligned}$$

Correspondingly, income is recognized as the amount realized upon sale. Each security is sold for .1 more than the acquisition price, so income is calculated.

$$\begin{aligned} 50\ a_1\ @\ .1 &= 5 \\ 50\ a_2\ @\ .1 &= 5 \\ 100\ a_3\ @\ .1 &= 10 \\ \text{total gain} &: 20 \end{aligned}$$

The numbers must all fit together in debit credit mechanics. One possible way to incorporate cash, inventory, and income effects is in a directed graph, as in figure 5.1. Note particularly the reported valuation of the derivative securities under historical cost: 735. The question arises: Is this an equilibrium number? Before bringing equilibrium into view, consider a mark to market accounting variation.

5.2 mark to market

"Mark to market" valuation is accomplished by recognizing unrealized gains and losses on inventory items by revaluing the inventory to an observed market price. Here the most recent observed market prices are the prices associated with the

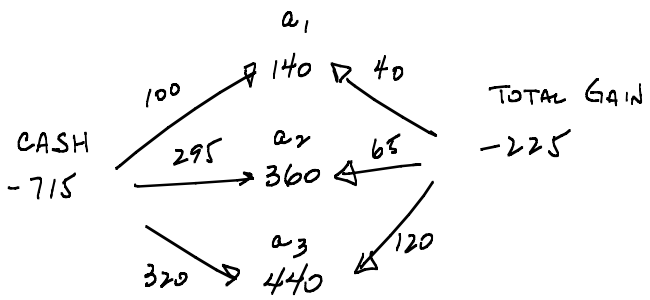


FIGURE 5.2
MARK TO MARKET ACCOUNTING
FOR ARROW-DEBREU SECURITIES

sales. The inventory valuation, then, is as follows.

$$\begin{aligned} 350\ a_1\ @\ .4 &= 140 \\ 600\ a_2\ @\ .6 &= 360 \\ 1100\ a_3\ @\ .4 &= 440 \\ \text{total assets} &: 940 \end{aligned}$$

As far as the income statement is concerned, there are two kinds of gains and losses: realized and unrealized. The realized gain is as with historical cost, the gain on the securities sold, here 20. The unrealized gains are the gains due to the revaluation of inventory totaling 205.

$$\begin{aligned} 350\ a_1\ @\ .1 &= 35 \\ 600\ a_2\ @\ .1 &= 60 \\ 1100\ a_3\ @\ .1 &= 110 \\ \text{unrealized gains} &: 205 \end{aligned}$$

Total gain is 225, and everything fits together as in the directed graph. The valuation number of interest is, under this method, 940. Notice that for both methods so far, there is no arbitrage or equilibrium concept in view. Equilibrium is the topic of the next section.

5.3 row space valuation

The question is whether or not we have succeeded in putting equilibrium values on the balance sheet. And, so far, it seems unlikely. There is another way to pose the question: if one could buy and sell portfolios at the balance sheet reported values, are there arbitrage opportunities? To address the question, we can write

down the Arrow-Debreu portfolio state outcomes, denoted by a below As they are Arrow-Debreu securities, the state outcomes are simply the number of the respective securities in inventory. Also included are the prices and state outcomes for the bond and the stock.

Example 5.2

x (price)		θ_1	θ_2	θ_3
.8	*bond*	1	1	1
50	*stock*	25	50	100
?	a	350	600	1100

Notice, for this example, the set of outcomes can be reproduced exactly with a portfolio of stocks and bonds, that is, in the rows of the state-act-outcome matrix. The portfolio weights are 100 for the bonds and 10 shares of stock.

	outcomes			
λ (weights)	θ_1	θ_2	θ_3	price
100 bonds	100	100	100	80
10 shares	250	500	1000	500
total	350	600	1100	580

The portfolio weights, λ, can be solved in the state equations which reproduce the Arrow-Debreu portfolio outcomes, a.

$$\begin{aligned} 1\lambda_1 + 25\lambda_2 &= 350 \\ 1\lambda_1 + 50\lambda_2 &= 600 \\ 1\lambda_1 + 100\lambda_2 &= 1100 \end{aligned}$$

Even though there are three equations, the two portfolio weights satisfy all three equations.

$$\lambda = \begin{bmatrix} 100 \\ 10 \end{bmatrix}$$

There is no ambiguity (multiple equilibrium problem) as far as the stocks and bonds are concerned. The no arbitrage equilibrium price to generate the noted state outcomes is 80 for the 100 bonds and 500 for 10 shares of stock. The conclusion is any value for the a portfolio other than 580 is not an equilibrium. In other words, arbitrage opportunities are available.

Suppose, for example, the a portfolio could be sold for some amount greater than 580, say for 735 as in historical cost reporting. Then buying 100 bonds and 10 shares of stock, along with selling the a portfolio is a perfect hedge, meaning the outcome in every state is zero. And arbitrage results as the price of the total portfolio is

$$580 - 735 = -155 < 0$$

The "hedger" is paid to hold the portfolio.

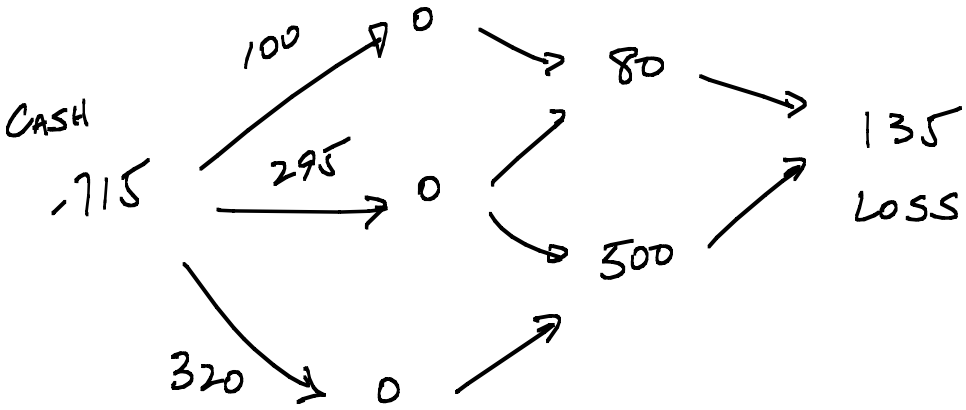


FIG. 5.3
NO ARBITRAGE VALUATION
FOR ARROW-DEBREU SECURITIES

To see the arbitrage opportunity in terms of the fundamental theorem of finance, augment the A matrix by the a portfolio outcomes, and x by the presumably available price of a, 735.

λ	x		θ_1	θ_2	θ_3
100	.8	bond	1	1	1
10	50	stock	25	50	100
-1	735	a	350	600	1100

It is easy to verify the portfolio weight vector, λ, results in a negative portfolio price, and a minimum return of zero in all states, that is, arbitrage profits.

$$\lambda^T x = \begin{bmatrix} 100 \\ 10 \\ -1 \end{bmatrix}^T \begin{bmatrix} .8 \\ 50 \\ 735 \end{bmatrix} = -155 < 0$$

$$A^T\lambda = \begin{bmatrix} 1 & 1 & 1 \\ 25 & 50 & 100 \\ 350 & 600 & 1100 \end{bmatrix}^T \begin{bmatrix} 100 \\ 10 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The only accounting valuation consistent with arbitrage free pricing is 580. One way to accomplish this in the accounting is to combine the three Arrow-Debreu securities into two equivalent bond and stock securities valued at 80 and 500, respectively. The purchase and recombination are depicted in the figure 5.3 directed graph.

To make the directed graph balance, a loss of 135 is recognized. (Notice the historical cost of the Arrow-Debreu securities implies a loss of 155, calculated

above. The 20 loss reduction arises from the realized gain of 20 from the cash sale.) The financial statement presentation has two line items on the balance sheet: one for the bond component of the outcome vector, and the other for the stock component. The bond component can be termed "risk-free" and the stock "risky," leading to the following balance sheet representation. Since the bond and stock are the rows of the state outcome matrix, this is valuation in the row space.

risk-free component of derivative portfolio	80
risky component of derivative portfolio	500

The numbers were chosen so that mark to market exaggerated the distortion from equilibrium. However, it is important to notice mark to market valuation is not necessary for the distortion to occur; distortions also occur with historical cost accounting. Multiple equilibrium prices is what causes the trouble.

5.4 null space valuation zero

In the previous example the Arrow-Debreu securities outcomes could be exactly reproduced by stocks and bonds. In other words, the Arrow-Debreu securities were positioned in the space of stocks and bonds, capable of being formed by a linear combination of stock and bond outcome vectors. That is, the derivative outcome portfolio vector, a, resides entirely in the rows of A where A is composed of the stock and bond outcome vectors. This need not always be the case, of course. Suppose there is a nullspace component: that is, a component, a_n, that is orthogonal to the rows of A.

$$\begin{aligned} a &= a_{\text{row}} + a_n \\ Aa_n &= 0 \end{aligned}$$

The question is how to compute an arbitrage free valuation in this case.

Two methods yield the same valuation result, and they are both valuations in the row space of A. That is, the null component, a_n, is not priced. The first method projects a into the rows of A. The resulting projection coefficients, β, are used to decompose a into bond and stock components, and the resulting valuation is $\beta^T x$. While allowing a risk-free and risky decomposition of the derivative portfolio, this method does not determine whether the row space valuation is an arbitrage free equilibrium. The second method can so determine.

The second method computes y_{row} by projecting any allowable state price vector, y, into the rows of A. y_{row} is then used to price the outcome vector, a: that is, compute the portfolio value $a^T y_{\text{row}}$. The computed valuation is the same as in method one - yielding a nice computational check.[1]

$$\beta^T x = a^T y_{\text{row}}$$

[1] For a derivation of this equality, see exercise 5.12.

In addition, according to the fundamental theorem of finance, since y_{row} solves $Ay = x$, y_{row} is an arbitrage free state pricing vector as long as

$$y_{\text{row}} \geq 0$$

The next example illustrates the case where the row space valuation is an equilibrium.

Example 5.3 *Use the data in the previous example, but this time assume no sales of Arrow-Debreu securities.*

		state outcomes		
price	*security*	θ_1	θ_2	θ_3
.8	*bond*	1	1	1
50	*stock*	25	50	100

$$\begin{aligned} \textit{Purchases} \quad & : \\ 400\ a_1\ @\ .3 \quad &= \quad 120 \\ 650\ a_2\ @\ .5 \quad &= \quad 325 \\ 1200\ a_3\ @\ .3 \quad &= \quad 360 \\ \textit{total} \quad &= \quad 805 \end{aligned}$$

As before denote the state returns for the Arrow-Debreu portfolio as a.

$$a = \begin{bmatrix} 400 \\ 650 \\ 1200 \end{bmatrix}$$

The row component of the a vector is computed by projecting (regression) into stocks and bonds, that is, the rows of A. The usual orthogonality conditions.

$$AA^T\beta = Aa$$

And here the two equations are

$$\begin{aligned} 3\beta_1 + 175\beta_2 \quad &= \quad 2250 \\ 175\beta_1 + 13,125\beta_2 \quad &= \quad 162,500 \end{aligned}$$

which imply the portfolio weights, β.

$$\begin{aligned} \beta_1 \quad &= \quad 125 \\ \beta_2 \quad &= \quad 10\frac{5}{7} \end{aligned}$$

That is, the "nearest" row portfolio consists of 125 bonds and $10\ 5/7$ shares of stock.

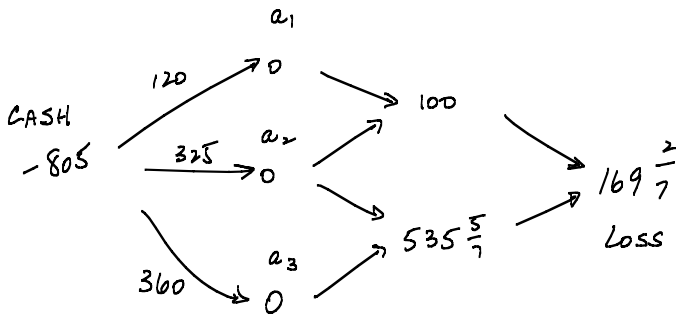


FIGURE 5.5
EXAMPLE 5.3 DIRECTED GRAPH

For balance sheet purposes the (no) arbitrage value of the A-D portfolio is

risk-free component of derivative portfolio	$125(.8) = 100$
risky component of derivative portfolio	$10\frac{5}{7}(50) = 535\frac{5}{7}$
total	$635\frac{5}{7}$

As the cost of the portfolio was 805, the unrealized loss is computed as

$$805 - 635\frac{5}{7} = 169\frac{2}{7}$$

The accounting numbers are presented in directed graph format in figure 5.5

There remains a question, however. Is the row space valuation so computed an arbitrage free equilibrium? The question is resolved by the second method: using y_{row} to price the derivative portfolio. Absent computational errors, the portfolio valuation will be the same as the computations for method one. Also, from the fundamental theorem of finance, if y_{row} is non-negative, it is a state price vector which does not allow for the formation of arbitrage portfolios. Hence, the valuation is arbitrage free

The orthogonality conditions to project y into the rows of A, and used to solve for β, are

$$AA^T\beta = Ay$$

It doesn't matter which y we choose for the computation, since for any y

$$Ay = x = \begin{bmatrix} .8 \\ 50 \end{bmatrix}$$

The orthogonality equations are

$$\begin{aligned} 3\beta_1 + 175\beta_2 &= .8 \\ 175\beta_1 + 13,125\beta_2 &= 50 \end{aligned}$$

which yield

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1/5 \\ 1/875 \end{bmatrix}$$

$$y_{\text{row}} = A^T\beta = \begin{bmatrix} 1 & 25 \\ 1 & 50 \\ 1 & 100 \end{bmatrix} \begin{bmatrix} 1/5 \\ 1/875 \end{bmatrix} = \frac{1}{35}\begin{bmatrix} 8 \\ 9 \\ 11 \end{bmatrix}$$

The first thing we notice is that y_{row} is non-negative, so using y_{row} to price the derivative portfolio ensures arbitrage free pricing. We should also check to see the pricing so computed gives us the same answer as method one, projecting a into the rows. Indeed, it does

$$a^T y_{\text{row}} = \begin{bmatrix} 400 \\ 650 \\ 1200 \end{bmatrix}^T \begin{bmatrix} 8 \\ 9 \\ 11 \end{bmatrix} \frac{1}{35} = 635\frac{5}{7}$$

5.5 null space valuation non-zero

We should do one more example, one in which zero is not an arbitrage free value for the null space component, a_n. In that case, the row component would not be arbitrage free, and a non-zero amount for the null component must be included in the balance sheet for arbitrage free pricing.

Example 5.4

x		θ_1	θ_2	θ_3
1	$bond$	1	1	1
$\frac{1}{5}$	$stock$	0	1	2
?	a	2	1	1

Notice for this example, the bond equation guarantees

$$\sum_i y_i = 1$$

Therefore, the state prices and the risk neutral probabilities are identical. For the computation of y_{row}, the orthogonality conditions are

$$AA^T\beta = Ay = x$$

where

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \text{ and}$$

$$x = \begin{bmatrix} 1 \\ \frac{1}{5} \end{bmatrix}$$

The orthogonality equations are

$$\begin{aligned} 3\beta_1 + 3\beta_2 &= 1 \\ 3\beta_1 + 5\beta_2 &= \frac{1}{5} \end{aligned}$$

which yield

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 11/15 \\ -2/5 \end{bmatrix}$$

Computing y_{row}

$$y_{\text{row}} = A^T\beta = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 11/15 \\ -2/5 \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 11 \\ 5 \\ -1 \end{bmatrix}$$

The negative number in the third element of y_{row} causes the problem according to the fundamental theorem of finance.[2] Our inclination to price out a using y_{row} might not be arbitrage free.

$$a^T y_{\text{row}} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 11 \\ 5 \\ -1 \end{bmatrix} \frac{1}{15} = \frac{26}{15} = 1.7\overline{33}$$

The question is whether the row space valuation is too high or too low to be arbitrage free. The *non-negative* state prices, or probabilities, satisfying $Ay = x$ can be used to establish bounds on the arbitrage free prices of the derivative portfolio.

$$\begin{aligned} 1y_1 + 1y_2 + 1y_3 &= 1 \\ 0y_1 + 1y_2 + 2y_3 &= \frac{1}{5} \end{aligned}$$

Setting $y_2 = 0$, we find one of the allowable y vectors.

$$y^1 = \frac{1}{10} \begin{bmatrix} 9 & 0 & 1 \end{bmatrix}^T$$

Similarly, setting $y_3 = 0$

$$y^2 = \frac{1}{10} \begin{bmatrix} 8 & 2 & 0 \end{bmatrix}^T$$

[2] Note there might be other y vectors which solve $Ay = x$. This issue is explored some in exercise 5.14.

Using y^1 and y^2 to price out the derivative portfolio, we find the arbitrage free bounds determined by the fundamental theorem of finance.

$$\begin{aligned} a^T y^1 &= \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 9 \\ 0 \\ 1 \end{bmatrix} \frac{1}{10} = \frac{19}{10} \quad \text{(upper bound)} \\ a^T y^2 &= \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 8 \\ 2 \\ 0 \end{bmatrix} \frac{1}{10} = \frac{18}{10} \quad \text{(lower bound)} \end{aligned}$$

So the row space valuation of $1.7\overline{33}$ is too low. In order to get the valuation in the range, an amount must be attached to the nullspace component of a: $1/15$ will do the trick.

derivative valuation:	
rowspace	$\frac{26}{15}$
nullspace	$\frac{1}{15}$
total	$\frac{27}{15} = 1.8$

1/15 is the minimum amount assigned to the nullspace in order to get the valuation within the arbitrage free bounds between 1.8 and 1.9.

It is also useful to check, once again, our understanding of the fundamental theorem of finance. According to the theorem, if the derivative portfolio is available for trading at any price outside the bounds of $1\frac{4}{5}$ and $1\frac{9}{10}$, then there exists a portfolio which generates arbitrage profits. Suppose the derivative portfolio is available at the row component price, $\frac{26}{15}$. We should be able to find portfolio weights, λ, which satisfy

$$\begin{aligned} A^T\lambda &\geq 0 \quad \text{and} \\ \lambda^T x &< 0 \end{aligned}$$

where A is now augmented by the derivative portfolio returns and x by the portfolio price.

λ	x	θ_1	θ_2	θ_3
-2	1	1	1	1
1	$1/5$	0	1	2
1	$26/15$	2	1	1

It is easily verified that $\lambda = \begin{bmatrix} -2 & 1 & 1 \end{bmatrix}^T$ satisfies the conditions of the theorem.

$$\begin{aligned} \lambda^T x &= \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 1 \\ 1/5 \\ 26/15 \end{bmatrix} = -\frac{1}{15} < 0 \\ A^T\lambda &= \begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \geq 0 \end{aligned}$$

The arbitrage portfolio consists of a sale of two bonds, purchase of one share of stock, and purchase of the derivative portfolio at a price of $\frac{26}{15}$. The investor is paid $\frac{1}{15}$ to hold the portfolio and, no matter what state occurs, the return is at least zero.[3] According to the theorem, an arbitrage portfolio can be formed whenever it is possible to trade the derivative portfolio outside the arbitrage free pricing bounds of 1.8 and 1.9.

In order to solve for the λ portfolio weights vector, first notice that, as the derivative portfolio is undervalued at $\frac{26}{15}$, the appropriate strategy is to buy, that is, set $\lambda_3 = 1$. Then notice the nearest equilibrium valuation is 1.8, which uses the y^2 probability vector to price out the derivative portfolio.

$$y^2 = \begin{bmatrix} 8 \\ 2 \\ 0 \end{bmatrix} \frac{1}{10}$$

y^2 puts no weight on state θ_3, so it seems sensible (or at least a good try) to use states 1 and 2 only in the $A^T\lambda$ condition as equalities:

$$\begin{aligned} \theta_1 &: 1\lambda_1 + 0\lambda_2 + 2\lambda_3 = 0 \\ \theta_2 &: 1\lambda_1 + 1\lambda_2 + 1\lambda_3 = 0 \end{aligned}$$

Using $\lambda_3 = 1$, and solving yields the following λ which, indeed, works, as was shown.

$$\lambda = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

A question naturally arises as to how, if at all, these valuation methods can be applied in the world of affairs. An easy, probably too easy, first step is to use historical data, if available, of the derivative, or derivative portfolio, returns. Running a regression of derivative returns on market returns yields a risk-free and risky decomposition of the derivative valuation, that is, the row component. As we have noted, the row component might not be an arbitrage free equilibrium. However, the greater danger is the implicit assumption that future returns are well predicted by historical data. It is easy to find stories where this approach has led to unhappy endings. There is no substitute for careful thinking about the structure and environment surrounding the construction of derivative securities.

5.6 summary

The chapter is about accounting valuation of derivative securities in a multiple equilibrium setting. Historical cost and mark to market accounting techniques

[3] Note that 1/15 is the amount of underpricing of the derivative relative to arbitrage free valuation.

are not, in general, consistent with (no) arbitrage equilibrium valuation. To guarantee valuation consistent with arbitrage pricing, the orthogonality tool, from the fundamental theorem of linear algebra, is utilized: the derivative portfolio is decomposed into row (bond and stock) and residual components. The mechanics of the decomposition process are identical to decomposing y_p into y_{row} and y_n from chapter 3.

The (no) arbitrage equilibrium value of the row component is unique. However, there is ambiguity in the orthogonal null component. Often zero lies in the arbitrage free pricing bound for the null component, so it can be ignored for pricing purposes. Sometimes, though, the null component requires a non-zero price to maintain arbitrage free pricing in the financial reporting.

It is a property of arbitrage free pricing that it is free of state probabilities. That is, there is no need to specify state probabilities when computing equilibrium state prices. That is a powerful, and somewhat surprising, feature of the theory. Perhaps it is also surprising, then, that it is possible, in some cases, to infer state probabilities from observed state prices, as we will see in chapter 8.

5.7 exercises

Exercise 5.1 *Here is a three state - two security setting.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	*1*	*1*	*1*	*1*
stock	*2*	*1*	*2*	*3*

The following purchases of Arrow-Debreu securities occur.

$$\begin{aligned} Purchases & : \\ 800\ a_1 \text{ @ } .25 &= 200 \\ 900\ a_2 \text{ @ } .5 &= 450 \\ 1600\ a_3 \text{ @ } .25 &= 400 \end{aligned}$$

For balance sheet presentation decompose the A-D securities holdings into risk-free (bond) and risky(stock) components.

Exercise 5.2 *Refer to the set-up in exercise 5.1. Are the observed exchange prices of the Arrow-Debreu securities consistent with a (no) arbitrage equilibrium? Are the row components (risk-free and risky) of the Arrow-Debreu portfolio holdings an arbitrage free price for the entire portfolio? Suppose the reported value of the derivative portfolio is 1400. Is this value arbitrage free? If not, form the arbitrage portfolio with the weight on the derivative portfolio of +1 or -1.*

Exercise 5.3 *Refer again to the set-up in exercise 5.1. Suppose only one a_1 is purchased for $.25. Decompose a_1 into risk-free and risky components.*

Exercise 5.4 *Here is the regular three state - two security setting.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	*1*	*1*	*1*	*1*
stock	*2*	*1*	*2*	*3*

The following purchases and sales of Arrow-Debreu securities occur; a_i.

$$\begin{aligned} \textit{Purchases} \quad & : \\ 100\ a_1 \text{ @ } .3 \quad &= \quad 30 \\ 100\ a_2 \text{ @ } .5 \quad &= \quad 50 \\ 100\ a_3 \text{ @ } .3 \quad &= \quad 30 \end{aligned}$$

$$\begin{aligned} \textit{Sales} \quad & : \\ 10\ a_1 \text{ @ } .4 \quad &= \quad 4 \\ 10\ a_2 \text{ @ } .6 \quad &= \quad 6 \\ 10\ a_3 \text{ @ } .4 \quad &= \quad 4 \end{aligned}$$

Using historical cost accounting, prepare a balance sheet inventory valuation for the portfolio valuing each A-D security separately.

Exercise 5.5 *Using the same data as in the previous exercise prepare market value accounting statements, where the market value is the most recent observed exchange price. All the security sales occurred after the purchases, and all the exchange prices were publicly observed. What is the balance sheet inventory valuation for the portfolio, valuing each A-D security separately?*

Exercise 5.6 *Use the same data in exercise 5.4 one more time. This time prepare a (no) arbitrage equilibrium balance sheet valuation for the derivative portfolio. Report two types of security assets (if necessary), risky and risk-free security.*

Exercise 5.7 *Here is the regular three state - two security setting with some purchases of Arrow-Debreu securities.*

		outcomes		
security	*price*	θ_1	θ_2	θ_3
bond	1	1	1	1
stock	5	2	4	9

$$\begin{aligned} Purchases \quad & : \\ 100\ a_1 \text{ @ } .25 \quad &= \quad 25 \\ 100\ a_2 \text{ @ } .40 \quad &= \quad 40 \\ 200\ a_3 \text{ @ } .25 \quad &= \quad 50 \end{aligned}$$

Using the (no) arbitrage equilibrium concept, report risk-free (bond) and risky (stock) components of the A-D portfolio. What is the balance sheet valuation of the securities portfolio?

Exercise 5.8 *Suppose in the previous exercise the financial statement valuation is done using historical cost accounting. What is the balance sheet valuation? Is the valuation arbitrage free? If not, specify portfolio weights for an arbitrage portfolio, restricting the weight on the Arrow-Debreu portfolio to be +1 or -1.*

Exercise 5.9 *Use the familiar state price set-up:*

		outcomes		
security	x	θ_1	θ_2	θ_3
bond	.8	1	1	1
stock	50	25	50	100

Find y_{row} using quadratic programming (and Solver).

Exercise 5.10 *Verify (or derive) the null space of the A matrix in the previous exercise is any scalar multiple of*

$$\begin{bmatrix} 1 \\ -3/2 \\ 1/2 \end{bmatrix}$$

Find y_{row} by projecting any y_p satisfying $Ay = x$ into the null space, and subtracting.

Exercise 5.11 *Revisit exercise 5.3. Suppose the derivative portfolio is valued at 620. Is this within the arbitrage free bounds? If not, find the λ weights for an arbitrage portfolio, where λ_3 is either plus or minus one.*

Exercise 5.12 *The two methods of the chapter for valuing a derivative portfolio always yield the same number. For projecting a into the rows of A, the orthogonality conditions are*

$$AA^T\beta = Aa$$

So the β coefficients can be written

$$\beta = \left(AA^T\right)^{-1} Aa$$

There is a linear algebra relationship that the transpose of a product is equal to the product of the transposes in reverse order. That is, for any matrices A and B

$$(AB)^T = B^T A^T$$

Using the relationship we have

$$\beta^T = a^T A^T \left(AA^T\right)^{-1}$$

The total value of the row component can, then, be written

$$\beta^T x = a^T A^T \left(AA^T\right)^{-1} x$$

Show the same expression is obtained when y is projected into the rows of A, and the resulting y_{row} is used to price the derivative portfolio:

$$a^T y_{\text{row}}$$

Exercise 5.13

		outcomes		
security	x	θ_1	θ_2	θ_3
bond	1	1	1	1
stock	1.8	0	1	2
a	*?*	1	1	0

What is the valuation of a in the row space of stocks and bonds? Is the row space value an arbitrage free equilibrium? If not, specify the portfolio weights for an arbitrage portfolio with λ_3, the weight on a, either plus or minus one. Also in the case that the row space value is not arbitrage free, make the minimum necessary adjustment to the null space value so the total reported value is arbitrage free.

Exercise 5.14 *Redo exercise 5.13 with the derivative outcomes*

$$a = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

Exercise 5.15 *Consider 3 securities and 3 states with the following state -act-outcome matrix and prices.*

prices		θ_1	θ_2	θ_3
1	*bond*	1	1	1
50	*stock*	25	50	100
?	*derivative*	1	2	3

Compute a value for the derivative by projecting into the bond and stock rows. Is the projection an equilibrium? If not, find the equilibrium nearest to the projection. Do the same, only this time set the stock price to 35. And once again, this time with a stock price of 91.

6
accounting stocks and flows - certainty

To this point we have confined our attention to static properties of the double entry system. An important part of accounting, however, is how the numbers march through time as events unfold. The passage of time is necessary if we are to study information. In this chapter there is no uncertainty, and, hence, no information. The discussion of information and accounting is forthcoming in later chapters, and is set up here with the passage of time, but without information. An important aspect of accounting is that balance sheet accounts remain open: the account balance at the end of one period is the beginning balance for the next period. In this chapter we will track the changes in account balances as time passes. Along the way some useful tools will be acquired.

The chapter begins with a discussion of the time value of money, which allows presentation of economic income and economic amortization. Particular attention is paid to continuous compounding of interest. The method is widely used, as it finesses the compounding period problem. But more important to us is the theoretical importance of continuous compounding when information is in view. Notably, continuous compounding is a central part of the mutual information theorem in chapter 8.

Other methods of amortization are considered: straight line, sum of the years' digits, and declining balance methods. In all cases attention is paid to the calculation of income and balance sheet effects.

The first set of examples are restricted to consideration of the life cycle of one asset. Typically, however, firms replenish long lived assets (and liabilities) on a regular basis. The latter part of the chapter examines accounting when old depleted assets are routinely replaced with new ones, referred to as a "steady state" situation. As usual particular attention is paid to balance sheet amounts and in-

come effects. The general result is, in steady state, income effects are independent of amortization method, while balance sheet numbers are sensitive to the amortization method in use. The expressions of some standard series in mathematics are very useful for the calculations.

Attention is paid to the data compression abilities of accounting. The calculations can be accomplished even though only two numbers are kept in memory: a stock (balance sheet) number, and a flow (income statement) number. This property turns out to be useful once an information question is posed in the next chapter.

6.1 time value of money

A dollar today is not the same commodity as a dollar one year from now. The rate of exchange between the two commodities is determined by the interest rate. A convenient way to think about it is in terms of a bank account. $100 deposited now can be exchanged for $100(1+r) in one year, where r is the annual interest rate. For r = 10%, the bank account will grow to $110 in a year. If the bank account is left undisturbed for 2 years, it will grow to $100(1+r)(1+r) = 100(1+r)^2$. For r = 10%, the amount is $121.

The general relationship is

$$P_n = P_0 (1+r)^n \tag{6.1}$$

where P_0 is the initial amount in dollars at time 0, and P_n is the amount it grows to after n periods at r interest rate per period. If we wish to solve for P_0, the formula can be rearranged.

$$P_0 = P_n (1+r)^{-n} \tag{6.2}$$

Often we will be interested in a sequence of payments. Once each payment is discounted to time 0, the amounts can be summed. The result is called the present value of the payment sequence.

$$P_0 = \frac{P_1}{(1+r)} + \frac{P_2}{(1+r)^2} + \frac{P_3}{(1+r)^3} + ...$$

Or, more parsimoniously using summation notation

$$P_0 = \sum_{n=1}^{N} P_n (1+r)^{-n} \tag{6.3}$$

6.1.1 continuous compounding

Equation (6.3) works fine if all the if all the cash flows, and all the requisite time value computations, occur at the end of the interest period, so far presumed to be a

year. That is, interest is compounded at the end of each year. As a first step, suppose we require a time value computation every six months. Semi-annual (twice a year) compounding is accomplished by halving the interest rate and doubling the number of periods.

$$P_1 = P_0 \left(\frac{1+r}{2} \right)^2$$

For the previous example:

$$100 \left(1 + \frac{.1}{2} \right)^2 = 110.25.$$

Notice there is a slight increase (.25) relative to the annual calculation; that is the compounding effect. Quarterly compounding yields a smaller increment.

$$100 \left(1 + \frac{.1}{4} \right)^4 = 110.38$$

What is the limit to which this process can be taken: that is, how much interest would accrue if interest were compounded every second, or even instantaneously? The answer involves taking the limit as the number of compounding periods becomes very large.

$$\underset{t \to \infty}{\text{Limit}} \left(1 + \frac{r}{t} \right)^t = e^r \tag{6.4}$$

e is an important and magical number in the history of mathematics. It first arose as an answer to an interest, that is to say, an accounting problem: a source of pride to at least some accountants. The numerical value of e is approximately 2.71828. The existence of e in the set-up has a way of simplifying the computations (as we will encounter later). Continuous compounding has the added benefit of not requiring a specification of a separate compounding period to go with the interest rate.

For the problem at hand

$$100e^r = 100e^{\cdot 1} = 110.52$$

So the incremental increase due to compounding, in this problem, reaches an upper bound at .52. When a continuous compounding process lasts for several years, n, the general formula is

$$P_n = P_0 e^{rn} \tag{6.5}$$

An important thing about the previous equation is that n need not be an integer. That is, we can (and will) slice a year into any fraction, and get a useful time value of money amount at any time, any day or night.

The limit expression in equation (6.4) does not converge very quickly (try it in a spreadsheet), so some cleverness was required to get an arithmetic representation

prior to the invention of high speed calculators. One useful way was to transform the expression into an infinite series

$$e = \underset{t\to\infty}{\text{Limit}}\ 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + ... + \frac{1}{t!}$$

which converges a bit faster.

One more thing about the arithmetic approximation of e. We have access to the first ten digits by contemplating a $20 bill. The picture is Andrew Jackson, the 7th president of the United States,[1] and he was elected in 1828. Stringing the numbers together, we have, to a ten digit approximation

$$e = 2.718281828...$$

For a discussion of this and other facts about e, see Maor, 1994.

6.1.2 perpetuities and annuities

There are a couple of concepts which will make the analysis of financial statement effects a bit easier.

Definition 6.1 *A perpetuity is a sequence of constant cash flows that lasts forever.*

It is easy to figure out how much cash can be withdrawn from a bank account every year forever without reducing the bank balance: it is the amount of the interest.

$$P_0 r = CF$$

where CF is the cash flow withdrawal each year. Rearranging we have the present value of a perpetuity.

$$P_0 = \frac{CF}{r}$$

The continuous compounding relationship derives from the fact that the principal must grow to the sum of the cash flow and the original principal amount.

$$\begin{aligned} CF + P_0 &= P_0 e^r \\ CF &= P_0(e^r - 1) \end{aligned}$$

And the present value of a perpetuity is

$$P_0 = \frac{CF}{e^r - 1}$$

Definition 6.2 *An annuity is a sequence of constant cash flows for a finite number of periods, say n.*

---

[1] After Washington (elected in 1788), Adams, Jefferson, Madison, Monroe, and J. Q. Adams. All were two term presidents, save Adams father and son. Counting the years gives us Jackson's election date.

It is easy to write down the expression for the present value of an annuity for n periods: subtract the present value of a perpetuity that begins in n periods from the present of a perpetuity beginning now.

$$\begin{aligned} P_0 &= \frac{CF}{r} - \frac{CF}{r}\frac{1}{(1+r)^n} \\ P_0 &= \frac{CF}{r}\left(1-(1+r)^{-n}\right) \end{aligned} \tag{6.6}$$

The corresponding relationship for continuous compounding is

$$P_0 = \frac{CF}{e^r - 1}\left(1 - e^{-rn}\right)$$

6.1.3 economic income

Armed with time value of money tools we are ready to prepare financial statements, where we are especially interested in multi-period effects. Start with an example.

Example 6.1 *An asset generates cash flows for three years as follows.*

year 1	*year 2*	*year 3*
1100	*605*	*133.10*

To calculate a time 0 price, use an interest rate of 10%.

$$P_0 = \frac{1100}{1.1} + \frac{605}{1.1^2} + \frac{133.10}{1.1^3} = 1600$$

The accounting task at hand is to prepare financial statements for the three year life. Central to the task, of course, is the determination of a value of the asset (for the balance sheet) and income attributed to the asset (for the income statement). The first time through the exercise, we will calculate the two important numbers using economic earnings.

Definition 6.3 *economic earnings (income) is the calculation of financial statement numbers using the following time value of money techniques.*

- Asset value is equal to the present value of the remaining cash flow.
- Income is equal to the beginning asset value times the interest rate.

One way to get the asset value is simply to do the discounting calculations for each period. The asset value at time 0 is the market price, P_0, already calculated to be 1600.

$$P_1 = \frac{605}{1.1} + \frac{133.10}{1.1^2} = 660$$

$$P_2 = \frac{133.10}{1.1} = 121$$

There is an alternative calculation that does not require all the cash flow numbers, only the current balance in the asset account (a stock number) and the current cash flow (a flow number). In subsequent chapters we will do a variety of accounting calculations using only two numbers: a stock and a flow. Here treat the asset like a bank account: increase the account by the income (interest) and decrease the account by the cash flow (withdrawal from the bank account). For the first year

asset	
1600	
160	1100
660	

The income is calculated by multiplying the beginning balance times the interest rate. Succeeding periods are calculated similarly.

asset	
660	
66	605
121	
12.10	133.10
0	

In general

asset	
beg bal	
r times beg bal	cash flow
end bal	

Algebraically, and with more parsimonious notation, the updating equation looks like this.

$$BB + rBB - CF = EB \tag{6.7}$$

There are a couple more things to be learned from the updating equation. The equation can be rearranged to solve for the beginning balance.

$$BB = \frac{EB + CF}{1 + r}$$

Therefore, the beginning balance for year 3 is

$$\frac{0 + 131.10}{1.1} = 121 = P_2$$

Work backward one year at a time to find P_1 and P_0.

$$\frac{121 + 605}{1.1} = 660 = P_1$$

$$\frac{660 + 1100}{1.1} = 1600 = P_0$$

Look one more time at the updating equation and solve for income.

$$rBB = EB - BB + CF$$

Written this way, income is composed of two parts: a cash part and an accrual part (EB - BB).

Definition 6.4 *The non-cash change in the value of an asset or liability is called an accrual.*

Definition 6.5 *The process of making systematic accrual entries over time is called amortization.*

Definition 6.6 *Depreciation expense is the amortization process applied to tangible assets.*

The accrual numbers for the example (called depreciation, in this case) can be calculated by taking the difference in the beginning and ending balances. Depreciation in year 1 is

$$1600 - 660 = 940$$

For year 2 the depreciation expense is

$$660 - 121 = 539$$

And, finally, year 3 depreciation is

$$121 - 0 = 121$$

This allows construction of the economic income statements.

income statements

	year 1	year 2	year 3
cash revenue	1100	605	133.10
depreciation expense	940	539	121.00
income	160	66	12.10

There is one more important check on the numbers. The total (3 year) income is

$$160 + 66 + 12.10 = 238.10$$

This is the total increase in cash due to interest. Total cash inflow is

$$1100 + 605 + 133.10 = 1838.10$$

which exceeds the initial outflow of 1600 by 238.10. Total income over the life of the asset is always equal to total cash flow regardless of the amortization technique.

