# Entropy for Smart Kids and their Curious Parents

## Arieh Ben-Naim

# Entropy for Smart Kids and their Curious Parents

# Entropy for Smart Kids and their Curious Parents

By

Arieh Ben-Naim

Cambridge
Scholars
Publishing

Entropy for Smart Kids and their Curious Parents

By Arieh Ben-Naim

This book first published 2019

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-2710-7
ISBN (13): 978-1-5275-2710-2

This book is dedicated to all
the kids, smart or not, in the world

# TABLE OF CONTENTS

Shannon's Measure of Information (SMI)

Entropy and the Second Law of Thermodynamics

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ABN | Arieh Ben Naim |
| 1D | One dimensional |
| 3D | Three dimensions |
| Pr | Probability |
| SMI | Shannon's Measure of Information |
| $2^{nd}$ Law | The Second Law of Thermodynamics |

# PREFACE

Entropy is one of the most interesting concepts in physics. Although it is a well-defined concept, it is still perceived by even well-known scientists as a concept cloaked in mystery. It is also the most misused and often abused concept in physics. Some scientists believe that entropy will forever remain a mysterious quantity, and for them demystifying entropy remains as elusive as ever.

This book's title "Entropy for Smart Kids" might be misleading. What it actually means is that even "Smart Kids" can understand the concept of entropy. The prerequisites for understanding entropy are:

1. You need to have a probability-sense (I will show you in Chapter 1 that you already have).

2. You need to have an information-sense (I will explain in Chapter 2, and show you that you already have).

3. You need common-sense (which I hope you have). You will need that in order to understand Chapter 3.

My aim in writing this book is to show you that if you have even a rudimentary sense of probability and of information, and if you are willing to use your common-sense, then you can understand what entropy is.

It is my conviction that in order to understand *entropy*, one needs to understand Shannon's measure of information (SMI), and in order to understand SMI one must be familiar with some basic concepts of probability.

Therefore, this book consists of three chapters. Chapter 1 discusses *probability*. You will find out in this chapter that you already know what probability is. Once you know what probability is, you can also understand what SMI means. This is discussed in Chapter 2. This knowledge will lead you to a straightforward understanding of entropy. You will see in Chapter 3 that entropy is nothing but a special case of SMI. Please memorize this acronym. It will appear many times in this book. A simple way of memorizing the meaning of this term is to think of a twenty-question (20Q) game. In this game there is always a minimal number of binary questions one needs to ask in order to find out one out of *N* possibilities.

In Chapter 3 we also briefly discuss the Second Law of Thermodynamics (2nd Law). We shall see that the 2nd Law is nothing but a law of probability. We shall also see under what conditions the 2nd Law is related to the concept of entropy. We shall conclude this book by mentioning a few misuses and misapplications of entropy and the Second Law. You will learn how, and why entropy, and the 2nd Law became so mysterious as a result of these very same misuses and misapplications, as well as the gross-exaggeration of the "power" of entropy.

You do not need to know any mathematics in order to understand this book. It is helpful to know though what a logarithm is, but in case you have no clue as to what it is, you can still understand both entropy and the Second Law. In Appendix A you will find a simple, qualitative discussion of logarithm. If you do not know what the symbol $\log_2\left(\frac{1}{2}\right)$ means you can simply look at the relevant graph in Appendix A.

I urge you to read this book slowly, carefully, and critically. You also need to do some exercises in order to test your comprehension. Once you do all these, I guarantee that upon reaching the end of the book, you would know what entropy is, you will understand the 2nd Law, and you will also understand why entropy has become such a mysterious concept in physics.

Furthermore, I promise with confidence, that once you read this book your understanding of entropy will surpass the understanding of entropy by many authors who write about entropy.

Arieh Ben-Naim

Department of Physical Chemistry
The Hebrew University of Jerusalem
Jerusalem, Israel
Email: ariehbook@gmail.com
URL: ariehbennaim.com

# ACKNOWLEDGMENTS

I am grateful to Ron Aharoni, Diego Casadei, Robert Engel, Robert Goldberg, Paul King, Jose Angel Sordo Gonzalo, Douglas Hemmick, Shannon Hunter, Richard Henchman, Maik Jacob, Mike Rainbolt and Erik Szabo for reading parts or all of the manuscript and offering useful comments. Special thanks are due to Alex Vaisman for some of the drawing he made for this book.

As always, I am very grateful for the graceful help I got from my wife, Ruby, and for her unwavering involvement in every stage of the writing, typing, editing, re-editing, and polishing of the book.

The cover designed is based on a picture taken by the author at the Villa Escudero Plantations and Resort located in Tiaong, Quezon Philippines, with permission from one of the resort owners, Mr. Aaron Escudero.

# CHAPTER 1

# PROBABILITY AND PROBABILITY DISTRIBUTIONS

*In this chapter we discuss the concept of probability. We will start by playing some very simple games. These games were designed in such a way that playing them will prove to you that you already know what probability means. To put it another way, you will be convinced that you already possess a probability-sense; you have an intuitive understanding of the concept of probability although you do not know how to define it.*

*Next, we shall briefly learn how the concept of probability evolved from gambling games into a well-respected branch of mathematics. This will lead us to attempt to define the concept of probability. As we shall see, all definitions of probability are circular, i.e. they use the concept of probability (or an equivalent one, like "chances," or "likelihood") in order to define probability.*

***Next, we will further train ourselves with some simple probability problems which will be important to understanding the concept of Shannon's measure of information discussed in Chapter 2, and the concept of entropy discussed in Chapter 3.***

***Finally, we will discuss a few probability distributions, and learn how to calculate average quantities. We shall use the concept of average to create a measure of an average uncertainty about the outcomes of an experiment. This average will have a similar meaning to Shannon's measure of information. It will also be indispensable for understanding entropy.***

## 1.1 Your probability-sense

In this section, we shall play a few simple games. These games were so designed in such a way that while playing, you will either consciously or subconsciously be using probabilistic reasoning even before knowing what it is, or how it is defined. We shall go back to discussing the "definitions" of probability later on.

Let us start with a very simple game.

Figure 1.1(see color centerfold) shows seven different dice. Each die has six faces. I tell you that the dice are "fair." You might ask "what does a fair die mean?" For the moment I will tell you that all the dice we will play with are perfect cubes, i.e. all faces have the same area, and all the edges have the same length. I also tell you that the density of the material of which the dice are made of (plastic, metal, or any other material) is the same at each point within the dice which means that there is no unsymmetrical mass distribution within the dice. If I throw the dice into the air, there is no *preferred* face on which the dice will land.

Can you explain why such a die is called a fair die?[1]

All the dice in Figure 1.1 are "fair, but instead of a regular die with different numbers of dots (as in Figure 1.2a (see color centerfold) on its face, we have different colors. The dice in Figure 1.1 are colored as follows:

die *a* has six blue faces

die *b* has one red, and five blue faces

die *c* has two red, and four blue faces die *d* has three red, and three blue faces

die *e* has four red, and two blue faces

die *f* has five red and one blue face

die *g* has six red faces

Altogether, we have seven different dice. All of these dice are *fair*, which means that any one of the six faces of any die has the same likelihood of appearing when I throw it.

Here are the rules of the game. Read them carefully before you accept, or refuse to play this game.

I choose a die from the seven dice in Figure 1.1. I tell you that it is fair, and that I will throw it high into the air in such a way that it will roll over several times before it falls on the ground, Figure 1.3. I ask you to please trust me, at least for this particular game – that I have no control on the outcome of this throw. In other words, I cannot affect, nor do I know on which face the die will land on the ground. If you do not trust me (why should you?), then just imagine that a machine or a robot will be throwing the die.

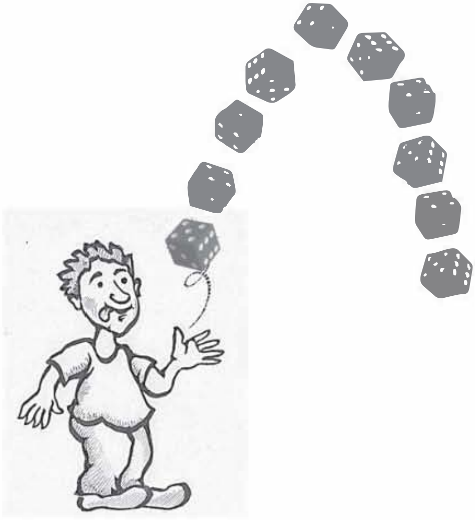


Figure 1.3. A die whirls in the air

You look at the die, count the number of blue and red faces, and see that the die is a perfect cube having no discernible defects, cuts, or irregularities.

Now, I offer you to play the following game:

I will choose one of the dice in Figure 1.1. You examine it, and choose either a blue or a red die.

Then I throw the die 100 times. Whenever the color you chose appears on the upper face of the die, you will get $1.00. If the color of the upper face is not the one you chose, you pay $1.00.

If you understand the rules of the game, repeat them before we continue playing this game.

***First game:***

I choose the die ***a***, from Figure 1.1. Which color will you choose?

Before you decide on the color, refresh your memory about the rules of the game (I chose the die, you chose the color after examining the die. I toss the die 100 times, and every time the outcome is the color you chose, you get $1.00, otherwise you pay $1.00). Are you ready?

Obviously, presuming that you understood the rules of the game, you will choose the color blue. (One important aspect which I would like to tell you is that I implicitly presume that in playing any of these games, your goal is to maximize your earnings).

Clearly, in case of die ***a*** you will choose the color *blue*, as this choice will ensure you that on each toss you will get $1.00. Altogether, you shall have earned $100.00 after I toss the die 100 times.

What an easy game! You do not have to think hard in order to choose the color of the die. You also do not have to use your *probability-sense.* You only need plain common-sense.

Can you use the word "probability" to explain why you chose the color blue?[2]

***Second game:***

I choose die *b*. You examine it, refresh your memory with the rules of the game, and choose a color. If you forgot the rules, read them again before you proceed to make a choice.

Which color will you choose?

I am sure you will choose blue again. Why am I so sure? In the previous game, I was sure you'd choose blue because in case ***a***, because you were certain to win $1.00 on each toss. Now, I am also *sure* that you will *choose* blue. But, I am *not sure* that you *will win* on each toss. In fact, I cannot even guarantee that you will win any money after 100 tosses.

Can you explain why you chose blue?

Here, unlike the previous game you must use your probability-sense to make the "right" choice. You see that die *b* has five blue faces, and one red. Your judgement tells you that there are “more chances” that the outcome blue will occur, therefore, it is in your interest to choose the color blue.

Can you explain why you chose blue by using the term probability? [3]

Note, at each toss of this die there is a chance that you will have to pay $1.00. However, your probability-sense tells you that if you play this game many times you will, “on average,” earn money. It is not *certain* that you will always earn, but it is *very likely*, or *highly probable*.

Although we did not define the concept of an average, I believe that you have a qualitative estimate of the average (or the expected, as mathematicians refer to it) earnings after 100 tosses? [4]

Remember that in this game there are some chances that you will earn $100.00. There are also some chances that you will lose $100.00. Your probability-sense tells you that the former is more likely than the latter. Knowing probability theory allows you to make a more precise statement on the probabilities of these two extreme events. [5]

***Third game:***

I chose die *c* from Figure 1.1, you examine it, count the number of faces having different colors, and choose a color.

Which color will you choose?

Your intuition, or your probability-sense tells you that it will be advantageous to choose blue. Can you explain why you chose blue in the term probability?[6]

Clearly, in this particular case the chances of earning in 100 tosses is less than in case *b*, but it is still to your advantage to choose the color blue. The argument favoring the color blue is not as powerful as in the previous case (*b*), and certainly less powerful than in case ***a***, yet it is still a good choice. Can you estimate your average net earnings in 100 tosses?[7]

Remember that in this case, you might earn $100.00. You might also lose $100.00 in 100 tosses. Can you estimate the probabilities of these two extreme events?[8]

***Fourth game:***

Next, I choose die ***d***. You look at the faces, count how many reds, and blues there are. Refresh your memory about the "rules of the game" before you choose a color.

Which color are you going to choose?

This is the most “difficult” game. It is difficult because you are clueless as to the preference of occurrence of either the red, or the blue. In Chapter 2, we shall see that in this case you are “given” the minimal “*information*” on how to make a choice. However, without knowing information theory, and without knowing even probability theory, your probability-sense tells you that there is no preferred color. In other words, you can choose either blue or red; there is no advantage in any of these choices.

Can you repeat the argument on the lack of advantage for any particular choice in terms of probabilities?[9]

Your intuition – or probability-sense tells you that whichever color you choose, and that no matter how many times you play this particular game, your expected “gain” is the same as your expected “loss,” and therefore your expected net gain is $0.0. This is another way of characterizing a fair die which has three red, and three blue faces.

Note again that even with this die (*d*), there is a chance that you will earn $100.00, but there is also a chance that you will lose $100.00. However, these two extreme cases are extremely improbable, and their probabilities of occurrence are equal.[10]

You can easily calculate that your average earnings in this case is $0.00. However, we do not need to calculate this average as we can trust our intuition, or our probability sense.

Although this particular case is relatively simple, it is instructive to pause, and calculate a few more probabilities. This will be important for understanding the Shannon measure of information in Chapter 2, and very important as well for understanding the Second Law in Chapter 3.

We saw in case ***d*** that the two outcomes; blue and red have equal probabilities; ½. We also saw that in 100 throws there is a small probability (very small but finite) that all of the outcomes will be red (and the same that all of the outcomes will be blue).

However, there is also a probability that any possible sequence of blues and reds will occur. For simplicity (to minimize on writing), suppose that we toss the die ***d*** ten times. A possible outcome in 10 tosses could be:

B, B, R, B, B, R, R, B, B, R

(B = *Blue and* R = *Red*). This is referred to as a *specific* sequence of 10 outcomes. By *specific*, we mean that we know which color occurred at which throw; first B, second B, third

R, and so forth. In this specific sequence, there are 6Bs, and 4Rs. We shall later learn that all *specific* sequence with 6Bs, and 4Rs have the *same probability* as the sequence of all Rs, or all Bs.

Sometimes we are interested in a *non-specific* sequence which we call a *generic* sequence. This means that we toss the die ***d*** ten times, and got, say 6Bs, and 4Rs, but we do not care about the *order* of the occurrence of the Bs, and the Rs, For instance, the sequences:

B, B, B, R, R, R, B, R, B, B

R, R, B, R, R, B, B, B, B, B

These two sequences have 6Bs, and 4Rs, therefore, they are two *different specific* sequences, but they are of the *same generic* sequence. I should mention at this point that the distinction between a *specific* and a *generic* sequence is essential for understanding the Second Law. As an exercise, write down all possible sequences of four tosses of die ***d***. Can you tell why a generic sequence of Bs and Rs will always have larger probabilities compared with a *specific* sequence, except for the extreme case of all Bs, or all Rs?[11]

**Table 1.1: All possible sequences of four outcomes of blues and reds**

| Sequence | Probability of the specific sequence | Probability of the generic sequence |
|---|---|---|
| BBBB | 1/16 | 1/16 |
| BBBR<br>BBRB<br>BRBB<br>RBBB | 1/16 | $4 \times \frac{1}{16} = \frac{1}{4}$ |
| BBRR<br>BRBR<br>RBBR<br>RBRB<br>RRBB<br>BRRB | 1/16 | $6 \times \frac{1}{16} = \frac{3}{8}$ |
| RRRB<br>RRBR<br>RBRR<br>BRRR | 1/16 | $4 \times \frac{1}{16} = \frac{1}{4}$ |
| RRRR | 1/16 | 1/16 |
|  | sum = 1 | sum =1 |

Note that the probability of each specific sequence is the same 1/16.

At this point, I mentioned the distinction between a *specific*, and a *generic* sequence only in the context of probability. As you can see from Table 1.1 (Note 11), except for the two extreme cases, the probability of the generic sequence is always larger than the probability of a specific sequence. Can you explain why?

Note also that each specific sequence has the same probability.[12]

***Fifth game:***

Next, I choose die *e*. You examine it, and as in the previous game, recall the rules of the game, and choose a color.

You realize that this case is "similar" to the case of die *c*. Which color will you choose now, and why? Can you estimate your net earnings in this case? What if you have chosen blue?

***Sixth game:***

Next, I choose die *f*, and you have to choose a color in order to play the same game with exactly the same rules.

Which color would you choose? Is this game similar to any of the previous games?

***Seventh game:***

I choose die *g*. By remembering the rules of the game, you should not have any problem in choosing the color which will guarantee that you will gain at each toss. This game is similar to the case of die *a*, only the color has changed.

Now that you have successfully played all these games, and you have hopefully earned (on average) in most of the games, you should realize that in all the games you have used probabilistic judgements, or probabilistic arguments. In each case you had to make a choice between two possibilities.

Briefly, I hope you were convinced that you have some degree of probability-sense. This probability-sense is not much different from what people refer to as "common sense."

In order to actually calculate the various probabilities of some events you should learn the *rules* of the Theory of Probability. We shall devote sections 1.3 and 1.4 to learn the elements of this theory.

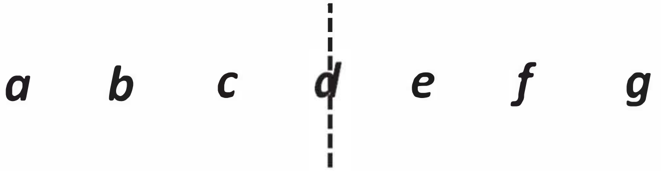


Figure 1.4. Symmetry in the properties of the seven dice about die *d*.

Before we go on, you should look again at the seven dice of Figure 1.1, and the seven games we played. You should notice an element of symmetry in these games. Look at Table 1.2 in order to aid you in finding this symmetry. Symmetry in the sense that game *a* is equivalent to *g*, *b* is equivalent to *f*, and *c* is equivalent to *e*. *d* is not equivalent to any other game, In Figure 1.4 we draw a line passing through *d*. The games on the right-hand side of this dashed line are the same as to the games on the left- hand side. Note carefully that die *a* is not the *same* as die *g*, and die *b* is not the *same* as die *f*. When I said symmetry, I meant that *playing* with *a* requires the same effort or judgement as *playing* with *g*, and *playing* with *b* is the same as *playing* with *f*, and *playing* with *c* is the same as *playing* with *e*.

## 1.2 Uncertainty-sense about the game we played

This section is actually an introduction to Chapter 2. We use the seven games we played in the previous section to study a new concept to which we refer to as the "uncertainty-sense."

In Table 1.2 we summarize the probabilities involved in these games as well as the *average probability* in each game. These are very special averages, where we take the average of the probabilities using the same probabilities as the weights for calculating the averages, see Note 4.

**Table 1.2: The probabilities of "blue" and "red," as well as the average probability for each of the dice in Figure 1.1***

| die name | die number | Pr(blue) | Pr(red) | Average of Pr(blue) and Pr(red) | Average of 1 – Pr(blue) and 1 – Pr(red) |
|---|---|---|---|---|---|
| die *a* | 1 | 1 | 0 | 1 | 0 |
| die *b* | 2 | 5/6 | 1/6 | 13/18 | 5/18 |
| die *c* | 3 | 4/6 | 2/6 | 5/9 | 4/9 |
| die *d* | 4 | 3/6 | 3/6 | ½ | ½ |
| die *e* | 5 | 2/6 | 4/6 | 5/9 | 4/9 |
| die *f* | 6 | 1/6 | 5/6 | 13/18 | 5/18 |
| die *g* | 7 | 0 | 1 | 1 | 1 |

* The reader is urged to read Note 4 to see how the averages were calculated.

In this section, we also introduce the uncertainty-sense (which will be changed to information-sense in Chapter 2). In Chapter 2, we shall generalize this concept of uncertainty. Once you have grasped this concept, you will understand what Shannon's measure of information is, and most importantly, once you understand the meaning of SMI, the meaning of entropy will fall right into your hands, ripe and ready for you to savor its meaning!

I therefore, urge you to follow carefully the concept of *uncertainty* with respect to a *single* event, and the concept of *uncertainty* with respect to the *entire* game.

In Figure 1.5a (see color centerfold), we plotted the probabilities of occurrence of the color blue for the various dice, numbered 1 to 7. The points in this plot are colored blue to remind you that these are the probabilities for the occurrence of the blue color in each of the dice in Figure 1.1. As you already know, in case *a*, you were *certain* that the blue color will appear. We denote this certainty by the number 1. We can also say that probability 1 is the *maximal certainty*, or the *maximal likelihood* that the color blue will occur.

In case *b* (number 2 in the figure) you were less certain about the occurrence of the blue. In other words, the likelihood of occurrence of blue was less than in case *a*. We can also say that the smaller the probability of the occurrence of blue is, the more *uncertain* you are about its occurrence, or that it is more unlikely to occur.

Thus, the *decreasing* values of the blue dots in Figure 1.5a (see color centerfold) represent either the *decreasing* values of the probabilities, the extent of certainty, or the likelihood of the occurrence of blue. When we say that we are "less certain" about an event, it is tantamount to saying "more uncertain." Similarly, less likelihood is the same as more unlikelihood. Please convince yourself that this is true.

In Figure 1.5b, we plotted the probabilities Pr(red), from the fourth column in Table 1.2. Here, the red dots increase from die *a* to *g* (or from 1 to 7). Figure 1.5c combines the data from both Figures 1.5a and 1.5b. Note the symmetry about the center of the figure.

Now, to make sure that you understand the meaning of certainty (or uncertainty) and likelihood (or unlikelihood), I suggest that you repeat the same discussion of the previous paragraph, but with respect to the red color.

In the fifth column in Table 1.2 we calculated the *weighted average* values of Pr(blue) and Pr(red), see note 4. In Section 1.9, we shall define more precisely what we mean by an average value. Here, we assume that we have an intuitive notion of an average between two numbers Pr(blue) and Pr(red) for each die which we write as: [13]

Average of probabilities $= \Pr(\text{blue})\Pr(\text{blue}) + \Pr(\text{red})\Pr(\text{red})$

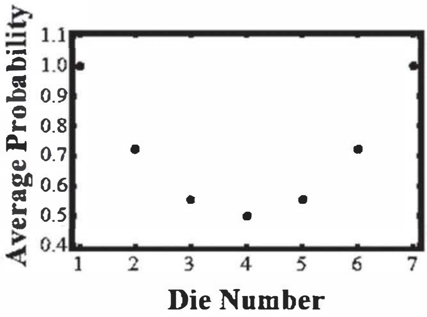


Figure 1.6. The *average probabilities* for the dice 1 to 7, corresponding to *a* to *g* in Table 1.2

Thus, the numbers we wrote in the fifth column are averages of the two numbers Pr (blue) and Pr (red) in each row. The average being calculated with the weights or the probabilities Pr (blue) and Pr (red.)[13] As we shall see in Chapter 2, this is a very special kind of average. It is *special average* since we calculate the average of Pr (blue) and Pr (red) using the "weights," or the probabilities Pr (blue) and Pr (red), see Note 13, Table 1.2 and Figure 1.6. Now, we see that this average starts with the value of 1. For die *a*, it is:

$$\Pr(\text{blue})\Pr(\text{blue}) + \Pr(\text{red})\Pr(\text{red})$$

$$= 1 \times 1 + 0 \times 0 = 1$$

It goes through a minimum value of ½ for die d (or die number 4 in Figure 1.6. For die d, the average is:

$$\mathrm{Pr(blue)Pr(blue)} + \mathrm{Pr(red)Pr(red)}$$

$$= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Then the value of the average climbs again to 1. Thus, for die *g* we have:

$$\mathrm{Pr(blue)Pr(blue)} + \mathrm{Pr(red)Pr(red)} = 0 \times 0 + 1 \times 1 = 1$$

Take note of the *symmetry* in the curve of the averages drawn in Figure 1.6. This is the symmetry we already saw when we played games *a* to *d*, then games *d* to *g*.

Note also that the values in the third and in the fourth columns measure the extent of *certainty* (or likelihood) of each outcome (blue or red). The values in the fifth column of Table 1.2 measure the *average* extent of certainty (or likelihood) over *all possible* outcomes of the relevant die.

If Pr(blue) is a measure of the extent of certainty of the occurrence of blue, then $1 - \mathrm{Pr(blue)}$ is a measure of the extent of uncertainty of the occurrence of the blue. For instance, for case *a*, the extent of certainty we have about the occurrence of blue is 1, and the extent of uncertainty of occurrence of blue is $1 - 1 = 0$. In the last column in Table 1.2 we have the average extent of *uncertainty* with respect to the entire distribution. Figure 1.7a.

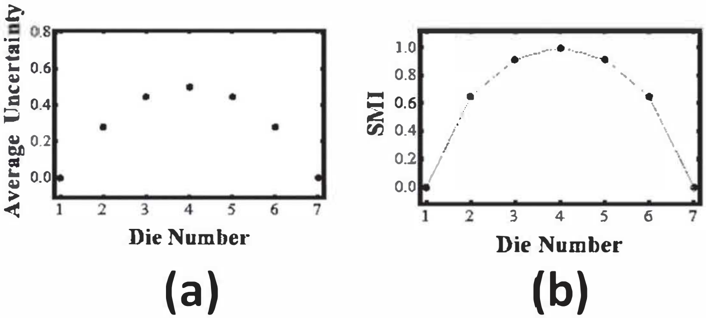


Figure 1.7. (a) The average *uncertainties* for the dice 1 to 7, corresponding to *a* to *g* in Table 1.2.
(b) The *average uncertainties*, as defined by the Shannon Measure of information (SMI), for the dice 1 to 7, corresponding to *a* to *g* in Table 1.2

This average uncertainty is calculated by:

The weighted average uncertainty=

$$\begin{aligned}\Pr(\text{blue})[1-\Pr(\text{blue})]+\Pr(\text{red})[1-\Pr(\text{red})]\\ &=\Pr(\text{blue})\Pr(\text{red})+\Pr(\text{red})\Pr(\text{blue})\\ &=2\Pr(\text{blue})\Pr(\text{red})\end{aligned}$$

Look carefully on how we calculate this average; it is the average of the two numbers: $[1 - \Pr(\text{blue})]$ and $[1 - \Pr(\text{red})]$, using the weights $\Pr(\text{blue})$ and $\Pr(\text{red})$.

This average is equivalent to “inverting” the graph in Figure 1.6 so that instead of minimum *uncertainty*, we get a maximum in the uncertainty. We define this “inversion” by:

$$1 - [\mathrm{Pr(blue)Pr(blue)} + \mathrm{Pr(red)Pr(red)}]$$

$$= 1 - \mathrm{Pr(blue)}^2 - \left(1 - \mathrm{Pr(blue)}\right)^2$$

$$= 2\mathrm{Pr(blue)Pr(red)}$$

which is the same as the average uncertainty associated with the distribution of the two outcomes. Pause and compare the definition of the average certainty and average uncertainty.

In Chapter 2, we will generalize the concept of certainty (or uncertainty) with respect to games (or any other experiment) with many outcomes. You will see that you can estimate in a qualitative way the average certainty (or uncertainty) of the entire experiment. The quantitative measure of this average uncertainty is the Shannon measure of information (SMI). We shall discuss this in detail in Chapter 2. Here, these games are shown in Figure 1.7b. Note the similarity of the two graphs in Figures 1.7a and 1.7b, in Chapter 3. We shall see that the same quantitative measure of uncertainty, when applied to a specific probability distribution is the entropy of a thermodynamic system.

We shall also learn in this chapter about a measure of information. Here, you have noticed that in playing either ***a***, or *g*, you feel that knowing the distribution provides a “lot of information," therefore, you could easily choose the right color. In case *b* or *f*, you feel that the distribution provided "less information," therefore it was more difficult to choose the right color.

Similarly, in cases *c* or *e*, you obtained "less information" so you were less certain about the choice of color. Finally, in case ***d***, you “feel” that you obtained "no information" as to which color to choose. I say “feel” because we did not quantify the measure of information. This will also be done in this chapter. At this stage, you should only reflect about the relative difficulty of making a choice of the color in each case. The easiest case was ***a*** (or *g*), and the most difficult case was ***d***.

Now that we know that we have a sense of probability, let us discuss some historical notes, and a few elements in the theory of probability.

## 1.3 The emergence of probability as a branch in mathematics

Probability theory is a branch of mathematics. It has uses in all fields of science, from physics and chemistry, to biology and sociology, to economics and psychology; in short, everywhere and anytime in our lives. For an elementary discussion of probability theory, see Ben-Naim (2015a). We do probabilistic "calculations" or "assessments," consciously or unconsciously, in many decisions we make, whether it be crossing the street, taking medicine, or serious decisions such as getting married. In many activities we try to estimate the chances of success or failure. Without this kind of probabilistic thinking, a doctor could not diagnose a disease from the symptoms, nor can he or she prescribe the best medication for the disease that has been diagnosed.

Historically, the theory of probability sprang from questions addressed to mathematicians by gamblers, presuming that the mathematicians have a better *knowledge* of how to *estimate* the chances of winning a game. Perhaps, some even believed that certain people have a "divine" power and that they could *predict* the outcome of a game. It is interesting to note that the Latin word for "guessing" is *adivinaré*, or in Spanish *adivinar*. The verb contains the root "divine." Today, when one says, "I guess," or when a Spanish speaking person says "*yo adivino*," it does not imply that one has some power to "predict" the outcome. Originally, the term *adivinaré* probably implied some "divine" power to "predict" the outcome of an experiment, or a game.

Basically, probability is a *subjective* quantity measuring one’s degree or extent of belief that a certain event will occur. For instance, I may estimate that there is only a 10% chance that it will rain in Jerusalem tomorrow. You might say that there is a 90% chance that it will rain in Jerusalem tomorrow. The reason that such an extreme discrepancy exists is mainly a result of different people having different *information* on the weather, and that they have varying assessments of this information.

Even when two people have the same information, they might process this information in such a way so as to reach different estimates of the chances, or the probability of the occurrence of an event (or the extent of plausibility of some proposition). Out of this highly vague, qualitative and subjective notion, a distilled, refined theory of probability has evolved which is quantitative and constitutes an *objective* branch of mathematics. Although it is not applicable to all possible events, probability is applicable to a very large body of events; for instance, games of chance and many “events” which are the outcomes of experiments in physics.

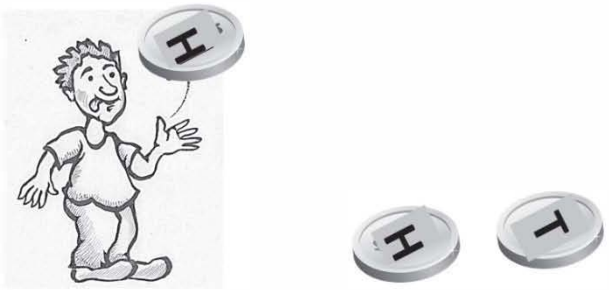


Figure 1.8. Tossing a coin, having two possible outcomes H and L.

When we toss a coin which we have no reason to suspect to be unbalanced or "unfair," we believe that the odds for the outcomes head (H) or tail (T) are 50%:50%, respectively, Figure 1.8. In essence, there is no proof that these are the "correct" probabilities. One can adopt a practical "experimental proof" based on actual, numerous tossing of a coin and counting of the frequencies of the outcomes. If we toss a coin a thousand times, there is a good chance that about 500 outcomes will turn out to be H and about 500 will turn out to be T; but there is also a chance that we will get 590 Hs and 410 Ts. In fact, we can get any sequence of Hs and Ts by tossing the coin a thousand times; there is no way to *derive* or to extract the probabilities from such experiments. We must accept the existence of probabilities for such well-defined experiments. The odds of 50:50 per cent, or probability half for H and half for T, must be accepted as something belonging to the event, much as a quantity of mass belongs to a piece of matter.

Today, the concept of probability is considered to be a primitive concept that cannot be defined in terms of more primitive concepts.

Thus, we all agree that if we toss a coin, and we have no reason to believe that the coin is unfair, then the chances of falling with either Head (H) or Tail (T) up are the same.

Similarly, when we throw a die, unless we have additional information, we assume that each of the possible outcome: 1, 2, 3, 4, 5, 6 have the same chances of occurring, Figure 1.2a.

Let us go back to the pre-probability theory era from the 16th and 17th centuries, when the concept of probabilities had just begun to emerge. An example of a question allegedly addressed to Galileo Galilei (1564–1642) was the following:

Suppose we play with three dice and we are asked to bet on the *sum* of the outcomes of tossing the three dice simultaneously. Intuitively, we feel that it would not be wise to bet our chances on the outcome of 3, nor on 18; our feeling is correct (in a sense discussed below). The reason is that both 3 and 18 have only one way of occurring; 1:1:1 or 6:6:6, respectively, and we intuitively judge that these events are relatively rare. Clearly, choosing the sum 7, is better. Why?

Because there are more *partitions* of the number 7 into three numbers (between 1 and 6), i.e. 7 can be obtained as a result of four possible partitions:

1:1:5,   1:2:4,   1:3:3,   2:2:3.

We also *feel* that the larger the sum, the larger the number of partitions, up to a point roughly at the center between the minimum of 3, and the maximum of 18. But how can we choose between 9 and 10? A simple count shows that both 9 and 10 have the same number of partitions, i.e. the same number of combinations of integers (from 1 to 6), the sum of which is 9 or 10. Here are all the possible partitions:

| | |
|---|---|
| For 9: | 1:2:6, 1:3:5, 1:4:4, 2:2:5, 2:3:4, 3:3:3 |
| For 10: | 1:3:6, 1:4:5, 2:2:6, 2:3:5, 2:4:4, 3:3:4 |

At first glance, we might conclude that since 9 and 10 have the same number of partitions, they should also have the same chances of winning the game. This conclusion is wrong as will be discussed below. The correct answer is that 10 has better chances of winning than 9. The reason is that, though the number of partitions is the same for 9 and 10, the total number of outcomes of the three dice that sum up to 9, is a little bit smaller than the number of outcomes for 10. In other words, the number of partitions is the same, but each partition has a different "weight," e.g. the outcome 1:4:4 can be realized in three different ways:

1:4:4, 4:1:4, 4:4:1

When we count all the possible partitions and all the possible weights, we get the following results. We quote here the final results. For a more detailed discussion of this problem, see Ben-Naim (2007).

