Lecture Notes on Mathematical Biology

Evolutionary Origins, Neural Communication, Population Dynamics, and Multiverse Chaos

Bo Deng

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Contents

Chapter One

Communication Theory Primer

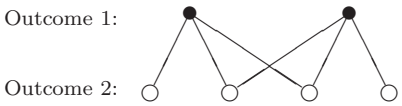
Genetics revolution and communication revolution took off at the same time last century. When the structure of DNA was discovered by James Watson and Francis Crick it was recognized immediately that information theory would play an important role in biology. It is only a recent realization that Claude Shannon's theory of communication may play a role in our understanding of the origins of DNA code and organism reproduction. It is mathematics that makes this connection possible. The purpose of this chapter is to introduce some necessary tools for that undertaking.

1.1 PROBABILITY MATTERS

The following simple arithmetic rule is frequently used in this section.

Rule of Product Processes: If a process can be reduced to a sequence of two processes for which process one has m possible outcomes and process two has n possible outcomes for every outcome of process one, then the process has $m \times n$ possible outcomes.

The diagram below gives an illustration of the rule.



Circles represent outcomes of individual processes. The number of edges represents the total outcomes of the product process.

Example 1.1.1 There are three villages, A, B, C, for which A and B are connected by 3 roads, B and C are connected by 2 roads. By the rule of product processes, there are $3 \times 2 = 6$ ways to go from A to C via B. ◎

The following three examples will be used often later.

Example 1.1.2 Let S be the set of all sequences, $s = s_1 s_2 \cdots s_k$, of a fixed length k, in n repeatable symbols: $1, 2, \ldots, n$, i.e., $s_i \in \{1, 2, \ldots, n\}$. Then the total number, $|S|$, of all sequences is n^k. The process to construct a sequence consists of picking the 1st element of the sequence, the 2nd, and so on, each has n possible outcomes, more accurately in this case, choices. By

the rule of product processes, there are n^k many ways to construct such a sequence, each producing a unique sequence. Such a sequence is also referred to as a *permutation* of length k in n (repeatable) elements. ◎

In the proceeding examples, process two is completely independent from process one, in the sense that not only the number of process two's outcomes remains the same but also the outcomes as well. Next example has the same number but not the same outcomes of process two.

Example 1.1.3 If the symbols in the example above are not allowed to repeat in the sequences, then $|S| = n(n-1)\cdots(n-k+1) = n!/(n-k)!$ because there are exactly one less choice to pick the next symbol down the sequences. Such a sequence is a permutation of length k in the n symbols. ◎

Example 1.1.4 Let S be the set of all subsets each having k symbols of the set $\{1, 2, \ldots, n\}$ with $n \geq k$. (Elements of a set are distinct.) Each permutation of length k gives rise to one set, disregarding the ordering, and each set can generate exactly $k!$ many permutations. That is, a permutation can be generated by first selecting the set that has the correct elements and then by arranging the elements in the desired order. Hence, by the rule of product processes, $|S| \times k! = n(n-1)\cdots(n-k+1)$, and thus $|S| = n(n-1)\cdots(n-k+1)/k!$. It is the number of ways to group k elements from a set of n elements. It is denoted by

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}. \tag{1.1}$$

It is also referred to as the number of *combinations* of k elements from n different elements. We note that they appear as the coefficients of the binomial expansion. ◎

Definition 1.1 *An outcome of a process is called a* ***sample****. The set of all samples is called the* ***sample space*** *of the process. Any subset of the sample space is called an* ***event****.*

Example 1.1.5 Consider the three village example, Example 1.1.1, from above. The sample space of the process of going from A to C via B has 6 ways. If 2 roads between A and B are paved and only 1 paved road between B and C, then travelling by paved roads from A to C is an event, by nonpaved-paved is another, which is also a sample because there is one such a path, and so on. The number of all-paved events is $2 \times 1 = 2$. ◎

Example 1.1.6 A communication transmitter emits symbols from the alphabet $\{1, 2, \ldots, n\}$, and each symbol takes 1 second to emit. The sample

space of 1-second emission is just the set of the alphabet, $S = \{1, 2, \ldots, n\}$ and $|S| = n$. In contrast, the sample space of 3-second emission is $S = \{s_1 s_2 s_3 : s_i = 1, 2, \ldots, \text{ or, } n\}$, and $|S| = n^3$. ◎

Definition 1.2 *Let $S = \{S_1, S_2, \ldots, S_n\}$ be the sample space of a process and $\{p_1, p_2, \ldots, p_n\}$ be a set of numbers satisfying conditions: (1) $0 \le p_k \le 1$ for all k, and (2) $\sum_{k=1}^{n} p_k = 1$. Then with the assignment $P(S_k) = p_k$, p_k is called the **probability** of sample S_k and $\{p_1, p_2, \ldots, p_n\}$ is called the **probability distribution** of S.*

One way to assign probabilities to the outcomes of a process is through repeated experimentations. For example, we can approximate the probability of tossing a coin with its head up as $P(\text{head}) \sim$ [# of times head is up]/[# of trials]. If $P(\text{head})$ is getting closer to 0.5 as the number of trials increases, then we can assign $P(\text{head}) = P(\text{tail}) = 0.5$ and make the assumption that the coin is fair. If the approximation discernibly differs from the equiprobability, then we can reasonably assume that the coin is loaded.

Another way is by sampling. For example, to assign probabilities to the English alphabet in mathematical writings, we can first randomly check out many books from a mathematics department's library. We then count each letter's appearances in all books. The ratio of the number and the total number of printed letters is an approximation of the probability of the letter. This approximation is called a **1st order approximation** of the probability of the sample space. We denote it by P^1. A **2nd order approximation** of the sample space is the same except that a string of two letters is a sample and there are $26^2 = 676$ many samples. An n**th order approximation** is generalized similarly. In contrast, the 0th order approximation is exact as defined below.

Definition 1.3 *The **0th order approximation** of the probability of a finite sample space $S = \{S_1, S_2, \ldots, S_n\}$ is the **equiprobability distribution**, denoted by $P^0 = \{1/n : k = 1, 2, \ldots, n\}$.*

Yet, a common practice in mathematical modelling is to assign the probability distribution based on assumptions made to the process. Such a assumption is often characterized by words such as "completely random", "uniformly distributed", "independent event", etc.

Definition 1.4 *Let E be an event of a sample space S with a probability distribution P. Then the **probability** of E, denoted by $P(E)$, is the sum of the probabilities of each outcome in the event. For $E = \emptyset$, it sets $P(E) = 0$.*

Example 1.1.7 Gregor Mendel was the first to demonstrate that heredity is the result of gene transfer between generations. By cross-pollination of pea plants of two colors, green (g) and yellow (Y), he showed that each pea has two copies of color gene and that the offspring receives one copy from

each parent. By the Rule of Product Processes, there are $2 \times 2 = 4$ color-gene combinations (not necessarily distinct) for the offspring. For example, if Yg and Yg are the parental gene pairs, then the sample space for the first generation offspring gene pair is $\{YY, Yg, gg\}$. Since Y and g are equal probable to appear in the offspring's gene make up, the sample frequency for the first offspring generation can be captured by the table below

Gregor Mendel (1822-1884), a monk-teacher-researcher, is considered the founding father of modern genetics. Based on many years of careful breeding experiments on cross-pollination of pea plant and sophisticated quantitative analysis by statistics, Mendel developed several fundamental laws of inheritance. He did not coin the word gene but called it factor instead.

$\times$	Y	g
Y	YY	Yg
g	gY	gg

with the top row representing the possible contribution from one parent and the left column representing that from the other parent. That is, the probability distribution is

$$p(YY) = 1/4,\ p(Yg) = 1/2,\ p(gg) = 1/4.$$

In this example, the yellow gene is dominant, and the green gene is recessive, meaning that a pea is yellow if it has at least one copy of Y-gene and green if both copies are green. Hence, given the Yg combination for parent peas, the probability of having a green pea is $P(\text{green}) = p(gg) = 1/4$ and a yellow pea is $P(\text{yellow}) = p(YY) + p(Yg) = 3/4$.

◎

Example 1.1.8 Consider the process of rolling a fair dice with the equiprobability $P^0 : p_k = 1/6$. Then the probability of the event that the outcome is a number no greater 4 is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} \sim 0.6667$ with the first $\frac{1}{6}$ for the outcome being 1, the second being 2, etc.

◎

This example has the following generalization to equiprobability distribution.

Theorem 1.5 *Let S be a finite sample space with equiprobability distribution. Then for any event E,*

$$P(E) = \frac{|E|}{|S|}.$$

Example 1.1.9 Consider the probability of at least two students in a class of 25 who have the same birthday. First we line them up or assign a seat to each or just give each a number from 1 to 25. There are 365 choices for each student's birthday. So the sample space S is the set of all sequences of length 25 in integers from 1 to 365. And $|S| = 365^{25}$. It is easer to consider the complementary event E^c that no two students share the same birthday. So E^c must be the set of all permutations of length 25, and $|E^c| = 365 \cdot 364 \cdots 341$. Consequently, $|E| = 365^{25} - 365 \cdot 364 \cdots 341$. If no two students are related, no twins for example, then we can reasonably

assume that the samples are equiprobable and conclude from Theorem 1.5 that

$$P(E) = \frac{|E|}{|S|} = 1 - \prod_{k=0}^{24}(1 - \frac{k}{365}) \sim 0.5860.$$

It is counter-intuitive that the odd is this high. With a group of 50 people, the odds goes up to a whopping 97%.

◎

Exercises 1.1

1. Verify the combinatoric formula
$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n-1 \\ k \end{array} \right) + \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right)$$
in two ways. One is to use the formula (1.1). The other is to use the narrative meaning of the formula.
2. Start with YY and gg color gene pairs for two pea plants. List in a table all possible pairings for the first offspring generation. List in another table all possible pairings of the second offspring generation. Find the probabilities of yellow and green peas of the second offspring generation. You should find the gg trait skips the first generation and reappears in the second.
3. There are numerous connections between New Jersey and Manhattan across the Hudson River, Brooklyn-Queens and Manhattan across the East River, The Bronx and Manhattan across the Harlem River. Use a map or search the Web to find the total number of connections between New Jersey and Brooklyn via Manhattan. What is the probability to go from New Jersey to Brooklyn via Manhattan by tunnel-bridge combination?
4. Prove Theorem 1.5.

1.2 COMMUNICATION CHANNEL

Every communication system is a **stochastic process** which generates sequences of samples from a sample space with a probability distribution. However, information systems rather than general stochastic processes are our primary consideration.

Definition 1.6 *An **information source** is a sample space of a process, $I = \{i_1, i_2, \ldots, i_m\}$ with probability distribution Q. A sequence of samples is a **message**. The sample space is called the **source alphabet** and a sample is called a **source symbol** or **signal symbol**.*

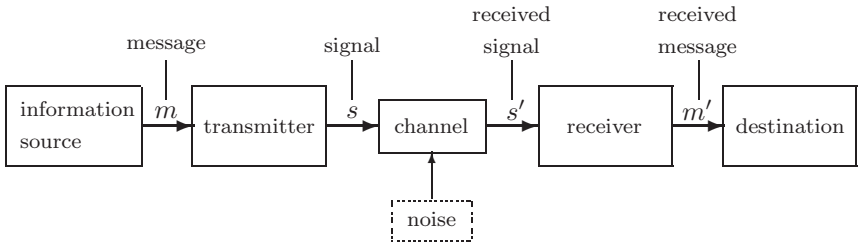


Figure 1.1 Schematic diagram of a communication system.

A **communication system** consists of essentially five parts:

1. An **information source** from which a message or a sequence of messages are produced to send to a receiver.
2. A **transmitter** which operates on the message in some way to produce a signal suitable for transmission over a channel. It has a **signal alphabet**, identified as $S = \{1, 2, \ldots, n\}$. For each information source, there is a bijective function f from I to a subset C of the set S^* of all words (finite sequences of S). The elements of C are called **codewords** and (C, f) is called the **encoding scheme** of the source I.
3. A **channel** which is merely a medium used to transmit the signal from transmitter to receiver.
4. A **receiver** which ordinarily performs the inverse operation of the encoding function f, reconstructing the message from the signal.
5. A **destination** which may be a person, a computer terminal for which the message is intended.

Figure.1.1 gives a conceptual representation of a communication system. Because it can be conceptualized, it can be generalized and applied to abstract formulations. For examples, a channel may be a physical embodiment such as a phone line, a coaxial cable, an optic fiber, a satellite, a nerve axon, a bridge, or just the traffic such as electromagnetic waves, beams of light, cars, etc. It may be just thought as some characteristics of its physical embodiment, such as the amplitudes of electrical pulses, a band of frequencies, the traffic volumes, etc. More abstractly, it can just be thought as a statistical process which subject its input to a time-consuming transformation and a random-state alteration. Immediately relevant to biology, DNA replication, cell division, sexual reproduction can indeed be thought as communication channels.

Definition 1.7 *The 1st order approximation of the source I, Q in terms of the signal alphabet S via an encoding scheme (C, f) is called the* ***signal distribution*** *of the source. Denote it by $P = f(Q)$.*

Example 1.2.1 Morse code was invented by Samuel Morse and Alfred Vail in 1835. It uses two states (on and off) composed into five symbols: dit (for dot, $\cdot$), dah (for dash, $-$), space between dits and dahs ($\wedge$), space between letters ($/$), space between words ($\sqcup$). Let 1 denote the "on" state and 0 the "off" state. Then the standard encoding scheme is as follows.

Morse Code	Codewords
$\cdot$	1
$-$	111
$\wedge$	0
$/$	000
$\sqcup$	0000000

For examples, the letter "S" is coded by three dots and the letter "O" by three dashes. Thus the Morse code for the International Distress Signal "S.O.S" is

$$\cdot \wedge \cdot \wedge \cdot / - \wedge - \wedge - / \cdot \wedge \cdot \wedge \cdot$$

and its binary representation is

$$10101\ 000\ 11101110111\ 000\ 10101.$$

Morse code can be thought as a 5-alphabet channel or a binary channel.

The world's first telegram was sent on May 24, 1844 by inventor Samuel Morse. The message, "What hath God wrought," was transmitted from Washington to Baltimore. In a crude way, the telegraph was a precursor to the Internet in that it allowed rapid communication, for the first time, across great distances. Western Union stopped its telegram service on January 27, 2006.

◎

Example 1.2.2 Consider a quaternary source $I = \{a, b, c, d\}$ with distribution $Q = \{\frac{1}{6}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6}\}$. Consider a binary channel with $B = \{0, 1\}$ (specifically reserved for binary code and $S = \{1, 2, \ldots, n\}$ for all others.)

- Let $C = \{00, 01, 10, 11\}$ be the codewords and $f(a) = 00, f(b) = 01, f(c) = 10, f(d) = 11$ be the encoding function. Then because of the apparent symmetry, the 1st order signal distribution P must be the equiprobability distribution P^0.
- Let $C = \{1, 10, 100, 1000\}$, an example of a **comma code** and $f(a) = 1, f(b) = 10, f(c) = 100, f(d) = 1000$. To find the signal distribution $P = \{p_0, p_1\}$, we proceed as follows.

  Let N be the length of a typical source message which contains $\frac{1}{6}N$ many a, $\frac{1}{3}N$ many b, etc. Then the length of the signal is

$$L = \left[\frac{1}{6}(1) + \frac{1}{3}(2) + \frac{1}{3}(3) + \frac{1}{6}(4)\right] N$$

with the numbers in parentheses being the full codeword lengthes. Of the signal the number of 0 is

$$L_0 = \left[\frac{1}{6}(0) + \frac{1}{3}(1) + \frac{1}{3}(2) + \frac{1}{6}(3)\right] N$$

with the numbers in parentheses being the codeword lengthes in 0s only. Hence $p_0 = L_0/L = 3/5$ and $p_1 = 1 - p_0 = 2/5$ respectively.

In fact, for any signal distribution $\bar{P}$ there is an encoding scheme (C, f) so that the signal distribution $P = f(Q)$ is as close to $\bar{P}$ as possible. Following this subject through we will enter the areas of data compression, encryption, and transmission efficiency. We will pick up the subject of transmission efficiency later but nothing else.

Since the subject of source-to-signal encoding is not the main concern of this chapter, we will regard a source simply in terms of its signal (or channel) alphabet S and its signal distribution P from now on. A word of caution: the coded source is only part of the signal source with same distribution $P = f(Q)$.

Exercises 1.2

1. For a source of distribution $Q = \{1/9, 2/9, 1/3, 2/9, 1/9\}$ in Morse code, find the signal distribution P in the binary code, using the encoding scheme from the main text.
2. Let $I = \{1, 2, 3, 4\}, Q = \{2/5, 3/10, 1/5, 1/10\}$ and (C, f) be an encoding scheme with $C = \{1, 10, 100, 1000\}$ and $f(k) = c_k$. Find the signal distribution P.

1.3 ENTROPY

For an information source, it is intuitive that the less frequent a signal symbol occurs, the more "information" the symbol carries because of its rarity. This section is about how to measure information precisely. As it turns out we need to start with an equivalent way to look at probabilities.

Definition 1.8 *For a sample space $S = \{S_1, S_2, \ldots, S_n\}$ with probabilities $P = \{p_1, p_2, \ldots, p_n\}$, the reciprocal $q_k = 1/p_k$ is called the* ***possibilities*** *of the sample event S_k.*

The possibilities of sample S_k literally mean the number of possible alternatives each time when S_k appears. More precisely, S_k is one of the $1/p_k$ many possibilities and there are precisely $1/p_k - 1$ many non-S_k possibilities that may appear.

Example 1.3.1 Consider a "black box" containing 2 red marbles and 5 blue marbles. The probability to randomly pick a marble from the box that is red is $p_r = 2/7$. Respectively $p_b = 5/7$. It means equivalently that every time a red is picked, it is just one out of $1/p_r = 3.5$ many possibilities, and there are $3.5 - 1 = 2.5$ many blues which might have been picked. ◎

For reasons of communication to be elaborated below, we will not use p_k nor q_k directly to measure information. Instead, we turn to another equivalent measurement of the possibilities as follows.

Definition 1.9 *Let $S = \{1, 2, \ldots, n\}$ be a sample space with distribution $P = \{p_1, p_2, \ldots, p_n\}$. The **information** of sample k is*

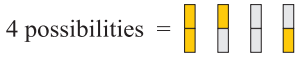
$$I(k) = \lg \frac{1}{p_k}$$

*in **bit** for unit, where* lg *is the logarithmic function of base 2:* $\lg x = \log_2 x$.

Remark.

1. The word "bit" comes from the phrase "binary digit". In fact, let $B = \{0, 1\}$ be the binary alphabet with equiprobability distribution $p = p_0 = p_1 = 1/2$. Then each binary symbol contains $\lg 1/p = 1$ bit of information, as anticipated.
2. The key reason to use bit rather than the number of possibilities to measure the symbol information is its property of **additivity**. Let $s = s_1 s_2 \cdots s_n$ be a binary sequence from a binary source with equiprobability P^0. Then the probability of the sequence according to Theorem 1.5 is $\frac{1}{2^n}$ because it is one sample out of the sample space $B^n := \{s_1 s_2 \cdots s_n : s_k = 0 \text{ or } 1\}$. By definition, its information is $I(s) = \lg 1/(1/2^n) = n$ bits. This is consistent with our expectation that a binary sequence of length n should contain n bits of information if each contains 1 bit.
3. For a signal alphabet $S = \{1, 2, \ldots, n\}$ with distribution $\{p_k\}$, in general, since $\frac{1}{p_k} = 2^{\lg 1/p_k} = 2^{I(k)}$, $I(k)$ can be interpreted as the "length" of a binary sequence. Here is another way to interpret the measurement. To make it simple let us assume $\lg 1/p_k$ is an integer. Now, if we exchange (as in "currency") each possibility of symbol k for one binary sequence of finite length, then we will have $1/p_k$ many sequences. Assume each sequence is a composite metal stick with the 0s and 1s representing alternating gold bars and silver bars respectively and all in one uniform width. (Visualizing any situation in terms of money helps most of us to find practical solutions quicker.) We are picky. We want our sticks to satisfy these conditions: (1) the gold bars and silver bars are equal in number after adding them up from all the exchanged sticks, (2) all sticks are in equal length, (3) no sticks of the same length are left behind that cannot be exchanged for one possibility of k. The only way to meet our criteria is to cut the stick in length $I(k)$.

4. In other words, just like the dimension of mass is measured in gram for example, the amount of information for each symbol is measured in bit unit, which in turns can be thought as the minimum length of binary sequences with equal symbol probabilities. For example, consider the quaternary source $S = \{a, b, c, d\}$ with equiprobability distribution $p = 1/4$. Each symbol has $4 = 1/(1/p)$ possibilities, the whole alphabet. Exchanging the 4 possibilities for binary sequences of minimum length in equiprobability distribution, $\{00,\ 01,\ 10,\ 11\}$ are the only choices. The minimum length $2 = \lg 4$ is the symbol's information.



5. The mathematical sense of information has little to do with the word's literal sense. Junk mails contain lots of information but little meaning. One person's information is another person's junk.

Example 1.3.2 A good $4'' \times 6''$ photo print has 800 horizontal pixels and 1200 vertical pixels (i.e. 200 pixels per inch). Assume that each pixel is capable to display $256 = 2^8$ colors. Then there are a total of $2^{8\times 800\times 1200} = 2^{7,680,000}$ many possible pictures, including those abstract arts someone will find it interesting. Hence each picture contains $I = 7,680,000$ bit information. Let us compare it to a story of 1000 words. Webster's Third New International Dictionary has about 263,000 main entries. Assume all the entries are equally probable. Then a story of 1,000 words contains $\lg 263,000^{1,000} = 18,005$ bit information. In fact, a black-white picture of the same resolution contains more information. 960,000 bits to be precise. ◎

A picture is worth a thousand words. So there will be a lot of them coming your way. Take another look at Fig.1.1. It has more features than we discussed it in text. Most pictures of this book are this way.