The same example can be done with continuous compounding. For the same cash flows, and the same nominal interest rate of 10%, the computed present value is slightly lower due to interest compounding occurring more often (actually continuously).

$$P_0 = \frac{1100}{e^{\cdot 1}} + \frac{605}{e^{\cdot 2}} + \frac{133.10}{e^{\cdot 3}} \simeq 1589.25$$

Using the same two rules for asset valuation and income determination, the financial statement numbers can be determined.

asset value on balance sheet:

	year 0	year 1	year 2	year 3
asset	1589.25	656.40	120.43	0

income statements:

	year 1	year 2	year 3
cash revenue	1100	605	133.10
depreciation	932.85	535.97	120.43
income	167.15	69.03	12.67

The total income for the three years is

$$167.15 + 69.03 + 12.67 = 248.85$$

which is also equal to the total cash flow less the present value of the asset.

$$1100 + 605 + 133.10 - 1589.25 = 248.85$$

6.2 alternative amortization techniques

With economic income, accounting income is calculated as r times the beginning balance. The accrual amount is determined from the income and cash flow amounts. An alternative amortization scheme is to specify the accrual amount first, then calculate income using the accrual and cash flow numbers.

6.2.1 straight line depreciation

One "calculate the accrual first" amortization method is straight line depreciation: simply divide the total amount to be depreciated (for our example, the cost of the asset, 1600) by the asset's useful life (here 3 years). Straight line depreciation in each of the three years is

$$\frac{1600}{3} = 533\frac{1}{3}$$

and the reported asset value and income statements follow.

asset value on balance sheet:

	year 0	year 1	year 2	year 3
asset	1600	$1066\frac{2}{3}$	$533\frac{1}{3}$	0

income statements:

	year 1	year 2	year 3
cash revenue	1100	605	133.10
depreciation	$533\frac{1}{3}$	$533\frac{1}{3}$	$533\frac{1}{3}$
income	$566\frac{2}{3}$	$71\frac{2}{3}$	(400.23)

Notice, once again, the total income is

$$566\frac{2}{3} + 71\frac{2}{3} - 400.23 = 238.10$$

the same total as with economic income. Indeed, any depreciation method with total three year depreciation of 1600 will result in total income of 238.10. Different depreciation methods merely rearrange income across periods; the total income over the life of the firm is invariant with respect to depreciation method choice.

6.2.2 sum of the years' digits depreciation

There exist accelerated depreciation methods in the sense that the depreciation expense is greater early in the life of the asset. For the sum of the years' depreciation method a depreciation rate is applied to the total depreciable value, and the rate declines over time. The denominator of the rate is the sum of the digits encountered when counting the life of the asset. For a 3 year life, the denominator is

$$1 + 2 + 3 = 6$$

In general, for an n year life the denominator is

$$\sum_{i=1}^{n} i = \frac{n\,(n+1)}{2} \tag{6.8}$$

It is not hard to derive the general expression, and, since we will use it later, let S equal the sum of the digits.

$$\begin{aligned} S &= 1 + 2 + 3 + ... + n \\ &= n + (n-1) + (n-2) + ... + 1 \\ 2S &= (n+1) + (n+1) + (n+1) + ... + (n+1) \\ &= n\,(n+1) \\ S &= \frac{n\,(n+1)}{2} \end{aligned}$$

The numerator of the depreciation rate is the year in reverse order. That is, the depreciation rate in the first year is $\frac{3}{6}$, the second year rate is $\frac{2}{6}$, and the third is $\frac{1}{6}$. And the depreciation expense for each year is calculated.

	year 1	year 2	year 3
depreciation expense	$\frac{3}{6}(1600) = 800$	$\frac{2}{6}(1600) = 533\frac{1}{3}$	$\frac{1}{6}(1600) = 266\frac{2}{3}$

And the sum of the years' digits income statements follow.

	year 1	year 2	year 3
cash revenue	1100	605	133.10
depreciation	800	$533\frac{1}{3}$	$266\frac{2}{3}$
income	300	$71\frac{2}{3}$	(133.57)

Notice that, once again, the total income is

$$300 + 71\frac{2}{3} - 133.57 = 238.10$$

6.3 steady state accounting

Typically firms don't behave as in the example: buy an asset, use it up, and then go out of business. There is usually a continuous investment in assets allowing the firm to stay in business for several periods, or even indefinitely. In order to explore the behavior of replenishing assets, augment the previous example so that the firm takes 1600 each year to invest in an asset with the same cash flow stream attached.

$$CF = [1100, 605, 133.10]$$

To keep the cash inputs and outputs separate, assume the asset purchases are made on January 1, and the cash inflows occur on December 31. The financials are prepared as of December 31 immediately subsequent to the cash inflow.

Steady state occurs when the number of assets is unchanged. For the example, after the third year the firm will have three assets, always acquiring and retiring one each year. The accounting questions are

- What is the steady state income?
- What is the steady state asset balance?

We'll attack each question using different depreciation methods, and introduce another accelerated method: declining balance depreciation. First calculate the steady state cash inflow. Once three machines are purchased, the cash inflow will consist of three parts (like Gaul). The newest machine will contribute 1100, the next newest 605, and the oldest machine 133.10. Hence, total cash inflow is 1838.10.

Reviewing the depreciation amounts for the various depreciation methods, similar calculations yield the steady state depreciation expense.

depreciation expense	newest	next	oldest	total
economic income	940	539	121	1600
straight line	$533\frac{1}{3}$	$533\frac{1}{3}$	$533\frac{1}{3}$	1600
sum of years' digits	800	$533\frac{1}{3}$	$266\frac{2}{3}$	1600

And the steady state income statement has a simple, and familiar, form.

income statement	
cash revenue	1838.10
depreciation	1600.00
income	238.10

The answer for income is straightforward: the steady state income number is the same regardless of amortization method. It would also be the same for cash basis accounting, by the way, as the income figure would be the net cash flow.

What about the reported asset value on the balance sheet? Is that also the same for the different depreciation methods? The answer turns out to be no. Let's explore. As a first step construct the asset T-accounts for the various depreciation methods, running the T-account out until the balance achieves steady state, that is, after three years of acquisitions. First do economic income.

asset-econ inc	
1600	
160	1100
660	
1600	
66	1100
160	605
781	
1600	1100
78.10	605
160	133.10
781	

The debits to the T-account are of two types: the 1600 purchase price and the accrual amount each year. The accrual amount is the interest rate times the balance in the account. As the balance in the account has two parts - the 1600 purchase and the beginning balance, the accrual entry has two parts each year. The credit entries are the cash flows, increasing each year until all three assets kick in, and the total cash flow is 1838.10. Since year 4, and all subsequent years, will be identical to year 3, the steady state asset amount, for economic income, is 781.

The T-accounts for the "calculate the accrual first" depreciation methods are a little simpler. The only debits are the beginning balance and the purchase, while

the credits are the depreciation expense amounts.

asset-straight line	
1600	$533\frac{1}{3}$
$1066\frac{2}{3}$	$533\frac{1}{3}$
1600	$533\frac{1}{3}$
1600	$533\frac{1}{3}$
1600	$533\frac{1}{3}$
	$533\frac{1}{3}$
1600	

asset-SYD	
1600	800
800	800
1600	$533\frac{1}{3}$
$1066\frac{2}{3}$	800
1600	$533\frac{1}{3}$
	$266\frac{2}{3}$
$1066\frac{2}{3}$	

It is evident that different depreciation methods imply different steady state asset amounts. For straight line the amount is 1600, while for SYD the steady state asset balance is $1066\frac{2}{3}$. To specify the different steady state amounts more generally, define some notation.

Let C be the cost of the asset

N: the useful life

$\sum CF$: the total cash flows generated over the life of the asset

B: steady state asset account balance

Using notation, the respective steady state asset balances follow.

	straight line	economic income	sum of years' digits
steady state asset balance	$C\left(\frac{N-1}{2}\right)$	$\frac{\sum CF - C(1+r)}{r}$	$C\left(\frac{N-1}{3}\right)$

We'll derive each expression in turn, saving economic income for last, as we'll use a trick there that will prove useful later.

6.3.1 straight line

To simplify the initial algebra, start with an asset with $C = 1$. The first few years of depreciation expense are $\frac{1}{N}$, $\frac{2}{N}$, $\frac{3}{N}$, and so forth, as one more asset is added each year. After N years steady state is attained; N assets have been purchased at a total cost of N. Calculate the steady state asset balance, B, by adding up

everything that happened in the T-account. The total credits are the total N year depreciation expenses: $\frac{1+2+3+...+N}{N}$. And the total debits are the asset purchases: N. The steady state balance is

$$B = N - \left(\frac{1 + 2 + 3 + ... + N}{N} \right)$$

We already know the sum of the first N integers from sum of the years digits:

$$\sum_{i=1}^{N} i = \frac{N\,(N+1)}{2}$$

Substituting

$$B = N - \frac{N\,(N+1)}{2N} = \frac{N-1}{2}$$

Finally, for an asset cost C, we have the general expression.

$$B = C\frac{N-1}{2} \tag{6.9}$$

6.3.2 sum of the years digits

Again start with $C = 1$. The first year of depreciation expense is the first year of the first asset purchased.

$$\frac{N}{\sum_{i=1}^{N} i}$$

For all the depreciation calculations, the denominator is the same. The numerator in the second year is $N-1$ for the first asset purchased plus N for the second asset. In the third year the numerator is $N-2$ plus $N-1$ plus N, and so forth. After N periods, and steady state is achieved, the total depreciation expense credited to the asset account is

$$\frac{N\,(N) + (N-1)\,(N-1) + (N-2)\,(N-2) + ... + 2\,(2) + 1\,(1)}{\sum_{i=1}^{N} i} = \frac{\sum_{i=1}^{N} i^2}{\sum_{i=1}^{N} i}$$

We won't derive the result here, but it is easy to verify that the sum of the squares of the first N digits is

$$\sum_{i=1}^{N} i^2 = \frac{N\,(N+1)\,(2N+1)}{6}$$

Substituting, we have

$$B = N - \frac{\sum_{i=1}^{N} i^2}{\sum_{i=1}^{N} i} = N - \frac{N(N+1)(2N+1)2}{N(N+1)6} = N - \frac{2N+1}{3} = \frac{N-1}{3}$$

The last step is to insert an arbitrary asset cost of C.

$$B = C\frac{N-1}{3} \tag{6.10}$$

6.3.3 economic income

With economic income it is easier to consider the asset T-account after steady state has been attained.

asset-econ inc	
.	
.	
.	.
B	
C	
$r(B+C)$	$\sum CF$
B	

The T-account equation is

$$B + C + r(B+C) - \sum CF = B$$

Solving for B yields the steady state relationship.

$$B = \frac{\sum CF - (1+r)C}{r}$$

A similar derivation gives the continuous compounding analog.

$$B = \frac{\sum CF - e^r C}{e^r - 1}$$

An alternative way to see the formula is to recall that for economic income the asset value is the discounted future cash flow. In steady state the net cash flow is the same every year:

$$\sum CF - (1+r)C$$

As the purchase amount, C, occurs at the beginning of the year, multiplying by $(1+r)$ transforms it into end of year dollars. Now simply apply the perpetuity formula.

The numbers in example 6.1 can be used to compute steady state financial statement numbers. For simple interest (annual compounding)

$$\begin{aligned} B &= \frac{\sum CF - (1+r)\,C}{r} \\ &= \frac{1838.10 - 1.1\,(1600)}{.1} = 781 \end{aligned}$$

which was computed earlier using the asset T-account.

For continuous compounding we have already computed the present value of the cash flows.

$$C = \frac{1100}{e^{.1}} + \frac{605}{e^{.2}} + \frac{133.10}{e^{.3}} \simeq 1589.26$$

The steady state asset account balance, then, is

$$\begin{aligned} B &= \frac{\sum CF - e^r C}{e^r - 1} \\ &= \frac{1838.10 - 1589.26\left(e^{.1}\right)}{e^{.1} - 1} \simeq 776.79 \end{aligned}$$

And the steady state income statement.

income statement	
cash revenue	1838.10
depreciation	1589.26
income	248.84

Note the income number checks with the interest calculations. Interest is the beginning balance in the asset account plus C multiplied times $e^r - 1$.

$$(776.79 + 1589.25)\left(e^{.1} - 1\right) = 248.84$$

6.3.4 declining balance depreciation methods

Another class of depreciation methods, called declining balance, calculates depreciation by multiplying the balance in the asset account by a rate that is unchanging over time. In the 3 year life example, a common declining balance rate is $\frac{2}{3}$; often the rate is twice the straight line rate, called double declining balance.

Use the $\frac{2}{3}$ rate in the numerical example and try to identify the steady state asset account balance. After the first few periods, the asset account looks as follows.

declining balance asset	
1600	$1066\frac{2}{3}$
$533\frac{1}{3}$	
1600	$1422\frac{2}{9}$
$711\frac{1}{9}$	
1600	1540.74
770.37	

It is not exactly clear where the asset balance and depreciation expense numbers are going. To see the steady state numbers, use the trick from the economic income derivation: assume the process has progressed far enough to converge.[2] Set up the steady state T-account in notation, where D is the depreciation rate.

asset declining balance	
.	
.	
.	.
B	
C	$D(B+C)$
B	

The T-account equation is

$$B + C - D(B + C) = B$$

and solving for B is straightforward.

$$B = C\frac{(1-D)}{D} \tag{6.11}$$

Revisiting the numerical example, it can be verified that, after 8 - 10 years, the asset balance gets pretty close to 800. Check this in a spreadsheet. Also notice the depreciation expense converges to 1600. In notation, the depreciation expense is

$$\begin{aligned} & D(B+C) \\ = \; & D\left[C\left(\frac{1-D}{D}\right)+C\right] \\ = \; & C(1-D)+CD \\ = \; & C \end{aligned}$$

So in steady state the income number (cash revenue less depreciation) is $\sum CF - C$, just as it is for all the other depreciation methods. The analysis of declining balance depreciation adds another data point to the conclusion: steady state income is independent of the depreciation method, but the steady state balance sheet amounts do depend on the method.

[2] Of course, this assumes the process does, indeed, converge, but for any D<1, the assumption holds, as can be verified.

6.4 summary

After doing the accounting for a single asset and a variety of depreciation methods, we explored the situation where a firm routinely replaces depleted assets. Some regularities were noted, in particular, while the steady state income statement does not depend on depreciation method, the balance sheet does.

The steady state calculations follow from the ability to predict all future cash flows with certainty. The question naturally arises as to what accounting looks like in a world of uncertainty. In particular, can the accounting numbers serve as good predictors of what will happen next? That is the question addressed in the next two chapters.

6.5 reference

Maor, Eli, *e: The Story of a Number*. Princeton University Press, 1994.

6.6 exercises

Exercise 6.1 *Here are the annual cash flows generated by an asset.*

CF_1	CF_2	CF_3
180	*470*	*330*

The market rate of interest is 10% per year. The asset is purchased for 800 on January 1 of the first year. The cash flows are received on December 31 of the respective years. The financial statements are prepared as of December 31 after the cash is received.

For the single asset calculate the reported asset value on the balance sheet for the first two years. Use three different amortization methods: economic income, straight line, and declining balance with a rate of $\frac{1}{3}$.

Also calculate the income amounts for all three years using the same amortization methods. For declining balance a convention is to depreciate the entire remaining balance in the last year; for simplicity use that convention here. (Note the convention does not come into play when assets are routinely replaced; see the subsequent exercise.)

Exercise 6.2 *This is a continuation of the previous exercise. Now presume an identical asset is purchased every January 1 for 800. Compute the balance sheet and income statement amounts after steady state occurs, that is, when the amounts don't change from year to year, or in the case of declining balance, when the change is very small. In the case of declining balance, use a spreadsheet to update the asset T-account and verify the answer.*

Exercise 6.3 *Pension liabilities are calculated as the present value of future cash flows to retired employees. The calculations are based on the pension contract and some actuarial estimates about what will happen in the future.*

the pension contract:

$$\text{annual retiree benefit } = 2\% \times \text{ service years } \times \text{ final salary}$$

actuarial estimates:

- employee will work 15 more years, then retire
- projected final salary: 100,000
- retirement will last 20 years

- discount rate is 8%

As of the beginning of the year the employee under consideration has worked ten years. Calculate the pension obligation as of the beginning of the year and the end of the year. (Under current rules for the calculation of annual benefit, the final salary is anticipated,and future service years are not. Hence, annual benefit after 10 service years is $2\% \times 10 \times 100,000 = 20,000$.)

Exercise 6.4 *The increase in a pension liability for a current employee is divided into two parts if there are no changes in the pension contract and the actuarial estimates. Service cost is the increase in the liability due to the employee working one more service year. Interest cost is the interest accrued during the year on the beginning liability.*

Revisit the previous exercise and calculate the two parts of the increase in the pension liability.

Exercise 6.5 *This exercise is a further continuation of the previous two. When a change is made to the pension contract, and the change is applied retroactively to service years already worked, the change in the liability is debited (if liability increases) to prior service cost which is an equity account that is a part of other comprehensive income. The amount in prior service cost is amortized over time to the income statement.*

Assume the pension contract was changed at the beginning of the year to

$$\text{annual retiree benefit } = 2\frac{1}{2}\% \times \text{ service years } \times \text{ final salary}$$

Update the beginning of the year pension liability for the prior service cost. Then compute the end of year liability and decompose the difference into service cost and interest cost.

Exercise 6.6 *When a firm enters into a lease obligation, and the lease is accounted for as a capital lease, it makes the following journal entry.*

lease asset	y_1	
lease obligation		y_1

The amount of the journal entry is the present value of the future cash payments. As time goes by, the lease obligation continues to be reported at the present value (economic income). The lease asset is typically amortized using an accrual first method like straight line or declining balance.

Let there be a three year capital lease obligation that requires a cash payment of 100 at the end of each of the next three years, at which point the lease expires. Further, let the firm enter into one of these obligations at the beginning of each year, so that, in steady state, the firm will hold three lease assets (leaseholds). Compute the steady state amount for the lease obligation on the balance sheet, as

well as the steady state interest expense on the lease. The continuous interest rate is 9%.

Assume the lease assets are amortized using declining balance at a rate of 60%. What is the steady state reported amount for the lease asset and amortization expense.

Exercise 6.7 *Start with the continuous compounding interest formula:*

$$P_1 = P_0 e^r$$

Derive an expression for the present value of an annuity amount, A, for n periods when interest is compounded continuously:

$$PV_A = \frac{A}{e^r - 1}\left(1 - \frac{1}{e^{rn}}\right)$$

Exercise 6.8 *What is the present value of an annuity of 100 for 3 years. The annual interest rate is 10% compounded continuously.*

Exercise 6.9 *Revisit the cash flow structure in exercise 6.1 The interest rate remains 10%, but interest is compounded continuously. What is the equilibrium price of the asset at January 1 of the first year? Using economic income amortization what is the reported asset value at the end of the years 1 and 2? What is the amortization expense for each of the three years?*

Exercise 6.10 *Redo the previous exercise under the assumption of uniform asset acquisition - that is, the firm acquires an identical asset (at identical cost) every January 1. What is the steady state asset value on the balance sheet? What is the steady state amortization expense?*

Exercise 6.11 *Here are the annual cash flows generated by an asset.*

$$\begin{array}{ccc} CF_1 & CF_2 & CF_3 \\ e^{.1} & e^{.2} & e^{.3} \end{array}$$

The market rate of interest is 10% per year compounded continuously. Compute the present value of the asset at time 0 which is the purchase price. Compute the asset values and the income amounts for the three periods. Also, verify total income is equal to total cash flow less the purchase price.

Exercise 6.12 *Here are the annual cash flows generated by an asset.*

$$\begin{array}{ccc} CF_1 & CF_2 & CF_3 \\ ae^r & be^{2r} & ce^{3r} \end{array}$$

The market rate of interest is r per year compounded continuously. Compute the present value of the asset at time 0 which is the purchase price. Compute the asset values and the income amounts for the three periods. Also, verify total income is equal to total cash flow less the purchase price.

Exercise 6.13 *The contractual cash flows in a lease agreement are due at the end of each year.*

$year$	1	2	3	4
$payment$	100	100	100	200

Let the firm enter into one of these obligations at the beginning of each year, so that, in steady state, the firm will hold four lease assets (leaseholds). Compute the steady state amount for the lease obligation on the balance sheet, as well as the steady state interest expense on the lease. The discount rate is 7%*, and interest is compounded continuously. Assume the lease assets are amortized using sum of the years' digits amortization. What is the steady state reported amount for the lease asset and amortization expense.*

Exercise 6.14 *Suppose the probability of winning a particular game is* $1/10$*. If the game is played ten (independent) times, what is the probability of winning at least once? Supposes the probability of winning is* $1/100$*, and the game is played* 100 *times? Now let the probability of winning be* $1/t$*, and play the game* t *times, where* t *gets very large. Hint: Consider the limit expression for continuous compounding*

$$\lim_{t\to\infty}\left(1+\frac{1}{t}\right)^t = e$$

to show the probability of winning at least one time approaches

$$1-e^{-1}$$

Exercise 6.15 *Refer to the previous exercise. This time let the probability of winning an individual game be* $p = 1/10,000$*, and the number of times the game is played be* $t = 1,000$*. What is the probability of winning at least one individual game? For* $p = 17/20,000$ *and* $t = 1,000$*? Now use the limit expression for continuous compounding with interest rate* r

$$\lim_{t\to\infty}\left(1+\frac{r}{t}\right)^t = e^r$$

to show the probability of winning at least one time is approximated by

$$1-e^{-pt}$$

when p *is small and* t *is large.*

Exercise 6.16 *Recently some study notes belonging to the young Abraham Lincoln were found. Here's an interest problem: "If 100 dollars in one year gain* $3\frac{1}{2}$ *dollars interest, what sum will gain $38.50 in one year and a quarter?" Presuming Lincoln did the problem with simple interest computations, what is the answer? Now do the problem using continuous compounding of the interest.*

Exercise 6.17 *Redo the Lincoln problem with slightly different numbers. "If 100 dollars in one year gain* 20 *dollars interest, what sum will gain $250 in one year and a quarter?" Do the problem with simple interest and with continuous compounding.*

Exercise 6.18 *An investment is available that grows at a 10% annual rate compounded continuously. How long will it take for the initial investment to double? What is the doubling time if the growth rate is 20%?*

Exercise 6.19 *An investor has $20 of investable funds. There is an investment available that returns 15% annually compounded continuously, but there is an initial fee of $5. How long will it take before the value of the investment is back to $20?*

Exercise 6.20 *Refer to the previous exercise. How much can the initial fee be if the investor requires the value of the investment be no less than $20 after 3 years?*

7

information and accounting stocks and flows

So far we have done the accounting for long-lived assets as if there is no uncertainty about the sequence of future cash flows. In this chapter uncertainty is introduced: an immediate implication is that we do not know for sure what the future cash flows will look like. The observed cash flows, however, contain some information about what will happen next. The first question: how can we use the observed cash flows to make our best prediction about what will happen next?

But this document is about accounting, and, while accounting writes down historical cash flows, it does much more than that. The historical cash flows are aggregated and reported as numbers on the current set of financial statements. The asset balance on the balance sheet, for example, is (or, at least, can be) the result of a history of cash flow numbers. So the second question arises: can accounting sequentially update so that a useful prediction about what will happen next is contained in a single financial statement number? The answer is yes, and can be accomplished by a judicious choice of amortization method.

The first step in dealing with the problem, however, is to access some ideas and results in probability, in particular Bayes' theorem about revising probabilities as new evidence arrives.

7.1 a little bit about probability

Definition 7.1 *a probability measure is a number (between zero and one) attached to each possible (uncertain) event.*

Recall the state act outcome representation in chapter 4. For arbitrage pricing we never had to assign probabilities to the states (which was, indeed, a powerful part of the theory). Now we can think quantitatively about the relative likelihood of different states. For concreteness, consider a four state example.

Example 7.1

state	θ_1	θ_2	θ_3	θ_4
probability	p_1	p_2	p_3	p_4

The states are defined as mutually exclusive and exhaustive; that implies the sum of the state probabilities is one. In notation, $\sum p_i = 1$. *Define event* A *as the occurrence of either state 1 or state 2, that is* $A = \{\theta_1, \theta_2\}$ *and, similarly* $B = \{\theta_2, \theta_3\}$.

The probability of event A occurring is the sum of the probabilities of its state components. In notation $P(A) = p_1 + p_2$ and, similarly, $P(B) = p_2 + p_3$. These are often referred to as the marginal probabilities.

Definition 7.2 *The joint probability of two (or more) events occurring.is denoted* $P(A, B)$*, and is called the joint probability.*

For the example $P(A, B) = p_2$. The relationship between marginal and joint probabilities is described in the sum rule.

Definition 7.3 *Sum rule: The marginal probability is equal to the sum of the logically possible joint probabilities.*

To get all the logically possible joint probabilities, it is convenient to define another event: the event not A denoted $\overline{A} = \{\theta_3, \theta_4\}$, and similarly $\overline{B} = \{\theta_1, \theta_4\}$. According to the sum rule

$$\begin{aligned} P(A) &= P(A, B) + P\left(A, \overline{B}\right) \\ &= p_2 + p_1 \end{aligned}$$

Similarly

$$\begin{aligned} P(B) &= P(A, B) + P\left(\overline{A}, B\right) \\ &= p_2 + p3 \end{aligned}$$

Definition 7.4 *A conditional probability is the probability of an event occurring given that some other event is known to occur, denoted, for example,* $P(B|A)$.

For the example

$$P(B|A) = \frac{p_2}{p_1 + p_2}$$

The only way $B = \{\theta_2, \theta_3\}$ can occur when A occurs (that is, either state θ_1 or θ_2) is θ_2. The conditional probability $P(B|A)$ is scaled by $p_1 + p_2$ so the total

conditional probabilities add to one.

$$\begin{aligned} & P(B|A) + P(\overline{B}|A) \\ = \quad & \frac{p_2}{p_1 + p_2} + \frac{p_1}{p_1 + p_2} = 1 \end{aligned}$$

Conditionals joints, and marginals obey the product rule.

Definition 7.5 *Product rule:*

$$P(A, B) = P(B|A)P(A)$$

For the example

$$P(A, B) = \frac{p_2}{p_1 + p_2}(p_1 + p_2) = p_2$$

Likewise,

$$\begin{aligned} P(A, B) &= P(A|B)P(B) \\ &= \frac{p_2}{p_2 + p_3}(p_2 + p_3) = p_2 \end{aligned}$$

Combining the two expressions

$$\begin{aligned} P(A, B) &= P(B|A)P(A) = P(A|B)P(B) \\ P(B|A) &= \frac{P(A|B)P(B)}{P(A)} \end{aligned}$$

The last equation is an important one and is sometimes referred to as Bayesian revision. In terms of the example,

$$P(A|B) = \frac{\left(\frac{p_2}{p_1+p_2}\right)(p_1 + p_2)}{p_2 + p_3} = \frac{p_2}{p_2 + p_3}$$

Some other useful probability concepts can be illustrated by augmenting the example with outcome values.

Example 7.2

state	θ_1	θ_2	θ_3	θ_4
probability	p_1	p_2	p_3	p_4
outcome	x_1	x_2	x_3	x_4

The mean (or expected value) of a probability distribution is the vector product of the probability vector with the outcome vector.

$$mean = \sum p_i x_i$$

The variance is another vector product.

$$variance = \sum p_i \left(x_i - mean\right)^2$$

Mean and variance are useful concepts for describing a distribution of probabilities. In fact, an important continuous probability distribution which we shall encounter shortly is the normal distribution which is entirely specified by the two parameters, mean and variance.

A final definition which will prove useful is precision of a probability distribution, defined as the reciprocal of the variance

$$precision = 1/variance$$

In chapter 8 we will continue to work with discrete (separate) states. For the remainder of this chapter we will move to a special continuous distribution often called the normal or Gaussian distribution.

7.2 Bayesian normal revision

The general problem we are interested in is how to best revise our prior beliefs (stocks) with evidence from the current period (flows) to arrive at posterior or conditional beliefs (new stock). Further, we would like to see if we could perform the revision computations in an accounting stock and flow framework. To that end we will be particularly interested in Bayesian revision for a special continuous distribution called the normal or Gaussian distribution.

With a continuous distribution the random variable can assume a continuous value, not just discrete values. In that case the probability measure p is replaced by a continuous density function f. The theorem for Bayesian updating with a general continuous distributions goes like the two event case with f replacing p. Suppose the variable we are interested is X and the evidence we have access to this period is Y.

Theorem 7.1 *Bayesian updating with continuous probability density functions is*

$$f(X|Y) = \frac{f(Y|X) f(X)}{f(Y)}$$

The normal distribution is specified as follows.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean of the distribution and σ^2 is the variance. At this stage we could substitute the normal distribution into the revision relationship and do some algebra. While the algebra is a little bit tedious, it is not impossible. In any event it is available is several places. See, for example, DeGroot. The result is slick and easily stated.

Theorem 7.2 *Consider a normally distributed random variable, X, with mean $E[X]$ and precision τ. Let observed evidence Y be available where $Y = X + \varepsilon$.*

ε has mean zero and precision r. The updated conditional distribution after Y is observed is also normal with mean

$$E[X|Y] = \frac{\tau E[X] + rY}{\tau + r}$$

and precision

$$\tau + r$$

Pause for a moment to appreciate how neat this result is. The updated mean is the precision weighted combination of the previous mean and the new evidence, Y. As only two numbers are required to do the updating, it appears quite promising that accounting can reproduce the computations. After all, accounting marches through time by updating a balance sheet number (stock) with a periodic flow number.

At this point we have sufficient machinery to attack the cash flow prediction problem. The next section introduces such a problem.

7.3 accounting set-up

Accountants process an observed sequence of cash flows. To keep a concrete example in mind, suppose the sequence is

$period$	$cash\ flow$
1	100
2	200
3	300
4	200
5	?

The first question: "What do we know (what can we predict) about the cash flow in period 5?" And then: "Will the answer, or some part of it, show up in the accounting?" The direction of the cash flow could be in or out. If outflow, the journal entries are

debit asset

credit cash

and, probably, an amortization entry.

debit expense

credit asset

Notice the accounting brings to the table its own stocks and flows as captured in the T-account representation.

Asset	
Beginning balance	
CF_t (evidence)	Amortization
Ending balance	

The beginning and ending balances are stocks, and the cash flow and amortization entries are, of course, flows.

To enrich the problem sufficiently to make it interesting (but not too complicated), we'll assume the cash flow is composed of two parts: a temporary and a permanent component. Further, the relationship between the components is linear, allowing the following representation.

$$\begin{aligned}
\overline{CF}_1 &= \overline{CF}_0 + \varepsilon_1 \\
CF_1 &= \overline{CF}_1 + s_1 \\
\overline{CF}_2 &= \overline{CF}_1 + \varepsilon_2 \\
CF_2 &= \overline{CF}_2 + s_2 \\
\overline{CF}_3 &= \overline{CF}_2 + \varepsilon_3 \\
CF_3 &= \overline{CF}_3 + s_3
\end{aligned}$$

CF_t is the observed (realized) cash flow in period t; $\overline{CF}_t$ is an *unobservable* parameter which will be convenient to treat as the underlying mean for the observed cash flow. The actual cash flow, CF_t, is the mean, $\overline{CF}_t$, plus an uncertain error (or shock) term, s_t. Furthermore, the underlying mean, $\overline{CF}_t$, is, itself, subject to random shocks, ε_t. The shocks to the mean, ε_t, have a permanent effect, as they will affect the cash flow in subsequent periods as the underlying mean gets updated. Conversely, the shocks, s_t, to the cash flow, CF_t, are temporary, as they do not affect the updating of the underlying mean.

To keep the problem from becoming too cumbersome, we further assume the error terms are independent of each other. This will simplify some of the computations while still keeping the main ideas in view.[1]

The problem is that only the CF_t are observable, but we would like to estimate the underlying (and unobservable) mean, and whatever else would be useful for predicting the next cash flow in the sequence. At least on the surface, this looks to be a fairly complicated problem. As it turns out though, we will be able to do the entire problem with regular accounting computations. The key is to find the appropriate (declining balance) amortization rate.

7.4 information stocks and flows

First restate the relationship for the revised mean using the cash flow notation.

$$E[\overline{CF}_t|CF_t] = \frac{\tau E[\overline{CF}_{t-1}] + rCF_t}{\tau + r}$$

[1] A reasonable way to deal with covariance in the error terms is to effect a transformation of variables, so the transformed variables exhibit no covariance. The analysis then proceeds as in our examples.

The first step is to get an expressions for the variance of the unconditional distribution $f\left(\overline{CF}_{t-1}\right)$ and the evidence distribution $f\left(CF_t\right)$ The variance of $f\left(\overline{CF}_{t-1}\right)$ is composed of two parts. We have a prior variance of $\overline{CF}_{t-1}$ depending on our prior beliefs: denote the prior variance as σ^2_{t-1}. Then the permanent error term ε is added, so the total variance is $\sigma^2_{t-1} + \sigma^2_{\varepsilon}$. The variance of two random variables is the sum of the variances as long as the variables are not related. In other words, the summation is appropriate if there is no systemic relationship between the two variables as we have assumed.

The variance of the evidence distribution $f\left(CF_t\right)$ is simply σ^2_s. In terms of precisions, we have

$$\begin{aligned} \tau &= \frac{1}{\sigma^2_{t-1} + \sigma^2_{\varepsilon}} \\ r &= \frac{1}{\sigma^2_s} \end{aligned}$$

Using theorem 7.2 for normal Bayes, we can write the precision of the revised distribution $f\left(\overline{CF}_t | CF_t\right)$ as

$$\begin{aligned} \frac{1}{\sigma^2_t} &= \tau + r \\ &= \frac{1}{\sigma^2_{t-1} + \sigma^2_{\varepsilon}} + \frac{1}{\sigma^2_s} \end{aligned}$$

At this stage simplify the notation a little bit. First, substitute a relationship about the relative size of the temporary and permanent shocks.

$$\sigma^2_s = v\sigma^2_{\varepsilon}$$

And, since, as far as the accounting is concerned, the relative size of the variance is all that matters set

$$\sigma^2_{\varepsilon} = 1$$

All the important derivations, as far as the accounting is concerned, are the same for general σ^2_{ε}; see exercise 7.8. Now the precision updating expression:

$$\frac{1}{\sigma^2_t} = \frac{1}{\sigma^2_{t-1} + 1} + \frac{1}{v}$$

Now make one further simplification: assume we have reached steady state. Steady state is when the variance is unchanged from one period to the next. In other words the decrease in the variance due to acquisition of more evidence, an observed cash flow, is offset by the variance increase due to the permanent shock term. Notationally,

$$\frac{1}{\sigma^2} = \frac{1}{\sigma^2 + 1} + \frac{1}{v}$$

where σ^2 is substituted for both σ_t^2 and σ_{t-1}^2. Now do some algebra and solve for an expression for σ^2. First put the right hand side over a common denominator.

$$\frac{1}{\sigma^2} = \frac{v^2 + \sigma^2 + 1}{(\sigma^2 + 1)\, v}$$

Crossmultiply and simplify.

$$\left(\sigma^2\right)^2 + \sigma^2 - v = 0$$

An expression for σ^2 can be acquired by using the quadratic formula.

$$\sigma^2 = \frac{-1 \pm \sqrt{1 + 4v}}{2}$$

As the variance must be positive, we are only interested in the positive root.

$$\sigma^2 = \frac{-1 + \sqrt{1 + 4v}}{2}$$

The expression for the variance of $f\left(\overline{CF}_t | CF_t\right)$ is not too bad, What we are really interested in, however, is the updated mean $E[\overline{CF}_t | CF_t]$, and for that we require the expressions for precision.

$$E[\overline{CF}_t | CF_t] = \frac{E[\overline{CF}_{t-1}]\tau + CF_t r}{\tau + r}$$

So we would like to have manageable expressions for

$$\frac{\tau}{\tau + r} \text{ and}$$
$$\frac{r}{\tau + r}$$

So far, our expressions for steady state τ, r and $\tau + r$ are as follows.

$$\begin{aligned} \tau &= \frac{1}{\sigma^2 + 1} \\ r &= \frac{1}{v} \\ \tau + r &= \frac{1}{\sigma^2} \end{aligned}$$

Substituting

$$\frac{r}{\tau + r} = \frac{\frac{1}{v}}{\frac{1}{\sigma^2}} = \frac{\sigma^2}{v} = \frac{-1 + \sqrt{1 + 4v}}{2v}$$

This expression is of particular interest to us, as it is the coefficient of cash flow in the updating formula. In the accounting the coefficient of cash flow is the amortization rate, so let's define the following as the declining balance amortization rate and see how it works.

$$D = \frac{-1 + \sqrt{1 + 4v}}{2v}$$

And since

$$\frac{\tau}{\tau + r} = 1 - \frac{r}{\tau + r} = 1 - D,$$

the updating formula is now

$$E[\overline{CF}_t | CF_t] = (1 - D)\, E[\overline{CF}_{t-1}] + DCF_t$$

Let's see how this formulation will work in the T-account.

The general T-account looks like this.

Asset	
BB	
CF_t	$(BB + CF_t)D$
EB	

The question before us is whether the T-account can produce the conditional expected cash flow: $E[\overline{CF}_t | CF_t]$. In particular, is the depreciation entry on the credit side of the T-account the number we are interested in? The answer is yes, as long as the beginning balance satisfies the following condition:

$$BB = \frac{1 - D}{D} E[\overline{CF}_{t-1}]$$

Then the T-account will look like this.

Asset	
$\frac{1-D}{D} E[\overline{CF}_{t-1}]$	
CF_t	$(1 - D)\, E\left[\overline{CF}_{t-1}\right] + DCF_t = E\left[\overline{CF}_t | CF_t\right]$
$\frac{(1-D)^2}{D} E[\overline{CF}_{t-1}] + (1 - D)\, CF_t$	

There are two things to notice about the T account. The first is that the depreciation expense amount is the updated conditional mean of the cash flow. That is, we have used the T account to do a relatively complicated Bayesian updating problem. The second thing to notice is that the ending balance satisfies the same condition we required for the beginning balance. So next period, and every period after that, Bayesian updating will be accomplished in the T account.

$$\begin{aligned} EB &= \frac{(1 - D)^2}{D} E[\overline{CF}_{t-1}] + (1 - D)\, CF_t \\ &= \frac{(1 - D)}{D} \left[(1 - D)\, E\left[\overline{CF}_{t-1}\right] + DCF_t\right] \\ &= \frac{(1 - D)}{D} E\left[\overline{CF}_t | CF_t\right] \end{aligned}$$

That is, the beginning balance every period maintains the correct relationship to the updated mean.