The total number of outcomes for 9 is 25.

The total number of outcomes for 10 is 27.

Therefore, the relative chances of winning with 9 and 10, is 25:27, i.e. favoring the choice of 10. Thus, the best choice of a winning number, presumably as suggested by Galilei, is 10.

But what does it mean that 10 is the "best" choice and that this is the "correct" winning number? obviously, I could choose 10 and you could choose 3 and you might win the game. Do our calculations guarantee that if I choose 10, I will always win?

Obviously not. So what does the ratio 25:27 mean? The theory of probability gives us an answer. It is not a precise, nor a fully satisfactory answer, and it does not guarantee winning; it only says that if we play this game many times, the probability that the choice of 9 will win is 25/216, whereas the probability that the choice of 10 will win is slightly bigger, 27/216 (216 being the total number of possible outcomes; $6^3 = 216$).

How many times do we have to play in order to guarantee my winning? On this question, the theory is mute. It only says that in the limit of an infinite number of games, the frequency of occurrence of 9 should be 25/216, and the frequency of occurrence of 10 should be 27/216. But an infinite number of games cannot be realized. So what is the meaning of these probabilities? At the moment, we can say nothing more than that the ratio 27:25, reflects our *belief* or our degree of confidence that the number 10 is more likely to win than the number 9.

We mentioned this example here because of its historical significance in the development of the theory of probability. Before we go on, I suggest that you do the calculation on a simpler game of two dice rather than three. This is a simple game and you should be able to do all the required calculations. The problem is simple. We throw two dice simultaneously. We record the *sum* of the two outcomes: The "events" here are: 2,3,4, … ,12. Can you calculate the probabilities of each of these outcomes? See Note 14.

## 1.4 The mathematical approach to probability

This section is not essential for understanding either the SMI, or entropy. The mathematical, or the so-called axiomatic approach was developed mainly by Kolmogorov in the 1930s. It consists of the following three basic concepts:

**1) *The sample space***

This is the set of all possible outcomes of a specific, well-defined experiment.

Examples: The sample space of throwing a die consists of six possible outcomes {1, 2, 3, 4, 5, 6}; tossing a coin has the sample space consisting of two outcomes {H: T} (H for head and T for tail). Each of these outcomes is called an *elementary event*. Note that we cannot write down the sample space for every experiment. Some sample spaces consist of an infinite number of events (e.g. shooting an arrow at a circular board). We are interested only in simple spaces where the counting of the outcomes, which are referred to as elementary events, is straightforward.

**2) *A collection of events***

A compound event, or simply an *event,* is defined as a union (or a sum) of elementary events.

Examples: (a) The result of tossing a die is “even.” This event consists of the elementary events {2, 4, 6}, i.e. either 2 or 4 or 6 has occurred, or will occur in the experiment of tossing a die.
(b) The result of tossing a die is “larger than or equal to 5.”

This event consists of the elementary events {5, 6}, i.e., either 5 or 6 has occurred.

**3) *Probability***

To each event, we *assign* a number, which is referred to as the probability of that event, which has the following properties:

(a) The probability of each event is a number between zero and one.

(b) The probability of the *certain* event. The certain event is the entire sample space. Since the sample space contains all possible outcomes, its probability is one.

(c) The probability of the *impossible* event is zero. The occurrence of "no event” is assigned the probability zero.

(d) If two events are disjoint or mutually exclusive (disjoint, or mutually exclusive events mean that there are no elementary events common to both events), then the probability of the sum (or union) of the two events is simply the sum of the probabilities of the two events.

Condition (a) simply gives the scale of the probability. In daily life, we might use the range 0–100% to describe the chances of, for example, raining tomorrow. In the theory of probability, the range [0,1] is used. The second condition simply states that if we *do* perform an experiment, one of the outcomes must occur. Therefore, we assign the value of *one* to the sample space, which is the collection of all events. this event is called the certain event. Similarly, we assign the probability zero to the impossible event. The last condition is intuitively self-evident. Mutual exclusivity means that the occurrence of one event excludes the possibility of the occurrence of the second. In mathematical terms, we say that the *intersection* of the two events is empty (i.e. contains no elementary event).

For example, the two events:

$A$ = {the outcome of throwing a die is *even*} = {2,4,6}

$B$ = {the outcome of throwing a die is *odd*} = {1,3,5}

The events $A$ and $B$ are disjoint; the occurrence of one *excludes* the occurrence of the other. If I threw a die and told you that the event "even" occurred, then you know that the event "odd" did not occur.

We define the event:

$C$ = {the outcome of throwing a die is larger than or equal to 5} = {5,6}

Note that $A$ and $C$, or $B$ and $C$ are not disjoint. $A$ and $C$ contain the elementary event 6. $B$ and $C$ contain the elementary event 5.

The events, "greater than or equal to 4," and "smaller than or equal to 2," are disjoint. We can calculate the probability of the first event {4, 5, 6} to be 3/6, and the probability of the second event {1, 2} to be 2/6; hence, the combined (or the union) event {1, 2, 4, 5, 6} has the probability 5/6, which is the sum of 2/6 and 3/6.

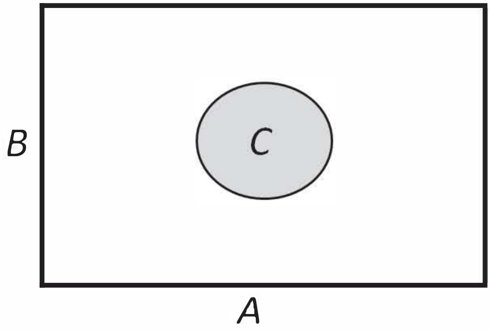


Figure 1.9. A board of total area of $S = A\times B$, and a circle with area $C$.

A very useful way of demonstrating the concept of probability and the sum rule is the Venn diagram. Suppose while blindfolded, we throw a dart at a rectangular board having a total area of $S = A\times B$. We assume that the dart *must* hit some point within the board (Figure 1.9). We now draw a circle within the board, Figure 1.9, and ask: What is the probability of hitting the area within this circle? We assume, by plain common sense, that the probability of the event "hitting inside the circle" is equal to the ratio of the area of the circle to the area of the entire board. (Actually, we are asking about the conditional probability that the dart hit the circle, given that the dart has hit the board).

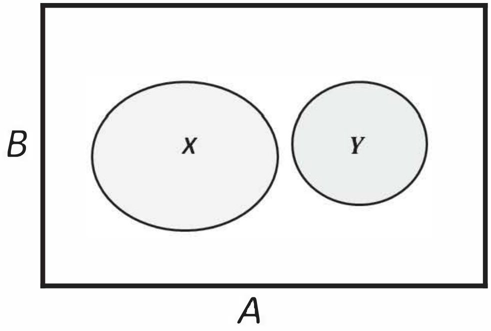


Figure 1.10. Two non-overlapping events; *X* and *Y*

Two regions drawn on the board are said to be disjoint if there is no overlap between the regions (Figure 1.10). We assume that the probability of hitting either one region or the other is the ratio of the area of the two regions, to the area of the whole board. This leads directly to the sum rules stated above. The probability of hitting either one of the regions is the sum of the probabilities of hitting each of the regions. This sum rule does not hold when the two regions overlap, i.e. when there are points on the board that belong to both regions, like the case shown in Figure 1.11.

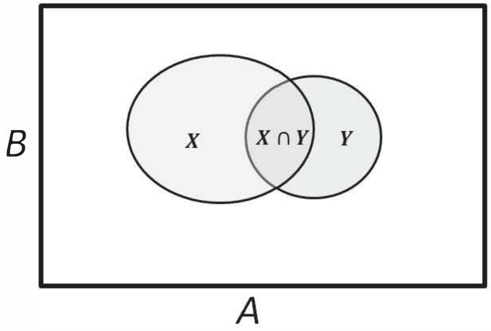


Figure 1.11. Two overlapping events, *X* and *Y*, and the corresponding intersection.

It is intuitively clear that the probability of hitting either of the regions is, in this case, the sum of the probabilities of hitting each of the regions, minus the probability of hitting the overlapping region. Another way of understanding this is to think of the area covered by the two regions; it is the sum of the two areas of the two regions—minus the area of the intersection.

On this relatively simple (axiomatic) foundation, the whole edifice of the mathematical theory of probability has been erected. It is not only extremely useful but also an essential tool in all the sciences and beyond. As you must have realized, the basics of the theory are simple, intuitive, and require no more than common sense.

In the mathematical theory of probability, the probabilities are said to be *assigned* to each event. These probabilities must subscribe to the four conditions *a*, *b*, *c*, and *d*. The theory does not *define* probability, nor does it provide a method for calculating or measuring probabilities. In fact, there is no way of calculating probabilities for any general event. It is still a quantity that measures our degree or extent of belief of the occurrence of certain events, and as such, it is a highly subjective quantity. However, for some simple experiments, say tossing a coin, throwing a die, or finding the number of atoms in a certain region of space, we have some very useful methods of calculating the probabilities. They have their limitations and they apply to "ideal" cases, yet these probabilities turn out to be extremely useful. What is more important, since these are probabilities based on common sense reasoning, we should *all* agree that these are the "correct" probabilities, i.e. these probabilities turn from being subjective quantities to objective quantities. We shall describe two very useful "definitions" that have been suggested for this concept in the next section.

## 1.5 How do we calculate probabilities?

I hope that you are comfortable with the intuitive meaning of probability. As we have noted earlier, there is no definition for the term probability. However, for some simple cases we have some methods of calculating probabilities. Two of these are discussed in this section.

### 1.5.1 The classical "definition" of probability

Note that I enclosed "definition" in quotation marks. You will understand the reason why I did so later. For the moment let us "invent" or "discover" this definition by ourselves.

Example: We throw a fair die in such a way that each outcome has the same likelihood of occurrence. What are the probabilities of the following events?

(a) The outcome is {4}.
(b) The outcome is an even number, i.e. it is one of the results {2}, {4} or {6}. We write this event as {2,4,6}.
(c) The outcome is greater than 4. This means it is either {5} or {6}. We write this event as {5,6}.

Note that we use here curly brackets for the event "4", {4}. This is consistent with the notation in set theory. In this book, we use either {4} or "4" to denote that the result "4" occurred.

Write down your answer before you check Note 15.

How did we assign probabilities to these events? We assumed that the die is “fair,” and that we threw it in such a way that each single possible outcome (elementary event) has the same likelihood. This means that the die is a perfect cube, and its mass density is evenly distributed, and that we throw the die in such a way that it spins in the air many times before it lands on the floor with one of its sides facing upwards.

For the event (a) in the example above, we reason out that since there are altogether six possible outcomes, and we assume that each outcome has the same likelihood of occurrence, therefore the probability of a single outcome, say, {4}, in case (a) is 1/6.

If you pause and think about the reasoning that led us to the number 1/6, you will find that we used the phrase “each of the outcomes has the same “likelihood,” which is tantamount to saying that each event has the same probability. Once we also fixed the value of the *certain* event to be one, we can calculate the probability of each elementary event as being 1/6. Thus, what we have done is not to *define* the probability of the event {4}, but to assume that we *know* the probability of the event to be 1/6. In other words, this “definition” is *circular*. It uses the concept of probability (or likelihood or chances or odds) to “define” the probability. Sometimes an argument based on symmetry, or equivalence of all possible results is used to reach the conclusion that the probability of each elementary event is 1/6.

Let us go to case (b). What is the probability of the occurrence of the (compound) event {2,4,6}, i.e. the outcome is an *even* number?

There are two ways of reasoning. First, we can argue that there are altogether six equally likely outcomes (i.e. elementary events), each having probability of occurrence 1/6. Therefore, the occurrence of either {2} or {4} or {6} must be larger than 1/6, and most likely to be the sum of these three probabilities, i.e.: $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6}=\frac{1}{2}$.

The second reasoning is to divide all possible outcomes into two groups of events; “even” and “odd” outcomes. Think of coloring all the faces of the die with even number of dots with red, and all the faces of the die having an odd number of dots with blue. The probability of the event “even” is equivalent to the probability of the event “red” face. Since there are two possible outcomes either “red” or “blue,” and since we believe that the die is fair, we conclude that the probability of the event "red" (or "even") is ½.

Note again that in calculating the probability of the "event" we used *probabilistic* arguments, i.e. we assumed that each elementary event has the same likelihood. Therefore, this method of calculation cannot be viewed as a *bona-fide* definition of the probability of the event "even."

Let us turn to the case (c). The event "greater" than 4 means that the outcome is either {5} or {6}. We write this event as {5,6}. Using the same type of argument as before we can conclude that the probability of the event {5,6} is the sum of the probabilities of the elementary events {5} and {6}, i.e. $\frac{1}{6}+\frac{1}{6}=\frac{2}{6}=\frac{1}{3}$.

Exercise: Suppose that the faces of the die are colored as follows:

Faces {1} and {2}, with red; faces {3} and {4}, with blue, and faces {5} and {6}, with green. What is the probability of the event "green"? [16]

If you have calculated correctly the probabilities of the events in (a), (b) and (c), based on the example above, you have *almost* discovered the so-called *classical definition* of probability.

The classical "definition" of probability is:

For an experiment which has *n* equally likely outcomes (i.e. *n* elementary events) denoted: $A_1, A_2, \dots, A_n$ the probability of a compound *event* B is calculated by the rule:

$$\Pr(\mathrm{B}) = \frac{Number\ of\ outcomes\ included\ in\ \mathrm{B}}{Total\ number\ of\ outcomes}$$

Remember that we assume that each single possible outcome is said to be an elementary event.

Let us first see that our calculations for the events in (a), (b) and (c), in the example given above are consistent with this rule.

In case (a), the event "B" consists of one outcome {4}. Applying the rule above, we get:

$$\Pr(a) = \frac{1}{6}$$

In the case (b), the event "B" contains three outcomes, hence

$$\Pr(b) = \frac{3}{6} = \frac{1}{2}$$

In case (c), the event “B” contains two outcomes, hence

$$\Pr(c) = \frac{2}{6} = \frac{1}{3}$$

We see that the classical “definition” is intuitively clear. You should realize however that this is not a *bona-fide definition* of probability. The “classical definition” already assumes that we *know* the *probabilities* of each elementary event. Therefore, this definition is circular.

Furthermore, this rule of calculating probabilities does not apply in general. First, it is not always clear what the *elementary events* are. For example, for the event: "It will be sunny in New York tomorrow," there are no simple elementary outcomes. For now, it is sufficient to say that in the case of throwing a die we assume that there are six possible outcomes (we neglect the possibilities that the die will fall on an edge or on a vertex, or perhaps it will fall and break into pieces so that no definite outcome is observed).

More importantly, sometimes the single outcomes are not equally probable. Therefore, the classical definition – or rather the method of calculating probabilities by the rule given above does not apply for all cases. Nevertheless, it is a very useful rule for calculating the probabilities of a large class of cases.

### 1.5.2 The relative frequency "definition" of probability

In all of the examples discussed in the previous section, we started with the assumption that there exist a finite number of elementary events, or elementary outcomes, and that these have equal likelihood (or chances, or probability) or occurrence. How do we calculate probabilities in cases where there are no obvious elementary events? How are the probabilities of such events defined? The general and honest answer is that there is no *bona-fide* definition of probability, nor a method of calculation which is satisfactory for every event.

What is the probability that a dormant volcano will erupt in the next hour? What is the probability that the sun will explode tomorrow? What is the probability that there are intelligent beings in some other planets?

There is no way of defining, let alone calculating the probabilities of these events. Yet, people do use the term probability in connection with such events. The only meaning that "probability" has in such context, is the extent of one's belief on the chances of occurrence of that event.

However, there is a large class of events for which one can offer an "experimental" way of calculating their probabilities. These are the cases when we can repeat an experiment many times, or certain events have occurred many times in the past, and we can collect "statistics" on specific events. For example, suppose we have a die which is known to be unfair, say a die with an asymmetric distribution of mass, or a partially broken or twisted die. Obviously, we cannot assume that each outcome has the same probability.

In this particular example, we apply the so-called *relative frequency* "definition" of probability.[17]

We throw the die many, many times, say a thousand times, and collect the "statistics" about the frequency of observing the result {1}, {2}, {3}, ⋯ {6}. By "frequency," we mean the ratio of the number of times a specific result occurred and the total number of throws.

Suppose we found that after a thousand throws the following results:

50 results showing {1}

100 results showing {2}

100 results showing {3}

200 results showing {4}

250 results showing {5}

300 results showing {6}

We might tentatively *assume* that the probabilities of the different outcomes are:

$$\frac{50}{1000}, \quad \frac{100}{1000}, \quad \frac{100}{1000}, \quad \frac{200}{1000}, \quad \frac{250}{1000}, \quad \frac{300}{1000}$$

The relative frequency definition states, that if we throw the die *infinite* number of times the fraction of times each outcome occurs is the "probability" of that event.[17]

This definition is problematic at best. In fact, it is also circular in the following sense. We believe that if we do the experiment infinite times the fraction of times each outcome will occur will tend to some constant value between zero to one. Unfortunately, we cannot perform an infinite number of experiments. In fact, no one can guarantee that if we calculate these fractions for many experiments the fractions will tend to some constant values.

In practice, we take a large but finite number of experiments, collect the "statistics," as we did above and *assume* that these results are the approximate probabilities of the outcomes. We believe that if we repeat the experiment many times (thousands, millions, billions…) it is *highly probable* that the fractions we get are the "true" probabilities. You see that we use the concept of "probable" to define the concept of probability. Therefore, this method cannot be considered as a *bona-fide* definition.

Yet, we practically use this method with finite number of experiments to estimate the most likely probabilities. It is not perfect, it does not guarantee that we will get the correct results, and it is not always applicable. Yet, *this is what we have*, and in many cases this method is very useful.

Based on these methods we can determine the approximate probabilities of outcomes of an unfair dice as we did before. Doctors and pharmaceutical companies determine the efficacy of certain drugs. Insurance companies estimate the likelihood that a certain person (within a specific age bracket, sex, education, marital status, etc.) will be involved in an accident, and with this estimate they will calculate the cost of one's insurance policy. In all of these cases, and in many others we do not have "exact" probabilities, but this is the best we have, and we use them because we *need* to use them.

Exercises: Suppose I threw an unfair die a thousand times, and found the following results:

The outcome {1} occurred 600 times.

The outcome {2} occurred 98 times.

The outcome {3} occurred 96 times.

The outcome {4} occurred 95 times.

The outcome {5} occurred 97 times.

The outcome {6} occurred 14 times.

Based on the gathering of these statistics, estimate the probabilities of each of the outcomes. What is the probability of the event "even?" What is the probability of the event "odd?" Can you guess how I prepared this die?

Based on the given data my estimate of the probabilities is:

| Event: | {1} | {2} | {3} | {4} | {5} | {6} |
|---|---|---|---|---|---|---|
| Probability: | $\frac{600}{1000}$ | $\frac{98}{1000}$ | $\frac{96}{1000}$ | $\frac{95}{1000}$ | $\frac{97}{1000}$ | $\frac{14}{1000}$ |

The probability of the event "even" is: $\frac{98+95+14}{1000} = \frac{207}{1000}$.

The probability of the event "odd" is: $\frac{600+96+97}{1000} = \frac{793}{1000}$.

You may guess that I took a regular die, and added a heavy metal to the face having six dots so that this face will, with high probability, land on the floor, while face "1" will face upwards. In this case the probability of the outcome {1} will be the largest, the probability of the outcome {6} will be the smallest, and all other faces have nearly equal probabilities of about 1/10.

Note that this method gives only a reasonable estimate of the probabilities. It does not guarantee that these are the "correct" probabilities. We believe that had we done one million throws or one billion throws, we shall get results which are closer and closer to the "correct" probabilities. But the truly correct probabilities are elusive. Even if we do infinite throws (whatever that means) we cannot be assured that we shall get the "correct" probabilities (whatever that means too). Yet, in spite of this uncertainty, this is the most useful and most used method of calculating probabilities. It is useful not because it is accurate, but because this is the *best* method available to us.

You might wonder how this method compares with the classical "definition." The latter sounds more accurate, more precise, more reliable, but this is only an illusion.

First, we can never be sure that the die is perfectly fair (whatever that means). If we are not sure, we must use the experimental method, and if we find that each outcome occurs with the same frequency, i.e. about $\frac{100}{600}$, or $\frac{1000}{6000}$, etc. then we can be reasonably sure that it is a fair die, and that the probabilities are 1/6. But what if we are (somehow) sure that the die is fair, how do we know that the probabilities of the outcome are equal to 1/6? In fact, we do not *know*. We believe that this is a reasonable assumption. If we doubt this assumption we can do the experiment, or we can imagine doing the experiment many times. Of course, our *imagination* allows us to repeat the same experiment *infinite* number of times, and *imagine* that the relative frequencies will all be equal to 1/6.

### 1.5.3 Conclusion

In this section, we saw that there is no bone-fide definition of the term probability. Yet, we also learned that we have some kind of a sense-of-probability which guides us in making some probabilistic decisions. We also learned that there is absolutely no reliable method of calculating probabilities. Instead, we rely on, or believe that if we repeat an experiment many times, the resulting frequencies will be a reasonable measure of the extent of certainty or uncertainty regarding the occurrence of one outcome or another. We live in a world of uncertainty, this is clear. This section taught us that we are also uncertain about the extent of our uncertainty of the occurrence of an event.

Yet, in spite of all these uncertainties, shortcomings, limitations and so on, we need probabilities in almost every aspect of our lives. Without a knowledge of probability theory, we might not be able to make prudent decisions, or we might be led to making the wrong decisions. Probability theory offers us the best it can, and in many cases it is extremely useful. It is difficult to find even one aspect of our lives which do not depend on probabilistic reasoning and decision making. For more details and examples, see Ben-Naim (2015b).

## 1.6 Independent Events and Conditional probability

The concept of dependence between events and conditional probability are central to probability theory and have many uses in sciences. In this section, we discuss briefly the concept of independence between the two events. Two events are said to be *independent* if the occurrence of one event has no effect on the probability of occurrence of the other.

For example, if two people who are far apart from each other throw a fair die individually, the outcomes of the two die are independent in the sense that the occurrence of say, "5" on one die, does not have any effect on the probability of occurrence of a result, say, "3," on the other, Figure 1.12a. On the other hand, if each die has a little magnet, then the outcomes of the two results would be dependent, Figure 1.12b. Intuitively, it is clear that whenever two events are independent, the probability of the occurrence of both events, say, "5" on one die, and outcome "3" on the other, is the *product* of the two probabilities. The reason is quite simple. By tossing two dice simultaneously, we have altogether 36 possible elementary events. Each of these outcomes have equal probability of $^1\!/_{36}$ which is also equal to $^1\!/_6$ times $^1\!/_6$.

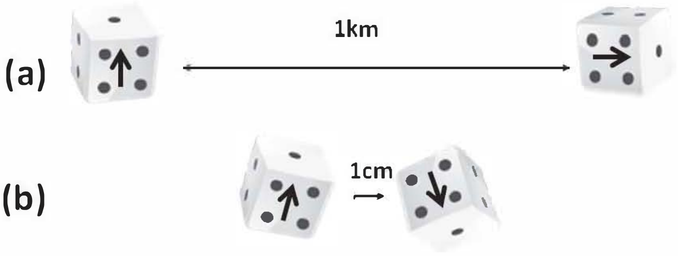


Figure 1.12. (a) Two independent outcomes of two dice, (b) The same two dice when they are at closed distance, their outcome would be dependent.

---

A fundamental concept is the conditional probability. This is defined as the probability of the occurrence of an event $A$ given that an event $B$ has occurred. We write this as $\Pr\{A|B\}$ (Read: Probability of $A$ given $B$).[18]

Whenever the two events are independent, then the occurrence of $B$ has no effect on the probability of the occurrence of $A$. We write this as $\Pr(A|B) = \Pr(A)$. The interesting cases are when the events are dependent, i.e., when the occurrence of one event does affect the occurrence of the other. In everyday life, we make such estimates of the conditional probabilities frequently.

Sometimes, the occurrence of one event enhances the probability of the second event, sometimes it could diminish it. Examples:

1. The probability that it will rain this afternoon, given that the sky is very cloudy at noon, is *larger* than the probability of "raining this afternoon."
2. The probability that it will rain this afternoon, given that the sky is clear at noon, is *smaller* than the probability of "raining this afternoon."
3. The probability that it will rain today, given that the outcome of tossing a die is "4," is the same as the probability of "raining today."

We can say that in the first example, the two events are *positively* correlated. In the second example, they are negatively correlated, and in the third example, they are uncorrelated or indifferent.

In the three examples given above, we *feel* that the statements are correct. However, we cannot quantify them. Different persons would have made different estimates of the probabilities of "raining today this afternoon." To make things more quantitative and objective, let us consider the following events:

$A$ = The outcome of throwing a die is “4”

$B$ = The outcome of throwing a die is “even,” (i.e. it is one of the following: 2, 4, 6)

$C$ = The outcome of throwing a die is “odd,” (i.e. it is one of the following: 1, 3, 5)

We can calculate the following two conditional probabilities

$$\Pr\{of\ A|given\ B\} = \frac{1}{3} > \Pr\{of\ A\} = \frac{1}{6}$$

$$\Pr\{of\ A|given\ C\} = 0 < \Pr\{of\ A\} = \frac{1}{6}$$

In the first example, the knowledge that $B$ has occurred *increases* the probability of the occurrence of $A$. Without that knowledge, the probability of $A$ is ⅙ (one out of six possibilities). Given the occurrence of $B$, the probability of $A$ becomes *larger*, ⅓ (one out of three possibilities). But *given* that $C$ has occurred, the probability of $A$ becomes zero, i.e. *smaller* than the probability of $A$ without that knowledge.

It is important to distinguish between *disjoint* (i.e. mutually exclusive events) and *independent* events. Disjoint events are events that are mutually exclusive; the occurrence of one excludes the occurrence of the second. Being disjoint is a property of the events themselves (i.e. the two events have no common elementary event). Independent events are not defined in terms of the elementary events comprising the two events but in terms of their probabilities. If the two events are disjoint, then they are strongly *dependent*. For more details and examples, see Ben-Naim (2015b).

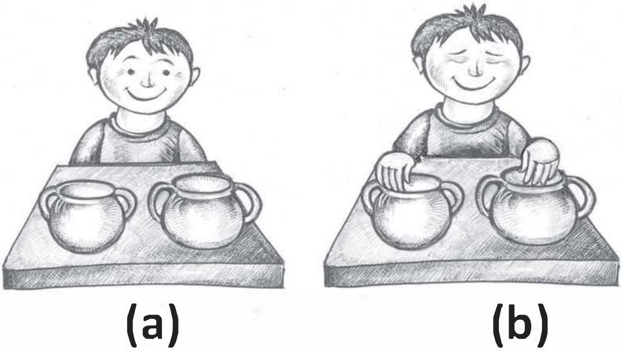
## 1.7 Children's perception of probability

Now that you have a good idea what probability means, and how people calculate or estimate probabilities, it is time to relax a little bit. In this section I will describe a research which examines how children of different ages develop a probability-sense.

I will show you only a few examples of experiments conducted on children ages 4 to 12. The experiment described here is a variation of a similar experiment carried out by Falk *et al* (1980). In the original publication you will find many variations of these experiments and many thoughtful and insightful conclusions. There are also some very surprising results.

Figure 1.13 (see color centerfold) shows two games denoted "easy" and "difficult."

The "easy" game consists of two urns. One on the left, contains four red and four blue marbles. The second on the right, contains four red and eight blue marbles.



(a) A child examining the contents of two urns.

(b) A child chooses while blindfold between two urns after examining their contents.

Children aged 4-11 were shown two urns and their contents. After examining the contents of the two urns they were told to close their eyes, choose an urn, and draw a marble from the urn they chose. They were also told that if the marble they drew is blue they will be rewarded with a prize (win), while if they drew a red marble they will get nothing (lose).

Which urn do you think did the children choose? Which urn will you choose if you were the child asked to draw a marble from the urns? Why did you choose that particular urn? Suppose you chose urn X and did not win, what urn will you choose the next time you play?

What is the probability of winning in this particular game if you choose the urn on the right?

What is the probability of not winning if you choose the urn on the left?

Write down the answers before you consult Note 19.

After contemplating the odds in the easy game shown in Figure 1.13 (see color centerfold), let us go to the more "difficult" game shown in Figure 1.14 (see color centerfold). It is not really difficult, but within the research on children's perception of probabilities, it was considered the more difficult one.

In this game, the urn on the left has the same number of blue and red marbles (4 and 4) as in the easy game in Figure 1.13.

The urn on the right contains 8 blue marbles and 12 red marbles (Figure 1.14).

Children who were exposed to this game responded differently according to their ages. The younger children made the wrong choices most of the time. They chose the urn on the right. When asked why they made that choice they answered "because there are *more* blue marbles in the urn on the right than in the urn on the left." Clearly, those children had no sense-of-probability. As mentioned in connection with the easy game in Figure 1.13, young children chose the urn according to the *absolute* number of blue marbles; the more blues, the more attractive the urn. In the easy game more blues coincided with larger probability. Therefore, the children chose the urn on the right for the wrong reason. In the more difficult game the young children again chose the one on the right. Here, they chose the wrong urn for the wrong reason; the *absolute* number of blue marbles.

Children aged 10 and above made better choices. Their probability-sense told them that the important quantity is not the *absolute number* of blue marbles (eight on right versus 4 on the left), but the rather the *ratio* of the numbers of blues and reds.

For the game in Figure 1.14, the ratios are:

For the urn on the left: $$\frac{\Pr(win)}{\Pr(lose)} = \frac{4}{4} = 1$$

For the urn on the right: $$\frac{\Pr(win)}{\Pr(lose)} = \frac{8}{12} = \frac{2}{3}$$

Clearly, the better choice in this case is the urn on the left.

Exercise: Calculate the probability of winning if you choose the urn on the right, and if you choose the urn on the left.

Suppose you chose the urn on the left, and you drew a red marble, would you change your choice the next time you play the game? Note that by playing the game the "second time" we mean *exactly* the same game. This means that whichever marble you draw in the first trial, it is returned to the urn before you play the next game, and so on.

Now that you are convinced that not-so-young children have a probability-sense, let us test your probability-sense.

Consider the following simple games:

I threw a fair die. I hope you remember what a fair die is. Can you define a fair die?

You see that the die in the air whirls before it lands on the floor.

## Game A

Choose a number between one and six, say “4.” If the die falls with the upper face up showing the number "4", then you get a dollar. If the outcome is different from “4” you get nothing.

Which number will you choose in the game?

Why did you choose this particular number?

How much do you expect to gain if you play this game 1,000 times?

## Game B

The same as in Game A, i.e. if the outcome is the same as the number you chose (say “4”), then you get a dollar. But if the outcome is different, you have to *pay* 21 cents.

Presuming you want to maximize your earnings which number would you choose in this game?

How much do you expect to win (or lose) if you play this game 1,000 times?

Answer these questions before you continue. These are easy questions. After answering compare your results with Note 20.

Now consider a slightly more complicated game.

You are shown an urn, Figure 1.15 (see color centerfold) containing 4 blue marbles, 6 red marbles and 10 green marbles.

You know the contents of the urn. You have to choose a *color* (not an urn as in the previous games); either blue, red or green. You shut your eyes and draw one marble from the urn. If you draw a blue one, you get $2, but you get nothing if it is not a blue marble.

If you choose a red marble and draw a red one, you get $3, but you get nothing if it is not a red marble.

If you choose a green marble and draw a green one, you get $1, but you get nothing if it is not a green marble.

Presuming you want to maximize your earnings, which color will you choose?

Explain why you chose that particular color.

In order to answer the question above you have to calculate first the following probabilities:

What is the probability of drawing a blue marble?

What is the probability of drawing a red marble?

What is the probability of drawing a green marble?

After calculating these probabilities, will you change the "color" of your choice?

The probability of drawing a blue is: $\frac{4}{20} = \frac{1}{5}$.

The probability of drawing a red is: $\frac{6}{20} = \frac{3}{10}$.

The probability of drawing a green is: $\frac{10}{20} = \frac{1}{2}$.

Clearly, the probability of drawing a green is the biggest simply because there are more green marbles. The probability of drawing a blue is the smallest, simply because there are fewer blue marbles. In making a color choice you have to consider both the probability of the color, and the prize you will get if you chose that color. Now, suppose you play the game 1,000 times. Each time you play you choose the same color, and after you draw a marble, and get whatever you earned, you return the marble to the urn, and draw a marble again from the same urn under the same *initial* conditions.

If you chose "blue" your average expected earning is $2 \times \frac{1}{5} \times 1000 = 400$ US$.

If you chose a "red" your average expected earning is $3 \times \frac{3}{10} \times 1000 = 900$ US$.

If you chose a "green" your average expected earning is $1 \times \frac{1}{2} \times 1000 = 500$ US$.

Now you see that in this game, the better choice is the red color. Although it has a smaller probability of being drawn compared to the green, the average earnings when the red marble is chosen is the highest.

If you chose red then you should stick to this color. However, if you chose a green because there is a higher probability of picking a green, then you should switch to red, which has higher expected earnings.

Now consider a more challenging problem. Suppose you chose red in the previous game, and suppose you picked a red and earned $3. Good luck!

Now you repeat the game but with one difference, you do not return the red marble to the urn. Which color will you choose next? Note that now, the initial conditions have changed, Figure 1.16 (see color centerfold). There are four blue, five red and 10 green. Do the calculation before you look at Note 21.

Let us try the calculation for the next game. Suppose that you drew in the second game a “red,” and removed it from the urn. Which color would you choose now? Note that now the initial conditions have changed, Figure 1.17 (see color centerfold) there are four blue, four red and ten green. Do the calculations before consulting Note 22.