Example 1.3.3 Consider Example 1.3.1 of red, blue marbles. Each red marble has $I(\text{red}) = \lg 1/p_r = \lg 7/2 = 1.8074$ bits of information and each blue marble has $I(\text{blue}) = \lg 1/p_b = \lg 7/5 = 0.4854$ bits of information. On average, each marble contains

$$p_r I(\text{red}) + p_b I(\text{blue}) = \frac{2}{7}\lg\frac{7}{2} + \frac{5}{7}\lg\frac{7}{5} = 0.8631 \text{ bits}.$$

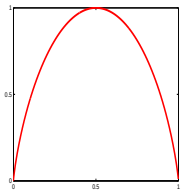
◎

We now introduce the main concept of this section, the averaged information of the source.

Definition 1.10 *Let $S = \{1, 2, \ldots, n\}$ be a sample space with distribution $P = \{p_1, p_2, \ldots, p_n\}$. Then the* ***entropy*** *of the sample space is defined as the*

average information per sample, denoted by $H(S)$ or interchangeably $H(P)$. That is

$$H(S) = \sum_{k=1}^{n} p_k I(S_k) = \sum_{k=1}^{n} p_k \lg(1/p_k) = -\sum_{k=1}^{n} p_k \lg p_k.$$



Graph of $H(p)$

Example 1.3.4 Consider a binary source of distribution $p, 1-p$ for $0 \leq p \leq 1$. The entropy function is $H(p) = p\lg(1/p) + (1-p)\lg(1/(1-p))$. Notices that the graph of H reaches the maximum at the equiprobability $p = 1/2$ where as $H(0) = H(1) = 0$, for which the signal sequences are either entirely 0s and respectively entirely 1s, containing no information for lack of variation.

In fact, the entropy always reaches its maximum, $H_n = \lg n$, at equiprobability distribution. Note that $\lg n = \sum_{k=1}^{n} \frac{1}{n} \lg(1/\frac{1}{n})$. We have the following.

Theorem 1.11 *For any probability distribution distribution $P = \{p_1, p_2, \ldots, p_n\}$, we have*

$$H(P) = \sum_{k=1}^{n} p_k \lg 1/p_k \leq H_n.$$

The equality holds if and only if $p_k = 1/n$ for all k.

Proof. We give here a somewhat elementary proof. A generalized result and a different proof by optimization will appear later. The proof is based on this simple fact from calculus: $\lg x \leq x - 1$ with equality at $x = 1$. Because $\lg n$ can be rewritten as $\sum p_k \lg n$, we have the following

$$\begin{aligned} H(P) - H_n &= \sum p_k(\lg \frac{1}{p_k} - \lg n) = \sum p_k \lg \frac{1}{np_k} \\ &\leq \sum p_k(\frac{1}{np_k} - 1) = \sum_{k=1}^{n}(\frac{1}{n} - p_k) = 0. \end{aligned}$$

Hence, $H(P) \leq H_n$ follows and the equality holds if and only if $1/(np_k) = 1$, the equiprobability condition. □

The equiprobability distribution P^0 is also referred to as the **mean distribution** because

$$\frac{1}{n} = \frac{p_1 + p_2 + \cdots + p_n}{n}$$

for any distribution P of n symbol space.

Exercises 1.3

1. According to the standard of NTSC (National Television Standards Committee), the resolution of your TV is capable to display images at a resolution of 720 pixels wide and 525 scanlines. Assume that each pixel is capable to display $256 = 2^8$ colors. Find the total information a still picture can have.
2. Let us assume on average your parents speak to you for 100 words a day. Let us also assume that your parents use a vocabulary of 5,000 words at home which must be an overestimate because they seem only use a few words like "eat your breakfast", "turn off the TV". Anyway, there are a total of $5,000^{100\times365\times18}$ possible talks when you reach 18, including all that are random streams of words. (What is the difference? No one is listening anyway. They might just as well mumble randomly only to increase the bits and to have your attention that way.) Nevertheless, find the information of such a life long lecture can have. This may give you an idea why kids like to watch TV and think their parents hopelessly boring.
3. Let $Q = \{1/4, 1/4, 1/4, 1/4\}$ be the distribution of an information source, $C = \{1, 10, 100, 1000\}$ be an encoding scheme to the binary system. Find the signal distribution P and entropies $H(Q), H(P)$.
4. Let $H(p) = -p \lg p - (1-p)\lg(1-p)$ be the entropy function for the a binary system. (a) Use l'Hôpital's Rule to show $H(0) = H(1) = 0$. (b) Use the First Derivative Test of optimization to show that it has a unique critical point and it is the absolute maximum in $[0, 1]$.

1.4 OPTIMAL MEAN RATE AND ALL-PURPOSE CHANNEL

An information source can be just one message, but usually is a type of messages, such as all emails, or all picture files, etc., and each type follows a particular industry encoding protocol. Thus in terms of its channel encoded signals, the source assumes a signal distribution. Certainly there is no need to pay special attentions to particular sources all the times. In such a situation, what may be important is the aggregated distribution of all sources.

Definition 1.12 *The* ***aggregated*** *distribution, denoted by P^*, of a channel $S = \{1, 2, \ldots, n\}$ is the 1st order approximation of all sources signal distributions.*

The question of practical importance when you design a channel as an engineer or buy an information service as a consumer is primarily about the transmission rate in bit per second (bps). But the notion of transmission rate can be ambiguous if it is not defined carefully. The first question should be about the transmission rate in general for the aggregated distribution P^*. If you are an Internet gamer, what matters more in particular perhaps is the rate for gaming whose source distribution may not be the aggregated distribution P^*.

Definition 1.13 *Let τ_k in second be the transmission time over a channel $S = \{1, 2, \ldots, n\}$ for symbol k, $k = 1, 2, \ldots, n$. Then the average **transmission time** of a source $P = \{p_1, p_2, \ldots, p_n\}$ is*

$$T(P) = p_1\tau_1 + p_2\tau_2 + \cdots + p_n\tau_n \text{ in second per symbol.}$$

*And the **transmission rate** of the source is*

$$R(P) = \frac{H(P)}{T(P)} = \frac{\sum p_k \lg 1/p_k}{\sum p_k\tau_k} \text{ in bps.}$$

Here $H(P)$ is the entropy of the source in bit per symbol.

A source distribution P is a problem of a particular interest. But the aggregated distribution P^* is the interest of all sources, and that is a subject of speculation and hypothesizing.

Definition 1.14 *A channel $S = \{1, 2, \ldots, n\}$ is called an **all-purpose channel** if the aggregated distribution P^* is the equiprobability distribution P^0, $p_k = 1/n$. Denoted by $T_n = T(P^0)$, $H_n = H(P^0)$, and $R_n = R(P^0) = H_n/T_n$, which is called the **mean transmission rate** or **mean rate** for short.*

Example 1.4.1 Consider the Morse code again from Example 1.2.1. Let $\tau_\bullet$ denote the transmission time of a dit. Then a dah is conventionally 3 times as long as a dit ($\tau_- = 3\tau_\bullet$). Spacing between dits and dahs is the length of one dit ($\tau_\wedge = \tau_\bullet$). Spacing between letters in a word is the length of 3 dits ($\tau_/ = 3\tau_\bullet$). Spacing between words is 7 dits ($\tau_\sqcup = 7\tau_\bullet$). Treating the Morse code as an all-purpose channel, the average transmission time per symbol is

$$T_5 = \frac{\tau_\bullet + \tau_- + \tau_\wedge + \tau_/ + \tau_\sqcup}{5} = \frac{\tau_\bullet}{5}\,[1 + 3 + 1 + 3 + 7] = 3\tau_\bullet$$

and the mean transmission rate is $H_5 = \lg 5/(3\tau_\bullet) = 0.5283/\tau_\bullet$. ◎

A few more important points to learn from this section.

First, entropy is maximized when a channel becomes all-purpose, allowing the most diversity and variability in information through. If you are an engineer assigned to design an Internet channel, your immediate concern is not about a particular source, nor an obscure industry's encoding protocol. Instead, your priority is to give all sources an equal chance. As a result, you have to assume the maximal entropy in equiprobability distribution for the aggregated source and use the mean rate R_n as the key performance parameter. All Internet connections thus can be reasonably assumed to be of all-purpose.

Secondly, the mean rate is an intrinsic property of a channel with which different channels can be compared. A channel may use electromagnetic waves or light pulses as signal carriers. As a consumer you cannot care less. But assuming money is not an issue, you want an optic fiber connection

Connection	Rate in Kbps
Dial-up	2.4 – 56
DSL	128 – $8{\cdot}10^3$
Cable	512 – $20{\cdot}10^3$
Satellite	– $6{\cdot}10^3$
Optic Fiber	$45{\cdot}10^3$– $150{\cdot}10^3$

because it is the fastest on the market. The key point is that you want to base your decision on what is the optimal solution to the transmission rate problem. If a straw could transmit information faster, the Earth would look like a giant wheat field. Realistically though it probably will be wired like an optic fiber ball in a not so distant future.

Thirdly, the symbol rate $T(R)$ is only an indirect measure of a channel. It plays a secondary role in communication theory. What makes a channel is its capability to handel the randomness that information sources inherit. Photons streaming down from the Sun does not make a channel no matter how fast the Sun can pump them out nor how fast they can travel through space.

Example 1.4.2 In one episode of the scifi TV series, *Star Trek: The Next Generation*, an alien species, *Bynars*, hijacked the spaceship *The U.S.S. Enterprise*. Like our computers, the Bynars brain and therefore its civilization are based on the binary system. Here is one reason why the binary system may be sufficient. Let Let $S = \{1, 2, \ldots, n\}$ be the signal alphabet of an all-purpose channel and $\tau_1, \tau_2, \ldots, \tau_n$ be the corresponding symbol transmission times. We know from Sec.1.3 that each signal symbol takes $\lg n$ many binary symbols to represent as a sequence. Thus, if we assume that its transmission time scales accordingly that $\tau_k = t \lg n$. Then the average time is $T_n = t \lg n$ and the mean rate is $R_n = H_n/T_n = 1/t = R_2$, the same as the binary system. Hence, there is no need for non-binary systems. ◎

Had our brain works like the Bynars, the following result would be unnecessary.

Theorem 1.15 *Let $S = \{1, 2, \ldots, n\}$ be the signal alphabet of an all-purpose channel and $\tau_1, \tau_2, \ldots, \tau_n$ be the corresponding symbol transmission times. Assume there is an increment $\Delta\tau$ so that*

$$\tau_k = \tau_1 + \Delta\tau(k-1), \text{ for } k = 1, 2, \ldots, n.$$

Then the mean transmission rate is

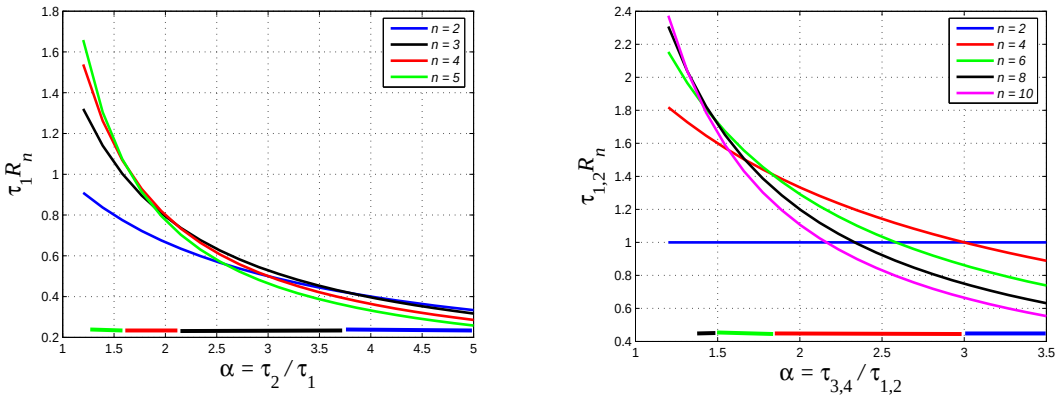
$$R_n = \frac{2 \lg n}{\tau_1[2 + (\alpha - 1)(n-1)]}, \tag{1.2}$$

where $\alpha = \tau_2/\tau_1$.

Proof. Use the identity $\sum_{k=1}^{n} k = n(n+1)/2$ and the relation $\Delta\tau = \tau_2 - \tau_1$. Then

$$\begin{aligned} T_n &= \frac{\sum_{k=1}^{n} \tau_k}{n} = \frac{1}{n} \sum_{k=1}^{n} (\tau_1 + \Delta\tau(k-1)) \\ &= \frac{1}{n} \left[n\tau_1 + (\tau_2 - \tau_1) \frac{(n-1)n}{2} \right] = \tau_1 \frac{2 + (\alpha - 1)(n-1)}{2}. \end{aligned}$$

The formula of H_n follows. □



(a) (b)

Figure 1.2 Numerical illustration (a) for Theorem 1.15, (b) for Theorem 1.16.

Optimal Mean Rate. All-purpose channels can be optimized. As shown in Fig. 1.2(a) (`ch1meanrate1.m`), each signal alphabet size n has its own parameter range in α over which its mean rate is the maximum. (The logarithmic plot for $\tau_1 R_n$ is used to give a better view to the separation of the curves.)

Example 1.4.3 Consider a comma code $1, 10, 100, \ldots, 1\underbrace{0\cdots0}_{n-1}$ as a channel. Assume each binary symbol $0, 1$ takes the same amount of time, τ_0, to transmit. Then $\tau_k = k\tau_0$ for $k = 1, 2, \ldots, n$. Thus, $\Delta\tau = \tau_0$ and $\alpha = \tau_2/\tau_1 = 2$, giving the fastest mean rate when $n = 4$. ◎

A similar result relevant to DNA replication from next chapter is as follows.

Theorem 1.16 *Let n be an even integer. Let $S = \{1, 2, \ldots, n\}$ be the signal alphabet of an all-purpose channel and $\tau_1, \tau_2, \ldots, \tau_n$ be the corresponding symbol transmission times with $\tau_{2k-1} = \tau_{2k}$ denoted by $\tau_{2k-1,2k}$. Assume there is an increment $\Delta\tau$ so that*

$$\tau_{2k-1,2k} = \tau_{1,2} + \Delta\tau(k-1), \text{ for } k = 1, 2, \ldots, n/2.$$

Then the mean transmission rate is

$$R_n = \frac{4 \lg n}{\tau_1[4 + (\alpha - 1)(n-2)]}, \tag{1.3}$$

where $\alpha = \tau_{3,4}/\tau_{1,2}$.

Notice again from the comparison plot Fig.1.2(b) that R_4 is the fastest rate when $\alpha = 2$.

Example 1.4.4 Consider a binary channel $\{0, 1\}$ with the symbol transmission times $\tau_0 = 1, \tau_1 = 2$, in second per symbol respectively.

- If a source's distribution is $P = \{\frac{1}{4}, \frac{3}{4}\}$, then its transmission rate is

$$R(P) = \frac{H(P)}{T(P)} = \frac{(1/4)\lg(1/(1/4)) + (3/4)\lg(3/(3/4))}{(1/4)(1) + (3/4)(2)} = 0.4636 \text{ bps}$$

- By formula (3.1), the mean rate is

$$R_2 == \frac{2\lg 2}{1+2} = 0.6667 \text{ bps}$$

since $n = 2, \tau_1 = 1, \alpha = 2$.

- For another source with $Q = \{0.6180, 0.3820\}$, its transmission rate is

$$R(Q) = \frac{H(Q)}{T(Q)} = \frac{0.618\lg(1/0.618) + 0.382\lg(1/0.382)}{0.618(1) + 0.382(2)} = 0.6942 \text{ bps}$$

◎

It shows that a source can go through the channel at a faster or slower rate than the mean rate. The rate 0.6942 bps is about the fastest this channel is capable of, the subject of next two sections.

Exercises 1.4

1. Let $Q = \{1/4, 1/4, 1/4, 1/4\}$ be the distribution of an information source, $C = \{1, 10, 100, 1000\}$ be an encoding scheme to the binary system. Find the mean rates $R(Q), R(P)$ when Q and $B = \{0, 1\}$ are treated as different channels.
2. Consider the Morse code of Example 1.4.1. Find the mean rate if the timings of the symbols are changed to: $\tau_{-} = 2\tau_{\bullet}$, $\tau_{\wedge} = \tau_{\bullet}$, $\tau_{/} = 2\tau_{\bullet}$, $\tau_{\sqcup} = 3\tau_{\bullet}$).
3. Prove Theorem 1.16.
4. Consider a comma code of 4 code words, $1, 10, 100, 1000$. Assume each binary symbol takes the same amount of time to transmit, t_0. Find the transmission rate $R(P)$ of a source whose probability distribution is $P = \{2/5, 3/10, 1/5, 1/10\}$.

1.5 LAGRANGE MULTIPLIER METHOD

A Prototypic Example — Get Out the Hot Spot

Consider an experiment consisting of an oval metal plate, a heat source under the plate, and a bug on the plate. A thought experiment is sufficient so that no actual animals are harmed. In one setup, Fig.1.3(a), the heat source is place at the center of the plate, and the bug is placed at a point as shown. Assume the heat source is too intense for comfort for the bug. Where will it go? By inspection, you see the solution right away. The coolest spots are the edge points through plate's major axis, and the bug should go to whichever

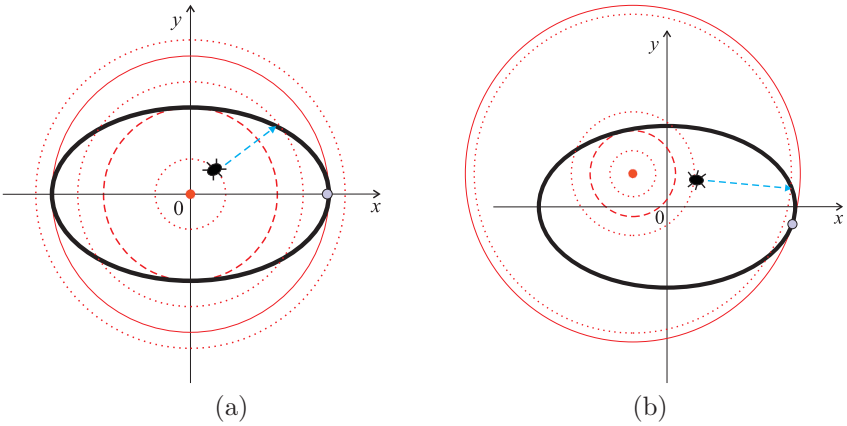


Figure 1.3 Thought experiment set-up.

is closer. However, the bug cannot see the "whole" picture. It has its own way to solve the problem. Translating the bug's solution into mathematics gives rise to the Lagrange Multiplier Method. We break its solution into two steps: the solution to reach plate's edge, and the solution to go to the coolest point on the edge.

Level Curve and Gradient

Place the plate in the coordinate system as shown. Let (x_0, y_0) be the bug's current position as shown. Let $T(x, y)$ denote the temperature in $^\circ F$ at any point (x, y) in the plate. So the bug experiences a temperature of $T(x_0, y_0)$ $^\circ F$. The bug's immediate concern is to move away from the heat source as fast as possible. The direction it takes is denoted by $-\nabla T(x_0, y_0)$ and the opposite direction $\nabla T(x_0, y_0)$ is called the gradient of T at the point.

To derive the gradient, let us give function T a specific form for illustration purposes. For simplicity, we assume that T is distributed concentrically, i.e., the temperature remains constant in any concentric circle around the heat source $(0, 0)$, and the closer to the source, the higher the constant temperature. Such curves are called isothermal curves for this situation and **level curves** in general. Obviously, the bug will not go into its current isothermal curve $T(x, y) = T(x_0, y_0)$, nor stay on it, walking in circle so to speak. It must go to another level curve whose isothermal temperature is immediately lower, $T(x, y) = k$ with $k < T(x_0, y_0)$. If we zoom in on the graphs, we see the answer quickly. At close-up, any smooth curve looks like a line, see Fig.1.4, the **tangent line** by definition. In fact, all tangent lines to level curves nearby are packed like parallel lines. Therefore, the shortest path from the tangent line of the current level curve to the tangent line of another level curve nearby is through the line perpendicular to these parallel tangent lines. The perpendicular direction leading to higher values of T is the **gradient** $\nabla T(x_0, y_0)$, and the opposite direction leading to lower values

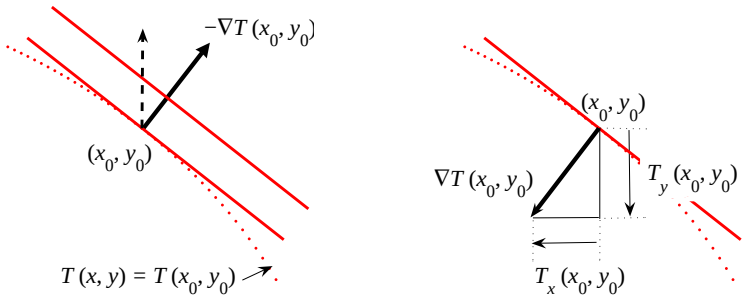


Figure 1.4 Dashed arrow is a non-gradient vector.

of T is $-\nabla T(x_0, y_0)$.