7.5 accounting stocks and flows

We are now ready to do a numerical example using the cash flow numbers from the beginning of the chapter. In particular, we can compute the updated mean of $f\left(\overline{CF}_t|CF_t\right)$ each time a new cash flow datum is observed.

Example 7.3 *Recall the cash flow numbers from the beginning of the chapter.*

$period$	$cash\ flow$
1	100
2	200
3	300
4	200

We finessed the issue of a starting point for the variance of $\overline{CF}_1$, as it does not appear in the updating. We do, however, require a starting point for the mean. For simplicity use $E[\overline{CF}_1] = CF_1$. For the equal variance case, $v = 1$, the amortization rate is

$$\begin{aligned} D &= \frac{-1+\sqrt{1+4v}}{2v} \\ &= \frac{-1+\sqrt{5}}{2} \simeq 1.618034 \end{aligned}$$

To get the T account started right, set the beginning balance to

$$\begin{aligned} BB &= \frac{(1-D)}{D} E[\overline{CF}_1] \\ &= \frac{(1-D)}{D}(100) \\ &\simeq 61.8034 \end{aligned}$$

Then the T account looks like this.

Asset	
61.8034	
200	$D\,(61.8034+200) = 161.8034$
100	
300	$D\,(100+300) = 247.2136$
152.7864	
200	$D\,(152.7864+200) = 218.034$

The numbers for the updated mean appear as credits to the asset account, that is, the amortization expense for each period.

Example 7.4 *Use the same cash flows, but this time let* $v = 2$*,and, hence,*

$$\begin{aligned} D &= \frac{-1+\sqrt{1+4v}}{2v} \\ &= \frac{-1+3}{2\,(2)} = \frac{1}{2} \end{aligned}$$

The numbers for the updated mean appear as credits to the asset account, that is, the amortization expense for each period.

Using the same cash flows and the same starting point, $E[CF_1] = CF_1$, the asset T-account is presented.

Asset	
100	
200	$(100 + 200)/2 = 150$
150	
300	$(150 + 300)/2 = 225$
225	
200	$(225 + 200)/2 = 212.5$

As before the updated mean numbers appear as amortization expense in the asset account. For this example, as D and $1 - D$ are the same, the ending balance is the same as $E[\overline{CF}_t|CF_t]$, which is the amortization amount.

Here's another thing worth pointing out. The balance in the asset account has no evident connection to any market value. There is, for example, no discount rate in view to compute a discounted cash flow number. The account balance is used in the computation of the mean of the posterior distribution for future cash flows, not to represent a market value.

To elaborate further on the role of the account balance in this exercise, recall how steady state accounting worked in the certainty case of chapter 6. In a trivial, but nonetheless correct, sense the "expected" cash flow is recorded as both a debit (cash flow) and a credit (amortization expense) in the T-account. The actual cash flow is the expected amount, as there is no uncertainty about the future. In the uncertainty case the cash flows are used to compute the posterior distribution mean, which then shows up as the amortization expense.

Notice also that the depreciation rate depends on v which is part of the economic context. Recall

$$\sigma_s^2 = v\sigma_\varepsilon^2$$

where σ_s^2 is the variance of the temporary shock term. If the temporary shock term has high variability relative to the permanent shock, σ_ε^2, then the most recent cash flow is a relatively poor predictor of the underlying mean: that is, the recent cash flow will tend to have a relatively high error embedded in it. In that case, its weight will decline as the prediction is made. The weight on the most recent cash flow is the depreciation rate: D. We would expect, then, to see the depreciation rate go down when v increases.

That is the way the numerical examples behaved. When v was one, the depreciation rate was approximately $.618034$. The rate declined to $.5$ when v increased to 2. To illuminate the general relationship invert the depreciation rate function and write v as a function of D..

$$v = \frac{1 - D}{D^2}$$

One way to verify the above function is to expand into quadratic form.

$$vD^2 + D - 1 = 0$$

Using the quadratic formula, the expression for D returns.

$$D = \frac{-1 + \sqrt{1 + 4v}}{2v}$$

So whichever way v goes, the depreciation rate goes in the other direction. Some D, v pairs are tabulated.

D	0	$1/4$	$1/3$	$1/2$	$.618$	$2/3$	$3/4$	1
v	∞	12	6	2	1	$3/4$	$4/9$	0

This result squares with casual empiricism. A high technology industry is one in which the permanent shock term has a relatively high variance, that is, low v We are used to seeing high depreciation rates in those settings, sometimes even no capitalization of assets in the first place. On the other hand, stable industries with unchanging technologies tend to have lower depreciation rates.

7.6 summary

The main idea of the chapter is how to rationally update prior beliefs with evidence; Bayes' is the main theorem. The version of Bayes' for the normal probability density function is remarkably slick. We brought it to bear on a stylized, but relatively general, revision problem. Further, accounting is well suited to organize the computations. All prior knowledge can be captured by a "stock" number, and consistently updated with new evidence: a "flow."

7.7 exercises

Exercise 7.1 *Ralph's firm pays cash every year for a similar long-lived asset. The amount of the cash payment is uncertain determined by a dynamic process. The observed cash flow contains both permanent and temporary shocks as in the chapter. On average, the two shocks are the same, that is the variances of the shock components are equal,* $v = 1$. *The observed cash flows for the first four years of operation are*

year	*cash flow*
1	*400*
2	*300*
3	*450*
4	*500*

Calculate the updated estimate of the mean of the cash flows after each cash flow is observed. Compute an amortization rate and a beginning balance adjustment so that the mean estimate appears as amortization expense each year. Assume

$$E\left[\overline{CF}_1\right] = CF_1 = 400$$

and the variance of $\overline{CF}_t$ *is steady state.*

Exercise 7.2 *Redo the previous exercise, but this time use* $v = 1.2$.

Exercise 7.3 *One more time - do the previous exercise with* $v = \sqrt{2}$.

Exercise 7.4 *Ralph owns and manages a hotel, called Hotel Ralph, near a major football stadium. Ralph's customers are virtually all institutions who pay in advance to gain access to the facilities - guest rooms, banquet rooms, and so forth - for special occasions like football weekends. Upon receipt of cash Ralph credits an unearned revenue account. At year end an adjusting entry is made to recognize revenue. This sequence of journal entries is identical to depreciating the cost of long-lived assets except that debits and credits are reversed. Ralph estimates the process generating cash inflows is linear and has both a permanent and transitory error component, and Ralph estimates the variances of the two shock terms are equal. The cash flows for the first three years are*

year	*cash flow*
1	*8*
2	*12*
3	*16*

Construct an accounting scheme so the updated estimate of the mean cash flow appears in the financial statements.

Exercise 7.5 *Show for general variance ratio* v *the following amortization scheme results in amortization expense equal to the expected cash flow*

$$E\left[\overline{CF}_t | CF_t\right]$$

amortization rate, D:

$$\frac{2}{1+\sqrt{1+4v}}$$

beginning balance adjustment, $\frac{1-D}{D}$:

$$\frac{-1+\sqrt{1+4v}}{2}$$

Exercise 7.6 *There are prior beliefs about probability density $f(x)$: x is normally distributed with expected value 10 and variance 5. x is unobservable. There is an observable variable y, where $y = x + \varepsilon$, where ε is a normally distributed mean zero shock with a variance of 10. (x and y are unrelated, that is, zero covariance.)*

The observation is $y = 3$. Describe the posterior distribution of x conditional on y.

Exercise 7.7 *Use the cash flow dynamics from the chapter where the shock terms are independent, normally distributed, and mean zero.*

$$\begin{aligned} \overline{CF}_t &= \overline{CF}_{t-1} + \varepsilon_t \\ CF_t &= \overline{CF}_t + s_t \end{aligned}$$

What is the steady state variance of $f\left(\overline{CF}_t \mid CF_t\right)$ if $\sigma_s^2 = \sigma_\varepsilon^2 = 1$?

Recompute the steady state variance for $\sigma_s^2 = 2$ and $\sigma_\varepsilon^2 = 1$.

Finally, recompute the steady state variance for $\sigma_s^2 = 1$ and $\sigma_\varepsilon^2 = 2$.

Exercise 7.8 *The assumption was made in the text that $\sigma_e^2 = 1$, and the expression for updated mean was derived.*

$$E[\overline{CF}_t | CF_t] = (1-D)\, E[\overline{CF}_{t-1}] + DCF_t$$

Show the same expression is derived for general σ_e^2.

8
the fundamental theorem of accounting

Many academic bodies of knowledge possess what are, at least sometimes, referred to as "fundamental theorems." In earlier chapters we have encountered a couple of these: the fundamental theorem of linear algebra and the fundamental theorem of finance. We will encounter some others, such as the fundamental theorem of arithmetic. There are also, of course, fundamental theorems of algebra and calculus.

These various theorems serve, among other things, to describe the nature of their respective sciences, as well as to provide a foundation upon which to construct other results and insights. At this stage of our exploration it might seem natural to ask if there exists a result which does some of the same things for the information science of accounting. Chapter 8 is devoted to a modest evaluation of a theorem which might so serve.

Some of the "fundamental theorems" are not developed in the context in which they are seen as fundamental. The fundamental theorem of finance is an example, being a general theorem in applied mathematics, but possessing strong implications when framed in an finance context. The proposed "fundamental theorem of accounting" has a similar history: the heavy lifting for developing the theorem is due mostly to Claude Shannon, John Kelly, and Steve Ross. The issue is whether it can be illuminating when framed in an accounting context.

Accounting is often characterized as information for economic decisions. There are, of course, multiple sources of information in any economic setting. Nonetheless, accountants are inclined to feel there is something special about accounting information, being structured as it is by the double entry system and using the language of economic stocks and flows.

The theorem under consideration connects accounting stocks and flows with other information denominated in probabilities. Indeed, the relationship is stronger than a simple connection: it is an equivalence. For any source of information (stated in probabilities) the amount of the information (using entropy measures developed by Claude Shannon) is equal to the rate of return generated by the information (using accounting stocks and flows). In other words, when an economic decision maker acquires information, the increase in the accounting rate of return is equal to the amount of information. The accounting reveals how much information is used without disclosing the information itself.

The remainder of the chapter goes like this. The theorem is presented along with a simple numerical example. Then the requisite conditions for the theorem to hold are discussed, of which there are three:

- arbitrage free pricing
- spanning of the state space by opportunities
- long run decision criterion

Accounting, in the structured way it marches through time, is inherently a long run activity, so it seems reasonable to be matched up with long run decisions. Spanning the state space is also connected with long run behavior, as the inability to trade in some states is routinely eased by multiperiod arrangements.

The implications of the theorem are discussed in terms of the framing of accounting choice problems. Also, some social welfare consequences are discussed. In what sense might paying attention to accounting numbers make society better off?

8.1 mutual informationtheorem

The mutual information theorem establishes an equivalence between rates of return (stocks and flows) and the information content of an information channel. The right hand side - that is, the information content - is denominated in probabilities and the content metric is based on the entropy concept of Claude Shannon. Start with the right hand side of the theorem.

8.1.1 right hand side (entropy and mutual information)

Definition 8.1 *The entropy of an random variable, X, is defined:*

$$H(X) = -\sum_{x} p(x) \ln p(x)$$

Entropy is a measure of the uncertainty in a random variable, or a system of random variables, and is, importantly, a function only of the probabilities of the

random variables. The greater the uncertainty, the greater the entropy number. For a degenerate random variable which can take only one value, the entropy is zero, since $\ln 1 = 0$. The entropy of a random variable which can attain two different values each with probability one-half is

$$H\left(\frac{1}{2},\frac{1}{2}\right) = -\left[\frac{1}{2}\ln\left(\frac{1}{2}\right)+\frac{1}{2}\ln\left(\frac{1}{2}\right)\right] = -\ln\left(\frac{1}{2}\right) = \ln 2 \simeq .693$$

Notice the negative sign ensures entropy is a positive number.

We are interested in the information properties of a random variable. That is, what one random variable tells us about another. To that end, we require definitions for the joint and conditional entropy of two random variables.

Definition 8.2 *The joint entropy of a system with two random variables:*

$$H(X,Y) = -\sum_x \sum_y p(x,y)\ln p(x,y)$$

Example 8.1 *Consider a simple two variable example with the following joint and marginal probabilities.*

$p(x,y)$	y_1	y_2	$p(x)$
x_1	$\frac{1}{3}$	0	$\frac{1}{3}$
x_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
$p(y)$	$\frac{2}{3}$	$\frac{1}{3}$	

The entropy of X is

$$\begin{aligned} H(X) &= -\left[\frac{2}{3}\ln\frac{2}{3}+\frac{1}{3}\ln\frac{1}{3}\right] \\ &= \ln 3 - \frac{2}{3}\ln 2 \\ &\simeq .6365 \end{aligned}$$

The joint entropy of X and Y is (where $0ln0$ is evaluated as zero)

$$\begin{aligned} H(X,Y) &= -\left[\frac{1}{3}\ln\frac{1}{3}+\frac{1}{3}\ln\frac{1}{3}+\frac{1}{3}\ln\frac{1}{3}\right] \\ &= \ln 3 \\ &\simeq 1.0986 \end{aligned}$$

Definition 8.3 *The conditional entropy of a random variable, Y, conditional on the value of another random variable, X, uses conditional probabilities:*

$$H(Y|X) = -\sum_x p(x)\sum_y p(y|x)\ln p(y|x)$$

The next step uses the sum and product rules for probabilities.

$$\begin{array}{rcl} \text{sum rule} & : & \\ p(x) & = & \sum_y p(x,y) \end{array}$$

$$\begin{array}{rcl} \text{product rule} & : & \\ p(x,y) & = & p(y|x)\,p(x) \end{array}$$

Back to the example conditional probabilities are given.

$p(y\|x)$	y_1	y_2
x_1	1	0
x_2	$\frac{1}{2}$	$\frac{1}{2}$

$$\begin{aligned} H(Y|X) &= -\frac{1}{3}[1\ln 1] - \frac{2}{3}\left[\frac{1}{2}\ln\frac{1}{2} + \frac{1}{2}\ln\frac{1}{2}\right] \\ &= \frac{2}{3}\ln 2 \\ &\simeq .4621 \end{aligned}$$

Again using the product and sum rules, the entropy expressions can be written:

$$\begin{aligned} H(X) &= -\sum_x\sum_y p(x,y)\ln p(x) \\ H(Y|X) &= -\sum_x\sum_y p(y|x)\,p(x)\ln p(y|x) \\ &= -\sum_x\sum_y p(x,y)\ln p(y|x) \end{aligned}$$

An important property of entropy is additivity, that is $H(X,Y) = H(Y|X) + H(X)$ which can be derived from the additive property of the ln function.

$$\begin{aligned} H(Y|X) + H(X) &= -\sum_x\sum_y p(x,y)\left[\ln p(y|x) + \ln p(x)\right] \\ &= -\sum_x\sum_y p(x,y)\left[\ln p(y|x)\,p(x)\right] \\ &= -\sum_x\sum_y p(x,y)\ln p(x,y) \\ &= H(X,Y) \end{aligned}$$

The additivity result is important enough to be stated in a theorem.

Theorem 8.1 *Conditional and marginal entropy add to joint entropy.*

$$H(X,Y) = H(Y|X) + H(X)$$

Verify additivity in the numerical example.

$$\begin{aligned} & H(Y|X) + H(X) \\ = & \left(\frac{2}{3}\ln 2\right) + \left(\ln 3 - \frac{2}{3}\ln 2\right) \\ = & \ln 3 \\ = & H(X,Y) \end{aligned}$$

Additivity implies the total uncertainty in a system, $H(X,Y)$ is the sum of the uncertainty in X, $H(X)$, plus the residual uncertainty in the system conditional on X, $H(Y|X)$. If X is thought of as an information variable, then conveniently when information goes up, uncertainty goes down by the same amount. This is the basis for defining mutual information as the amount of information in X about Y, that is, the amount the uncertainty declines when X is available.

Definition 8.4 *Mutual information:* $I(X;Y) = H(Y) - H(Y|X)$.

$H(Y)$, of course, is the total uncertainty in Y. If the residual uncertainty in Y conditional on X, $H(Y|X)$, is lower than $H(Y)$, then the decrease in uncertainty is due to the information in X.

It will be convenient for computational purposes to substitute the additivity condition in mutual information to get

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

which demonstrates that $I(X;Y)$ is symmetric. That is, the amount of information in X about Y is the same as the amount of information in Y about X, as both $H(Y) - H(Y|X)$ and $H(X) - H(X|Y)$ reduce to the expression on the right.

For the ongoing example, note that $H(Y) = H(X)$, and we can compute mutual information.

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \left(\ln 3 - \frac{2}{3}\ln 2\right) + \left(\ln 3 - \frac{2}{3}\ln 2\right) - (\ln 3) \\ &= \ln 3 - \frac{4}{3}\ln 2 \\ &\simeq .1744 \end{aligned}$$

The right hand side of the theorem is $I(X;Y)$.

Note $I(X;Y)$ is composed entirely of probabilities: there are no dollars and no decisions. The left hand side has both dollars and decisions. And the theorem establishes the equivalence of the (change in) rate of return and the amount of information acquired.

8.1.2 left hand side (Kelly criterion)

For the left hand side of the theorem we require a decision problem denominated in dollars. A general problem can be represented in state-act-outcome format,

here with two states, θ_1 and θ_2, and two acts with prices $x_1$and x_2.

price	θ_1	θ_2
x_1	a_{11}	a_{12}
x_1	a_{21}	a_{22}

The outcomes, a_{ij}, are jointly determined by act and state. If the states are spanned by the acts - there are as many independent acts as there are states - the problem can be converted to Arrow-Debreu format.

price	θ_1	θ_2
ad_1	1	0
ad_2	0	1

ad_i is the price of the Arrow-Debreu security returning 1 dollar in state θ_i, and zero elsewhere. Finally, the problem can be scaled so the act prices are unity.

price	θ_1	θ_2
1	y_1	0
1	0	y_2

Here, y_i is the reciprocal of the Arrow-Debreu price.

$$y_i = \frac{1}{ad_i}$$

At this stage the problem is entirely specified by a vector of scaled Arrow-Debreu payoffs, y, and a vector of probabilities, $p(y)$. Suppose, for simplicity, we have \$1 to invest, and we choose b_i to be the amount invested in scaled Arrow-Debreu payoff y_i. That is, we can purchase a fraction of the act opportunity with the corresponding fractional payoff. The payoff, then, if state θ_i occurs is $y_i b_i$. We usually need not require that each b_i be between zero and one (although typically, and surprisingly, that will occur at the optimal investment strategy). We do, however, require the b_i's sum to one.

Our objective is to choose the b vector so as to maximize the expected rate of return. But which rate of return shall we maximize? We choose here to maximize the expected continuously compounded rate of return in the time value of money formula, that is, maximize $E[r]$ in

$$P_t = P_0 e^{rt}$$

This choice of what to maximize has implications which will arise later.

In our setting we have

$$y_i b_i = e^{r_i}$$

where r_i is the continuously compounded return in state i, and can be written

$$r_i = \ln y_i b_i$$

The quantity we wish to maximize is *expected* rate of return

$$E[r_i] = \sum_i p(y_i) \ln y_i b_i$$

The problem is easy to state, and the optimal solution is simple enough to state, as well. To simplify and standardize the notation a little, let $W(Y)$ be the expected continuous rate of return for payoffs y and marginal probabilities $p(y)$, so

$$W(Y) = E[r_i] = \sum_i p(y_i) \ln y_i b_i$$

Theorem 8.2 *The Kelly*[1] *criterion: For*

$$\begin{aligned} W^*(Y) &= \max_b \sum p(y_i) \ln y_i b_i \\ \textit{subject to } \sum b_i &= 1 \end{aligned}$$

The solution is

$$W^*(Y) = \sum p(y_i) \ln y_i - H(Y)$$

and the optimal investment (betting) strategy is

$$b_i^* = p(y_i)$$

It is relatively easy to see that, if $b_i = p(y_i)$, then the expected rate of return is

$$W(Y) = \sum p(y_i) \ln y_i - H(Y)$$

since

$$\begin{aligned} W(Y) &= \sum p(y_i) \ln y_i b_i \\ &= \sum p(y_i) \ln y_i p(y_i) \\ &= \sum p(y_i) \ln y_i + \sum p(y_i) \ln p(y_i) \\ &= \sum p(y_i) \ln y_i - H(Y) \end{aligned}$$

To see

$$b_i^* = p(y_i)$$

consider the two state problem so p and b can be interpreted as scalars.

$$\underset{b}{Maximize} \quad p \ln b y_1 + (1-p) \ln (1-b) y_2$$

[1]The Kelly criterion and the mutual information relationship we will call the fundamental theorem were developed by John Kelly, a colleague of Claude Shannon's at Bell Labs. Kelly's important paper is entitled "A New Interpretation of Information Rate."

The first order condition with respect to b is

$$\begin{aligned} \frac{p}{b} + (-1)\frac{1-p}{1-b} &= 0 \\ \frac{p}{b} &= \frac{1-p}{1-b} \\ p &= b \end{aligned}$$

For n states a Lagragian method will yield the same condition.

Example 8.2 *Consider a risky opportunity which, along with a risk-free opportunity, appears as follows in state-act-outcome format.*

price	θ_1	θ_2
1	1	1
1	$\frac{1}{2}$	2

The first step is to get the problem into the form for which the Kelly criterion can be applied. That is, we want the opportunities to be scaled Arrow-Debreu securities with unit price. Solving for the state prices, the problem has Arrow-Debreu form.

state price	θ_1	θ_2
$2/3$	1	0
$1/3$	0	1

(Here's where spanning is used; without sufficient independent investment opportunities, there would be ambiguous state prices as in chapter 4.)

Rescaling gives unit price and scaled Arrow-Debreu payoff.

price	θ_1	θ_2
1	$3/2$	0
1	0	3

and this can be presented in parsimonious form with unit prices and zero off diagonal payoffs implied.

	θ_1	θ_2
y	$3/2$	3

The next step is to assign state probabilities and use them to operationalize the Kelly criterion. Start with as simple and instructive set of state probabilities: the state prices, themselves. That is, let p_i be the state price for state i.

$$\begin{aligned} p_1 &= \frac{2}{3} \\ p_2 &= \frac{1}{3} \end{aligned}$$

Now use the Kelly criterion to maximize $E[r]$ by setting the investment shares, b, equal to the probabilities, p, sometimes referred to as "betting one's beliefs."

$$\begin{aligned} E[r] &= W(Y) = \frac{2}{3}\ln(\frac{2}{3})(\frac{3}{2}) + \frac{1}{3}\ln(\frac{1}{3})(3) \\ &= \ln 1 = 0 \end{aligned}$$

Here the maximum rate of return is zero. A general result which we will encounter in the exercises is that when the probabilities are scaled state prices, the maximum rate of return available is the risk free rate. For the example the risk free rate is zero.

Different state probabilities yield different rates of return. Anticipating a numerical example for the fundamental theorem, use two more probability vectors: $p_1 = 1$ and $p_2 = 0$, as well as $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{2}$. For $p_1 = 1$

$$\begin{aligned} W(Y|p_1 = 1) &= 1\ln(1)(\frac{3}{2}) + 0\ln(0)(3) \\ &= \ln\frac{3}{2} \\ &\simeq .4055 \end{aligned}$$

And one more set of probabilities

$$\begin{aligned} W\left(Y|p_1 = \frac{1}{2}\right) &= \frac{1}{2}\ln(\frac{1}{2})(\frac{3}{2}) + \frac{1}{2}\ln(\frac{1}{2})(3) \\ &= \frac{1}{2}\ln(\frac{3}{4}) + \frac{1}{2}\ln(\frac{3}{2}) \\ &= \ln 3 - \frac{3}{2}\ln 2 \\ &\simeq .05889 \end{aligned}$$

The left hand side of the mutual information theorem is $W(Y)$. More particularly, it is the change in $W(Y)$ that comes about when information is available. The theorem establishes an equality between expected rate of return and mutual information $I(X;Y)$.

8.1.3 the mutual information theorem

When the decision maker has access to information, x, the investments can be made using the conditional probabilities, $p(y|x)$. Then, to compute expected rate of return, the conditional rates of return are weighted by the probability of receiving the signal, $p(x)$. Conditional on the information the optimal rate of return is written

$$\begin{aligned} W(Y|X) &= \sum_x p(x) \sum_y p(y|x)\ln p(y|x)\, y \\ &= \sum\sum p(x,y)\ln p(y|x) + \sum\sum p(x,y)\ln y \\ &= \sum\sum p(x,y)\ln y - H(Y|X) \end{aligned}$$

From the Kelly criterion, we already have an expression for the rate of return without the information, x.

$$\begin{aligned} W(Y) &= \sum p(y_i)\ln y_i - H(Y) \\ &= \sum\sum p(x,y)\ln y - H(Y) \end{aligned}$$

To obtain an expression for the increase in return due to the information, call it ΔW, take the difference, and derive the important result.

$$\begin{aligned}\Delta W &= W(Y|X) - W(Y) \\ &= \sum\sum p(x,y)\ln y - H(Y|X) - \left(\sum\sum p(x,y)\ln y - H(Y)\right) \\ &= H(Y) - H(Y|X) \\ &= I(X;Y)\end{aligned}$$

That is the theorem: the increase in the rate of return due to information is equal to the mutual information. The conditions for the theorem to hold will be discussed a bit more in the subsequent section.

Theorem 8.3 *The mutual informatioin theorem. When the following three conditions hold*

- *arbitrage free pricing*
- *opportunities span the state space*
- *long run decision criterion*

then

$$\Delta W = I(X;Y)$$

Example 8.3 *Combine the information structure from example 8.1 with the decision problem in example 8.2*

$p(x,y)$	y_1	y_2	$p(x)$
x_1	$\frac{1}{3}$	0	$\frac{1}{3}$
x_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
$p(y)$	$\frac{2}{3}$	$\frac{1}{3}$	

price	θ_1	θ_2
1	1	1
1	$\frac{1}{2}$	2

The question is how much can the decision maker increase the rate of return with the information.

We already have the no information solution.

$$W(Y) = 0$$

If signal x_1 is received, the conditional probabilities are $p(y_1|x_1) = 1$ and $p(y_2|x_1) = 0$. And we already have the computations for that probability vector from the previous example.

$$W(Y|x_1) = \ln 3 - \ln 2$$

And from the same example, we have the computations for $p(y_1|x_2) = 1/2$ and $p(y_2|x_2) = 1/2$.

$$W(Y|x_2) = \ln 3 - \frac{3}{2}\ln 2$$

Weighting by the marginal probabilities of the signals we have

$$\begin{aligned} W(Y|X) &= \frac{1}{3}W(Y|x_1) + \frac{2}{3}W(Y|x_2) \\ &= \frac{1}{3}(\ln 3 - \ln 2) + \frac{2}{3}\left(\ln 3 - \frac{3}{2}\ln 2\right) \\ &= \ln 3 - \frac{4}{3}\ln 2 \end{aligned}$$

The increase in rate of return is

$$\begin{aligned} \Delta W &= W(Y|X) - W(Y) \\ &= \ln 3 - \frac{4}{3}\ln 2 \\ &= I(X;Y) \end{aligned}$$

The mutual information computation was from the initial example for the right hand side of the theorem.

Example 8.4 *Do one more example, this time with "perfect" information. That is, signal x_i always occurs when y_i occurs.*

$p(x,y)$	y_1	y_2	$p(x)$
x_1	$\frac{2}{3}$	0	$\frac{2}{3}$
x_2	0	$\frac{1}{3}$	$\frac{1}{3}$
$p(y)$	$\frac{2}{3}$	$\frac{1}{3}$	

For the perfect information setting $H(X) = H(Y) = H(X,Y)$. Therefore,

$$\begin{aligned} I(X;Y) &= -\left[\frac{2}{3}\ln\frac{2}{3} + \frac{1}{3}\ln\frac{1}{3}\right] \\ &= \ln 3 - \frac{2}{3}\ln 2 \\ &\simeq .6365 \end{aligned}$$

For the decision problem

$$\begin{aligned} W(Y|x_1) &= \ln\frac{3}{2} \\ W(Y|x_2) &= \ln 3 \\ \Delta W &= W(Y|X) = \frac{2}{3}\ln\frac{3}{2} + \frac{1}{3}\ln 3 \\ &= \ln 3 - \frac{2}{3}\ln 2 \\ &\simeq .6365 \end{aligned}$$

And we have another numerical example of the mutual information theorem.

8.2 accounting and information

Now it is time to connect entropy and mutual information to accounting numbers. Recall that when economic income accounting is used, the ratio of income to assets available is the interest rate.

$$\frac{inc}{B+C} = r_p$$

where r_p is the periodic rate of interest.

While the above relationship is true by definition of the accounting method, it won't hurt to verify it using the relationships derived in chapter 6 as a check on our acounting thinking. In chapter 6 we had in steady state

$$\begin{aligned} inc &= \Sigma CF - C \\ B &= \frac{\Sigma CF - (1+r_p)\,C}{r_p} \end{aligned}$$

Substituting

$$\begin{aligned} \frac{inc}{B+C} &= \frac{\Sigma CF - C}{\frac{\Sigma CF-(1+r_p)C}{r_p} + C} \\ &= r_p \frac{\Sigma CF - C}{\Sigma CF - C - r_p C + r_p C} \\ &= r_p \end{aligned}$$

And, since periodic and continuous interest rates are related

$$\begin{aligned} e^r &= 1 + r_p \\ r &= \ln(1+r_p) \end{aligned}$$

where r is the continuous rate of interest.

From the entropy relationships we have derived the mutual information theorem.

$$E[r] = W(Y) + I(X;Y)$$

It is very tempting to specify conditions under which the two relationships can be combined. The first step is to consider when the observed r converges to $E[r]$ yielding

$$\ln\left(1 + \frac{inc}{B+C}\right) = W(Y) + I(X;Y)$$

That will be done in the next subsection on the law of large numbers.

The second step is to replace $W(Y)$ with r_f the risk free rate of return. Then the relationship looks like

$$\ln\left(1+\frac{inc}{B+C}\right)=r_f+I(X;Y)$$

As this has accounting numbers on one side, and information numbers on the other (along with a presumably observable interest rate, r_f), we will refer to this form as the fundamental theorem of accounting. The substitution of r_f for $W(Y)$, the expected return without information, will be considered in a subsequent subsection on maximum entropy probability assignment.

8.2.1 the law of large numbers

The law of large numbers states that, as the number of observations gets large, the relative frequency of each observation converges to its probability. For example, imagine flipping a fair coin: one where the probability of observing heads is the same as the probability of tails, i. e. $p=1/2$. It is certainly possible to get more heads than tails, for example. But as the number of trials increases, the number of heads divided by the total number of trials (that is, the relative frequency of heads) will approach $1/2$. That is, the relative frequency approaches the probability. It is a useful exercise to verify the empirical phenomenon by performing multiple coin flips. Or, to save time, use a computer simulation.

The accounting numbers are generated by actual rates of return. As there are more and more observed returns, the effect is as if the observed returns were all the expected value. This effect, of course, requires many observations. But that is consistent with our long run accounting frame. The firm is a "going concern" whose expected life is long relative to individuals who comprise the firm. Also, the firm is composed of many assets, each of which generates a separate rate of return observation.

For a derivation of the effect of the law of large numbers in our context, recall wealth in time t, P_t, is a function of t observed returns: $r_1, r_2, ..., r_t$.

$$P_t=e^{r_1}e^{r_2}\cdots e^{r_t}$$

Setting P_0 to one for convenience.

$$\begin{aligned}\ln P_t &= \ln e^{r_1}+\ln e^{r_2}+...+\ln e^{r_t}\\ &= \sum_{i=1}^{t} r_i\end{aligned}$$

And dividing by t

$$\frac{1}{t}\ln P_t=\frac{1}{t}\sum_{i=1}^{t} r_i$$

Now access the law of large numbers. The mean of the r_i observations (sum of the r_i divided by t) converges to the expected value, $E[r]$, as the relative frequencies converge to the probabilities.

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}r_i = E[r]$$

So we have

$$\begin{aligned}\frac{1}{t}\ln P_t &= E[r_i]\\ P_t &= e^{E[r]t}\end{aligned}$$

The mutual information theorem expresses rate of return as a function of mutual information.

$$E[r] = W(Y) + I(X;Y)$$

The question is whether we can compute mutual information using accounting numbers. The answer is in the affirmative as long as the relative frequencies approximate the probabilities. That is, we will rely on the law of large numbers. And use the numerical examples to illustrate the idea.

Recall the running example (without additional information).

	θ_1	θ_2
y	$\frac{3}{2}$	3
p	$\frac{2}{3}$	$\frac{1}{3}$

Without information, the Kelly rate of return is the same in both states.

$$\begin{aligned}r_1 &= \ln\frac{2}{3}\left(\frac{3}{2}\right) = \ln 1 = 0\\ r_2 &= \ln\frac{1}{3}(3) = \ln 1 = 0\end{aligned}$$

So the observed end of period asset amount in dollars is

$$P_1 = P_0 e^r = 1e^{\ln 1} = 1$$

And observation of the accounting numbers reveals the rate of return of zero.

When information is included in the example, the observed end of period asset amount is differs by state. Start with the perfect information setting.

$p(x,y)$	y_1	y_2
x_1	$\frac{2}{3}$	0
x_2	0	$\frac{1}{3}$

And recall

$$I(X;Y) = \ln 3 - \frac{2}{3}\ln 2$$

The question is whether cash accounting can reveal mutual information.

When state one occurs, the perfect information rate of return is $r_1 = \ln \frac{3}{2}$; and in state two, $r_2 = \ln 3$. So the observed asset amount is computed for each state

$$\begin{aligned} \theta_1 \quad &: \qquad P_1 = P_0 e^{r_1} = 1e^{\ln \frac{3}{2}} = \frac{3}{2} \\ \theta_2 \quad &: \qquad P_1 = P_0 e^{r_2} = 1e^{\ln 3} = 3 \end{aligned}$$

These are both for $P_0 = 1$. As the observations, and accounting, march through time, ending balance in one period is the beginning balance the next. So for a sequence of .θ_1 then θ_2, the asset balances are as follows.

$$\begin{array}{ccccc} t & 0 & 1(\theta_1) & & 2(\theta_2) \\ P_t & 1 & \frac{3}{2} & & \frac{3}{2}(3) = \frac{9}{2} \end{array}$$

Now access the law of large numbers by setting the relative frequency of state occurrence equal to its probability. For the example $p(\theta_1) = 2/3$ and $p(\theta_2) = 1/3$, so a sequence of two θ_1's and one θ_2 will satisfy the relative frequency requirement.

$$\begin{array}{ccccccc} t & 0 & 1(\theta_1) & & 2(\theta_1) & & 3(\theta_2) \\ P_t & 1 & \frac{3}{2} & & \frac{3}{2}\left(\frac{3}{2}\right) = \frac{9}{4} & & \frac{9}{4}(3) = \frac{27}{4} \end{array}$$

When the relative frequencies are equal to the probabilities, the multiperiod cash accounting rate of return is equal to the growth rate $E[r]$ and, importantly, also equal to the mutual information $I(X;Y)$.

$$\begin{aligned} \frac{P_3}{P_0} &= \frac{27}{4} = e^{3E[r]} = e^{3I(X;Y)} \\ 3I(X;Y) &= \ln 27 - \ln 4 \\ &= 3\ln 3 - 2\ln 2 \\ I(X;Y) &= \ln 3 - \frac{2}{3}\ln 2 \end{aligned}$$

And the mutual information is revealed by the accounting numbers.

Do one more cash accounting example, this time with the imperfect information set-up.

$$\begin{array}{ccc} p(x,y) & y_1 & y_2 \\ x_1 & \frac{1}{3} & 0 \\ x_2 & \frac{1}{3} & \frac{1}{3} \end{array}$$

with the same scaled Arrow-Debreu payoffs as before.

$$\begin{array}{ccc} & \theta_1 & \theta_2 \\ y & \frac{3}{2} & 3 \end{array}$$

And recall $I(X;Y) = \ln 3 - \frac{4}{3}\ln 2$.

When x_1 is observed, the rate of return is $\ln 3/2$. When x_2 is observed, the Kelly strategy is to invest half in each security, so the rates of return are $\ln 3/4$

and $\ln 3/2$, each with probability one-half. Combining with the observation probabilities, with probability $2/3$ the rate of return is $\ln 3/2$, and with probability $1/3$ the rate of return is $\ln 3/4$. Tabulating the observed asset realizations with these relative frequencies

$$\begin{array}{cccccc} t & 0 & 1\left(r=\ln\frac{3}{2}\right) & & 2\left(r=\ln\frac{3}{2}\right) & 3\left(r=\ln\frac{3}{4}\right) \\ P & 1 & \frac{3}{2} & & \frac{3}{2}\left(\frac{3}{3}\right)=\frac{9}{4} & \frac{9}{4}\left(\frac{3}{4}\right)=\frac{27}{16} \end{array}$$

Now retrieve mutual information from the accounting numbers.

$$\begin{aligned} e^{3I(X;Y)} &= P_3 = \frac{27}{16} \\ 3I\left(X;Y\right) &= \ln 27 - \ln 16 \\ &= 3\ln 3 - 4\ln 2 \\ I\left(X;Y\right) &= \ln 3 - \frac{4}{3}\ln 2 \end{aligned}$$

The law of large numbers also provides some additional conceptual background to the Kelly criterion. Previously we stated the objective of the Kelly criterion is to maximize the expected continuously compounded rate of return. Recall from the law of large numbers, we have

$$P_t = e^{E[r]t}$$

So maximizing $E\left[r\right]$ is the same maximizing P_t. When t is large, we can think of P_t as terminal wealth, and the Kelly criterion can be stated as maximizing terminal wealth. And the Kelly criterion is seen as a long term wealth maximization strategy, providing another connection to the long run accounting frame.