Again, you’d better choose the red. However, be careful in the next step. If you drew red in the previous game, and did not return it, the urn now contains a total of 17 marbles, Figure 1.18 (see color centerfold), and the expected earnings are:

For “blue,” $\Pr = \frac{4}{17}$, with expected earning $\frac{4}{17} \times 2 = \frac{8}{17}$

For “red,” $\Pr = \frac{3}{17}$, with expected earning $\frac{3}{17} \times 3 = \frac{9}{17}$

For “green,” $\Pr = \frac{10}{17}$, with expected earning $\frac{10}{17} \times 1 = \frac{10}{17}$

In order to calculate the expected earnings for 1000 draws under the same condition, you have to multiply these numbers by 1000.

At this stage you can see that it is better for you to switch to the "green." Although the prize for the "red" is bigger, the probability of the "green" is now much higher so that the expected earning is bigger for the choice of green.

I hope you did all the proposed calculations and compared them with the notes. You will realize that although we did not *define* what probability is, you were able to calculate some simple probabilities by relying on your probability-sense. In fact you have also calculated some *average* quantities, even though we did not define the term *average*. We shall return to discuss this term in Section 1.9.

## 1.8 Probability Distributions

In this section, we discuss some of the most important *probability distributions*.

Whenever we have an experiment (or a game) which we know has $n$ possible outcomes, and that the probability of the $i$th outcome is $\Pr(i)$, we say that we know the probability distribution of that experiment (or the game).

Example 1: Tossing a coin

The two possible outcomes are Head (H) and Tail (T). Knowing that $\Pr(H)$, and $\Pr(T)$ is equivalent to knowing the probability distribution.

Example 2: Throwing a die

The six possible outcomes are: 1, 2, 3, 4, 5, 6. Normally, we assume that the die is fair. This means that the distribution is $\left(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right)$, i.e. there is equal probability for each of the possible outcomes of this die.

Example 3: Throwing a two-color die

This kind of die was discussed in Section 1.1. Here, there are six faces to the die, but we were interested only in the color of the outcomes. Therefore, in this experiment we have only two outcomes; blue and red. We discussed various distributions of this game in Section 1.1.

In the rest of this section, we discuss a few more distributions which are important in physics in general, and in understanding entropy in particular. We shall describe qualitatively the various distributions.

### 1.8.1 The uniform distribution

Have you even wondered how the fragrance from a small drop of perfume dropped at the corner of the room will, after a short time permeate the entire room. Of course, the reason the molecules of the perfume reach all parts of the room is that they possess *kinetic energy*. Indeed, the kinetic energies of the molecules propel them to move randomly in all kinds of directions and all kinds of velocities. But why do they not stay within the drop of perfume, or in its vicinity? The fact is that after a very short time the molecules of the perfume will reach a *uniform distribution* throughout the entire room. This means that the probability of finding any of these molecules in a small region anywhere in the room is the same, Figure 1.19.

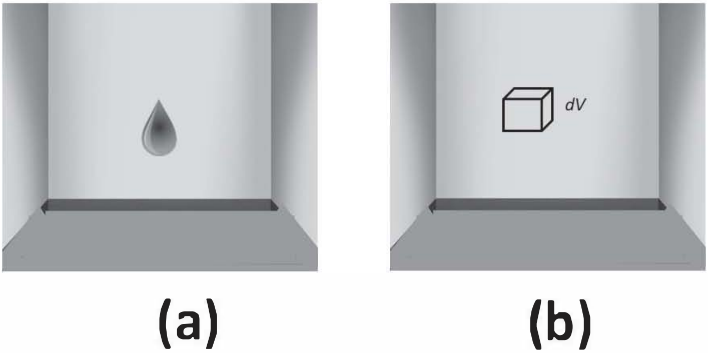


Figure 1.19. (a) A drop of perfume is placed initially at the corner of a room. (b) After sometime the perfume evaporates and occupies the entire room. The density of the perfume at each small element of volume $dV$, is the same at any point in the room.

We take this fact for granted, but understanding it is far from trivial. If you ask someone who learned thermodynamics you might get an answer that the reason for this uniform distribution is the Second Law of thermodynamics. This is true. However, underlying the Second Law is a more fundamental principle. The *uniform probability distribution* is the distribution which has the largest *probability* (provided we neglect the interactions between particles and the gravitational field). I will explain that in the next few pages. I will also return to this aspect of the uniform distribution in Chapter 3. By uniform distribution we mean that a molecule can be found in any small element of volume with equal probability. More precisely, the probability of finding any molecule in the element of volume $dV$ in Figure 1.19 is $\rho dV$, when $\rho$ is the total density of the molecules $(\rho = N/V)$, $N$ is the total number of molecules, and $V$ the total volume of the room.

Let us do a simple experiment. Suppose a box is divided into say, 10 cells. Within one of these cells we place 10 marbles. The partitions between the cells are high enough so that all the 10 marbles can be contained in one of the cells. Initially, we place all the 10 marbles in one cell as shown in Figure 1.20a.

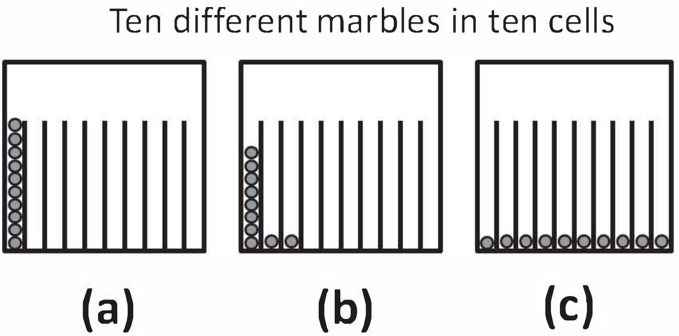


Figure 1.20. (a) The initial configuration of ten marbles in ten cells. (b) The configuration of the ten marbles after a short period of time. (c) The configuration of the ten marbles after a long period of time.

We then *shake* the box vigorously in all directions in such a way that the marbles can cross over the partitions between the cells. While we shake the box, we take snapshots of the *distribution of the marbles* in the various cells. A distribution of the ten marbles in the ten cells is the set of numbers:

$$N_1, N_2, N_3, \dots, N_{10}$$

where $N_1$ is the number of marbles in cell "1", $N_2$ is the number of marbles in cell "2," and so on. We sometimes call the distribution of the marbles in the cell a *configuration*. The initial configuration is written as:

$$N_1 = 10, \quad N_2 = 0, \quad N_3 = 0, \dots, N_{10} = 0$$

This means that all the marbles are in cell number “1.” When we start shaking the box we expect that some marbles from cell “1” will cross over to cell “2,” or “3,” and so forth. For instance, after we shake the box a few times we might find the configuration:

$$N_1 = 8, \quad N_2 = 1, \quad N_3 = 1, \quad N_4 = 0, \dots, N_{10} = 0$$

This means that two of the marbles have crossed over the partitions; one reached cell “2” and one reached cell “3,” Figure 1.20b.

Imagine that while shaking the box you took millions of snapshots. You will see that in most of these snapshots the distribution of marbles in the cells will be uniform or nearly uniform, Figure 1.20c. Note that we use here the term "distribution" for the arrangement of the particles. In most cases we use the same term as shorthand for probability distribution. The same is true when we start with *N* particles in one cell, and remove the partitions between the cells. After sometime the distribution of particles will be uniform in the entire volume *V*. As we shall see in Chapter 3, the reason for obtaining this particular distribution is because this distribution is the most probable one. [23]

### 1.8.2 The Bernoulli distribution and the binomial distribution

Consider an experiment for which there are *only two* possible outcomes; H or T in tossing a coin, hitting the right (R) or the left (L) area on a board, or finding a molecule in one or the other compartments, Figure 1.21. We already discussed some aspects of this probability distribution in Section 1.1.

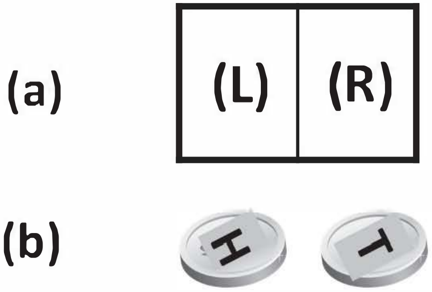


Figure 1.21. Two equivalent problems:
(a) A particle being in the left or right compartment, and
(b) two outcomes of a coin.

We assume that the probability of the occurrence of one of the two outcomes is $p$, and the second, $q$. Since we always assume that the experiment was done, and one of the outcomes has occurred, we must have $p + q = 1$.

Now we repeat the same experiment 10 times. What is the probability of the *specific* sequence of outcomes?

H H T H T T H T H H

given that the probability of H is $p$, and of T, is $q$ (we use the language of H and T, but you can use R or L, or any other notation for the two events), and given that the sequence of experiments are independent? We have:

$$\Pr(\text{a } specific \text{ sequence of six Hs and four Ts})$$
$$= p \times p \times p \times p \times p \times p \times q \times q \times q \times q$$
$$= p^6 q^4$$

We used here the rule that the probability of the ten *independent* events is a *product* of all the individual events.

In general, for $N$ experiments of the same kind as before, the probability of obtaining a *specific sequence* of $n$ outcomes H and $N - n$ outcomes T, is:

$$\Pr(\text{a specific sequence of } n \text{ Hs and } (N-n) \text{ Ts}) = p^n q^{N-n} \qquad (1)$$

This is the Bernoulli distribution.

Note that we underlined the words *specific sequence*. There is a subtle point to pay attention to before we proceed to the next step. A *specific* sequence is when we specify the first outcome, the second outcome, and so on until the $N$th outcome.

For example, for four experiments the *specific* example HTTT has the probability:

$$\Pr(\mathrm{HTTT}) = pq^3$$

The *specific* sequence THTT has also the same probability:

$$\Pr(\mathrm{THTT}) = pq^3$$

Any *specific* sequence having one H, and three Ts has the same probability $pq^3$. When we say *specific* sequence, we mean we are given the specific order of the events; first T, second H, third T, fourth T. We see that for *each specific* sequence having one H, and three Ts, the probability is $pq^3$, independently of the specific order. This result follows from the assumption of independence of the event, and from the application of the multiplication rule for the probabilities of independent events.

Although it sounds a little paradoxical, each *specific* sequence having one H, and three Ts has the same probability $(pq^3)$ no matter what the specific *order* is.

If you are not convinced write down a few *specific* sequences of four results and calculate the probability for each sequence.

Now, we ask a slightly different question. What is the probability of *any sequence* of four outcomes, one of which is H and three of which are Ts? The emphasis is now on the words *any sequence*. We call this a *generic* sequence of one H, and three Ts. To calculate this probability we write first all possible *specific sequences*. These are:

HTTT, THTT, TTHT, TTTH

Check that these are all the possible *specific* sequences with one H, and three Ts. The question we ask is about the probability of finding *either* the first, the second, the third or the fourth of these specific sequences. To calculate this probability, we have to use the rule of *summing* the probabilities of *disjoint* events. The *sequence* can *either* be the first, the second, the third, or the fourth. We cannot obtain two of these at the same time. Hence, the listed four events are *disjoint*, and the probability of obtaining any of these events in this list is:

$$\Pr(\text{any sequence of one H and three Ts}) = 4pq^3$$

The probability of a *generic* sequence of $n$ Hs, and $N-n$, Ts is called the *Binomial* distribution. We shall firther study the Binomial distribution in Chapter 3.

*Exercise*: Calculate the probability of the specific sequence HHHTTT, and the specific sequence HTHTHT. What is the probability of a *specific* sequence of six outcomes having three Hs, and three Ts? What is the probability of the *generic* sequence of six outcomes, having three Hs and three Ts?[24]

### 1.8.3 The normal distribution

The Normal distribution is a very important distribution. It occurs everywhere, not only in physics, but almost anywhere where statistics applies. This is why it is referred to as the *Normal* distribution. It is the *norm* rather than the exception. This distribution is referred to by different names, such as the bell-shaped function (because of its typical shape), the Maxwell-Boltzmann distribution of velocities in one dimension, and the Gaussian distribution after Carl Friedrich Gauss (1777-1855). This distribution may be obtained from the Binomial distribution for large *N*. We will see this trend in Chapter 3. One can also prove it mathematically. See Ben-Naim (2015b).

There are many ways of deriving this particular distribution. All require some degree of mathematics. However, we shall see in Chapter 3 that the probability distribution $\Pr(n, N - n)$ of finding $n$ particles on the left (or $n$ Hs in $N$ series of coin tossing), and $N - n$ on the right tend to have a bell-shaped form. One can prove mathematically that in the limit of very large $N$, the Binomial distribution will tend to the Normal distribution. See Ben-Naim (2015b).

A more elegant derivation based on Shannon's measure of information is available, but this again requires some mathematics. We shall mention this method in Chapter 3.

There is also an "experimental" way of obtaining the Normal distribution which you can do by either imagining an experiment with marbles in cells or simulating the experiment on a computer. This is also described in Ben-Naim (2010).

The Normal distribution was first discovered in analyzing the distribution of errors in an experiment. If you measure any quantity, say people's heights in a given city, the concentration of sugar in your blood, or the weights of newly born babies, you will find a distribution which is similar to the bell-shaped curve.

The simplest way of visualizing the Normal distribution is in the distribution of velocities of particles in a *one dimensional* system. We assume that particles have kinetic energy of motion and that the total energy of all the particles is constant. We also assume that there are no external fields that will affect the location or the velocities of the particles. Due to random collision between the particles, we expect that the motion to the right is as probable as the motion to the left, and whatever the distribution of the velocities is, it will be symmetrical about the value of $v = 0$. Figure 1.22.

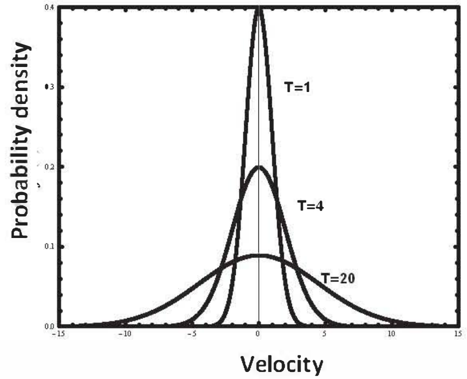


Figure 1.22. Equilibrium distribution of velocities in one dimension, at different temperatures.

The larger the deviation from the center, the lower the probability of finding a molecule with such velocity. As we have pointed out earlier, the width of the curve as a measure by the standard deviation is proportional to the temperature, the higher the temperature, the greater the spread of the molecules on a larger range of velocities.

It should be noted that each of the curves shown in Figure 1.22 is referred to as the probability ***density***. For our purpose we can think of this curve as a series of points. Each point represents the probability of obtaining a velocity value in the vicinity $v_x$.

Finally, it should be noted that so far we have discussed the probability distribution of *velocities* in a one-dimensional system. In this case we have a symmetrical Normal curve. In Chapter 3, we will discuss the distribution of *speeds* of the particles. The speed is the absolute value of the velocity of the particle in any direction in space. This speed is by definition, always positive and its distribution can be derived from the Normal distribution in one-dimension. In this case the distribution is not symmetrical. It has the form as shown in Figure 1.23, and it is referred to as the Maxwell Boltzmann distribution.

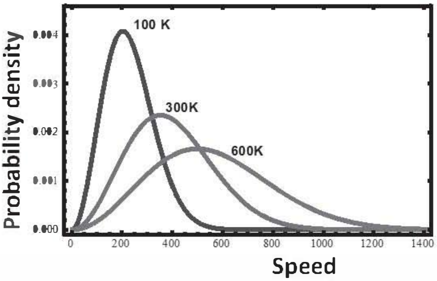


Figure 1.23. Equilibrium distribution of speeds (absolute velocities) in three dimensions, at different temperatures.

### 1.8.4 Conclusion

In this section, we got acquainted with the most important distributions and the most frequently found distributions in natural phenomena. We did not provide any proof about the attainment of these distributions. However, the "proofs" are available both experimentally and theoretically. The experimental proof is easy to carry out, whereas the theoretical proof is more difficult to obtain. This requires some sophisticated mathematics. In Chapter 3, we shall point out that the Uniform and the Normal distributions are related to the Second Law of Thermodynamics.

## 1.9 Average quantities

Before we define an average quantity, let us see if you have an intuitive sense of what an average means.

Suppose you measure the sizes of ten balls, and you obtain the following results:

$$4, 6, 6, 4, 3, 4, 4, 4, 6, 6$$

The units we chose for these measurements are not important, it could be in centimeters, millimeters, or any other units. What is the average size of the balls? We have here all the outcomes of the ten balls, and we calculate the average as:

$$\text{average size} = \frac{\text{sum of all results}}{\text{number of balls}} = \frac{47}{10} = 4.7$$

Thus, we say that the average size of the balls is 4.7. Note carefully that the average is a property of the *entire* set of results, and not of any particular outcome.

Now, suppose that we put the balls inside an urn. While we are blindfolded, we pick up a ball at random, and ask for the probability distribution of this experiment. Looking at the sequence of numbers listed above, we can conclude that:

Probability of size "4" is: $\frac{5}{10} = \frac{1}{2}$

Probability of size "3" is: $\frac{1}{10}$

Probability of size "6" is: $\frac{4}{10}$

Note that the sum of these probabilities must be one. Can you explain why?

Now, we can calculate the average size by the formula:

$$\Pr(4) \times 4 + \Pr(3) \times 3 + \Pr(6) \times 6$$

$$= \frac{1}{2} \times 4 + \frac{1}{10} \times 3 + \frac{4}{10} \times 6 = \frac{47}{10} = 4.7$$

We can now generalize this definition of the average. Suppose we perform an experiment, call it E. We know that there are $n$ possible outcomes (size of balls, people's heights, temperatures of the day, etc.). We denote these outcomes by the sequence of numbers (again, we do not care for the units; whatever units we choose, the average will have the same units):

$$E(1), E(2), E(3), \ldots, E(n)$$

where $E(i)$ is the value of the $i$th outcome. Next, suppose we also know the probability distribution:

$$\Pr(1), \Pr(2), \ldots, \Pr(n)$$

where $\Pr(i)$ is the probability of the $i$th outcome. The average of the outcomes of this experiment is defined by:

$$\begin{aligned}&\text{average of E}\\&= \text{sum of the products } \Pr(i)E(i)\end{aligned}$$

This sum is written as:

$$\sum_{i=1}^{n} \Pr(i)E(i)$$

where the symbol Σ stands for the "sum over all possible indices $i$."

We already encountered an average quantity when we played with various dice. We shall encounter a very special kind of an average in Chapter 2. This same average will also be the basis on which we define and interpret the entropy in Chapter 3.

To check your understanding of the concept of averages, consider the following short stories:

### 1.9.1 Can an average Grade be Higher than the Highest Grade?

We are informed that the average grade of all the students in the Average-State University (ASU) in 1970 was 83.4 (100 is the maximal grade). Not bad compared with the grades of all the students in the country. In the years that followed, the average grade of the students of ASU declined each decade, from 83.4 in 1970, dropping down to 82.1 in 1980, and dropping further to 79.1 in 1990. Finally, in 2000 the ASU's local newspaper proudly announced that grades of ***all*** the students of ASU were *above the average*.

Is that good news for a change?[25]

### 1.9.2 How can one increase the average IQ of the professors in two universities?

In another publication, it was reported that in Highiq State University (HSU), the average IQ of university professors was 130. In that same publication, it was also reported that the average IQ of university professors at the Lowiq State University (LSU), which is located in the next town, was only 80. (Please do not take any offense if you belong to LSU. The numbers I quoted here are purely fictitious).

Two years ago, a professor from the HSU took a position in LSU. At the end of the academic year, it was published that as a result of this transfer of the professor, the *average* IQ of the professors of *each* of the two universities had increased!

Could this be possible?

Yes, it is possible. You can easily find an example.[26]

Does the increase of the average IQ of the professors of each university imply that the average IQ of *all the professors* in the two universities had also increased?

No, that is not possible. Can you explain why?

### 1.9.3 Average speed and average of two speeds

You drive from Jerusalem to Tel Aviv at a constant speed of 40 km per hour. You drive back from Tel Aviv to Jerusalem at constant speed of 100 km per hour. What is the average speed in the round trip to Tel Aviv and back?[27]

If you have difficulty in answering the previous question try the following "easier" one, but with a surprising result.

Suppose you travel from Jerusalem to Tel Aviv on a donkey. The speed of the donkey is $v$ (disregard the units). On the way back you fly with nearly the speed of light, call it $c$. You know that $c \gg \mathbf{v}$. What is the average speed of the round trip, and what is the average of two speeds? A rough estimate will be accepted.[28]

This section was introduced to "formalize" the concept of average (or mean, or expected value). In fact, we have used this concept several times in this chapter. We shall further encounter average quantities in the next two chapters.

## 1.10 Do animals have a probability-sense?

At this stage of reading the book I am confident that you know enough about the concept of probability so that you will be able to understand the two concepts of entropy and the Second Law.

I am confident because I have taught probability theory for many years, and I also know that researchers have come to the conclusion that young children aged 12-14 already possess a probability-sense.

During all this time, it never crossed my mind to ask myself the question posed in the heading of this section. The purpose of this section is to share with you some of my thoughts, as well as some of my findings regarding the question posed above.

You do not need to read this section in order to understand the rest of the book. However, if you are curious – as I was, please read the rest of the section. I believe that aside from satisfying your curiosity, perhaps reading this book will prepare you to ask new questions, design experiments, and conduct research leading to interesting results.

Let me suggest three questions for you to ponder about:

1. Do you believe that animals (those which are considered intelligent) have a probability-sense?

2. Can you think of a possible experiment, the results of which will provide the answer to the first question?

3. Can you design an experiment to measure to what extent some animals make probabilistic judgements, or have a probability-sense?

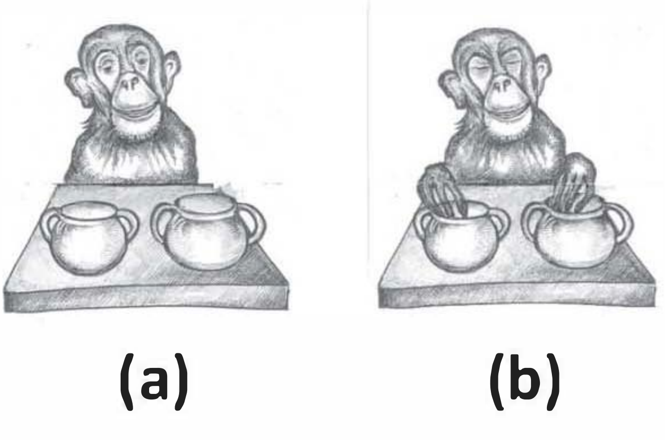
I urge you to pause and think about these questions. Do not rush to read the rest of this section, or skip it, and continue to the next chapter. I also suggest that you write short answers to these questions. This will train you to do original thinking and research. You will realize that sometimes asking the right questions, and designing the right experiments are the most important and exciting activities in any scientific research.

Before I tell you about an experiment which was actually carried out with animals, think of the experiment described in Section 1.7 on children's perception of probability. Can one repeat the same experiments with animals instead of young children?

My answer to the first question is that perhaps some animals have a certain degree of probability-sense. The reason for my belief is that animals, like humans must make probabilistic judgements in their search for food (or prey), finding mates, etc. A hungry lion who seeks an easy prey "knows," based on his prior experience that the chances of finding a deer near the lake are much larger than in arid areas. Similarly, a male seeking for a female mate who happens to be close to another male must assess his chances of winning in case he has to fight the second male.

The main difficulty in conducting a research to measure the extent of having a probability-sense is how to explain to an animal the "rules of the game" as in the case of children who have to choose between two urns having different numbers of colored marbles.

A group of scientists in Germany[29] did this type of research and reported their results in an article entitled: "Apes are intuitive statisticians." Of course, apes or any other animals are not "statisticians," in the sense we use them for professional statisticians. What the researchers did was to design an experiment to check the extent of probability-sense of apes.



If you tried to answer questions 2 and 3 posed above, you might think of a set-up similar to Figure 1.24(see color centerfold) in which an animal has to choose between two urns. (Of course, you might have invented a completely different set-up for such an experiment. If that is the case, I will be delighted to hear from you).

Obviously, it would be difficult to do exactly the same experiment with marbles in urns as described in Section 1.7. It would be difficult, though not impossible, to train monkeys to choose, say a blue marble which will reward them a prize, and to select the urn which has a relatively larger ratio of blue to red marbles. The more difficult part would be to train them to do the selection of the right urn, i.e. the urn with a better ratio of blue to red while blindfolded.

The researchers have solved this problem in an ingenious manner. First, instead of blue and red marbles, they put banana pellets and carrot slices in transparent buckets. The banana pellets were found to be the preferred food for all the ape subjects (chimpanzees, orangutans, and bonobos).

The replacement of marbles with different appealing pieces of food eliminated the need to explain to the subjects that a blue marble will reward them, while a red marble would not. Here, the subjects could *see directly* the rewards, not through the intermediate marble. The more clever aspect of the experiment is to eliminate the need to instruct the animal to choose the right urn while it is blindfolded (after seeing the contents of the two urns). The ape was first shown the two urns, each having the same total number of slices of bananas and carrots, but with varying ratios.

After the animals examined the contents of the two buckets (you do not need to explain to them which the preferred color is, they see the preferred food without explanation). The experimenter drew one slice of food from each bucket in such a way that the animal could see from which bucket each slice was taken, but the slices in the experimenter's hands were not visible to the animal, Figure 1.24c (see color centerfold).

After the experimenter drew two slices, the subject knew which hand drew from which bucket. The subjects were then offered closed fists, concealing the contents of the fists to the subjects. This brilliant trick removed the necessity of training the animal to draw the slice while in a blindfold. Instead, the animal had to choose from two closed fists. Sometimes the animal pointed to both of the experimenter's hands, but after some training they learned to choose only one hand.

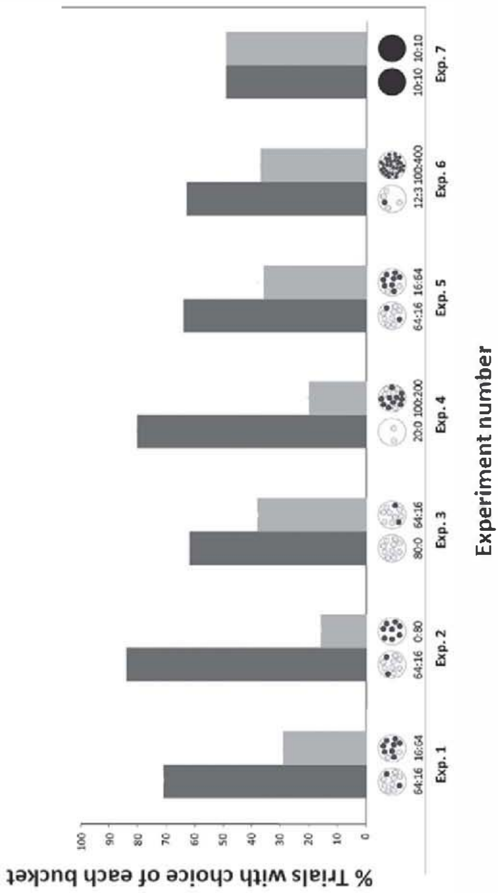


Figure 1.25. Results of the different experiments discussed in Section 1.10.

Once the animal pointed with their finger as to the hand of their choice, they instantly received the food hidden as a reward.

The results are shown in Figure 1.25. There are altogether seven experiments, and the percentage of trials in which the subjects chose the correct/incorrect buckets (shown in dark and light grey, respectively).

The contents of each bucket is shown below the graphs. The numbers refer to the banana to carrot ratio.

Let us go through the results and see what they imply about what the author refers to as "intuitive statistician."

Experiment 1: left 64:16 right 16:64

In this case it is obvious that most subjects chose the left bucket; it contains larger, absolute number of banana pellets, as well as larger probability of obtaining the banana. About 70% got the left bucket (which is the right choice).

Experiment 2: left 64:16 right 0:86

Again, this is a simple experiment (for the animal). Indeed, over 80% chose the left bucket. Explain why.

Experiment 3: left 80:0 right 64:16

This is the simplest case, there are no carrots on the left.

Experiment 4: left 20:0 right 100:200

This one is a little more difficult. In all of the previous experiments the subject could choose the left bucket for the wrong reason, i.e. because of the larger absolute number of the favored food. In experiment 4, the absolute number of banana pieces on the right hand bucket is much larger than in the left one. Therefore, those who choose by the absolute number would have chosen the right bucket. However, since the left bucket contains only banana pellets, it is more attractive to the subject. Indeed, most subjects chose the left bucket.

Experiment 5 is identical to experiment 1. This was designed to rule out the "clever Hans effect" which we will not discuss here.

Experiment 6: left 12:3 right 100:400

Again, this is a more difficult choice. The right bucket has *more* banana pellets, but the ratio on the left is overwhelmingly larger than the ratio on the right. Therefore, most subjects chose the left.

Experiment 7: The two buckets contain equal amounts of the two food choices, and we see that on average about 50% chose the left, and 50% chose the right bucket.

Note that while experiments 4 and 6 may be considered more difficult relative to experiments 1, 2, and 3, they are not as difficult as the ones presented to children as we discussed in

Section 1.7. My suggestion for a more difficult choice is:

left 6:4 right 10:9

Can you explain why I consider this one more difficult than either experiment 4, or experiment 6?

## 1.11 Summary of Chapter 1

In this chapter, we discussed the concepts of probability, probability distributions, and averages. I hope you were convinced that you already possess a probability-sense which means that you have an intuitive understanding of these concepts even though you might not have learned the theory of probability. As you will see in the next chapter, these concepts are essential in order to understand the concept of the Shannon measure of information (SMI) which is discussed in the next chapter. The SMI will turn out to be essential in understanding entropy, which shall be discussed in Chapter 3. As we shall see, entropy may be interpreted as a special case of SMI. However, I want to tell you right now that in order to understand the Second Law of Thermodynamics, you do not need to know anything about either entropy or SMI. The only thing you will need to understand the Second Law is probability.

# CHAPTER 2

# SHANNON'S MEASURE OF INFORMATION (SMI)

*In this chapter we shall learn one of the most important concepts in modern science. I say "modern science," and not in any specific science because this concept was developed in the "Mathematical Theory of Communication," but it is useful in many branches of science. In this book, we bring this concept as the precursor for the concept of entropy.*

Claude Shannon

*Shannon's measure of information* (SMI) *was developed by Shannon in 1948 in connection with communications theory. Shannon (1916-2001), was an American mathematician, electric engineer and cryptographer. He is considered as the father if information theory. Shannon was interested in transmitting information along* *channels. He sought a measure of the information, and he found a specific measure of a specific type of information.*

*In this chapter, we shall define the* SMI, *examine some of its properties and interpretations. Two of the central interpretations of SMI are:*

*A measure of the average uncertainty, and a measure of information, both associated with a probability distribution. We shall see that even young children have a sense of such uncertainty, or of such a measure of information.*

*It is my conviction that familiarity with the SMI is indispensable for understanding the concept of entropy. As we shall see in Chapter 3, the entropy is proportional to a special case of* SMI.

## 2.1 Your uncertainty-sense

In Section 1.1, we discussed your probability-sense. I hope you were convinced that you already have a rudimentary sense of probability. In order to calculate probabilities of various events you need to learn more about probability theory. However, to understand entropy and the Second Law you do not need more than the basic concept of probability.

In this section, we discuss another intuitive sense that you have, which I will refer to as the uncertainty-sense. Of course, you are uncertain about many things; about the weather tomorrow, or whether or not you will understand this book, or your uncertainty could be about the meaning of entropy, or any other matter. We shall discuss one *specific kind of uncertainty*. We have mentioned this kind of uncertainty at the end of Section 1.1. I want to emphasize here that when we say that the Shannon measure of information is also a measure of *average uncertainty,* we mean uncertainty about a specific distribution, not any uncertainty.

Consider an *unfair* die. This means that the various outcomes: 1, 2, 3, 4, 5, 6 have different probabilities. A possible probability distribution is:

| | | | | | | |
|---|---|---|---|---|---|---|
| Outcomes: | 1, | 2, | 3, | 4, | 5, | 6 |
| Probabilities: | 0.55 | 0.1 | 0.1 | 0.1 | 0.1 | 0.05 |

You noticed that the outcome "1" has the largest probability of occurrence. The outcome "6" has the smallest probability. This can happen if the face having six dots is much heavier than all the other faces, as in Figure 1.2b. Therefore, whenever we throw this die it will fall with high probability with the number

"1" facing upwards. There is a small probability that it will fall on all other faces.

You can say that you have a large degree of *certainty* about the outcome when throwing this die. If I do not show you the die you also might have an uncertainty regarding the color, the size, the form, and many other properties of the die. In this and the following sections we shall discuss a very special kind of uncertainty; the average uncertainty regarding the occurrence of outcomes of the experiment, given the *entire* probability distribution of this die, or any other experiment.

Remember the games with the seven different dice that we played in Section 1.1? At the end of Section 1.1, we discussed very briefly one kind of average uncertainty. In all of those dice, we only had two outcomes; blue and red.

If you do not remember those games we played in Section 1.1, please re-read the section to refresh your memory. Recall that I chose a die (*a* to *g*), then you had to choose a color. I tossed the die a hundred times, and every time the color of your choice came up, you gained a dollar. In those games, you noticed that there are different degrees of certainty with respect to the choice of the color (you had complete certainty in case *a*, or *g*, less certain in *b*, or *f*, and least certain in case *d*).

At the end of Section 1.1, we summarized all the probabilities and the average certainties (or uncertainties) of all the games in a very qualitative way.

In this chapter we place the concept of *uncertainty* in the focus of our discussion. This will not be done in a qualitative – intuitive manner, as we did in Section 1.1, but rather in a more quantitative way. However, before proceeding to the quantitative discussion I want to convince you that you already have an uncertainty-sense, in the specific sense mentioned above.

We use again the seven dice shown in Figure 1.1. In Section 1.1, I chose a die and you had to choose a color. By choosing the right color in each game in Section 1.1, you have shown that you have a probability-sense.