Next, we derive the direction analytically. Let $y = y(x)$ represent the level curve $T(x, y) = T(x_0, y_0)$ near (x_0, y_0), i.e., $T(x, y(x)) \equiv T(x_0, y_0)$ for all x near x_0 and $y(x_0) = y_0$. Then $\frac{dy}{dx}(x_0)$ is the slope of the tangent line $y = y_0 + \frac{dy}{dx}(x_0)(x - x_0)$. The slope is found by the Chain Rule and Implicity Differentiation. Formally we have,

$$\begin{aligned}
\frac{d}{dx}T(x, y) = \frac{d}{dx}T(x_0, y_0) &= 0 \\
T_x(x, y) + T_y(x, y)\frac{dy}{dx} &= 0 \\
\frac{dy}{dx} &= -T_x(x, y)/T_y(x, y) \\
\frac{dy}{dx}(x_0) &= -T_x(x_0, y_0)/T_y(x_0, y_0),
\end{aligned}$$

where $T_x(x, y) = \partial T(x, y)/\partial x$ is the partial derivative of T with respect to variable x, and similarly for $T_y(x, y)$. Since the gradient direction is perpendicular to the tangent of the level curve, its slope, m, is the negative reciprocal of the tangent slope,

$$m = -\frac{1}{dy/dx} = \frac{T_y(x_0, y_0)}{T_x(x_0, y_0)},$$

for which $T_x(x_0, y_0)$ can be thought to be a run of the gradient line and $T_y(x_0, y_0)$ its rise. Finally, we have

---

Definition 1.17 *Let $F(x, y)$ be a differentiable function, the* ***gradient*** *of F at (x, y) is the vector*

$$\nabla F(x, y) = (F_x(x, y), F_y(x, y)).$$

Important Properties of Gradient

1. Function value increases the fastest in the gradient direction.
2. Function value decreases the fastest in the opposite gradient direction.
3. The gradient is perpendicular to its level curve.

Constrained Optimal Solutions

Driven by instinct, the bug runs in the negative gradient direction to escape the heat. This strategy will only lead it to the edge of the plate, see Fig.1.3(a). Its next move is *constrained* by the boundary curve of the plate. In one direction, heat increases to its highest on the y-axis. In the other, heat decreases to the lowest on the x-axis. The critical clue for the bug as to which direction to move next when reaching the boundary is the fact that the isothermal curve is transversal to the boundary curve. At the minimum temperature point, both the isothermal curve and the boundary curve are tangent, and moving in either direction does not lower the heat further.

For the situation of Fig.1.3(a), we see the exact solution right away. We are even able to derive the solution *graphically* in general such as Fig.1.3(b). That is, when the first isothermal curve, radiating outwards from the source, touches the boundary it gives rise to the constrained maximum temperature point, and when the last curve to leave the boundary it gives rise to the constrained minimum temperature point. The question we address below is how to translate this picture into *analytical* terms.

To this end, let us derive the analytical method for the trivial case first. To be specific, let $\frac{x^2}{9} + \frac{y^2}{4} = 1$ be the boundary curve of the plate so that $(3, 0)$ is the constrained coolest point that the bug comes to rest. We further introduce a notation $g(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$. Thus the boundary can be thought as the level curve of the function g: $g(x, y) = 1$, and the plate the region: $g(x, y) \leq 1$. Therefore the condition for both level curves, $T(x, y) = T(3, 0)$ and $g(x, y) = 1$, to be tangent at $(3, 0)$ is to have the same gradient slope:

$$\frac{T_y(3,0)}{T_x(3,0)} = \frac{g_y(3,0)}{g_x(3,0)}.$$

An equivalent condition is

$$T_x(3,0) = \lambda g_x(3,0), \quad T_y(3,0) = \lambda g_y(3,0),$$

where λ is a scaler parameter. In fact, if $g_x(3,0) \neq 0$, $\lambda = T_x(3,0)/g_x(3,0)$ as one can readily check. Hence, the constrained optimal point $(3, 0)$ is a solution to the equations

$$T_x(x,y) = \lambda g_x(x,y), \quad T_y(x,y) = \lambda g_y(x,y), \quad g(x,y) = 1. \tag{1.4}$$

This problem has three unknowns x, y, λ and three equations. For the trivial case, it should have 4 solutions: 2 constrained maximums and 2 constrained

minimums. The Lagrange Multiplier Method is to solve these equations for the extrema of function $T(x, y)$ subject to the constraint $g(x, y) = 1$.

Example 1.5.1 The heat source is located to a new point $(0, 1)$. Let $T(x, y) = \frac{1000}{1+x^2+(y-1)^2}$ be the temperature. (It is easy to check that the isothermal curves are concentric circles around the source $(0, 1)$.) The constrained extrema (x, y) satisfy equations (1.4), which in this particular case are

$$T_x = -\frac{2000x}{[1 + x^2 + (y-1)^2]^2} = \lambda g_x(x, y) = \lambda \frac{2}{9} x$$
$$T_y = -\frac{2000(y-1)}{[1 + x^2 + (y-1)^2]^2} = \lambda g_y(x, y) = \lambda \frac{2}{4} y$$
$$g(x, y) = \frac{x^2}{9} + \frac{y^2}{4} = 1$$

Divide the first equation by the second to eliminate λ,

$$\frac{x}{y-1} = \frac{4}{9}\frac{x}{y}.$$

This equation has 2 types of solutions: (1) $x = 0$, then $y = \pm 2$. (2) $x \neq 0$, then $1/(y-1) = 4/(9y)$. Solving it gives $y = -4/5$ and $x = \pm 3\sqrt{1 - y^2/4} = \pm 3\sqrt{21}/5$. Last, run a comparison contest among the candidate points, we find that $(0, 2)$ is the constrained maximum with $T = 500°F$, $(\pm 3\sqrt{21}/5, -4/5)$ are the constrained minimums with $T = 80.5°F$. Point $(0, -2)$ is a constrained local maximum with $T = 100°F$. ◎

Generalization — Lagrange Multiplier Method

In general, let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote an n-vector variable, $f(\mathbf{x})$ be a real-valued function. Then the n-vector

$$\nabla f(\mathbf{x}) = (f_{x_1}(\mathbf{x}), f_{x_2}(\mathbf{x}), \ldots, f_{x_n}(\mathbf{x}))$$

is the **gradient** of f at $\mathbf{x}$. In the gradient direction f increases most rapidly and opposite the gradient f decreases most rapidly. In addition, the gradient vector $\nabla f(\mathbf{x}_0)$ is perpendicular to the hypersurface (level surface) $f(\mathbf{x}) = f(\mathbf{x}_0)$.

A constrained optimization problem is to find either the maximum value or the minimum value of a function $f(\mathbf{x})$ with $\mathbf{x}$ constrained to a level hypersurface $g(\mathbf{x}) = k$ for some function g and constant k. The following theorem forms the basis of the Lagrange Multiplier Method.

Theorem 1.18 *Suppose that both f and g have continuous first partial derivatives. If either*

- *the maximum value of $f(\boldsymbol{x})$ subject to the constraint $g(\boldsymbol{x}) = 0$ occurs at $\boldsymbol{x}_0$; or*

- *the minimum value of $f(\boldsymbol{x})$ subject to the constraint $g(\boldsymbol{x}) = 0$ occurs at $\boldsymbol{x}_0$,*

then $\nabla f(\boldsymbol{x}_0) = \lambda \nabla g(\boldsymbol{x}_0)$ for some constant λ.

To find the candidate points $\mathbf{x}_0$ for the constrained extrema, the method calls to solve the system of equations

$$\begin{cases} \nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \\ g(\mathbf{x}) = 0. \end{cases} \tag{1.5}$$

There are $n+1$ variables $(\mathbf{x}, \lambda)$ and $n+1$ equations since the vector equation for the gradients has n scalar equations. We expect to have non-unique solutions since they are candidates for both constrained maximum and constrained minimum.

Example 1.5.2 We now give an alternative proof to Theorem 1.11. It is to solve the constrained optimization problem

$$\begin{aligned} \text{Maximize:} \quad & H(p) = -(p_1 \lg p_1 + p_2 \lg p_2 + \cdots + p_n \lg p_n) \\ \text{Subject to:} \quad & g(p) = p_1 + p_2 + \cdots p_n = 1. \end{aligned}$$

By Lagrange Multiplier Method,

$$\begin{aligned} \nabla H(p) &= \lambda \nabla g(p) \\ g(p) &= 1 \end{aligned}$$

In component, $H_{p_k} = -(\lg p_k + 1) = \lambda g_{p_k} = \lambda$. Hence, $p_k = 2^{-\lambda - 1}$ for all k. Since $2^{-\lambda-1}$ is a constant for any λ, the distribution must be the equiprobability distribution $p_k = 1/n$. Since there are distributions, for example $p = (1, 0, \ldots, 0)$, at which $H = 0$, and since the equiprobability distribution is the only constrained solution, it must be the constraint maximum because its value is positive $H = \lg n > 0$. ◎

Exercises 1.5 (Give numerical approximations if exact solutions are not feasible.)

1. Consider the prototypic example from the main text. Suppose the heat source is relocated to a new point $(1, 1)$ and $T(x, y) = \frac{1000}{1+(x-1)^2+(y-1)^2}$ is the temperature function. Assume the same boundary $g(x, y) = \frac{x^2}{9} + \frac{y^2}{4} = 1$ for the metal plate. Find both the hottest and coolest spots on the plate.
2. Find the minimum distance between the point $(3, 3)$ and the ellipse $x^2/4 + y^2/9 = 1$.
3. The Lagrange Multiplier Method can be generalized to multiple constraints. For example, to find optimal values of a function $f(\mathbf{x})$ subject to two constraints $g(\mathbf{x}) = 0, h(\mathbf{x}) = 0$, we solve the following equations

$$\begin{cases} \nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) + \mu \nabla h(\mathbf{x}) \\ g(\mathbf{x}) = 0 \\ h(\mathbf{x}) = 0 \end{cases}$$

for candidate extrema. Here λ and μ are multipliers. Let an ellipse be the intersection of the plane $x+y+z=4$ and the paraboloid $z=x^2+y^2$. Find the point on the ellipse that is closest to the origin.

4. **The Least Square Method:** Let $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ be a collection of data points. Assume x is the independent variable and y is the dependent variable, and we wish to fit the data to a function $y=f(x)$ from a class of functions. Each possible fitting function has a deviation from each data point, $d_k = y_k - f(x_k)$. Function f is a **least square fitting** function if

$$\sum_{k=1}^{n} d_k^2 = \sum_{k=1}^{n} [y_k - f(x_k)]^2 = \text{ a minimum}$$

in the class. Here $\sum d_k^2$ can be thought as the square of a distance between the data points and the candidate function, and hence the name for the method. For example, if we wish to fit the data to a function of the linear class $y=f(x)=b+mx$ with b the y-intercept, m the slope to be determined, we want to minimize the following function of b,m:

$$F(b,m) = \sum_{k=1}^{n} [y_k - f(x_k)]^2 = \sum_{k=1}^{n} [y_k - (b+mx_k)]^2.$$

Use the Lagrange Multiplier Method to show that the least square fit satisfies

$$b = \frac{\sum x_k^2 \sum y_k - \sum x_k \sum x_k y_k}{n\sum x_k^2 - (\sum x_k)^2}, \quad m = \frac{n\sum x_k y_k - \sum x_k \sum y_k}{n\sum x_k^2 - (\sum x_k)^2}.$$

(*Hint:* Use the trivial constraint $g(b,m)\equiv 0$.)

5. Let $(1,1.5),(2,2.8),(3,4.5)$ be three data points. Use the formula above to find the least square fit line $y=b+mx$.

1.6 OPTIMAL SOURCE RATE — CHANNEL CAPACITY

For a channel S, the mean rate R_n is a measure for all sources on average. But a particular source P can go through the channel at a slower or faster rate $R(P)$ than the mean rate. Once an optimal channel is chosen with respect to R_n, individual source can take advantage of the channel to transmit at a rate as fast as possible.

Definition 1.19 *The fastest source transmission rate over all possible sources is called the* ***channel capacity*** *of the channel. Denote it by* $K = \max_P R(P)$.

The channel capacity indeed exists and it is unique.

Theorem 1.20 *The source transmission rate $R(P)$ has a unique maximum K constrained to $\sum p_k = 1$. For the optimal source distribution, p_k^{1/τ_k} is a constant for all k, and $K = -\lg p_1/\tau_1 = -\lg p_k/\tau_k$. In particular, $p_k = p_1^{\tau_k/\tau_1}$, $\sum_{k=1}^n p_1^{\tau_k/\tau_1} = 1$.*

Proof. We use the Lagrange Multiplier Method to maximize $R(P) = H(P)/T(P)$ subject to the constraint $g(P) = \sum_{k=1}^n p_k = 1$. This is to solve the joint equations:

$$\begin{cases} \nabla R(P) = \lambda \nabla g(P) \\ \quad g(P) = 1. \end{cases}$$

The first n component equations are

$$R_{p_k} = \frac{H_{p_k} T - H T_{p_k}}{T^2} = \lambda g_{p_k} = \lambda,\ k = 1, 2, \ldots, n.$$

Write out the partial derivatives for $H = -\sum p_k \lg p_k$ and $T = \sum p_k \tau_k$ and simplify, we have

$$-(\lg p_k + 1)T - H\tau_k = \lambda T^2,\ k = 1, 2, \ldots, n.$$

Subtract the first equation ($k = 1$) from each of the remaining $n-1$ equations to eliminate the multiplier λ and to get a set of $n-1$ new equations:

$$-(\lg p_k - \lg p_1)T - H(\tau_k - \tau_1) = 0$$

which solves to $\lg \frac{p_k}{p_1} = (H/T)(\tau_1 - \tau_k) = R(\tau_1 - \tau_k)$ and hence $p_k = 2^{R(\tau_1 - \tau_k)} p_1$. Denote by

$$\mu = 2^R = 2^{H/T} \text{ equivalently } H = T \lg \mu, \tag{1.6}$$

we have

$$p_k = \mu^{\tau_1 - \tau_k} p_1, \text{ for all } k. \tag{1.7}$$

Next we express the entropy H in terms of μ and p_1, τ_1:

$$\begin{aligned} H &= -\sum_{k=1}^n p_k \lg p_k = -\sum_{k=1}^n p_k[(\tau_1 - \tau_k)\lg \mu + \lg p_1] \\ &= -[\tau_1 \lg \mu - \sum_{k=1}^n p_k \tau_k \lg \mu + \lg p_1] \\ &= -[\tau_1 \lg \mu + \lg p_1] + T \lg \mu, \end{aligned}$$

where we have used $\sum_{k=1}^n p_k = 1$ and $T = \sum_{k=1}^n p_k \tau_k$. Since $H = T \lg \mu$ from (1.6), cancelling H from both sides of the equation above gives

$$\lg p_1 + \tau_1 \lg \mu = 0 \implies \mu = p_1^{-1/\tau_1}. \tag{1.8}$$

From (1.7) it follows

$$p_k = \mu^{\tau_1 - \tau_k} p_1 = p_1^{\tau_k/\tau_1} \text{ for all } k.$$

Last solve the constraint equation

$$f(p_1) := g(p) = \sum_{k=1}^{n} p_1^{\tau_k/\tau_1} = 1$$

for p_1. Since $f(p_1)$ is strictly increasing in p_1 and $f(0) = 0 < 1$ and $f(1) = n > 1$, there is a unique solution $p_1 \in (0,1)$ so that $f(p_1) = 1$ by the Intermediate Value Theorem. From (1.6) and (1.8) we derive the channel capacity

$$K = R = \lg \mu = -\lg p_1/\tau_1 = -\lg p_k/\tau_k.$$

This completes the proof. □

Notice that if there is no time difference, i.e., $\tau_k = \tau_1$ for all k, then $p_k = p_1$, the equiprobability distribution and the capacity is the mean rate $K = H_n$.

Channels Cannot Be Optimized By Channel Capacity. A proof of the following result is left as an exercise.

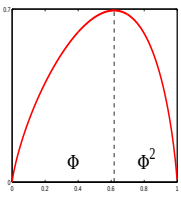
Corollary 1.21 *For a channel $S = \{1, 2, \ldots, n\}$, the channel capacity K increases in the size of the alphabet n.*

This means that optimal channel does not exist if the design criterion is the channel capacity. A channel cannot be designed to chase the absolute maximum rate for one individual source at the expense of all others, see Exercise 4. Every economy produces one mega-rich individual like Bill Gates, but each revolves around all ordinary individuals like you and me.

Example 1.6.1 Consider the Morse code from Example 1.4.1 for which $\tau_-/\tau_\bullet = 3, \tau_\wedge/\tau_\bullet = 1, \tau_/\tau_\bullet = 3, \tau_\sqcup/\tau_\bullet = 7,$. The probability, $p_\bullet$, of the dit for the channel capacity satisfies

$$p_\bullet + p_\bullet^3 + p_\bullet + p_\bullet^3 + p_\bullet^7 = 1.$$

Solving it numerically (`ch1morsecapacity.m`), we have $p_\bullet = 0.4231$ and the channel capacity is $K = 1.2410/\tau_\bullet$. Recall that its mean rate is $R_5 = 0.5283/\tau_\bullet$.



Graph of $R(p)$

Example 1.6.2 Consider again the comma code channel of Example 1.4.3. For $n = 2$, the capacity probability p_1 for symbol 1 satisfies

$$p_1 + p_1^2 = 1.$$

The solution

$$p_1 = \Phi := \frac{\sqrt{5} - 1}{2} = 0.6180$$

is the **Golden Ratio**. In fact, since $1 - p_1 = p_{10} = p_1^2$,

$$\frac{p_1}{1} = \frac{1-p_1}{p_1},$$

a standard way to define the Golden Ratio: *the ratio of the long segment to the whole is the ratio of the short segment to the long one.* See a graphical illustration below.

The channel capacity is $K = -\lg p_1/\tau_1 = 0.6943/\tau_1$. A similarly calculation as Example 1.4.4 shows $R_2 = 0.6667/\tau_1 < K$ as expected. Alternatively, the source rate function $R(p) = -(p\lg p + (1-p)\lg(1-p))/(p+2(1-p))/t_0$ with $t_0 = 1$ is shown in the margin. It reaches the maximum at $p = \Phi$. ◎

Exercises 1.6

1. Consider the Morse code of Example 1.4.1. Find the channel capacity if the timings of the symbols are changed to the following: $\tau_- = 2\tau_\bullet$, $\tau_\wedge = \tau_\bullet$, $\tau_/ = 2\tau_\bullet$, $\tau_\sqcup = 3\tau_\bullet$.
2. (*Hint:* Show p_1 decreases in n.)
3. Consider a comma code of 4 signal symbols, $1, 10, 100, 1000$. Assume each binary bit takes an equal amount of time to transmit, t_0. Numerically approximate the channel capacity of the comma code.
4. Consider a comma code of infinitely many signal symbols, $1, 10, 100, \ldots$. Assume each binary bit takes equal amount of time to transmit, t_0. Show that the channel capacity is $K = 1/t_0$ and the mean rate is $R_\infty = 0$.

Chapter Two

DNA Replication and Sexual Reproduction

All life forms are coded in 4 DNA bases. Why is it not in 2 or 6 bases? Without exception, sexual reproduction takes 2 sexes. Why does not it take 3 or more? Since these are the ways Nature is, science alone is inadequate to tackle these questions. There can be endless possibilities in the numbers of replication bases and sexes that no amount of observation and experimentation can check out them all. Here is where mathematics may play a bigger role. We will approach these questions by mathematical modelling. The immediate results therefore lie between science and mathematics. The ultimate goal of course is to gain new knowledge entirely through scientific methodology, a task of future exploration, hence outside the scope of this book. The theory presented is tentative. But the method used is a well-tested, and useful tool for scientific exploration.

2.1 MATHEMATICAL MODELLING

It is unlikely that we will be able to reduce scientific discovery into an algorithm. However there are some common aspects of this unique process. Figure 2.1 gives a flow chart illustration about these commonalities.

Acquisition of new knowledge starts with observations and experiments which may answer some old questions or lead to new questions. Sometime a question may require further clarifications by keeping out secondary factors of the problem, leading to an approximate question. For example, rather than asking how tides form we should ask how Moon's gravity influences a body of water the size of Earth's ocean. At any point of this process, a conceptual model may appear to give an explanation to an observation or experiment, or an answer to the questions. In such a case, the process of discovery does not go down to the next level to form a mathematical model. That is, the arrows leading to the Conceptual Model box from atop also reciprocate. It can lead to new knowledge in the forms of scientific principles, rules, laws. James Watson and Francis Crick's discovery of the DNA structure follows this path in two ways. First, mathematics plays little role. Secondly, only after their discovery of DNA's copying mechanism did they realize they have also answered the question of how genetic traits are passed between generations in addition to the structural question they always had in mind from the beginning. Good models tend to give answers before the questions are asked. The reciprocal arrows and the arrow from the

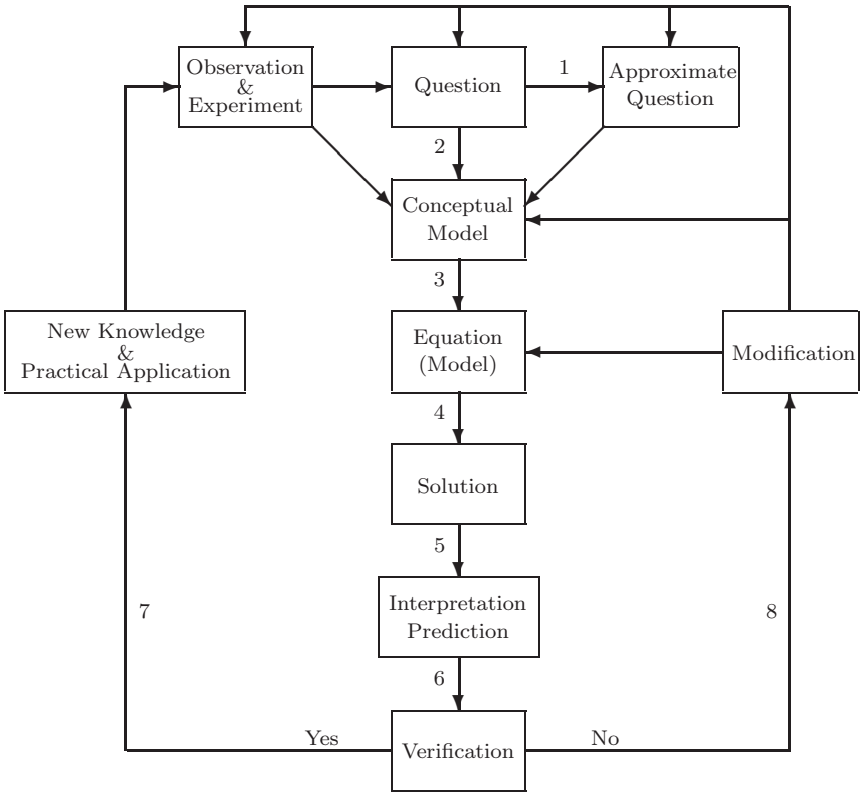


Figure 2.1 Schematic diagram of mathematical modelling.