8.2.2 maximum entropy probability assignment

One purpose of this subsection is to replace $W\left(Y\right)$ with the risk-free rate on the information side of the fundamental theorem.

$$\begin{aligned} \ln\left(1+\frac{inc}{B+C}\right) &= W\left(Y\right)+I\left(X;Y\right) \\ \ln\left(1+\frac{inc}{B+C}\right) &= r_f+I\left(X;Y\right) \end{aligned}$$

We are, indeed, headed in that direction, but first we confront a more general problem: How to assign probabilities in an uncertain setting. In the problems we have been doing so far, the probabilities have been part of the problem statement. A natural question arises: Where do the probabilities come from? In this section probability numbers will be derived from the statement of the problem. i

The general idea if fairly straightforward. When the setting is uncertain, we are allowed to use all our experience and all our knowledge to asses the likelihood of various events. Using entropy as a measure of uncertainty is a way to make systematic that we effective use all we know, but be careful not to use *more* than we know.

Roughly speaking, the additivity of entropy impliesour knowledge of uncertain events breaks cleanly like a saltine into two additive components: what we know, and what we don't know. Notationally, let Z be our background knowledge associated with the problem context. (Use Z to distinguish from the information system X which we might be able to access depending on the setting.) Then we have our additivity relation.

$$H(Y,Z) = H(Y|Z) + H(Z)$$

where $H(Z)$ is the entropy of our background knowledge (what we know), and $H(Y|Z)$ is the remaining uncertainty (what we don't know). When one quantity goes up, the other declines. $H(Y,Z)$, the state of the system, does not change; what is subject to change is our knowledge or understanding of the system.

This suggests two possible ways of attacking the probability assignment problem. We could try to pick probabilities which minimize the residual entropy, $H(Y|Z)$, until we bum,p constraints specifying things we don'e know. Or, we could pick probabilities which maximize residual entropy until we bump constraints describing what we know. Because of additivity, both ways will get to the same place. Both ways look (and are) difficult. But we might at least have a shot with the second approach. It might be feasible to write down constraints which capture what we know. Writing down constraints to represent what we don't know seems a hopeless task.

We have, then, constrained optimization problem. The choice variable is the probability vector, p. The objective to maximize is the residual entropy $H(Y|Z)$. And, since Z is our background knowledge which we propose to capture in the constraint formulation, we can write $H(Y)$ in the objective function and put Z in the constraints.[2]

$$\max_{p} H(Y)$$
$$subject\ to\ Z$$

with another constraint to make sure the probabilities are positive and sum to one.

While maximum entropy probability assignment is a powerful theoretical tool, actual formulation and solution in the world of affairs is typically complicated. The particular problem we are interested in, however, is not too bad. Relief is sometimes available by changing the frame of the problem, and the mutual information theorem supplies a different frame.

---

[2]In a book entitled Probability Theory: The Logic of Science E. T. Jaynes offers extensive explanations for maximum entropy probability assignment, including some mathematically rigorous derivations.

Mutual information is defined as the reduction in entropy.

$$I(Z;Y) = H(Y) - H(Y|Z)$$

So $H(Y|Z)$ and $I(Z;Y)$ move together, but with the opposite sign. Furthermore, from the mutual information theorem, $I(Z;Y)$ and $W(Y|Z)$ move together with the same sign.

$$W(Y|Z) = W(Y) + I(Z;Y)$$

Therefore, we can replace $H(Y|Z)$ in our thinking with $W(Y|Z)$ with a different sign. Now the constrained optimization can be written as follows, with Z moved into the constraints, as before.

$$\min_{p} W(Y)$$
$$subject\ to\ Z$$

Now consider the background knowledge, Z, for our problem. The only background knowledge available is the state-act-outcome matrix, A, and the price vector, x. From A and x we can compute the state prices vector s. And we can get the scaled Arrow-Debreu security outcome in state i, y_i, as the reciprocal of the state price.

$$y_i = \frac{1}{s_i}$$

A Kelly investor will "bet his beliefs" so the objective function is

$$\begin{aligned} W(Y) &= \sum p_i \ln p_i y_i \\ &= \sum p_i \ln \frac{p_i}{s_i} \end{aligned}$$

Then the maximum entropy program can be written as a program to minimize expected rate of return.

$$\begin{aligned} \underset{p}{Min} W(Y) &= \sum p_i \ln \frac{p_i}{s_i} \\ s.t. \sum p_i &= 1 \end{aligned}$$

Notice the background knowledge, Z, consisting of A and x, is incorporated into the objective function by the use of the state prices, s. The only remaining constraint is about the probabilities.

The problem in this form is not particularly difficult. Form the Lagrangian

$$\mathcal{L} = \sum p_i \ln \frac{p_i}{s_i} - \lambda \sum p_i$$

And the first order conditions are

$$\frac{\partial \mathcal{L}}{\partial p_i} = 1 + \ln \frac{p_i}{s_i} - \lambda = 0$$

So the ratio of the assigned probabilities to the state prices is a constant.

$$\frac{p_i}{s_i} = \frac{p_j}{s_j} \quad \forall i, j$$

As the probabilities must sum to one, they are equal to the scaled state prices: scaled so they sum to one.assignment

$$p_i = \frac{s_i}{\sum s_i}$$

To ensure the probability numbers are positive, we rely on the fundamental theorem of finance, and the no arbitrage condition on A and x, which implies the state prices, and, hence, the probabilities, are non-negative.

This is the solution to the maximum entropy probability assignment problem: when the only background knowledge is A and x, the maximum entropy probabilities are the scaled state prices. We're still after the maximum return available, and that can be computed by substituting the probabilities into the objective function.

$$\begin{aligned} E[r] &= W(Y) \\ &= \sum p_i \ln \frac{p_i}{s_i} \\ &= \sum \frac{s_i}{\sum s_i} \ln \frac{s_i}{\sum s_i} \frac{1}{s_i} \\ &= \ln \frac{1}{\sum s_i} \\ &= r_f \end{aligned}$$

where r_f is the risk-free rate. To get the last step, recall $\sum s_i$ si the price of a risk-free security; that is, one which pays one dollar in all states of the world. The continuously compounded risk-free rate, then, satisfies

$$1 = \sum s_i e^{r_f}$$

Hence,

$$r_f = \ln \frac{1}{\sum s_i}$$

Example 8.5 *Recall the Cox, Ross, Rubinstein example with state-act-outcome matrix and price vector.*

price	θ_1	θ_2
.8	1	1
50	25	100

Note the periodic risk free rate is

$$1 - \frac{1}{.8} = .25$$

The state prices are computed as

$$s_1 = .4 = s_2$$

and the scaled A-D outcomes are

$$y_1 = y_2 = \frac{1}{s_i} = \frac{5}{2}$$

Therefore, the maximum entropy probabilities are

$$\begin{aligned} p_i &= \frac{s_i}{\sum s_i} \\ p_1 &= \frac{.4}{.8} = \frac{1}{2} = p_2 \end{aligned}$$

Using these probabilities, the maximum possible Kelly return is

$$\begin{aligned} W(Y) &= \sum p_i \ln p_i y_i \\ &= \frac{1}{2} \ln \frac{1}{2}\frac{5}{2} + \frac{1}{2} \ln \frac{1}{2}\frac{5}{2} \\ &= \ln 1.25 \\ &= r_f \end{aligned}$$

To transform to the periodic risk free rate

$$e^{w(Y)} - 1 = .25$$

which we noted was the return on the given risk-free security.

As the example illustrates, setting the probabilities to minimize the expected return implies the best a Kelly investor can do is to purchase the risk free security. As it is minimum return, it is also maximum entropy by the mutual information theorem, and we have accomplished maximum entropy probability assignment.

We have also established the maximum rate of return achievable using only the information available in observed prices and outcomes, that is, in A and x. Any extra return implies access to additional information; and by the fundamental theorem the extra return is equal to the mutual information. So we have our theorem relating accounting numbers and information.

Theorem 8.4 • *When accounting is done using economic income accounting, and the four long run conditions hold: (1) arbitrage free prices, (2) opportunities span the state space, (3) Kelly decision-makers, and (4) enough observations to access the law of large numbers, then*

$$\ln\left(1 + \frac{inc}{B + C}\right) = r_f + I(X;Y)$$

Recall how the three "long run" conditions are used. Arbitrage free prices are necessary for the state prices, and hence, the probabilities to be non-negative. If the opportunities span the state space, then all Arrow-Debreu securities are available for trading, and the scaled Arrow-Debreu return, y, is uniquely specified for every state. Long run Kelly decision makers maximize the expected return for the information available, and sufficient observations allow the interchangeability of $E[r]$ and r.

8.2.3 spanning

Of the long run conditions, it might be appropriate to offer a few words about the spanning condition at this point. In order to invoke the Kelly criterion, we require unique (scaled) state prices. Uniqueness is a consequence of the number of independent investment opportunities equal to the number of states. In other words, the investment opportunities span the state space. In the absence of spanning, that is, strictly fewer independent investment opportunities than states, there will be multiple equilibrium prices, and non-uniqueness issues as in chapter 4 will arise.

Lack of spanning implies that there are some states in which trade can not occur; that is, there is no Arrow-Debreu security for some states. The standard (and perhaps at some level the only) reason for trade to fail is asymmetry of information. Consider, for example, the sale of a used car. The seller has private information about the car's history: how well it has been maintained, and collisions, and so forth. The buyer, of course, is in the dark, and, hence, is likely reluctant to buy something the privately informed seller would like to dispose of. The is the famous "market for lemons" problem posed by George Akerlof.

Attempts to alleviate the private information problem naturally, and perhaps inevitably, turn to ways of extending the time horizon. Warranties, maintenance contracts, and the corporate form, itself, are examples of multiperiod contracts. If the relationship is an enduring one, the uninformed buyer becomes more willing to purchase, made more secure in the knowledge that there will be future interactions down the road. And the privately informed seller, aware of potential benefits in the future, will be less apt to make an early exploitative decision.

The accounting process is an important part of many enduring relationships like, for instance, corporations. So while a multiperiod perspective is required for the accounting fundamental theorem to hold, accounting, itself, enables the existence of enduring relationships.

8.3 financial statements and information

In this section we will use the fundamental theorem to construct some example financial statements. To ease financial statement preparation, we will revisit continuous replacement of assets and steady state. In the first subsection economic income accounting will be utilized. We know, however, from chapter 6 that other

accounting methods can reproduce economic income. The subsequent subsection will use declining balance depreciation methods to do so.

8.3.1 steady state financials

We start with the same bakground knowledge as in previous examples, that is, the state act outcome matrix, A, and the price vector, x. Earlier we appended state probabilites to the statement of the problem. Now utilizing maximum entropy probability assignment, we can compute the state probability vector, p, as well as the scaled Arrow Debreu payoff, y, from the state prices, s.

Now recall how continuous replacement of assets and steady state works. As in chapter 6 let the assets be acquired in a continuous replacement fashion. That is, a new asset is acquired at the beginning of *each* period for the same acquisition price, C. Each asset then generates periodic cash flows at the end of each of the next n periods after acquisition: $CF_1, CF_2, ..., CF_n$. All assets generate the same cash flow sequence. Since one asset is acquired each period, the accounting entity will hold n productive assets for any period $\geq n$. The acquisition cost, C, is the asset's discounted cash flow.

$$C = \sum_i \frac{CF_i}{e^{r_i}}$$

As in chapter 6 the asset valuation, B, converges to a stable amount using economic income accounting after n periods. Then we have only one steady state asset amount B to keep track of. The steady state amount is relatively easy to compute using the asset T-account.

Asset	
$\vdots$	
B	
1	
$(e^r - 1)(B + C)$	$\sum CF$
B	

Economic income for the period is $(e^r - 1)(B + C)$ and from the T-account, this is equal to $\sum CF - C$. And $\sum CF$ while equal to the total cash inflows over n periods from one asset is also, conveniently, equal to one period's total cash flow from n assets. Set T-account inflows and outflows equal, and solve for B:

$$\begin{aligned} C + (e^r - 1)(B + C) &= \sum CF \\ Be^r - B + Ce^r &= \sum CF \\ B &= \frac{\sum CF - e^r C}{e^r - 1} \end{aligned}$$

Recall from chapter 6 that B is the present value of a perpetuity of $\sum CF$ less the adjusted cost $e^r C$.

Since asset, B, and income are specified, the accounting rate of return is

$$\begin{aligned}\frac{income}{B+C} &= \frac{\sum CF - C}{B+C} \\ &= \frac{(e^r-1)(B+C)}{B+C} \\ &= e^r - 1\end{aligned}$$

And the economic income relationship in the theorem is verified for the special case of continuous asset replacement. Time for a numerical example.

Example 8.6 *Use the same A and x from previous examples.*

$$A = \begin{bmatrix} 1 & 1 \\ 1/2 & 2 \end{bmatrix}; \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Previously we assigned state probabilities exogenously, but now we can use maximum entropy assignment. The state prices are

$$s = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

so the state probabilities, scaled state prices, are the same.

$$p = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

The only other thing we need is a specification of the cash flows.

t	0	1	2	3
CF	-1	0	0	e^{3r}

so C is scaled to 1. r is determined by the information environment: for this example we'll have no information with $r_f = 0$, an imperfect information system, and perfect information where the signal, x, specifies the state without error. Use the imperfect system specified as before with the joint probability matrix:

$p(x,y)$	y_1	y_2
x_1	$1/3$	0
x_2	$1/3$	$1/3$

To solve for B use the derived expression:

$$\begin{aligned}B &= \frac{\sum CF - e^r C}{e^r - 1} \\ &= \frac{e^{3r} - e^r}{e^r - 1}\end{aligned}$$

The expression for B can be simplified to a polynomial[3] *in e^r.*

$$\begin{aligned} B &= e^r + e^{2r} \\ \textit{since } e^{3r} - e^r &= (e^r - 1)\left(e^r + e^{2r}\right) \end{aligned}$$

The following table has the accounting numbers for the three information environments. The numbers are computed using the derived relationships.

information	*null*	*imperfect*	*perfect*
$r_f + I(X;Y)$	0	$ln3 - 4/3ln2 \simeq .1744$	$ln3 - 2/3ln2 \simeq .6365$
B	2	2.60796	5.4615
$income$	0	.6875	5.75
$\frac{income}{B+C}$	0	.19055	.88988
$ln\left(1 + \frac{income}{B+C}\right)$	0	.1744	.6365

One purpose of the table is to show the numerical equivalence of the right hand side of the theorem, $r_f + I(X;Y)$, with the left hand side, $\ln(1 + inc/(B+C))$.

8.3.2 declining balance depreciation

The question arises whether other accounting techniques, besides economic income, connect mutual information and accounting rate of return in the manner of the theorem. The answer is in the affirmative. In this section we will draw the connection with declining balance depreciation.. The depreciation rate, D, however, must be chosen judiciously.

Recall from chapter 6, a choice of rate, D, will yield a steady state valuation, B, and the relationship can be derived using the same T-account technique as before.[4]

Asset	
$\vdots$	
B	
C	$D(B+C)$
B	

[3] Similarly, if the cash flow is e^{4r} in period 4,

$$B = e^r + e^{2r} + e^{3r}$$

sand, in general, if the cash flow is e^{tr} in period t

$$B(r,t) = \sum_{n=1}^{t-1} e^{rn}$$

This expression allows for valuation of more general cash flow patterns as developed in the exercises. Actually, when $r = 0$, there is a "division by zero" problem,but the result does go through in the limit.

[4] As before the derivation requires convergence of B, but that is assured as long as $D < 1$.

$$\begin{aligned} D(B+C) &= 1 \\ D &= \frac{1}{B+C} \\ &\text{or} \\ B &= C\frac{1-D}{D} \end{aligned}$$

When the asset balance is the same for the depreciation method as it is for economic income, the accounting rate of return is also the same, as steady state cash income is independent of the accounting method. Returning to the numerical example for $E[r] = I(X;Y) = 0$, we had $C = 1$ and $B = 2$, so a declining balance rate of $D = 1/3$ will work. For the first few periods

	Asset
1	$\frac{1}{3}$
$\frac{2}{3}$	
1	$\frac{1}{3}\left(\frac{5}{3}\right) = \frac{5}{9}$
$\frac{10}{9}$	
1	$\frac{1}{3}\left(\frac{19}{9}\right) = \frac{19}{27}$
$\frac{38}{27}$	

It takes a while to converge to $B = 2$. After 5 periods the asset balance is 1.74; after 15 periods the balance is 1.995.

As mutual information increases, the asset value increases, as well, and the corresponding depreciation rate declines. Tabulating the asset values, B, and the depreciation rate, D, for the numerical example[5]

$I(X;Y) = E[r]$	0	$ln3 - \frac{4}{3}\ln 2$	$ln3 - \frac{2}{3}\ln 2$
Asset value, B	2	2.60796	5.4615
Depreciation rate, D	$\frac{1}{3}$	$.2772$	$.1548$

The depreciation rate can be computed more directly, and, at the same time, a little intuition can be added. For depreciation rate, D, the asset balance converges to

$$B = \frac{1-D}{D}$$

For mutual information $I(X;Y) = r$, the economic income asset balance is

$$B = \frac{e^{3r} - 1}{e^r - 1}$$

[5]The lower D's take longer to converge. For example, with imperfect information, $I(X;Y) = \ln 3 - \frac{4}{3}\ln 2$, the asset value is 2.5879 after 15 periods, 2.6072 after 25.

Setting the two expressions equal yields[6]

$$D = \frac{e^r - 1}{e^{3r} - 1}$$

The expression for the declining balance rate of depreciation has a certain intuitive appeal. It is a fraction; the numerator is the amount of interest revenue recognized in the first year of the asset life; the denominator is the total amount of interest revenue over the life of the asset. If the two amounts are equal, the rate is unity and the full amount of the asset is depreciated in the first year, that is, cash accounting. As the total interest revenue increases over the first year amount, the depreciation rate declines, and more of the asset appears on the balance sheet.

Returning to the numerical example, there's a bit of a problem applying the depreciation rate formula with the no information case where $I(X;Y) = 0$. The expression for D is an indeterminate form of zero divided by zero. Skip that for now and evaluate the imperfect and perfect information cases. For $I(X;Y) = \ln 3 - \frac{4}{3}\ln 2$

$$D = \frac{e^r - 1}{e^{3r} - 1} \simeq \frac{.19055}{.6875} \simeq .2772$$

And for $I(X;Y) = \ln 3 - \frac{2}{3}\ln 2$

$$D = \frac{e^r - 1}{e^{3r} - 1} \simeq \frac{.8899}{5.75} \simeq .1548$$

Both depreciation rates are as tabulated previously.

One more numerical example.

Example 8.7

$$A = \begin{bmatrix} 7 & 4 \\ 1 & 4 \end{bmatrix};\ x = \begin{bmatrix} 4.4 \\ 2 \end{bmatrix}$$

Let the cash flow sequence be

t	0	1	2
CF	-1	0	e^{2r}

Compute accounting asset value, B, and income for two information structures: null and perfect information. Also compute an equivalent declining balance depreciation rate, D.

---

[6] The more general expression from chapter 6 for economic income accounting is

$$B = \frac{\sum CF - e^r C}{e^r - 1}$$

where $\sum CF$ is the sum of the cash inflows, and C is the asset cost.

The state prices are

$$s_1 = s_2 = .4$$

And the implied state probabilities:

$$p_1 = p_2 = .5$$

The risk free rate, then, is

$$r_f = \ln \frac{1}{\sum s} = \ln \frac{1}{.8} = \ln 5 - \ln 4 \simeq .22314$$

So the null information accounting numbers are

$$\begin{aligned} B &= e^{r_f} = \frac{e^{\ln 5}}{e^{\ln 4}} = 1.25 \\ income &= e^{2r_f} - 1 = \frac{e^{\ln 25}}{e^{\ln 16}} - 1 = .5625 \\ D &= \frac{C}{B+C} = \frac{1}{2.25} = \frac{4}{9} \end{aligned}$$

For perfect information the joint probabilities are

	y_1	y_2
x_1	$1/2$	0
x_2	0	$1/2$

And, hence„

$$I(X;Y) = \ln 2$$

So the perfect information rate of return is

$$r = \ln 5 - \ln 4 + \ln 2 = \ln 5 - \ln 2 \simeq .91629$$

And the accounting numbers:

$$\begin{aligned} B &= \frac{e^{\ln 5}}{e^{\ln 2}} = 2.5 \\ income &= \frac{e^{\ln 25}}{e^{\ln 4}} - 1 = 5.25 \\ D &= \frac{1}{2.5+1} = \frac{2}{7} \end{aligned}$$

8.4 social welfare

As shown in the previous section, it is possible for accounting to capture the long run effects in decision making. That is, accounting rate of return, with judicious attention to accrual details, can be made the same as the long run rate of return

in a particular decision context. Maximizing one is equivalent to maximizing the other. The issue in this section is the social welfare implications of long run return maximization.

One immediate imlication of the long view is the decision maker wishes to avoid going bankrupt, sometimes referred to as "gamblers' ruin." That is, being asked to leave the casino even though profitable investment opportunities are still available. The Kelly criterion guards against gamblers' ruin by spreading the investment amount over all possible states with non-zero probability.

When a decision maker does go bankrupt, there are social welfare implications. If an important player can not honor their debts, there are economy-wide effects: creditors of the bankrupt entity might also be so unable, with deleterious consequences for the credit markets. Recent financial crises have been framed in this narrative. The government has, on occasion, felt compelled to step in and bail out "too big to fail" entities.

These consequences never arise when the decision making is driven by long run consideration. That is not to say Kelly investors never go into debt. They may certainly borrow (issue debt securities). What they will not do is hold a portfolio which might have non positive value in some state of the world. Whatever borrowing which does occur can be paid back whatever the state occurrence. In other words, a Kelly investor will never go short in an Arrow-Debreu security.

To illustrate modify the ongoing example so that borrowing occurs.

Example 8.8 *The same two securities (with one alteration) are presented in state-act-outcome format.*

price	θ_1	θ_2
1	1	1
1	$2/3$	3
probability	.5	.5

The example is the same as before, except the "triple or nothing" security pays more than nothing ($2/3$) when θ_1 occurs. $2/3$ is sufficient to induce borrowing in the risk free asset.

The state prices are

$$y = \begin{bmatrix} 6/7 \\ 1/7 \end{bmatrix}$$

And the problem can be reframed in scaled Arrow-Debreu format.

	θ_1	θ_2
y	$7/6$	7
$p(y)$	.5	.5

Kelly investing yields an expected rate of return

$$\begin{aligned} W(Y) &= \frac{1}{2}\ln\frac{1}{2}\left(\frac{7}{6}\right)+\frac{1}{2}\ln\frac{1}{2}(7) \\ &= \frac{1}{2}\ln\left(\frac{7}{12}\right)+\frac{1}{2}\ln\frac{7}{2} \end{aligned}$$

The Kelly investor chooses a portfolio that returns $\frac{7}{12}$ in θ_1 and $\frac{7}{2}$ in θ_2. The weights on the Arrow-Debreu securities are positive equal to the state probabilities.

However, when the portfolio is formed using the original securities, the weight on the risk free security is negative: borrowing occurs. The security weights, λ, solve the state equations.

$$\begin{aligned} 1\lambda_1+\frac{2}{3}\lambda_2 &= \frac{7}{12} \\ 1\lambda_1+3\lambda_2 &= \frac{7}{2} \\ &\Rightarrow \quad \lambda=\begin{bmatrix} -1/4 \\ 3/4 \end{bmatrix} \end{aligned}$$

The Kelly investor sells 1/4 of the risk free security in order to invest more in the "triple" security. But borrowing of this type does not risk gamblers' ruin or any of the associated negative social welfare consequences. Basing decisions on carefully prepared accounting statements is equivalent to long run decision making. It might be the case that accountants can argue preparing and using stock and flow financial statements mitigates the "too big to fail" social welfare problem.

8.5 concluding remarks

This chapter is a speculation that accounting is an information science that might merit a fundamental theorem, and a modest proposal along those lines is offered. The theorem establishes the equivalence between accounting rate of return and the amount of information held by the reporting entity. The financial statements, then, are a statement about information.

From the work of Shannon, information is the thing that reduces uncertainty. The theorem implies information is likewise the thing that increases rate of return. The double entry accounting system is well designed to capture rate of return. Judicious use of accruals is required so that accounting rate of return equals the information amount.

Reconsider a table occupied by the leading information scientists in the University. Does accounting have sufficient intellectual tools to deserve a seat at the table? The fundamental theorem connects the idea of the amount of information

with accounting concepts including double entry, accruals and deferrals, and even depreciation schedules. Looked at from this angle, accountants at least have their own unique set of tools to bring to the table.

The fundamental theorem establishes the equivalence of an information frame and an accounting frame. Any accounting problem, such as asset valuation or income determination, can be reframed as an information problem. And, as illustrated in the final section, any information problem can be reframed as an accounting problem.

The fundamental theorem sets us up for further study of accounting related information problems. The information metric (mutual information) is central to the discussion of self-correcting codes in chapter 9. Cryptography, or the design of secret codes, is the topic of chapter 10: here the objective is to disclose the *least* possible amount of information. Information as characterized in the fundamental theorem is basic to these discussions, as well as the topics of subsequent chapters including quantum encryption and quantum information.

8.6 references

Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 1991.

Jaynes, E. T.,*Probability Theory: The Logic of Science. Cambridge University Press, 2003.*

Kelly, John, "A new interpretation of information rate." *Bell Sys. Tech. Journal*, 35: 917-926, 1956.

Ross, Stephen A.,*Neoclassical Finance*. Princeton University Press, 2005.

Shannon, Claude E., "A mathematical theory of communication." *Bell Sys. Tech. Journal*, 27: 379-423, 623-656, 1948.

8.7 exercises

Exercise 8.1 *Consider a game wherein a fair coin is flipped. If the result is heads, you receive \$1. If tails, the coin is flipped again; a result of heads on the second trial yields a payment of \$2. A head on the third trial, with no heads preceding, yields \$4. In general, if the first head is on the t^{th} trial, the payment is 2^{t-1}. The payments, x, along with their probabilities are tabulated.*

t	1	2	3	4	...	t	...
x	1	2	4	8	...	2^{t-1}	...
$p(x)$	$1/2$	$1/4$	$1/8$	$1/16$	...	$1/2^t$	...

The game is called the St. Petersburg paradox, and was invented by one of the Bernoulli's in the 18th century. The reason it is called a paradox is that the expected value of the payments is unbounded, and it seems unlikely a rational individual would be willing to pay an infinite amount to play the game.

Suppose the game is played over and over; that is, played long enough to invoke the law of large numbers. Compute the expected continuous compounded rate of return of the game, and show $E[r] = \ln 2$ implying the value of one play of the game is \$2. Is this a more reasonable amount? More importantly, is this how an accountant would approach an analysis of the game?

Exercise 8.2 *Consider a risk-free and a risky opportunity with prices and discounted payoffs given in state-act-outcome format. The state probabilities without information are equal to the scaled state prices.*

price	θ_1	θ_2
1	1	1
1	0	2

The time sequence of scaled cash flows is as follows.

period	0	1	2	3
cash flow	-1	0	$\frac{1}{2}e^{2r}$	$\frac{1}{2}e^{3r}$

One dollar is invested at the beginning of each period, and the cash inflows occur at the end of the period. There are sufficient assets and cash flows to access the law of large numbers, so r is used, rather than $E[r]$. Suppose the decision maker uses the Kelly criterion, and accounting is economic income with continuous compounding. Without information compute the steady state accounting numbers: asset value, B; income; accounting rate of return; and the appropriate declining balance depreciation rate, D.

Exercise 8.3 *Redo the previous exercise with a perfect information structure available to the decision maker before each period's investment is made. Here is the information structure in joint probability format.*

$p(x,y)$	y_1	y_2
x_1	$\frac{1}{2}$	0
x_2	0	$\frac{1}{2}$

What if the information is imperfect with these joint probabilities?

$p(x,y)$	y_1	y_2
x_1	$\frac{1}{4}$	0
x_2	$\frac{1}{4}$	$\frac{1}{2}$

Exercise 8.4 *Redo the previous two exercises with only one change in the set-up: the time sequence of scaled cash flows is now*

period	0	1	2	3	4
cash flow	-1	$\frac{1}{10}e^{r}$	$\frac{2}{10}e^{2r}$	$\frac{3}{10}e^{3r}$	$\frac{4}{10}e^{4r}$

Exercise 8.5 *Repeat the exercise one more time with another sequence of cash flows. This time, however, the cash flows are not scaled so the initial cash outflow is unitary.*

period	0	1	2
cash flow	*?*	1	1

Using a rate of return, r, consistent with each of the three information structures in the previous exercises, rescale the cash flow numbers so the initial cash outflow is one, and the cash outflows are scaled as powers of e^{r}. Then, for each rate of return, compute steady state B, income, rate of return, and depreciation rate.

Exercise 8.6 *Monty Hall was the master of ceremonies of an old TV game show wherein a contestant was sometimes faced with this opportunity. There are three doors, behind one of which is a new car; the other two doors hide a goat. The contestant chooses a door, and then - at least sometimes - Monty Hall opens one of the other doors to disclose a goat. The contestant is then asked whether they would like to switch choices before the chosen door is opened. In order to talk in terms of probabilities, here is some notation.*

D_i : *the event that door number i is opened, i goes from 1 to 3.*

C_i : *the event that the car is behind door i.*

The initial placement of the car is uniform, that is $p(C_i) = 1/3$ *for all* i*. The contestant chooses a door which we will designate as door* 1*. For purposes of this problem, Monty Hall's behavior will follow these rules: he always opens one of the other doors to disclose a goat, and chooses uniformly if the car is behind door* 1*.*

$$\begin{aligned} p(D_1) &= 0 \\ p(D_2|C_1) &= \frac{1}{2};\ p(D_3|C_1) = \frac{1}{2} \\ p(D_2|C_2) &= 0;\ p(D_3|C_2) = 1 \\ p(D_2|C_3) &= 1;\ p(D_3|C_3) = 0 \end{aligned}$$

Using the sum and product rules for probability, compute all the joint probabilities $p(C, D)$*, and the conditional probabilities* $p(C|D)$*. In particular, pay attention to the conditional probability of the position of the car once a door is opened. Does it makes sense, in this game, for the contestant to switch doors?*

Exercise 8.7 *Consider a four state setting with the following outcomes.*

$states$	θ_1	θ_2	θ_3	θ_4
$outcomes\,(\varepsilon)$	-1.5	$-.5$	$.5$	1.5

Find the maximum entropy probability vector for two separate cases:

a. There is no other constraint on p *(other than sum to one);*
b. An additional constraint is $E\left[\varepsilon^2\right] \leq .75$*.*

Exercise 8.8 *Think of a setting with three possible states.*

state	θ_1	θ_2	θ_3
probability	p_1	p_2	p_3
outcome	-1	0	1

Suppose the only other belief about the random process is that the variation is bounded. In particular, the expected value of the squared outcome is no greater than .4.

$$\begin{aligned} E\left[x^2\right] &= p_1\left(-1^2\right) + p_2\left(0^2\right) + p_3\left(1^2\right) \\ &= p_1 + p_3 \leq .4 \end{aligned}$$

Find the probability vector which maximizes entropy subject to the variational constraint. The problem can be easily completed with paper and pencil. However, it might be useful to run a Solver program for comparison purposes.

Exercise 8.9 *Redo the previous exercise where the expected value of the squared outcome is no greater than 1.0.*

$$E\left[x^2\right] \leq 1$$

Exercise 8.10 *Conduct the entropy maximization exercise in a five state setting.*

state	θ_1	θ_2	θ_3	θ_4	θ_5
probability	p_1	p_2	p_3	p_4	p_5
outcome	-2	-1	0	1	2

Let the upper bound on the expected squared outcome be one. This time the paper and pencil computations are a little bit complicated. The suggestion is to run a Solver program, and verify with paper and pencil as much of the solution as possible.

Exercise 8.11 *The familiar expected value calculation is sometimes called the "arithmetic" mean.*

$$\begin{aligned} E[X] &= p_1x_1 + p_2x_2 + ... + p_nx_n \\ &= \sum_{i=1}^{n} p_ix_i \end{aligned}$$

An alternative concept is the "geometric" mean wherein, instead of multiplying payoffs by probabilities, payoffs are raised to a probability power, and then, instead of adding, the terms are multiplied.

$$\begin{aligned} G[X] &= x_1^{p_1} \cdot x_2^{p_2} \cdots x_n^{p_n} \\ &= \prod_{i=1}^{n} x_i^{p_i} \end{aligned}$$

For example, consider a two state setting

	θ_1	θ_2
X payoff	1	2
probability	.5	.5

$$\begin{aligned} E[X] &= .5(1) + .5(2) = 1.5 \\ G[X] &= 1^{.5} \cdot 2^{.5} = \sqrt{2} \approx 1.414 \end{aligned}$$

Using the notation in the chapter, the expected continuous rate of return for a random payoff X is

$$W(X) = p_1 \ln x_1 + p_2 \ln x_2 + ... + p_n \ln x_n$$

Show

$$e^{W(X)} = G[X]$$

Exercise 8.12 *Suppose we have available an investment opportunity that pays off $6 for every $1 invested half the time. The other half of the time the invested $1 is completely lost. What fraction of our investable capital should be invested in the opportunity (assuming we wish to maximize expected continuous rate of return, and we have a risk-free alternative investment)? What is the maximum expected continuous rate of return available?*

Exercise 8.13 *Suppose an information system, X, is available to go with the opportunities in the previous exercise. The joint probabilities of the signals (x) with the two states (y) are*

$p(x,y)$	y_1	y_2
x_1	.25	.2
x_2	.25	.3

y_1 is the lose everything state, y_2 is six-fold return. How much will access to the information system increase the expected continuous rate of return?

Exercise 8.14 *Consider a "double or nothing" gamble: the probability of winning is .55. Using the Kelly criterion what fraction of available capital should be invested in the gamble? (A risk-free alternative exists.) What is the expected continuous rate of return? Suppose an information system, X, has these joint probabilities: y_2 is "double," and y_1 is "nothing."*

$p(x,y)$	y_1	y_2
x_1	.25	.25
x_2	.2	.3

Using X, how much will the expected continuous rate of return increase?

Exercise 8.15 *Redo the previous exercise when a winning outcome is four times the amount bet. That is, the gamble is "quadruple or nothing."*

Exercise 8.16 *Consider an investment opportunity with three Arrow-Debreu securities with the following payoffs.*

	y_1	y_2	y_3
payoffs	5	$\frac{10}{3}$	2
probabilities	.2	.3	.5

What is the maximum expected continuous rate of return achievable? Suppose there is an information system, X, with these joint probabilities.

$p(x,y)$	y_1	y_2	y_3
x_1	.1	.1	.1
x_2	.1	.2	.4

What is ΔW? That is, what is the increase in the expected continuous rate of return if information system X is used?

Exercise 8.17 *Consider a firm (or individual) with the following uncertain prospects presented in Arrow-Debreu format.*

	θ_1	θ_2
y	2	2
$p(y)$	$\frac{1}{2}$	$\frac{1}{2}$

What is the maximum expected continuously compounded rate of return?

Verify the preceding answer with a two period income calculation: first θ_1 occurs, then in the second period θ_2 occurs. Does it matter if θ_2 occurs first?

Now assume the firm has access to an information partition X: $\{\{\theta_1\},\{\theta_2\}\}$. That is the joint probabilities are

$\mathrm{p}(x,y)$	y_1	y_2
x_1	$\frac{1}{2}$	0
x_2	0	$\frac{1}{2}$

What is the maximum expected return available when the information partition is available?

Exercise 8.18 *Consider a firm (or individual) with the following uncertain prospects presented in Arrow-Debreu format.*

	θ_1	θ_2
y	4	1
$p(y)$	$\frac{1}{3}$	$\frac{2}{3}$

What is the maximum expected continuously compounded rate of return? To verify with value and income computations, three periods are required to get the state realizations in line with the probabilities. Let θ_1 happen first, then two realizations of θ_2. Does the order matter?

Exercise 8.19 *Refer to the previous exercise. What expected return is available with perfect information? That is, the joint probabilities are as follows.*

$p(x,y)$	y_1	y_2
x_1	$\frac{1}{3}$	0
x_2	0	$\frac{2}{3}$

Also compute the maximum expected return with these joint probabilities.

$p(x,y)$	y_1	y_2
x_1	$\frac{1}{3}$	$\frac{1}{3}$
x_2	0	$\frac{1}{3}$

Exercise 8.20 *Reconsider the prospects and the perfect information system in the previous exercises.*

	θ_1	θ_2
y	4	1
$p(y)$	$\frac{1}{3}$	$\frac{2}{3}$

$p(x,y)$	y_1	y_2
x_1	$\frac{1}{3}$	0
x_2	0	$\frac{2}{3}$

How much can the individual pay for the information system, and still achieve \$2 in expectation after three periods? Assume $P_0 = \$1$.

Exercise 8.21 *Consider a perfect information system.*

$p(x,y)$	y_1	y_2
x_1	$\frac{1}{3}$	0
x_2	0	$\frac{2}{3}$

Suppose the planning horizon is three periods. How much can be paid for the information system and still break even in expectation at $n = 3$? Assume $P_0 = \$1$. By "breakeven" is meant the expected value of the firm (or portfolio) is the same if the information system is purchased as it would have been without it. Notice nothing is said in this exercise about the prospects in the absence of the information system.

Exercise 8.22 *Let the information system X be defined in joint probability format.*

$p(x,y)$	y_1	y_2
x_1	$1/4$	$1/6$
x_2	$1/12$	$1/2$

Suppose the cost of the information system is $P_I = 1/2$, and beginning wealth P_0 is scaled to one without loss of generality. Find the breakeven number of periods, n, where the investment value is the same with or without the information system. Alternatively, set $n = 4$, and find the maximum breakeven P_I.

Exercise 8.23 *Consider the following gamble: bet \$1 on the flip of a fair coin. When the bet is heads, and heads occurs, the payoff is \$2; if tails occurs, the payoff is zero. Same thing for tails except the winning payoff is \$3. It's okay to "spread" the bet; that is, bet, for example, \$.75 on heads and \$.25 on tails. In that case the payoff would be \$1.50 if heads, and \$.75 if tails.*

Opportunity one -

A benefactor gives you \$1 to invest in the gamble. You get to put the winnings, if any, in the bank with zero interest. The opportunity is repeated 50 times: each time you receive \$1 and bet that amount. After 50 times you get to keep all the winnings. Notice the benefactor gives you a total of \$50.

Opportunity two -

The benefactor only supplies \$1, but you have access to the gamble 50 times. You are not restricted to \$1 bets; rather, you can bet whatever winnings you accumulate. The payoffs are scaled so, for example, if \$2 is bet on heads, the payoff is \$4 if heads occurs. Which opportunity do you prefer?

Information -

A coin flip predictor supplies imperfect predictions about the result. The joint probabilities for the signals "H" and "T," and the results heads and tails are

	heads	tails
"H"	.3	.2
"T"	.2	.3

How much are you willing to pay for the predictor in the opportunity one setting? Opportunity two?