Now we use the same dice as in Figure 1.1, but we shall modify the games we play. The rules of the modified games are as follows:

I choose a *color*, say blue, and then you choose one of two dice I select from Figure 1.1. Then, I (or you, if you wish) throw the die you chose a hundred times. Whenever a blue color appears you gain a dollar, when red appears you pay a dollar. Ready?

***First game:***

I choose the blue color, and offer you to choose one of the two dice; *a* or *b*. Which one will you choose?

Obviously, you will choose die *a*. I will not ask you why you chose *a*. Instead, I will ask you to explain in terms of certainty (or uncertainty) why you made that choice.[1]

***Second game:***

I choose the color blue, and offer you to choose one of the two dice; *b* or *c*. Which die will you choose? Can you explain your choice using the term certainty (or uncertainty)? Remember, when we played the games in Section 1.1, I chose the die, and then you chose the color. The choice you made (hopefully, the correct choice) was based on the *relative probabilities* of the two events; blue or red for that specific die. In that game, your probability-sense guided you in choosing the better color.

Now you have to choose between two dice; *b* or *c*. Each of these dice has a different probability distribution. See third and fourth columns in Table 1.2. Therefore, your choice should be guided by the *entire* probability distribution of each die. As we noted in the fifth column in Table 1.2, you are *more certain* (or less uncertain) about your earning in the case of die *b*, than in die *c*. We can say that the probability distribution of die *b* $\left(\frac{5}{6},\frac{1}{6}\right)$, provides more certainty (or less uncertainty) than the distribution of die *c* $\left(\frac{4}{6},\frac{2}{6}\right)$.

We can also say that the ***average uncertainty*** in die *b* is less than the average uncertainty in die *c*. Another way of saying the same thing is that the ***average unlikelihood*** (about the outcomes) in die *b* is less than the average unlikelihood in die *c*. Later, we shall also say that in case *b*, you are provided with more information than in case *c*. But now, let us proceed to the next game.

***Third game:***

I choose the color blue, and I offer you to choose one of the two dice; *c* or *d*. Which die will you choose? Explain your choice in terms of certainty or uncertainty.

This is an easy choice, however, the implication of this choice is most profound in understanding your uncertainty-sense, as well as the measure of information which will be discussed later in this chapter.

Clearly, in playing die *c* you are less certain about your earning than in *b*, and in *b* less than in *a*. However, in playing with die *c*, you are more certain than in die *d*. Of course, you can choose die *c*, and after 100 throws you will lose money. Remember, there is a small chance that all throws will result in red, and you will lose $100.00. You can also choose die *d*, and earn money. You can even earn $100.00, if in each throw the outcome will be blue. However, these events have extremely low probability.

By choosing *c* rather than *d*, you exercise your uncertainty-sense. You will on average earn more money by playing with the die *c*, than with the die *d*.

We can say that the probability distribution in die *c* provides more certainty (or less uncertainty) than in die *d*.

For the next few games I choose the red color to be the gaining one. All the other rules are unchanged.

***Fourth game:***

I choose the red color, and offer you to play either with die *d*, or *e*. Which die will you choose?

Remember that the color red is the winning color. Compare this game with the third game.

***Fifth game:***

I choose the red color, and offer you to play either with die *e*, or *f*. Which die will you choose? Explain your choice in terms of certainty or uncertainty. Compare with the second game.

***Sixth game:***

I choose the red color, and offer you to play with either die *f* or *g*. I will not ask you which die you will choose, but I will ask you to explain your choice in terms of certainty or uncertainty. Also, if you wish, compare this with the first game.

In each of the last three games I asked you to compare the game with one of the previous games. We started with die ***a***, where we had maximum certainty about the outcomes (or minimum uncertainty), then the extent of certainty decreases from ***a*** to *b*, to *c* to ***d***. Once you get to ***d***, you have minimum certainty or maximum uncertainty with respect to the outcomes. The extent of certainty will now increase from ***d*** to *e*, to *f* to *g*. At *g*, you have again maximum certainty or minimum uncertainty. Qualitatively, the extent of certainty changes as in the graph in Figure 1.6.

If you have correctly chosen the better die in each game, you have demonstrated that you have an uncertainty-sense. Let us proceed to the related concept of the amount of information contained in, or belonging to a probability distribution.

## 2.2 The amount of information contained in a probability distribution

In this section we shall discuss the amount of information one gets by asking a binary question. A binary question is answerable by Yes, or No.

For instance: What is your name? This is not a binary question. You do not answer this question with a Yes, or a No. At least, I never gave this kind of answer.

Is your name, Dan? This is a binary question. I can answer either with a Yes, or a No. Now, we want to quantify the amount of information given to a binary question. Consider the following questions I ask you, and to which you will answer with a Yes, or a No.

Q1: Will it rain tomorrow in New York?

A1: Yes.

Q2: Did you enjoy reading this book?

A2: Yes.

Q3: Will the sun rise tomorrow morning?

A3: Yes.

Q4: Did die *a* (in Figure 1.1) fall with the blue face up?

A4: Yes.

Q5: Did die *b* (in Figure 1.1) fall with the blue face up?

A5: Yes.

Q6: Did die *d* in Figure 1.1 fall with the blue face up?

A6: Yes.

Let us assess qualitatively the *amount* of *information* I obtained from you when I asked you these questions, and your answer was a Yes.

In the first question, I had no previous information (in the colloquial sense of the word) what the weather in New York will be tomorrow, so I feel that I got a lot of information from your answer.

In the second question, I believe that you have enjoyed reading this book (so far, at least), but I am not sure about it. Therefore, by answering with a Yes, you have confirmed my guess. Therefore, you gave me some information, perhaps less than in case of question 1.

In question 3, I am almost sure (not absolutely sure!) that tomorrow morning, the sun will rise. Therefore, with your "Yes" answer, you did not give me much new information.

So far these estimates of the amount of information I got by asking a binary question were very qualitative. Let us be more quantitative.

In asking question 4, I already knew that all the faces of die *a* are blue, therefore, by giving me the answer Yes to my binary question, you did not provide me with any new information. I knew that the answer must be a Yes, therefore, no information was given to me by asking question 4.

In asking question 5, I know with relatively high probability that the answer will be a Yes. But I am not totally sure. It might also be a No. Therefore, by answering with a Yes, you gave me some information. Certainly more than in answer 4.

What about question 6? Here, I am totally uncertain about the outcome by throwing die *d*. By giving me the answer Yes (or No), you gave me the *maximum* information for a binary question. The amount of information given to me in answer 6 is called one *bit* of information.

Exercise: Note that in the list of questions, I skipped die *c*. As an exercise, suppose that I asked the same question, and you can answer with a Yes for die *c*. Can you tell how much information was given to me from this answer?

Two comments are now in order. First, the term *bit* is short for *bi*nary dig*it*. In transmitting message over channels, the term bit is defined as either "1" or "0." Thus, if we transmit a sequence of zeros and ones such as:

1, 0, 1, 1, 0, 0, 1, 0, 1,0,

we have transmitted 10 symbols (of "1s" and "0s"), and therefore we have transmitted ten bits.

In information theory, and in this book, we use a *different* meaning of the *bit*. This is the amount of information given to a binary question, when we know that the two possible answers have *equal probabilities*.

The condition of "equal probabilities" is essential. You might find in many popular science books a "definition" of the *bit* as the amount of information given to a binary question.

This definition is obviously incorrect. To understand why the qualification of "equal probabilities" is essential, we shall consider the following examples of binary questions:

Table 2.1 shows seven coins. For each coin, we give its probability distribution; Pr (H) and Pr (T) = 1 − Pr(H). As you will notice, I have chosen the same distribution in Table 2.1 as in Table 1.1. In fact, I could use the same dice of Section 1.1

for this discussion. However, I decided to choose the language of coin distribution, first; because we had enough of the discussion on dice having two colors, and second; to show you that the discussion in this section does not depend on the particular system we use (game with dice or with coins, or any experiment having only two possible outcomes with equal probabilities).

Shannon showed that the amount of information given to a binary question about any experiment having two outcomes, say H and T with probability Pr (H) and Pr (T), respectively, is given by the formula:

$$\text{SMI(for binary question)} \\ = -\text{Pr(H)}\log_2\text{Pr(H)} - \text{Pr(T)}\log_2\text{Pr(T)}$$

We can denote by $p = \text{Pr(H)}$, hence $\text{Pr(T)} = 1 - \text{Pr(H)} = 1 - p$, and this formula is actually a function of one parameter $p$ which ranges from zero to one.

$$\text{SMI}(p) = -\text{p}\log_2 p - (1 - p)\log_2 p$$

Here, $\log_2 p$ is the logarithm to the base 2. If you do not know what logarithm means, you can look at Appendix A, Figure A.2. For any $p$ we plot the corresponding value of $\log_2 p$. For instance, for $p = 1, \log_2 p = 0$. For $p = 0.5, \log_2 p = -1$.

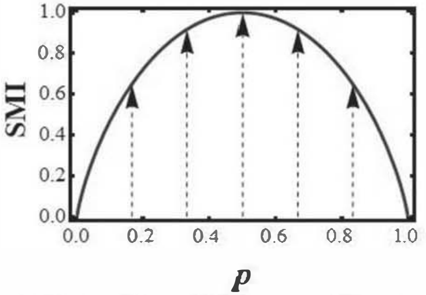


Figure 2.1. The values of the SMI for the different dice in Table 2.1.

Some further discussion of the logarithm function is given in Appendix A. The values of SMI ($p$) are also shown in Table 2.1, and plotted in Figure 2.1[2]

**Table 2.1: Probabilities and Shannon's measure of information for coins with different probability distributions**

| Coin | Pr (H) = $p$ | Pr (T) = 1-$p$ | SMI ($p$) |
|---|---|---|---|
| a | 1 | 0 | 0.00 |
| b | 5/6 | 1/6 | 0.65 |
| c | 4/6 | 2/6 | 0.92 |
| d | 3/6 | 3/6 | 1.00 |
| e | 2/6 | 4/6 | 0.92 |
| f | 1/6 | 5/6 | 0.65 |
| g | 0 | 1 | 0.00 |

We now play a simple game with the coins in Table 2.1. I tell you that I threw one of the coins in Table 2.1 I also tell you the probability distribution of this coin, i.e. the corresponding Pr (H) and Pr (T). After I threw the coin you have to make a guess as to what the result was. If the answer is Yes, you get a prize, if the answer is No, you get nothing.

Now I throw each of the coins in Table 2.1. What is your guess regarding the outcome? Clearly, in case *a*, your guess will be H, and the answer will be a Yes with certainty. In this case, you already know the outcome from the knowledge of the probability distribution of *a*, therefore by giving you the answer Yes, I did not give any information. We can also say that the amount of missing information is zero.

In case *b*, you should also guess H, and there is a good chance that the answer is a Yes. In this case, you "almost" know the answer, therefore you get a little more information than in case *a*. The amount of information you get is given in the fourth column in Table 2.1, and in Figure 2.1 In case *c*, you get more information than in case *b*, and in case *d*, you get the maximum amount of information, which is one *bit*. We can also say that in case *d*, the amount of missing information is one bit. Once you go to coins *e*, *f*, and *g*, the amount of information decreases until you get zero information in case *g*.

In the following sections, we shall generalize the concept of SMI (as well as the concept of average uncertainty and average unlikelihood) for the general experiment with any number of outcomes. Before doing so we shall spend some time playing the familiar twenty-question (20Q) game. Playing these games will convince you that you already have an intuitive sense of what the SMI means. Or, if you prefer, you can call it informational-measure-sense.

## 2.3 Forget about SMI and let's play some 20Q game

In this section, we will play the well-known parlor game of "20 questions" (20Q), Figure 2.2. This game was very popular on radio in the 1940s, and later on, on TV in the 1950s. We begin by playing the "Person" version of this game.

I choose a person from a given list, and you have to find out who I have chosen by asking binary questions only. Remember, a binary question is one which is answerable only with a Yes, or a No. Suppose, I chose a person, say Einstein, and you have to find out which person I chose. The rules of the games are;

you have to pay me $1.00 for each answer you receive. When you discover the person I chose, you will get $20.00.

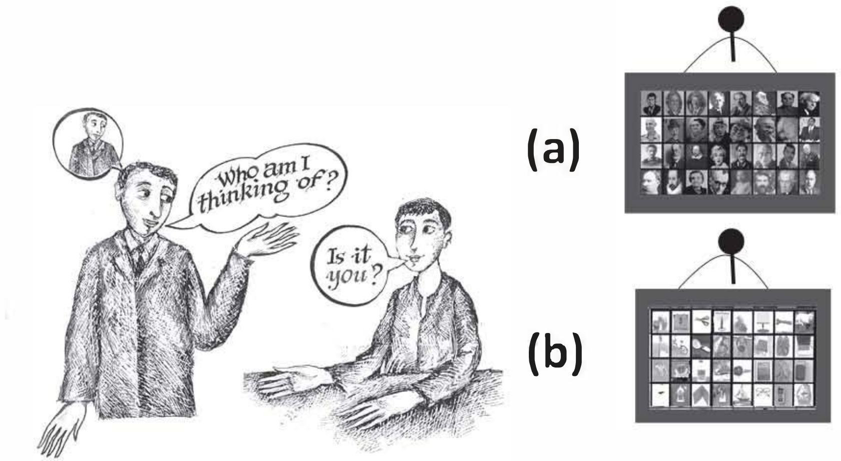


Figure 2.2. Two persons playing the 20Q game either with, (a) 32 people, or (b) 32 objects.

In this game, you know what you do not know; the person I have chosen. you also want to acquire that unknown information (or the missing information), by asking the fewest number of questions. Otherwise, you will be spending more than $20.00 to get the $ 20.00 prize when you discover the person I chose. Therefore, you have to carefully plan your *strategy* of asking questions.

You might be wondering why 20Q? Why not 25 questions, or 100 questions? The answer will probably surprise you.

Twenty questions are more than enough for all practical purposes. By "practical purposes" I mean games in which I choose an object, a person, an animal or whatever, from a pool of objects that we both know. Obviously, I cannot choose someone you never heard of. Therefore, in practice we implicitly limit the range of persons to those that we are both familiar with.

Suppose you play this game with me. How many persons do you think we could find that we both know, perhaps 1000, 10,000, or 100,000? I bet you cannot find more than 100,000 names to choose from.

If you are convinced that the "size of the game" is not larger than 100,000, or even 1,000,000, then you will be surprised to learn that 20 questions are more than enough to win the game (i.e. to spend less than $20.00 on questions to get the $20.00 prize). In order to achieve this, you only have to be smart! If you are smart enough to play this game correctly then you will also be able to understand the Second Law, as well as entropy.

Let us go back to the 20Q game. Remember, I chose a person, say from a group of persons, such as those in Figure 2.3. You have to find out which person I chose by asking binary questions.

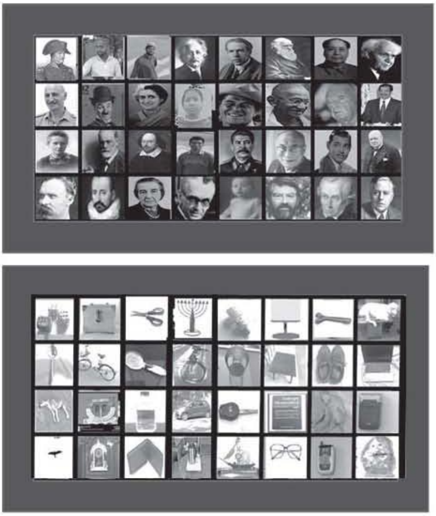


Figure 2.3. (a) The 20Q game with 32 people, and (b) with 32 objects.

There are many possible *strategies* of asking questions. The simplest is to *guess* the name of the person you believe I have chosen. For instance, you can ask: "Is the person Einstein?" Another method is to ask if that person has some physical features,– say, "does the person have blue eyes?" Another strategy is similar to the previous one, but now you divide all possible persons into roughly two halves, and ask: "Is the person a male?" or "Is the person alive?"

What strategy will you choose to play this game?

If you choose the first strategy, you might hit the right person on the first question and win. However, since there are so many possible persons to choose from, your winning on the first question, although *possible* is extremely unlikely. This means that you *can* win after one question, but the probability of winning is very small. Note also that the larger the group of persons that we agreed upon, the harder your task is. Guessing the name of the person would be an inefficient way of asking questions. The reason is that with each No answer that you get, you have eliminated one person. This means you are still left with almost the same *missing information* that you began with (i.e. the original number of unknown persons minus one). As we shall see later on, very young children do indeed choose this strategy as they are apparently impatient to receive the prize and they feel that this is the only way a Yes answer will terminate the game in their favor and thereby win the prize. This "strategy" of asking questions is referred to as the "dumbest" strategy.

However, if you choose the second strategy, you cannot possibly win on the first question. If you ask, "Is the person living in London?" and you get a Yes answer, you still have to continue asking questions until you establish the identity of the person. If you get a No answer, which is more probable, you exclude all people residing in London. This is certainly much better than excluding only one person as done in the first strategy.

At this stage you feel that the best strategy is to divide all possible persons roughly into two groups. Again, you cannot win on the first question, or on the second, or the third. However, at each step of the game, whatever answer you get whether it be a Yes, or a No answer, you exclude a very large number of people, and you narrow down your range of possibilities into almost half of the original number of possibilities.

Intuitively, you feel that in choosing the "smarter strategy" you gain more *information* from each answer you ask. Therefore, you get the maximum information for each dollar that you spend. It pays to be patient and choose the best strategy rather than rushing impatiently to guess the right person.

Exercise: Look at Figure 2.3. There are some pictures of famous people and some not so famous. Suppose that we agree that I choose a person only from this collection. Is there a smartest strategy in this case? You either say "is there a smart strategy in this case," or "What is the smartest strategy in this case?" What is the minimal number of questions that you need to *guarantee* that you will get the required information?

Remember, this chapter is not about the 20Q game itself but about a *measure* of the game. Roughly, the size of the 20Q game can be measured by the amount of money you will spend in asking questions. Qualitatively, the larger the pool of persons from which we agree to choose from, the larger the size of the game. If we agree to select only a person from a certain city, then the size of the game is smaller compared with the game where we agree to choose a person from a country. If we agree to choose only persons who are in the same room where we are attending the party, then the size of the game is much smaller.

Now that you know how to play the 20Q game correctly, you have implicitly admitted that you know what information means, and that this information has some measure or size. The "*information*" in this game is "which person I have chosen." You also know that if there are more people from where I specifically choose one for the game, it will be *harder* to find that person.

## 2.4 Strategies for playing the 20Q game

We start with the simplest 20Q game. Instead of choosing a person from an unspecified number of people, we have eight boxes (Figure 2.4). I hide a coin in one of the boxes and you have to find where I hid the coin by asking binary questions.

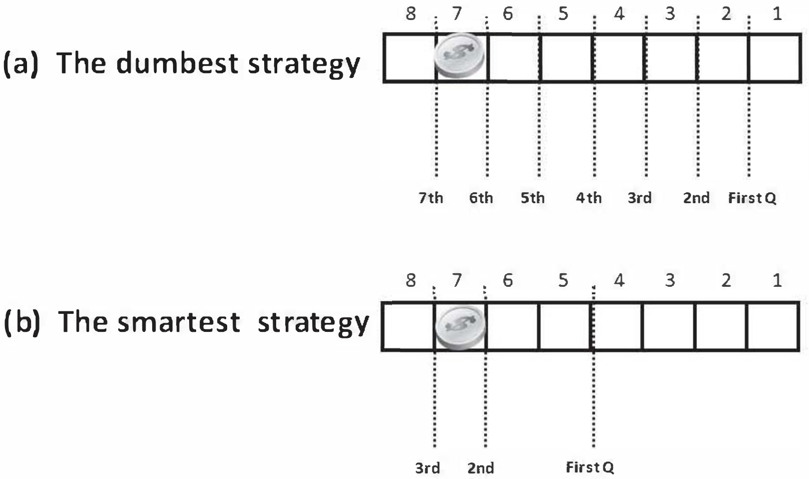


Figure 2.4. Two strategies of asking questions, (a) The dumbest, and (b) The smartest.

In this game, the *information* you are seeking to find is "where the coin is." What we are interested in is not the *information* itself, but some measure of the *size* of the *missing information.* Before we choose a measure of the size of the missing information, we note that the more boxes there are the more difficult it will be to find the missing information. What is meant by "more difficult" is that we need more questions to ask in order to obtain the missing information. One way to measure the size of the game is simply by the number of boxes. Clearly, the larger the *N*, the "bigger" the problem and the more questions we shall need to ask. This measure is fine, but it will be difficult to generalize for the case of non-uniformly distributed games which we shall discuss in Section 2.6.

Another measure we can adopt is the number of questions we have to ask in order to find where the coin is. However, there is a difficulty with this measure. We already know that the number of questions depends on the *strategy* we choose in asking questions.

Because of this difficulty, let us devote some time to playing this game with different numbers of boxes.

We start with eight boxes ($N = 8$). We can choose many strategies in asking questions. Figure 2.4 shows the two extreme strategies. We call them "the dumbest" and "the smartest" strategies. The dumbest is to make a specific guess where the coin is, and ask "Is the coin in box 1?" Is it box 2?" and so on.

I should also mention here that in playing this game, you are informed about the total number of boxes, and also that I have no preference for any specific box which means that I have randomly placed the coin in one of the boxes; that is, each box has a probability of 1/8 of containing the coin. Note that in this game we have completely removed any traces of subjectivity. You cannot use any information you might have about me or about my personality to help you in finding or guessing where I placed the coin. The information we need here is "where the coin is." The "hiding" of the coin can be done by a computer which chooses a box at random.

The strategies here are well defined and precise, whereas in the general 20Q game I could not define them precisely. In this game, with the dumbest strategy you ask, "Is the coin in box $k$ ?" where $k$ runs from one to eight. The smartest strategy is different: Each time, we divide the entire range of possibilities into two halves.

Note that in this case I use the adjectives "dumbest" and "smartest" strategies. The reason is that here one can *prove* mathematically that if you choose the *smartest* strategy and play the game many, many times, you will outperform any other possible strategy, including the worst one denoted the "dumbest." Since we cannot use the tools of mathematical proof, let me try to convince you why the "smartest" strategy is far better than the "dumbest" one (and you can also "prove" for yourself by playing this game with a friend or against a computer).[3]

Qualitatively, if you choose the "dumbest" strategy, you might hit upon the right box on the first question. But this could happen with a probability of $\frac{1}{8}$ and you could fail with a probability of $\frac{7}{8}$. Presuming you failed on the first question (which is more likely and far more likely with a larger number of boxes), you will have a chance of a correct hit with a probability of $\frac{1}{7}$ and to miss with a probability of $\frac{6}{7}$, and so on. If you miss on six questions, after the seventh question you will *know* the answer; that is, you will have the information as to where the coin is. If, on the other hand, you choose the "smartest" strategy, you will certainly *fail* on the first question. You will also fail on the second question but you are *guaranteed* to have the required information on the third question.

The qualitative reason for preferring the smartest strategy is the same as in the general 20Q game (but now can be made more precise and quantitative). By asking, "Is the coin in box 1?" you might win on the first question but with very low probability. If you fail after the first question, you have eliminated only the first box and decreased slightly the number of remaining possibilities, from eight to seven. On the other hand, with the smartest strategy the first question eliminates ***half*** of the possibilities, leaving only four possibilities. The second question eliminates another half, leaving only two, and in the third question, you get the information! We can also say that with the smartest strategy we reduce the "size" of the game each time by half.

In information theory, the amount of missing information – the amount of information one needs to acquire by asking questions – is *defined* in terms of the probability distribution. In this example, the probability distribution is simple: $\{^1/_8, ^1/_8, ^1/_8, ^1/_8, ^1/_8, ^1/_8, ^1/_8, ^1/_8\}$. In asking the smartest question, one gains from each answer the maximum possible information which is one bit. See Section 2.2. One can prove that maximum information is obtained in each question when you divide the space of all possible outcomes in two *equally probable* parts.[4]

You can check yourself that if at each step of the smartest strategy I gain *maximum* information, then I will get the information I want in a *minimum* number of questions. Again, we stress that this is true *on average*; that is, if we play the same game many, many times, the smartest strategy provides us with a method of obtaining the required information with the smallest number of questions.

Note also that the *amount* of information that is required is fixed for a given game, and it is independent of the strategy you choose. The choice of the strategy allows you to obtain the same amount of information with different number of questions. The smartest strategy guarantees that you will get it on average, with the minimum number of questions.

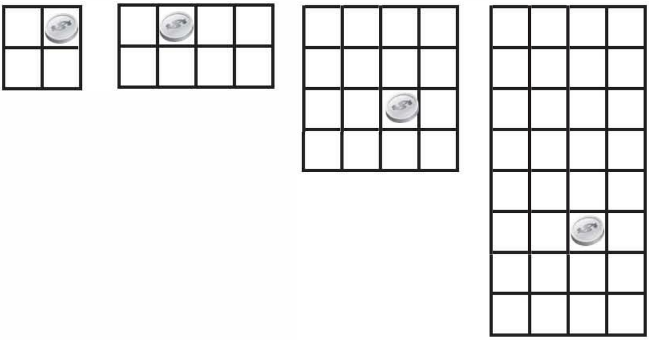


Figure 2.5. A coin hidden in one of *n* boxes, with *n* = 4, 8, 16, 32.

In Figure 2.5 we have several games with different numbers of boxes. Whenever we doubled the number of boxes the number of questions one needs to ask in the smartest strategy is increased by only one! The average number that one needs to ask in the dumbest strategy needs is far larger.

If you are still unconvinced, think of a game having 1,048,576 boxes. Using the smartest strategy, you are guaranteed to find the required information in 20 questions! Can you imagine how many questions you will need, on average, in the dumbest strategy?[5]

The important point to be noted at this stage is that the larger the number of boxes, the greater the amount of missing information, hence the greater the number of questions needed to acquire that information. This is clear intuitively. The amount of information is *determined* by the probability distribution.

If you use the dumbest method, with larger $N$ the average number of questions increases *linearly* with $N$. This means that the average number of questions is proportional to the number of boxes. One can show that for very large $N$, the average number of questions is about $N/2$, (Figure 2.6).

On the other hand, if you use the smartest strategy you will need only $\log_2 N$ questions on the average. Why $\log_2 N$?[6]

We have already seen that if we *double* the number of boxes, the number of (smart) questions increases by only one! This is a very important observation. This is also the reason why we shall adopt $\log_2 N$ as a *measure* of the size of the game, rather than $N$ itself.

Now let us play the 20Q game with money. Suppose that we have 16 boxes, and you pay \$1.00 for each question that you ask, when you find where the coin is, you get \$5.00. Surely you would not like to ask specific questions. If you do so, you will have to ask about eight questions on average, which means you will be paying \$8.00 to earn \$5.00. However, if you choose the smartest strategy then you will be spending only \$4.00 to earn \$5.00.

If this game did not impress you, think of having 1,000,000 boxes. You pay \$1.00 per question but you gain \$1000.00 once you find the coin. You would not even dare to use the dumbest method in this case, since you will lose on average about \$500,000.00! However, astonishing as it may sound to you, with the smartest strategy you will need to spend *less* than \$20.00 to gain the amount of \$1000.00.

If you are astonished by this, do not be intimidated. In fact, with 20 questions you can find the coin not in 1,000,000 boxes but in 1,048,576 boxes. If you do not believe me, try it out yourself. It is very simple. Each time, divide the total number of boxes into two halves – in 20 questions you are guaranteed to find the coin $\left(\log_2 1{,}048{,}576 = \log_2 2^{20} = 20\right)$.

Figure 2.6 shows how the average number of questions depends on the number of boxes in the two strategies. Look how fast the average number of questions increases when you choose the dumbest strategy. Be impressed by the fact that with the smartest strategy, when the number of boxes becomes *huge*, the number of questions you need to ask is large, but still manageable. It is impressive that with the smart strategy, as the number of boxes become huge, the number of questions you need to ask is large, but still manageable.

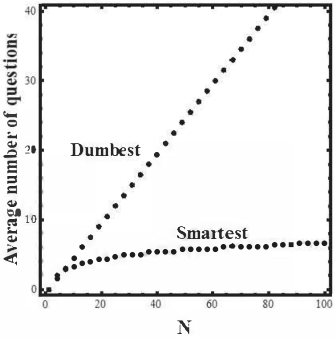


Figure 2.6. Average number of questions in the two strategies.

The number of molecules in a glass of water is about $10^{23}$. If you have this number of boxes you will almost never be able to find the coin with the dumbest strategy. But with the smartest strategy you are guaranteed to find the coin in less than 80 questions!

Let us adopt the logarithm to the base 2 of the number of boxes as a measure of the size of the missing information in the game of the type discussed above. We shall refer to this measure as the Shannon Measure of Information and use the abbreviation SMI. If you are not comfortable with the concept of logarithm, forget it. Just remember that the smartest strategy of asking questions is indeed *the smartest strategy*, and take the average number of questions one needs to ask in the smartest strategy as a *measure* of the size of the missing information.

You should also take note that the type of information for which the SMI is applicable is but a tiny fraction of all possible types of information. We cannot apply this measure for most types of information. For instance, the probability that "a snow storm is expected at 3 p.m. tomorrow in the New York area" cannot be measured by the SMI. However, information theory can deal with the *size* of this message, regardless of its meaning or the information it conveys. This type of SMI is important in the field of communication and transmission of information for which Shannon constructed his measure. For more details, see Ben-Naim (2017b).

## 2.5 How young children play the 20Q game

In Section 1.7, we encountered children who had to decide between two possibilities of unequal probabilities. In this section, we are interested in the way children ask questions to obtain information. This task involves not only a sense of probability, but also choosing the best strategy of obtaining the required information in the most efficient manner. This is what we refer to as information-sense, or uncertainty-sense.

Many investigations were carried out on the way children play the 20Q game. We shall discuss only a few examples.

Children aged 6-11, in grades 1-6 were shown 42 pictures of familiar objects. Each child was first asked to identify the pictures to ensure that he or she was familiar with them. The following instructions were given to each child:

*"Now, we're going to play some question-asking games. I'm thinking of one of these pictures, and your job is to find out which one it is that I have in mind. To do this you can ask any questions at all that I can answer by saying "yes" or "no", but I can't give any other answer but "yes" or "no." You can have as many questions as you need, but try to find out in as few questions as possible."*

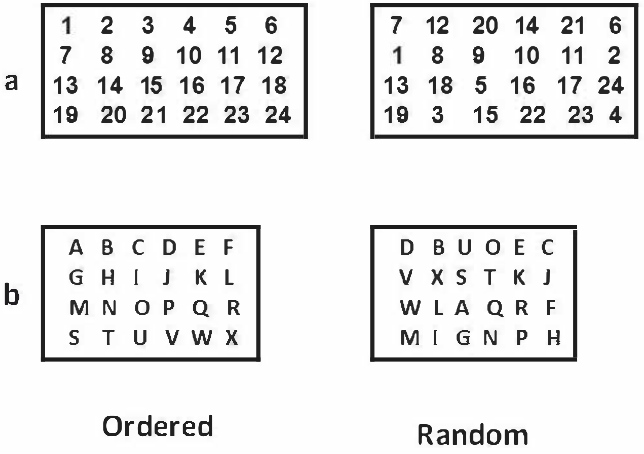
The questions asked by the children were classified into two groups. The first, where the child asked about a *specific* object, such as: "Is the object the dog?" This is what we have referred to as the dumbest strategy. The second, called "constraint seeking," is similar to what we have referred to as the smart strategy (not necessarily the smartest). In this strategy, the child divides the entire range of possibilities into two groups. The best strategy (the smartest) is to divide the entire range into two parts of equal number of objects each time. In practice, the children used some criterion for grouping, such as, "Is the object an animal?" or "Is it food?" Clearly, this is better than the dumbest strategy but short of being the smartest.

As is expected, the first-graders (aged about 6) almost always chose the "specific questions." Out of the 30 children, only 5 asked some constraint questions. From the third-graders (aged about 8), only about one-third asked "specific questions," and of the sixth-graders (aged about 11), almost all asked constraint questions.

No doubt, the choice of the constraint type of question requires some cognitive skills, planning, and patience. Some mature thinking is required to invest in questions that certainly cannot provide them with the required information immediately, but that proves to be more efficient on average. Whereas young children are in a hurry to ask specific questions, hoping to be instantly successful, the older ones invest in thinking and planning before asking.

In another study, children aged 13-14 were successively presented two matrices, each with six rows and four columns.

One matrix contained the numbers 1-24, while the other one contained letters A-X, Figure 2.7.



|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 7 | 12 | 20 | 14 | 21 | 6 |
| 1 | 8 | 9 | 10 | 11 | 2 |
| 13 | 18 | 5 | 16 | 17 | 24 |
| 19 | 3 | 15 | 22 | 23 | 4 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| A | B | C | D | E | F |
| G | H | I | J | K | L |
| M | N | O | P | Q | R |
| S | T | U | V | W | X |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| D | B | U | O | E | C |
| V | X | S | T | K | J |
| W | L | A | Q | R | F |
| M | I | G | N | P | H |

a
b
Ordered
Random

Figure 2.7. Several choices of 20Q games.

Although the game with 24 *numbers* is equivalent to the game with 24 *letters*, the results obtained for the two games were different. Also, the results were different when the numbers (or letters) were arranged in the natural order, or randomly.

Interestingly, the results show that children did systematically better (i.e. fewer questions) on the number matrix (Figure 2.7a) than on the letter matrix (Figure 2.7b). Also, they did better when the numbers (or letters) were in their natural order than when the numbers (or letters) were randomly placed. Can you think of an explanation to this result?

Finally, I want to mention an experiment that I did myself both on children, as well as on audiences in my lectures (which included students and professors). I showed them Figures 2.3a and 2.3b. One contained only pictures of people, while the other picture contained pictures of simple objects. I offered to play the 20Q game on either one of those figures. I told them that I would choose a person or an object at random from either Figure 2.3a or Figure 2.3b. You have to find out which person or object I chose by asking binary questions. You have to pay $1.00 for each answer you get. Once you guess the person or object I chose, you will get a $20.00 prize. Which game will you prefer to play?