Conceptual Model box to the New Knowledge box are not shown because the diagram is intended as a clean illustration of knowledge acquisition through mathematics.

Asking the right questions and formulating useful conceptual models are all important. For example, it was an ill-posted question to ask why does the Sun revolve around Earth, or a wrong model to think that Earth is the center of the universe. Whatever mathematics, however beautiful, is useless. The process eventually cycled back to a brand new start. Such episodes do serve useful purposes beside history. It is part of the process and often necessitates the eventual arrival of correct models. Having a simple and correct conceptual model that embody all the essentials of a theory is extremely important to new discoveries. Fundamental breakthroughs often require unconventional thinking which usually cannot be put in the framework of existing models, and therefore creating an original model of its own. Watson-Crick's double helix DNA model is one example. Here is another. At his teens, Einstein

often pondered the question that would he be able to catch a photon if he could travel as fast as the light?. Out from his conceptual model that light travels at the same speed to an observer inside a speeding train and to an observer in a train station Einstein developed his theory of special relativity. The equations used to capture his conceptual picture is the mathematical model of the problem. His model became a physics law after it was verified experimentally.

Unwilling to go to the next level to find a conceptual model often delays a new discovery, and brings out disappointment to its would-be discoverers. Rosalind Franklin had the DNA structure right in her X-ray diffraction (crystallography) image of DNA. She did not embrace Watson and Crick's approach of model-building, considered unconventional then. The rest was history because three persons held two different views on modelling. In the scheme of things it is paradoxically small and huge.

Let us now discuss in more details the passages (arrows) connecting the boxes of the diagram.

1. Prioritize factors and consider only those immediately fundamental to the question to obtain an approximation question.
2. Formulate a conceptual model for the observation/experiment, the question, or the approximate question. Using existing models, such as following bodies, oscillating springs, cooling bodies, etc., may be considered routine, but doing it right is challenging nevertheless. Discovering new ones can be the most original part of the process, and thus the least describable. For this stage, you may find one of Einstein's quotes enlightening, "Imagination is more important than ...."
3. Identify the variables, both independent and dependent, and known parameters of the conceptual model, especially the dependent or otherwise referred to as the unknown variables that are directly tied to the questions. Make necessary hypotheses on those aspects of the model that are either unknown to existing laws or only approximates of such. Incorporate them all to formulate a set of mathematical equations which is considered as a representation of the conceptual model—the mathematical model.
4. Solve the equations, in whichever ways possible: analytical, graphical, numerical, to obtain solutions as provisional answers to the questions.
5. Interpret the solutions, make predictions for experiments. Unintended questions and answers often appear in the process. Model conceptualization sows the seed, solution interpretation reaps the fruit.
6. Check if the solutions are consistent with known properties of the problem ("Does it make sense?"), and check if the predictions as well as the hypotheses are backed up by experiments.
7. If the model is validated, then new knowledge in the forms of theories, scientific principles, operating rules, or physical laws, are obtained. Theoretical foundation is also set for practical applications. In addition, another cycle of knowledge acquisition starts anew.

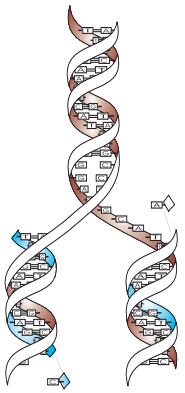
8. However, any mismatch between what is predicted and what is known triggers another iterate of the process. Modification to the model maybe required at any level. The original observation or experiment may not be set up as intended. Or the question may be ill-posted, or the conceptual model is utterly wrong. In such cases, we have to start it over from scratch, avoiding the old mistake in the doing. In another scenario, the question is correct, but the approximate question is incomplete. Therefore the new iterate may require the incorporation of some secondary factors that was kept out originally. In another, it requires the reformulation of some hypotheses made in the stage of model making, or recalibration of some parameter values. The possibilities can be endless, not including the trivial kinds such as debugging "dump algebraic mistakes" made during the solution process which we don't count into the iterative process of modelling.

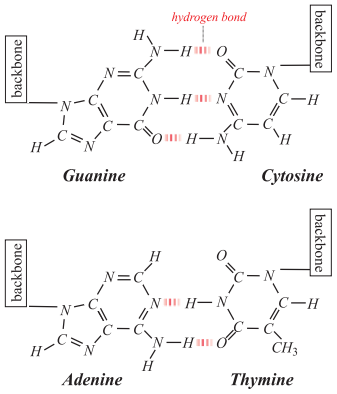
The practice of mathematical modelling is both a science and an art. A friend of mine attributes the following quote to Henrich Hertz. It summarizes the art part of mathematical modelling well. It says, "Mathematical modelling is to construct pictures so that the consequences of the pictures are the pictures of the consequences." It is extremely helpful to be able to visualize in every aspects of the process. It is the most important to be able to translate the visualizations in mathematics. Without mathematics there would be no Einstein's special relativity theory however elegant his conceptual model is. It won't hurt to keep in mind this rule of thumb: "when all are equal, the simplest explanation may well be the answer." One possible reason is that in the physical world things tend to stay in some minimized energy states which often are associated with simplicity and robustness. After discovering the double helical structure of DNA, James Watson and Francis Crick made the prediction that DNA replication is **semi-conservative**, that is, each strand serves as a template for one daughter double helix. They took the simplest possibility of probable replications. When it was experimentally verified, the whole cycle of knowledge acquisition they started was nothing short of spectacular.

Henrich Hertz (1857–1894), a German physicist, is credited to lay the foundation to all modern wireless communications and validate James Clerk Maxwell's electromagnetic theory which in turn led to Einstein's special relativity. The unit measurement for radio wave frequency was changed in 1960 to "hertz" (Hz, 1 cycle per second) in his honor.

Because Nature tends to be a minimalist, we tend to perceive science and mathematics truths beautiful. When James Watson and Francis Crick finally put together their DNA model, they thought it had to be right because it was so elegant. Physicists marvel at Einstein's relativity theories the same way. Aristotle's Earth-centric model was beautiful before its eventual withering because it was a bad model and its beauty was only artificial and transient. Real beauty can only be found in true science and useful mathematics, which often are governed by optimization principles.

DNA Replication. Deoxyribonucleic acid (DNA) molecule has the structure of a helical ladder with **nucleotide bases** forming the ladder rungs. The nucleotides are paired: adenine (A) with thymine (T), guanine (G) with cytosine (C), called **complimentary bases**. During replication, the double helix is separated into two single strands each is the base-complementary image of the other. Each single strand is then used as a template to complete its base-complementary daughter strand. Complementary bases are replicated in opposite directions of the separated strands one at a time. This process is called by Watson and Crick the **semi-conservative replication** and the base pairing is called the **Watson-Crick base paring principle**.





Hydrogen Bond. Although each nucleotide as a whole is charge neutral, the electrons tend to be unevenly distributed. As a result some oxygen and nitrogen atoms of one base bear partial negative charges to draw electrons toward them, and some hydrogen atoms of a complementary base bear a partial positive charge to be drown toward the oxygen and nitrogen atoms, forming the **hydrogen bonds**. The heat energy needed to separate the hydrogen-oxygen bond is about half of the amount to separate the hydrogen-nitrogen bond (3 kcal/mol *v.s.* 6 kcal/mol).

2.2 DNA REPLICATION

The question of this section is why all life forms are coded in four bases but not 2 or six bases? We will construct a mathematical model for this problem, using the steps outlined in the previous section as a guide.

Step 1. Approximate Question.

Prioritizing important factors to a problem usually takes the form of making reasonable assumption. In this case, we assume DNA replication is the most fundamental function to life. The city morgue has as much DNA information as when all its corps were alive. The difference between living and dead is whether or not their DNA is replicated.

	H	T	R
Living organism	+	+	+
Dead organism	+	∞	0
Photon from the Sun	0	+	0
Robotic motion	0	+	0

Step 2. Conceptual Model.

We treat the DNA replication as signal transmission. Specifically, the genome of an organism is thought as a message sequence coded in the nucleotides bases, adenine (A), thymine (T), guanine (G), and cytosine (C), and the instantaneous event at which a base is replicated along the 2 single strands of the separated double helix is considered as the moment when the signal symbol, which is thought to represent the base, is transmitted.

Step 3. Mathematical Model.

Since the question is about the number of bases, we must assume the possibility that DNA replication may be done in bases other than 4. More precisely, a set of assumptions essential for replication as a communication channel is put forth. They are an extension of known mechanisms of DNA replication to base numbers other than 4. See the company text and diagram for an illustration.

HYPOTHESES

1. DNA replication is an all-purpose channel.
2. There is an even number n of nucleotides bases with $n \geq 2$.
3. Replication is done one base a time on two separated strands of the mother helix.
4. The bases can be paired as b_{2k-1}, b_{2k} according to the number of their hydrogen bonds, $k+1$. They are **complementary bases**.
5. Replication occurs when a base bonds to its complementary base by their hydrogen bonds.
6. The paring times of complementary bases are equal, i.e., $\tau_{2k-1} = \tau_{2k}$.
7. The pairwise paring times progress like the natural number with the A-T pairing time as the progression unit, i.e., $\tau_{2k-1,2k} = k\tau_{1,2}$ for $k = 1, 2, \ldots, n/2$.

Remarks on Hypotheses. For Hypothesis 1, we almost get it for free. In fact, upon the completion of replication, the number of bases are doubled, half of which is from the mother helix and the other half is newly acquired. Of the newly acquired, base A and base T have the same number, and so do base C and base G by Watson-Crick's base pairing principle. So the only unknown is whether or not the A-T pair and C-G pair are equally probably. Of course we do not expect them to be so for a particular organism just as we do not expect all our Internet traffics are equally distributed. The assumption is about the aggregated distribution of all genomes. This hypothesis simply assume that the DNA replication machinery is for all possible base combinations—the maximal genomic diversity. Hypotheses 2–6 are straightforward extension of the known 4-base process.

Hypothesis 7 deserves an undivided attention. Intuitively, it seems reasonable in light of Hypothesis 4. The more bonds there to pair the longer

time it takes to do so. An indirect empirical finding strongly suggests this assumption. In fact, we know that RNA transcription slows down in *G*-*C* rich regions.[1] More precisely, the hypothesis is based on the hydrogen bond structure of the bases. It is a biochemistry fact that the DNA backbone bonds are covalent bounds. Measured in terms of bonding energy, covalent bounds are about 20 times greater than hydrogen bonds. We also know that for DNA bases, the bonding energy of the hydrogen-nitrogen, $H\cdots N$, bond is twice the amount of the hydrogen-oxygen, $H\cdots O$, bond. One important consequence of having low bonding energy is that the lower the energy the longer the pairing time. Hence, the $H\cdots O$ bond takes the longest time to pair. Its bonding time is taken as the "0th" order approximation of the base pairing time. The *A*-*T* pair has one $H\cdots O$ bond, so its base pairing time, $\tau_{1,2}$, is one $H\cdots O$ pairing time. Hence, the *G*-*C* pairing time is twice as long for having two pairs of $H\cdots O$ bond. The assumption assume this pattern to persist: the 3rd base pair has three $H\cdots O$ bonds and so on. ◎

For this all-purpose model of DNA replication, the conditions of Theorem 1.16 are satisfied. Therefore, the mean rate solution to the model is given by

$$R_n = \frac{4 \lg n}{\tau_1[4 + (\alpha - 1)(n - 2)]},$$

with $\alpha = \tau_{3,4}/\tau_{1,2}$. We know from the theorem that having four bases gives the best mean replication rate for $\alpha = 2$.

Step 4. Interpretation.

The model only gives a provisional solution to the 4-base replication problem. It may advance our understanding on evolution in the following ways.

If the model is right, it will imply that life on Earth is where it should be in time. This is because information entropy measures how diverse the genomes in the pool of life is, and the mean transmission rate measures how much of the maximal diversity can go through the time bottleneck set up by the channel. Equivalently, the reciprocal of the mean rate, $1/R_n$, measures the time needed to replicate one bit of the maximal diversity. At the minimum replication time needed, each bit of the maximal diversity moves through time the fastest, leading to the greatest mutation rate and consequently the fastest adaptation rate. It also leads to the greatest consumption (metabolism) rate, thus out competing and wiping out all non-quaternary systems.

If the model is right, it will explain why DNA evolved away from a base-2 system if indeed life started with a protogenic form of base-2 replication in the G, C bases. From Fig.1.2(b), R_2 for the base pair *A*-*T* is $R_2 = \lg 2/\tau_{A,T}$. Thus, $R_2(G, C) = \lg 2/\tau_{G,C}$ for the base pair *G*-*C* is half as much since

[1] Uptain, S.M., C.M. Kane, and M.J. Chamberlin, Basic mechanisms of transcript elongation and its regulation, *Annu. Rev. Biochem.* **66**(1997), pp.117-172.

$\tau_{G,C} = 2\tau_{A,T}$. However, $R_4 = 1.35/\tau_{A,T} = 2.7R_2(G,C)$. Hence, the **mean evolutionary year** of the G-C system in terms of the quaternary system's year is

$$\frac{R_2(G,C)}{R_4} \times 3.5 \text{ billion years} = 1.2963 \text{ billion years},$$

assuming life started some 3.5 billion years ago. That is, the evolutionary clock would be set back by 2.2037 billion years in a G-C coded world.

Beside the standard four bases of nucleotides, there are others such as uracil (U) for RNA and inosine (I) occasionally found in DNA. If the model is right, it will suggest the existence of these bases to be the result of Nature's relentless, memoryless, and so far unsuccessful attempts to better the quaternary system.

Step 5. Modification.

The model presented above is susceptible to modifications due to its considerable number of hypotheses. The most susceptible of all is Hypothesis 1. An alternative is to consider the aggregated genome distribution P^*. However, it is impossible to know P^* exactly. It will be a long time before we can sequence all known species genomes. Genomes of some long ago extinct species may have lost forever. One exercise problem gets around this problem by suggesting that without the dynamical process of replication DNA prefers a static state correlated to the hydrogen bounding energies of the bases. Other modifications are also left to the Exercises.

Extension of this model to RNA transcription is also left to the Exercises.

Exercises 2.2

1. Compare the mean rates between the quaternary system and the binary system in A-T bases. Find the number of years that evolution would be set back by the A-T system assuming life started 3.5 billion years ago. Do the same for the base-6 system of the model.
2. Instead of the equiprobability hypothesis, assume (1) the aggregated distribution P^* is proportional to its hydrogen bonding energy, i.e., $p_{2k-1,2k}/p_{2\ell-1,2\ell} = E_k/E_\ell$, where $p_{2k-1,2k}$ denotes the probability of the kth base pair, and E_k denotes the pairs total hydrogen bond energy; (2) the base pair k has one $H\cdots N$ bond and k $H\cdots O$ bonds; and (3) $E_{H\cdots N} = 2E_{H\cdots O}$. Write a numerical program to find the mean rate $R(P^*)$ for the same base paring condition, Hypothesis 7. Show that $R_4(P^*)$ is still the optimal solution of mean rate.
3. Consider the same conditions of the previous exercise but assume the pairing times progress with increment $\Delta\tau$. Write a numerical program to show that the interval in $\alpha = \tau_{3,4}/\tau_{1,2}$ for which the mean rate $R(P^*)$ is maximal is approximately $[1.825, 3]$.
4. For the previous 2 exercises, further assume the following:

- The base pairing time $\tau_{2k-1,2k}$ is the sum of the bonding times of the base's hydrogen bonds.
- The bonding time of a hydrogen bond $H\cdots X$ is inversely proportional to its bonding energy $E_{H\cdots X}$ and the proportionality is the same for different hydrogen bonds, i.e., $\tau_{H\cdots X}/\tau_{H\cdots Y} = E_{H\cdots Y}/E_{H\cdots X}$.

Show that the corresponding ratio $\tau_{3,4}/\tau_{1,2}$ is 1.667. Numerically show also that $R_4(P^*)$ is still the optimal solution. (More generally, the corresponding optimal interval in $\alpha = \tau_{3,4}/\tau_{1,2}$ is approximately $[1.65, 2.7]$.) Is it the optimal solution as an all-purpose channel? (It will fall short by 9.5%.)

5. Like DNA replication, RNA transcription can also be considered as a communication channel. Do a research on RNA transcription. Identify a phase of RNA transcription as a communication channel. Construct an all-purpose channel model. Write a one-paragraph commentary on the speculation that the DNA code originated from RNA replication.

2.3 REPLICATION CAPACITY

Nature may prefer the quaternary system because it can produce the best evolutionary result for all species and on average, allowing the most species diversity to pass through the time. The focus of this section is instead on the individual genomic type which can naturally replicate at the channel capacity, referred to as **the replication capacity**. Such a species makes out the most that the replication mechanism offers, rushing through time the most genomic information ahead of all others.

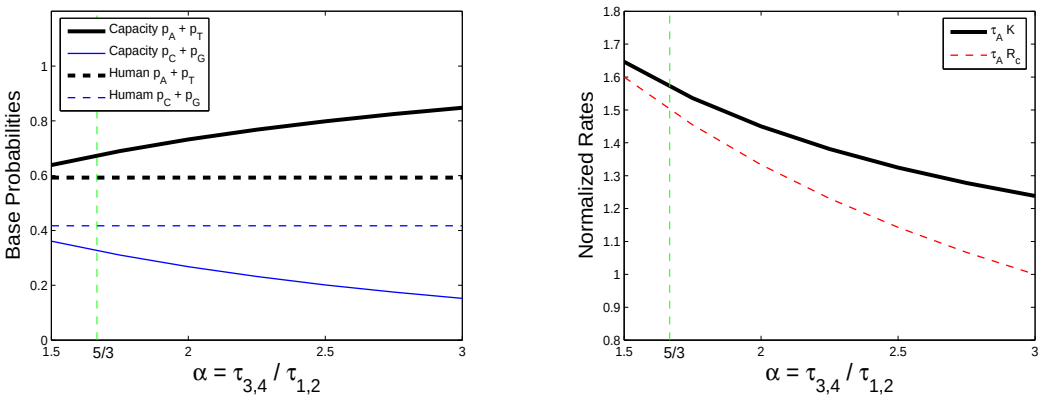
It is a straightforward application of Theorem 1.20 to find the replication capacity, K, of the DNA replication model.

Recall that $\tau_{_A} = \tau_1,\ \tau_{_T} = \tau_2,\ \tau_{_C} = \tau_3,\ \tau_{_G} = \tau_4;\ \tau_1 = \tau_2,\ \tau_3 = \tau_4$; and $\tau_{_{G,C}} = \alpha\tau_{_{A,T}}$ with $\alpha = 2$ for the model considered in the main text of Sec.2.2 and $1.65 < \alpha < 3$ in Exercises 2.2. Then, the capacity generating distribution $p_{_A} = p_{_T}, p_{_C} = p_{_G}$ satisfy:

$$p_{_G} = p_{_A}^{\alpha},\ 2(p_{_A} + p_{_A}^{\alpha}) = 1,\ K = \lg\frac{1}{p_{_A}}.$$

Fig.2.2(a) (`ch2capacityrate.m`) shows the graphs of pairwise probabilities $p_{_{A+T}},\ p_{_{C+G}}$. Fig.2.2(b) (`ch2ratecomparison.m`) shows the normalized capacity $\tau_{_{A,T}}K$ as well as the normalized mean rate $\tau_{_{A,T}}R_4$, all as functions of the parameter $\alpha = \tau_{_{G,C}}/\tau_{_{A,T}}$ from $[1.5, 3]$. Notice that the slower C, G bases pair with each other at the moment of replication (larger α value), the smaller the replication capacity becomes, and the less frequent the C-G pair should be to achieve the capacity rate.

Organisms that shrive near volcanic vents on deep-ocean floor maybe rich in C, G bases probably because of their thermal stability. The genome of



(a) (b)

Figure 2.2 (a) Comparison of the capacity-generating base distribution and the Human genome distribution. (b) Comparison of the normalized replication capacity and the diversity rate.

E. coli has a near equiprobability base distribution. The Human genome is rich in A-T bases as shown below (with the 1% discrepancy duely duplicated from the source).

Human genome base frequency[1]

Base pairs	Percent
$A + T$	54
$G + C$	38
undetermined	9

[1]Venter, J.C., *et al.*, The sequence of the human genome, *Science*, **291**(2001), pp.1304–1351.

Shaped by their particular constraints, each replicates at a genomic rate not necessarily at the mean. However some may replicate as much information as possible either by chance or by evolutionary pressure. Human's asymmetrical base distribution is curiously near the capacity distribution as shown in Fig.2.2(a). The apparent heterogeneities are qualitatively the same. In fact, the p_{A+T}, p_{C+G} graphs for the Human genome are produced by allocating the undetermined percentage in proportion to the determined percentages. That is,

$$p_{A+T} = 0.54 + 0.09 \times \frac{0.54}{0.54 + 0.38} = 0.5923,$$
$$p_{C+G} = 0.38 + 0.09 \times \frac{0.38}{0.54 + 0.38} = 0.4172.$$

Given the fact that the C-G pair is structurally more stable, the A-T portion of the undetermined is probably disproportionately larger, making the base frequency even closer to the capacity-generating distribution.