Exercise 8.24 *Here are three securities in state-act-outcome format.*

x		θ_1	θ_2	θ_3
.8	*bond*	*1*	*1*	*1*
. 50	*stock*	*25*	*50*	*100*
1.7	*derivative*	*1*	*2*	*3*
	probabilities	*.25*	*.25*	*.50*

First, check for arbitrage opportunities. If none, compute the maximum expected growth rate achievable. What are the portfolio weights for the maximum growth portfolio?

Exercise 8.25 *Consider the following state-act-outcome matrix along with the price vector,* x*.*

x		θ_1	θ_2	θ_3
1	$bond$	1	1	1
2	$stock$	0	2	4
2	$derivative$	0	0	8

Verify whether arbitrage opportunities exist by computing the state prices. (There are no such opportunities.) Suppose a Kelly investor uses the scaled state prices as probabilities. What is the maximum expected return available? Now suppose there is extra information.

	θ_1	θ_2	θ_3
x_1	.25	.25	0
x_2	0	.25	.25

What is the maximum expected return when the extra information is used?

Exercise 8.26 *Suppose the risk free rate is zero, so the risk free security has price equal to one, and the state prices sum to one, so they can be used as probabilities. Then the scaled Arrow-Debreu format looks like this for a two state setting.*

y	$\frac{1}{SP_1}$	$\frac{1}{SP_2}$
p	SP_1	SP_2

And the Kelly return is

$$W(Y) = \ln \frac{SP_1}{SP_1} + \ln \frac{SP_2}{SP_2} = 0$$

Now suppose the risk free rate is non-zero. Then the state prices must be rescaled to serve as probabilities. Show the Kelly return in that case is still equal to the risk free rate.

Exercise 8.27 *Here is a two-state two-security problem in state-act-outcome format.*

$price$	θ_1	θ_2
1.25	4	1
50	25	100

What is the risk free rate of return? What are the maximum entropy state probabilities? What is the largest possible expected rate of return achievable with the maximum entropy probabilities?

Exercise 8.28 *Suppose the scaled cash flows associated with an asset are distributed across 5 years as follows.*

	0	1	2	3	4	5
$cashflow$	-1	$.2e^{r}$	$.3e^{2r}$	$-.1e^{3r}$	$.5e^{4r}$	$.1e^{5r}$

The initial cash outlay occurs at the beginning of year 1; the remaining cash flows are at the end of the respective year. The firm buys an asset at the beginning of every year. If the growth rate (information amount) is $r = 0$, what is the steady state asset valuation, using economic income accounting with continuous compounding? What is the associated declining balance depreciation rate? Recompute the steady state asset and depreciation rate for $r = \ln 2$.error detecting and correcting codes

9
error correcting codes

A main idea of this sequence of notes is that accounting is a legitimate field of academic study independent of its vocational implications and its obvious importance in the economic environment. For example, the study of accounting promotes careful and disciplined thinking important to any intellectual pursuit. Also, studying how accounting works illuminates other fields, both academic and applied. In this chapter we begin to connect the study of accounting with the study of codes. Coding theory, while theoretically rich, is also important in its applications, especially with the pervasiveness of computerized information transfer and storage. Information integrity is inherently of interest to accountants, so exploring coding can be justified for both academic and applied reasons.

9.1 kinds of codes

We will study three kinds of codes: error detecting, error correcting, and secret codes. The first two increase the reliability of message transmission. The third, secret codes, are designed to ensure that messages are available only to authorized users, and can't be read by the bad guys. The sequence of events for all three types of codes is as follows.

1. determine the message to transmit. For our purposes, we will treat messages as vectors.

2. The message is encoded. The result is termed a codeword, or cyphertext, also a vector.

3. The codeword is transmitted through a channel. The channel might inject noise into the codeword, or it might be vulnerable to eavesdropping.

4. The received vector is decoded. If the received vector is no longer the transmitted codeword, the decoding process may detect, or even correct, the error.

Setting aside secret codes until chapter 10, this chapter will concentrate on error detecting and correcting codes. Applications of these types of codes are numerous. In a business environment codes are used to design account numbers, inventory part numbers, and all sorts of identification numbers. Codes enable efficient transmission of television pictures as well as pictures sent from the farthest parts of the solar system. The cause of noise in a codeword may be as mundane as a typing mistake, or as cosmic as interference from sunspots.

We will study a popular class of error detecting and correcting codes called linear codes. Linear codes employ the same techniques we have used for understanding the linear transformations in accounting. Both encoding and decoding are accomplished by matrix multiplication. Furthermore, decoding is a direct application of the concept of a nullspace.

Consider decoding. Every linear code can be specified by a matrix called a parity check matrix, denoted H. Decoding is accomplished by multiplying the received vector by H. If the received vector is in the nullspace of H, then the received vector is a legal codeword. Notice the connection with accounting. H is "like" the accounting transformation matrix A. The legal codewords are "like" the set of looping transactions which leave the account balances unchanged. Similar to the accounting applications, the received vector, denoted y, is in the nullspace of H if H times y is a vector of zeros.

Definition 9.1 *The matrix product Hy is termed the syndrome.*

If the syndrome contains a non-zero element, an error has been detected. If we are clever in our analysis of the syndrome, we may be able to infer the position and amount of the error, thereby allowing error correction.

Before moving to examples of codes, we need to acquire another tool; the notion of a finite field.

9.2 modular arithmetic

We will restrict ourselves to a finite set of messages which will, in turn, imply a finite number of errors. This allows for efficient error detection and correction. The mechanism to accomplish this is modular arithmetic. The main idea is pretty simple, and the notation for it (devised by Gauss) is straightforward. An excellent introduction to modular arithmetic and number theory is in Ore, 1948.

Definition 9.2 *Modular arithmetic reduces the set of integers under consideration to a finite number.*

$$a = b + cm \Longleftrightarrow a \equiv b(\mathrm{mod}\ m) \text{ where } a, b, c, \text{ and } m \text{ are integers.}$$

The second equivalence is read "a is congruent to b modulo (or simply 'mod') m."

Any integer has a corresponding element in a finite set. Simply divide by m and report the remainder. There will only be m integers in the reduced set.

A convenient example is "binary." The arithmetic is done modulo 2; all integers are equivalent (congruent) to either 0 or 1. For example,

$$\begin{aligned} 2 &\equiv 4 \equiv 6 \equiv 0(\mathrm{mod}\ 2) \\ 1 &\equiv 3 \equiv -1 \equiv 1(\mathrm{mod}\ 2) \end{aligned}$$

Multiplication and addition work pretty much the way we are used to. The key is the answer is always a member of the original set. For binary the answer is either 1 or 2. For example,

$$1 + 1 \equiv 0\,(\mathrm{mod}\ 2)$$

Matrix multiplication also works in modular arithmetic.

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix} \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} (\mathrm{mod}\ 2)$$

In preparation for the first error detecting example consider arithmetic modulo 11, in which every number is divided by 11, and the remainder reported.

$$\begin{aligned} 7 + 6 &= 13 \equiv 2\,(\mathrm{mod}\ 11) \\ 7 \times 6 &= 42 \equiv -2 \equiv 9\,(\mathrm{mod}\ 11) \end{aligned}$$

Here's the entire multiplication table modulo 11.

	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	2	4	6	8	10	1	3	5	7	9
3	3	6	9	1	4	7	10	2	5	8
4	4	8	1	5	9	2	6	10	3	7
5	5	10	4	9	3	8	2	7	1	6
6	6	1	7	2	8	3	9	4	10	5
7	7	3	10	6	2	9	5	1	8	4
8	8	5	2	10	7	4	1	9	6	3
9	9	7	5	3	1	10	8	6	4	2
10	10	9	8	7	6	5	4	3	2	1

There are two important things to notice about the multiplication table.

1. There are no zeros. When two non-zero numbers are multiplied, the result is non-zero. This property does not hold for all moduli. For example,

$$6 \times 6 \equiv 0 \pmod{9}$$

2. There are no "repeats." Each row (and column) consists of all the possible 10 numbers.

The two properties have important applications in coding, and they follow directly from the concept of a prime number and what is known as the fundamental theorem of arithmetic.

Definition 9.3 *A prime number is divisible (evenly, that is, leaving no remainder) only by 1 and the number itself.*

Theorem 9.1 *Any integer can be factored into a product of prime numbers. Furthermore, the factorization is unique.*

Consider property 1 for some prime modulus, p. A zero entry in the multiplication table means

$$rs \equiv 0 \pmod{\text{p}} \Longleftrightarrow rs = pn$$

where r and s are integers less than p and n is some integer. The equality on the right implies two different factorizations of rs, violating unique factorization. Hence, the fundamental theorem implies property 1 is true for all p.

Property 2 follows from similar logic. "Repeats" imply the first two equations below which, in turn, imply the third.

$$\begin{aligned} rs &= pn + k \\ rt &= pm + k \\ r(s-t) &= p(n-m) \end{aligned}$$

And the third equation is not allowed by the fundamental theorem.

With these preliminaries we are ready for an example code.

9.3 isbn - an error detecting code

ISBN stands for "international standard book number." Virtually every book published in the world is assigned an ISBN. The numbers are assigned in a way so that some typographical errors which might occur in typing or transcribing an order can be detected. The probability that the wrong book is delivered is thereby reduced.

In 2007 the design of the ISBN was altered slightly. A visible manifestation is that the length of the number (codeword) increased from 10 to 13. We'll start with ISBN 10 as it supplies a nicer illustration of the two properties in the previous section, and catch up with ISBN 13 later.

9.3.1 isbn 10

The ISBN is a linear code in the sense that it is completely specified by its parity check matrix. For ISBN 10 the parity check matrix has one row and 10 columns.

$$H = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$

All ISBN codewords reside in the nullspace of H where arithmetic is conducted modulo 11. That is, for an ISBN y vector,

$$Hy \equiv 0 \pmod{11}$$

If y does not satisfy the above equation, an error has been made.

Example 9.1 *The ISBN for The Norton History of Mathematics (published prior to 2007) is 0-393-04650-8.*

The ISBN is divided into 4 parts, possibly of various lengths across countries and companies. The first number is the official language of the country in which the book is published: zero is English. The second set of numbers is the number assigned to the publisher: W. W. Norton Publishing is 393. The next set is an internal inventory number chosen by the publisher. The last number is a check digit which ensures the ISBN resides in the nullspace of H.

Do the arithmetic.

$$\begin{aligned} Hy &= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ 9 \\ 3 \\ 0 \\ 4 \\ 6 \\ 5 \\ 0 \\ 8 \end{bmatrix} \\ &= 231 = 11 \times 21 \equiv 0 \pmod{11} \end{aligned}$$

The ISBN code is designed to detect any single error and any transposition error, not necessarily of adjacent digits. Check to see that any such error for Norton results in a non-zero syndrome. More generally, the two error types are always caught because of the two noted properties of the multiplication table.

Consider a single error. Let the received ISBN vector be the sum of the correct ISBN and an error vector with a non-zero entry in position i.

$$y = y_{ISBN} + e = y_{ISBN} + \begin{bmatrix} 0 \\ . \\ . \\ . \\ e_i \\ . \\ . \\ . \\ 0 \end{bmatrix}$$

The syndrome is calculated.

$$Hy = H\left(y_{ISBN} + e\right) = Hy_{ISBN} + He = 0 + He$$

He is e_i times the ith element of H. As both numbers are non-zero, the syndrome is non-zero by property 1, and the error is detected.

Consider a transposition error, not necessarily of adjacent digits. Suppose in the received y that elements y_j and y_k are transposed. The received y vector appears as follows.

$$
\begin{aligned}
y &= y_{ISBN} + \begin{bmatrix} 0 \\ . \\ . \\ . \\ y_k - y_j \\ . \\ . \\ . \\ y_j - y_k \\ . \\ . \\ . \\ 0 \end{bmatrix} \\
&= y_{ISBN} + \begin{bmatrix} 0 \\ . \\ . \\ . \\ y_k - y_j \\ . \\ . \\ . \\ -(y_k - y_j) \\ . \\ . \\ . \\ 0 \end{bmatrix}
\end{aligned}
$$

For the syndrome He to be zero, two distinct elements of H, H_j and H_k, must yield the same answer when multiplied by the same number:

$$y_k - y_j$$

But that is impossible by property 2, so the transposition error is detected.

One more note before moving on to ISBN 13. Sometimes to ensure the ISBN resides in the nullspace of H, the check digit, the 10th element of y_{ISBN}, must be the number 10. Rather than writing the two digit number, the ISBN assigns the Roman numeral X. From watching the super bowl, we know X stands for 10.

9.3.2 isbn 13

ISBN 13 is similar to the universal product code used for all kinds of inventory items. It has a different looking parity check matrix.

$$H = \begin{bmatrix} 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 \end{bmatrix}$$

Furthermore, the arithmetic is done modulo 10, the multiplication table for which follows.

	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	0	2	4	6	8
3	3	6	9	2	5	8	1	4	7
4	4	8	2	6	0	4	8	2	6
5	5	0	5	0	5	0	5	0	5
6	6	2	8	4	0	6	2	8	4
7	7	4	1	8	5	2	9	6	3
8	8	6	4	2	0	8	6	4	2
9	9	8	7	6	5	4	3	2	1

It is noticed right away that the two properties of no zeros and no repeats do not hold, in general. They do hold, however, for the rows and columns associated with 1 and 3, among others. And, since those are the only numbers used in H, a single error remains detectable by the syndrome. Further, some transposition errors are detectable, as well. For example, transposition of adjacent digits will usually yield a non-zero syndrome, as the same number multiplied by a 1 and a 3 usually gives a different answer. The exception is 5. So a transposition error of adjacent digits which differ by 5 will not be caught.

$$\begin{aligned} 3\,(y_j - y_k) + 1\,(y_k - y_j) &\equiv 0\,(\mathrm{mod}\,10) \\ \text{when } y_j - y_k &\equiv 5\,(\mathrm{mod}\,10) \end{aligned}$$

Furthermore, transposition of digits removed by two places will not be caught; the syndrome will still be calculated as zero. Presumably, the designers of the code are less worried about this particular error type occurring often.

Example 9.2 *The ISBN for Managerial Uses of Accounting Information by Joel Demski (published post 2007) is 978-0-387-77450-3.*

Do the arithmetic.

$$\begin{aligned} Hy &= \begin{bmatrix} 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 9 \\ 7 \\ 8 \\ 0 \\ 3 \\ 8 \\ 7 \\ 7 \\ 7 \\ 4 \\ 5 \\ 0 \\ 3 \end{bmatrix} \\ &= 120 \equiv 0\,(\mathrm{mod}\ 10) \end{aligned}$$

The universal product code (UPC) is a ubiquitous variation on ISBN 13. The UPC is particularly visible at supermarkets where the checkout scanners read the bar codes on the inventory items. Typically, a UPC code has 12 digits, and the parity check matrix is

$$H = \begin{bmatrix} 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 & 3 & 1 \end{bmatrix} (\mathrm{mod}\ 10)$$

9.4 an error correcting code

So far we have been able to discern when an error exists in the codeword. We have not, however, been able to fix the error, at least not with the information in the syndrome alone. It is possible, by expanding the parity check matrix, to not only detect, but also correct errors in the codeword. The following example, along with other linear codes can be found in Hill, 1986.

Example 9.3 *Define the parity check matrix as follows with modulus 2.*

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Now the syndrome, Hy, has 3 elements, and the extra information can be used for error correction. The logic is not complicated. An error means the element y_i is a one instead of a zero, or vice-versa. For no errors, the syndrome is all zeros, that is, in the nullspace of H. If y_i is one instead of zero, then by the rules of matrix multiplication, the syndrome will be the ith column of H. And, since $-1 \equiv 1 \,(\mathrm{mod}\, 2)$, if y_i is a zero instead of a one, the syndrome will likewise be the ith column of H. Check with an example.

Example 9.4 *Suppose the received codeword is y=*$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^T$. *Syndrome decoding yields*

$$Hy = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \equiv \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} (\mathrm{mod}\ 2)$$

As the syndrome is the 4th column of H, it indicates y_4 should be corrected to a 1 from 0. The corrected y

$$y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

And calculation of the syndrome verifies the new y is now a legal codeword.

$$Hy = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \pmod 2$$

It would be convenient to generate legal codewords directly, rather than pick a random vector and correct it using syndrome decoding as above. There must be a simpler way, and, indeed, there is: use of the generator matrix.

9.4.1 generator matrix

In the example under consideration, the original vector (inventory number, employee id, etc.) is 4 elements long. Call it x. The matrix which multiplies the original message, x, in order to generate the codeword is called the generator matrix, G.

$$Gx = y$$

In this case, the generator matrix adds three redundant elements to x, enabling syndrome decoding to correct any single error.

It turns out to be fairly easy to construct G given H. Write H in the following block matrix format.

$$\begin{aligned} H &= \begin{bmatrix} B & I_3 \end{bmatrix}, \text{ where} \\ B &= \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} \text{ and} \\ I_3 &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

For G to be a legitimate generator matrix, the matrix product Gx must yield the zero vector when multiplied by H.

$$HGx = Hy = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ for all possible } x\text{'s}$$

This implies G is 7×4 such that

$$HG = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Here's the one that does the trick.

$$\begin{aligned} G &= \begin{bmatrix} I_4 \\ -B \end{bmatrix} \\ HG &= \begin{bmatrix} B & I_3 \end{bmatrix} \begin{bmatrix} I_4 \\ -B \end{bmatrix} = BI_4 - I_3B \\ &= B - B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Substituting for B, and recalling that -1 $\equiv 1 \pmod 2$, the generator matrix is as follows.

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

Example 9.5 *Let the original message be* $x = \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}^T$.

Calculate the codeword by multiplying by the generator matrix, G.

$$Gx = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \equiv \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \pmod 2 = y$$

Notice that G keeps the original message, x, intact and adds three redundant elements. Decoding verifies that y is a legal codeword (no noise has been injected).

$$Hy = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \pmod 2$$

9.4.2 perfect codes

The example code has another property which is convenient, on occasion.[1]

Definition 9.4 *A perfect single error correcting code is one in which the number of possible received vectors is equal to the number of legal codewords plus the number of vectors which can be changed into a legal codeword with exactly one correction.*

For a perfect code, in other words, there are no wasted vectors. Every vector is either a legal codeword, or just one element removed from a legal codeword.

To demonstrate the example code satisfies the definition for perfect requires two steps. First, it is verified there is no vector which can simultaneously be corrected to two legal codewords. That is, no vector is one change from two different legal codewords. But that can't happen. Inspection of the parity check matrix verifies there is no ambiguity about which element should be corrected: the columns of H are distinct. Whichever column is equal to the syndrome specifies the element of the received vector to correct.

Now it is a matter of counting the number of vectors in each of the categories.

- the total number of possible received vectors: $2^7 = 128$.
- the total number of legal codewords: $2^4 = 16$.
- the total number of ways a received vector can be one off from a legal codeword: $16\,(7) = 112$.

The last calculation is the number of legal codewords times the number of positions available for an error to occur. Perfectness is verified by the sum: $112 + 16 = 128$.

[1] It is particularly easy, for example, to construct examination questions.

9.5 another set of examples

The example code of the previous section, while illustrating some nice properties, is not a very large code. That is, if the codewords is meant to characterize different inventory items, for example, the size of the inventory is limited to 16 units. It is relatively straightforward, however, to increase the size of the code by increasing the modulus. As usual, the way to specify the code is to write down the parity check matrix. For the next example, arithmetic is done modulo 5.

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 0 & 1 \end{bmatrix} \pmod 5$$

With modulo 5 there are 5 possible values for each element. The syndrome, then, must supply information, not only about the position of an error, but the amount of the error, as well. When there is a zero in the syndrome, the error is easily positioned as in element 5 or 6.

Example 9.6 *Let the received vector be* $y = \begin{bmatrix} 2 & 1 & 2 & 1 & 3 & 1 \end{bmatrix}^T$. *Perform syndrome decoding.*

$$\begin{aligned} Hy &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 9 \\ 15 \end{bmatrix} \\ &\equiv \begin{bmatrix} 4 \\ 0 \end{bmatrix} \pmod 5 \end{aligned}$$

The syndrome can be fixed by subtracting 4 times the 5th column of H. And that can be accomplished by subtracting 4 from the 5th element of y. Since $-4 \equiv 1 \pmod 5$, the correction is to add 1 to the 5th element.

$$\begin{aligned} \text{corrected } y_c &= \begin{bmatrix} 2 & 1 & 2 & 1 & 4 & 1 \end{bmatrix}^T \\ Hy_c &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \end{bmatrix} \\ &\equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix} \pmod 5 \end{aligned}$$

And the codeword is corrected.

When there is no zero in the syndrome, the error resides in one of the first 4 elements.

Example 9.7 *Let the received vector be* $y = \begin{bmatrix} 3 & 2 & 4 & 1 & 2 & 3 \end{bmatrix}^T$. *Perform syndrome decoding.*

$$\begin{aligned} Hy &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 12 \\ 26 \end{bmatrix} \\ &\equiv \begin{bmatrix} 2 \\ 1 \end{bmatrix} (\mathrm{mod}\, 5) \end{aligned}$$

As the first row of H is all 1's, the first element of the syndrome is the amount of the error. The syndrome can be fixed by finding the column of H, when multiplied by 2, yields the syndrome. Because of the "no zeros - no repeats" properties of a prime modulus, only one column of H will satisfy the condition. There is, then, no ambiguity about the position and amount of the correction.

Searching the columns reveals the syndrome is 2 times the 3rd column of H.

$$2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \end{bmatrix} \equiv \begin{bmatrix} 2 \\ 1 \end{bmatrix} (\mathrm{mod}\, 5)$$

The correction is to subtract 2 from the 3rd element of y.

$$\begin{aligned} \text{corrected } y_c &= \begin{bmatrix} 3 & 2 & 2 & 1 & 2 & 3 \end{bmatrix}^T \\ Hy_c &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ 20 \end{bmatrix} \\ &\equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix} (\mathrm{mod}\, 5) \end{aligned}$$

And the correction results in the appropriate syndrome.

A generator matrix can be constructed using the methods of the previous section.

$$G = \begin{bmatrix} I_4 \\ -B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 4 & 4 & 4 & 4 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

Example 9.8 *Consistent with the prior example, let the original message* $x = \begin{bmatrix} 3 & 2 & 2 & 1 \end{bmatrix}$. *Calculate the redundant digits.*

$$\begin{aligned} Gx &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 4 & 4 & 4 & 4 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \\ 32 \\ 23 \end{bmatrix} \\ &\equiv \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \\ 2 \\ 3 \end{bmatrix} \pmod 5 \end{aligned}$$

And the appropriate redundant digits are added, as consistent with the syndrome analysis of the problem.

The final thing to do with this example is to verify the perfectness of the code. We already resolved there is no ambiguity in the correction, so just count the vectors in the categories.

- the total number of possible received vectors: $5^6 = 15,625$.
- the total number of legal codewords: $5^4 = 625$.
- the total number of ways a received vector can be one off from a legal codeword: $625\,(6)\,(4) = 15,000$.

The last calculation is the number of legal codewords times the number of positions available for an error to occur times the number of possible error amounts. Perfectness is verified by the sum: $15,000 + 625 = 15,625$.

One more example demonstrates the code can become as large as desired.

Example 9.9 *Perform arithmetic modulo 11, and let the parity check matrix be*

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 0 & 1 \end{bmatrix} \pmod{11}$$

The code has $11^{10} = 25,937,424,601$ legal codewords. This coding system could handle, for example, 12 digit phone numbers. If an individual dialed the number with one mistake, the system can correct the error, and the call can still go through to the intended party. There are no wasted phone numbers, as the code is a perfect one. The number of "one off" phone numbers is $11^{10}\,(12)\,(10) = 11^{10}\,(120)$, the number of legal phone numbers times the number of positions available for an error times the number of possible error amounts. The total of 12 digit phone numbers is $11^{12} = 11^{10}\,(1 + 120) = 11^{10}\left(11^2\right)$.

9.6 double error correction

To do *single* error correction the idea is to solve two linear equations for the two unknowns: the amount and position of the error. For *double* error correction there are four unknowns: two error amounts and their respective positions in the received vector. That requires four independent equations which, in turn, implies a parity check matrix with four rows. This double error correcting code is also in Hill, 1986.

Here is a 4 row parity check matrix which defines a linear code in what is called the BCH class.[2]

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 & 7^2 & 8^2 & 9^2 & 10^2 \\ 1 & 2^3 & 3^3 & 4^3 & 5^3 & 6^3 & 7^3 & 8^3 & 9^3 & 10^3 \end{bmatrix}$$

The arithmetic is done modulo 11.

Suppose vector y is sent, but the received vector is $y + e$ where the error vector, e, has 2 errors, a and b, in positions i and j.

$$e^T = \begin{bmatrix} 0 & ... & 0 & a & 0 & ... & 0 & b & 0 & ... & 0 \end{bmatrix}$$

The syndrome $Hy \equiv s$ looks like the following, where s_i is the i^{th} element in the syndrome vector, s.

$$\begin{aligned} s_1 &\equiv a + b \\ s_2 &\equiv ai + bj \\ s_3 &\equiv ai^2 + bj^2 \\ s_4 &\equiv ai^3 + bj^3 \end{aligned}$$

That's the set of equations we have to deal with.

Before embarking on the 4 equation 4 unknown problem, notice that no errors and a single error are fairly simple to deal with. For an error free transmission $Hy \equiv 0$ (i. e., $s_1 \equiv s_2 \equiv s_3 \equiv s_4 \equiv 0$). Then $a \equiv b \equiv i \equiv j \equiv 0$ solves the four equations.

If there is only one error, then $b = 0$, and the equations are

$$\begin{aligned} s_1 &\equiv a \\ s_2 &\equiv ai \\ s_3 &\equiv ai^2 \\ s_4 &\equiv ai^3 \end{aligned}$$

The first two equations can be easily solved for a and i which will then satisfy the other equations.

---

[2] See, for example, Hill, *A First Course in Coding Theory*, chapter 11.

For the general 2 error case return to the 4 non-trivial equations.

$$\begin{aligned} s_1 &\equiv a+b \\ s_2 &\equiv ai+bj \\ s_3 &\equiv ai^2+bj^2 \\ s_4 &\equiv ai^3+bj^3 \end{aligned}$$

The first step is to eliminate one unknown, a, by multiplying the first equation by i and subtract the second equation. Similarly, multiply the second by i and subtract the third; finally, multiply the third and subtract the fourth. We now have 3 equations in3 unknowns, numbered (1), (2), and (3).

$$\begin{aligned} is_1 - s_2 &= b(i-j) \qquad (1) \\ is_2 - s_3 &= bj(i-j) \qquad (2) \\ is_3 - s_4 &= bj^2(i-j) \qquad (3) \end{aligned}$$

Now convert 3 equations into one quadratic equation. To do this, multiply (1) times (3), and, also, (2) times itself.

(1) times (3):

$$\begin{aligned} b^2j^2(i-j)^2 &\equiv (is_1 - s_2)(is_3 - s_4) \\ &\equiv i^2 s_1 s_3 - i(s_1 s_4 + s_2 s_3) + s_2 s_4 \end{aligned}$$

(2) times (2):

$$\begin{aligned} b^2j^2(i-j)^2 &\equiv (is_2 - s_3)^2 \\ &\equiv i^2 s_2^2 - 2is_2 s_3 + s_3^2 \end{aligned}$$

Combining:

$$i^2\left(s_2^2 - s_1 s_3\right) + i(s_1 s_4 - s_2 s_3) + \left(s_3^2 - s_2 s_4\right) \equiv 0 \,(mod\ 11)$$

We have a quadratic equation modulo 11 which, with a few adjustments for modular arithmetic, we can solve for position i of one of the errors. Actually, as the quadratic equation has two roots, we will get *both* error positions. To simplify the notation let

$$\begin{aligned} p &= s_2^2 - s_1 s_3 \\ q &= s_1 s_4 - s_2 s_3 \\ r &= s_3^2 - s_2 s_4 \end{aligned}$$

And we can specify the two roots as

$$i, j \equiv \frac{-q \pm \sqrt{q^2 - 4pr}}{2p} \,(\mathrm{mod}\, 11)$$

We notice at this point that square root and division are not defined in modular arithmetic, but we can handle that as we shall see in the numerical example. Once i and j are known, use equation (1) to get b:

$$b \equiv \frac{is_1 - s_2}{i - j}$$

And use the first syndrome equation to get a:

$$a \equiv b - s_1$$

Example 9.10 *Suppose* $e^T \equiv \begin{bmatrix} 0 & 2 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$. *Do syndrome to find* a, b, i, *and* j. *Looking ahead we should find*

$$\begin{aligned} i &\equiv 2 \\ j &\equiv 4 \\ a &\equiv 2 \\ b &\equiv 3 \end{aligned}$$

First compute $Hy \equiv s \pmod{11}$.

$$\begin{aligned} s_1 &= 2 + 3 = 5 \\ s_2 &= 2\,(2) + 3\,(4) = 16 \equiv 5 \\ s_3 &= 2\,(4) + 3\,(16) \equiv 8 + 15 \equiv 1 \\ s_4 &= 2\,(8) + 3\,(64) \equiv 5 + 3\,(-2) \equiv -1 \equiv 10 \end{aligned}$$

Plugging into the expressions for p, q, and r:

$$\begin{aligned} p &= s_2^2 - s_1 s_3 = 25 - 5 \equiv 9 \\ q &= s_1 s_4 - s_2 s_3 = 50 - 5 \equiv 1 \\ r &= s_3^2 - s_2 s_4 = 1 - 50 \equiv -5 \equiv 6 \end{aligned}$$

So the solution to the quadratic equation is

$$\begin{aligned} i, j &\equiv \frac{-q \pm \sqrt{q^2 - 4pr}}{2p} \\ &= \frac{-1 \pm \sqrt{1 - 4\,(9\,(6))}}{2\,(9)} \\ &\equiv \frac{-1 \pm \sqrt{1 - 3\,(6)}}{7} \\ &\equiv \frac{-1 \pm \sqrt{5}}{7} \end{aligned}$$

Now we have the two issues of how to divide , and how to extract a square root in modular arithmetic. For division by 7, solve

$$7x \equiv 1 \pmod{11}$$

Then we can multiply times the multiplicative inverse, x, instead of dividing by 7. The modular equation has has a unique solution by the no zeros, no repeats property of a prime modulus. Here

$$7x \equiv 1 \pmod{11} \Longrightarrow x = 8$$

So instead of dividing by 7, multiply times 8.[3]

Similarly for the square root, we can solve

$$x^2 \equiv 5 \pmod{11}$$

for which there are two solutions: $x = 4$ and $x = 7$. But either solution will produce the same positions i and j, just relabeled.

For $x = 4$:

$$\begin{aligned} i &= (-1+4)\,8 \equiv 2 \\ j &= (-1-4)\,8 = -40 \equiv 4 \end{aligned}$$

For $x = 7$

$$\begin{aligned} i &= (-1+7)\,8 = 48 \equiv 4 \\ j &= (-1-7)\,8 = -64 \equiv 2 \end{aligned}$$

Use either pair to get the error amounts a and b.

$$b = \frac{is_1 - s_2}{i - j} = \frac{2\,(5) - 5}{2 - 4} \equiv \frac{5}{9}$$

Once again to divide by 9 use

$$9x \equiv 1 \Longrightarrow x = 5$$

So

$$b = 5\,(5) = 25 \equiv 3$$

And finally

$$a = s_1 - b = 5 - 3 = 2$$

The solution is as we predicted: an error of 2 in position 2, and error of 3 in position 4.

9.7 generator matrix and the fundamental theorem of linear algebra

As it is a little bit more difficult to construct than in the earlier examples, the generator matrix for the double error correcting code has yet to appear in the discussion. Some consideration of orthogonality and vector subspaces may ease the

[3]Here the equation for the multiplicative inverse can easily be solved by inspection. For larger moduli, Euclid's algorithm, as presented in chapter 10, will yield a solution.

analysis. The generator matrix is composed of the null space to the parity check matrix. The fundamental theorem of linear algebra, then, is useful for describing the construction of the two matrices.

Reconsider the single error correcting code specified by the parity check matrix, H.

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} (\mathrm{mod}\, 2)$$

H consists of 3 linearly independent y element vectors. According to the fundamental theorem the null space will have 4 independent 7 element vectors. That is, the dimension of the nullspace (here, 4) plus the dimension of the row space (3) will equal the 7 dimensions of the vectors. G, as constructed earlier in the chapter, has that form.

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} (\mathrm{mod}\, 2)$$

Orthogonality of the two spaces is assured by the parity-generator relationship.

$$HG \equiv 0$$

(For the syndrome $Hy \equiv 0$, we must have $HGx \equiv 0$ for all possible messages, x. Therefore, $HG \equiv 0$ must hold.)

As a tangential observation, there is a difference between orthogonal vectors in Euclidean spaces and spaces defined by modular arithmetic. A vector in the latter space can be orthogonal to itself; that is, the vector product can be zero. For example, all 3 vectors in H have this property. Hence, they belong to both the row space and the nullspace.

One implication of this perhaps seemingly odd state of affairs is that there are not a total of 7 independent vectors in the combination of the rows of H and the columns of G; the rows of H, for example, are not independent of the columns of G. In other words, the space of 7 element vectors is not spanned by the H,G combination, and not all vectors in the 7 element space can be formed by linear combinations of the row and null vectors.

Spanning is not an issue that arises in the coding problem. But it is central for the understanding of other topics we have considered: decomposition of journal entries, uniqueness of arbitrage free prices, and the mutual information theorem, for example. So it doesn't hurt to confront the concept of spanning on occasion.

Return to the double error correction example with

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 & 7^2 & 8^2 & 9^2 & 10^2 \\ 1 & 2^3 & 3^3 & 4^3 & 5^3 & 6^3 & 7^3 & 8^3 & 9^3 & 10^3 \end{bmatrix} \pmod{11}$$

A generator matrix composed of the appropriate number of orthogonal vectors (6 by the fundamental theorem) is

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 4 & 10 & 9 & 2 & 1 & 7 \\ 7 & 8 & 7 & 1 & 9 & 6 \\ 9 & 1 & 7 & 8 & 7 & 7 \\ 1 & 2 & 9 & 10 & 4 & 1 \end{bmatrix} \pmod{11}$$

The vectors in G can be found using standard linear techniques. After the 6×6 identity matrix component of G, each vector has 4 unknown elements, for a total of 24 unknowns. Each vector, in turn, must satisfy 4 orthogonality conditions with the rows of H, for a total of 24 equations. 24 independent linear equations with 24 unknowns is a little bit tedious, but once the answer is reduced to modulo 11, the above G appears.

Example 9.11 *Let the message be* $x^t = \begin{bmatrix} 5 & 3 & 0 & 6 & 8 & 0 \end{bmatrix}$. *Use the* G *matrix to append 4 redundant digits rendering the message amenable to double error correction.*

$$y = Gx = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 4 & 10 & 9 & 2 & 1 & 7 \\ 7 & 8 & 7 & 1 & 9 & 6 \\ 9 & 1 & 7 & 8 & 7 & 7 \\ 1 & 2 & 9 & 10 & 4 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \\ 0 \\ 6 \\ 8 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 0 \\ 6 \\ 8 \\ 0 \\ 70 \\ 137 \\ 152 \\ 103 \end{bmatrix} \equiv \begin{bmatrix} 5 \\ 3 \\ 0 \\ 6 \\ 8 \\ 0 \\ 4 \\ 5 \\ 9 \\ 4 \end{bmatrix} \pmod{11}$$

It can be verified that the syndrome $Hy \equiv 0$.

9.8 error correction and mutual information

The noisy channel theorem is typically characterized as Shannon's most important result. It laid the foundation for the technical devices of the information age, including smart phones, the internet, and on and on. The theorem is the reason Shannon, himself, if often referred to as the father of the information age.

For the plethora of electronic devices which communicate with each other, the communication channel is inherently error prone; disturbances in the message are common, even typical. Sending the message again doesn't appear to help, as errors occur in the repeated message, as well. And other redundant schemes are subject to the same problem: more symbols in a message means more errors.

Given this state of affairs, it didn't seem like a good idea to even try for error free communication through a noisy channel. But Shannon's noisy channel theorem changed everything. First, he *proved* error free transmission was possible through a noisy channel. He showed redundancy could work to bring the error rate arbitrarily close to zero. And he also proved how much redundancy was necessary to eliminate the errors.

His proof, however, was abstract and not constructive. that is, it didn't specify how to construct an error free code; it just said some must exist. But that was enough to get coding theorists to explore the field, and several efficient codes have been developed. Since Shannon had shown the minimum redundancy needed, actual codes are compared to the Shannon bound, and many, indeed, approach the bound.

We won't go through the theorem here, but we will show how error correcting codes like we have studied can significantly reduce error rates. Mutual information is the relevant measure of error rates. In particular, we know

$$0 \leq I(X;Y) = H(Y) - H(Y|X) \leq H(Y)$$

Dividing both sides by $H(Y)$ we have

$$0 \leq \frac{I(X;Y)}{H(Y)} \leq 1$$

The closer the ratio to one, the closer the communication is to error free. While Shannon showed a ratio of one is (virtually) possible, we will be content with illustrating how redundant codes can move the ratio closer to one.

Example 9.12 *Consider three possible messages, x, say 0, 1, or 2, with a 10% error rate. That is, if x=1 is sent, the received message, y, will be 1 (correct) with probability .9; 0 or 2 will be received each with probability .05. Messages of 0 or 2 will behave the same way. Compute the mutual information ratio*

$$\frac{I(X;Y)}{H(Y)}$$

if one symbol at a time is sent. Then compute the ratio sending 4 symbols at a time using the linear code with parity check matrix

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} \bmod 3$$

(Looking ahead, the ratio for one symbol at a time is about 64%. For the 4 symbol block code the ratio increases to 84%.)

First one symbol at a time. The conditional probabilities are as follows where x is the message sent, and y the message received.

		y		
	$p\,(y\|x)$	0	1	2
x	0	.9	.05	.05
	1	.05	.9	.05
	2	.05	.05	.9

The joint probabilities are used for the entropy computations; and to access the joints, the marginal probabilities, $p(x)$, are used. But the marginals are a decision variable in the code design problem. That is, the code designer can choose the states of the world which generate the message x_i, effectively choosing the probabilities $p(x)$. The objective here, then, is to choose the marginal probabilities which yield the largest mutual information ratio $I(X;Y)/H(Y)$.