Write your answer before you continue and explain why you chose that particular game.

I was not surprised to learn that young children of any age choose preferentially the game in Figure 2.3b. When asked why, they simply answered: "I do not know all the persons in Figure 2.3a, but I recognize all the figures in Figure 2.3b. This is understandable.

I was shocked to learn that even in audiences of about 50 people (students, graduate students, and professors), there were always a few people who preferred the game with Figure 2.3b rather than that of Figure 2.3a. It is shocking because anyone who understands the rules of the games should know that the two games are completely equivalent whether or not they recognize the people, or the objects.

Pause and think. In Section 1.10, I discussed the experiments done on animals to test their sense of probability. While writing this chapter, and in particular the section on children playing 20Q games, I was wondering whether animals can be trained to play this game, if not the game itself perhaps to learn whether animals have or do not have a sense of uncertainty as discussed above.

I checked in Google and asked researchers in biology who do such experiments with animals, but I could not find any research which was specifically designed to learn about the uncertainty-sense of animals.

Can you think of an experiment similar to the ones discussed in this section, but with animals instead of children?

Can you train an animal to play the 20Q game?

If you have a good idea, I will be glad to hear from you. Perhaps, we could convince some biologists to carry out such an experiment.

At this point I would like to emphasize again that when we say "uncertainty" in connection with SMI, we mean a very specific kind of uncertainty. If I ask you, "Do animals have a sense of uncertainty? Your answer should be, "Yes, of course." Animals have many uncertainties; they are uncertain about almost everything they do. But I was asking a question about the average uncertainty in the sense discussed in Section 2.9.1. This kind of uncertainty involves the ability to "calculate" or to estimate an average over many different outcomes. I doubt whether animals have this kind of uncertainty. When we discuss entropy in the next chapter, we shall see that entropy is a very special case of SMI. Therefore, entropy can also be interpreted as an average uncertainty. But now the uncertainty is even more restricted than the uncertainty involved in the interpretation of SMI.

## 2.6 The amount of information contained in a uniform 20Q game

Before we discuss the quantitative relationship between the number of objects, and the minimum number of questions you have to ask in a 20Q game, look again at Figures 2.3a and 2.3b. How many questions do you need to ask in order to guarantee that you will get the required information?[7]

Now, suppose I double the number of objects from 32 to 64. How many questions will you need now? [8]

Let us next discuss a simpler example where the meaning of the SMI (not the meaning of the information) becomes clear and easy to grasp.

Suppose that I show you a board with four equal areas as in Figure 2.8. I throw a dart and I tell you that it hit one of the squares on the board. Your task is to find the square in which the dart hit by asking *binary questions*, that is: I can answer with a Yes, or a No. To make the game more dramatic and more exciting, and perhaps also to encourage you to think harder, let us assume that you pay \$1.00 for each answer you get. When you find the square in which the dart hit you get \$4.00.

If you play this game many times you will find that on average, you earn more money by adopting the smartest strategy instead of any other strategy. In this particular game the difference between the two strategies is not too dramatic. But when we increase the number of areas, say from 4 to 8 to 16, and so on, you will find that on average, the number of questions you will have to ask using the "dumbest strategy" increases with the number of squares roughly as $N/2$, i.e. *proportional* to $N$.

On the other hand, adopting the "smartest" strategy of asking questions; that is, dividing each time the entire $N$ squares into two halves, then again into two halves until you find the dart, the number of questions you need to ask is approximately $\log_2 N$, which is much smaller than $N/2$. You can verify this "law" by checking the number of questions you need to ask in the "smartest" strategy for the following cases, Figure 2.8.

$N = 4 \quad N = 8 \quad N = 16 \quad N = 32 \quad N = 64$

The corresponding number of questions are: 2, 3, 4, 5, 6, respectively.

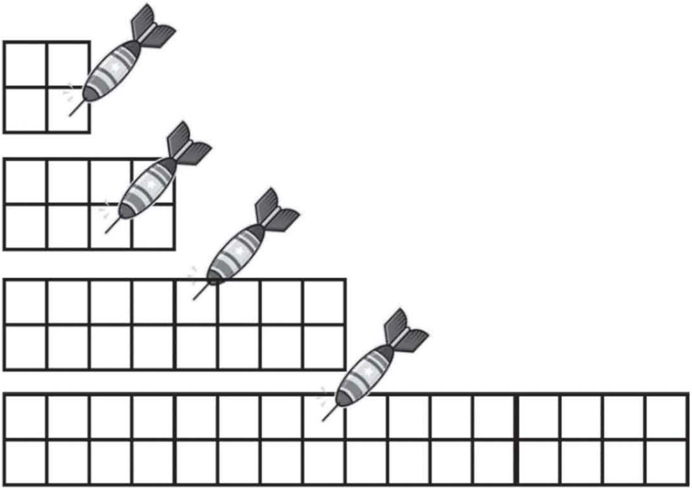


Figure 2.8. A dart hitting a board with different regions.

Note that in this series of games, when we *multiply* the number of squares by 2, the number of questions you need to ask increases only by one! In general, for $N$ of the form $2^{\mathrm{n}}$ ($n$ integer), the number of questions will be $\log_2 N = \log_2 2^n = n$, (provided you play the game "smartly.")

Let us summarize what we have learned so far. You are told that a dart has hit one of the $N$ squares. You also know that the areas of the small squares are equal. Therefore, the probability that the dart hit any specific area is $1/N$. You do not know the *information*: "Where is the dart?" So you ask questions in order to get this *information*. Here, we are not really interested in the *information* itself, but rather on how to get it by spending less as possible; that is, earning the maximum dollars in each game. We find that the number of questions you need to ask if you are smart enough to adopt the "smartest" strategy is of the order of $\log_2 N$.

As an exercise suppose that you pay 1 US$ for each answer you get, and you receive $N/2$ dollars when you find the information (on where the dart hit). How much do you expect to gain on average if you play this game 1000 times by adopting the "dumbest" strategy (i.e. asking about *specific* squares), and the "smartest" strategy (i.e. dividing each time into two equal halves). The games are with the following:

$$N = 8, 16, 32, 64, 2^{10}, 2^{100}, 2^{1000}$$

Up to this point we found the relationship between the number $N$, of equally probable events (the squares in which the dart hit) and the number of questions, for $N$ of the form $N = 2^n$ where $n$ is an integer. We found that

$$number\ of\ smart\ questions = \log_2 N$$

One can show that this relationship is valid for any $N$. We shall not discuss the proof here.[9]

Now that we know the relationship between the number of events $N$, and the number of questions one must ask to find which of the $N$ events has occurred (e.g. where the dart hit), we can ask what this has got to do with *information*.

The information that we do not have is "which event occurred?" The number of questions we must ask is a measure of the *size* of that *information*. The larger the number of events, the larger the number of questions we must ask to obtain this information. Hence, we use this number as a measure of the *size* of the *missing information*.

## 2.7 The amount of information contained in a non-uniform 20Q game

Before we discuss the general case let me offer you again to play the 20Q game, but now with a non-uniform distribution. The conditions are the same as before. Which game would you prefer to play in Figure 2.9, game *a* or *b*? If you have difficulty in choosing, try the simpler choice: Which game would you prefer to play, *a* or *c*? Explain why this choice is simpler.

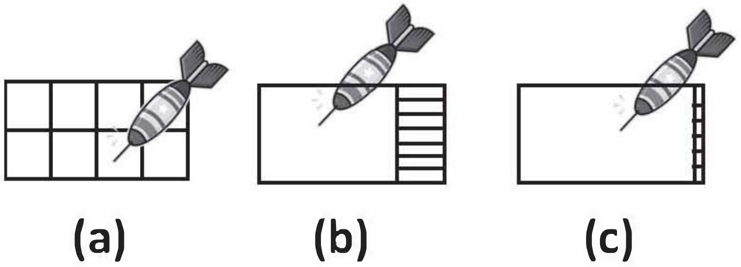


Figure 2. 9. Uniform and non-uniform games.

So far we have discussed events with equal probability, or a game with a *unif•rm* probability distribution. The next step is to study the case of a non-uniform distribution of events. Suppose we have an experiment with *N* outcomes, each with different probability. This defines a probability distribution $p_1, p_2, \dots, p_N$ with $\sum_{i=1}^{N} p_i = 1$. On this distribution we define the SMI as we did above:

$$\mathrm{SMI} = -\sum_{i=1}^{N} p_i \log p_i$$

One can prove mathematically that for any given distribution $p_1, \dots, p_N$, the quantity SMI defined above is equal (up to an accuracy of $\pm 1$) to the average number of questions one has to ask in order to find out which event has occurred by asking binary questions.

Thus, the larger the number of questions we need to ask the larger the size of the information contained in the game. Again, we stress that the *information* itself is not important. It can be the location of a dart on the board, the location of a coin in $N$ boxes, or any other experiment having $N$ outcomes with a given probability distribution $p_1, \dots, p_N$.

Note that in calculating the SMI of the experiment, or of the game, we used only the given distribution and not any details of the experiment or the game. Therefore, it is more appropriate to refer to the SMI as a measure of the information *contained in* or *belonging to* a given *distribution*, leaving the details of the experiment unspecified. The distribution can pertain to dice, coins or cards. It does not matter. The only thing that matters is the probability distribution itself. This is why the SMI is so general, as well as being a useful quantity.

We shall examine here only a few examples. It is convenient to use again the dart hitting a board divided by $N$ regions. But now the areas of the regions are unequal. The probability of hitting the $i$th region in the board having area $A_i$ is assumed to be equal to $p_i = A_i/A$, where $A$ is the total area of the board. The information we need is "where the dart hit" or for a more general experiment "which event occurred?" The size of this information is measured by the minimal number of questions one needs to ask in order to obtain this information.

Sometimes the SMI is referred to a "measure of information," and sometimes as a "measure of the missing information." This distinction is not important. In one, we assign the "information" to the *experiment*, in the other we assign the "missing information" to the one who has to ask the question. Either way this quantity is positive and an objective quantity associated with a given experiment with a given probability distribution.

Consider the following two divisions *a* and *b* of the same board, Figure 2.9.

Both boards are divided into eight regions. In Figure 2.9a, we have eight *equal* areas, whereas in Figure 2.9b, we have eight *unequal* areas. Which of these games is easier to play? Putting it differently, suppose you have to find where the dart hit by asking binary questions. Again, assume that you pay $1.00 for each question you ask, and you will get $10.00 when you find the region in which the dart hit. The question is given the two distributions, in which game can we get the information on "where the dart hit" with the fewest number of questions? The exact answer to this question is *contained* in the values of

SMI for these two cases. The larger SMI, the larger the *size* of the *information*.

Intuitively, we feel that in this particular example, playing game $b$ requires fewer numbers of questions. The reason is that with game $a$ we must ask at least three questions. On the other hand, with game $b$ we can ask on average fewer questions. We start by asking "is it on the left half?" There is probability of ¾ to get a Yes answer on the first question. If we get a No, we have to ask on average three more questions.

The value of SMI for these two cases, are:

$$\mathrm{SMI}(a) = \sum_{i=1}^{8} \frac{1}{8} \log 8 = 3$$

$$\mathrm{SMI}(b) = -\frac{3}{4}\log\frac{3}{4} - \sum_{i=1}^{7} \frac{1}{28}\log\frac{1}{28} = 1.513$$

Finally, consider the game in Figure 2.9c. You have the same number of regions but now one area is $A_1 = 0.999A$, and all the other areas are within 0.01 of $A$. Which game would you prefer to play, $b$ or $c$? This is an easy question. I am sure you will choose game $c$. If you play this game many times you will need on average not more than one question to ask.

## 2.8 The birth of Information Theory

The concept of information is a very general concept. It includes subjective ("this book is interesting") and objective ("this book contains 200 pages") information. Information can be either interesting or dull, it can be meaningful or meaningless, it can be helpful, beautiful, reasonable – add any adjective that pops up in your mind, and you will find examples of such information.

The idea that information is something *measurable* was not widely appreciated until 1948, when Norbert Wiener's book, *Cybernetics*, was published, and Claude E. Shannon published "*The Mathematical Theory of Communication*."

One can think of many measures of information, much like there are many measures of physical objects (weight, volume, surface-area, etc.), and of events (probability, length of the event, the number of objects the event is concerned with, etc.).

Consider the two following items of information that I have heard in the news today:

"The snow storm that hit New York this morning has left a trail of devastation with 60 people injured, 4 dead, and thousands left with no electricity."

"A bomb in Iraq killed 80 people."

Clearly, the first is longer than the second. It has more letters and more words. It is also about more people affected by the snow storm in New York. On the other hand, the second is shorter in length, but it reports on a *larger* number of people killed. The second might also be more important and relevant for those who live in Iraq, but the first is more important for those who live in New York.

As you must realize, it is even more difficult to assign a measure to *any* information than it is to assign weight to *any* object, or to assign probability to *any* event. Therefore, we have to find a "distilled" form of information on which we can define a precise and objective measure.

In 1948, Shannon published a landmark paper entitled "A mathematical theory of communication." In Section 6 of this paper, Shannon writes:

*"Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \dots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event, or how uncertain we are of the outcome?*

1. *H should be continuous in the* $p_i$.
2. *If all the* $p_i$ *are equal,* $p_i = \frac{1}{n}$, *then H should be a monotonically increasing function of n. With equally likely evens there is more choice, or uncertainty, when there are more possible events.*
3. *If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.*

Then Shannon proved that the only quantity satisfying the three assumptions above is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

Shannon denoted his quantity by the letter *H*. We shall refer to it as SMI.

Shannon referred to the quantity he found as the amount of "choice" or "uncertainty" about the outcome. This quantity later became the central concept in Information Theory and was sometimes referred to simply as "information." This has caused considerable degree of confusion, mainly because "information" in general, can have *meaning*, *value*, *importance*, etc. but the SMI is a purely objective quantity depending only on the probability distribution and not on the specific experiment or game.

To highlight the gist of Shannon's achievement let us go back to the 20-question game. I hid a coin in one of $N$ boxes and I tell you that I chose the box where I placed the coin with probability distribution $p_1, \ldots, p_N$. If you prefer, think of the game with the dart hitting a board having $N$ regions with relative areas $p_1, \ldots, p_N$. Suppose that I pay a dollar for each answer I get, and when I find the outcome of the experiment (where the coin is hidden, or where the dart hit), I get a prize of X dollars.

How should one play this game with a minimum number of questions so that one maximizes his/her earnings? Offhand, it is not clear that such a "maximizing-earning-method" exists, and even if it does, it is not clear how to find out the number of questions that will give us the maximum returns in the game.

Shannon formulated an equivalent problem, having no idea if a solution to his problem exists. He further assumed that if such a measure exists, it must fulfill some plausible properties. With these set of plausible properties he proved that there is only one quantity that fulfills these properties, and that was how he found the formula for $H$, which we can now refer to as

SMI. The details of the proof are highly mathematical and will not concern us here. I wanted to convey to you the flavor of the type of problem Shannon faced; to find a solution to a problem without knowing if such a solution exists.

Once the SMI was found, people realized that it can be used in many other fields of research far from communication theory. It was found useful in physics, mathematics, biology, psychology, sociology and even in literature, music and other arts. This huge scope of applicability is probably the reason why this is deemed so powerful, and for some even an awesome quantity.

As we have seen the SMI is purely a *probabilistic quantity* depending only on the given probability distribution, but independent of the experiment or the game that provided this distribution.

Notwithstanding Shannon's enormous achievement, he committed a small semantic mistake – he called his quantity Entropy. Of course, one can call the SMI by any term one wants. Unfortunately, the specific choice of the term Entropy which was already used in physics has caused great deal of confusion, and a vigorous debate ensues on the very meaning of the SMI, as well as the meaning of Entropy. We shall see in

Chapter 3 that entropy is a particular case of SMI, but the SMI in general is not entropy.

## 2.9 Interpretations of SMI

There are several interpretations of SMI. The most important and useful ones are:

1. Average uncertainty about all the outcomes of an experiment;
2. Average unlikelihood about all the outcomes of an experiment;
3. A measure of the amount of information contained in a probability distribution, or the amount of missing information in playing the 20Q game.

> You might be wondering; the Shannon measure of information is, by definition, a measure of information. Why does one need to *interpret* the SMI as a measure of information?

This is indeed a very valid question. As we shall soon see the interpretation of SMI as a measure of information is the least straightforward compared with the first two interpretations. It is ironic that the "informational" interpretation of SMI is the most difficult to see, and as a result it is also the one which is commonly misused. Note that the SMI has the form of average quantity. However, this is a very special average. It is an average of the quantity $-\log p_i$ using the probability distribution $p_1, \dots, p_N$. (Note that Shannon used the letter $n$ for the total number of outcomes. We shall use the letter $N$ instead).

### 2.9.1 The uncertainty meaning of SMI

The interpretation of SMI as an *average uncertainty* is very popular. This interpretation is derived directly from the meaning of the probability distribution.

Suppose that we have an experiment yielding $N$ possible outcomes with probability distribution $p_1, \dots, p_N$. If, say, $p_1 = 1$, then we are *certain* that the outcome "1" occurred or will occur. In general, for any other value of $p_i$ we are *less certain* about the occurrence of the event $i$. *Less certainty* can be translated to *more uncertainty*. Therefore, the smaller the value of $p_i$, the larger the value of $-\log p_i$, the larger the extent of uncertainty about the occurrence of the event $i$. Multiplying $-\log p_i$ by $p_i$, and summing over all $i$ , we get an *average uncertainty* about *all* the possible outcomes of the experiment.[10] See also Appendix A, Figure A.3.

Finally, we note that whenever one says that SMI is a measure of uncertainty, we mean uncertainty with respect to *all* the *outcomes* of an experiment in the sense discussed above. For instance, when we throw a die we can talk of many uncertainties about the color, the mass, the form, etc. of the die. Unfortunately, you can find in many popular science books a description of SMI (as well as entropy) as "uncertainty," without specifying what that uncertainty refers to.

### 2.9.2 The unlikelihood interpretation

A slightly different but still useful interpretation of SMI is in terms of *likelihood* or *expectedness*. These two are also derived from the meaning of probability. When $p_i$ is small, the event *i* is unlikely to occur, or its occurrence is less expected. When $p_i$ approaches 1, the occurrence of *i* becomes more likely, or more expected. Since $\log p_i$ increases when $p_i$ increases, we can say that the larger the value of $\log p_i$, the larger the *likelihood* or the larger expectedness for the event. Since $0 \leq p_i \leq 1$, we have $-\infty \leq \log p_i \leq 0$. The quantity $-\log p_i$ is thus a measure of the *unlikelihood* or the *unexpectedness* of the event *i*. See Appendix A for a plot of $-\log p$ as a function of $p$. Therefore, the quantity SMI $= -\sum p_i \log p_i$ is a measure of the *average* *unlikelihood*, or *unexpectedness* of the entire set of the outcomes of the experiment.

### 2.9.3 The meaning of the SMI as a measure of information

As we have seen, both the uncertainty and the unlikelihood interpretation of SMI are derived from the meaning of the probabilities $p_i$. The interpretation of SMI as a measure of information is a little trickier and less straightforward. It is also more interesting since it conveys a different kind of *information* on the Shannon measure of *information*. As we already emphasized the SMI is not *information*. Also, it is not a measure of every piece of information, but of a very particular kind of information. Confusing SMI with information is almost the rule, not the exception by scientists and non-scientists alike.

Some authors assign to the quantity $-\log p_i$ the meaning of information (or self-information) associated with the event $i$. This is not a valid interpretation of $-\log p_i$. For details, see Ben-Naim (2017b).

Both $p_i$ and $-\log p_i$ are measures of the uncertainty about the occurrence of an event. They do not measure *information* about the events. Therefore, we do not recommend to refer to

$-\log p_i$ as information (or self-information) associated with the event $i$.

It is sometimes said that removing the *uncertainty* is tantamount to obtaining *information*. This is true for the *entire* experiment; that is, the entire probability distribution, not to individual events.

Suppose that we have an unfair die with probabilities $p_1 = \frac{1}{10}, p_2 = \frac{1}{10}, p_3 = \frac{1}{10}, p_4 = \frac{1}{10}, p_5 = \frac{1}{10}$ and $p_6 = \frac{1}{2}$. Clearly, the uncertainty we have regarding the outcome $i = 6$ is less than the uncertainty we have regarding any specific outcome, say, $i = 4$. When we carry out the experiment and find the result, say $i = 3$, we removed the uncertainty we had about the outcome before carrying out the experiment. However, it would be wrong to argue that the *amount* of information we got is larger or smaller than if another outcome had occurred. Note also that we talk here about the *amount* of information, not the information itself. If the outcome is $i = 3$, the information we got is: “The outcome is “3.” If the outcome is $i = 6$, the information is: “The outcome is “6.” The information you receive is different in these two cases, but one cannot claim that one is larger or smaller than the other.

We emphasize again that the interpretation of SMI as *average uncertainty* or *average* unlikelihood is derived from the meaning of each term $-\log p_i$. The interpretation of SMI as a measure of information is not associated with the meaning of each probability $p_i$, but with the entire distribution $p_1, \dots, p_N$.

As everyone who has played the 20-question game knows, the number of questions you need to ask depends on the *strategy* of asking questions. It turns out that the quantity SMI to which we referred to as Shannon's Measure of Information (SMI) provides us with a measure of this information in terms of the minimum number of questions one needs to ask in order to find the outcome of an experiment, given the *probability distribution* of the various outcomes.

For a general experiment with *N* possible outcomes, having probabilities $p_1, \dots, p_N$, the SMI is a measure of how "difficult" it is to find out which outcome has occurred given that an experiment was carried out, and that we know the probability distribution of the outcomes. As we have seen, experiments having the same total number of outcomes *N* but with different probability distributions, the amount of information (measured in terms of the number of questions) is different. In other words, knowing the probability distribution gives us a "hint" or some partial information on the outcomes. This is the reason we refer to SMI as a measure of the amount of information *contained* in, or *associated* with a given probability distribution. We emphasize again that the SMI is a measure of information associated with the *entire* distribution, not with the individual probabilities.

## 2.10 Conclusion regarding “uncertainty” and “measure of information”

In this chapter we introduced the concept of SMI, its interpretations, and some of its properties. Once you have grasped the meaning of SMI, you will effortlessly understand the meaning of entropy – which is nothing but a special case of SMI multiplied by a constant.

Before we discuss entropy and the Second Law in the next chapter, we must pause and ponder on the meaning of the SMI. This is essential since there are many pitfalls lurking along the paths leading from SMI to entropy in which many scientists, and even Nobel prize winners, no less, have fallen into. You will be surprised that some respected authors fill entire books with such confusion of information with SMI, and SMI with entropy.[11]

The first thing to understand is that the SMI is not *information*, but rather a specific measure of a specific kind of information. When I tell you that "today there is a snowstorm in New York," I provide you with some information. This information can be important to you (if you live in New York, and plan to go out today), or totally unimportant and irrelevant (if you live in China, and have no plans to visit New York). This information could either be exciting, boring, irrelevant, and many other attributes you might wish to attach to it. All these attributes have nothing to do with the SMI associated with this particular message.

The SMI associated with this message has to do with the frequencies of the letters in the English alphabet. It has to do with how efficiently you can code this message, transmit it through some channels, and decode it. This is exactly what Shannon was interested in when he sought a *measure* of information.

The following three messages carry different information:

John loves Ruth and Linda

Linda loves Ruth and John

Ruth loves John and Linda

All of these messages are *different*. However, from the point of view of SMI they are identical. In fact, there are many more messages, even meaningless messages which have the same SMI. You can construct some examples by jumbling the letters of these sentences.

Now let us discuss the relationship between the concept of uncertainty and a measure of information. Both of these concepts are in the context of this book associated with a probability distribution. Some people will tell you that entropy (as a special case of SMI) is information, or it is uncertainty. This is not the case, either for the SMI, or for the entropy.

If I hold a die, you might have *uncertainties* about its color, size, weight, number of faces and many more. I can also provide you with *information* about its color, size, or about the number of faces, and much more. All these uncertainties and all this information are irrelevant to SMI.

To clarify the meaning of the uncertainty and the meaning of the SMI, consider the following three dice with their corresponding probability distributions:

**Table 2.2: Three different dice and the corresponding probability distributions**

| Face | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| die *a* | 0.98 | 1/400 | 1/400 | 1/400 | 1/400 | 0.01 |
| die *b* | 0.8 | 1/40 | 1/40 | 1/40 | 1/40 | 0.1 |
| die *c* | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

I show you the three dice and I ask you the following questions:

1. In which of these dice are you most certain (or least uncertain) about the outcome?
2. I throw each of the three die. I know on which face each one fell. You have to ask binary questions to find out which outcome occurred. Can you estimate how many questions you will need to ask in each case to find out in which outcome occurred?
3. I throw each of the dice. I tell you the outcome of each die. Can you tell in which case you got more information, or less information? (Note that the information I gave you is about the outcome, or about the face on which the die landed).

Before answering these questions, you should realize that the answers you are going to provide depend on the probability distribution of each die. Without having the information on the probability distribution there is no way you can answer these questions.

Remember, it is essential to know the distribution before you attempt to answer these questions.

**Answers to the questions**

Clearly, for die *a* you are most certain about the outcome. You look at the three distributions, and you can tell that the result "1" has the largest probability in case *a*. Therefore, you are most certain, or least uncertain about the outcome.

On the other hand, you are most uncertain (or least certain) about the outcome in case *c*. It is also clear that in case *a*, you need to ask fewer questions than in case *c*. The average *uncertainty*, as well as the average number of questions you will need in each case is given by the SMI. One can show that for these three dice, we have: $\mathrm{SMI}(a) < \mathrm{SMI}(b) < \mathrm{SMI}(c)$. For details see Ben-Naim (2017a).

Regarding the last question, it is often said that by asking questions you get information, and getting information removes your uncertainty about the outcomes. Therefore, in case *a*, your uncertainty is the least. Hence, by getting the information about the outcomes you did not get much information. You got a little more information in case *b*, and you got the maximum information in case *c* (since your uncertainty was maximal for these dice).

In a more general experiment, when there are $N$ possible outcomes the SMI measures the average uncertainty about the outcomes of the experiment or about the minimum number of questions you need to ask in order to find out which outcome occurred. Equivalently, the SMI is a measure of the amount of information contained in, or belonging to the distribution.

The maximum value of the SMI is when all the outcomes are equally probable, in which case, the SMI is $\log_2 N$ (If you know how to handle logarithms you can check that whenever all the outcomes are equally probable then the SMI as defined by Shannon is equal to $\log_2 N$ ). The minimal value of the SMI is obtained when one outcome has probability 1, and all others have probability zero. In this case you are certain about the outcome. Your uncertainty is zero, and you do not lack any information about the outcomes.

In the above discussion we used the terms “I know,” or “I have the information” about the outcome. This language might lead you to think that the SMI is a subjective quantity. In fact, many scientists have reached such a conclusion. However, this conclusion is wrong. The SMI is an objective quantity defined on any given distribution. It does not depend on who has, or does not have information on the distribution. When we interpret the SMI as the number of questions one needs to ask in order to obtain the information given the distribution, we mean *anyone* who has that distribution, and it can be you, or me, or the computer.

Another very common mistake is the following: Suppose there are $N$ equally probable outcomes. In this case, the SMI is $\log_2 N$. Some people reach the conclusion that if I (or you) know the outcome, the SMI you will calculate is $\log_2 1 = 0$, which means you do not have to ask any question – you already know the answer. This conclusion is wrong. The SMI does not depend on how much information you have on the outcomes. It only depends on the distribution, and it does not depend on who has that information.

This kind of conclusion is also mistakenly reached about entropy. If there are W equally probable microstates of a

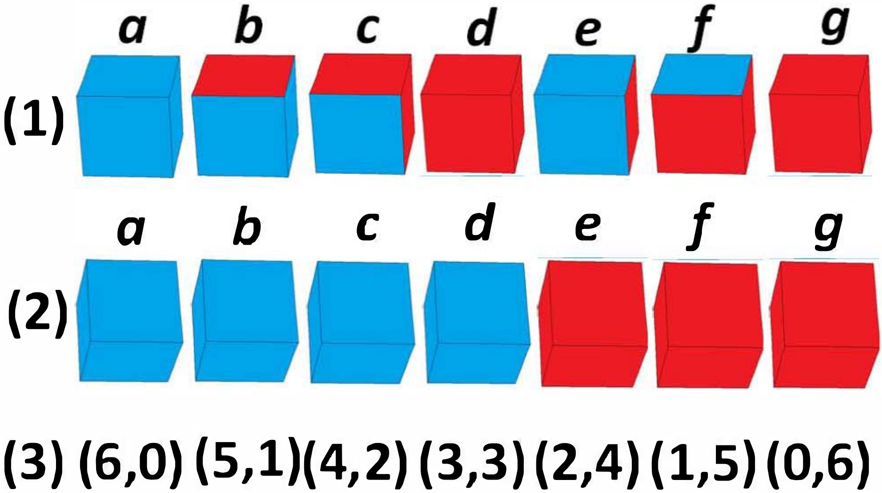


Figure 1.1. Seven dice, each having a different number of blue and red faces. Panels (a) and (b) show each die from a different angle. To get the number of blue or red faces, you need to count, the number of blue or faces in both panels.

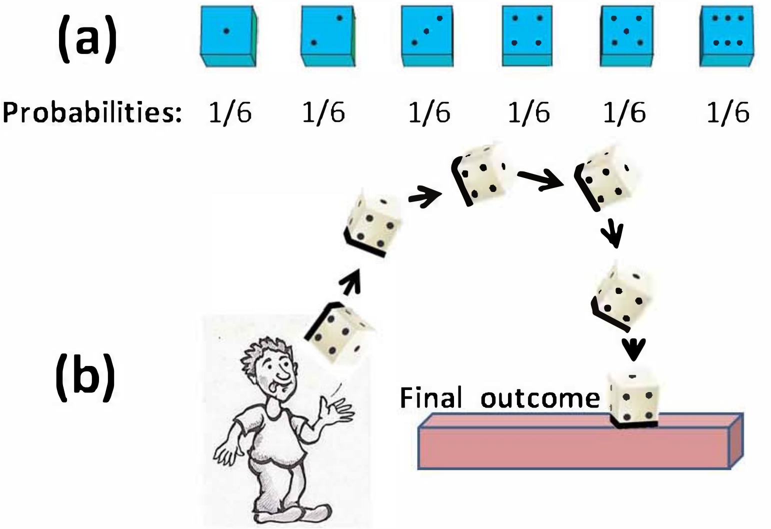


Figure 1.2. (a) A regular fair die with different numbers of dots on its faces. (b) An "unfair" die having additional weight on one of its faces.

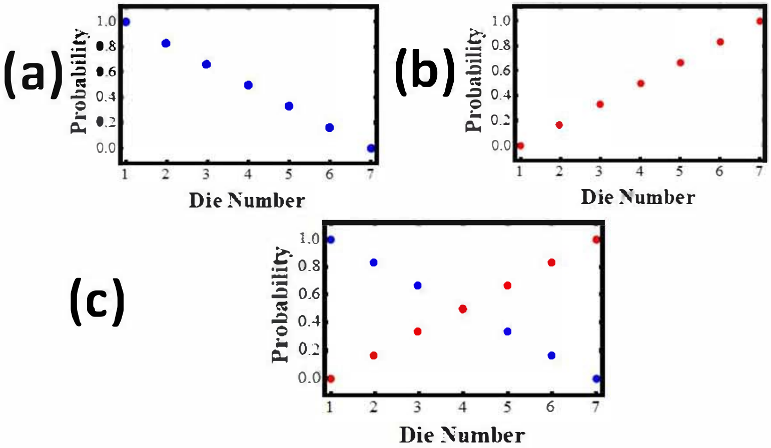


Figure 1.5. (a) Probabilities of occurrence of blue for the dice 1 to 7, corresponding to *a* to *g* in Table 1.2. (b) Probabilities of occurrence of red. (c) Probabilities of occurrence of blue and red for the dice 1 to 7. Note the symmetry of the figure.

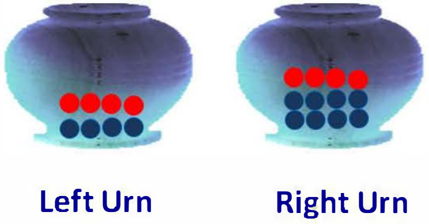


Figure 1.13.  Easy game; choice between the two urns.

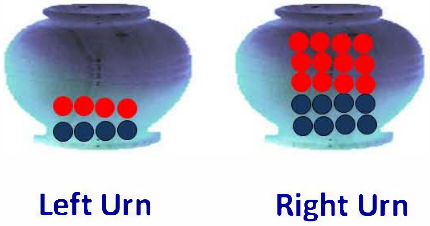


Figure 1.14.  Difficult game; choice between the two urns.

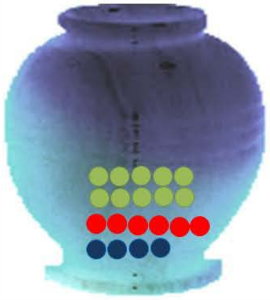


Figure 1.15.  An urn containing four blue, six red, and ten green marbles.

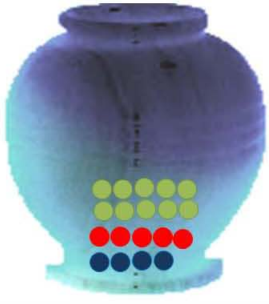


Figure 1.16. An urn containing four blue, five red, and ten green marbles.

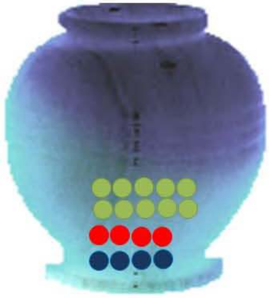


Figure 1.17. An urn containing four blue, four red, and ten green marbles.