To summarize, both the mean rate and the capacity rate are about DNA replication. The former is about the best average for all and the latter the absolute best of one. According to the model, Nature's quaternary replication system is the optimal solution to the first problem and the Human genome is near the optimal solution to the second.

Exercises 2.2

1. Consider the binary C-G system. Find the replication capacity and its corresponding base probability distribution.
2. Find the exact replication capacity distribution for the DNA replication model when $\alpha = \tau_{G,C}/\tau_{A,T} = 2$.
3. If we use the DNA replication as a communication channel for which information is coded by the base pairs A-T and G-C rather than the 4 bases, and assume that the transmission times satisfy $\tau_{G,C}/\tau_{A,T} = 2$, then what is the channel capacity distribution? Find the channel capacity as well as the mean transmission rate.
4. Consider the replication system of even bases $2n$. Assume the replication times satisfy $\tau_{2k-1} = \tau_{2k} = k\tau_{1,2}$ for $k = 1, 2, \ldots, n$. Find the replication capacity probability $p_{1,2}$. Show that $\lim_{n\to\infty} p_{1,2} = 1/3$.

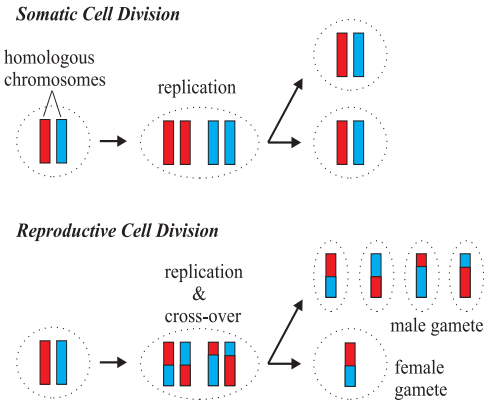
2.4 SEXUAL REPRODUCTION

DNA replication is a stochastic process by which genomes mutate over time. Mutation is recognized as a driving force of evolution. However, there is another equally important but taken-for-granted factor—the time scale alignment between DNA replication and environmental changes, climate change for one as a principle component. The former must operate at a fast time scale and the latter at a slow time scale. Otherwise evolution would be like pushing a marble through a syringe needle. Since these dynamical processes are otherwise uncoordinated, DNA replication must generate more mutations than evolution actually needs. Thus over time an organism will accumulate too many unusable mutations to keep its DNA replication running indefinitely. Therefore, reproduction is the logical and, obviously, the practical solution to this necessary problem of evolution—leave a working copy behind to continue the DNA replication process.

Step 1. Question.

Asexual reproduction works perfectly well in Nature. Why sexual reproduction, and why sexual reproduction by 2 rather than 3 or more sexes? We will again use mathematical modelling to quantify these questions.

Cell Divisions. DNA replication characterizes living. But living big requires cell division. There are two types of cell division: **meiosis** for reproductive cells and **mitosis** for non-reproductive cells, referred to as **somatic** cells. Reproductive cells produce sperms in male animals or spores in male plants, and ova in female animals and ??? in female plants. They are respectively referred to as **male gamete** and **female gamete**. Mitosis produces genetically identical daughter cells. Meiosis produces genetically varying gametes. Both are schematically shown here.

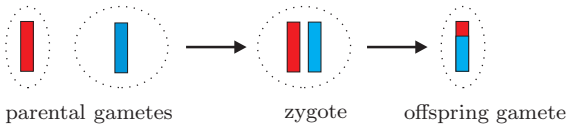


Chromosomes are numbered according to their length except for the sex-determining chromosomes, which are called the X, Y chromosomes. Parental chromosomes of the same number are called **homologous**. For mitosis, the diagram only highlights three critical phases: congregation of homologous chromosomes, replication, and cell division after chromosome recombination. Meiosis differs from mitosis in two critical ways. First, after replication, homologous chromosomes cross over and exchange parts of themselves. If homologous chromosomes were lottery tickets, sex-determining chromosomes would be the printing machine, an equipment that is best kept in the same condition ticket after ticket. Secondly, meiosis division produces 4 male gametes and one female gamete respectively from one male and one female reproductive cell, and each gamete contains only half as many chromosomes as the original reproductive cell. In human, each cell contains 23 pairs of chromosomes, or a total of 46 chromosomes. Each sperm and ovum on the other hand contains 23 chromosomes of all chromosome numbers plus either an X or a Y chromosome. The number of gamete's chromosomes is called the **haploid** number and that of cell's chromosomes is called the **diploid** number. The union of male and female gametes forms a cell called **zygote**, which then has the full set of chromosomes of a cell. Critical to our mathematical modelling is the fact that *each gamete is a genetically distinct mixture of both parents DNAs.*

Genes are segments of DNA that code proteins, the building blocks of cell. A gene can have different working versions, called **alleles**. In terms of genes, it is the parental alleles that are exchanged during meiosis.

Step 2. Conceptual Model.

Asexually reproductive species give their offspring one working copy of genome. Sexually reproductive species give their offspring a combination of two copies. The obvious advantage of sexual reproduction is to enhance its practitioners genetic diversities. Like DNA replication, we can think sexual reproduction as a communication system in a grander scale, as depicted below.



In this picture, each fertilizing gamete is considered as one packet transmission, and each packet contains a total information $\lg 2 \times L$, where $\lg 2$ is the entropy per exchanged segment of homologous chromosomes and L is the total number of exchanged segments along gamete chromosomes. The gain is the information of chromosome exchange. However, unlike the narrow definition of a communication channel, the cost to this information gain is not only in time but also in energy throughout the whole process of courtship $\longrightarrow$ conception $\longrightarrow$ reproductive maturity, as illustrated in the diagram. In other words, it is more accurate to think sexual reproduction as a stochastic process of which the product is a fertilizing gamete and the cost is in time or/and energy.

Step 3. Mathematical Model.

As a communication channel on one hand, reproduction is characterized by the mean transmission rate. On the other hand, it is more accurate to think it as a stochastic process with cost not just in time alone. Thus, it is more general and accurate to characterize the process by a payoff-to-cost ratio. The payoff is the information gain from parental chromosome exchange in offspring's gametes and the cost is in time or/and energy. Similar to the mean rate of communication channels, we like to think the 2-sexes reproduction as a constrained optimal solution to the entropy-to-cost ratio. The set of constraints is given by the following hypotheses.

HYPOTHESES

1. There are n sexes and reproduction requires the recombination of gamete chromosomes from all sexes.
2. Each gamete autosome (non-sex-determining chromosome) is a mixture of its contributing sex's parental homologous chromosomes by the crossover process.
3. The mixing probability at any exchanging site along any gamete chromosome is the same for all parental sexes, i.e., the equiprobability $1/n$ from each parent.
4. The sex ratio of any pair of sexes is 1:1.
5. The time and energy required to produce a fertilizing gamete is proportional to the average number of randomly grouping n individuals that has exactly one sex each, called a **reproductive grouping** below. ◎

Remarks on Hypotheses. Hypothesis 1 is necessary to fix the unknown variable. Hypotheses 1 to 4 are true for $n = 2$ as discussed in the company text. More specifically, for Hypothesis 3, an exchanging segment can be a sequence of many bases or genes. The model applies to whatever length a segment may actually be. The equiprobability part of Hypothesis 3 follows from the following facts. First, when a pair of mixed homologous chromosomes split, at any mixing segment one copy is from one sex and the other copy is from the opposite sex. Thus, there is always an equal num-

ber of exchanged copies from all sexes at any site and in any population of gametes. Secondly, the exchange of parental alleles is believed to be independent from segment to segment so that each gamete contains a unique mix of its contributing sex's parental DNAs. Further factoring the fact that it usually takes an overwhelming number of gametes for each fertilization, we can indeed assume the mixing to be completely thorough and thus the equiprobability. As a consequence, the information entropy of the chromosome mixing is maximal, denoted by $H_n = \lg n$ in bits per segment and referred to as the **reproductive entropy** or **reproductive diversity**. It is the same for all gamete autosomes. As expected, the more parental sexes there are, the greater the reproductive diversity per exchanging site is. As a result of this hypothesis, the model does not discriminate against any sex's genetic contribution to reproduction.

Hypothesis 4 can be considered as the 0th order approximation to the sex ratio. In fact, for $n = 2$ it is a genetic consequence to the fact that the sex-determining chromosomes, X and Y, are equally distributed in male gametes. It is not hard to concoct hypothetical schemes to maintain the equiratio for $n \geq 3$ cases, which is left as an Exercise. Alternatives to the equal sex ratio will not be pursued because of its molecular origin from the more fundamental process of meiosis.

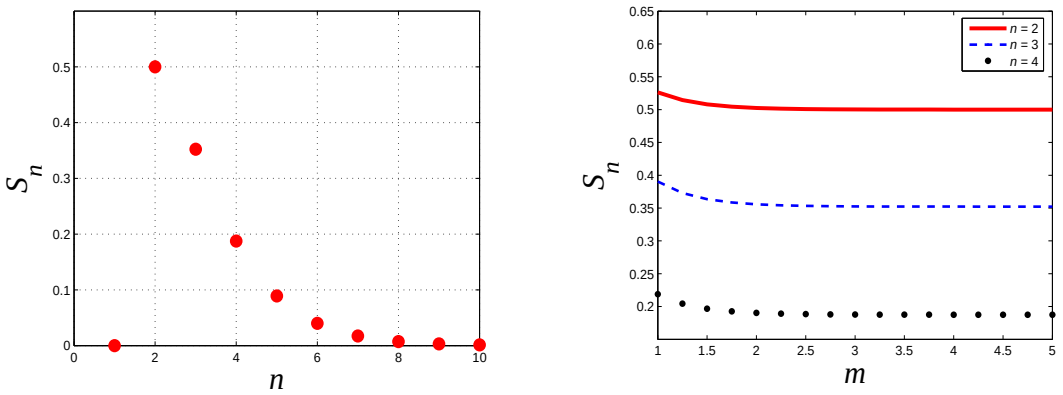
At this moment we can regard the assumed mating process by Hypothesis 5 as a simple chance encounter, i.e., a "double-blind" model ("blindfold blind date"). Further justification will be given after its mathematical representation is derived below.

We are ready to translate the narrative model into mathematics. We already have $H_n = \lg n$, the per-site reproductive entropy. Let E_n be the dimensionless cost (without the proportionality from Hypothesis 5) in time or/and energy over the whole reproduction cycle from parental to offspring fertilizing gametes. The ratio, $S_n = H_n/E_n$, is called the **reproductive entropy-to-cost ratio** over the number n of possible sexes. The optimization objective is to maximize S_n over n.

Derivation of Cost. Without loss of generality from Hypothesis 4, assume each sex has a same number, M, of individuals. Then there are $\binom{nM}{n} = \frac{(nM)!}{n!(nM-n)!}$ many ways to choose a group of n individuals from the total nM many individuals of all sexes. Of which only M^n many are reproductive groupings by Hypothesis 1. Hence, the reproductive probability is $p_{n,M} = M^n/\binom{nM}{n}$ and

$$\begin{aligned} p_n &= \lim_{M\to\infty} p_{n,M} = \lim_{M\to\infty} \frac{M^n}{\binom{nM}{n}} = \lim_{M\to\infty} \frac{n!M^n}{nM(nM-1)\cdots(nM-n+1)} \\ &= \lim_{M\to\infty} \frac{(n-1)!}{(n-\frac{1}{M})(n-\frac{2}{M})\cdots(n-\frac{n-1}{M})} = \frac{(n-1)!}{n^{n-1}} \end{aligned}$$

for an infinite population.



(a) (b)

Figure 2.3 (a) Infinite population. (b) Finite population with equal sex population $M = 10^m$.

Let $P(k)$ be the probability that the kth try is the first reproductive grouping. Then $P(k) = (1 - p_{n,M})^{k-1} p_{n,M}$, valid for both finite and infinite M, and the average number of tries to have a reproductive grouping is $E_n = \sum_{k=1}^{\infty} kP(k) = p_{n,M} \sum_{k=1}^{\infty} k(1 - p_{n,M})^{k-1}$. Using the identity that $\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$ (see Exercises), we have

$$E_n = \frac{p_{n,M}}{(1 - (1 - p_{n,M}))^2} = \frac{1}{p_{n,M}},$$

which is the possibility of each conception.

The reproductive entropy-to-cost ratio is

$$S_n = \frac{H_n}{E_n} = \frac{\lg n}{1/p_{n,M}}.$$

Step 4. Interpretation.

Fig.2.3(a) shows the graph of S_n when $M \sim \infty$. Clearly, S_2 is the optimal solution. That $S_1 = 0$ is expected since asexual reproduction has zero reproductive entropy. Fig.2.3(b) shows some graphs of S_n as a function of finite equal sex population $M = 10^m$. The limiting ratios are good approximations beyond a modest size $M = 100$. Surprisingly, the 2-sexes reproductive strategy remains optimal even when the population size is small, $M = 10$.

The dimensionless cost function $E_n = 1/p_{n,M}$ is the possibility of each successful fertilization. For $n = 2$, $E_2 \sim 2$. That is, for each reproductive interaction between two opposite sexes, there is one non-reproductive interaction between like sexes. Like-sex interactions can be in the forms of competition for mating or cooperation for offspring rearing. Similar inter-

pretation applies to $n > 2$ cases. Thus, E_n is a reasonable functional form for reproductive cost at the population level.

The prediction that S_2 is the optimal solution is expected from any reasonable model. Some immediate implications are nevertheless surprising. For species having the same average number, L, of chromosome exchange sites, the unit reproductive entropy-to-cost ratio S_n can be used as an intrinsic measure for comparison. Since $S_2/S_3 \sim 1.43$, a 3-sexes reproductive strategy will reduce the part of biological diversity that is due to sexual reproduction by 43% at every evolutionary stage, given the same amount evolutionary time and energy. Equivalently, since the reciprocal $1/S_n$ measures the minimal time or/and energy required for each bit of reproductive diversity, a 3-sexes strategy will set back the evolutionary clock that is due to sexual reproduction by

$$3.5 - \frac{S_3}{S_2} \times 3.5 = 1.0525 \text{ billion years},$$

assuming life started some 3.5 billion years ago. All these are good reasons why a 3-sexes reproductive machinery is unknown to be invented by Nature.

There are two ways to compare and contrast the asexual and 2-sexes reproductive strategies since Hypotheses 1–5 can be thought either to apply trivially to the asexual case or not at all. In the first case, take for an example multiparous mammals which could have their litters effortlessly cloned from one fertilized egg but did not. For them, each gamete's reproductive entropy-to-cost remains at $S_2 = 0.5$ in bits per exchange segment per cost *v.s.* $S_1 = 0$ for the would-be cloned embryos. Since $1/S_n$ measures the minimal time or/and energy required at the organismic level for each bit of sexual reproductive diversity, that $1/S_1 = \infty$ implies that such species, all mammals included, would never appeared if they adopted the asexual reproductive strategy. In this regard, our model is consistent with these known sexual reproductive realities.

In the second case, S_n cannot be used to quantify differences between the asexual and 2-sexes reproductive strategies because there is no parental diversity to begin with for the former. Therefore, they must be treated as two distinct categories second only to the supreme purpose of DNA replication. Nevertheless, our model offers another insight into the asexual reproductive strategy. Its continued usage can be explained by the principle reason that sexual reproduction is not a necessary but only a sufficient way to increase genetic diversities. Less complex organisms, such as bacteria, may be able to generate enough genetic diversities by DNA replication alone to compensate for their lack of reproductive diversity. In this regard, asexual reproductive realities do not contradict our model.

Step 4. Modification.

The cost function derived from Hypothesis 5 may be considered as a 0th approximation of sexual reproductive process. It is a reasonable functional form at the population level for species interactions as pointed out before.

However, it may need higher order corrections. For examples, at the cellular level, somatic maintenance is a necessary reproductive cost expenditure, and at the organismic level, growth before sexual maturity is another. The question is will such modifications alter the conclusion easily?

To answer this question, observe from Fig.2.3 that the optimal solution S_2 is quite robust against the next best solution S_3. In fact, the difference between S_2 and S_3 is about 30% and 43% of S_2 and S_3 respectively: $|S_2 - S_3|/S_2 = 0.3$, $|S_2 - S_3|/S_3 = 0.4285$. This implies that the S_2 optimal solution can indeed tolerate high order corrections of considerable magnitude and persist.

Exercises 2.4

1. Explain why Theorem 1.20 does not apply to sexual reproduction.
2. Verify the formula $\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$, using the formula $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ for $|x| < 1$, and one of the following two ways: (a) by multiplication of infinite series; or (b) by differentiation of infinite series, that if $f(x) = \sum a_k x^k$ then $f'(x) = \sum k a_k x^{k-1}$.
3. Construct an (artificial) equal sex ratio scheme for a 3-sexes reproductive scenario.

2.5 DARWINISM AND MATHEMATICS

Biology is the branch of natural science in which mathematics plays the least role. Darwin's theory of evolution is even more removed from mathematics. However, the mathematical models presented in this chapter for the origins of DNA replication and sexual replication suggested otherwise. A careful examination reals a far more closer parallel between these two fields than conventionally thought, as illustrated in the table below.

	Evolution	**Mathematics**
Darwinism	Survival of the Fittest	Constrained Optimization
	Common Descent	Hierarchical Reduction
	Punctuated Equilibrium	State Modelling
Modern Genetics		Stochastic Process

First, Darwin's theory of survival-of-the-fitness mirrors the central dogma of constrained optimization in mathematics: maximizing payoff, minimizing cost with constraints. This point is intuitively and conceptually trivial—every species, extinct or present, can be considered as a constrained optimal solution. Imagine if we could travel back in time and observe evolution at every stage, we would be able to write down all fitness objective functions and all constraint conditions for every species so that the optimizing solution to the functions over the constraints is a mathematical representation of the

Darwinism. According to Darwin, evolution is driven by **natural selection** and **sexual selection**. Both are processes by which species change over time, as a result of the propagation of heritable traits that affect the capability of individual organisms to survive and reproduce. Sexual selection depends on the success of certain individuals over others of the same sex, in relation to the propagation of the species; while natural selection depends on the success of both sexes, in relation to the general ecological conditions. Paradoxically, however, sexual selection does not always seem to promote survival fitness of an individual. Some contemporary biologists prefer to make a distinction between the two by referring to the former as "ecological selection" and referring the combination of both as "natural selection" which we will adopt. It is commonly characterized by the phrase "survival of the fittest". Darwin's overall theory of evolution also includes: **common descent**, **gradualism**, and **pangenesis**. Darwin's natural selection principles give rise to a mechanism by which new species emerge, leading to his theory of universal common descent. The last universal common ancestor is believed to have appeared about 3.5 billion years ago. Modern genetics indeed supports his common descent theory. The theory of gradualism can be summarized by a direct quote of Darwin's, "the periods during which species have undergone modification, though long as measured in years, have probably been short in comparison with the periods during which they retain the same form." This leads to the contemporary concept of "punctuated equilibrium", which we will use for gradualism. Pangenesis was Charles Darwin's hypothetical mechanism for heredity, which is considered flawed. It was replaced by Gregor Mendel's laws of heredity. Bringing together Darwin's theory of natural selection and Mendel's theory of genetics give rise to the **modern evolutionary synthesis**, often referred to as **neo-Darwinism**. Today we think evolution to be driven by random mutation of DNA bases and natural selection.

The following quote from various online sources summarizes the philosophical impact of Darwinism on Western thinking: "Darwin's evolutionary thinking rests on a rejection of essentialism, which assumes the existence of some perfect, essential form for any particular class of existent, and treats differences between individuals as imperfections or deviations away from the perfect essential form. Darwin embraced instead what is called population thinking, which denies the existence of any essential form and holds that a class is the conceptualization of the numerous unique individuals. Individuals, in short, are real in an objective sense, while the class is an abstraction, an artifact of epistemology. This emphasis on the importance of individual differences is necessary if one believes that the mechanism of evolution, natural selection, operates on individual differences."

species at the corresponding stage of evolution. The non-triviality however is of practicality, resulting in far too many missing links between biology and mathematics in the past, perhaps forever for most species. Simply stated, for any process shaped by evolution, how it works should tell us why it does so by reasons of optimization.

Secondly, Darwin's common descent theory implies that evolutionary problems can be prioritized and compartmentalized so that a reductionistic modelling approach can be effective. This point is based on the following arguments. If we believe that new and higher order organisms evolved from simpler ones, then we must held his evolutionary principles to be applicable to life's all levels: molecular, genetic, cellular, organismic, and phenotypic. Although it is extremely difficult to model evolution by constrained optimization as a whole—not only the number of fitness functions and con-

straints of a given species is no doubtably enormous but also they always change with time, the same practical challenge at the genetic and molecular level may not be as acute. All species share a surprisingly few fundamental commonalities: 4-base replication, 3-codonization for amino acids, 20 amino acid groups, 2-sex reproduction scheme, etc. Making the matter seemly easier is the observation that these fundamental commonalities seem to form an inverted hierarchy with the 4-base replication problem at the root, thus opening themselves to a reductionistic treatment.

Thirdly, Darwin's theory of gradualism together with its implied theory of punctuated equilibrium give the reductionistic approach a theoretical ground to construct "punctuated equilibrium" or state models one level a time, each decouples from its upper level "equilibria". The 4-base DNA replication model is an equilibrium model oblivious to evolutionary features build atop long after its origination. The 2-sexes model is another, disentangled from other reproductive features appeared long thereafter, but building on top of the replication model as one of its optimization constraint. Without punctuated equilibriumization, constrained optimization models would be extremely complex, having to link many if not all levels at once.