The example has a symmetric conditional probability matrix, and, for that case, it is particularly simple to specify the optimizing marginal distribution $p(x)$. The answer is to choose p(x) to be uniform over the possible messages. For the example, $p(x_i) = 1/3$ for all i.

Theorem 9.2 *For a symmetric channel, that is, one where the conditional probability matrix $p(y|x)$ is symmetr4ic, the maximum information ratio is achieved with a uniform distribution on the messages, x.*

A version of the theorem is in Cover and Thomas on page 190 (theorem 8.2.1). Its application in numerical examples can easily be illustrated using a spreadsheet optimizer: when asked to maximize the information ratio, the optimizer returns a uniform marginal distribution. It is also instructive to query the optimizer when the channel conditional probabilities are not symmetric. Then, in general, uniformity is not optimal.[4]

Converting to joint probabilities using uniform marginals:

		y		
	$p\,(x,y)$	0	1	2
x	0	.3	$\frac{.05}{3}$	$\frac{.05}{3}$
	1	$\frac{.05}{3}$	.3	$\frac{.05}{3}$
	2	$\frac{.05}{3}$	$\frac{.05}{3}$	.3

[4]It is tempting to refer to the theorem as the "dog theorem," as it is illuminated by a semi famous Far Side cartoon by Gary Larson in which all dog barks are decoded as "hey." The dogs could be more informative if they "spread out" their messages.

Using the joint probabilities, $p(x,y)$, mutual information $I(X;Y)$ is computed the usual way.

$$\begin{aligned} H(X) &= H(Y) = \ln 3 \\ H(X,Y) &= -\left[.9\ln\frac{3}{10} + .1\ln\frac{1}{60}\right] \\ &= .9\ln 2 + .9\ln 5 - .9\ln 3 + .2\ln 2 + .1\ln 5 + .1\ln 3 \\ &= 1.1\ln 2 - .8\ln 3 + \ln 5 \end{aligned}$$

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= 2\ln 3 - 1.1\ln 2 + .8\ln 3 - \ln 5 \\ &= 2.8\ln 3 - 1.1\ln 2 - \ln 5 \end{aligned}$$

The mutual information ratio is

$$\frac{I(X;Y)}{H(Y)} = \frac{2.8\ln 3 - 1.1\ln 2 - \ln 5}{\ln 3} \simeq .641$$

Approximately 64% of the uncertainty in the sent message, y, is reduced by the receipt of the signal, x.

Before proceeding to using an error correcting code in the noisy channel, let's simplify the information ratio expression for the symmetric channel case. The general conditional probability matrix is

$$\begin{matrix} p & (1-p)/(n-1) & \cdots \\ (1-p)/(n-1) & p & \\ \vdots & & \ddots \end{matrix}$$

where p is the error free rate of transmission, and n is the number of possible messages. Invoking uniform marginals, the general joint probability matrix is

$$\begin{matrix} p/n & (1-p)/n(n-1) & \cdots \\ (1-p)/n(n-1) & p/n & \\ \vdots & & \ddots \end{matrix}$$

By symmetry

$$H(X) = H(Y) = \ln n$$

And the joint entropy:

$$\begin{aligned} H(X,Y) &= -\left[p\ln\frac{p}{n} + (1-p)\ln\frac{1-p}{n(n-1)}\right] \\ &= \ln n + H(p) + (1-p)\ln(n-1) \end{aligned}$$

$$\text{where } H(p) = -\left[p\ln p + (1-p)\ln(1-p)\right]$$

Therefore,

$$\begin{aligned}\frac{I(X;Y)}{H(Y)} &= \frac{H(X)+H(Y)-H(X,Y)}{H(Y)} \\ &= \frac{\ln n - H(p) - (1-p)\ln(n-1)}{\ln n}\end{aligned}$$

For the example so far we have $n = 3$ and $p = .9$. Plugging into the expressions yield

$$H(p) = H(.9) \simeq .325$$

and

$$\frac{I(X;Y)}{H(Y)} = \frac{\ln 3 - .325 - .1\ln 2}{\ln 3} \simeq .641$$

as before.

The next step is to use the error correcting code, and see if the amount of information passing through the noisy channel increases. Instead of one symbol at a time, the linear code will send blocks of 4 symbols: 2 symbols in the original message and 2 redundant. The generator matrix for the code is

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 2 \\ 2 & 1 \end{bmatrix} \bmod 3$$

So, for example, if the original message is

$$y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

the 4 symbol block sent is

$$x = Gy = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix} \bmod 3$$

The error correcting code will correct any single error, so the probability the message gets through the channel unchanged is the sum of the probability of no errors plus the probability of one error.

$$\begin{aligned}\text{Prob(no error)} &= .9^4 = .6561 \\ \text{Prob(one error)} &= .9^3.1^1(4) = .2916\end{aligned}$$

The general expression for t errors is

$$\text{Prob(t errors)} = .9^{4-t}.1^t\binom{4}{t}$$

where the third term in the expression is the number of ways the error can occur in the 4 symbol block.

The nine by nine conditional probability matrix can be constructed. On the diagonal is the probability of error free transmission

$$p = .6561 + .2916 = .9477$$

And, continuing with the symmetry frame, assume there is no information about y if more than one error occurs. Then the off-diagonal consists of

$$\frac{1 - .9477}{8} = .0065375$$

The entropy computations can be made from the symmetric probability matrices, or the more general relationships can be used for

$$p = .9477 \text{ and } n = 9$$

$$H(p) = H(.9477) \simeq .205$$

and

$$\frac{I(X;Y)}{H(Y)} = \frac{\ln 9 - .205 - .0523 \ln 8}{\ln 9} \simeq .857$$

As predicted in the statement of the example, the fraction of information getting through the channel increased from 64% to approximately 86% by using the single error correcting code. The redundancy rate was 50% since each 4 element block has 2 redundant symbols.

Example 9.12 began with a symbol error rate of 10% or, equivalently an error free transmission rate of $p = .9$. For different error free transmission rates, an improvement in the mutual information ratio also occurs. Below are tabulated some symbol error free rates along with information ratios for single symbol transmission and block transmission using the single error correcting of the example.

symbol error free rate p	.99	.975	.95	.9	.8
$\frac{I(X;Y)}{H(Y)}$ single transmission	.943	.878	.788	.641	.418
$\frac{I(X;Y)}{H(Y)}$ block transmission	.997	.986	.953	.857	.614

Shannon's powerful theorem states there exists a block code which achieves an error rate arbitrarily close to zero, but does not supply the code. For our purposes we will be satisfied with illustrating error correction and documenting improvements in information transmission rates.

9.9 summary

This chapter is the first to deal with the issue of how to preserve data integrity. The error detecting and correcting codes presented herein rely on two primary

academic tools. One is orthogonality and the concept of the nullspace, a tool used extensively in prior chapters. The other tool is number theory which will prove quite useful in subsequent chapters. In this chapter we got our first exposure to prime numbers and the fundamental theorem of arithmetic. We also presented a numerical example which improved the information transfer in a noisy channel, though not as much as Shannon's noisy channel theorem states is possible.

9.10 references

Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 1991.

Hill, Raymond, *A First Course in Coding Theory*. Oxford University Press, 1986.

Ore, Oystein, *Number Theory and Its History*. McGraw-Hill Book Company, 1948.

9.11 exercises

Exercise 9.1 *Here are some ISBN's - possibly in error. Check for accuracy.*

1. *Probability Theory* by E. T. Jaynes. 978-0-521-59271-0
2. *Essays in Accounting Theory in Honour of Joel S. Demski*. 0-387-30397-9 and 978-0387-30397-0
3. *Number Theory and Its History* by Oystein Ore. 0-486-65620-9

Exercise 9.2 *Fill in the missing digits in the following UPC numbers.*

$$\begin{array}{cccccccccccc} 0 & 7 & 0 & 5 & 5 & 4 & 0 & 0 & ? & 2 & 3 & 6 \\ 0 & 2 & 8 & 4 & 0 & 0 & 0 & 9 & ? & 9 & 0 & 1 \end{array}$$

Exercise 9.3 *Consider a "perfect" single error correcting code with parity check matrix*

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 5 & 6 & 0 & 1 \end{bmatrix} (\mathrm{mod}\ 7)$$

For the following received codewords, detect and correct a single error, if necessary.

$$\begin{array}{cccccccc} 1 & 5 & 3 & 6 & 1 & 3 & 1 & 6 \\ 3 & 2 & 2 & 0 & 0 & 3 & 1 & 2 \end{array}$$

Exercise 9.4 *For the same modulo 7 code as in the previous problem, append the appropriate redundant digits.*

$$\begin{array}{cccccc} 1 & 2 & 4 & 3 & 6 & 6 \\ 5 & 6 & 1 & 1 & 2 & 1 \end{array}$$

Exercise 9.5 *Using the same modulo 7 code in the previous exercises, verify the "perfectness" of the same code by computing the number of possible received codewords, the number of legal codewords, and the number of codewords within one change of a legal codeword.*

Exercise 9.6 *Consider Matthew 5:37. "Let your communication be Yea, yea; Nay, nay: for whatsoever is more than these cometh from evil." The communication is an implied binary code. What is the implied parity check matrix, and the generator matrix? Is the code error correcting? What about error detection?*

Exercise 9.7 *Consider a noisy symmetric channel which transmits one bit at a time, either a zero or a one. The error rate is 10%, or, alternatively, the error free transmission rate is 90%. What is the mutual information ratio for the channel? Now use the single error correcting with parity check matrix*

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \bmod 2$$

What is the mutual information ratio for the 7 element block?

Exercise 9.8 *Consider the double error correcting code in the chapter with parity check matrix*

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 & 7^2 & 8^2 & 9^2 & 10^2 \\ 1 & 2^3 & 3^3 & 4^3 & 5^3 & 6^3 & 7^3 & 8^3 & 9^3 & 10^3 \end{bmatrix} \pmod{11}$$

Suppose the syndrome is $Hy = s \equiv \begin{bmatrix} 3 & 3 & 3 & 3 \end{bmatrix}^T$. *What are the error(s) and position(s) thereof? Suppose* $Hy = s \equiv \begin{bmatrix} 10 & 1 & 2 & 8 \end{bmatrix}^T$.

Exercise 9.9 *Redo example 9.12 with symbol error rate of .025. That is, with probability .0125 one of the incorrect symbols is received, and the same probability for the other incorrect symbols.*

Exercise 9.10 *Consider a double error correcting code with the following parity check matrix.*

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 \\ 1 & 2^3 & 3^3 & 4^3 & 5^3 & 6^3 \end{bmatrix} \pmod{7}$$

Suppose the syndrome is $Hy = s \equiv \begin{bmatrix} 4 & 0 & 2 & 1 \end{bmatrix}^T$. *What are the error(s) and position(s) thereof?*

10
secret codes

Secret codes as a way to communicate privately has been an important activity for centuries. Modern commerce relies on the integrity of large data bases, and a common method of ensuring integrity is encryption (encoding the data so that it is meaningful only to someone familiar with the coding technique). Accountants have as much interest in maintaining data integrity as anyone. Further, given the skills acquired in the analysis of accounting problems so far, we are well positioned to study cryptography, the science of secret codes.

A fundamental result for encryption was worked out over 300 years ago by Pierre de Fermat. Another central result is even older, over 1,000 years from Euclid.[1] But the application to encryption had to await the development of powerful, high-speed computers. Cryptography is an intriguing mixture of old and new science.

10.1 Fermat's theorem

Pierre de Fermat was an amateur mathematician who explored the mysteries of numbers for the pure joy of discovery. It is ironic that some of his results should turn out to be of such great practical application. The practical use of Fermat's theorem is due to the fact that the relationship is true for all numbers, no matter how big they are. Computer encryption uses extremely large numbers.

Theorem 10.1 *Fermat's theorem*

[1] A discussion of Fermat's theorem and Euclid's algorithm is in Ore, 1948.

$$a^{p-1} \equiv 1 \pmod{p}$$

where p is any prime number, and a is any number not a multiple of p.

Example 10.1 *For an easy example, 3 is a prime number. Then checking Fermat*

$$2^{3-1} = 2^2 = 4 \equiv 1 \pmod{3}$$

Example 10.2 *17 is a prime number. Calculate* 12^{16}. *By Fermat when* 12^{16} *is divided by 17, the remainder is* 1.

12^{16} is already a fairly large number, but fortunately we can verify the example without actually calculating 12^{16}. We can go in small stages. $12^2 = 144$, and when 17 is divided into 144, the answer is 8 with a remainder of 8. The remainder is what we want.

$$12^2 \equiv 8 \pmod{17}$$

The next step is to calculate 12^4.

$$12^4 = \left(12^2\right)^2 \equiv 8^2 \pmod{17} = 64 \equiv 13 \pmod{17}$$

Similarly,

$$12^8 = \left(12^4\right)^2 \equiv 13^2 \pmod{17} \equiv (-4)^2 = 16 \equiv -1 \pmod{17}$$

Notice when it came to squaring 13, it was easier to use $13 \equiv -4 \pmod{17}$ and square -4 instead. One more step gets us to 12^{16}.

$$12^{16} = \left(12^8\right)^2 \equiv (-1)^2 = 1 \pmod{17}$$

The last line verifies Fermat for the example.

As Fermat is true for all prime numbers, there are an infinite number of examples, and it is a good idea to complete several of them. For now, though, it would be nice to demonstrate Fermat with a little more generality. That can be done, since the theorem follows directly from the two properties of multiplication using a prime modulus: there are no zeros and no repeats. For example, consider modulo 5. Multiply all the numbers less than 5 by 2, and then multiply the answers together.

$$\begin{aligned} & (2 \times 1)(2 \times 2)(2 \times 3)(2 \times 4) \\ \equiv \quad & 2 \times 4 \times 1 \times 3 \pmod{5} \end{aligned}$$

We know because of no zeros and no repeats, the numbers $(1, 2, 3, 4)$ will return after multiplying by 2, or indeed any non-zero number in the multiplication table. Only the order will change.

We can rewrite the result using the factorial $(!)$ operator.

$$2^4 (4!) \equiv (4!) \pmod 5$$

We can effectively divide out the factorial term.[2]

$$2^4 \equiv 1 \pmod 5$$

The above algebra verified Fermat for $p = 5$ and $a = 2$. The logic for any p and a is similar. Each step is allowed because of the two properties. The product never disappears because there are no zeros, and the two expressions are congruent because they are composed of numbers in the same set without repeats (although probably in a different order).

10.2 an encryption technique

Fermat's theorem suggests a secret coding technique. Express the message in numbers. Encode the message by raising to a high power. That, then, is the cyphertext sent to the receiver. Decoding is accomplished by raising the cyphertext to another power until we reach 1, as Fermat guarantees we will. Then multiply one more time and the message will return.

Denote the message by a, and let p be the prime modulus.

$$\begin{aligned} a^{p-1} &\equiv 1 \pmod p \qquad \text{(Fermat's theorem)} \\ a^{(p-1)n} &\equiv 1 \pmod p \\ a^{(p-1)n+1} &\equiv a \pmod p \end{aligned}$$

Encryption involves accomplishing the above in two steps. Let e be an encryption parameter, so the cyphertext is a^e. What is required is a decryption parameter, d, such that

$$a^{e^d} = a^{ed} = a^{(p-1)n+1} \equiv a \pmod p$$

The designer of the code has the freedom to pick any p and e. But d must satisfy

$$ed = (p-1)\, n + 1$$

Or, in congruence notation, the problem is, given p and e, solve the following congruence for d.

$$ed \equiv 1 \pmod{p-1}$$

[2]Actually, we can't strictly divide. However, we know (because of no zeros and no repeats) that there is some number, when multiplied by 4!, the product is congruent to 1 modulo 5.

The problem is actually not a difficult one to solve; a solution is achieved using Euclid's algorithm, one of the oldest (perhaps the oldest) algorithm in mathematics. Euclid's algorithm is the subject of the next section, so we will temporarily skip that part. For now, here's a small numerical example.

Example 10.3 *Let $p = 67$, $e = 35$, and $d = 17$. It is easy to verify that e and d will work as the encryption and decryption parameters.*

$$35 \times 17 = 595 = 9 \times 66 + 1 \equiv 1 \pmod{66}$$

With this small code any number less than 67 can be encrypted. The following tabulates the cyphertext form of messages from 2 to 29.

a	a^e	a	a^e
2	63	16	55
3	58	17	21
4	16	18	11
5	42	19	26
6	36	20	2
7	18	21	39
8	3	22	15
9	14	23	60
10	33	24	40
11	13	25	22
12	57	26	6
13	32	27	8
14	62	28	20
15	24	29	37

It is not too hard to verify individual entries, and they should be verified. For example, the calculations for $a = 2$ and $e = 35$ follow. (Notice the use of negative numbers when they are easier to raise to a power.)

$$\begin{aligned} 2^8 &= 256 \equiv 55 \equiv -12 \pmod{67} \\ 2^{16} &\equiv (-12)^2 = 144 \equiv 10 \pmod{67} \\ 2^{32} &\equiv (10)^2 = 100 \equiv 33 \pmod{67} \\ 2^{35} &= \left(2^{32}\right)\left(2^3\right) \equiv 33 \times 8 = 264 \equiv 63 \pmod{67} \end{aligned}$$

Decoding is accomplished with $d = 17$.

$$\begin{aligned} 63 &\equiv -4 \,(\mathrm{mod}\, 67) \\ 63^4 &\equiv (-4)^4 = 256 \equiv -12 \,(\mathrm{mod}\, 67) \\ 63^8 &\equiv 144 \equiv 10 \,(\mathrm{mod}\, 67) \\ 63^{16} &\equiv 100 \equiv 33 \,(\mathrm{mod}\, 67) \\ 63^{17} &\equiv 33 \times (-4) = -132 \equiv 2 \,(\mathrm{mod}\, 67) \end{aligned}$$

And the original message, $a = 2$, is returned.

The cyphertext, a^e, bears no systematic relationship to the message a. Because of this, it is difficult (probably impossible) to "break" the code, that is, invert from a^e to a. Another way to state the relationship is that there is no information in a^e about a. Indeed, many random number generators work by raising a large number to a large power and then reporting the remainder. In this sense a^e will look to the bad guys like random numbers, and will prove no use to them.

It will be easier to visualize the randomness in a^e when we get to examples with large numbers. For now, we have some unfinished business: solving for the decryptor, d.

10.3 Euclid's algorithm

The problem in the last section was to find a decryptor for the prime modulus, $p = 67$, and encryptor $e = 35$. It was easy to verify $d = 17$ does the trick. It satisfied the congruence

$$ed \equiv 1 \,(\mathrm{mod}\, p - 1)$$

Or, in this case,

$$\begin{aligned} 35d &\equiv 1 \,(\mathrm{mod}\, 66) \\ &\quad \text{or} \\ 35d &= 66n + 1 \end{aligned}$$

Euclid's algorithm provides a method to discover the answer in the first place. The method of Euclid starts with 66 and 35; then divides the smaller into the larger to get a remainder (here it is 31). Then the division is repeated with the smaller numbers 35 and 31. And repeated still further until the remainder is as small as it can get. (For there to be a solution to the congruence, the final remainder must be 1.) The equations are as follows.

$$\begin{aligned} 66 &= 1 \times 35 + 31 \\ 35 &= 1 \times 31 + 4 \\ 31 &= 7 \times 4 + 3 \\ 4 &= 1 \times 3 + 1 \end{aligned}$$

The last equation is in the form we want; we can substitute into that equation for 3, 4, and 31 to get everything in terms of 66 and 35. The substitutions to use are restatements of the first 3 equations above.

$$\begin{aligned} 31 &= 66 - 35 \\ 4 &= 35 - 31 \\ 3 &= 31 - 7 \times 4 \end{aligned}$$

Now substitute into the last equation for 3, 4, and 31 sequentially.

$$\begin{aligned} 4 &= 3 + 1 \\ 4 &= (31 - 7 \times 4) + 1 \\ 35 - 31 &= (31 - 7 \times (35 - 31)) + 1 \\ 35 - (66 - 35) &= ((66 - 35) - 7 \times (35 - (66 - 35))) + 1 \end{aligned}$$

Combining terms to get the coefficients of 66 and 35:

$$\begin{aligned} 2\,(35) - 66 &= 8\,(66) - 15\,(35) + 1 \\ 17\,(35) &= 9\,(66) + 1 \end{aligned}$$

The last expression is in the form we want; we can see that $d = 17$ and, also, we get $n = 9$ for free.

The algorithm works, but parts of it, especially the substituting part, can get tedious. It would be cool if there was a parsimonious way to conduct the algorithm, and, indeed, there is. It amounts to keeping three columns of numbers. The two left hand columns accomplish taking successively smaller remainders (going down), and the right hand column does the substitution (coming back up). Start the procedure with the two numbers of interest.

66	
35	

Divide the smaller number into the larger. The middle column has the number of time 35 goes into 66, and the remainder is placed in the first column.

66	
35	1
31	

Repeat the procedure until the remainder is 1.

66	
35	1
31	1
4	7
3	1
1	

Start the third column at the bottom by placing 1there.

66		
35	1	
31	1	
4	7	
3	1	
1		1

Now work up the third column. Each entry is calculated by multiplying the entry in the second column times the third column entry one row lower. Then add the element in the third column two rows lower. For the first calculation there is only one element in the third column, so the calculation is $1 \times 1 + 0 = 1$.

66		
35	1	
31	1	
4	7	
3	1	1
1		1

Repeat. The next calculation is $7 \times 1 + 1 = 8$.

66		
35	1	
31	1	
4	7	8
3	1	1
1		1

The next calculations are $1 \times 8 + 1 = 9$, and $1 \times 9 + 8 = 17$.

66		
35	1	17
31	1	9
4	7	8
3	1	1
1		1

This scheme goes pretty fast with a little practice as in another example or two.

Example 10.4 *Let the prime modulus p=103 and encryptor e=91. Find d.*

The first two columns yield successively smaller remainders.

102		
91	1	
11	8	
3	3	
2	1	
1		1

The elements in the third column are calculated by multiplying the associated second column element times the element in the third column down one row, then add the third column element down two rows.

102			102			102			102		
91	1		91	1		91	1		91	1	37
11	8		11	8		11	8	33	11	8	33
3	3		3	3	4	3	3	4	3	3	4
2	1	1	2	1	1	2	1	1	2	1	1
1		1	1		1	1		1	1		1

The verification is $37 \times 91 = 3367 = 33 \times 102 + 1 \equiv 1 \pmod{102}$. The example can be further verified by encrypting and decrypting to see if the original message returns.

Example 10.5 *Let $p = 103$, $e = 91$, and $d = 37$. Encrypt and decrypt $a = 2$.*

Encrypt:

$$\begin{aligned}
2^9 &= 512 \equiv 100 \equiv -3 \pmod{103} \\
2^{54} &= \left(2^9\right)^6 \equiv (-3)^6 = 729 \equiv 8 \pmod{103} \\
2^{36} &= \left(2^9\right)^4 \equiv 81 \pmod{103} \\
2^{91} &= \left(2^{54}\right)\left(2^{36}\right)(2) \equiv (8)(81)(2) = 1296 \equiv 60 \pmod{103}
\end{aligned}$$

Decrypt:

$$\begin{aligned}
60^2 &= 3600 \equiv -5 \pmod{103} \\
60^8 &= \left(60^2\right)^4 \equiv 625 \equiv 7 \pmod{103} \\
60^{24} &= \left(60^8\right)^3 \equiv 7^3 = 343 \equiv 34 \pmod{103} \\
60^5 &= \left(60^4\right)(60) \equiv (25)(60) = 1500 \equiv 58 \pmod{103} \\
60^{37} &= \left(60^{24}\right)\left(60^8\right)\left(60^5\right) \equiv (34)(7)(58) = 13,804 \equiv 2 \pmod{103}
\end{aligned}$$

One more example illustrates a potential problem that might arise and its solution.

Example 10.6 *Let $p = 103$ and $e = 29$. The regular calculation for d:*

102		
29	3	7
15	1	2
14	1	1
1		1

In this case things don't quite work: $7(29) = 203 = 2(102) - 1 \equiv -1 \pmod{102}$, but they are easily fixed. Simply change the sign:

$$-7\,(29) = -203 = -2\,(102) + 1 \equiv 1\,(\mathrm{mod}\,102)$$

d, then, is -7; a negative d is a little bit uncomfortable, so recall $-7 \equiv 95\,(\mathrm{mod}\,102)$, and $d = 95$ works fine.

$$95\,(29) = 2755 = 27\,(102) + 1 \equiv 1\,(\mathrm{mod}\,102)$$

10.4 a real example

Sending a meaningful message requires a large prime modulus, which, in turn, implies use of a computer. A large modulus, encryptor, and decryptor increase the security of the code, as well. In the example to follow the programming language is Mathematica, and none of the commands requires more than a second or two to execute, at least on the machine I am using.

Start with a sample message: "Transfer $1,000,000 to Nish's account." And we require the message in numerical form. It is not hard to write a crude routine which, in this case, changes each symbol into a two digit number.

a_1=582811242916152837826575747474757474743730253752192918932937111 3132531243038

"T" is transformed to '58', "r" to '28', "a" to '11', and so forth. Simple substitution cyphers such as this one are popular in literature: see, for example, "The Gold Bug" by Edgar Allan Poe and "The Adventure of the Dancing Men" by Arthur Conan Doyle. They are also common on the puzzle pages of daily newspapers. In any event, they are not very secure. To access Fermat's theorem we require some large numbers. The Mathematica command "Random[Prime,{1,10^100}]" returns a prime number between 1 and 1 followed by 100 zeros.

p=576563504677725908429689266756324378566695789080916336292327464555239249898664893571248350463104064 3

e=157097701213889009824735169518631225161867667059827638689881028117224089167923903477514008626613434 3

Mathematica's command for Euclid's algorithm is ExtendedGCD[p-1,e], and a decryptor is easily obtained.

d=156583223228157477640122193155489418367493398237803340740291317316071379698942880113164904441649483 1

A patient reader can type the numbers into a computer and verify that $ed \equiv 1\,(\mathrm{mod}\,p-1)$. It's more instructive, and much more fun, however, to use p, e, and d to encrypt and decrypt some messages. The Mathematica command to generate a cyphertext is PowerMod[a_1, e, p]; in other words, raise a_1 to the power e, and take the remainder using the modulus p. Denote the cyphertext as c_1.

c_1=253797145363964263562275011010487918931745091763108208427178164150295859043016764977086031552766065 7

To return the message a_1, it is simply a matter of running PowerMod once more, this time with d and c_1: PowerMod[c_1, d, p]. While it is certainly possible to

verify the example with the numbers given, it would be better to make up different examples.

The cyphertext, c_1, is essentially a random number, and does, indeed, look that way to an eavesdropper. A way to emphasize its random character is to make a trivial change in the original message: "Transfer $1,000,000 to Oish's account." The numeric version is hardly distinguishable from the original; indeed, it's hard to find the single digit at variance.

582811242916152837826575747474757474743730253753192918932937 1113132531243038

Using p and e to generate cyphertext c_2, however, results in a sequence with no apparent relationship to c_1.

c_2=411732176127370657160452505374305474075914705595495763853 8186298827008076055619317468196994629382843

Even if the bad guys had the original message in both encrypted and unencrypted forms, it would be of no help in deciphering cyphertext for a very similar message. This is a powerful property of Fermat encryption, and one which strengthens the security of the codes.

There is, however, one area where Fermat encryption is vulnerable. The vulnerability is known as the private key problem. In order to establish a code, it is necessary to communicate the prime modulus and the encryption number so that the sender and the receiver can coordinate the transmissions. But this original communication can not be sent over the derived secure Fermat channel, as the secure channel did not come into existence until after the communication of p and e (or d). The next section, and the next chapter, are different ways of confronting the private key problem.

10.5 public key encryption

One way of avoiding the private key problem is to publicly announce the modulus and the encryptor, but do so in a way that is of no use to people wishing to intercept coded messages. This is called public key encryption and relies on two things: a theorem from Euler, and the fact that some problems are hard to solve, even by high speed computers. First Euler's theorem.

Fermat's theorem works only for prime numbers, the logic requiring that there be no repeats or zeros in the multiplication table. Euler generalizes Fermat's result to non-prime numbers, as well, by restricting consideration to rows of the multiplication table which satisfy the no zeros and no repeats conditions. Recall

the multiplication table modulo 10 from chapter 9 reproduced here.

	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	0	2	4	6	8
3	3	6	9	2	5	8	1	4	7
4	4	8	2	6	0	4	8	2	6
5	5	0	5	0	5	0	5	0	5
6	6	2	8	4	0	6	2	8	4
7	7	4	1	8	5	2	9	6	3
8	8	6	4	2	0	8	6	4	2
9	9	8	7	6	5	4	3	2	1

The rows (and columns) associated with 1, 3, 7, and 9 satisfy the necessary properties; the other rows do not. The difference is that 1,3, 7, and 9 are relatively prime to the modulus 10.

Definition 10.1 *Two numbers are relatively prime if, when comparing their respective (unique) factorizations into prime numbers, there are no common factors.*

The factorization of 10 is 2×5. All the numbers less than 10 except for 1, 3, 7, and 9 have either a 2 or 5 in their factorization into prime numbers. Euler denoted the size of the set of numbers relative prime numbers to m and less than m as $\phi(m)$, called "phi" or the totient.

Theorem 10.2 *Euler's theorem: for any modulus m, and any number a which is relatively prime to m,*

$$a^{\phi(m)} \equiv 1 \pmod{m}$$

Verifying the theorem for $m = 10$ we have $\phi(10) = 4$ and

$$\begin{aligned} 3^4 &= 81 \equiv 1 \pmod{10} \\ 7^4 &= 2401 \equiv 1 \pmod{10} \\ 9^4 &= 6561 \equiv 1 \pmod{10} \end{aligned}$$

Euler's generalization allows for a way around the public key problem. When the modulus is prime, and the modulus and encoder are known, the bad guys can solve for the decoder using the congruence $ed \equiv 1 \pmod{p-1}$. Because of Euclid this is an easy congruence to solve. On the other hand, when the modulus is not prime the congruence to solve for the decoder is $ed \equiv 1 \pmod{\phi(m)}$.[3] This is also an easy congruence to solve but only when $\phi(m)$ is known. This is the part that trips up the bad guys: finding $\phi(m)$ requires knowing the prime factors of m.

[3] Note that for a prime number $\phi(p) = p - 1$.

Example 10.7 *Suppose $m = r \times s$, where r and s are both prime numbers. Then any multiple of r less than m is not relatively prime to m, since they have common prime factor r. There are $s - 1$ multiples of r less than m. Similarly there are $r - 1$ multiples of s less than m. The total of relatively prime numbers $\phi(m)$ is*

$$\begin{aligned} \phi(m) &= (m-1) - (r-1) - (s-1) \\ &= (rs-1) - r + 1 - s + 1 \\ &= (r-1)(s-1) \end{aligned}$$

Finding the prime factors of a large number, say 100 digits or more, is computationally hard and beyond the capacity of present day computers to find in a reasonable time. Anyone who wishes to receive secret messages can disclose publicly a modulus and an encoder, and be (sort of) confident that the decoder can not be inferred. The receiver can easily derive their own decoder by constructing the modulus as the product of large prime numbers. This is the essence of public key encryption.

Example 10.8 *Let $r = 911$ and $s = 1009$. also, let the encryptor, e, be 601. The code modulus, m, then, is*

$$m = rs = 911\,(1009) = 919,199$$

And the number of relatively prime numbers less than m is

$$\begin{aligned} \phi(m) &= (r-1)(s-1) = (910)(1008) \\ &= 917,280 \end{aligned}$$

Finding $\phi(m)$, of course, is the difficult part for an outsider unaware of the prime factors of m. (It is even a little bit difficult for this small example.) But once $\phi(m)$ is known, it is relatively simple to find the decryptor, d, that solves

$$ed \equiv 1 (\mathrm{mod}\, \phi(m))$$

The calculations using Euclid's algorithm are tabulated.

$917,280$		
601	$1,526$	$244,201$
154	3	160
139	1	41
15	9	37
4	3	4
3	1	1
1		1

d is computed to be 244,201 and is verified by the calculations.

$$\begin{aligned} 244,201\,(601) &= 146,764,801 \\ 160\,(917,280) &= 146,764,800 \end{aligned}$$

With some computer assistance, the code can be tested. For example,

$$\begin{aligned} 121^e &\equiv 405,061 \pmod{m} \\ 405,061^d &\equiv 121 \pmod{m} \end{aligned}$$

Technology advances, however, and factoring large numbers might not always be difficult enough to support public key encryption.

> How many computational steps are needed to find the prime factors of a 300-digit number? The best classical algorithm known would take about 5×10^{24} steps, or about $150,000$ years at terahertz speed. By taking advantage of innumerable quantum states, a quantum factoring algorithm would take only 5×10^{10} steps, or less than a second at terahertz speed. (M. Nielsen, *Scientific American*, May 31, 2003)

The ability to harness quantum processes has already begun. The effect of quantum capabilities on encryption is the subject of the next chapter.

10.6 infinitude of primes

Encryption techniques use as raw material large prime numbers. Further, the existence of lots of prime numbers is necessary both for designing a secret code and ensuring the code does remain, in fact, secret. This section has a couple of theorems verifying the existence of a sufficient number of large primes; in fact, there are infinitely many. Besides being relevant to our coding activities, there is something satisfying about peering into the domain of very large numbers and identifying the regularities, and even beauty in that domain.

As primes become sparse as the numbers get large (there is, on average, more and more distance between two primes), the concern might be that we'll run out of prime numbers. Fortunately, for coding and a variety of other applications, that concern is unfounded. The number of primes is infinite, and the proof thereof, from Euclid, is elegant.[4]

Theorem 10.3 *There is an infinity of prime numbers.*

First assume there exists a largest prime number, and then derive a contradiction, thereby showing the original assumption to be false. Suppose the largest prime number is P. Construct the product of all the primes up to and including P, and to the product add 1.

$$Q = (2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdots P) + 1$$

Q is not divisible by any of the primes used in its construction: the remainder is always 1. Then Q must itself be prime, or divisible by primes larger than P.

[4] See G. H. Hardy's discussion of the proof in *A Mathematician's Apology*.

Either way we have a contradiction to the assumption of a largest prime, and that assumption must then be false.

Another important theorem concerning the behavior of primes is known as the prime number theorem and was first stated by Gauss. The actual number of primes less than a particular number, n, is conventionally denoted as $PrimePi(n)$. For example, $PrimePi(10) = 4$,and the four prime numbers are 2, 3, 5, and 7. Gauss' insight was that $PrimePi(n)$ behaves a lot like $n/\ln n$.

Theorem 10.4 *Prime number theorem:*

$$\underset{t\to\infty}{Limit}\frac{PrimePi(t)}{t/\ln t} = 1$$

The prime number theorem gives another verification of the infinitude of primes, as the ration $t/\ln t$ is always increasing.

$$\begin{aligned}\frac{d}{dt}\frac{t}{\ln t} &= \frac{\ln t - 1}{(\ln t)^2} \\ &= \frac{1}{\ln t} - \frac{1}{(\ln t)^2} > 0\end{aligned}$$

Furthermore, we can see there are plenty of large primes. Just as 90% of the numbers below 10^t exceed 10^{t-1}, the same is approximately true for primes. 90% of the primes less than 10^{100}, say, exceed 10^{99}. Doing approximate computations

$$\begin{aligned}\frac{\frac{10^{100}}{100\ln 10} - \frac{10^{99}}{99\ln 10}}{\frac{10^{100}}{100\ln 10}} &\approx \frac{10^{100} - 10^{99}}{10^{100}} \\ &= 1 - .1 = .9\end{aligned}$$

The computation presumes $99\ln 10$ is "close enough" to $100\ln 10$; being a little more careful yields a result of 89.9%.

10.7 cyphertext entropy

A necessary condition of a good secret code is that the cyphertext (that is, the message that is sent after encryption takes place) should not be useful to the bad guys. One way to evaluate this characteristic of the cyphertext is to compute its entropy.

To investigate the entropy of messages in words, consider the name "kyle kerner." The name is composed of three e's, two k's and two r's, as well as one apiece of y, l, n, and a space, for a total of 11 symbols. The relative frequency of e, for example, is 3/11. Using relative frequencies instead of probabilities, applying the

entropy operator yields

$$\begin{aligned} & Entropy[\text{"kyle kerner"}] \\ = & -\left(\frac{3}{11}\ln\frac{3}{11}+\frac{4}{11}\ln\frac{2}{11}+\frac{4}{11}\ln\frac{1}{11}\right) \\ \simeq & \ 1.8462 \end{aligned}$$

Here is a quotation which appeared as an encrypted puzzle in the daily newspaper, presented for simplicity without upper case and punctuation.

> "the details vanish in the birdseye view but so does the birdseye view vanish in the details william james"

Using the same relative frequency technique as before, the computed entropy of the quotation is 2.649. Some perspective is added by comparing the actual entropy to the most the entropy could be . In this case the maximum possible entropy is ln 27, as there are 27 possible characters (26 letters and a space). And the ratio of actual to maximum is approximately 80%.

For the puzzle in the newspaper the cyphertext appears as follows.

> "axl hladgcj tdvgjx gv axl pgyhjlul tgle pka ji hilj axl pgyhjlul tgle tdvgjx gv axl hladgcj egccgdf odflj"

"t" in plaintext becomes "a" in cyphertext, "h" becomes, "x", and so forth. The entropy ratio of the cyphertext is the same 80%, of course. It is instructive to compute the ratio for a random sequence of symbols of the same length: for this example there are 103 characters.

A computer simulation experiment generating the ratio for 27 possible numbers generated randomly in lengths of 103 routinely returns an entropy ratio of about 95%, greater than the 80% entropy ratio in the puzzle. So the codebreaker (or newspaper puzzler) has a significant advantage just with the relative frequency of individual symbols. There are, of course, other useful patterns besides individual letters, including word patterns, letter patterns within words, and so forth.

The next part of the experiment involves a computer, as well. The idea is to check the entropy ratio for cyphertexts using the encryption systems under consideration in this chapter. Here we generate cyphertext using a prime modulus and an encryptor of 100 digits. The entropy ratio is computed as the entropy of the cyphertext divided by ln 10, as there are 10 possible digits. For the cyphertext the ration is routinely about 98%. For a list of 100 random integers, again the ration is about 98%. Indeed, many computerized random numbers are generated by raising a seed number to a very high power, divide by another large number, and report the remainder: effectively this reproduces the transformation of plaintext to cyphertext.