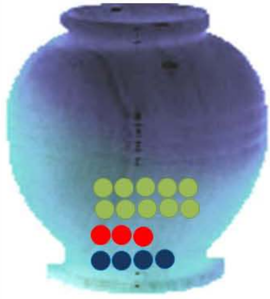


Figure 1.18. An urn containing four blue, three red, and ten green marbles.

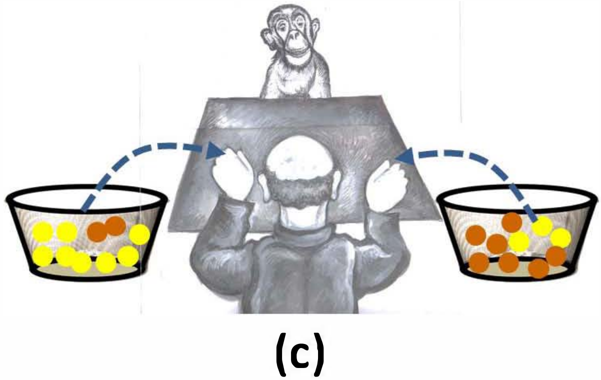


Figure 1.24. (a) A monkey examining two urns, then;
(b) choosing while blindfolded between the two urns.
(c) The actual experiment with banana and carrot pieces.

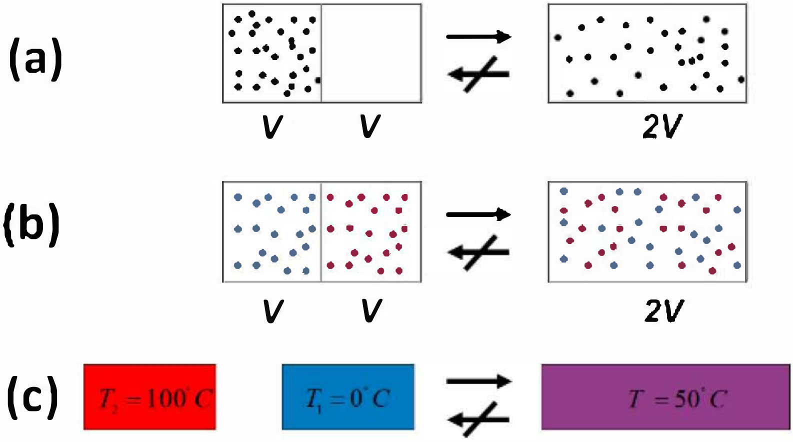


Figure 3.1. Three typical spontaneous processes occurring in isolated systems. (a) Expansion of an ideal gas. (b) Mixing of two ideal gases, (c) Heat transfer from a cold to a hot body.

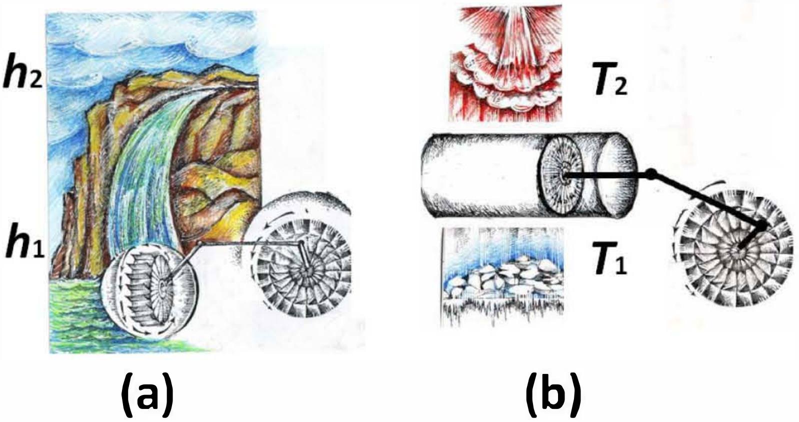


Figure 3.2. (a) An artist rendition of (a) “Water fall,” and (b) “Heat fall.”

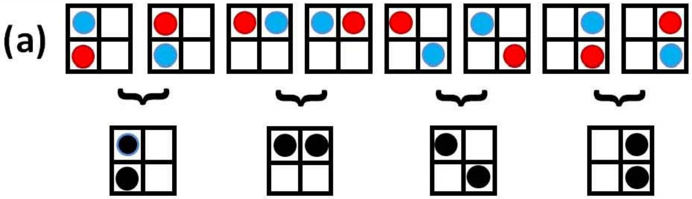


Figure 3.5. (a) Eight of the 12 possible configurations of two different particles in four cells. These become four configurations when the particles are indistinguishable.
Can you draw the missing four configurations?
See Figure 3.5 (b) below.

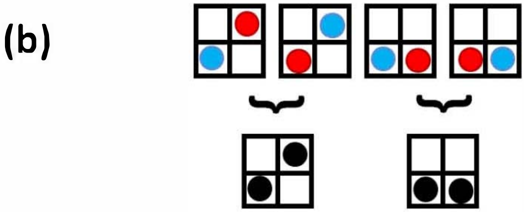


Figure 3.5 (b) Four of 12 possible configurations of two different particles in four cells. These become two configurations when the particles are indistinguishable.

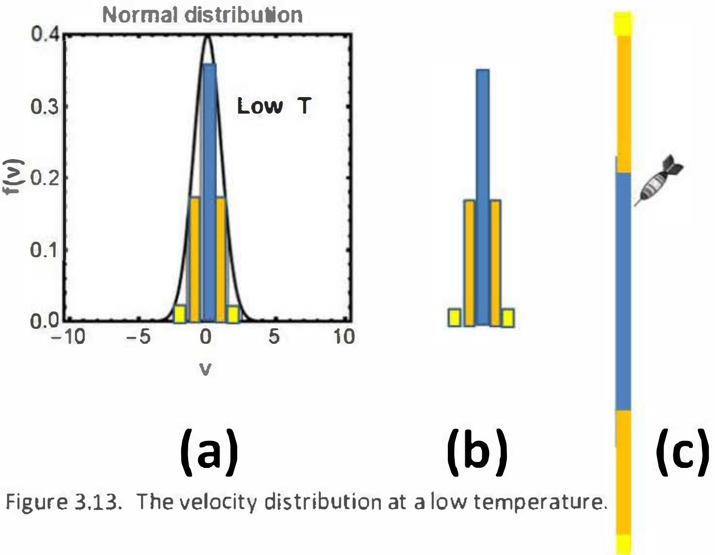


Figure 3.13. The velocity distribution at a low temperature.

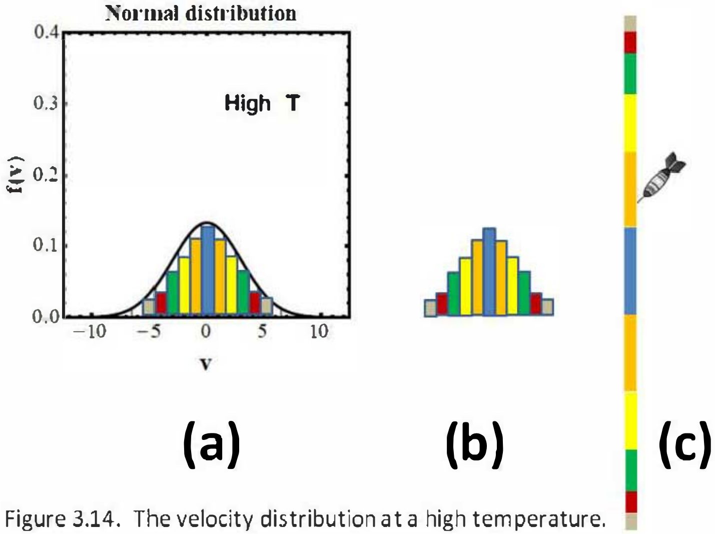


Figure 3.14. The velocity distribution at a high temperature.

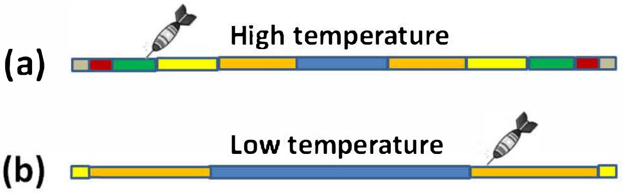
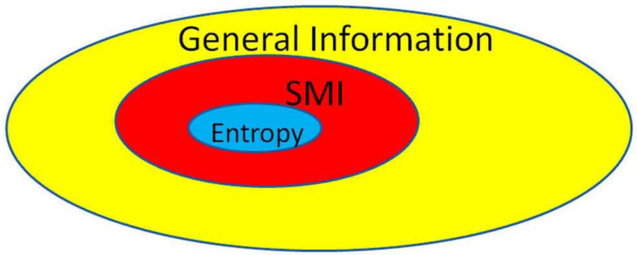


Figure 3.15. Two different 20Q games with a board divided into different numbers of regions.



A qualitative relationship between the "sizes" of the three concepts: General Information, SMI and Entropy.

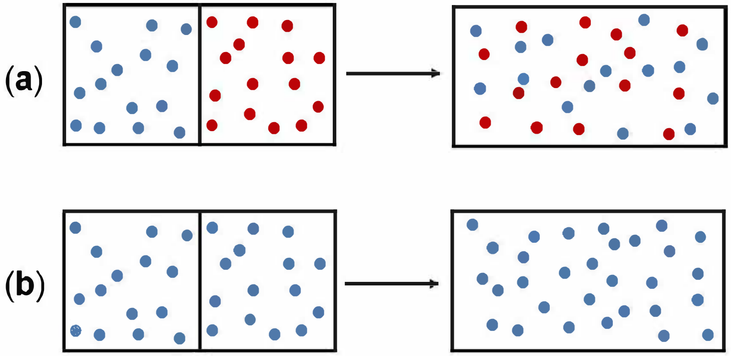


Figure 3.26. (a) Mixing of two different kinds of molecules. (b)"Mixing" the same kind of molecules.

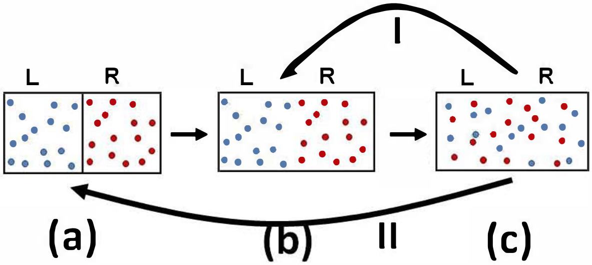


Figure 3.27. The initial, the final, and an intermediate state in the mixing process. Process I is possible, but highly improbable. Process II is impossible.

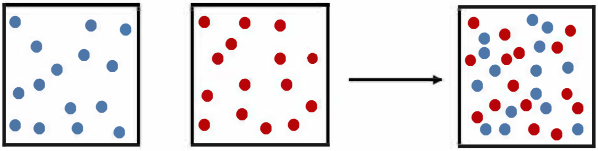


Figure 3.28. Mixing of two gases with zero change in entropy.

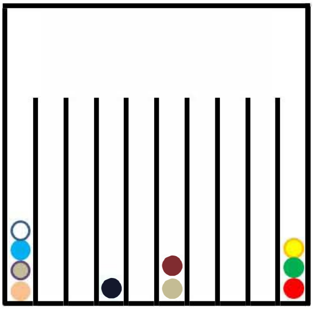


Figure 3.30. A possible distribution of ten marbles in ten cells.

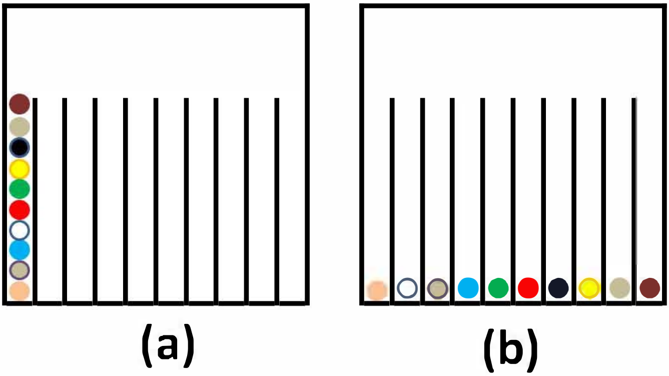


Figure 3.31. Two extreme distributions of ten marbles in ten cells.

thermodynamic system, the entropy is given by $S = k\ln W$. Some people say that if they know in which of the W states the system is, then the entropy of the system from their point of view would be zero ($k\ln 1 = 0$). We shall further discuss this case in the next chapter.

## 2.11 Summary of Chapter 2

In this chapter we learned about the Shannon measure of information, some of its properties and some of its interpretations. We also saw that young children have some sense of the *average uncertainty*, or the measure of information, in the sense discussed in this chapter.

I hope that you now have a qualitative idea of the meaning of SMI. In the next chapter, we shall see that the entropy – the concept considered to be the most mysterious one in physics – is nothing but a quantity which, up to a multiplicative constant, is a special case of SMI. It is a measure of information associated with a very special distribution.

# CHAPTER 3

# ENTROPY AND THE SECOND LAW OF THERMODYNAMICS

*Finally, we arrived at the most important chapter of this book. In this chapter we discuss the concept of entropy and its association with the Second Law of Thermodynamics (2$^{nd}$ Law). We start with some historical notes on the concept of entropy and the Second Law. We then outline the derivation of the entropy of an ideal gas based on the SMI. We shall see that entropy is nothing but a special case of SMI. It turns out that this is the simplest and the only valid and proven interpretation of entropy. We shall also see that entropy is defined for a thermodynamic system only at equilibrium states. The Second Law deals with the changes that occur when we remove some constraints from a system at equilibrium. We shall see that the reason for the seemingly one-way processes that we observe is probabilistic. We shall also discuss the connection between the various thermodynamic formulations of the Second Law, on one hand, and the probabilistic formulation, on the other hand.*

## 3.1 The birth and the early evolution of the concept of entropy

In most textbooks on thermodynamics one finds that the concepts of entropy and the Second Law are intimately intertwined. We shall discuss these two concepts separately. We shall see that entropy may be defined, interpreted, and applied without ever mentioning the Second Law. Likewise, one can formulate and understand the Second Law without ever mentioning entropy. As we shall see these two concepts are related to each other, but this relationship holds only for processes occurring in an *isolated* system.

Even before the Second Law was enunciated people noticed that some processes occur in an apparent one-direction only.

Look at the three processes depicted in Figure 3.1 (see color centerfold). We assume that each of these processes occur in isolated systems which means that there are no interactions between the system and its surroundings.

In Figure 3.1a, we start with a gas initially confined to a volume *V*. For simplicity we assume that the gas molecules do not interact between themselves. This is an approximation. Any pair of real molecules always interacts with each other. However, if the density of the gas is very low we can assume that the average distance between pairs of molecules is very large and hence, the interactions are negligible. We call such a gas an *ideal gas*.

We remove the partition and we see that the gas *expands* to occupy the entire volume $2V$. This experiment is so obvious that you can imagine the outcome even without actually performing the experiment.

We never observe the reverse of this process. If I tell you that I started with a gas, which fills the entire volume of some container, then suddenly all the molecules moved to occupy half of the container, you would not believe me. Right?

Next, look at the process 3.1b. We start with two different ideal gases separated by a partition. When we remove the partition separating the two gases, we will observe mixing of the two gases.[1] We never observe the reverse of this process. If I tell you that I started with a mixture of two gases as on the right-hand side of Figure 3.1b, then suddenly the two gases were separated such as on the left-hand side of the Figure 3.1b, you would not believe me.[2]

Finally, look at Figure 1.3c. Initially, we have two blocks of metals at two different temperatures. We bring them to thermal contact and we observe that the temperature of the hot body will go down, while the temperature of the cold body will go up. Thermal contact here means that the insulating wall separating the two bodies is replaced by heat-conducting walls so that heat can flow between the two bodies.

In this experiment we always observe that heat will flow from the higher-temperature body to the lower-temperature body. After some time we reach an equilibrium state at which the temperature of the two bodies will be equal.

We never observe the reverse of this process. If we start with a piece of metal at a uniform temperature, we never observe a spontaneous heating of one part (that is, raising its temperature), and cooling a second part (that is, lowering its temperature).

These simple processes, as well as many others seem to occur in one direction (from left to right in Figure 3.1), never in the reverse direction (from right to left in Figure 3.1). Why?

In all of these processes we start with a system at equilibrium, we remove a constraint (e.g. a partition), and the system evolves into a new equilibrium state.

In all of these examples we *never* observe the reverse process spontaneously; the gas *never* condenses into a smaller region in space, the two gases *never* un-mix spontaneously after being mixed,[3] and heat *never* flows from the cold to the hot body. Note carefully that I italicized the word *never* in the previous sentences. Indeed, we *never* observe any of these processes occurring spontaneously in the reverse direction. For this reason, the processes shown in Figure 3.1 (as well as many others) are said to be *irreversible*. However, one should be careful with the use of the words "reversible" and "irreversible" in connection with the Second Law. There are several, very different definitions assigned to these words.[4]

Here, we point out two possible meanings of the term *irreversible*.

1. We *never* observe that the final state of any of the processes in Figure 3.1(see color centerfold) returns to the initial state (on the left hand side of Figure 3.1) spontaneously.

2. We *never* observe that the final state of any of the processes in Figure 3.1 returns to the initial state, and *stays* in that state.

In case 1, the word *never* is used in "practice." The system can go from the final to the initial state. In this case, we can say that the initial state will be *visited*. However, such a reversal of the process would occur once in many ages of the universe.

Therefore, this is *practically* an irreversible process; we will "never" observe such a reversal in practice.

In case 2, the word *never* is used in an absolute sense. The system will never go to its initial state and stay there!

The distinction between these two meanings of "irreversibility" is important in connection with the formulation of the Second Law.

Rudolf Clausius

In 1865, the German physicist Rudolf Clausius (1822-1888) introduced a new concept which he called entropy. Clausius himself ***did not***, and ***could not*** understand this concept on a molecular level. In fact, even the well-known concepts of temperature and pressure were not understood on a molecular level. It was much later that temperature was interpreted in terms of the kinetic energy of the atoms and the molecules. This interpretation is far from being trivial. There is nothing in our sense of temperature which indicates that it is a result of the molecular velocities of the particles.

The situation with the molecular interpretation of entropy is similar to that of the temperature. However, while everyone can feel the temperature of a body, there is no way to perceive entropy with our senses (excluding our *common sense*). This has led to many speculations about the meaning of entropy – speculations that continue to this day.

Originally, scientists in the 19th century were interested in heat engines. Heat engines were supposed to do the *work* which people used to do with their bare hands and raw muscles. Basically, you can think of a "heat engine" as an analog of waterfall engines. In waterfalls, water cascades from higher to lower levels. On its way down, the water can rotate a turbine. We can harness this rotation to our advantage like plowing a field or generating electricity. Likewise, ***heat*** flows spontaneously from a high level of temperature to a low level of temperature. On its way down, this flow of heat can also be harnessed for doing some work, like running a train or lifting weights from low to high levels.

In the 19th century, scientists believed that ***heat*** is a kind of fluid called *caloric* that flows from a higher to a lower temperature. Today, this "caloric theory" is considered obsolete. We shall still use the term "heat flow" meaning heat transferred.

Qualitatively, think again of a waterfall. You can imagine that for some quantity of water falling from $h_2$ to $h_1$ you can do some useful work. Likewise, for a given amount of heat "falling" from the higher temperature $T_2$ to a lower temperature $T_1$, you can do some useful work, Figure 3.2 (see color centerfold).

Note, however that "falling of water," and the "flow of heat" are governed by very different laws. The first is governed by Newton's Gravitation Law, the second, by the Second Law of Thermodynamics.

Traditionally, the birth of the 2nd Law is associated with the name Sadi Carnot (1796-1832). Carnot was a French physicist and engineer. Although Carnot himself did not formulate the Second Law, his work laid the foundations on which this law was formulated. Carnot was interested in the *efficiency* of a heat engine; how much useful work one can get from a given amount of heat that flows from the higher temperature $T_2$ to the lower temperature $T_1$. Carnot found, somewhat unexpectedly, that there is a limit on the efficiency of a heat engine operating between two temperatures $T_2$ and $T_1$. This finding was not a formulation of the Second Law, but it sowed the seeds for the inception of the Second Law.

The seeds sowed by Carnot sprouted in different directions which led to the different formulations of the Second Law, as well as different definitions of entropy. We shall first discuss the different definitions of entropy and then present a few formulations of the Second Law.

## 3.2 Two older definitions of entropy

In this and in the next section, we present three different definitions of entropy.

### 3.2.1 Clausius' definition of entropy

Basically, Clausius observed as every one of us does, that there are many processes that occur in nature spontaneously and always in one direction.

Going back to the three processes shown in Figure 3.1 (see color centerfold), we can ask *why* these processes always occur in one direction. Is there a law of Nature that dictates the direction of the unfolding of these processes? Look again at the three processes depicted in Figure 3.1. Take note that these are quite *different* processes and that it is far from being clear that they are all governed by the same law. Perhaps, there is one law for the spontaneous expansion of a gas, another for the spontaneous mixing of two gases, and still another for the spontaneous flow of heat from the hot to the cold body.

It was Clausius who realized the common principle underlying all these processes, and postulated that there is only one law that governs all these processes. Even before formulating the Second Law, Clausius' postulate was an outstanding achievement considering the fact that none of these processes was understood. You watched how a colored gas expands and fills a larger volume. You watch a drop of blue ink mixing with a glass of water coloring the entire liquid. You watch the hot body cooling and the cold body heating. You *watch* all of these with your *macroscopic eyes*, but you have no idea what *drives* these processes, what goes on *inside* the system you are watching. Such an insight was not even possible before the atomic nature of matter was embraced by the scientific community which allowed us to use our "*microscopic eyes*" to "see" what goes on when such processes occur. "Seeing," even with our microscopic eyes, is one thing, and *understanding* what we see is quite another. As we shall learn later the explanation of all these processes is probabilistic. This is exactly the reason why we dedicated Chapter 1 to discuss the concept of probability.

Clausius started from one particular process; the spontaneous flow of heat from a hot to a cold body. Based on this specific process, Clausius *defined* a new quantity which he called *entropy*. From here on, I will describe Clausius' "definition" of entropy. Nothing in this definition will help your understanding of entropy. If you like you can skip this part, and continue to Section 3.2.2. Just remember that Clausius defined entropy in connection with heat engines, heat transfer, and temperature. This definition is important for engineering, but it contributes nothing to understanding the meaning of entropy.

Denote by $dQ$ a *small quantity* of heat flowing *into* a system, being at a given temperature $T$. $T$ is referred to as the *absolute temperature*. $T = 0\ K$, or zero Kelvin, is the lowest possible absolute temperature, which is about $-273$°C. The *change* in *entropy* was defined as:

$$\text{(Clausius' definition)} \qquad dS = \frac{dQ}{T}$$

$Q$ has the units of *energy*, and $T$ has the units of *temperature*. Therefore, the entropy change has the units of *energy* divided by units of *temperature*. The quantity of heat, $dQ$, *must* be very small, such that when it is transferred into the system, or out of the system, the temperature $T$ does not change. If $dQ$ is a large quantity of heat, and you transfer it to a system which is initially at a given $T$, the temperature of the system might change, and therefore the change in entropy will depend on both the initial and the final temperature of the system[5]. (As you can understand intuitively, suppose you have a pot of food, initially at a temperature of say, 30°C. You put it for a very short time on the stove. A small amount of heat will flow into the pot, but its temperature will not change much. However, when you leave the pot for a long period of time, the pot's temperature will change.) Note carefully that this equation does not define *entropy* but only *changes in entropy* for a particular process, i.e. a *small* exchange of heat ($dQ > 0$ means heat flows into the system, $dQ < 0$ means heat flows out of the system). There are many processes, two of which are shown in Figure 3.1, which do not involve heat transfer. Yet, from Clausius' definition and the postulate that the entropy is a *state function*, one could devise a *path* leading from one state to another, for which the entropy change can be calculated. A *state function* means that for any well-defined system at equilibrium the value of its entropy is determined. These details are not relevant to us in this book.

It is not uncommon to refer to the equation $dS = dQ/T$ as Clausius' *definition* of entropy. In fact, this equation does not define entropy, nor *changes* in entropy for a general process (e.g. expansion of an ideal gas).

Initially, Clausius formulated a “restricted” Second Law, namely that heat does not flow spontaneously from a cold body to a hot body. However, later he postulated that there exists a quantity which he called entropy that is assigned to any macroscopic system, such that when a spontaneous process occurs *the entropy always increases*. This was the birth of the Second Law of Thermodynamics. This law introduced a new quantity to the vocabulary of physics, and at the same time brought together many processes under the same umbrella.

The extraordinary achievement of Clausius was the enormous generalization from a few spontaneous processes to *any* spontaneous process. It should be stressed here that the formulation of the Second Law in terms of entropy is valid only for *isolated systems*, i.e. systems having a constant energy, volume and number of particles. For other systems, say, at constant pressure and temperature, the entropy can either go up or down.

Look again at the three processes depicted in Figure 3.1. These are very different processes, but they are governed by the same law, the Second Law of Thermodynamics. Today, we can calculate the change in entropy, and we find that whenever a spontaneous process occurs in an *isolated* system, the entropy of the system always increases. We shall discuss the various formulations of the Second Law in Section 3.5. Before we continue, we must emphasize that by "entropy changes" we mean difference in entropy between two *equilibrium states*. We do not know how to calculate the entropy change for every spontaneous process.

At this point, we pause to discuss the important concept of equilibrium state. Experimentally, we know that systems consisting of a huge number of particles can be described by a few parameters. For instance, a gas consisting of $N$ atoms of argon can be described by its pressure and its temperature. This description is referred to as the *thermodynamic* or *macroscopic* state of the system. Obviously, a macroscopic state does not specify the *microscopic* states of the system. For these, we need to know the locations and velocities of a huge number of particles, $N \approx 10^{23}$.

We also know that there exist states for which the thermodynamic parameters, say temperature, pressure, or density do not change with time. These states are called equilibrium states.

It should be stressed however, that there is no general definition of equilibrium which applies for all systems. Callen (1985) introduced the existence of an equilibrium state as a *postulate*. He also emphasized that any definition of an equilibrium state is necessarily circular.

In practice, we find many systems, for which the parameters describing the system seem to be unchanged with time. Yet, they are not equilibrium states. But for all intents and purposes, we assume that every well-defined system, say having a fixed energy $E$, volume $V$, and number of particles $N$, will tend to an equilibrium state. At this state the entropy of the system, as well as many other thermodynamic relationships are applicable.

We mentioned above the entropy for an *isolated* system. There exists no strict isolated system. However, such systems are convenient for the construction of thermodynamics, as well as for statistical mechanics.

As for the choice of the term "entropy," Clausius explained:

> *"I propose, accordingly, to call S the* ***entropy*** *of a body, after the Greek word '****transformation.****' I have designedly coined the word entropy to be similar to* ***energy****, for these two*

*quantities are so analogous in their physical significance, that an analogy of denominations seems to me helpful."*

The choice of this term was not entirely appropriate.[6] However, during the time it was chosen the meaning of entropy was not clear. It was a well-defined quantity, and one could calculate changes in entropy for many processes without giving a second thought to the *meaning* of entropy. Perhaps, there is no "deeper" meaning to entropy. Perhaps, entropy is just another physical quantity, such as volume and energy which do not have any "deeper" meaning. In fact, there are many scientists who use the concept of entropy successfully and who do not care for the meaning of entropy, that is if it has a meaning at all.

Notwithstanding the enormous success and the generality of the Second Law, Clausius made one further generalization of the Second Law:

**The entropy of the universe always increases**

This formulation can be said to be an unwarranted *over-generalization*. We shall further discuss the fallacy of this over-generalization in Section 3.5. Here, suffice it to say that even Clausius' definition of entropy does not apply to the entire universe. First, because the universe is not at constant temperature, and second, because the universe does not have an "environment," from which heat can be transferred. Similarly, in some popular science books you might read that "children's rooms tends to be disordered," or that "kitchens tend to get messy," and ascribed these phenomena erroneously to the Second Law. We shall further discuss such misapplications of the Second Law in the Epilogue.

### 3.2.2 Boltzmann's definition of entropy

Towards the end of the 19th century, the atomistic theory of matter was firmly consolidated. The majority of scientists believed – yes, it was still a belief – that matter consists of small units called atoms and molecules. A few scientists persistently rejected that idea arguing that there is no proof of the existence of atoms and molecules; no one has seen any atom or a molecule! Therefore, they justifiably claimed that the existence of atoms and molecules was a mere speculation.

On the other hand, the so-called *kinetic theory of heat* which was based on the assumption of the existence of atoms and molecules had scored a few impressive gains.

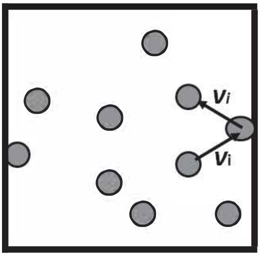


Figure 3.3. Pressure is a result of particles colliding with the wall. Here we show one particle hitting the right wall with velocity $\boldsymbol{v}_i$ and reflected with the same velocity.

First, the pressure of a gas was successfully explained as arising from the molecules bombarding the walls of the container. (Figure 3.3). Then came the interpretation of temperature in terms of the kinetic energy of the molecules,[7] that was a remarkable achievement that supported and lent additional evidence for the atomic constituency of matter.

Remember that both pressure and temperature are measurable quantities. We can feel both of them on the tip of our fingers. Neither the measurements, nor our feelings give us any hint that these quantities are manifestations of the motions of a huge number of tiny particles.

Furthermore, the concept of *heat* which was believed to be a kind of fluid that flows from a hot to a cold body was also interpreted in terms of the energies of all the individual molecules.

Thus, while the kinetic theory of heat was successful in explaining the concepts of pressure, temperature and heat, entropy was left lagging behind. It was Boltzmann who first suggested a new definition of the entropy in terms of the *total of number microstates* of a system consisting of a huge number of particles, but characterized by the macroscopic parameters of energy *E*, volume *V* and number of particles *N*.

What are these “microstates,” and how are they related to entropy? Basically, a *microstate* of the system is a detailed description of the *locations* and the *velocities* of all the particles in the system. Here is a more "technical" description.

Consider a gas consisting of *N* simple particles in a volume *V*; each particle may be at some location $\boldsymbol{R}_i$, and have some velocity $\boldsymbol{v}_i$, Figure 3.4. By simple particles we mean particles with no internal structures or "degrees of freedom." Atoms such as argon, neon, and the like are considered as simple. They all have internal degrees of freedom, but are assumed to be unchanged in all the processes we discuss here. Assuming that the gas is very dilute so that interactions between the particles can be neglected, then all the energy of the system is simply the sum of the kinetic energies of all the particles.

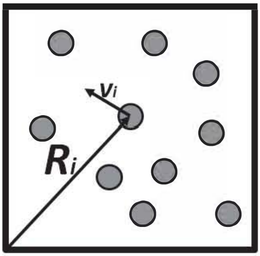


Figure 3.4. Ten particles in a box of volume $V$.
Each particle, $i$ has a locational and a velocity vector.

Imagine that you could have *microscopic eyes*, and you could see the particles rushing incessantly, colliding with each other, and with the walls from time to time. You will be impressed that there are infinite *configurations* or *arrangements* of the particles which are consistent with the requirements that the total energy is a constant, and that they are all contained within the box of volume $V$. One such configuration is shown in Figure 3.4. Each particle is specified by its location $\boldsymbol{R}_i$, and its velocity $\boldsymbol{v}_i$. Here, $\boldsymbol{R}_i$ is a *vector*, meaning it consists of the three coordinates $(x_i, y_i, z_i)$ of the location of the particle along the three axes. Similarly, $\boldsymbol{v}_i$ consists of the three components of the velocities of the particles along the three axes. We bring here a simple example of calculation of the number of configurations of two particles in 4 cells in a two-dimensional system, Figure 3.5 (see color centerfold). There are altogether 12 configurations or arrangements when the particles are different, and six configurations when they are identical.

Without getting bogged down with the question of how to define or how to calculate the total number of arrangements, it is clear that this is a huge number, far “huger” than the number of particles which is of the order $N \approx 10^{23}$. Boltzmann postulated the relationship which is now known as the Boltzmann entropy:

$$S = k_B \log W$$

where $k_B$ is a constant, now known as the Boltzmann constant, and $W$ is the total number of microstates of the system. Here, log is the natural logarithm. See Appendix A. This formula is inscribed on Boltzmann’s tombstone in Vienna.

Ludwig Boltzmann (1844-1906) was an Austrian physicist and a philosopher. His ideas were not accepted by his contemporaries.

At first glance, Boltzmann's entropy seems to be completely different from Clausius' entropy. Nevertheless, it was found that in all cases for which one can calculate changes of entropy, one obtains agreements between the values calculated by the two methods.

Ludwig Boltzmann

Exercise: Remember the definition of SMI? Look again at Section 2.8. Suppose you have $W$ microstates or outcomes of any experiment. You are told that all these microstates have equal probability ($1/W$). Plug these probabilities in the SMI and see what you get.

Boltzmann's entropy was not easy to accept, not only by those who did not accept the atomic theory of matter, but also by those who accepted it. The criticism was not focused so much on the definition of entropy, but more on the formulation of the Second Law of Thermodynamics. Boltzmann explained the Second Law as a probabilistic law. In Boltzmann's words:

*"... a system...when left to itself, it rapidly proceeds to disordered, most probable state."*

This statement, especially the phrase "most probable state" was initially shocking to many physicists. Probability was totally foreign to physical reasoning. Physics was built upon deterministic and absolute laws; there were no provisions for exceptions. The macroscopic formulation of the Second Law was absolute – no one has ever observed a single violation of the Second Law. Boltzmann, on the other hand, insisted that the Second Law is only *statistical*; entropy increases *most* of the time, not ***all*** the time. The decrease in entropy is not an *impossibility* but is only highly improbable.[8]

At the time when Boltzmann proclaimed the probabilistic approach to the Second Law, it seemed as if this law was somewhat *weaker* than the other laws of physics. All physical laws were absolute and no exceptions were allowed. The Second Law, as formulated by Clausius, was also absolute. On the other hand, Boltzmann's formulation was not absolute – exceptions were allowed. It was much later realized however, that the admitted non-absoluteness of Boltzmann's formulations of the Second Law, is in fact more *absolute* than the absoluteness of the macroscopic formulation of the Second

Law, as well as of any other law of physics for that matter. We shall see why this is true in Section 3.5. At this point, I suggest that you pause and try to understand what this last sentence means.