Lastly, modern genetics imposes stochastic formulation on most if not all state models of constrained optimization. Without the stochasticity we will have to model evolution's every creation as a deterministic trajectory, a mathematical nightmare. Without optimization, we will have to appeal to God's intervention to impose arbitrary orders, the end of reasoning.

To summarize the Darwinism and mathematics parallel, the constrained optimization approach deconstructs evolution level by level, down to its molecular root, only to build it up in stochastic state models level upon level from the bottom. ◎

Further implications of this stochastic constrained optimization to evolution (SCOTE) are as follows.

To reconcile with the paradoxical observation that sexual selection does not always optimize, the explanation lies in the fact that constrained optimal solutions are not necessarily absolute—an optimal solution at one level may not be so at a higher level, or when its underlining constraints are no more. Consider the problem of equal sex ratio. As a 0th order approximation it is very accurate even among those mammals, such as African elephant and lion, for which a disproportionate number of males never reproduce. It suggests that sexual evolution does not optimize. From the view of SCOTE, such a reproductive inefficiency is only a secondary effect to the more fundamental priority of maximizing the reproductive entropy-to-cost ratio and to the equiratio's origin as a fundamental consequence to the meiosis evolution. Recall that the entropy-to-cost ratio can be the mean reproductive rate if the cost is in time. Thus, it must be a secondary consideration if indeed the mean reproductive rate is the priority but not the entropy-to-energy. Further strengthening the argument is the possibility that the reproductive diversity may be sufficiently fulfilled from the contribution of a few mating

males because hundreds of million of distinct male gametes from one male take part in one conception and adding more males to the gene pool does not raise the already saturated pool level from the few. In short, changing it from an effect to the cause to override the constraints (Hypotheses 2, 3, Sec.2.4) will violate the hierarchical order of evolution. By the same reason, the group's reproductive priority cannot supersede its individual's DNA replication priority. This may explain why those male species fight but not fight to death for the reproductive privilege. Nature has won a jackpot in 2-sexes procreation. It allows some species to waste some of the fortune to engineer social structures in which the strong is defined by a mass of the week.

To further expand on the point of hierarchical and punctuated equilibrium modelling, consider the problem of disproportionality in male and female gamete size. A conceptual optimization model, probably not from the communication theory, would view Nature's way of having small male gamete and large female gamete as an optimal strategy to maximize the hit probability between the two. That is, sexual reproduction essentially employs a "dumb bomb" strategy for long range haul and a "smart bomb" strategy for near range hit, using small male gametes in hundreds of millions to hit one stationary and large female gamete. Whatever constraints maybe the goal of such a model is to guarantee a sure hit at a minimum cost, and everything else should be compartmentalized at higher hierarchical levels. The Pentagon may come up such a model faster than any mathematics or biology department.

The stochastic formulation of both replication and reproduction models paints the following picture of Darwin's theory of evolution. First, since parents cannot choose the genetic composition of their offspring (Hypothesis 3, Sec.2.4) and offspring cannot choose its parents (Hypothesis 5, Sec.2.4), the notion of individual ownership of DNA cannot be well-defined. Thus each organism is only an accidental and temporary carrier of the protogenic DNAs at the origin or origins of life. This point is not new since it comes from the stochastic nature of Mendel's inheritance theory. Secondly however, the maximal diversity entropies in bits per DNA base or per exchange segment of chromosome reside in each organism in a suspended probabilistic state all the time, and it is by the dynamical processes in replication and reproduction that the maximal entropies are expressed through time, expanded at length, and multiplied in space, giving rise to an information entropy explanation of the common descent theory, the evolutionary equivalence of the Big Bang. Thirdly, because the 4-base replication and 2-sex reproduction are optimal strategies, evolution is where it should be in time and diversity, supplementing Darwin's mechanisms of natural selection at a more fundamental level.

Chapter Three

Spike Information Theory

Neurons are fundamental units of animal brains. Human consciousness and intelligence are physical states of our brain. Someday when we fully understand both we will have understood neurons in all nitty-gritty details. Of which we will have had answers to these simple questions: Does a neuron have an optimal information mean rate? Does it have a preferred information distribution to realize its channel capacity?

3.1 SPIKE CODES

Billions of neurons light up right at this moment you read these words. One popular conceptual model in the area of Artificial Intelligence treats the firing mechanism of a neuron as a binary devise having two modes: "on" (1) and "off" (0). That is, if you can hook up this binary model to an oscilloscope, you will only get one type of output in either voltage or current, both looked the same as below:

Proposition 3.1 *Let $B = \{0, 1\}$ be the binary alphabet for a channel. Assume the symbol transmission times are equal, $\tau_0 = \tau_1 = \theta_0$. Then the mean rate R_2 and the channel capacity K are equal*

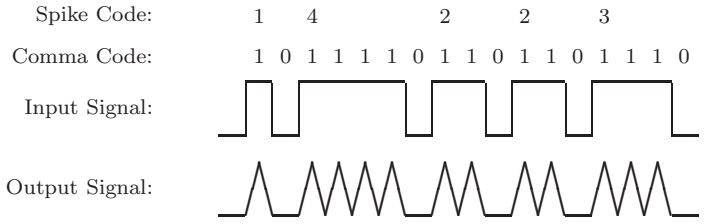
$$R_2 = K = 1/\theta_0$$

and the capacity distribution is the equiprobability distribution.

A proof is based on Theorems 1.15 and 1.20. It is left as an exercise.

Presumably the brain of the alien species *Bynars* from the scifi TV series, *Star Trek: The Next Generation*, is build this way. It is too crude a model for our brain.

As a first modification of the binary artificial neuron, a neuron is thought as a black box having one input and one output. If you hook it up to a voltmeter or ampmeter, the signals look like the following:



That is, the signals are characterized by a comma code: 10, 110, 1110, 11110, ... Let θ_1 be the process time for the "on" symbol 1 and θ_0 be the process time for the "off" symbol 0. Then the code symbol process times are

$$\tau_k = k\theta_1 + \theta_0 \text{ for all } k.$$

Since each code symbol can be identified by its number of 1s, i.e., 1 for 10, 2 for 110, etc., we have the following definition.

Definition 3.1 *A **spike code** of size n consists of the channel alphabet $S_n = \{1, 2, \ldots, n\}$ together with its corresponding transmission times of the form*

$$\tau_k = k\theta_1 + \theta_0 \text{ for all } k = 1, 2, \ldots, n,$$

*where θ_0, θ_1 are nonnegative parameters. θ_1 is called the **spike interval** and θ_0 is called the **intersymbol interval**. S_n is called the **spike alphabet** and $\{\tau_n\}$ is called the sequence of **spike times**.*

We can think such a phenomenological neuron either as an encoder or a decoder. As an encoder, the input signal can be thought as a representation of some information sources internal or external, and the output signal as a discretization of the source. As a decoder, the input signal can be thought from a channel and the output signal as a decoded representation of the source. For this dichotomy, we call it a **spike coder**.

Proposition 3.2 *For a spike code of size n with symbol transmission times*

$$\tau_k = k\theta_1 + \theta_0 \text{ for all } k = 1, 2, \ldots, n,$$

the mean transmission rate R_n is

$$R_n = \frac{2 \lg n}{\tau_1[2 + (\alpha - 1)(n - 1)]}, \tag{3.1}$$

where $\alpha = (2\theta_1 + \theta_0)/(\theta_1 + \theta_0) = 1 + 1/(1 + \theta_0/\theta_1)$.

A proof is obtained by Theorem 1.15 using $\Delta\tau = \theta_1$.

By definition, an **ideal spike code** is one so that the intersymbol interval is zero $\theta_0 = 0$. A coder can be considered near ideal if θ_0 is sufficiently small relative to the spike interval θ_1, i.e., $0 < \theta_0/\theta_1 \ll 1$ (meaning sufficiently near zero of the quotient). In such a case, $\alpha \sim 2$ for which R_4 is the optimal rate, see Fig. 1.2(a). That R_5 is an optimal mean rate is considered in the Exercises.

By definition, a **block spike code** is one so that the first spike symbol has a fixed number, n_0, of spikes which is used as a base, and the second spike symbol has $2n_0$ spikes, the third spike symbol has $3n_0$ spikes, and so on. In this way, we can substitute $n_0\theta_1$ for θ_1 in the analysis above, and the block spike coder for large n_0 is close to being ideal: $0 < \theta_0/(n_0\theta_1) \ll 1$. The same result of the Proposition applies.

Here are some examples that a quaternary spike code or block spike code *may* play a role.

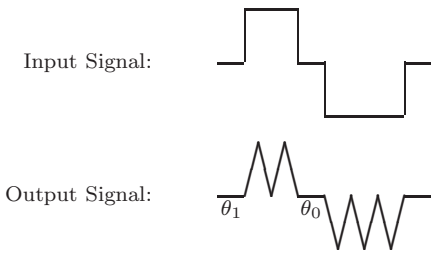
- Most spoken languages have 4 to 5 vowels to carry the conversation.
- Studies show that human brains can handle up to 4 tasks simultaneously without a significant comprise in efficiency and accuracy.
- The common tabulating method using one line increment up to a group 5 as illustrate below:

  Example:  = the number 7.

  In this example, if the tabulation is going on indefinitely, then the distribution of number 1, 2, 3, 4, 5 is in fact the equiprobability distribution.
- Children in India learn counting by a base 4 system. ◎

Exercises 3.1

1. Prove Proposition 3.1.
2. Prove Proposition 3.2.
3. Consider a block spike code of a base number n_0. Assume $\theta_1 = \theta_0$. Find the smallest n_0 such that the mean rate R_4 is optimal.
4. What is the size of the alphabet if a spike coder has the optimal mean rate when $\theta_1 = \theta_0$?
5. Consider a **polar-spike coder** whose output not only can produce up spikes but also down spikes as shown.



Let the corresponding polar-spike code take an up or down designation. For example, 2_u, 3_d would be the up 2-spike and the down 3-spike signal output as shown. The up and down spike symbols of the same spike number are naturally paired in their processing times: $\tau_{2k-1} = \tau_{2k} = k\theta_1 + \theta_0$ for the pair of k spikes, with θ_1 being the unit spike interval and θ_0 the intersymbol interval.

(a) For n even, find R_n.
(b) For the ideal case when $\theta_0 = 0$, show that R_4 is the optimal rate.
(c) Comparing with the mean rate from Proposition 3.1, which system is better, the conventional binary system or the ideal polar-spike system?
(d) Show that in terms of the mean rate, the polar-spike system is always better than the conventional binary system if $\theta_0 < \theta_1/2$.

3.2 GOLDEN RATIO DISTRIBUTION

Neural information that has the fastest information transmission rate through a spike coder is given by the following result.

Proposition 3.3 *For a spike code of size n with symbol transmission times*

$$\tau_k = k\theta_1 + \theta_0 \text{ for all } k = 1, 2, \ldots, n,$$

the channel capacity is $K = -\lg p_1/\tau_1 = -\lg p_k/\tau_k$ with the source distribution satisfying

$$p_k = p_1^{(k+\beta)/(1+\beta)} \text{ and } \sum_{k=1}^{n} p_1^{(k+\beta)/(1+\beta)} = 1$$

where $\tau_k/\tau_1 = (k+\beta)/(1+\beta)$ and $\beta = \theta_0/\theta_1$.

A proof follows from Theorem 1.20 with $\Delta\tau = \theta_1$. For a binary spike code $n = 2$, the capacity distribution is listed below

$$p + p^{(2+\beta)/(1+\beta)} = 1$$

with p for the probability of the 1-spike symbol and $1 - p$ for the 2-spike symbol, where as above $\beta = \theta_0/\theta_1$.

For the ideal spike code with $\theta_0 = 0$, the capacity distribution is the Golden Ratio Distribution

$$p = \frac{\sqrt{5}-1}{2} = 0.6180,\ 1 - p = p^2 = 0.3820.$$

In this book, we use $\Phi = \frac{\sqrt{5}-1}{2}$ to denote the Golden Ratio. In the literature however you will likely find that the notation Φ is reserved for $\frac{\sqrt{5}+1}{2} = 1.6180$.

Example 3.2.1 Golden Sequence. Here is an example of a sequence message from a Golden Ratio distribution. The sequence is generated by the following algorithm. Star with the number 1. Replace 1 by 10. From the segment 10 onwards, replace each 1 by 10 and each 0 by 1 to generate the next sequence. The first 6 sequences are

1
10

101
10110
10110101
1011010110110

and so on. Note that the numbers of 1s in the sequences are 1, 1, 2, 3, 5, 8, ..., and the numbers of 0s starting at the second sequence repeat those of the 1s: 1, 1, 2, 3, 5, ... Let $N_{n,1}$ be the number of 1s of the n-sequence and $N_{n,0}$ be the number of 0s of the nth sequence. It is left as an exercise to show that $N_{n,0} = N_{n-1,1}$, and $\{N_{n,1}\}$ is the Fibonacci sequence, $N_{n,1} = F_n = F_{n-1} + F_{n-2}, F_1 = F_2 = 1$. In addition, the probability, p, of symbol 1 satisfies

$$p = \lim_{n\to\infty} \frac{N_{n,1}}{N_{n,0} + N_{n,1}} = \Phi.$$

This means that if we use the 1-spike symbol to encode 1 and the 2-spike symbol to encode 0, then the encoded Golden Sequence will reach the capacity information rate when it goes through the ideal binary spike coder. ◎

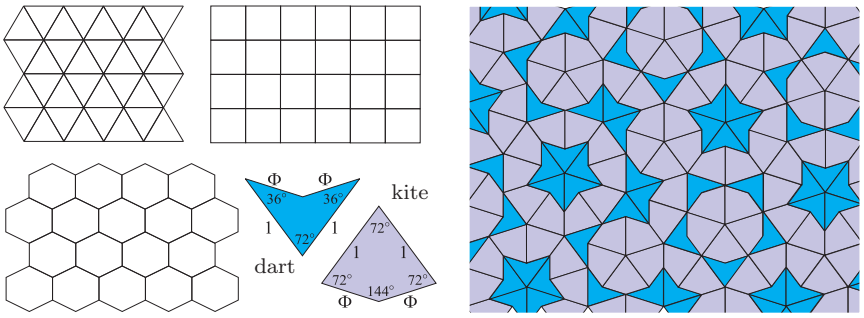
Example 3.2.2 Penrose Tiling. The equilateral triangle, the square, the pentagon, and the hexagon are geometric objects of symmetry. If you rotate the equilateral triangle 120° at its center, you get the same object. The same applies to the other objects by rotating them 90°, 36°, and 30° respectively. However, the pentagon is markedly different. You can tile your kitchen floor using square tiles only, or equilateral triangles or hexagons. But you cannot tile it with pentagons without gaps or holes. Sir. Roger Penrose discovered in 1974 that you can cover your floor with two shapes each, like the pentagon, returns to itself after five rotations of 36° each at any of its vertices. The company "dart" and "kite" give one example of the Penrose tiling pairs. The Golden Ratio is the defining ingredient. Unlike your ordinary kitchen floor however, a Penrose floor is **aperiodic**, it never repeats. As long as two like tiles do not touch along an edge and form a parallelogram at the same time, you can cover an infinite plane without gaps or holes. Here is another interesting property: let $N_{\text{dart}}(n)$ and $N_{\text{kite}}(n)$ be the numbers of darts and kites in the square, $Q_n = [-n, n] \times [-n, n]$. Then as the size of the square increases to infinity, we have

$$\lim_{n\to\infty} \frac{N_{\text{dart}}(n)}{N_{\text{kite}}(n)} = \Phi.$$

Alternatively, let $p_{\text{dart}}(n) = N_{\text{dart}}(n)/(N_{\text{dart}}(n) + N_{\text{kite}}(n))$ and $p_{\text{kite}}(n) = N_{\text{kite}}(n)/(N_{\text{dart}}(n) + N_{\text{kite}}(n))$ be the approximating probabilities, then

$$p_{\text{kite}} = \lim_{n\to\infty} p_{\text{kite}}(n) = \Phi,\ p_{\text{dart}} = \lim_{n\to\infty} p_{\text{dart}}(n) = \Phi^2,$$

the Golden Ratio probability distribution.



More specifically, take the 2-element set of kite and dart as an information source alphabet. Pick a tile from an infinite or very large Penrose floor as the center. Form a sequence of kites and darts in anyway you want from the square Q_1. Add new kites and darts to the sequence in any way as long as they are from the enlarged square Q_2 that have not be included already. Continue this way for the square $Q_3, Q_4, \ldots$ and so on. This will generate an infinite sequence or very long sequence. Different choices of the initial point and different ways to concatenate the sequence as new kites and darts become available from the expanding squares produces different sequences. This defines an information source and the probability distribution is the Golden Ratio distribution.

Furthermore, if we encode the kite by the 1-spike symbol and the dart by the 2-spike symbol, then the encoded Penrose floor will reach the capacity rate in an ideal spike coder. What is about your ordinary kitchen floor? The information rate is zero for being predictably periodic!

One important conclusion of this section is: for an ideal binary spike coder, the channel capacity distribution of an information source is the Golden Ratio distribution.

Exercises 3.1

1. Prove Proposition 3.3.
2. For the Golden Sequence example, use induction to verify
   (a) $N_{n,0} = N_{n-1,1}$.
   (b) $\{N_{n,1}\}$ is the Fibonacci sequence,
   $$N_{n,1} = N_{n-1,1} + N_{n-2,1}, N_{1,1} = N_{2,1} = 1.$$
   (c) $\lim_{n\to\infty} \dfrac{N_{n,1}}{N_{n,0} + N_{n,1}} = \Phi.$

3. For the Penrose Tiling example, assume

$$\lim_{n\to\infty} \frac{N_{\rm dart}(n)}{N_{\rm kite}(n)} = \Phi$$

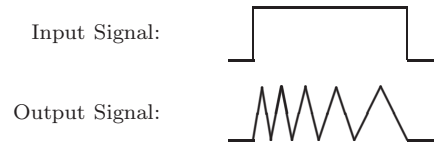
is true. Verify the following formulas

$$p_{\rm kite} = \lim_{n\to\infty} p_{\rm kite}(n) = \Phi,\ p_{\rm dart} = \lim_{n\to\infty} p_{\rm dart}(n) = \Phi^2.$$

4. Consider a binary block spike code of a base spike number $n_0 = 10$. If $\theta_1 = \theta_0$, find the channel capacity distribution p_1, p_2. Show that its deviation, $(p_1 - \Phi)/\Phi$, from the Golden Ratio is less than 2%.

3.3 OPTIMIZING SPIKE TIMES

We have considered the case in the previous section that the constitutive spike interval θ_1 is constant for all spike symbol. It should be taken as a "0th" order approximation of such. In reality, the progression of spikes in a spike symbol may not be as neat as the natural number progression. For example, even if the intersymbol interval θ_0 is ignored (the ideal spike coder), τ_1 may be longer or shorter than half of τ_2. In fact, for a spike symbol of large spikes, the time difference between adjacent spikes elongates towards the termination of spikes. This phenomenon is called **spike adaptation** in the literature and it is illustrated as follows.



By definition, an **adaptive spike code** of size n consists of the channel alphabet $S_n = \{1, 2, \ldots, n\}$ together with its corresponding transmission times of the form

$$\tau_k = k\theta_1 + \theta_0 + a_k \text{ for all } k = 1, 2, \ldots, n,$$

where θ_0, θ_1 are nonnegative parameters, and $\{a_k\}$ is a monotone increasing sequence. $\{\tau_k\}$ is the sequence of **adaptive spike times**. A spike coder endowed with an adaptive spike code is called an **adaptive spike coder**.

In conclusion, the spike interval θ_1 varies, but assuming it a constant is a good "0th" order approximation. This is particularly relevant for block spike codes.

However, the consequence to varying θ_1 is important. It leads to varying time ratios $\frac{\tau_k}{\tau_1}$, which can be different from the natural number progression $\mathbb{N}$ even if the spike coders is ideal, $\theta_0 = 0$. For example, for a binary spike coder, if $\alpha = \tau_2/\tau_1 \neq 2$, then the capacity distribution will be different from the Golden Ratio distribution. Neurologically, if an information source of

capacity distribution is somehow related to the psychological phenomenon of preference, for example, then differences in capacity distribution would accommodate individual preferences. An example of this sort is considered in the Exercises.

We will not purse adaptive spike coders further for the following reasons:

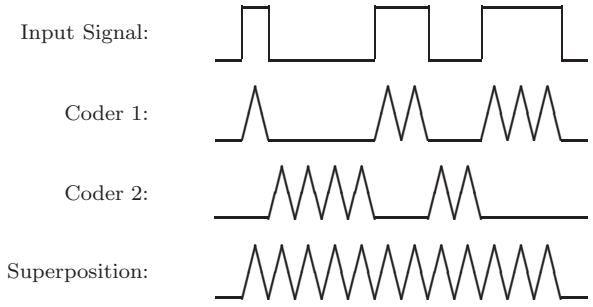
- We will consider primarily spike codes of small alphabet size, no more than 5
- The effect of adaptation is lessened for block spike codes, for which the progression in the spike symbol times τ_k is close to the natural number progression for which Propositions 3.2, 3.3 are good approximations.
- Most important of all, there seems little room to optimize the spike interval θ_1 other than physical limit to how short θ_1 can be. Varying τ_k gives different capacity rates for individual spike codes.

Parallel Cluster and Ideal Spike Coder

Unlike the spike interval θ_1, the intersymbol interval θ_0 should be made as short as possible. An ideal spike coder is one for which $\theta_0 = 0$. Here is one optimal strategy to achieve such a goal.

Definition 3.2 *Assume the input signal, I, to a spike coder takes two discrete value, 0, 1, and that the spike coder outputs spikes only when $I = 1$ and no spikes when $I = 0$. Two spike coders are* ***in-phase*** *coders if either both output spikes (not necessarily the same number) or no spikes for the same on-or-off signal I. Two spike coders are* ***out-phase*** *coders if the spiking phase and the non-spiking phase of the two coders are exactly opposite for the same on-or-off signal I. That, one spike coder outputs spikes if and only the other spike coder does not. Two out-phase spike coders are referred to as* ***complementary*** *spike coders.*

A schismatical illustration of the definition is given below.