So the relative frequency wedge available to the codebreaker with substitution cyphers does not exist for the "Fermat encryption" secret codes in this chapter. Of course, as mentioned earlier, there are a variety of other possible patterns. But they can be checked in a similar fashion, and the interested reader with access to some computer capability is encouraged so to do.

10.8 summary

The process of sending secret messages, or maintaining the secrecy and integrity of databases, has changed as technology has changed. Centuries old mathematical results from Fermat, Euler, and Euclid have become central to secret codes as computational technology has improved. Technology continues to improve; encryption follows along, and different mathematical results become important. In the next chapter quantum processes and their effects on encryption are discussed.

10.9 references

Doyle, Arthur Conan, "The Adventure of the Dancing Men," in The Return of Sherlock Holmes, 1903.

Hardy, G. H., *A Mathematician's Apology*.

M. Nielsen, *Scientific American*, May 31, 2003

Ore, Oystein, *Number Theory and Its History*. McGraw-Hill Book Company, 1948.

Poe, Edgar Allan, "The Gold Bug," 1843.

10.10 exercises

Exercise 10.1 *Consider a private key secret code with prime modulus $p = 2633$ and encoder $e = 43$. What is the decoder, d? Suppose the received cyphertext is 2477. What is the message? (Use an Excel spreadsheet and the function "mod.")*

Exercise 10.2 *Using Fermat's theorem, verify that 7387 is not a prime number. Consider a multiplication table with modulus equal to 7387. Find some entries in the multiplication table which are zero. (An Excel spreadsheet will be useful.) Using what you learned from the position of the zero entries, estimate a value for Euler's totient (phi) by eliminating the affected numbers. Use Euler's theorem to verify (or disprove) your estimate of phi.*

Exercise 10.3 *For $p = 97$ and $e = 37$, find the decoder, d. Verify that the code works by encoding and decoding a message, say $a = 2$. Excel might be useful for the verification.*

Exercise 10.4 *Repeat the previous exercise for $p = 1009$ and $e = 97$.*

Exercise 10.5 *Consider modulus $m = 33,277 = (107)(311)$. The code encryptor is $e = 2003$. What is the decryptor, d?*

Exercise 10.6 *The code parameters are $p = 79$ and $e = 41$. Compute d. Compute the cyphertext for plaintext $a = 2$.*

Exercise 10.7 *From the proof of the infinitude of primes, when the first seven primes are multiplied together, and 1 is added to the product, the result is either prime or divisible by some prime number greater than the seventh prime. Which is it?*

How hard would it be to repeat the exercise for the first ten primes?

Exercise 10.8 *Consider a secret code modulus $m = 221 = (13)(17)$. Let the code encryptor be $e = 97$. Find the decryptor d, and encrypt, and then decrypt, a few messages. Do you see any weakness(es) with this code? Does the phrase "fixed point" have any relevance?*

Exercise 10.9 *Using the prime number theorem, approximately how many prime numbers are between* 10^{200} *and* 10^{201}*?*

Exercise 10.10 *What is the relative frequency entropy of "sandy koufax"? "leonhard euler"?*

11
quantum cryptography

In the preceding chapter we confronted the private key dilemma: once the private key was in the receiver's hands, the code is quite secret, but how to communicate the key itself? The progress made by using Euler's theorem could be erased by the power of quantum computers to factor very large numbers. But quantum thinking, itself, offers a solution. As stated by Nielsen and Chuang, "what quantum mechanics takes away with one hand, it gives back with the other: a procedure known as quantum cryptography ... exploits the principles of quantum mechanics to provide provably secure distribution of private information."[1]

Quantum physics allows a return to Fermat encryption, as it offers a secure communication scheme, so the secret code parameters, p and e, can be sent even though the code, itself, is not in place. Given the ability to send the code parameters, why not go ahead and send the secret message, itself? The answer resides in the probabilistic nature of quantum communication, and will, hopefully, become clear as the discussion proceeds.

The probabilistic nature of quantum communication is determined by three axioms; linear algebra foundations will prove to be quite useful.

11.1 quantum axioms

Quantum physics can be stated in an axiomatic structure; that is, all the results can be shown to follow logically from a limited number of axioms. There are only four

[1] Page 582, Quantum Computation and Quantum Information, Nielsen and Chuang.

axioms necessary, as in Nielsen and Chuang, but three will be enough for us in this chapter. The fourth deals with combining quantum units, and our cryptography excursion only requires us to work with one unit at a time; combination will be important, however, in the next chapter on synergy and production. The axioms take us to some strange places, and, for that reason, some physicists argue the structure is somehow incomplete. Nevertheless, the axioms are *very* predictive for physical phenomena and applications such as cryptography.

11.1.1 superposition

A quantum unit can be thought of as an electron or a photon or any other subatomic particle. The first axiom states that a quantum unit can be represented by a two element vector.

Definition 11.1 *a qubit is a two element vector with unit length.*

"Qubit" is short for quantum bit, and is distinguished from a classical bit (one or zero) in that it requires two numbers to describe.

Example 11.1 *Some examples of qubits are*

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Notice the example qubits all possess the unit length property, that is, vector multiplication of the transpose times the vector yields one.

$$\begin{aligned} \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= 1 \\ \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}^T \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} &= \frac{1+1}{2} = 1 \end{aligned}$$

Actually, the length of a vector is defined as the square root of the transpose times the vector. Taking the square root doesn't matter when the answer is one, but sometimes it will, and we will have to be more careful.

Definition 11.2 *A complex number is written in the form $a+bi$ where a and b are real numbers and $i = \sqrt{-1}$ and is called an imaginary number.*

Qubits can be described using complex numbers.[2] For example, valid qubits are

$$\begin{bmatrix} i \\ 0 \end{bmatrix} \quad \frac{1}{\sqrt{2}}\begin{bmatrix} i \\ -i \end{bmatrix}$$

[2]We won't really need complex qubits to accomplish quantum cryptography, but we will use them later when Euler's formula is in view. Besides they are just too beautiful to ignore.

For the calculation of the length of a vector with imaginary elements to make sense, a modification of vector multiplication is required.

Definition 11.3 *When transposing a vector (or matrix) with complex elements, change the sign on the imaginary part; this is called the complex conjugate of the vector.*

Use the complex conjugate to compute the length (actually the squared length) of the "imaginary" qubits.

$$\begin{aligned} \begin{bmatrix} i \\ 0 \end{bmatrix}^H &= \begin{bmatrix} -i & 0 \end{bmatrix} \\ \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ -i \end{bmatrix}^H &= \frac{1}{\sqrt{2}} \begin{bmatrix} -i & i \end{bmatrix} \end{aligned}$$

The superscript H means take the conjugate transpose; H stands for "Hermitian." when the vector (or matrix) has no complex elements, H means just the regular transpose. Now the computations of the length make sense.

$$\begin{aligned} \begin{bmatrix} i \\ 0 \end{bmatrix}^H \begin{bmatrix} i \\ 0 \end{bmatrix} &= \begin{bmatrix} -i & 0 \end{bmatrix} \begin{bmatrix} i \\ 0 \end{bmatrix} = -i^2 = 1 \\ \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ -i \end{bmatrix}^H \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ -i \end{bmatrix} &= \frac{1}{\sqrt{2}} \begin{bmatrix} -i & i \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ -i \end{bmatrix} = \frac{-i^2 - i^2}{2} = 1 \end{aligned}$$

The first axiom is called "superposition" because, as hard as it is to visualize, the qubit actually can be described by both numbers simultaneously. The numbers can describe the position, the charge, the angle or some other property of a subatomic unit. The simultaneous position, for example, is only resolved when the qubit is observed or "measured." Measurement is the third axiom; the next axiom describes how the quantum state is transformed to another state.

11.1.2 transformation

Matrix algebra and, in particular, matrix multiplication, describes the transformation of a quantum state.

Definition 11.4 *Transformation of a qubit occurs by multiplication by a 2×2 matrix; the operation maintains the unit length property of the qubit.*

Example 11.2 *Three important matrix transformations are known as the Pauli matrices (for Wolfgang Pauli).*

$$\begin{aligned} X &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ Y &= \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \\ Z &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \end{aligned}$$

Example 11.3 *Another important transformation is the Hadamard matrix (for Jacques Hadamard).*[3]

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Each example matrix is multiplied by the $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ qubit to check that unit length is preserved. Other qubits work the same way, and are worth checking, as well.

$$\begin{aligned} X \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ Y \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= \begin{bmatrix} 0 \\ i \end{bmatrix} \\ Z \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ H \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

The X transformation is known as a "bit flip," as it flips the position of the bit. In the laboratory an X flips the polarity or some other property of a sub-atomic particle. Similarly H is known as a "beam splitter," and in the lab a beam of photons, for example, can be split into two parts. Z represents a "phase flip," as it changes the sign on part of the qubit, more easily seen in the following transformation.

$$Z \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

11.1.3 measurement

Measurement of a qubit is accomplished by projection (regression) of the qubit vector into orthonormal vectors called a basis. For encryption purposes we require

[3] We've used the symbol H for at least four things so far: the parity check matrix, Hermitian (on the previous page), entropy, and Hadamard transformation. However, the usage is standard, and usually (we hope always) the meaning is clear from the context.

only 2 bases; we will use the standard basis, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, as well as the Hadamard basis, $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

The result of measuring a qubit is one or the other of the orthonormal vectors. That is, measurement changes the qubit being measured. This surprising axiom is very important for quantum processes, in general, and quantum cryptography, in particular.

The big question, then: which of the orthonormal vectors does the measured qubit turn into? The answer is in terms of probabilities, and that is another mysterious property of quantum physics. We are used to computing R^2 and interpreting it, when vector b is projected into vector a, as the component of b that resides in a. But now the vector b we are projecting is a quantum unit, and can not be physically decomposed any further. Quantum units are the basic elements of the universe. So we can't speak of a component of b, therefore we interpret R^2 as a probability that a particular measurement occurs. The probability interpretation is allowed, since, as we have seen, R^2 is a number between zero and one.

Definition 11.5 *Measurement of a qubit by orthonormal vectors changes the qubit into one of the orthonormal vectors. The probability of a particular vector being the result is the vector product of the resulting projection with itself (squared length of the projection).*

Example 11.4 *Measure the qubit* $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ *by the standard basis. That is, use the orthonormal vectors of the standard basis.*

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

The result of the measurement will be one of the orthonormal vectors. To compute the probabilities, project the measured qubit into the two vectors. First, project $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$*, call it* b*, into the first eigenvector* $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$*, called* a*. The orthogonality conditions when vector* b *is projected into* a *is*

$$\begin{aligned} a^T (b - a\beta) &= 0 \\ \beta &= \frac{a^T b}{a^T a} \end{aligned}$$

Solving for the regression coefficient β*:*

$$\beta = \frac{a^T b}{a^T a} = \frac{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}}{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}} = \frac{1}{\sqrt{2}}$$

Notice the denominator is always one, as the orthonormal vector is unit length by definition. This saves us some work: only the vector product of the qubit and the basis vector needs to be calculated.

The resulting projection, then, is

$$a\beta = \frac{a^T b}{a^T a} a = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The probability the result of the measurement is $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is the squared length of the projection which is R^2, since the denominator is the hypotenuse of the right triangle, and, in this quantum case, has unit length.

$$R^2 = (a\beta)^T (a\beta) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{1}{\sqrt{2}} = \frac{1}{2}$$

More generally,

$$(a\beta)^T (a\beta) = \beta^T a^T a \beta = \beta^2$$

since a has unit length. So, for the quantum case, where the vectors have unit length, the computations associated with projecting vector b into a are quite simple.

$$\begin{aligned} \beta &= a^T b \\ R^2 &= \beta^2 \end{aligned}$$

Similar computations show the measurement resulting in the other basis vector is also one-half, as it must be for the probabilities to sum to one.

R^2, as we said, is now a probability because of the quantum properties of the vectors. That is, it is certainly possible to write mathematically the initial vector as a weighted sum of the two basis vectors.

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

But this is not possible physically. The result of the measurement can not be the combination of two indivisible units. It must, instead, be one of the basis vectors or the other. So R^2 does not describe the proportion, but the probability.

Example 11.5 *Measure* $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ *by the standard basis.*

The regression coefficient β when the qubit is projected into the first basis vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is

$$\beta = a^T b = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1$$

The squared length of the projection is

$$R^2 = \beta^2 = 1$$

So the result of the measurement is the first basis vector with certainty. Whenever the qubit is identical to one of the basis vectors, measurement will *always* yield that basis vector.

Example 11.6 *Measure* $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ *by the Hadamard basis.*

The Hadamard basis vectors are

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Projecting the qubit into the first basis vector

$$\beta = a^T b = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}}$$

Compute the squared length of the projection to find the probability.

$$R^2 = \beta^2 = \frac{1}{2}$$

Each basis vector has equal probability of being the result of the measurement.

11.2 Dirac notation

Dirac notation was developed by Paul Dirac. It is disconcerting, at least, to change the notation at this stage. Nonetheless, there are real advantages to doing so. The notation simplifies writing things down, especially when the problems get complicated. And sometimes the notation actually illuminates the process. But no matter how we feel about it, the notation is standard in the area, so, if we wish to read about what people are up to, we'll need to follow the notation: everyone uses it.

The standard qubit $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is written as $|0\rangle$, and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ as $|1\rangle$. A general qubit $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ can be written

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \alpha |0\rangle + \beta |1\rangle$$

To transpose a qubit, or complex conjugate transpose, simply reverse the brackets.

$$\begin{aligned}\begin{bmatrix} 1 \\ 0 \end{bmatrix}^T &= \langle 0| \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T &= \langle 1| \end{aligned}$$

It is a bit easier to write vector products.

$$\begin{aligned}\begin{bmatrix} 1 & 0 \end{bmatrix}^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} &= \langle 1|\, 0\rangle = 0 \\ \begin{bmatrix} 0 & 1 \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= \langle 0|\, 1\rangle = 0 \end{aligned}$$

Whenever the two qubits in a vector product are the same, the result is one.

$$\begin{aligned}\langle 1|\, 1\rangle &= 1 \\ \langle 0|\, 0\rangle &= 1 \end{aligned}$$

The notation for the Pauli transformations are simplified a little bit.

$$\begin{aligned} X\,|0\rangle &= |1\rangle & X\,|1\rangle &= |0\rangle \\ Y\,|0\rangle &= i\,|1\rangle & Y\,|1\rangle &= -i\,|0\rangle \\ Z\,|0\rangle &= |0\rangle & Z\,|1\rangle &= -\,|1\rangle \end{aligned}$$

The Hadamard transformation introduces two new symbols, $|+\rangle$ and $|-\rangle$.

$$\begin{aligned} H\,|0\rangle &= \frac{|0\rangle + |1\rangle}{\sqrt{2}} = |+\rangle \\ H\,|1\rangle &= \frac{|0\rangle - |1\rangle}{\sqrt{2}} = |-\rangle \end{aligned}$$

Notice that $|+\rangle$ and $|-\rangle$ are the orthonormal vectors of the Hadamard basis.

Projections are accomplished in Dirac notation. Project $|0\rangle$ (which we denote as vector b) into $|0\rangle$ (denoted as a).

$$\begin{aligned} a\beta &= a^T b a \\ &= \langle 0\,|0\rangle\,|0\rangle = |0\rangle \end{aligned}$$

Project $|0\rangle$ into $|+\rangle$.

$$\begin{aligned} \langle + \,|0\rangle \,|+\rangle &= \frac{1}{\sqrt{2}} \,|+\rangle \text{ since} \\ \langle + \,|0\rangle &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \end{aligned}$$

The squared length of the projection of $|0\rangle$ into $|+\rangle$ is

$$\frac{1}{\sqrt{2}} \langle + | + \rangle \frac{1}{\sqrt{2}} = \frac{1}{2}$$

Therefore, when $|0\rangle$ is measured using the Hadamard basis, there is an equal probability (equal one-half) of either $|+\rangle$ or $|-\rangle$ being the result of the measurement. This and similar measurements are the basis of the quantum encryption solution to the private key problem.

11.3 quantum encryption

Recall the private key problem is the inability to communicate the code parameters, p and e, in a secure fashion. Quantum encryption offers a solution to the private key problem because of the way quantum measurement works: measurement changes the qubit. Even if bad guys intercept the key, they can't help but change it. If evidence of tampering shows up, the key is discarded, and the procedure starts again.

Here's how the sender of the key proceeds. For simplicity start with a lot of $|0\rangle$ qubits. Then process each one using one of four possible procedures. So as to not reveal anything about the chosen preparation on an individual qubit, the sender can generate a string of random numbers to decide which procedure to use on each qubit. There are four possible procedures.

	procedures:
1	unchanged = $\|0\rangle$
2	$X \,\|0\rangle = \|1\rangle$
3	$H \,\|0\rangle = \|+\rangle$
4	$HX \,\|0\rangle = \|-\rangle$

The sender, then, has a string of qubits, each qubit is one of four possible states. The string is sent to the receiver. The receiver then measures each qubit in the string using either standard or Hadamard basis. The receiver might as well generate a string of random numbers to decide which measurement to use on which qubit, as he (and any bad guys) have no information about how the qubits were prepared.

Notice that if the receiver measures with the standard basis for sender transformation procedures 1 and 2, the receiver sees exactly the qubit that was sent. Recall the standard orthonormal vectors are $|0\rangle$ and $|1\rangle$, and that, when the measured

qubit equals the basis vector, the basis vector is the result of the measurement with certainty.

However, still with sender procedures 1 and 2, if the receiver uses the Hadamard basis, then the qubit will be distorted. The result of the measurement will be $|+\rangle$ or $|-\rangle$, each with probability one-half, as demonstrated at the end of the previous section.

After the measurements are concluded the receiver announces publicly the measurement schedule employed That is, the receiver states which qubits were measured using standard and which using Hadamard. The receiver most assuredly does *not* announce the result of the measurement; the whole point is to keep that information from the bad guys.

The sender then compares the qubit preparation procedures with the announced measurement schedule.. If the measurement was standard, and the preparation either procedure 1 or 2, then the sender says (publicly) keep the qubit as part of the key. Likewise, Hadamard measurement will be matched with either procedure 3 or 4. The other qubits are discarded. In this fashion as many qubits as desired can be communicated, and the key, that is p and e, can be arbitrarily long. Notice the public communications are of no use to the bad guys. All that can be learned is what measurements are appropriate for what qubits, *not* the results of the measurements. It is too late, at this point, to actually conduct the measurements as the qubits have already been transmitted.

An important part of the procedure is what happens if a bad guy attempts to eavesdrop on the communication, that is, intercept the qubits while in transit. What happens, of course, is that with high probability the eavesdropper will alter the qubit. Even if the eavesdropper knows which *set* of measures to use, standard and Hadamard in this case, with probability one-half the wrong one is chosen. And repeated measurements increase the probability pretty quickly. If only ten qubits are measured, for example, the probability that none of them are altered is

$$\left(\frac{1}{2}\right)^{10} = .000977$$

The alteration free probability can be made as small as desired by sending more qubits. Using the communicated code parameters, p and e, to send a test message is one way to check to see if the qubits arrived without tampering. As seen in the previous chapter, even a minute variation in one of the code parameters results in a significant distortion to the message.[4]

[4]There is another issue not covered here: qubits are fragile and could become altered naturally, without interference from eavesdropping. That complicates the problem somewhat, but there are possible remedies; there exist quantum self-correcting codes, for example.

To summarize the procedures -

		receiver	
	sender procedures	measure	basis vectors
1	unchanged = $\|0\rangle$		
2	$X\|0\rangle = \|1\rangle$	std	$\|0\rangle$ and $\|1\rangle$
3	$H\|0\rangle = \|+\rangle$		
4	$HX\|0\rangle = \|-\rangle$	Had	$\|+\rangle$ and $\|-\rangle$

The string of qubits received and verified can serve as encryption parameters. The string can be read as a binary number, for example, and the next highest prime number chosen as p, or e. We have accomplished the objective of specifying a quantum private key communication procedure. Now the code parameters, p and e, can be used confidently in a regular Fermat encryption scheme.

Example 11.7

sender	*qubit*	*receiver*	*msmt*	*public communication*	
procedure	*sent*	*procedure*	*result*	*rec to sender*	*sender to rec*
null	$\|0\rangle$	*std*	$\|0\rangle$	*std*	*yes*
X	$\|1\rangle$	*Had*	$\|+\rangle/\|-\rangle$	*Had*	*no*
H	$\|+\rangle$	*std*	$\|0\rangle/\|1\rangle$	*std*	*no*
HX	$\|-\rangle$	*Had*	$\|-\rangle$	*Had*	*yes*

The sender prepares, and sends, four qubits as in the table. The receiver (randomly) chooses the measurement procedures. (A random choice prevents any eavesdropper from using the same measurement schedule.) The measurement result for the first qubit yields the qubit as sent. For the second qubit the measurement result is probabilistic: the notation $|+\rangle/|-\rangle$ implies either $|+\rangle$ or $|-\rangle$ will result, each with probability one-half. The last two columns are public announcements. The receiver sends the measurement schedule; the sender can then tell which qubits were measured appropriately, and tells the receiver which measurements to keep. Notice the public communications are worthless to an eavesdropper, as it is not possible to infer the measurement of the qubit.

Example 11.8 *Inject an attempt by an eavesdropper to intercept the quantum communication.*

sender	qubit	e'dropper	e'dropper	receiver	msmt	public comm.	
proc.	sent	procedure	msmt result	proc.	result	r to s	s to r
null	$\|0\rangle$	std	$\|0\rangle$	std	$\|0\rangle$	Z	yes
X	$\|1\rangle$	std	$\|1\rangle$	Had	$\|+\rangle/\|-\rangle$	X	no
H	$\|+\rangle$	std	$\|0\rangle/\|1\rangle$	std	$\|0\rangle/\|1\rangle$	Z	no
HX	$\|-\rangle$	std	$\|0\rangle/\|1\rangle$	Had	$\|+\rangle/\|-\rangle$	X	yes

Notice particularly the fourth qubit: the eavesdropper's attempt to intercept garbles the qubit, so what reaches the receiver is different from what is originally sent.

Even though the sender directs the receiver to use the fourth qubit, it will not work as part of a Fermat encryption technique, as a test message will determine.

11.4 summary

The quantum solution to the private key problem relies on quantum measurement in which the qubit being measured ends up as one of the eigenvectors of the measurement matrix. So any attempt to eavesdrop changes the message in a discernible fashion. In this way the code parameters, p and e, can be communicated securely, and Fermat encryption is back in.

In general the measurement result is probabilistic. For encryption purposes we only had use for discrete probabilities equal to zero, one-half, and one. Euler's formula allows us to examine settings with continuous probability results, and that will be employed in the next chapter.

11.5 reference

Nielsen, Michael A. and Isaac L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

11.6 exercises

Exercise 11.1 *Measure $|0\rangle$ using Hadamard basis. Report the possible vector results and the associated probabilities.*
Measure $HX|0\rangle$ using standard basis.
Measure $HX|0\rangle$ using Hadamard basis.

Exercise 11.2 *Fill in the missing elements of the following quantum communication table.*

sender procedure	qubit sent	receiver procedure	msmt result	public communication rec to sender	public communication sender to rec
null	$\|0\rangle$	Had	?	?	?
X	?	std	?	?	?
H	?	Had	?	?	?
HX	?	std	?	?	?

Exercise 11.3

sender proc	*qubit sent*	*e'dropper proc*	*e'dropper msmt result*	*receiver proc*	*msmt result*	*public comm. r to s*	*public comm. s to r*
null	$\|0\rangle$	*Had*	*?*	*std*	*?*	*?*	*?*
X	*?*	*Had*	*?*	*Had*	*?*	*?*	*?*
H	*?*	*std*	*?*	*std*	*?*	*?*	*?*
HX	*?*	*std*	*?*	*Had*	*?*	*?*	*?*

Exercise 11.4

sender procedure	*qubit sent*	*receiver procedure*	*msmt result*	*public communication rec to sender*	*public communication sender to rec*
null	$\|0\rangle$	*std*	*?*	*?*	*?*
X	*?*	*Had*	*?*	*?*	*?*
H	*?*	*std*	*?*	*?*	*?*
HX	*?*	*Had*	*?*	*?*	*?*

12
synergy and information

Synergy occurs when activities are combined in a particularly efficient way: when two activities are combined the output is strictly more than the sum of the outputs when the two activities are performed separately. As business combinations occur, information systems become available which, in turn, enhance productivity and efficiency. While information can be the cause of significant synergies, there are potential pitfalls, as well. The history of Enron provides examples of both.

In this chapter we will examine the role of information in the creation of synergies. The first examples will be based on the material covered in chapter 8 about entropy, in particular, how the mutual information theorem affects business combinations. The latter part of the chapter utilizes a production function observed in nature, wherein the information effects are particularly powerful and elegant. The illustrations are based on the material from chapter 11 on quantum encryption. Information synergy is a delicate phenomenon, in the quantum world as well as in a production environment, and improper measurement can easily lead to its destruction.

12.1 synergy and the mutual information theorem

Enron, at least in the early years, is a good example of synergy in business combinations based on information. Enron was originally an owner and operator of natural gas pipelines, and their first trading venture in natural gas contracts was quite successful. Their initial attempt to replicate that success in electricity trading did not work so well, since they did not have access to the information available

to a producer of electricity. As explained in "The Smartest Guys in the Room" by Bethany McLean and Peter Elkind (page 106):

> "And through their access to the nationwide electronic grid, utilities could tell in an instant when a plant anywhere in the country had gone down, a move that might spike a region's price in a matter of minutes. To put it another way, they had precisely the kind of information advantage Enron had in natural gas."

Subsequently, Enron acquired an electric utility (Portland General) and other physical assets to provide informational support for trading activities. Also from McLean and Elkind (page 110):

> "Enron's pipeline system revealed the secrets for making a fortune trading natural gas; the Portland General acquisition provide entree into the new world of electricity trading. Later Skilling bought paper mills so Enron could start trading pulp and paper, and a billion-dollar fiber network launched broadband trading. ... Enron would buy the infrastructure needed to crack the code."

Let's see if a numerical example will be helpful in explicating the information effects of combining activities in the Enron fashion.

Example 12.1 *Suppose there are two equiprobable states of the world, and there are two firms, P and S, with the following prospects in Arrow-Debreu format.*

P:	θ_1	θ_2
y	4	1
$p(y)$	$\frac{1}{2}$	$\frac{1}{2}$

S:	θ_1	θ_2
y	1	1
$p(y)$	$\frac{1}{2}$	$\frac{1}{2}$

In addition firm S has access to an information system in joint probability form.

$p(x,y)$	θ_1	θ_2
x_1	$\frac{1}{2}$	0
x_2	0	$\frac{1}{2}$

Using the Kelly criterion for investment, both firms have, on average, zero rate of return.

$$W(Y_P) = \frac{1}{2}\ln 4\left(\frac{1}{2}\right) + \frac{1}{2}\ln 1\left(\frac{1}{2}\right) = 0$$

For firm S the rate of return without the information is

$$W(Y_S) = \frac{1}{2}\ln\frac{1}{2} + \frac{1}{2}\ln\frac{1}{2} = -\ln 2$$

The mutual information of the information system is easily computed as

$$I(X;Y) = \ln 2$$

So with the information the rate of return for S is

$$W(Y_S|X) = W(Y_S) + I(X;Y) = 0$$

Append a cash flow sequence assumption so we can prepare steady state statements as in chapter 8. Assume the cash inflow for both firms is in period 3 along with our regular scaled investment of one dollar.

period	0	1	2	3
cash flow	-1	0	0	e^{3r}

If each firm invests one dollar in an asset every year, the steady state asset balance, B, for each is as computed in chapter 8:

$$B = e^r + e^{2r} = e^0 + e^0 = 2$$

And the steady state income for each firm is

$$inc = \sum CF - 1 = e^{3r} - 1 = 0$$

Now contemplate what happens if the two firms are combined. In particular, let firm P acquire firm S, and in so doing, P also acquires the information system. The P part of the combined firm will increase in value, as its growth rate has now increased to $ln2$.

$$\begin{aligned} B_P &= e^r + e^{2r} \\ &= e^{\ln 2} + e^{\ln 4} \\ &= 6 \end{aligned}$$

The acquisition of firm S, as it comes equipped with a useful information system, is a valuable proposition for P. Suppose the owners of P pay the owners of S the market value, that is the value of the discounted cash flow to P. That will be 4, as P's value will increase from 2 to 6. That will be the amount of "goodwill," and it is thought of as the excess paid over the market value of the assets acquired. P also pays for the assets of S with market value of 2 as computed above. The total amount paid is then 6, which shows up on the equity side of the combined balance sheet below as "new equity" as if it were a stock transaction. (It could, of course, be debt financed.) Notice the amount of enhanced value of P due to the information advantage shows up on the combined balance sheet as goodwill, so the total value of P is now 6.

combined balance sheet

assets P	2		
assets S	2	"new equity"	6
goodwill	4	equity	2
	8		8

But the information story does not end with goodwill being added to the consolidated balance sheet. It is also instructive to prepare divisional steady state income statements for divisions P and S. For division P steady state income is

$$\begin{aligned} inc_P &= \sum CF_P - 1 \\ &= e^{3r} - 1 \\ &= e^{\ln 8} - 1 \\ &= 7 \end{aligned}$$

and steady state rate of return is

$$\frac{inc_P}{B_P + C} = \frac{7}{6+1} = 1$$

which is $e^r - 1$ for $r = ln2$. For division S the steady state income remains at its precombination level of zero. But the information supplied by S is essential for P's enhanced income. Care must be taken, therefore, with performance evaluation. Compensation based on relative income might be problematic. If all the bonuses go to P's executives, for example, S might become reluctant, or even unwilling, to share crucial information.

As Enron's success was built upon the sharing of information, there was a possibility that relative performance problems might occur. Indeed, the following quote is from a book about Enron called "Conspiracy of Fools" by Kurt Eichenwald (page 462). The topic is Enron's practice of "forced rankings" whereby compensation, and even employment, is conditional on how an employee does relative to others.

> "...forced ranking was destroying the company. If everyone did a good job, the only way to move ahead was by undermining a colleague, but analysts needed to work as a team to get the best answers."

The themes of information synergies and performance evaluation will be further explored in the remainder of the chapter. The frame will be changed to a more explicit production setting: nature's production function using the quantum ideas from chapter 11 on quantum encryption.

12.2 Euler's formula

Before proceeding to a production setting using nature's production function, we require a really quite remarkable result known as Euler's formula.

Theorem 12.1 *Euler's formula:*

$$e^{i\theta} = \cos\theta + i\sin\theta$$

$$\text{where } i = \sqrt{-1}$$

We noted an expansion for e^x in chapter 6.

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + ...$$

There exist similar expansions for the trigonometric functions sine and cosine.

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ... \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + ... \end{aligned}$$

Euler worked out that using i raised to the appropriate powers allowed fitting the expressions together. The angle θ is typically denominated in radians rather than degrees. For example, 90 degrees is $\pi/2$ radians, 180 degrees is π radians, and so forth. A familiar case of Euler's formula is when it is evaluated at

$$\theta = \pi$$

The expression then contains the five most important numbers in mathematics, 0, 1, e, π, and i, as well as addition, exponentiation, and equality.

$$e^{i\pi} + 1 = 0$$

Euler's formula describes a circle when graphed in the complex plane with real numbers on the horizontal axis and imaginary numbers on the vertical axis, as in Graph 12. 1.

As complex numbers of the form $a + bi$ can be thought of as vectors in the complex plane of the form $\begin{bmatrix} a & bi \end{bmatrix}^T$, to keep the (squared) length of the vector positive, the sign of i is changed when doing the vector multiplication.

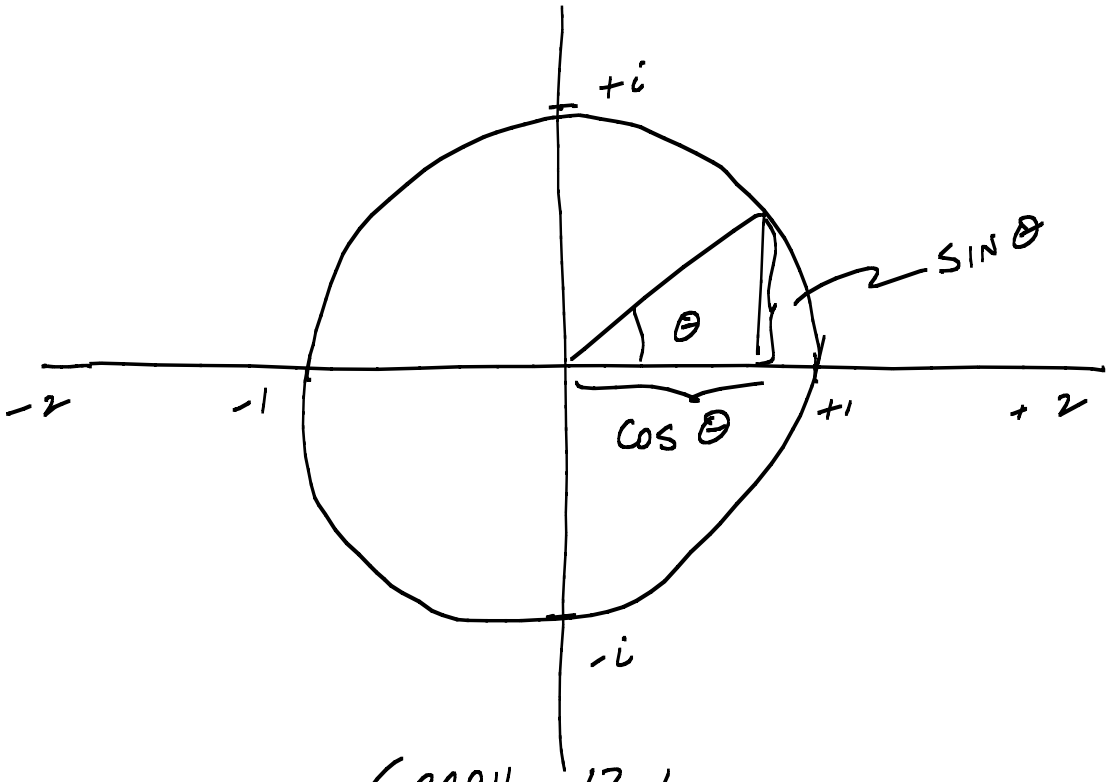
$$(a + bi)(a - bi) = a^2 - b^2 i^2 = a^2 + b^2$$

(Keeping the length positive by changing the sign on the imaginary term is the same problem, and solution, discussed in chapter 11 using the Hermitian transformation and the complex conjugate.) Accessing the Pythagorean theorem, we can see all the vectors in graph 12.1 described by Euler's formula have unit length.

$$\begin{aligned} & (\cos\theta + i\sin\theta)(\cos\theta - i\sin\theta) \\ = \;& \cos^2\theta - i^2\sin^2\theta \\ = \;& \cos^2\theta + \sin^2\theta = 1 \end{aligned}$$

12.3 quantum operations

Quantum units (qubits) are the basic building blocks of nature, and include particles like electrons and photons. In the previous chapter we defined qubits as two



GRAPH 12.1
EULER'S FORMULA IN THE COMPLEX PLANE

element vectors. In this chapter we exploit the Dirac notation where the two basic qubits are denoted as $|0\rangle$ and $|1\rangle$. The basic operations on qubits are transformation and measurement.

In the previous chapter quantum processes were derived from an axiomatic development. In this chapter the development is simplified somewhat, and it *might* be possible to do this chapter without studying the previous one, although checking with chapter 11 on occasion certainly would not hurt.

12.3.1 transformation

Quantum objects can be transformed using operations, some of which are denoted in the table below.

	quantum objects	
transformations	$\|0\rangle$	$\|1\rangle$
X	$\|1\rangle$	$\|0\rangle$
Y	$i\|1\rangle$	$-i\|0\rangle$
Z	$\|0\rangle$	$-\|1\rangle$
H	$(\|0\rangle + \|1\rangle)/\sqrt{2}$	$(\|0\rangle - \|1\rangle)/\sqrt{2}$
Θ	$e^{i\theta}\|0\rangle$	$\|1\rangle$

The transformations X, bit flip, Y, phase flip, and H, Hadamard or beam splitter, were introduced in chapter 11, as was Z. Θ is new for this chapter. We will require it for our production functions. The angle θ will be interpreted as the amount of labor put into production: the larger the angle, the greater the amount of labor.

Some of the transformations are written below in algebraic form.

$$\begin{aligned} X|0\rangle &= |1\rangle \\ Y|0\rangle &= i|1\rangle \\ H|0\rangle &= (|0\rangle + |1\rangle)/\sqrt{2} \end{aligned}$$

Sometimes, as in chapter 11, the result of $H|0\rangle$ is denoted $|+\rangle$, and the result of $H|1\rangle$ is $|-\rangle$.

12.3.2 measurement

In the previous chapter measurement of a qubit was accomplished by projection into a pair of orthogonal vectors, often $|0\rangle$ and $|1\rangle$. The result of the measurement was uncertain, and the probabilities were computed from the projection. Dirac notation simplifies the measurement calculation in a large variety of cases. Physically, measurement can establish the position of an electron which can be uncertain before the measurement is conducted.

A general qubit, before it is measured (or looked at carefully), can be a combination of the elementary objects.

$$\alpha|0\rangle + \beta|1\rangle$$

Once measurement occurs, the qubit will reduce to one or the other elementary object. It can no longer be an uncertain combination of the two. All qubits have unit length, meaning

$$\alpha^2 + \beta^2 = 1$$

So the probability that the result of the measurement will be $|0\rangle$ is α^2 and the probability a $|1\rangle$ will be the result is β^2. Since the numbers sum to one, this is a sensible probability measure.