Boltzmann's formulation of the Second Law created two "paradoxes," the so-called "reversibility paradox," and the "recurrence paradox." We shall not discuss these paradoxes here.[9]

As we noted in Section 3.2.1, Clausius did not define the entropy function, nor did he provide a method of calculating the *value* of the entropy for any system at equilibrium. This could be achieved by using the Third Law of Thermodynamics. The application of the Third Law to calculate the absolute values of the entropy from experimental data (on heat capacity and heat of phase transitions) will not be discussed here.[10]

## 3.3 The new definition of entropy based on Shannon's measure of information

### 3.3.1 Introduction

In this section, we present a new and relatively recent definition of entropy. However, it is superior to both the Clausius and the Boltzmann definitions. Unlike the Clausius definition which provides only a definition of changes in entropy, the present one provides the *entropy* itself. Unlike the Boltzmann's definition which is strictly valid for isolated systems and does not provide a simple intuitive interpretation, the present one is more general and provides a clear, simple, and intuitive interpretation of entropy. It is more general in the sense that it relates the entropy to probability distributions, rather than to the number of microstates. One final "bonus" afforded by this definition of entropy is that it not only removes any traces of mystery associated with entropy, but it also expunges the so-called irreversibility paradox (See Note 9).

We present here in a very qualitative way, the procedure of defining entropy starting with the concept of SMI. If you are not sure what the SMI means you should refresh your memory by reading Chapter 2.

Having done with the interpretation of entropy, i.e. answering the question: "*What* is entropy?" we turn to discuss the meaning of the Second Law which essentially answers the question: *Why* does entropy increase in a spontaneous process occurring in an isolated system? We also show the intimate relationship between the "*what*" and the "*why*" questions. Finally, we discuss the extension of these relationships to systems other than isolated ones.

Before we describe the procedure for obtaining the entropy we have to discuss two principles which will be used in this procedure. These are the *uncertainty* and the *indistinguishability* principles.

Suppose you have a single atom in a one-dimensional box of unit length. The *state* of the atom can be described by its *location* and its *velocity* at each instance of time. Let $l$ be its location and $v$ its velocity. The pair $(l, v)$ describes the state of this atom. It is also clear that there are infinite number of such states. Therefore, if I know the *exact* state of the atom, and you have to find its state by asking binary questions, you will need on average, an *infinite* number of questions to ask.

Fortunately, there is the uncertainty principle in physics. This principle states that you cannot determine both the exact location and the exact velocity of the atom, but there is a limit to the "size of the box" in which you can determine the state of the atom.

This passage from the infinite number of states in the *continuous* range of locations and velocities to the finite number of possibilities is demonstrated schematically in Figure 3.6. Here, we reduce the *infinite* number of points in the range [0,1], Figure 3.6a, to a *finite* number of small intervals, Figure

3.6b. Similarly, the *state* of an atom in terms of its location and velocity may be defined in terms of a small box in the (*l*, *v*) space.

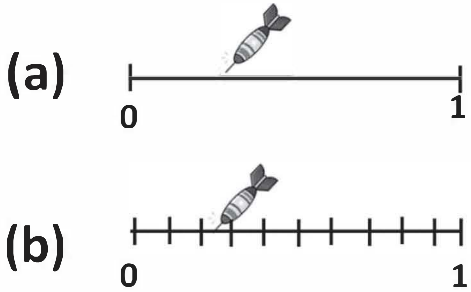


Figure 3.6. (a) A dart hits a one-dimensional segment of length 1. There are infinite possible locations for the dart.
(b) Passage from the infinite to the discrete description of the states.

Now, if I know the state of the atom and you have to ask binary questions, you will need to ask only a *finite* number of questions. This game is no different from playing the 20Q game that you are familiar with. Next, we move from one atom to a huge number of atoms, say $10^{23}$ atoms in a cubic box of edge length 1. The problem is now to find the "state" of this huge number of atoms – not the exact state, but an approximate state as is imposed by the uncertainty principle.

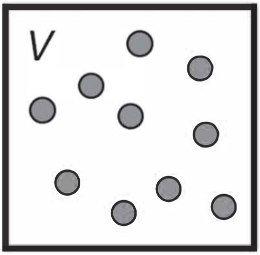


Figure 3.7. A specific (locational) configuration of ten particles in a box of volume *V*.

Figure 3.7 shows a configuration of ten particles. Each atom is described by its location and velocity along each of the axes (*x*,y,z). Such a complete description of the *state o*f all atoms is called a *microstate*. Again, there is a huge number of such microstates of the system. We can imagine playing the 20Q game on *all* these microstates. Note that the microstates are changing with time. Here, we play the game with *all the microstates,* of the system. There should be no difficulty in playing this game on this huge number of microstates. You will need to ask many questions, far too many than you can achieve in your lifetime, or during the whole age of the universe. However, there is no difficulty *in principle* in imagining playing such a game with such a huge number of ojects (the microstates). There will be a finite number of questions – finite, albeit a huge number. As we shall soon see, this number of questions will be related to the entropy of the system.

Next, we need to introduce another principle from physics. The particles are *indistinguishable*. This means that if you interchange the locations of two identical atoms, you get the same configuration.

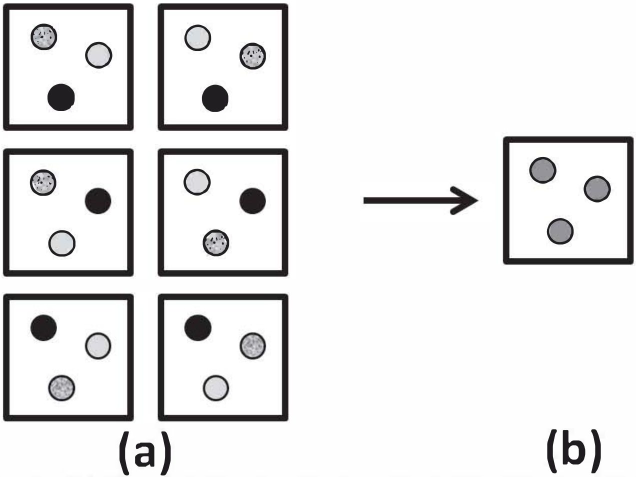


Figure 3.8. (a) Six different configurations are reduced to one when the particles a become indistinguishable (b).

Figure 3.8 illustrates this reduction in the number of configurations for three particles. As can be seen on the left-hand side, there are six *different* configurations. The six configurations coalesce to one configuration when the particles are indistinguishable.

The road leading from the SMI to entropy is highly mathematical. We shall outline here only the general steps in a qualitative way. More details on the procedure is available in Ben-Naim (2008, 2012, 2015a).

The next few subsections will be the most *difficult* part of this book. They will be "difficult" from the technical, mathematical point of view, but will be very simple *conceptually*. If you have followed me in playing the 20Q games in Chapter 2, you should not have any difficulty in understanding the concept of entropy. Entropy is nothing but a special case of SMI. Only the units of entropy are different – you should multiply the SMI by a constant ($k_B$ ln 2, where $k_B$ is the Boltzmann constant, and ln 2 is the natural logarithm of 2).

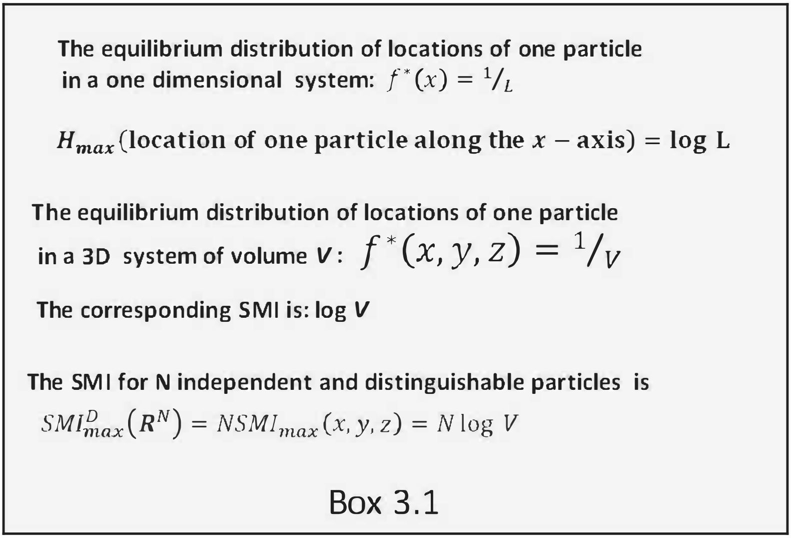
Also, the game is the same as the 20Q game except for the huge number of objects – which in our case will be the various microstates, or configurations of the molecules.

The overall plan of obtaining the entropy of an ideal gas from the SMI consists of four steps. First, we calculate the *locational* SMI associated with the *equilibrium* distribution of *locations* of all the particles in the system. Second, we calculate the *velocity* SMI associated with the *equilibrium* distribution of velocities (or momenta) of all the particles. Third, we add a correction term due to the quantum mechanical *uncertainty principle*. Fourth, we add a correction term due to the fact that the particles are *indistinguishable*. Thus, the final result is an SMI based on the probability distribution of all the locations and velocities of all the particles in the system at equilibrium. Once you understand what this probability distribution is, you can define the corresponding SMI, and since you already know the *meaning* of the SMI, you will also know the meaning of entropy. The entropy is simply obtained by multiplying this SMI by a constant.

### 3.3.2 The locational SMI of one particle in one dimension

Consider first the case of one particle which freely moves in a one-dimensional (1D) "box" of length $L$. There are *infinite* numbers of points in which the center of the particle may be found. However, we are never interested in the *exact* point in which the particle is, but rather in which small interval of length $dx$ the particle is. In Figure 3.9, we draw the probability density for the uniform distribution. This means that $f(x)dx$ is the probability of finding the particle in a small interval of length $dx$. In Figure 3.9b we divided the entire range [0,L] into 20 intervals. As we can see the probability of being in any interval

$dx$ is simply $dx/L$. The corresponding SMI is $\log_2 20$. For the continuous case it is $\log_2 L$. (For details see Note 11, box 3.1 and Ben-Naim 2017a).



The equilibrium distribution of locations of one particle in a one dimensional system: $f^*(x) = {}^1/_L$

$H_{max}$(location of one particle along the $x$ − axis) = log L

The equilibrium distribution of locations of one particle in a 3D system of volume $V$: $f^*(x, y, z) = {}^1/_V$

The corresponding SMI is: log $V$

The SMI for N independent and distinguishable particles is
$SMI^D_{max}(\boldsymbol{R}^N) = NSMI_{max}(x, y, z) = N \log V$

Box 3.1

We note here that log $L$ is to be understood as the SMI associated with the *location* of a particle in the 1D system of length $L$. The larger the $L$, the larger is the SMI. One should keep in mind that the SMI for the continuous case is actually infinity. In passing from the continuous case, Figure 3.9a to the discrete case, Figure 3.9b, we reduce the number of outcomes to 20.

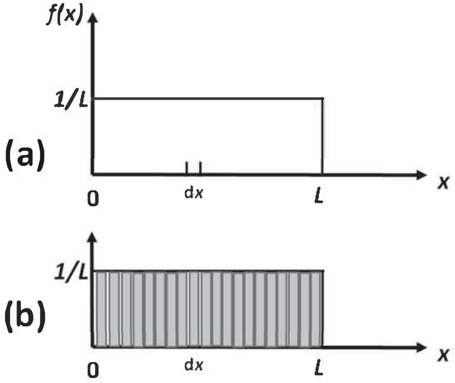


Figure 3.9. (a) The continuous uniform distribution of location.
(b) The probability offinding a particle in a small interval d*x*, *is:* dx/L.
Altogetherwe have twenty "events" in (b)
compared with infinite events in (a).

Note that log here stands for the logarithm to the base 2. We shall treat *L* as a pure number (i.e. dimensionless). We shall ignore the units of length, as well as the units of any other quantity under the logarithm sign. In the final expression, we must have a pure number under the logarithm. To summarize, we started with one particle moving in 1D "box" of length L. We assume that the probability of finding the particle in any small interval is constant. This is actually the equilibrium distribution of locations in 1D. For this distribution, we calculate the corresponding SMI. You can interpret this SMI as the number of questions you need to ask to find out in which of the 20 "boxes" in Figure 3.9b, the particle is. We skipped one important mathematical detail. The calculated SMI is the maximum possible SMI for the locations of one particle in the 1D box.

### 3.3.3 The SMI of one particle in a box of volume $V$

The generalization of this result to the three-dimensional (3D) case is straightforward. Suppose the particle is confined to a cubic box of edge $L$ and volume $= L^3$. Because of the equivalence of the three axes, the SMI associated with the $y$-axis and the $z$-axis will be the same as the SMI associated with the $x$-axis. Furthermore, we assume that the events "being at a location $x$," "being at a location $y$," and "being at a location $z$" are independent events. Therefore, the SMI associated with the location $x$, $y$, $z$ within the cube of volume $V$ is the sum of the SMI associated with the three axes,[12] see Box 3.1.

The final result is that the probability of finding a particle in any small volume $dV$ is the same for any point in the system. This is the equilibrium locational distribution of one particle in a 3D system of volume $V$. The corresponding SMI is given in Box 3.1. Again, note that the final SMI is the *maximum* value of the SMI which corresponds to a system at equilibrium. For more details, see Ben-Naim (2017b).

### 3.3.4 Extending to N distinguishable particles

We next extend this result to the case of *N-independent* and *distinguishable* ($D$) particles. We also use the short hand notation: $\boldsymbol{R}_i = (x_i, y_i, z_i)$ for the locational vector of particle $i$, and $\boldsymbol{R}^N = \boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ for the locational vector of all the $N$ particles. $\boldsymbol{R}^N$ is a shorthand notation for the *locations* of all the $N$ particles.

Since the particles are assumed to be *independent*, the SMI of the $N$ particles is simply the sum of the SMI of all the single particles. In terms of the 20Q game this means that the number of questions one needs to ask in order to find the locations of all the $N$ particles is simply $N$ times the number of questions one needs for one particle in the box. The SMI for $N$ independent and distinguishable particles, denoted $\mathrm{SMI}^{\mathrm{D}}$, is given in Box 3.1.

Note that we added the superscript $D$ for *distinguishable* particles. We shall soon see that the fact that the particles are *indistinguishable* ($ID$) causes a *reduction* in the SMI of $N$ particles.

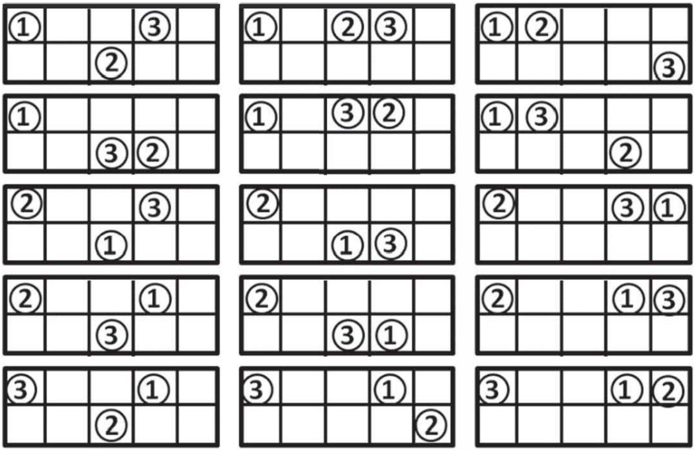


Figure 3.10. A few configurations of three different particles in ten cells. Try to draw some more of these configurations.

Before we continue with the ideal gas we discuss a simpler case of three particles in a 2D system of 10 sites, Figure 3.10. We assume that each site can accommodate one particle only. Therefore, the total number of configurations may be calculated as follows:

To place the first particle we have ten possibilities. To each of these possibilities we have nine possibilities for placing the second particle. Once we placed the two particles we have eight possibilities for the third particle. Altogether, we have:

number of possibilities

for three *distinguishable* particles on 10 sites $= 10 \times 9 \times 8 = 720$

Now, we introduce the principle of indistinguishability of the particles. This reduces the total number of configurations; instead of 720, we have $720/6 = 120$ configurations. Note, that every six distinguishable configurations reduce to one, when we remove the labels (numbers, colors, etc.). An example is shown on Figure 3.11.

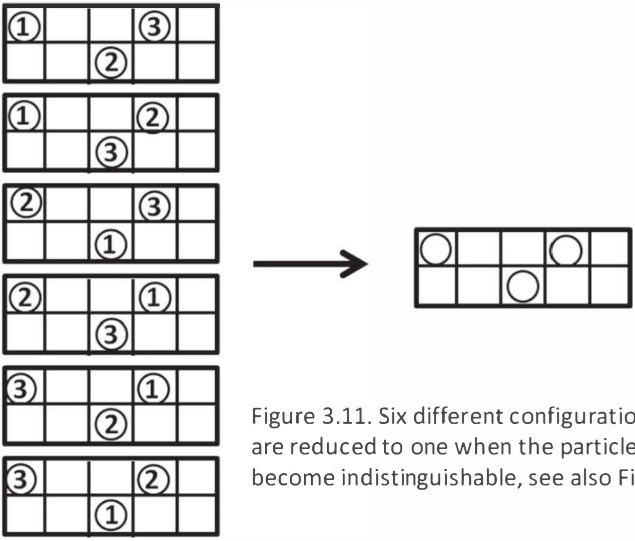


Figure 3.11. Six different configurations are reduced to one when the particles become indistinguishable, see also Figure 3.8.

In the case of gas molecules, we have to take into account both the *locations* and the *velocities* of the particles. Therefore, we shall next briefly and qualitatively add the distribution of velocities (or momenta, the momentum $p$ is related to the velocity by $p = mv$ where $m$ is the mass of the particles).

### 3.3.5 The velocity SMI of an ideal gas

Again, we start with particles moving along the 1D system of length $L$. We are interested in the probability density of finding a specific particle with velocity between $\nu_x$ and $v_x + dv_x$. We assume that the particles can have any value of $\nu_x$ from $-\infty$ to $+\infty$, but we require that the *average* kinetic energy of the particles is constant. Note, the kinetic energy is proportional to the *square* of the velocity, which is always positive, independently of the sign of the velocity. We shall skip the details of the derivation. We shall refer to $f^*(v_x)$ as the *equilibrium density distribution* of the velocities in one dimension.

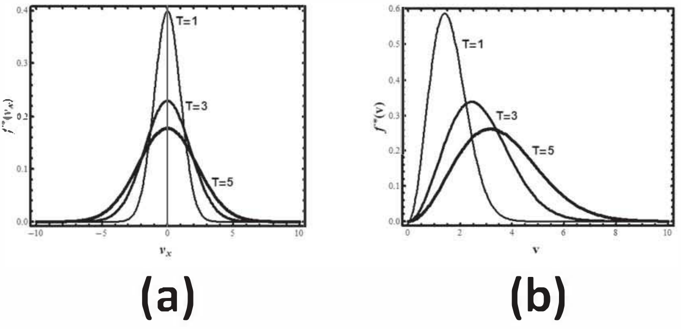


Figure 3.12. (a) The velocity distribution of particles in one dimension at different temperatures.
(b) The speed (or the absolute velocity) distribution of particles in 3D at different temperatures.

Figure 3.12 shows the distribution $f^*(v_x)$ for various values of $T$ (for this illustration we take $m = 1, k_B = 1$). We see that the larger the temperature, the larger the *spread* of the distribution of the velocities. Thus, the average "width" of the distribution may be described either by the variance $\sigma^2$ or by the temperature $T$.[13] As we have commented in the case of locations, we are not interested in the *exact* velocity of the particle, but only in the probability of finding a velocity in a small interval, say $dv_x$. The SMI associated with the equilibrium distribution $f^*(v_x)$ is given in Box 3.2.

**The equilibrium distribution of velocities of one particle in a one dimensional system:**

$$f^*(v_x) = \sqrt{\frac{m}{2\pi k_B T}}\exp\left[-mv_x^2/2k_B T\right]$$

**The corresponding SMI is:**

$$SMI_{\max}(v_x, v_y, v_z) = SMI_{\max}(v_x) + SMI_{\max}(v_y) + SMI_{\max}(v_z)$$

$$= 3H_{\max}(v_x) = \frac{3}{2}\log\left(2\pi\, e\, k_B T / m\right)$$

**The correction due to indistinguishability of the particles is:**

$$SMI^{ID}(1,2,\ldots,N) = SMI^{D}(1,2,\ldots,N) - \log N!$$

**The correction due to uncertainty principle is:**

$$SMI_{\max}(x, y, z, p_x, p_y, p_z) =$$

**Box 3.2**

$$SMI_{\max}(x, y, z) + SMI_{\max}(p_x, p_y, p_z) - 3\log h$$

**Pause and think:**

Try to understand: Why do we expect the SMI to be larger when the temperature is higher?

Think of the SMI as a measure of information as we discussed in sections 2.3 and 2.7. We shall try to understand this observation in terms of 20Q game.

In Figure 3.13 (see color centerfold), we show the velocity distribution at low temperature (T). We approximate the area under the curve by drawing a few rectangular bars. In Figure 3.13b, we show only the bars which were removed from under the curve. Next we add all the bars and form a single long bar as in Figure 3.13c. We do the same for the distribution of velocities at higher temperature, Figure 3.14 (see color centerfold). Note the different numbers of bars in the two figures.

Now consider the two bars in Figures 3.13c and 3.14c. Think of these two bars as two boards on which I throw a dart, and you have to find out in which region the dart hit by asking binary questions.

Which game do you think is easier to play? (by easier, I mean a fewer number of questions). See Figures 3.15a and 3.15b (see color centerfold).

If you have any difficulty in deciding between games (a) and (b) in Figure 3.15, look at the two extreme cases in Figure 3.16.

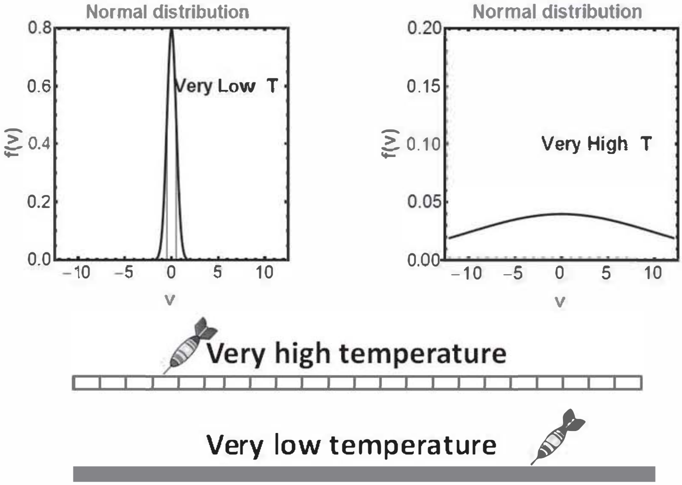


Figure 3.16. The velocity distribution at a very low and a very high temperature.

The important result we have obtained is that the larger the temperature (or equivalently the average kinetic energy of the particles) the larger is the SMI or the uncertainty associated with the distribution of the velocities.

We next assume that the velocities along the three axes $\nu_x, \nu_y, \nu_z$ are independent. Therefore, the SMI for a single particle moving with velocities $v_x, v_y, v_z$ is given by the sum of the SMI for each axis, see Box 3.2.

For the purpose of constructing the entropy of an ideal gas, we need the distribution of the momenta. This is simply obtained by the definition of the three components of the momentum: $p_x = mv_x$, $p_y = mv_y$, $p_z = mv_z$. We shall skip the details here. The interested reader can find the mathematical details in Ben-Naim (2008, 2017b, 2018).

### 3.3.6 A correction due to the indistinguishability of the particles

We have already seen in Figure 3.11 that indistinguishability of the particles *reduces* the total number of configurations of the system. This can be cast in the form of *mutual information*.

This term is a measure of the correlation between information we have on different experiments. Basically, if the two experiments are *independent*, then knowing the result of one experiment does not tell us any information on the second experiment. When knowing the result of one experiment provides information on the probability of the other experiment we say that there is correlation between the two experiments. In Box 3.1 we wrote the SMI of *N* particles as a sum of the SMI of all the particles presuming that they are independent. For the case of $N$ *indistinguishable* particles, we have to add a correction term, Log $N!$ which is given in Box 3.2.

We conclude that the indistinguishability of the particles introduces a correction which causes a reduction of the SMI. The reader should be convinced by checking a few examples, that whenever we "un-label" the particles, the number of configurations or arrangements is always reduced, and as a result the value of the SMI of the system is also reduced.

### 3.3.7 A correction due to the uncertainty principle

In Section 3.3.3, we calculated the locational SMI for a single particle. In Section 3.3.5, we calculated the velocity or the momentum SMI of a single particle. We now wish to find out the SMI associated with *both* the location and the momentum.

Classical thinking would have led us to conclude that the SMI associated with both the location ***and*** momentum of a particle should be the sum of the two SMIs. However, quantum mechanics tells us that the accuracies in determining the location and the momentum of a particle are *not independent*. This is the well-known Heisenberg *uncertainty* principle. For our case the uncertainty principle states that we cannot determine *both* the location and the momentum within an accuracy of the order of $h$, here $h$ is the Planck constant, $\boldsymbol{h} =$

$6.626 \times 10^{-34}\, Js$. Therefore, we must add a correction to account for the uncertainty principle. Again, the mathematical details are not important, just remember that this correction makes the total number of the configuration finite.

Once we take into account the uncertainty principle, we get the SMI for one particle in one dimension. For the three-dimensional case we have to add the correction term $-3\log \boldsymbol{h}$, i.e. we subtract log $\boldsymbol{h}$ for each coordinate $x, y, z$.

Finally, for $N$ indistinguishable and non-interacting particles, we have the final expression for the SMI of $N$ indistinguishable particles.

$$\mathrm{SMI}^{ID}(1,2,\cdots,N)$$

$$= \mathrm{SMI}(\boldsymbol{R}^N) + \mathrm{SMI}(\boldsymbol{p}^N) - \log N! - 3N\log \boldsymbol{h}$$

This is an important result. To obtain the SMI of $N$ particles described by their locations $(\boldsymbol{R}^N)$, and momenta $(\boldsymbol{p}^N)$, we first treat the particles as being *distinguishable* and *classical*. In this case, we can sum the SMI associated with all the locations SMI $(\boldsymbol{R}^N)$, and all the momenta SMI$(\boldsymbol{p}^N)$, of the particles. Then, we add two corrections, one due to the indistinguishability of the particles, and the other, due to the Heisenberg uncertainty principle. These two corrections *change* the total SMI by the amount $\log N! + 3N \log h$. Note again that at this stage we consider $h$ to be a pure number. We can fix its units in the final expression for the entropy, see below. I am well aware of the fact that since we skipped all the mathematical details, the meaning of the SMI in the above formula might not be clear. Therefore, I suggest to the reader to think of the total number of microstates or configurations and imagine playing the 20Q game on these microstates. The SMI will be simply the number of questions you need to ask in this game.

### 3.3.8 The entropy of a classical ideal gas

In the previous section, we calculated the value of the SMI of a system of $N$ simple, non-interacting, and indistinguishable particles at equilibrium.

Recall that the SMI may be defined for *any* distribution. It can be defined for any distribution of locations and any distribution of momenta, not necessarily at equilibrium. It can be defined for any number of particles and can be defined for distinguishable or indistinguishable particles. All these have nothing to do with the entropy. Up to this point you can rightfully regard the SMI as a quantity that measures the size of a huge 20Q game about all the locations and the momenta of all the particles in the system.

Here we are interested in *very special distributions*. These are the distributions of locations and momenta that *maximize* the corresponding SMI. We denoted these special distributions with asterisks, and the corresponding SMI by $H_{max}$. However, we also know that starting with any arbitrary distribution of locations and momenta, the system will tend to a limiting equilibrium distribution; the uniform distribution for locations, and the Normal distribution for the momenta. Therefore, we shall refer to these distributions as the *equilibrium distributions*. We shall soon see that these distributions are also the distributions which maximize the super-probability.[14] Later we shall see below that maximum SMI is related to entropy, and the corresponding maximum super-probability will be related to the Second Law.

In this section we make a huge conceptual leap, from SMI to a fundamental concept of thermodynamics. As we shall soon see, this leap is rendered possible by recognizing that the SMI associated with the equilibrium distribution of locations and momenta of a large number of indistinguishable particles is *identical* (up to a multiplicative constant) with the statistical mechanical or Boltzmann's entropy of an ideal gas. Since the statistical mechanical entropy of an ideal gas has the same properties as the thermodynamic entropy as defined by Clausius, we can declare that this special SMI is identical to the entropy of an ideal gas. This is a very remarkable achievement. Unfortunately, some people still refer to the SMI as entropy. This was (and still is) a mistake. In general, SMI is not entropy. Only when you apply the SMI to a *special* distribution does it become identical with entropy.

We recall that the SMI of a system of $N$ particles at equilibrium has two contributions due to location and momentum, and two corrections due to the indistinguishability of the particles and the uncertainty principle.

In order to obtain the expression for the entropy of an ideal gas, all we have to do is to use the natural logarithm and multiply SMI by the Boltzmann constant $k_B$, i.e. $S = (k_B \ln 2)\text{SMI}$. Thus, we define the entropy of an ideal gas of simple particles as the value of the SMI associated with the equilibrium distribution of locations and momenta. See Box 3.3.

$$
\begin{aligned}
&SMI^{ID}(1,2,\cdots,N)\\
&= SMI_{max}(locations) + SMI_{max}(momenta)\\
&- I(uncertainty\ principle)\\
&- I(indistinguishable)\\
&\qquad = N\log\left[\frac{V}{N}\left(\frac{2\pi mk_BT}{h^2}\right)^{3/2}\right] + \frac{5N}{2}
\end{aligned}
$$

$$
S(ideal\ gas) = Nk_B \ln\left[\frac{V}{N}\left(\frac{2\pi mk_BT}{h^2}\right)^{3/2}\right] + \frac{5Nk_B}{2}
$$

Box 3.3

The multiplication by a constant $k_B$, as well as changing to the natural logarithm determines the *units* in which we measure entropy. It does *not* affect the meaning of entropy, as the SMI associated with the location and the momenta of a system of *N* non-interacting particles at equilibrium. We normally apply this identity between entropy and SMI for a thermodynamic system, i.e., when *N* is very large. This is important when we formulate the Second Law of Thermodynamics.

### 3.3.9 Conclusion: What is entropy?

Before you read this section remind yourself of the following concepts: The SMI, the probability distributions of the locations and velocities of all particles at equilibrium and the corresponding SMI. To obtain the entropy of a system of non-interacting particles we start with the SMI defined on the locational and the velocity distributions of all the particles. Since the locations and the velocities of all the particles determine the *microscopic* state of the system, we can say that the SMI is built on the probability distribution of the *microscopic* states of the classical system.[15]

To obtain the entropy of the system from the SMI, we first calculated the specific distributions that *maximizes* the SMI. These are referred to as the *equilibrium distributions.* This is done by using the calculus of variations (see Ben-Naim, 2008). In order to obtain the entropy of the system, we simply multiply the resulting SMI by a constant. This constant consists of the Boltzmann constant $(k_B)$, and the conversion from the logarithm base 2, to the natural logarithm. Once we do this we get the *entropy* of an ideal gas.

This is the most amazing result. Starting with a quantity defined by Shannon in *communication theory*, a quantity which has nothing to do with physics, we got the *entropy* of a system defined in thermodynamics.[16]

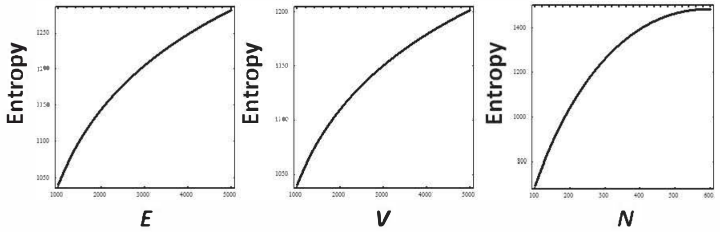


Figure 3.17. The dependence of the entropy on *E*, *V* and *N*.

In Figure 3.17, we show that the entropy as a function of the energy *E*, the volume *V*, and the number of *N* particles obtained from the SMI. Note that the entropy is an increasing function of *E*, *V* and *N*. Also, it has a negative *curvature*. This means that the slope of the curve becomes smaller as *E* (or *V or N*) increases. This is an important characteristic property of the entropy. For more details, see Ben-Naim (2012, 2016b).

After learning about probabilities in Chapter 1, and about SMI in Chapter 2, and after deriving the entropy of a classical ideal gas of simple particles in this chapter,[17] we are ready to discuss the *meaning* of the entropy of the system. Recall that if you have an experiment (or a game) with *n* outcomes having a probability distribution $p_1, \cdots, p_n$, the SMI of such a system is a measure of the average uncertainty associated with all these outcomes. It is also a measure of the average unlikelihood associated with all the outcomes. It is also a measure of the amount of information associated with the distribution of the outcomes. All these interpretations are equivalent. Also remember that the value of the SMI does not depend on who plays the game, or who performs the experiment. It is a quantity which is *associated* with, or *belongs* to the experiment, when all we know about the experiment is the distribution of outcomes.

Now, for an ideal gas (of simple particles) the *microscopic* description of the system comprises the sequence of 6N numbers. There are three locations $(x, y, z)$, and three velocities $(v_x, v_y, v_z)$ for each particle. Therefore, we have six numbers to describe the microstate of each particle. For $N$ particles we need $6N$ numbers to describe the *microstate* of the entire system.