Coder 1 and 2 are out-phase of each other, forming a complementary pair. Two complementary spike coders are in the **parallel configuration** when an input signal is duplicated to go through both simultaneously. The resulting

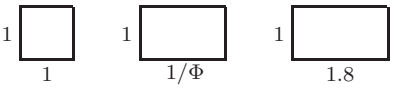
encoder or decoder from the two complementary coders in parallel is referred to as a **parallel cluster** of two out-phase spike coders. For the superposed output signal, the intersymbol interval θ_0 is zero. As a result, such a parallel cluster is an ideal spike coder.

Three important lessons of this section are:

1. Spike interval θ_0 cannot be optimized.
2. Varying spike symbol processing times τ_k gives rise to different channel capacity distributions for individual spike coders.
3. An optimal strategy to have an ideal spike coder is to cluster two out-phase spike coders in parallel.

Exercises 3.3

1. It has been a on-going debate whether or not the Golden Rectangle is aesthetically more appealing to our brain than the square or rectangle of other length ratio. Shown below are a square, a Golden Rectangle (with length ratio 1:1/Φ =1:1.6180), and a rectangle with length ratio 1:1.8.



   Some informal survey showed that on average people prefer a rectangle of length ratio around 1:1.8, about the ratio of a wide screen. It is not clear how our brain encode a rectangle. But strictly from the point view of spike codes, we know that if the ratio of the symbol processing times is $\alpha = \tau_2/\tau_1 = 2$, then the binary capacity distribution is the Golden Ratio distribution $p_2/p_1 = \Phi$. Find the ratio α so that the capacity distribution $\{p_1, p_2\}$ of a binary spike coder satisfies $p_2/p_1 = 1.8$.
2. Consider a binary spike coder with $\theta_1 = \theta_0 > 0$.
   (a) Find numerically the capacity distribution if it is not adaptive $a_1 = a_2 = 0$.
   (b) Find the exact capacity distribution if it is adaptive with $a_1 = 0, a_2 = \theta_1$.
   (c) Assume the spike coder is adaptive and $a_1 = 0$, for what value a_2 does the capacity distribution gives the 1:1.8 ratio: $p_2/p_1 = 1.8$?

3.4 COMPUTER AND MIND

(perspective.)

Chapter Four

Circuits

(... introduction here ...)

4.1 CIRCUIT BASICS

Objects may possess a property known as electric charge. By convention, an electron has one negative charge (-1) and a proton has one positive charge ($+1$). In the presence of an electric field, the electromagnetic force acts on electrically charged objects, accelerating them in the direction of the force.

We denote by Q the net positive charge at a given point of an electrical circuit. Thus $Q > 0$ means there are more positive charged particles than the negatives ones, whereas $Q < 0$ means the opposite. Through a point of (or an imaginary plane with a fixed direction that is perpendicular to) a circuit, Q may change (or flow) with time and we denote it by $Q(t)$ with t being the time variable. By definition, its rate of change (or the *flux* through the plane in the direction of the imaginary plane) is defined as the *current*

$$I = \frac{dQ}{dt}. \tag{4.1}$$

In reality though, it is usually the negative charged electrons that actually move. In such cases, the net 'positive' charge Q would move opposite to the actual direction of the free electrons. Electric charges can move in one direction at one time and the opposite direction at another. In circuit analysis, however, there is no need to keep track the actual direction in real time. Instead, we only need to assign the direction of the current through a point *arbitrarily* once and for all (or to be the same direction as the imaginary plane through the point and perpendicular to the circuit). Notice that this assignment in turn mutually and equivalently designates the *reference* direction of the movement of the net positive charge Q. In this way, $I > 0$ means the net positive charge Q flows in the reference direction while $I < 0$ means it flows in the opposite direction. Again, in most cases, free electrons actually move in the oppositive direction of I when $I > 0$ and the same direction of I when $I < 0$. To emphasize, we state the following rule.

Rule of Current Direction Assignment:

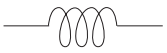
> The reference direction of the current through a point or an electrical device of a circuit is assigned arbitrarily and kept the same throughout the subsequent analysis of the circuit.

This does not mean that all assignments are equally convenient. In practice, some assignments make the analysis easier than others. Fundamentally, however, all choices are equally valid and equivalent.

Between two points in an electric field, we envision the existence of a potential difference, and the field to be the result of the gradient of the potential. More precise, the electric field is the negative gradient of the potential. The potential difference, measured in voltage, or volt, can be thought as "electric pressure" that push electric charges down the potential gradient. However, we do not need to keep track of the actual potential, but instead use the reference direction of the current to define the reference voltage. More precisely, at a given point of a circuit, the side into which the reference direction of the current points is considered to be the high end of the potential whereas the other side from which the current leaves is the low end of the potential. Hence, $I > 0$ at the point if and only if $V \geq 0$ with V denoting the voltage across the point.

The case with $V \equiv 0$ is special — it represents a point on an ideal conductor without resistance. Here below, however, we will consider various elementary cases for which the mutual dependence between V on I, which we refer to as the **IV-characteristic curve** or **IV-characteristic**, is not this trivial kind.

By **device** in this chapter, we mean an inductor, or a capacitor, or resistor, which are uniquely defined by their IV-characteristics.



Inductor
$V = L\frac{dI}{dt}$

An **inductor** is an electrical device that stores energy in a magnetic field, typically constructed as a coil of conducting materials. The voltage across an inductor with inductance L with the current through it is described by the equation

$$V = L\frac{dI}{dt}. \tag{4.2}$$

It is the IV-characteristic of inductors. In one interpretation, a positive change of the current results a voltage drop across the inductor. Alternatively, a potential difference either accelerates (when $V > 0$) or decelerates (when $V < 0$) the positive charge Q through the inductor. It is this IV-characteristic that allows us to model processes and devices as an inductor even though they are not made of coiled wires.

Capacitor
$V = \frac{1}{C}Q$

A **capacitor** is another electrical device that stores energy in the electric field. Unlike an inductor, it does so between a thin gap of two conducting plates on which equal but opposite electric charges are placed. Its capacitance (C) is a measure of the conducting material as to how much of a voltage between the plates for a given charge

$$V = \frac{Q}{C}. \tag{4.3}$$

Since the rate of change in charge is the current $I = \frac{dQ}{dt}$, we also have

$$\frac{dV}{dt} = \frac{1}{C}\frac{dQ}{dt} = \frac{I}{C},$$

which can be considered to define the IV-characteristic of the capacitor.

A word of clarification is needed for capacitors. For a given direction of the current I, the side that I goes into a capacitor is the "positive" side of the capacitor, and the voltage is fixed by the relation $V = Q/C$ with Q being the "net positive" charge. It means the following, let Q^+ be the net positive charge on the positive side of the capacitor and Q_+ be net the positive charge on the opposite side of the capacitor, then $Q = Q^+ - Q_+$. In this way, the current "flows" from the "net positive" charged side of the capacitor plate to the "net negative" charged side of the plate. In reality, however no actual charges flow across the plates which are insulated from each other. Instead, positive charges are repelled from the "net negative" charged side or equivalently negative charges are attracted to the "net negative" charged side, giving the effect that it is as if the current "flows" across the plates.

Resistor
$V = RI$ or
$I = GV$

A **resistor** is an electrical device that is used to create a simpler voltage-to-current relationship. It has two broad categories: linear and variable resistors. A linear resistor has a constant voltage to current ratio, the resistance R, by the Ohm's law

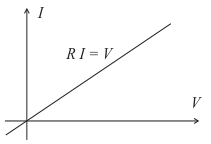
$$V = RI. \tag{4.4}$$

Variable Resistor
$g(V, I) = 0$

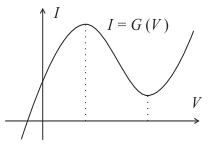
Alternatively, a resistor is also referred to as a conductor, with the conductance

$$G = \frac{1}{R} \text{ and } I = GV. \tag{4.5}$$

For a variable resistor, however, the relationship between V and I is nonlinear in general. There are three cases. For one case, V is a nonlinear function of I, $V = f(I)$, often referred to as a variable (or nonlinear) resistor. For



linear characteristics



nonlinear characteristics

another case, I is a nonlinear function of V, $I = f(V)$, referred to as a variable (or nonlinear) conductor. For the third category, neither V nor I can be expressed as a function of the other, but nonetheless constrained by an equation, $F(V, I) = 0$. The corresponding curve defined by such an equation is still called the IV-characteristic the nonlinear resistor (or nonlinear resistive or passive network as referred to in literature). It can also be thought as the level curve of $F = 0$. Notice that the first two cases can also be expressed as $F(V, I) = \pm[I - f(V)]$ or $F(V, I) = \pm[V - f(I)]$ with the exception that either I or V can be solved explicitly from $F(V, I) = 0$.

Quantity	Unit	Conversion
Elementary Charge		1 elementary charge = the charge of a single electron (-1) or proton $(+1)$
Charge Q	coulomb (C)	1 C $= 6.24 \times 10^{18}$ elementary charges
Current $I = \frac{dQ(t)}{dt}$	ampere (A)	1 A = 1 C/s (s = second)
Potential Difference V or E	volt (V)	1 V = 1 Newton-meter per C
Inductance L	henry (H)	1 H = 1 Vs/A
Capacitance C	farad (F)	1 F = 1 C/V
Resistance R	ohm Ω	1 Ω = 1 V/A
Conductance $g = 1/R$	siemens (S)	1 S = 1 A/V

A linear resistor is characterized by its resistance R, which is the constant, positive slope of the IV-characteristic curve. Heat dissipates through the resistor resulting the potential drop in voltage. The situation can be different on a nonlinear resistor. Take the nonlinear IV-characteristic curve, $I = f(V)$, shown as an example. It has two critical points that divide the V-axis into three intervals in which $I = f(V)$ is monotone. In the voltage range defined by the left and right intervals, the slope of f is positive. As result, the device behaves like a resistor as expected, dissipating energy. However, in the voltage range of the mid interval, the slope of f is negative. It does not behave like a resistor as we think it would — a faster current results in a smaller voltage drop. That is, instead of dissipating energy, energy is drown to the device, and hence the circuit that it is in. In electronics, nonlinear resistors must be constructed to be able to draw and store energy, usually from batteries.

Useful Facts about Elementary Devices:

- For each device, we only need to know one of the two quantities, the current or the voltage, because the other quantity is automatically defined by the device's IV-characteristics (4.1–4.5).
- Once the direction of a current is assigned, the corresponding voltage is automatically fixed by the device relations (4.1–4.5).
- The change of voltage and current through an inductor and capacitor are gradual, transitory in time.
- In contrast, the relationship between the voltage and current across a resistor is instantaneous, at least for an idealized resistor or conductor.
- Unit and conversion for these basic devices are given in the company table. As an example for interpretation, consider the unit of voltage. According to its unit, one can visualize that one volt is the amount of energy that is required to move one coulomb of charge by a force of one Newton for one meter.

Exercises 4.1

1.
2.

4.2 KIRCHHOFF'S LAWS

In a circuit, the voltage and current across a given device change with time. As functions of time, they satisfy a set of differential equations, which in turn are considered as a mathematical model of the circuit. All equations are based on two laws of conservation, one for the conservation of energy — *Kirchhoff's voltage law*, and the other for the conservation of mass — *Kirchhoff's current law*.

Kirchhoff's Voltage Law

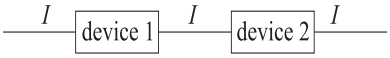
> The directed sum of the electrical potential differences around a circuit loop is zero.

Kirchhoff's Current Law

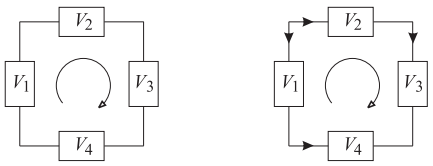
> At any point of an electrical circuit where charge density is not changing in time, the sum of currents flowing towards the point is equal to the sum of currents flowing away from the point.

Remark.

1. Two devices are connected in **series** if the wire connecting them does not branch out in between and connect a third device. As one immediate corollary to Kirchhoff's Current Law, the current through all devices in a series is the same. This negates the need to assign a current to each device in a series.



2. Although individual devices in a circuit loop can be assigned arbitrary current directions, and their voltages are thus fixed thereafter by the formulas (4.1–4.5), the "directed sum" stipulation in Kirchhoff's Voltage Law *does not* mean to simply add up these arbitrarily defined voltages. It means the following. We first pick an orientation for the loop, for example, the counterclockwise direction. We then either add or subtract the voltage V of a device in the "directed sum". It is to add if V's current direction is the same as the orientation of the loop. It is to subtract if the current direction is against the orientation of the loop. Two examples are illustrated below. For the left loop, the currents through the V_i devices with voltage V_i are implicitly taken to be the same counterclockwise orientation of the loop. In contrast, each device from the right loop is given an individual reference direction of its current.

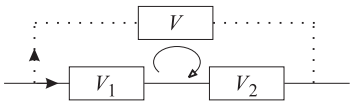


Voltage Law: $V_1 + V_2 + V_3 + V_4 = 0$ $-V_1 + V_2 + V_3 - V_4 = 0$

3. **Voltage Additivity in Series.** As an immediate corollary to the Voltage Law, the voltages over devises in series are additive. That is, if $V_1, V_2, \ldots, V_n$ are voltage drops across devices $1, 2 \ldots, n$, then the voltage drop V over all devices is

$$V = V_1 + V_2 + \cdots + V_n.$$

The diagram below gives an illustration of the required argument.



Voltage Law: $-V_1 - V_2 + V = 0$ or $V = V_1 + V_2$

We wire a resistor with infinite resistance $R = \infty$ in parallel with the serial devices as shown. The new wire draws zero current from the existing circuit. By Kirchhoff's Voltage Law (and the previous remark), we see that the voltage across the infinite resistor is the numerical sum of the existing serial voltages.

Because of the second last remark, we can fix the direction of a device in a circuit as follows for convenience and definitiveness.

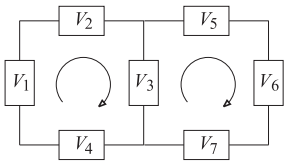
Current Direction Convention:

Unless specified otherwise, the following convention is adopted:

- If a circuit consists of only one loop, then all device currents are counterclockwise oriented.
- If a circuit consists of a network of parallel loops, all device currents in the left most loop are counterclockwise oriented. Device currents of subsequent addition to the existing loop that forms the new loop are counterclockwise oriented as well.

Remark. Useful tips to note: the current of a device that is in two adjacent loops is counterclockwise oriented with respect to the left loop but clockwise oriented with respect to the right loop. Its voltage takes different signs in the directed sums of the loops.

Example 4.2.1 According to the Current Direction Convention, the circuit below has three equations by Kirchhoff's Laws as shown.



Voltage Law:

left loop: $V_1 + V_2 + V_3 + V_4 = 0$

right loop: $-V_3 + V_5 + V_6 + V_7 = 0$

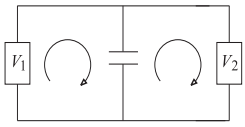
Current Law:

top and bottom nodes: $I_2 = I_3 + I_5$

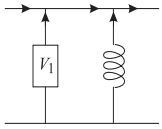
Notice that both the top and bottom node points give rise to the same equation of currents by the Current Law because $I_2 = I_1 = I_4$ and $I_5 = I_6 = I_7$.

Exercises 4.2

1. In each of the circuits below, label appropriate currents, use Kirchhoff's Current Law and Voltage Law to write the equations governing the voltages and the currents through the devices.

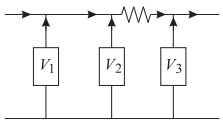


(a)

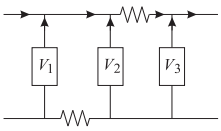


(b)

2. In each of the circuits below, label appropriate currents, use Kirchhoff's Current Law and Voltage Law to write the equations governing the voltages and the currents through the devices.



(a)



(b)

4.3 CIRCUIT MODELS

Circuit analysis is to study how devices currents and voltages change in a circuit. The main reason why a circuit can be completely understood is because of the following principles.

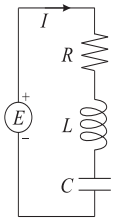
- Each device has only one independent variable, either the current or the voltage, and the other is determined by the device relations (4.1–4.5).
- If a circuit has only one loop, then all the devices are in series, and we only need to know the current variable. Since the circuit loop gives rise to one equation from the Voltage Law, we should solve one unknown from one equation.

- If a circuit has two parallel loops, such as the circuit diagram of Example 4.2.1, then there are three serial branches—left, middle, and right branches—each is determined by its current variable, or one device voltage variable in general. However, the node points give rise to only one independent equation from the Current Law, and each loop gives rise to one equation from the Voltage Law. Together we have three equations for three variables, and it is solvable in principle.
- The same argument above applies to multiple parallel loops.

To illustrate these principles, we consider the following examples for which only the equations as mathematical models of circuits are derived. The equations inevitably contain derivatives of variables. Such equations are not *algebraic* rather than *differential* equations. Methods to solve the differential equations are considered in later sections.

Linear Circuits

Example 4.3.1 *RLC* **Circuit in Series.** Consider the *RLC* circuit as shown. We only need to know the current variable I. From the Voltage Law we have



$$V_R + V_L + V_C - E = 0,$$

with the subscript of the voltage V_i corresponding to the devices (R for the resistor, L for the inductor, and C for the capacitor). Since batteries have a designated positive terminal, its corresponding current is fixed as well. Thus, in this case it is opposite the reference current direction chosen, and hence the negative sign of E in the directed sum equation. In principle, this equation together with the device relations (4.1–4.5),

$$V_R = RI,\ V_L = L\frac{dI}{dt},\ V_C = \frac{1}{C}Q,\ \text{ and } I = \frac{dQ}{dt}$$

completely determines the circuit.

However, circuit analysis does not stop here. We usually proceed to reduce the number of equations as much as possible before we solve them by appropriate methods for differential equations.

To illustrate the point, we express all voltages in terms of the current I by the device relations and substitute them into the voltage equation. More specifically, we first have

$$L\frac{dI}{dt} + RI + \frac{1}{C}Q = E. \tag{4.6}$$

Differentiating the equation above in t yields,

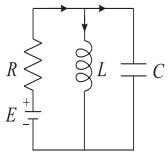
$$L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}I = \frac{dE}{dt}. \tag{4.7}$$

Alternatively, (4.7) can be directly expressed in term of Q as

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E. \tag{4.8}$$

Either of the last two equations models the circuit mathematically. They are now ready to be solved by appropriate methods for differential equations which is a different issue from the model building completed above. ◎

We note that the equation above is a **second order**, **linear**, **ordinary differential equation**(ODE). The term "order" refers to the highest derivative taken in the equation. The term "linear" refers to the fact that the expression on the left side is is a linear combination of the unknown I in derivatives: $d^2I/dt^2, dI/dt, I$. The term "ordinary" is given to equations for which the unknown is a function of only one variable. If an equation has an unknown which is a function of more than one variables and contains partial derivatives of the unknown, then it is called **partial differential equation**. We will discuss numerical methods to solve differential equations in later sections.



Example 4.3.2 *RLC* **Circuit in Parallel.** Consider the RLC circuit in parallel as shown. By the Current Law, we have

$$I_R = I_L + I_C,$$

for both points. By the Voltage Law, we have

$$V_R + V_L - E = 0 \text{ and } V_C - V_L = 0$$

for the left and right loop respectively.

Like the previous example, we can derive a second order, linear ODE of the inductor current I_L for the circuit, which is left as an exercise. Here we derive alternatively a system of two equations of two variables, V_C and I_L.

From the capacitor relation $V_C = Q/C$, the current definition $dQ/dt = I_C$, and the node point current equation $I_C = I_R - I_L$, we have

$$\frac{dV_C}{dt} = \frac{1}{C}I_C = \frac{1}{C}(I_R - I_L). \tag{4.9}$$

Replacing I_R by the resistor relation $I_R = V_R/R$ and $V_R = E - V_L = E - V_C$ from both loop voltage equations gives rise to the first equation

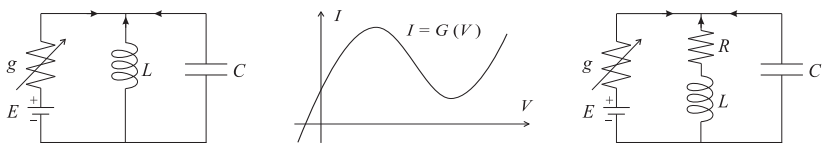
$$\frac{dV_C}{dt} = \frac{1}{C}\left(\frac{1}{R}(E - V_C) - I_L\right).$$

From $V_C = V_L$ of the right loop and the inductor relation $LdI_L/dt = V_L$, we have the second equation

$$L\frac{dI_L}{dt} = V_C.$$

Together, we have obtained a system model of the circuit

$$\begin{cases} C\dfrac{dV_C}{dt} = \dfrac{1}{R}(E - V_C) - I_L \\ L\dfrac{dI_L}{dt} = V_C. \end{cases} \tag{4.10}$$



(a) (b) (c)

Figure 4.1 (a) van der Pol circuit. (b) IV-characteristics. (c) FitzHugh-Nagumo circuit.

The main difference between Eqs.(4.7, 4.8) and Eq.(4.10) is that the former are second order equations whereas the latter is a system of two first order equations. Geometric and numerical methods work the best for systems of first order equations. For this reason, all higher order ODEs are converted to systems of first order ODEs before geometric and numerical methods are applied.

Nonlinear Circuits

Circuits above contain only linear resistors. We now consider circuits with nonlinear resistors (or conductors).

Neuron Circuit Current Convention.