A particular problem arises when a coefficient contains the imaginary number i. As $i^2 = -1$, negative numbers show up, and, of course, are unacceptable as probabilities. The problem is solved by using $i\,(-i)$ to "square" the i term, called the complex conjugate.

$$i\,(-i) = -\,(-1) = 1$$

In a production function a qubit of the following form will occur.

$$\frac{e^{i\theta}+1}{2}|0\rangle + \frac{e^{i\theta}-1}{2}|1\rangle$$

Euler's formula is used to compute the probability of $|0\rangle$ being the result from measurement. When "squaring" the coefficient, change the sign on the i term.

$$\begin{aligned} & \left(\frac{e^{i\theta}+1}{2}\right)\left(\frac{e^{-i\theta}+1}{2}\right) \\ = \quad & \frac{1+e^{i\theta}+e^{-i\theta}+1}{4} \end{aligned}$$

Now substitute Euler.

$$\frac{1+\cos\theta+i\sin\theta+\cos\left(-\theta\right)+i\sin\left(-\theta\right)+1}{4}$$

and recalling that

$$\begin{aligned} \cos\left(-\theta\right) &= \cos\theta \quad \text{and} \\ \sin\left(-\theta\right) &= -\sin\theta \end{aligned}$$

The squared coefficient is

$$\frac{2+2\cos\theta}{4} = \frac{1+\cos\theta}{2}$$

The derived expression is a fine probability measure, as it is always positive between 0 and 1. We can, however, convert to a more convenient form using Euler

once more.

$$\begin{aligned} e^{i(2\theta)} &= \left(e^{i\theta}\right)^2 \\ \cos(2\theta) + i\sin(2\theta) &= (\cos\theta + i\sin\theta)^2 \\ &= \cos^2\theta + 2i\sin\theta - \sin^2\theta \end{aligned}$$

Restricting attention to the real terms (terms without an i),

$$\cos(2\theta) = \cos^2\theta - \sin^2\theta$$

As a side note, it is easy to generate trigonometric identities like the above using Euler's formula. In this case, we know from Pythagoras that, for all angles

$$1 = \cos^2 + \sin^2$$

Adding the two equations together

$$1 + \cos(2\theta) = 2\cos^2\theta$$

Or rescaling the angle

$$\frac{1+\cos\theta}{2} = \cos^2\left(\frac{\theta}{2}\right)$$

And that is a nice expression for the probability that $|0\rangle$ is the result of the measurement. If the angle θ is zero, $|0\rangle$ will always be the result; as θ increases the probability $|1\rangle$ appears as the result increases. As the probabilities add to one, the probability of $|1\rangle$ being the result is

$$\sin^2\left(\frac{\theta}{2}\right)$$

We will use these results in the production function in the next section.

12.4 single unit production

In this section we develop a quantum production function. There are two productive inputs: material and labor. For the simplest process, let the material input be the $|0\rangle$ qubit. Further, let the labor input be an angle, θ, where, up to a point, the greater the angle, the greater or more productive the labor input. The one qubit production function is $H\Theta H|0\rangle$.

$$H\Theta H|0\rangle = \frac{\left(e^{i\theta}+1\right)|0\rangle + \left(e^{i\theta}-1\right)|1\rangle}{2}$$

If there is no labor input ($\theta = 0$), there is no production, and the output is $|0\rangle$: the input unit is what comes out. Successful production is when a different unit

$|1\rangle$ emerges from the production process. Recall from the previous section the measurement probabilities are

$$\begin{array}{cc} qubit & probability \\ |0\rangle & \cos^2(\theta/2) \\ |1\rangle & \sin^2(\theta/2) \end{array}$$

Restricting attention to angles between zero and 180 degrees,[1] the measurement is more likely to be $|1\rangle$ for greater θ. $\theta = 0$ yields no productive change: the measurement result is always $|0\rangle$. A 180 degree angle will produce $|1\rangle$ with certainty. For intermediate angles the probability is monotonically increasing. If production and measurement are conducted several times, greater effort will in probability yield a larger number of success ($|1\rangle$) measurements. An example allows using the production function to generate some accounting numbers.

Example 12.2 *Let the cost of labor be $.6\theta^2$, and, for simplicity, let the cost of the material input $|0\rangle$ be zero. Furthermore, let the revenue be 3 for a successful production $|1\rangle$ (and revenue of zero if output is $|0\rangle$). For an input angle of 60 degrees ($\frac{\pi}{3}$) the expected profit is computed as follows.*

direct labor cost :

$$.6\theta^2 = .6\left(\frac{\pi}{3}\right)^2 = \frac{\pi^2}{15}$$

expected revenue :

$$3\sin^2\frac{\theta}{2} = 3\left(\frac{1}{4}\right)$$

expected profit :

$$\frac{3}{4} - \frac{\pi^2}{15}$$

12.5 multiple qubits and entanglement

Dirac notation eases the transition to multiple qubit analysis. In fact, moving to two qubits is relatively painless. Instead of two objects, $|0\rangle$ and $|1\rangle$, they are now combined into four objects of two qubit pairs: $|00\rangle$, $|01\rangle$, $|10\rangle$, and $|11\rangle$. It is possible to operate on one of the qubits and leave the other unchanged. For example, H_1means conduct a Hadamard operation on the first qubit only, leaving the second part of the pair unchanged. Some two qubit operations are tabulated

[1] Notice this leaves unexplored the possibility of supplying "too much" labor.

below.

	objects		
transformations	$\|00\rangle$	$\|01\rangle$	$\|10\rangle$
H_1	$(\|00\rangle + \|10\rangle)/\sqrt{2}$	$(\|01\rangle + \|11\rangle)/\sqrt{2}$	$(\|00\rangle - \|10\rangle)/\sqrt{2}$
$CNOT$	$\|00\rangle$	$\|01\rangle$	$\|11\rangle$

	object
transformations	$\|11\rangle$
H_1	$(\|01\rangle - \|11\rangle)/\sqrt{2}$
$CNOT$	$\|10\rangle$

$CNOT$, called controlled not, is an important two qubit operation. When $CNOT$ operates on a two qubit system, the first qubit in the pair is the control qubit, and the second is the target. The target is flipped if, and only if, the control qubit is $|1\rangle$.

The use of $CNOT$, in conjunction with H_1, can produce a two qubit system so important it has its own name.

$$\begin{aligned} CNOT\ H_1|00\rangle &= CNOT\frac{|00\rangle + |10\rangle}{\sqrt{2}} \\ &= \frac{|00\rangle + |11\rangle}{\sqrt{2}} \\ &= |\beta_{00}\rangle \end{aligned}$$

In general,

$$CNOT\ H_1|ij\rangle = |\beta_{ij}\rangle$$

The resulting two qubit system $|\beta_{ij}\rangle$ is referred to as a Bell or EPR[2] state. It is also known as "entangled" qubits.

It is also possible to measure one qubit at a time. Consider the two qubit system

$$H_1|00\rangle = \frac{|00\rangle + |10\rangle}{\sqrt{2}}$$

and measure the two qubits in sequence. When the first qubit in the pair is measured, the result is either $|0\rangle$ or $|1\rangle$, each with probability of one-half, as that is the coefficient squared. Then, the second qubit is measured as $|0\rangle$ with probability one. The result of the second measurement is independent of the first measurement, as seems natural. The measurement sequence is illustrated in figure 12.1.

Now conduct the sequential measurement exercise on $|\beta_{00}\rangle$. The results are similar, but there is a striking and surprising difference as illustrated in figure 12.1.

[2]For Einstein, Podalsky, and Rosen, the authors of a paper describing the mysterious and (to them) unlikely properties of these mysterious qubits.

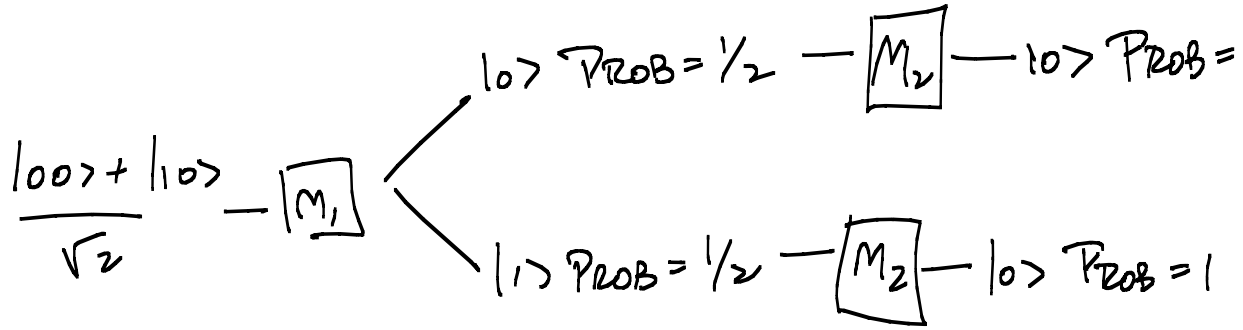


FIGURE 12.1
SEQUENTIAL MEASUREMENT

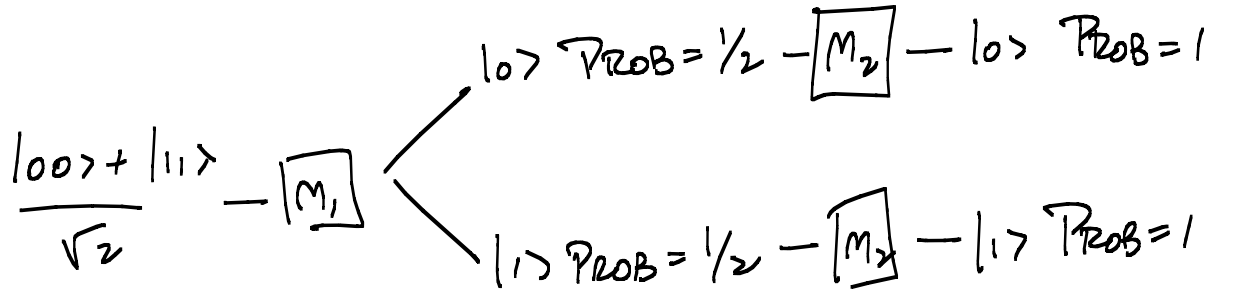


FIGURE 12.2
SEQUENTIAL MEASUREMENT
ENTANGLED QUBITS

The surprising thing is that the second measurement is no longer independent of the first. In fact, if the result of the first measurement is $|0\rangle$, then that is the result of the second with certainty. Similarly, a result of $|1\rangle$ in the first guarantees a $|1\rangle$ from the second. Einstein, among others, was skeptical of what he called "spooky action at a distance." In particular, the "at a distance" part was disconcerting, since there is nothing in the theory which requires the two qubits to be anywhere near each other. Something seems to happen to the far away one when measurement is applied to the close by one. But this and other surprising phenomena associated with entangled qubits have been consistently verified empirically. Nature appears to possess efficient resources to exploit interactions, and this one, termed entanglement, is a quite powerful one.

As mysterious as entanglement is in nature, perhaps it is not quite so surprising in a production context. It is, after all, the transfer of information from one unit to another. When production occurs, it is not beyond comprehension that when one person or department learns something about the production process, that knowledge is quickly available to other related units. In fact, the production process can be designed with fast and efficient transfer of information in mind. Entanglement is a mechanism to incorporate this phenomenon into the production process in a formal way.

12.6 synergy and multiple unit production

To get synergy to appear, we turn, not surprisingly, to entangled qubits, wherein exist powerful interaction effects. Consider a production process with two labor inputs denoted, as before as angles, θ_1 and θ_2, as well as direct material inputs of two $|0\rangle$ qubits, that is $|00\rangle$ This time, however, there is a common input cost which entangles the two qubits prior to production.

$$|\beta_{00}\rangle = CNOT\ H_1|00\rangle$$

The cost of entanglement is another cost pool in addition to the direct costs of labor and material. In the directed graph in figure 12.3 the common input factor is denoted cost pool 3.

A directed graph depicting parallel production of two outputs without the common input factor is in figure 12.4.

The entire production function has twice the direct costs as in the single product process, as well as the entanglement cost.

$$H_2\Theta_2 H_2 H_1\Theta_1 H_1 CNOT\ H_1|00\rangle$$

Successful production occurs when the entangled input,

$$|\beta_{00}\rangle = CNOT\ H_1|00\rangle$$

is transformed to an orthogonal state, namely $|\beta_{01}\rangle$. Of course, the probability of success is a function of the input angles.

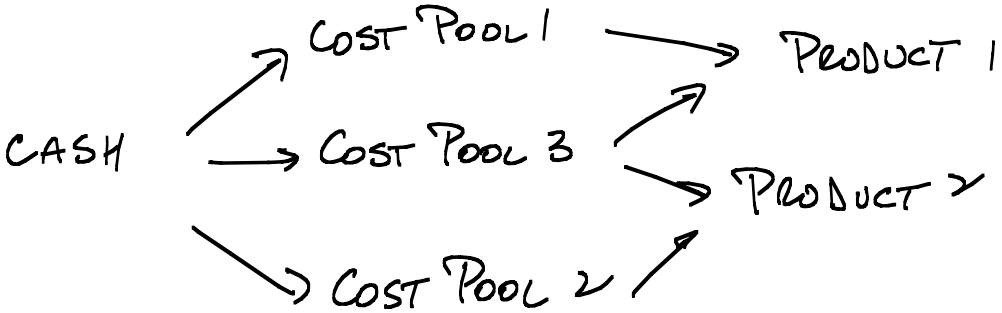


FIGURE 12.3
PRODUCTION WITH
COMMON INPUT FACTOR

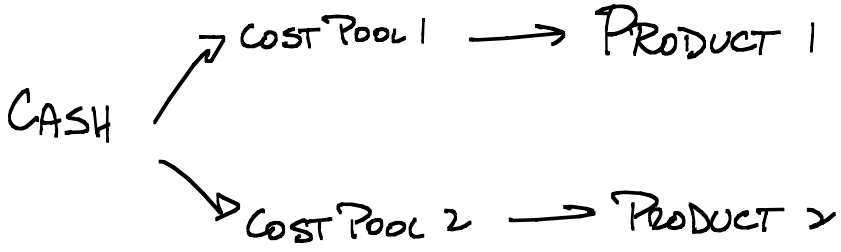


FIGURE 12.4
PRODUCTION WITHOUT
COMMON INPUT FACTOR

The computation of the success probability is not complicated, but a little bit tedious. The first part is derived below.

$$\begin{aligned}
& H_1\Theta_1 H_1|\beta_{00}\rangle \\
= \;& H_1\Theta_1 H_1 \frac{|00\rangle + |11\rangle}{\sqrt{2}} \\
= \;& H_1\Theta_1 \frac{|00\rangle + |10\rangle + |01\rangle - |11\rangle}{2} \\
= \;& H_1 \frac{e^{i\theta_1}\left(|00\rangle + |01\rangle\right) + |10\rangle - |11\rangle}{2} \\
= \;& \frac{e^{i\theta_1}\left(|00\rangle + |10\rangle + |01\rangle + |11\rangle\right) + \left(|00\rangle - |10\rangle - |01\rangle + |11\rangle\right)}{2\sqrt{2}} \\
= \;& \frac{\left(e^{i\theta_1}+1\right)}{2}\left(\frac{|00\rangle + |11\rangle}{\sqrt{2}}\right) + \frac{\left(e^{i\theta_1}-1\right)}{2}\left(\frac{|01\rangle + |10\rangle}{\sqrt{2}}\right) \\
= \;& \frac{\left(e^{i\theta_1}+1\right)}{2}|\beta_{00}\rangle + \frac{\left(e^{i\theta_1}-1\right)}{2}|\beta_{01}\rangle
\end{aligned}$$

Similarly we have

$$H_2\Theta_2 H_2|\beta_{00}\rangle = \frac{\left(e^{i\theta_2}+1\right)}{2}|\beta_{00}\rangle + \frac{\left(e^{i\theta_2}-1\right)}{2}|\beta_{01}\rangle$$

and

$$H_2\Theta_2 H_2|\beta_{01}\rangle = \frac{\left(e^{i\theta_2}+1\right)}{2}|\beta_{01}\rangle + \frac{\left(e^{i\theta_2}-1\right)}{2}|\beta_{00}\rangle$$

Using the three relationships we can evaluate the production process.

$$\begin{aligned}
& H_2\Theta_2 H_2 H_1\Theta_1 H_1|\beta_{00}\rangle \\
= \;& H_2\Theta_2 H_2\left[\frac{\left(e^{i\theta_1}+1\right)}{2}|\beta_{00}\rangle + \frac{\left(e^{i\theta_1}-1\right)}{2}|\beta_{01}\rangle\right] \\
= \;& \frac{\left(e^{i\theta_1}+1\right)}{2}\frac{\left(e^{i\theta_2}+1\right)}{2}|\beta_{00}\rangle + \frac{\left(e^{i\theta_1}+1\right)}{2}\frac{\left(e^{i\theta_2}-1\right)}{2}|\beta_{01}\rangle \\
& + \frac{\left(e^{i\theta_1}-1\right)}{2}\frac{\left(e^{i\theta_2}+1\right)}{2}|\beta_{01}\rangle + \frac{\left(e^{i\theta_1}-1\right)}{2}\frac{\left(e^{i\theta_2}-1\right)}{2}|\beta_{00}\rangle \\
= \;& |\beta_{00}\rangle\left(\frac{e^{i(\theta_1+\theta_2)} + e^{i\theta_1} + e^{i\theta_2} + 1 + e^{i(\theta_1+\theta_2)} - e^{i\theta_1} - e^{i\theta_2} + 1}{4}\right) \\
& + |\beta_{01}\rangle\left(\frac{e^{i(\theta_1+\theta_2)} - e^{i\theta_1} + e^{i\theta_2} - 1 + e^{i(\theta_1+\theta_2)} + e^{i\theta_1} - e^{i\theta_2} - 1}{4}\right) \\
= \;& \left(\frac{e^{i(\theta_1+\theta_2)}+1}{2}\right)|\beta_{00}\rangle + \left(\frac{e^{i(\theta_1+\theta_2)}-1}{2}\right)|\beta_{01}\rangle
\end{aligned}$$

Notice the similarity in form to the single input production function. In both cases the input qubit remains the same or is transformed to an orthogonal qubit.

The coefficients for the single and multiple qubit production processes are of the same form: for two agent production the coefficients are written with the sum of the two agent angles. The probabilities, then, are of the same form as the single input production. The only difference is the sum of the angles appears in the probabilities.

$$\begin{array}{cc} measurement & probability \\ |\beta_{01}\rangle & \sin^2\{(\theta_1+\theta_2)/2\} \\ |\beta_{00}\rangle & \cos^2\{(\theta_1+\theta_2)/2\} \end{array}$$

The orthogonal flip to $|\beta_{01}\rangle$ is considered a success in the multi-input (group) production function. The question is whether synergy exists, and that is easy to check. Whenever the individual angles sum to less than 180 degrees, the group success rate exceeds the sum of the individual probabilities. Some example angles are tabulated below.

θ_1	probability success θ_1	θ_2	probability success θ_2	sum individual probabilities	group success probability
30°	.067	30°	.067	.134	.250
30°	.067	45°	.146	.213	.371
30°	.067	60°	.250	.317	.500
30°	.067	90°	.500	.567	.750
45°	.146	30°	.067	.213	.371
45°	.146	45°	.146	.293	.500
45°	.146	60°	.250	.396	.629
45°	.146	90°	.500	.646	.854
60°	.250	30°	.067	.317	.500
60°	.250	45°	.146	.396	.629
60°	.250	60°	.250	.500	.750
60°	.250	90°	.500	.750	.933
90°	.500	30°	.067	.567	.750
90°	.500	45°	.146	.646	.854
90°	.500	60°	.250	.750	.933
90°	.500	90°	.500	1.00	1.00

Another way to see the synergy effect is to compute expected income in an extension of the production example.

Example 12.3 *Extend the previous example to a joint production setting. Append a cost of entanglement (common factor cost) K. Let the cost of labor be $.6\theta_1^2 + .6\theta_2^2$, and, for simplicity, let the cost of the material input $|00\rangle$ be zero. Furthermore, let the revenue be 3 for a successful production $|\beta_{01}\rangle$ (and revenue of zero if output is $|\beta_{00}\rangle$). For two input angles of 60 degrees ($\theta_1 = \theta_2 = \frac{\pi}{3}$) the expected profit is computed as follows.*

$$\begin{array}{rcl} \textit{direct labor cost} & : & \\ .6\theta_1^2 + .6\theta_2^2 & = & 1.2\left(\frac{\pi}{3}\right)^2 = \frac{2\pi^2}{15} \end{array}$$

expected revenue :

$$3\sin^2\frac{(\theta_1+\theta_2)}{2} = 3\sin^2\frac{(\pi)}{3} = 3\left(\frac{3}{4}\right)$$

expected profit :

$$3\left(\frac{3}{4}\right)-\frac{2\pi^2}{15}-K$$

Recall expected profit in the single product setting was

$$\frac{3}{4}-\frac{\pi^2}{15} \quad \textit{or for two products} \quad 2\left(\frac{3}{4}\right)-\frac{2\pi^2}{15}$$

So joint production is superior to two applications of single production as long as

$$K<\frac{3}{4}.$$

Recall K is the common cost of entangling the two qubits.

12.7 measurement implications

Now we come to the problem of assessing the responsibility (or blame) for the results of the production process. That is, we are particularly interested in the labor input angles performed by the production workers. If they are directly observable, there is no real problem. However, it is often the case that the labor inputs are not easily observable, and inferences about them must be made by measuring the observable outputs of production. As we have seen, when output is measured using the entangled Bell states, the probability a particular state is the measurement result is a function of the sum of the input angles: $\theta_1+\theta_2$. We can not, in other words, determine the individual input angles.

It is tempting to consider measuring individual qubits. After all, the first input angle θ_1 operates on the first qubit only, and similarly for θ_2. So if we observe the individual qubits, we might be able to infer something about the individual input angles. But, when synergy is important, this is a very dangerous path. In the first place, measuring individual qubits tells us *nothing* about the individual input angles. But it gets worse than that. Individual measures corrode the production process, and the synergy benefits are lost.

To examine the effects of individual measures, restate the production output in terms of basis qubits $|00\rangle$, $|01\rangle$, $|10\rangle$, and $|11\rangle$, as opposed to entangled Bell qubits.

$$\text{Let } \theta_1+\theta_2=\theta$$

$$\begin{aligned} & H_2\Theta_2 H_2 H_1 \Theta_1 H_1 |\beta_{00}\rangle \\ = & \frac{e^{i\theta}+1}{2}|\beta_{00}\rangle + \frac{e^{i\theta}-1}{2}|\beta_{01}\rangle \\ = & \frac{e^{i\theta}+1}{2}\frac{|00\rangle+|11\rangle}{\sqrt{2}} + \frac{e^{i\theta}-1}{2}\frac{|01\rangle+|10\rangle}{\sqrt{2}} \\ = & |00\rangle\frac{e^{i\theta}+1}{2\sqrt{2}} + |01\rangle\frac{e^{i\theta}-1}{2\sqrt{2}} + |10\rangle\frac{e^{i\theta}-1}{2\sqrt{2}} \\ & +|11\rangle\frac{e^{i\theta}+1}{2\sqrt{2}} \end{aligned}$$

Now measure the first qubit by "squaring" the coefficients. The probability the first qubit is $|0\rangle$ is $\frac{1}{2}$, computed as follows.

$$\begin{aligned} & \left(\frac{e^{i\theta}+1}{2\sqrt{2}}\right)\left(\frac{e^{-i\theta}+1}{2\sqrt{2}}\right) + \left(\frac{e^{i\theta}-1}{2\sqrt{2}}\right)\left(\frac{e^{-i\theta}-1}{2\sqrt{2}}\right) \\ = & \frac{1+e^{i\theta}+e^{-i\theta}+1+1+e^{i\theta}+e^{-i\theta}+1}{8} = \frac{4}{8} = \frac{1}{2} \end{aligned}$$

Likewise for the second qubit. It does not matter what θ_1 or θ_2 are; the measurement results are entirely independent of the input angles chosen.

And it gets even worse. When the measurements are completed, we have qubits in the form $|00\rangle$ or $|11\rangle$. Entanglement has disappeared. To restart the production process, the cost of entanglement must be incurred. If the measures yielded Bell states in the first place, the post measurement qubits are already entangled: there is no need entangle once again.

The conclusion is clear: measuring individual contributions in a synergistic environment, at least the synergy under consideration here, is a silly exercise. To some extent that squares with experience. An example which is fun to think about is Abbott and Costello's classic comedy routine entitled "Who's on First." Bud Abbott and Lou Costello were both funny individuals, but combined they were far funnier than the sum of two funny guys. Trying to determine their individual contributions to the act is futile. On average Costello got bigger laughs, but the set-up from Abbott was indispensable. But assigning results, in this case laughs, to individuals is almost certainly worse than futile: it can be corrosive if it causes the individuals to compete for greater individual measures. Attempting to eke out an extra laugh from a set-up can seriously degrade the punch line.

There are many examples of inappropriate use of individual measures in a synergy environment. For example, in the early days of Enron, real assets (like natural gas pipelines, paper mills, etc.) were combined with trading desks. Synergy seemed to be the result, possibly from information transfers across activities. As time went by, however, Enron invoked ruthless individual measures, people become reluctant to share information with colleagues, and troubles ensued.

Closer to home, contemplate a scholarly environment. It is often easier and more efficient to learn in a group. That's synergy. However, the imposition of individual measures like grades or pay raises, combined with the perception that

relative performance is important, might very well inhibit an individuals inclination to contribute to the group. If we want people to act as a team, then team measurements are sensible; individual measures are not sensible, and endanger teamwork. The larger message is synergy is both powerful and delicate, and can be destroyed by ill-advised individual measures.

12.8 summary

Efficient (synergistic) combination of resources is an important and difficult problem. Efficiency in combination often relies on efficient use of information. Nature is particularly clever, and mysterious, in the use of information when resources are combined. In particular, entanglement can create synergies.

We apply the thinking an formalism of combinations in nature, with a view toward illuminating both physics and accounting. Accounting double entry (directed graph representation) is useful for organizing combinations, and loops appear naturally in the presence of synergy. Entanglement, whether in nature or in an economic environment, is powerful though delicate, and can easily be destroyed with excessive individual measurements.

12.9 references

Eichenwald, Kurt, *Conspiracy of Fools*. Broadway Books, 2005.

McLean, Bethany and Peter Elkind, *The Smartest Guys in the Room*. penguin Books, 2003.

Nielsen, Michael A. and Isaac L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

12.10 exercises

Exercise 12.1 *a. Use the production function*

$$H\Theta H|0\rangle$$

The input costs are

	cost
raw material $|0\rangle$	0
labor θ	$.8\theta^2$

Revenue from an output of $|1\rangle$ *is 5. Suppose the labor input* θ *is 90 degrees (*$\frac{\pi}{2}$*). What is the expected income?*

b. Suppose θ *is 60 degrees (*$\frac{\pi}{3}$*). What is the expected income?*

Exercise 12.2 *Use the multiproduct production function*

$$H_2\Theta_2 H_2 H_1\Theta_1 H_1 CNOT\ H_1|00\rangle$$

The input costs are the same as in exercise 12.1. Output $|\beta_{01}\rangle$ *can be sold for 5, and output* $|\beta_{00}\rangle$ *can not be sold. Suppose* θ_1 *and* θ_2 *are both 90 degrees* $(\frac{\pi}{2})$*. Not counting the cost of entanglement, what is the expected income? Suppose the cost of entanglement is .7. Comparing with single production in exercise 12.1, is single or multiple production preferred?*

Exercise 12.3 *a. Redo the previous exercise with*

$$\theta_1 = \frac{\pi}{3} \text{ and } \theta_2 = \frac{\pi}{2}$$

What is the expected income? If the cost of entanglement is still .7, is single or multiple production preferred?

b. Finally, use

$$\theta_1 = \theta_2 = \frac{\pi}{3}$$

What is the expected income?

Exercise 12.4 *a. Refer to exercise 12.1. What is the optimal angle θ to maximize expected profit?*

b. Refer to exercises 12.2 and 12.3. What are the optimal angles, θ_1 and θ_2 to maximize expected profit?

Exercise 12.5 *Verify the math involved in the construction of EPR entangled qubits.*

$$CNOT\ H_1|ij\rangle = |\beta_{ij}\rangle$$

Exercise 12.6 *Show*

$$H_1\Theta_1 H_1|\beta_{11}\rangle = \frac{1-e^{i\theta_1}}{2}|\beta_{10}\rangle + \frac{1+e^{i\theta_1}}{2}|\beta_{11}\rangle$$

Exercise 12.7 *Show*

$$H_2\Theta_2 H_2|\beta_{10}\rangle = \frac{1+e^{i\theta_2}}{2}|\beta_{10}\rangle + \frac{-1+e^{i\theta_2}}{2}|\beta_{11}\rangle$$

Exercise 12.8 *Show*

$$\begin{aligned} & H_2\Theta_2 H_2 H_1\Theta_1 H_1|\beta_{11}\rangle \\ = \quad & \frac{-e^{i\theta_1}+e^{i\theta_2}}{2}|\beta_{10}\rangle + \frac{e^{i\theta_1}+e^{i\theta_2}}{2}|\beta_{11}\rangle \end{aligned}$$

The following exercises utilize some material from chapter 11 as well as chapter 12.

Exercise 12.9 *Measure $\Theta H|0\rangle$ using the standard basis where*

$$\Theta = \begin{bmatrix} e^{i\theta} & 0 \\ 0 & 1 \end{bmatrix}$$

Let $\theta = \pi/2$. Report the possible vector results and the associated probabilities.

Exercise 12.10 *Measure $\Theta H|0\rangle$ using the Hadamard basis with $\theta = \pi/3$.*

Exercise 12.11 *Measure $H\Theta H|+\rangle$ using Hadamard basis.*

Exercise 12.12 *Redo the previous exercise, but measure using $\mathrm{standard}$ instead of Hadamard.*

Exercise 12.13 *Measure $H\Theta H|0\rangle$ using Hadamard.*

Exercise 12.14 *Using Euler's remarkable formula, evaluate each of the following. Express the answer as a complex number with a real and an imaginary coefficient.*

$$\ln i$$
$$\sqrt{i}$$
$$i^i$$

Exercise 12.15 *Does Euler's formula allow taking the logarithm of a negative number? If so, what is the natural logarithm (base e) of -1? What is $\ln(-100)$?*

Exercise 12.16 *Find the real and imaginary coefficients of each of the following when expressed as a complex number.*

$$\pi^i$$

$$i^e$$

$$e^i$$

$$i^\pi$$

Exercise 12.17 *Express $\ln(3+4i)$ as a complex number. Using the result, express $(3+4i)^{1+i}$ as a complex number.*

Exercise 12.18 *Using Euler's remarkable formula, verify the sum of angle trigonometric identities.*

$$\cos(x+y) = \cos x \cos y - \sin x \sin y$$

$$\sin(x+y) = \sin x \cos y + \cos x \sin y$$

Exercise 12.19 *Firm A faces these opportunities and prices in state-act-outcome format.*

price	θ_1	θ_2	θ_3
1	*1*	*1*	*1*
2	*1*	*2*	*3*
1	*3*	*0*	*0*

State probabilities are consistent with maximum entropy probability assignment. An investment of 1 is made at the beginning of every year. The time sequence of cash flows associated with one investment is

t	0	1	2	3
cash flow	-1	0	0	e^{3r}

Where r is the maximum expected long run rate of return. What is r? What is the steady state valuation of firm A using economic income with continuous compounding (B)? What declining balance depreciation rate will converge to the economic income asset valuation?

Exercise 12.20 *Firm B faces the same opportunities as firm A above.*

price	θ_1	θ_2	θ_3
1	*1*	*1*	*1*
2	*1*	*2*	*3*
1	*3*	*0*	*0*

Firm B also invests one dollar every year, although firm B's cash flow sequence is slightly different.

t	0	1	2
cash flow	-1	0	e^{2r}

In addition firm B has access to an information system presented in joint probability format.

$p(x,y)$	θ_1	θ_2	θ_3
x_1	$\frac{1}{3}$	$\frac{1}{6}$	0
x_2	0	$\frac{1}{6}$	0
x_3	0	0	$\frac{1}{3}$

What is r? What is the steady state asset valuation (B) for firm B? What declining balance depreciation rate (D) will converge to the asset valuation?

Exercise 12.21 *Consider firm A and firm B from the two previous exercises. Suppose firm A acquires firm B for market value to A, that is, for discounted cash flow. Note that firm A acquires the information system as well as the assets of B. How much of the acquisition price is for the assets of B? How much is goodwill?*

Exercise 12.22 *Redo the preceding three exercises with the following change: the opportunity set for both firms is*

price	θ_1	θ_2	θ_3
1	*1*	*1*	*1*
$\frac{95}{3}$	*20*	*35*	*40*
$\frac{175}{3}$	*85*	*55*	*35*

Exercise 12.23 *There are two firms, A and B. Both face the same opportunity set.*

price	θ_1	θ_2	θ_3
1	*1*	*1*	*1*
$\frac{95}{3}$	*20*	*35*	*40*
$\frac{175}{3}$	*85*	*55*	*35*

State probabilities are maximum entropy. Each firm invests one dollar at the beginning of every year. Firm A's sequence of cash flow is

t	0	1	2	3
cash flow	-1	0	0	e^{3r}

For firm B the sequence of cash flow is

t	0	1	2
cash flow	-1	0	e^{2r}

In addition firm B has an information system in joint probability format.

$p(x,y)$	θ_1	θ_2	θ_3
x_1	$\frac{1}{4}$	$\frac{1}{8}$	0
x_2	0	$\frac{1}{8}$	0
x_3	0	0	$\frac{1}{2}$

If firm A acquires firm B, what is goodwill? What is the purchase price? (While it can, and should, be checked, you may assume the prices are arbitrage free, and the maximum entropy state probabilities fit with the information system.)

13
brief concluding remarks

Accounting in this manuscript is treated as a scientific, as opposed to vocational, discipline. That is not to say that vocational issues do not arise, as indeed they do. Asset valuation and income determination, for example, are central goals of vocational accounting. But the development herein followed a scientific framework.

13.1 examples of vocational issues arising from scientific frame

The first accounting question we pursued was decomposing a journal entry vector into orthogonal row (T-account) and null components. While the setting is inarguably one of debits, credits, and financial statements, it is difficult to imagine a vocational setting which requires such a decomposition. Nonetheless, the idea of orthogonal decomposition proved useful in later discussions where the vocational implications were more immediate: accounting valuation of derivative portfolios, for example. Orthogonal decomposition arose in other contexts: error correcting correcting codes, for instance, and the use of quantum processes in secret codes and information synergies. An advantage of adopting the scientific frame is that a number of related, and perhaps interesting, topics are naturally opened up.

Another problem was to find a non-negative set of journal entries which generate given financial statements, and, once again, it is difficult to imagine this problem arising in the vocational world of affairs. The accounting setting, however, is an excellent way to demonstrate the logic and underlying relationships. Once the ideas are understood, then the general result can be applied in the form of

the fundamental theorem of finance with profound implications for arbitrage free pricing and the accounting for various securities. Additionally, and probably of more accounting importance, is its use in the fundamental theorem of accounting.

Continuous compounding, while useful for dealing with instruments of various cash flow sequences, does not appear at first glance to be a vocationally motivated problem. It, too, finds its important application to be in the fundamental theorem of accounting. The fundamental theorem, itself, is not primarily a vocational tool. It does, indeed, supply reasonable and insightful asset valuation methods, surely a vocational task. But the theorem does not supply a recipe for the task - often the preferred vocational format. Rather judgment about the information environment is necessary for informed valuation. Indeed, the valuation task is not the primary contribution of the theorem. Rather, it is the idea that (under the three conditions) an information frame is equivalent to an accounting frame. That is, any information problem can be framed as an accounting one, and vice-versa.

Valuation of goodwill as in chapter 12, for example, is also an equivalent problem in information analysis. And the apparently non-accounting problem of state probability assignment can be reframed in accounting terms as in chapter 8.

This brings us back to the discussion of the metaphorical table for the University's best information scientists, and whether accounting is entitled to a seat. The examples enumerated in the manuscript are meant as a modest case that accounting does belong. For one thing accounting is a good context for developing insight and understanding of some important theorems. But more than that, there's the fundamental theorem of accounting. Just as accounting problems like valuation be reframed as information problems, so can information problems be reframed as accounting.

13.2 another example - Euler and polyhedrons

This last section is an example of what might have been. If accountants had been paying attention to the double entry system, they could have derived a famous relationship about polyhedrons. It was actually discovered by Leonhard Euler (with whom we are already acquainted) in the 18th century, well after the widespread use of double entry accounting. Euler's discovery is credited as an important result in the history and development of mathematics, particularly the opening of the field of topology. See Richeson, 2008.

A polyhedron is a three dimensional shape, the faces of which are polygons. A familiar and easy to visualize example is a cube. The cube has 6 faces, all of which are square. The fact that all the faces are the same size and shape makes the cube a special kind of polyhedron known as "regular" or "Platonic." There are lots of other regular and non-regular polyhedron shapes, of course, and Euler's formula applies to all of them as long as they are convex. Convexity means if a straight line connects any 2 points on the surface of a polyhedron, the line must reside entirely on or in the structure. That is, the polyhedron can't look as though a bite has been taken out of it.

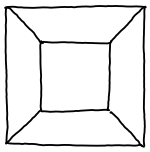


FIGURE 13.1
PROJECTION OF CUBE

The edges of the faces are called, appropriately enough, edges. A point where edges meet is called a vertex. A cube, for example, has 6 faces, 12 edges, and 8 vertices.

Theorem 13.1 *Euler's formula:*

$$V - E + F = 2$$

for any convex polyhedron, where E is the number of edges, V the number of vertices, and F the number of faces.

The formula obviously holds for a cube, but one reason the formula is so important is that it holds for all possible convex polyhedrons. In a general way the formula connects shape with algebra, and opens up new fields of mathematics.

Here's a way to demonstrate why the formula works using our knowledge of double entry accounting. Imagine the poyhedron is hollow and remove one of the faces. Then put a camera where the removed face was, and take a picture of all the other faces and edges from the inside. That effectively projects the 3-dimensional polyhedron onto a 2-dimensional surface, showing all the faces, edges, and vertices with one exception: the exception is the face removed to allow the camera to take the picture.

The projection of a cube, for example, looks as in Figure 13.1. The small square in the middle is the face opposite; the large square consists of the edges of the removed face. All the other faces are in view as long as the polyhedron is convex.The projected form is a graph with arcs, nodes, and loops, and can be described using the concepts of double entry from chapters 2 and 3.

journal entries = # T-acc'ts -1 + # loops

In terms of the polyhedron, the number of journal entries are the number of edges, E. The T-accounts are the vertices, V, and the loops are the faces, F, less the one removed. Substituting into the double entry relationship we have

$$E = V - 1 + (F - 1)$$

Which is a rearrangement of Euler's formula.

13.3 reference

Richeson, David S., *Euler's Gem: The Polyhedron Formula and the Birth of Topology*. Princeton University Press, 2008.

13.4 exercise

The dimple pattern on some modern golf balls appears at first glance to be all hexagons. Upon closer inspection, there are a few pentagons interspersed, actually 12 pentagons. What are the pentagons for? Does there have to be exactly 12? (This exercise is more fun with an example golf ball in hand, like the Callaway warbird 2.0.)