> Unless specified otherwise, all reference directions of electrical currents through the cell membrane of a neuron are from inside to outside. All circuit diagrams for neurons are placed in such a way that the bottom end of the circuit corresponds to the interior of the cells and the top end to the exterior of the cell.

Example 4.3.3 van der Pol Circuit. The circuit is the same as the RLC parallel circuit except for a nonlinear resistor, both are shown in Fig.4.1. For the nonlinear conductance g, we use the following polynomial form for the nonlinear resistive characteristics

$$I = F(V) = V(aV^2 + bV + c)) + d, \tag{4.11}$$

where $a > 0, b, c, d$ are parameters.

Applying Kirchhoff's Voltage and Current Laws, we have

$$V_R - E - V_L = 0,\ V_L - V_C = 0,\ I_C + I_R + I_L = 0. \tag{4.12}$$

From the capacitor and current relation $V_C = Q/C, dQ/dt = I_C$, and the equations above, we have

$$\frac{dV_C}{dt} = \frac{1}{C} I_C = \frac{1}{C}(-I_R - I_L). \tag{4.13}$$

Replacing I_R by the nonlinear resistor relation $I_R = F(V_R)$ and R_R by $V_R = E + V_L = E + V_C$ from Eq.(4.12) gives rise to the first equation

$$C\frac{dV_C}{dt} = -F(E + V_C) - I_L.$$

From $V_C = V_L$ of Eq.(4.12) and the inductor relation $LdI_L/dt = V_L$, we have the second equation

$$L\frac{dI_L}{dt} = V_C.$$

Finally, we have the model system of equations

$$\begin{cases} C\dfrac{dV_C}{dt} = -F(E + V_C) - I_L \\ L\dfrac{dI_L}{dt} = V_C. \end{cases} \tag{4.14}$$

◎

Example 4.3.4 FitzHugh-Nagumo Circuit. The circuit is the same as the van der Pol circuit except that a linear resistor is in series with the inductor as shown in Fig.4.1. The nonlinear conductance g has the same form as van der Pol's circuit.
It is left as an exercise to derive the following system of equations for the circuit, referred to as **FitzHugh-Nagumo equations**.

$$\begin{cases} C\dfrac{dV_C}{dt} = -F(E + V_C) - I_L \\ L\dfrac{dI_L}{dt} = V_C - RI_L. \end{cases} \tag{4.15}$$

◎

Notice that the only difference from the van der Pol equations is on the right hand side of the second equation. As result, the van der Pol equations is a special case of the FitzHugh-Nagumo equations with $R = 0$. For this reason, the latter reference will be used often. Another reason that will become apparent later is because FitzHugh-Nagumo equations are better suited for a class of neuron circuits.

Forced Circuits

The circuits considered so far are "closed" networks of devices. In applications, they may be a part of a larger circuit, having their own incoming and outgoing ports through which connections to the larger circuit are made. The current going into the compartment is thus considered as an external forcing. The circuit including the two ports is therefore considered as a forced circuit of the closed one without the interphase ports.

Example 4.3.5 Forced FitzHugh-Nagumo Circuit. The circuit is the same as the FitzHugh-Nagumo circuit except for the additional I_{in} and I_{out} branches as shown in Fig.4.2. It is left as an exercise to show that

$$I_{in} = -I_{out} = -(I_g + I_R + I_C).$$

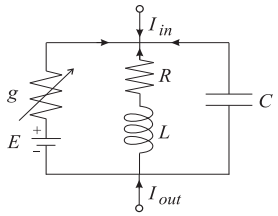


Figure 4.2 Forced FitzHugh-Nagumo circuit.

It is also left as an exercise to derive the **forced FitzHugh-Nagumo equations** for the circuit,

$$\begin{cases} C\dfrac{dV_C}{dt} = -F(E+V_C) - I_L - I_{in} \\ L\dfrac{dI_L}{dt} = V_C - RI_L. \end{cases} \tag{4.16}$$

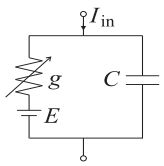
We will also interchangeably refer to it as FitzHugh-Nagumo equations because the closed circuit version corresponds to the case when $I_{in} = 0$.

Exercises 4.3

1. Show that the equation for the RLC circuit in parallel from Example 4.3.2 can also be written as the following second order, linear, ODE,

$$RCL\frac{d^2I_L}{dt^2} + L\frac{dI_L}{dt} + RI_L = E.$$

2. Use the current relation $dQ/dt = I$ to transform the equation Eq.(4.6) for the RLC serial circuit into a system of two first order ODEs.
3. Derive the FitzHugh-Nagumo equations (4.15).
4. Show that for the forced FitzHugh-Nagumo circuit from Fig.4.2, $I_{in} = -I_{out}$.
5. Derive an ODE for the RC circuit shown. The nonlinear conductance g is given by the IV-characteristics $I = F(V)$.



Problem 5

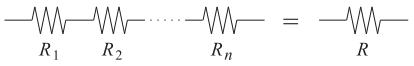
6. Derive the forced FitzHugh-Hagumo equations (4.16).
7. Show the forced FitzHugh-Nagumo equations (4.16) can be transformed into the following second order, nonlinear ODE,

$$L\frac{d^2I_L}{dt^2} + R\frac{dI_L}{dt} + \frac{1}{C}I_L + \frac{1}{C}F(E + L\frac{dI_L}{dt} + RI_L) = -\frac{1}{C}I_{in}.$$

4.4 EQUIVALENT CIRCUITS

A circuit is a network of devices. If we are only interested in a current, I_{in}, that goes into the circuit at one point, called the **input node**, another current I_{out}, that come out at another point, called the **output node**, and the voltage drop between these points, $V_{in,out}$, then whatever happen in between is not important. Loosely speaking, two circuits are **equivalent** if they are interchangeable between the input and output node points without altering the quantities $I_{in}, I_{out}, V_{in,out}$.

Example 4.4.1 A set of n linear resistors in series is equivalent to a resistor whose resistance is equal to the sum of individual resistors resistance $R_k, k = 1, 2, \ldots, n$.



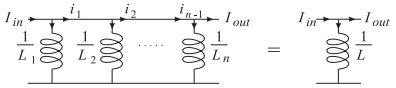
A verification follows from the property of Voltage Additivity in Series from the remarks after Kirchhoff' Voltage Law in Sec.4.2 that the voltage drop across devices in series is the sum of individual devices voltage. In particular, the input and output currents I_{in}, I_{out} are the same, I. Since the voltage drop across all resistors is

$$V = V_1 + V_2 + \cdots + V_n = R_1 I + R_2 I + \cdots R_n I = (R_1 + R_2 + \cdots + R_n)I,$$

it is equivalent to one resistor of the resistance

$$R = R_1 + R_2 + \cdots + R_n.$$

Example 4.4.2 A set of n linear inductors in parallel is equivalent to an inductor whose resistance in reciprocal is equal to the sum of individual inductors inductance in reciprocal: $\frac{1}{L} = \frac{1}{L_1} + \frac{1}{L_2} + \cdots + \frac{1}{L_n}$.



By the Current Direction Convention and Kirchhoff's Voltage Law, we have

$$-V_1 = V_2 = V_3 = \cdots = V_n = V$$

for the equivalence conditions for voltages, with the first equation from the left most loop, the second equation from the second left loop, etc. The last equation is the condition for an equivalent inductor. Let $i_k, k = 1, 2, \ldots, n-1$ be the currents running on the top edges of the $n-1$ loops. Then, by Kirchhoff's Current Law, we have

$$I_{in} + I_1 = i_1,\ i_k = I_{k+1} + i_{k+1},\ i_{n-1} = I_n + I_{out},$$

with the first equation from the top left node, the $(k+1)$th node from left, and the nth or the last node. Use the formulas recursively, we have

$$I_{in} = -I_1 + i_1 = -I_1 + I_2 + i_2 = \cdots = -I_1 + I_2 + \cdots + I_n + I_{out}.$$

Since for the equivalent circuit we have

$$I_{in} = I + I_{out},$$

we must have

$$I = -I_1 + I_2 + \cdots + I_n$$

as the equivalence condition for current. Differentiating this equation in time, and using the device relations $dI_k/dt = V_k/L_k$, we have

$$\frac{1}{L}V = -\frac{1}{L_1}V_1 + \frac{1}{L_2}V_2 + \cdots + \frac{1}{L_n}V_n.$$

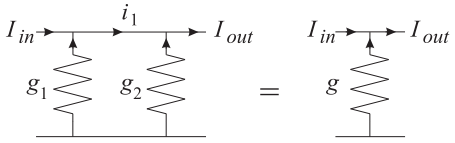
Since $V = V_k$ for $k \geq 2$ and $V = -V_1$ from the equivalence conditions for voltages, we have

$$\frac{1}{L}V = \left[\frac{1}{L_1} + \frac{1}{L_2} + \cdots + \frac{1}{L_n}\right] V.$$

Canceling V from both sides yields the equivalence condition for the inductance.

Equivalent Nonlinear Conductors. For linear devices, the equivalent characteristics can be expressed in simple algebraic relations. For nonlinear devices, it is usually not as simple. However, the principle reason why equivalence exists at all is the same. It all depends on Kirchhoff's Voltage and Current Laws.

Consider two nonlinear resistors in parallel as shown with IV-characteristics $g_1(V, I) = 0$ and $g_2(V, I) = 0$ respectively.



By Kirchhoff's Voltage Law, $V_1 = V_2$. By the voltage equivalence condition,

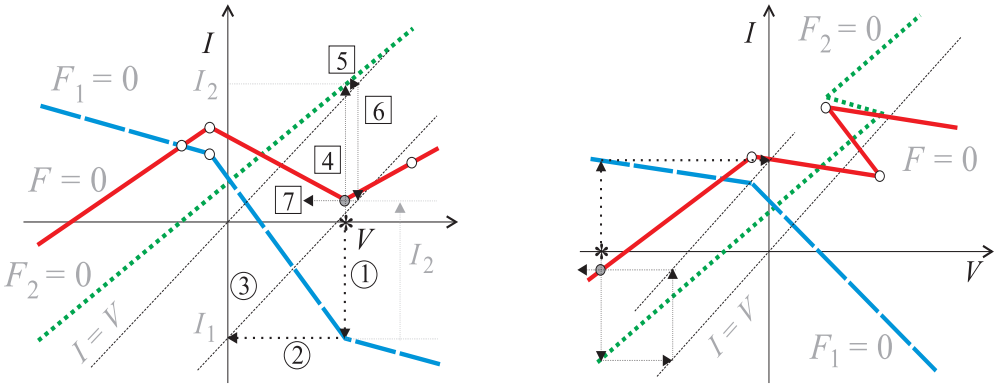
$$V_1 = V_2 = V. \tag{4.17}$$

By Kirchhoff's Current Law, $I_{in} + I_1 = i_0$ and $i_0 + I_2 = I_{out}$. Eliminating i_0, $I_{in} + I_1 + I_2 = I_{out}$. By the current equivalence condition, $I_{in} + I = I_{out}$. Together, they imply

$$I = I_1 + I_2. \tag{4.18}$$

Equations.(4.17, 4.18) mean the following. For a given voltage V, if I_1 and I_2 are corresponding currents on the resistors in parallel, $g_k(V, I_k) = 0, k = 1, 2$, then the current I on the equivalent nonlinear resistor, $g(V, I) = 0$, must be the sum of I_1 and I_2. Geometrically, all we have to do is for every pair of points $(V, I_1), (V, I_2)$ from the constituent IV-curves, plot the point $(V, I_1 + I_2)$ on the same vertical line $V = V$ in the plane to get the equivalent IV-characteristic curve.

There are two ways to do this.



Example 4.4.3 Example 4.4.4

Figure 4.3 Graphical Method for equivalent resistors of resistors in parallel.

Methods of Deriving Equivalent IV-Characteristics in Parallel

Graphical Method. This method will produce the equivalent IV-characteristics for two resistors in series or in parallel as long as the constituent IV-curves are given. Illustration is given for the parallel equivalent resistor case, and the serial case is left as an exercise.

Example 4.4.3 Consider two nonlinear resistors in parallel with their nonlinear IV-characteristics $F_1(V, I) = 0, F_2(V, I) = 0$ as shown. For clarity, piecewise linear curves are used for both curves. The problem is to place the point $(V, I_1 + I_2)$ graphically on each vertical line $V = V$ which intersects the constituent IV-curves, i.e. $F_1(V, I_1) = 0, F_2(V, I_2) = 0$.

Here are the steps to follow.

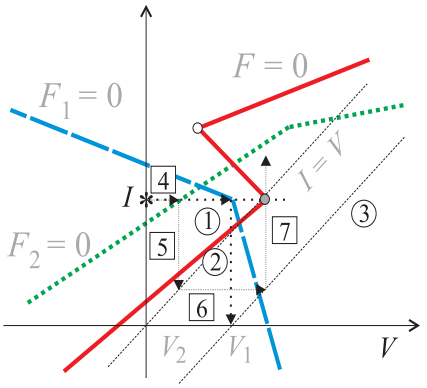
- *Step 1.* Plot both constituent IV-curves in one IV-plane. Draw the diagonal line $I = V$ in the plane (the line through the origin with slope $m = 1$).
- *Step 2.* Locate all **critical points** of the constituent IV-curves. A point is critical if it is not a smooth point of the curve, or it has a horizontal tangent line, or a vertical tangent line. Any point at which the curve changes its monotonicity in either I or V direction is a critical point.
- *Step 3.* To fix notation, let $(V, I_1), (V, I_2)$ be a pair of points from the F_1 and F_2 characteristic curves respectively, i.e., $F_k(V, I_k) = 0$, so that one of the two is a critical point and both are on the same vertical line $V = V$. The method now proceeds in sub-steps that are numbered as illustrated. For example, start at the critical value V labelled as $*$. Choose one of the two constituent IV-curves we want to start. Our illustration takes F_1 as an example. Draw a vertical line from $*$ to the graph F_1. If there are multiple intersections, do one intersection at a time. In sub-step number 2, draw a horizontal line from the intersection to the I-axis. From the intersection, draw a line parallel to the diagonal

to complete sub-step 3. Sub-step 4 starts from the critical point $*$ again, and do the same for the other IV-curve, $F_2 = 0$ for the illustration. Sub-step 7 draws a horizontal line to intersect the vertical line $V = V$. The intersection point is $(V, I_1 + I_2)$.

- *Step 5.* Repeat the same procedures for all critical points, and two far end "anchoring" points chosen arbitrarily.
- *Step 6.* Since the sum of two lines is another line, (here is the algebra: $(b_1 + m_1x) + (b_2 + m_2x) = (b_1 + b_2) + (m_1 + m_2)x$), we now connect the "dots" which are produced branch by branch. The resulting curve is the *exact* equivalent IV-characteristics. If the constituent IV-curves are not lines, we only need to "smooth" the dots to get an *approximate* of the equivalent curve, which is often sufficient for qualitative purposes.

Example 4.4.4 Figure 4.3 shows another example of the equivalent IV-characteristics for two nonlinear resistors in parallel. Again, the equivalent curve is constructed by the graphical method. A noncritical point is used to illustrate the steps which are not numbered.

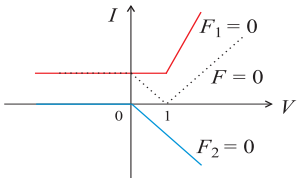
Example 4.4.5 The graphical method can be extended to nonlinear resistors in series. Let the IV-characteristics of nonlinear resistors be $F_1(V, I) = 0, F_2(V, I) = 0$ respectively. Then it is similar to argue by Kirchhoff's Current Law that for each fixed current I, if V_1, V_2 are the voltages on the two constituent characteristics, i.e., $F_k(V_k, I) = 0, k = 1, 2$, then the equivalent resistor's characteristics must be $V = V_1 + V_2$ for the same current level I. To locate the equivalent voltage $V = V_1 + V_2$, we only need to interchange vertical and horizontal constructing lines in the graphical method illustrated above for nonlinear resistors in parallel. The figure gives an illustration of the graphical method.



Analytical Method.

- If each I_k can be written as a function of V, $I_k = f_k(V)$, then the equivalent characteristics is simply
$$I = f_1(V) + f_2(V) = f(V).$$
- Suppose I_1 can be written piece by piece as a set of n functions of V, $I_1 = f_k(V), 1 \le k \le n$, and I_2 can be written piece by piece as a set of m functions of V, $I_2 = g_j(V), 1 \le j \le m$. Then the equivalent characteristics can be pieced together by no more than mn functions of V, $I = f_k(V) + g_j(V)$ for $k = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. The single functions case above is just a special case.

Example 4.4.6 Consider two nonlinear resistors in parallel whose IV-characteristics g_1, g_2 are as shown. For g_1, I_1 can be written piecewise as



$$I_1 = g_1(V) = \begin{cases} 1, & \text{for } V \le 1 \\ 2V - 1, & \text{for } V \ge 1. \end{cases} \tag{4.19}$$

For g_2, I_2 can be written piecewise as

$$I_2 = g_2(V) = \begin{cases} 0, & \text{for } V \le 0 \\ -V, & \text{for } V \ge 0. \end{cases} \tag{4.20}$$

The equivalent characteristic g can be put together with no more than $2 \times 2 = 4$ pieces. In fact, in the interval $(-\infty, 0]$, the piece of f_2 overlaps the left component of f_1, eliminating one unnecessary combination with the right component. The equivalent characteristics in that interval is $I = I_1 + I_2 = 1+0 = 1$. For the remaining interval $[1, \infty)$, the right piece of f_2 overlaps with both pieces of f_1 in range. Hence we break it up into two pieces, one over the interval $[0, 1]$ and the other $[1, \infty)$. In the mid interval, $I = I_1 + I_2 = 1 - V$, and in the right interval, $I = I_1 + I_2 = 2V - 1 - V = V - 1$. To summarize, we have

$$I = g(V) = \begin{cases} 1, & \text{for } V \le 0 \\ 1 - V, & \text{for } 0 < V \le 1 \\ V - 1, & \text{for } 1 \le V. \end{cases} \tag{4.21}$$

The equivalent IV-curve consists of the dotted lines. ◎

We note that this method can be implemented computationally.

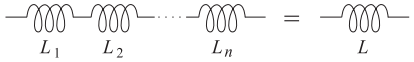
In general, there is no method to write down the equivalent IV-characteristic, $F(V, I) = 0$, for a parallel or serial network of resistors whose IV-characteristics are implicitly defined. However, this can be done for some special cases. For example, if one resistor's IV-curve is given implicitly as $F_1(V, I) = 0$ and another resistor's IV-curve is given explicitly with I as a function of V: $I = f_2(V)$. Then, the equivalent IV-characteristic $F(V, I) = 0$ for the two resistors in parallel is

$$F(V, I) = F_1(V, I - f_2(V)) = 0,$$

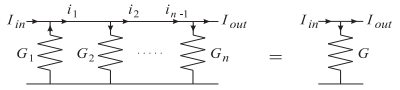
where $I = I_1 + I_2$ with $I_2 = f_2(V)$ and $F_1(V, I_1) = 0$. This is left as an exercise to the reader.

Exercises 4.4

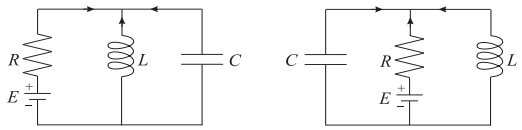
1. Show that a set of n inductors in series with inductance $L_k, k = 1, 2, \ldots, n$ is equivalent to an inductor whose inductance is the sum of individual inductors inductance.



2. Show that a set of n resistors in parallel with conductance $G_k = 1/R_k, k = 1, 2, \ldots, n$ is equivalent to resistor whose conductance is the sum of individual resistors conductance.

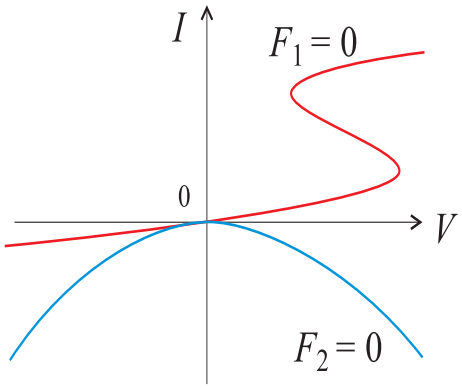


3. Show that the following circuits are equivalent in the sense that both have the set of equations. That is, the order by which circuit loops are laid out is not important. (When electrons moves in a circuit, it follows physical laws of energy conservations, not arbitrary order of circuit paths we lay out on a planar diagram.)

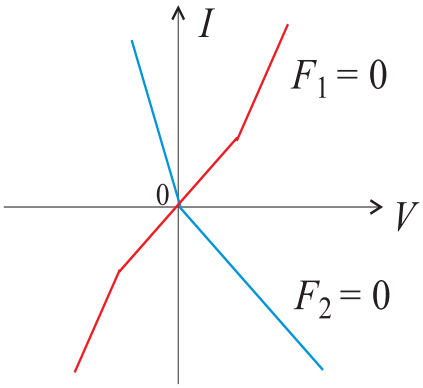


For the following problems, label the sub-steps in constructing one critical point of your choice.

4. Use the graphical method to sketch the equivalent IV-characteristics of Example 4.4.6.
5. Use the graphical method to sketch the equivalent IV-characteristic curve of two resistors in parallel whose IV-characteristic curves are as shown.
6. Use the graphical method for serial nonlinear resistors to sketch the equivalent IV-characteristic curve of two resistors in series whose IV-characteristic curves are as shown.



Problem 5



Problem 6

7. For problem 5 above, if $F_1(V, I) = V - I(I^2 + 4I + 5)$ and $F_2(V, I) = V^2 + I$, find the function form $F(V, I)$ so that $F(V, I) = 0$ defines the IV-characteristic of the equivalent resistor with F_1, F_2 in parallel.