We shall call this microstate a *configuration* of the system. Recall that because of the uncertainty principle we have only a finite number of configurations. Also, remember that the particles are indistinguishable, therefore, the total number of configurations is reduced when we erase the labels on the particles. [18]

We can calculate the corresponding SMI of the system for any distribution of locations and velocities. This is the average uncertainty, or the amount of information associated with the particular distribution of the configurations of the system. In order to obtain the entropy we need to calculate the SMI associated with the distribution at *equilibrium*, then multiply by a constant factor $(k_B \ln 2)$. The multiplication by a constant factor does not affect the meaning or the interpretation of entropy as an SMI. However, it is most important to remember that the entropy is associated with the *equilibrium distribution*. As we have seen the equilibrium distribution of locations is the uniform distribution,[19] and the equilibrium distribution of velocities is the Normal distribution.

Thus, the entropy of a system is the SMI of the *N* (simple) particles at equilibrium (after multiplications by the constant $k_B \ln 2$).

If you like you can imagine playing the 20Q game on this system. You have 3*N* boxes (We have altogether *6N* coordinates, but *each* "box" consists of one coordinate of location and one coordinate of velocity. Therefore, the number of "objects" for the 20Q game is 3*N*). When *N* is of the order of $10^{23}$, this is already a huge number, but the number of configurations is far greater than $10^{23}$. The number of configurations is so large that you cannot possibly play this game in your lifetime, nor if you were to live billions and billions of years – far longer than the estimated age of the universe. However, no matter how big the game is, you can always imagine playing it, and you can also imagine how many questions you will need in order to find out in which microstate the system is, when you are given the distribution of these microstates. I should add here that you do not actually need to play this huge 20Q game. You only have to know that there is a relationship between the distributions of the "objects," and the minimal number of binary questions.

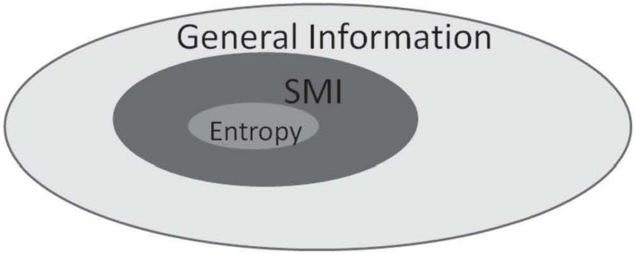
At this point it is important to emphasize again that the SMI is defined for any distribution of the microstates. The entropy (as an SMI multiplied by a constant) is defined only for the equilibrium distribution.

Exercise: Suppose that I tell you that the distribution of all the microstates is uniform (meaning, that each microstate has the same probability), and that altogether there are $W$ microstates to the system. Can you estimate its entropy? [20]

Note however, that no matter how you calculate the entropy of the ideal gas it can still be interpreted as an SMI, that is, it is still a measure of the *average* uncertainty (or unlikelihood) with respect to the distribution of all microstates. Equivalently, it is the amount of information associated with the distribution of microstates. The *amount* of *information* should not be confused with *information* about the state of the system. Such confusion is the rule, rather than the exception in many popular science books dealing with entropy.

Pause and think: Are you sure you understand the last two sentences?

Now that you know the difference between "Information" and SMI on one hand, and the difference between SMI and entropy, on the other hand, it is time to pause and "orient" the entropy within the general concept of information.



A qualitative relationship between the "sizes" of the three concepts: General Information, SMI and Entropy.

Look at the schematic diagram above. The outer (yellow in the color centerfold) region in this figure represents all possible pieces of information, all what you know and what everyone knows, all what is written in books or recorded in any device –

in short, all kinds of information. This is a vast region, whose boundaries no one knows. Yet, it does not include *all things* that exist. (I am insinuating here the famous slogan "it from bit" meaning that everything, every "it" is information, i.e. bit. For criticism of this slogan, see Ben-Naim (2015a, 2016a).

Within this vast region denoted "General Information," I drew a sub-region (colored red, in the color centerfold) which is denoted SMI. This region does not include *values* of SMI, but all possible information associated with well-defined probability distribution. In this region you will find the information associated with the distribution of the outcomes of dice, coins, and many more. As we have seen for each of these distributions one can define a SMI. This is the reason I denoted this region by SMI.

Next, we further narrow the set of all possible types of information; information associated with the distributions of the locations and velocities of *N* particles at equilibrium. As we have seen on these particular distributions we can define an SMI. Therefore, the corresponding information is a subset (blue region in the color centerfold) of the subset denoted SMI (red region, see color centerfold).

If we take the distributions included in the blue region and the corresponding SMI, then multiply each of these SMI by a constant $[k_B \ln 2]$ we get the entropy of that particular distribution. This is the reason for denoting the blue region, Entropy.

Thus, entropy is defined on a small subset of all possible distributions for which the SMI can be defined. The SMI is associated with any well-defined distribution – which is a subset of all possible information.

Until now we have discussed only systems of simple, non-interacting particles. One can extend the concept of SMI, as well as the entropy to also take into account the interaction energies among the particles, as well as the internal states of each particle. All these are technical details which are not relevant here. What is important to remember is that the *meaning* of entropy as a special case of SMI (after multiplication by a constant) is unchanged by these generalizations.

Before we turn to the Second Law, we should note that all we have said so far is valid for any number of particles *N*. *N* could be one, ten, one thousand, or $10^{23}$ particles. However, if we want the system of the *N* particles to obey the Second Law as described in Section 3.5, we should define the entropy only for macroscopic systems, i.e. when $N$ is a very large number.

Note carefully that we have discussed the concept of entropy and interpreted entropy without any reference to the Second Law. It is unfortunate that many authors define entropy in terms of the Second Law, and formulate the Second Law in terms of the entropy.

## 3.4 Examples

Until now we discussed the entropy of a system at equilibrium. Before we discuss the Second Law which applies to *processes*, we discuss in some details a few processes for which we can easily calculate the corresponding entropy changes.

### 3.4.1 Expansion of an Ideal Gas in an Isolated System

Consider the simplest spontaneous process. We start with an ideal gas confined to a volume $V$, we remove a partition, and observe that the gas will expand and fill the entire new volume $2V$. Figure 3.18.

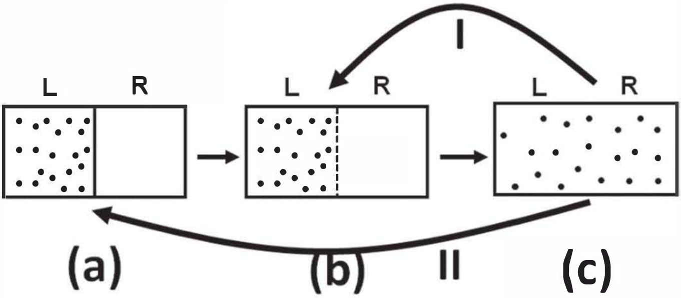


Figure 3.18. The initial, the final, and an intermediate state in the expansion process.

Ask any student who has learned thermodynamics: Why did the gas expand from $V$ to $2V$, and why it will never go back to the original volume $V$? The answer you are most likely to hear is that the *cause* of that expansion is the tendency of the entropy to increase. Equivalently, one would invoke the Second Law as the *cause* of this spontaneous process.

If you ask why the entropy tends to increase, the immediate answer would be: That is exactly what the Second Law states!

Does the tendency of the entropy to increase *drive* the spontaneous process, or does the spontaneous process drive the entropy upwards?

We will answer this question by examining the expansion process with different number of particles. In all the following examples we shall assume that the particles do not interact with each other (or that interactions are negligible), and that in each case only the locational distribution of the particles is changed. Each particle is initially located within the boundaries of a volume $V$, and in the final state it is located in the larger volume $2V$.

There are essentially two questions associated with the Second Law that are oftentimes muddled. The first, why a system evolves spontaneously from one state to another, the second, why entropy increases. Although the answers to these two questions are different, they are related to each other. They are related *only* in processes which occur in isolated systems, i.e. systems that do not interact with their surroundings.

The entropy formulation of the Second Law of Thermodynamics states that in any spontaneous process occurring in an isolated system the entropy increases. We shall further discuss this formulation of the Second Law in Section 3.5. The Second Law does not state anything as to why the entropy increases, nor addresses the question as to why a spontaneous process occurs at all. Note again that by any "process," we mean a process from an initial equilibrium state to a final equilibrium state, e.g. from Figure 3.18a to Figure 3.18b. We shall soon discuss the question of reversing the process from (c) to either (b) (possible but very improbable) or to (a) (absolutely impossible).

Accepting the relative frequency interpretation of probability, we can conclude that a system will be found more frequently in states which have higher probability. Specifically, for a thermodynamic system we will see that the probability of finding the system in a state of equilibrium is almost one. Thus, we can say that a thermodynamic system at any initial state will always evolve towards a state of higher probability, and eventually reach a state we call *equilibrium state*, the probability of which is almost one. (For details, see Ben-Naim, 2008, 2012).

Again, we stress here that we are talking about spontaneous processes in an isolated system, having a fixed energy, volume, and number of particles.

Answering the question of *why* the system evolves towards the state of equilibrium leaves the question of why entropy increases unanswered. Nevertheless, because of the intimate relationship between the SMI of the system and the probability of the state of the system, the answers to the two questions are also related to each other. This is discussed in the next subsections.

### 3.4.2 What drives the system to an equilibrium state?

In this section, we shall answer the question of what drives the system to an equilibrium state by examining a few simple examples.

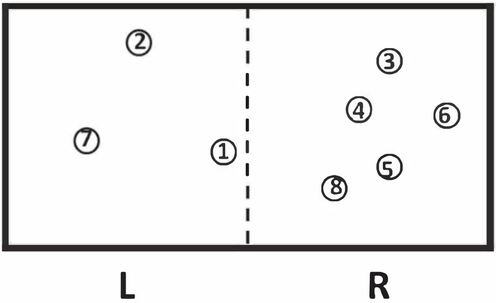


Figure 3.19. Specific configuration of eight particles in the two compartments. Here, particles 1, 2, and, 7 are in L, and 3,4,5,6, and 8 are in R. A generic configuration is obtained when we have three particles in L and five in R.

Consider a system of $N$ non-interacting particles (ideal gas) in a volume $2V$ at constant energy $E$. We divide the system into two compartments L and R, each of volume $V$ (see Figure 3.19). We define the *specific* microscopic state of the system when we are given $E, 2V, N$, and in addition we know which *specific* particles are in the right compartment (R), and which specific particles are in the left compartment (L). The generic description of the same system is $(E, 2V, N; n)$ where $n$ is the ***number*** of particles in the compartment $L$. Thus, in the specific description, we are given a *specific* configuration of the system as if the particles were labeled $1,2,\cdots,N$. Here, by configuration we mean only which particles are in R and which are in L. (This is different from the more detailed microscopic configuration we discussed in terms of all locations and velocities). In the *generic* description, we are given the information only on the *number* of particles in each compartment.

Clearly, if we know only that there are $n$ particles in L, and $N-n$ particles are in R, we have many *specific* configurations that are consistent with the requirement that there are $n$ particles in L.[25]

We denote by $W(n)$ the number of *specific* configurations consistent with $n$ particles in L. The first postulate of statistical mechanics states that all specific configurations of the system are equally probable. The total number of specific configurations is $2^{N}$, i.e. each particle can be in either one of the two compartments. Using the classical definition of the probability, we can calculate the probability of finding $n$ particles in L and $(N-n)$ particles in R. We denote this probability by $P_N(n)$. It is easy to show that both $W(n)$ and $P_N(n)$ have a maximum as a function of $n$ at the point $n^* = \frac{N}{2}$

(see below). The maximum value of the probability $P_N(n)$ (obtained at $n^* = \frac{N}{2}$), and is denoted by $P_N(n^*)$.

Thus, for any given $N$, there exists an $n$, such that the number of configurations, $W(n)$, or of the probability, $P_N(n)$ is maximal. Therefore, if we prepare a system with any initial distribution of particles $n$, and $N - n$ in the two compartments, and let the system evolve, the system's state will change from a state of lower probability to a higher probability. As $N$ increases, the value of the maximum number of configurations $W(n^*)$ *increases* with $N$. However, the value of the maximal probability $P_N(n^*)$ *decreases* with $N$.

To appreciate the significance of this fact, we will examine the "evolution" of systems with small numbers of particles. We shall soon see in what sense the spontaneous process of expansion proceeds in "one direction only," or is "irreversible." Later, we shall also follow the changes in the SMI in the process of expansion and finally, we shall calculate the entropy change for this process. For some simulations, see Ben-Naim (2008, 2010).

### The case of two particles: N = 2

Suppose we have the total of $N = 2$ particles. In this case, we have the following possible configurations and the corresponding probabilities:

$$\begin{array}{ccccc} n = 0 & , & n = 1 & , & n = 2, \\ P_N(0) = \frac{1}{4} , & & P_N(1) = \frac{1}{2} , & & P_N(2) = \frac{1}{4} \end{array}$$

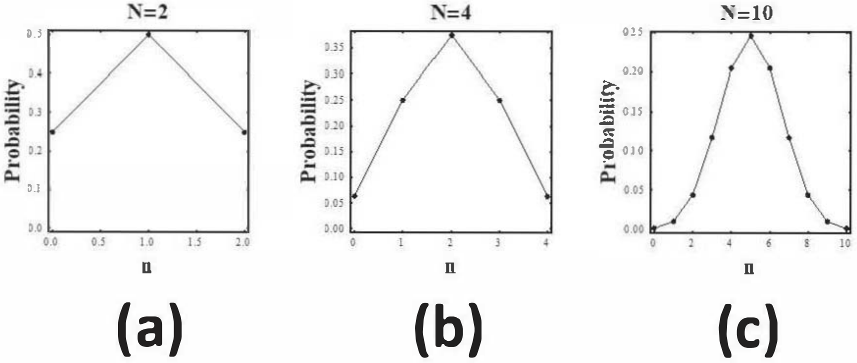


Figure 3.20. Probability of observing *n* particles in one compartment and *N*-*n* in the other for different numbers *N*.

This means that on the average, we can expect to find the configuration $n = 1$ (i.e. one particle in each compartment) about half of the time, but each of the configurations $n = 0$ and $n = 2,$ only a quarter of the time (see Figure 3.20a). If we start with all the particles in the left compartment, and remove the partition between the two compartments, we shall find that the system will "expand" from $V$ to $2V$. However, once in a while the two particles will be found in one compartment. In this case, you will have no sense that the process of "expansion" occurs in one direction; if you do this experiment and take a video of all the configurations, then run the video either forward or backward, you will not be able to tell the difference.

### The case of four particles: N = 4

For the case $N = 4$, we have the distribution as shown in Figure 3.20b. The maximal probability is $P_N(2) = \frac{6}{16} = 0.375$ which is *smaller* than $\frac{1}{2}$. In this case, the system will spend only $\frac{3}{8}$ of the time in the maximal state $n^* = 2$. Again, if we start with all particles in one compartment, the system will "expand" from $V$ to $2V$, but once in a while we shall see all the particles in one compartment. Again, you will not feel the process running in one direction even if you run the video either forward or backward, you will not be able to tell the difference.

### The case of ten particles: N = 10

For $N = 10$, the distribution is shown in Figure 3.20c. We calculate the maximum at $n^* = 5$ which is $P_{10}(n^* = 5) = 0.246$.

In all of the three examples we examined above, the system *expands* from $V$ to $2V$. However, there is nothing in this process which can be termed "irreversible." In each case, the initial state will be visited once in a while.

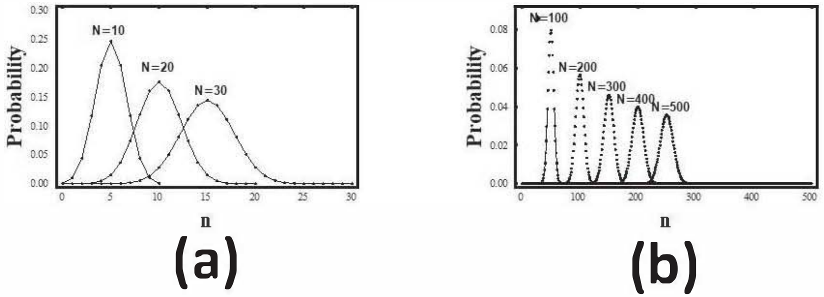


Figure 3.21. Probability of observing *n* particles in one compartment, and *N*-*n* in the other.

### Very large number of particles

Let us proceed with larger *N*. Figure 3.21 shows the probabilities $P_N(n)$ for larger number of particles. It is seen that the maximum value of $P_N(n)$ *decreases* as *N increases*.[21] You can skip the next paragraph on the first reading.

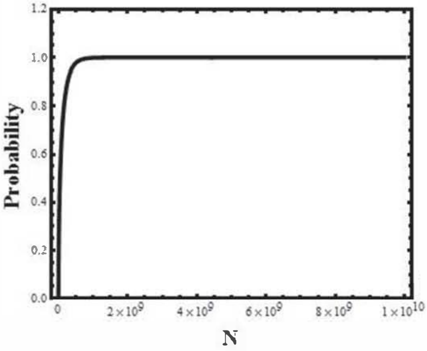


Figure 3.22. The probability of finding *n* particle in the neighborhood of *n**= *N*/2, in one compartment as a function of ***N***.

In Figure 3.22, we show the probability of finding $\boldsymbol{n}$ between $n^* - 0.0001\, N \leq n \leq n^* + 0.0001\, N,$ as a function of *N*. Plotting the probability $P_N(\, n^* - 0.0001\, N \leq n \leq n^* + 0.0001\, N)$ as a function of *N* shows that this probability tends to *one* as *N* increases. When *N* is on the order of $10^{23}$, we can allow deviations of $\pm 0.00001\%$ of *N*, or even smaller, yet the probability of finding $\boldsymbol{n}$ at or near $n^*$ will be almost one. It is for this reason that when the system reaches $n^*$or near $n^*$, it will stay in the vicinity of $n^*$ for most of the time. For *N* on the order of $10^{23}$, "most of the time" means practically ***always***.

The abovementioned specific example provides an explanation for the fact that the system will "always" evolve in "one direction," and "always" stay at the equilibrium state once that state is reached. The tendency towards a state of larger probability is equivalent to the statement that events which are supposed to occur more frequently will occur more frequently. This is plain common sense. The fact that we do not observe deviations from either the monotonic climbing of $n$ towards $n^*$, or staying close to $n^*$, is a result of our inability to detect small changes in $n$-(or equivalently small changes in the SMI, see below). Note that in this section we did not say anything about the entropy changes. Before moving on to calculate the entropy changes we repeat the main conclusion of this section. For each $N$ the probability of finding a distribution of particles; $(n, N - n)$ in the two compartments L and R has a maximum at $n^* = \frac{N}{2}$. For a very large number of particles the probability of obtaining the *exact* value of $n^* = \frac{N}{2}$ is not very large. However, the probability of finding the system at a small vicinity of $n^* = \frac{N}{2}$ is almost one!

When we say that the system has reached an equilibrium state, we mean that we do not *see* any changes that occur in the system. In this example, we mean changes in the *density* of the particles in the entire system. In other experiments when there is heat exchange between two bodies we characterize the equilibrium state as the one for which the temperature is uniform throughout the system and does not change with time.

At equilibrium the macroscopic density we measure at each point in the system is constant. In the particular system we discussed above the measurable density of the particles in the two compartments is $\rho^* \cong N/2V$. Note that fluctuations always occur. Small fluctuations occur very frequently, but they are so small that we cannot measure them. On the other hand, fluctuations that could have been measured are extremely infrequent, and practically we can say that they never occur. This conclusion is valid for very large $N$.

### The evolution of the SMI in the expansion process

Next, we will discuss the relationship between the probabilities of finding a particular generic state and the formulation of the Second Law in terms of the entropy. We rewrite the essential quantities of the example discussed above in a slightly different way. Instead of $n$ and $N-n$, we define the fractions $p = \frac{n}{N}$ , $q = (1-p) = \frac{N-n}{N}$. $p$ is the fraction of particles in the L compartment and $q = (1-p)$ is the fraction in the R compartment. Clearly, the pair of numbers $(p, 1-p)$ is a probability distribution.

We now quote an important relationship between the SMI of the system and the super-probabilities $\Pr(p, q)$. See Box 3.4 and Ben-Naim (2016a, 2016b, 2017b, 2017c).

$$\Pr(p,q)=\left(\frac{1}{2}\right)^{N}\frac{2^{N\times SMI(p,q)}}{\sqrt{2\pi Npq}}$$

**Box 3.4**

This is the relationship between the SMI defined *on the distribution* $(p,q)$, and the probability $\Pr(p,q)$ defined *on the same distribution*. We see that there exist a monotonic relationship between Pr and SMI. This means that whenever SMI increases, the Pr also increases, and at equilibrium both SMI and Pr attain a maximal value. We have seen that the maximal value of SMI is related to the entropy of the system. Therefore, the answer to the question of *why* the entropy increases (in a spontaneous process in an isolated system), is the same as the answer to the question of why the state of the system evolves towards equilibrium, namely; it is because the probability Pr of the equilibrium state is maximum.

Note carefully the two "levels" of probabilities. One is the probability distribution of a state described by $(p,q)$. Pr is the probability of finding a state described by the probability distribution $(p,q)$. To distinguish between the two probabilities, I refer to Pr as a *super probability*. Note also that the answer to the question: "Why the system evolves towards equilibrium?" is provided by the probability Pr. Because of the monotonic relationship between Pr and the SMI the answer to the question "why the entropy increases" is also probabilistic. It is easy to generalize this conclusion to the case of any number of compartments. See Ben-Naim (2008, 2012).

We can calculate the change in the SMI for this process, and find that it is $N$ bits. Initially, we were certain that all particles are in L. After removing the partition, we lost one bit per particle; we know that the particle is in either in L, or in R. From this value of the change in SMI, we can calculate the entropy change by multiplying $N$ by $k_B \ln 2$.

## Summary of facts

We summarize what we have found so far from the simple examples of expansion of $N$ particles from volume $V$ to $2V$.

For any $N$, right after removing the partition we follow the evolution of the system with time. In all the examples, we observed that the particles which were initially confined to one compartment of volume $V$ can access the larger volume $2V$. We can ask the following questions:

1. Why do the particles occupy the larger volume?
2. Does the number of particles in the left compartment change monotonically with time? (monotonically means that the quantity either increases, all the time, or decreases all the time)
3. Does the number of particles in the left compartment reach an equilibrium state?
4. How fast does the system reach the equilibrium state?
5. How does the SMI of the system change with time?
6. How does the entropy change with time?

I urge the reader to try to answer these questions before continuing. Note that the answers to all these questions depend on $N$. Here are the answers to the questions:

1. The reason the particles will occupy the larger volume $2V$ rather than $V$ is that the probability of the states where there are about $N/2$ in each compartment is larger than the probability of the state where all the particles are in one compartment. This is true for any $N$. When $N$ is very small there is a relatively large probability that the particles will be found in one compartment. For these cases we cannot claim that the process is irreversible, in the sense that it will *never* go back to the initial state. For large $N$, even of the order $10^6$, the probability to return to the

initial state becomes so small, that it is practically zero. However, there is always a *finite* probability that the system will visit the initial state. For $N$ or the order of $10^{23}$, the probability of visiting the initial state is so small (but still non-zero) that we can safely say that the system will *never* return to the initial state. Never, in the sense of *billions* of *ages* of the universe.

2. The number of particles in L, *n,* does not change monotonically from $N$ to $N/2$ (or from zero to $N/2$ if we start with all particles in the right compartments). Simulations show that for large values of $N$ the number $n$ changes *nearly* monotonically towards $N/2$. The larger the $N$, the more monotonic is the change of $n$. (For simulated results, see arienbennaim.com, books, Entropy Demystified, and simulated games.) For $N$ on the order of $10^6$ or more, you will see nearly perfect, smooth, monotonic change in $n$.

3. The answer to this question depends on how one defines the equilibrium state of the system. If we define the equilibrium state when the value of $n$ is *equal* to $N/2$, then, when $n$ reaches $N/2$ it will not stay there "forever." There will always be fluctuations about the value of $n^* = N/2$. However, if one defines the equilibrium state as the state for which $n$ is in the

neighborhood of $n^* = N/2$, then, we will find that once $n$ reaches this neighborhood, it will stay there for a longer period of time than in any other state. For $N$ of the order of $10^6$ or more, the system will stay in this neighborhood “forever.” Forever, again means here many ages of the universe.

4. The answer to this question depends on the temperature and on the size of the aperture we open between the two compartments (in the experiment of Figure 3.18 we remove the partition between the two compartments. However, we could do the same experiment by opening a small window. In such an experiment, the speed of reaching the equilibrium state would depend on the size of the aperture of the window). In any case thermodynamics does not say anything about the speed of attaining equilibrium.

5. For each *distribution* of *particles* $(n, N - n)$ we can define a *probability* distribution $(p, 1 - p)$, and the corresponding SMI. As the system evolves from the initial to the final state, $n$ will change with time, hence, $p$, as well as the SMII, will also change with time. For simulations, see Ben-Naim, 2010.

Can you please explain what this SMI means?

For small $N$, the SMI will start from zero (all particles being in one compartment) and will fluctuate between zero to $N$ bits. When $N$ is very large, say $10^6$ or more the value of SMI will change nearly monotonically from zero to $N$ bits. There will always be some fluctuations in the value of SMI, but these fluctuations will be relatively smaller the larger $N$. Once the system reaches the equilibrium state it will stay there *forever*. Note carefully that the SMI is defined here on the probability distribution $(p, 1-p)$. For the initial distribution $(1{,}0)$ the SMI is zero. The SMI defined on the distribution of locations and momenta is not zero.

6. The answer to this question is the simplest, yet it is the most misconstrued one. It is the simplest because entropy is a *state* function, it is defined for a well-defined macroscopic (or thermodynamic) *state* of the system. For the expansion process, the macro-state of the system is defined initially by $(E, V, N)$. The corresponding value of the entropy is $S(E, V, N)$. The final macro-state is characterized by $(E, 2V, N)$, and the corresponding value of the entropy is $S(E, 2V, N)$. In between the two macro-states $(E, V, N)$, and $(E, 2V, N)$ the macro-state of the system is not well defined.

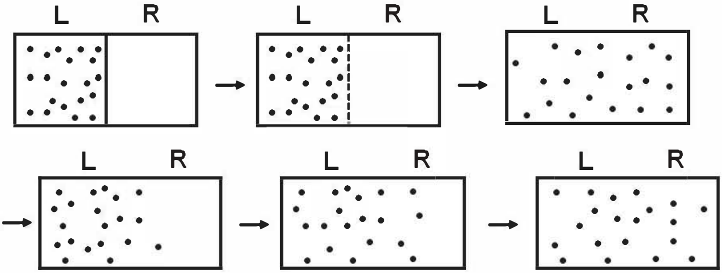


Figure 3.23. The initial, the final, and some intermediate states in the expansion process.

A few intermediate states are shown in Figure 3.23. While $E$ and $N$ are the same as in the initial state, the "volume" during the expansion process of the gas is not well defined. It becomes well defined only when the system reaches an equilibrium state. Therefore, since the volume of the system is not well defined when the gas expands, the entropy is also not well defined. We can say that the entropy changes abruptly from $S(E, V, N)$ to $S(E, 2V, N)$, and that this change occurred at the moment the system reaches a final equilibrium state. This is shown schematically in Figure 3.24a.

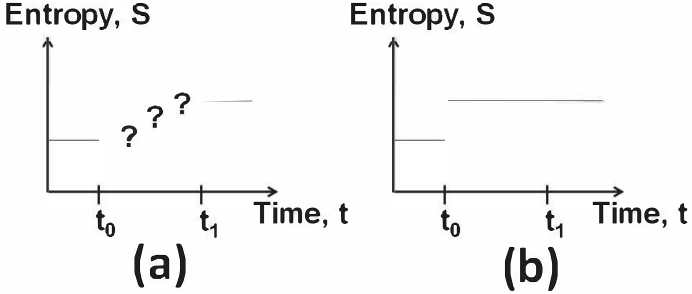


Figure 3.24. Two views of the entropy change after the removal of the partition.

One can also adopt the point of view that when we remove the partition between the two compartments, the volume of the gas changes abruptly from $V$ to $2V$, although the gas is initially still in one compartment, the total volume $2V$ is *accessible* to all particles. If we adopt this view, then at the moment we removed the partition, the volume changes from $V$ to $2V$, and the corresponding change in entropy is $S(E,2V,N) - S(E,V,N)$. This change occurs abruptly at the time we remove the partition, see Figure 3.24b. Both of these two views are acceptable. Personally, I used to prefer the first point of view. Initially, it has the value $S(E,V,N)$ before the removal of the partition, and it reaches the value of $S(E,2V,N)$ when the systems reach the new, final equilibrium state. In all of the intermediate states the entropy is not defined. Note however, that the SMI is defined for any intermediate states between the initial and the final states. However, the entropy is the maximal value of the SMI (multiplied by the Boltzmann constant and change of the base of the logarithm), reached at the new equilibrium state.[22]

It should be noted however that we could devise another expansion (referred to as quasi-static) process by gradually moving the partition between the two compartments. In this process the system proceeds through a series of equilibrium states, and therefore the entropy is well-defined at each of the points along the path leading from $(E, V, N)$ to $(E, 2V, N)$. In this process, the entropy of the gas will gradually change from $S(E, V, N)$ to $S(E, 2V, N)$, Figure 3.25. The length of time it takes to proceed from the initial to the final state depends on how fast, or how slow we carry out the process.

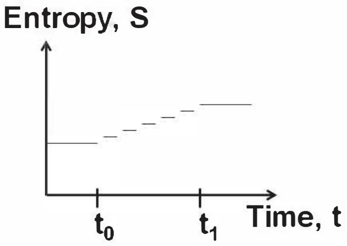


Figure 3.25. The entropy change in a quasi-static process.

Note that the sequences of states in the spontaneous process are different from the quasi-static process. In the latter, the states as well as the entropy of the gas are well-defined along the entire path from the initial to the final equilibrium states, whereas in the spontaneous expansion neither the *states*, nor the entropy are defined along the path leading from the initial to the final state.

### 3.4.3 The spontaneous mixing of two ideal gases

This example is brought up frequently in textbooks to demonstrate that the mixing of two gases is irreversible, and that a positive change in entropy corresponds to increase in disorder. We shall first analyze this process then we shall see that both of these contentions are incorrect.

The process is quite simple. We have two different gases A and B in two different compartments, Figure 3.26a (see color centerfold). We remove the partition between the two compartments and observe *mixing*. We always observe mixing from *a* to *b*, to *c*. We *never* observe the reversal of this process, i.e. the mixture of A and B will never un-mix. For simplicity, we assume that there are $N$, A-particles in a volume $V$, and $N$ B-particles in volume $V$.

We now redraw Figure 3.26, in Figure 3.27 (see color centerfold). and ask the following three questions:

1. Can the system in state *c* of Figure 3.27 return to state *b*?
2. Can the SMI in state *c* attain the value as in state *b*, or *a*?
3. Can the entropy of the system decrease from the value it has in *c* to the value it had in *a*?

Although we shall not prove it here, let me tell you that the process in Figure 3.26a is nothing but two expansion processes; each gas expands from the initial volume $V$, to the final volume $2V$.

The answer to the first question is a Yes. The system can visit the initial state *b*, i.e. process $c \rightarrow b$ can occur. This will occur with so small probability that in *practice* we can say that the process $b \rightarrow c$ is irreversible. This is irreversible, in *practice*, not absolute!

Regarding the change in the SMI, again as we noted in connection with the expansion process, since the system can be reversed from *c* to *b* the SMI can be reduced. Since the mixing in the process $a \rightarrow b \rightarrow c$ involves the expansion of two ideal gases from $V$ to $2V$, we can calculate that the change in the SMI will be 2*N*, i.e. one bit per particle (in the process of expansion, we had *N* particles, here we have 2*N* particles). Although this reversal is extremely improbable, it can occur and the SMI can attain its initial value at state $b$ (note that here we refer to the SMI based on the distribution of locations and velocities of all the 2*N* particles in the system). The entropy change in this process of mixing is the same as the change in the SMI, multiplied by $k_B$ ln 2.

Again, we emphasize that the entropy change in this process $(a \rightarrow c)$ refers to the two states $\boldsymbol{a}$ and $c$, which are equilibrium states (in either Figure 3.26a or Figure 3.27 see color centerfold). Once we are in state $c$, the system can return to state $b$ (it can, but with a very small probability), but it cannot return to state $\boldsymbol{a}$. For this to occur the system not only has to return to $b$ (with negligible probability), but also to stay there as in $\boldsymbol{a}$ (which has a zero probability).

Thus, we see that the mixing in Figure 3.26a between two different ideal gases A and B always involves positive change in SMI (2*N* bits, or one bit per particle), and a positive change in entropy (2*N* multiplied by the conversion constant $k_B$ ln 2).

At this stage I would like to comment on two aspects of the mixing problem which caused a great deal of confusion in the literature.

First, mixing is not always associated with increase in entropy. Therefore, the general statement that mixing is disordering, and disordering involves positive change in entropy is not true.

Figure 3.28 (see color centerfold), shows a process of mixing two different ideal gases. Initially, we have *N*, A-particles in volume *V*, and *N*, B-particles in volume *V*. We mix the two in the same volume *V*. One can show that the entropy change in this process is 0. Thus, although we did *mix* the two gases, and this mixing which can be viewed as disordering of the system, the entropy change in this process is 0. The reason is that neither the locational, nor the velocity distribution of each particle changes in this process. Therefore, the change in the SMI is 0. It follows that the entropy change is also 0.

It can be shown that a spontaneous process of demixing, can occur involving a positive change in entropy.[23]

Thus, the conclusion that mixing is always an irreversible process involving positive change in entropy is, in general, not true.

The second, subtler point related to the mixing problem is the following: Whenever we have two different ideal gases A and B in Figure 3.26a (see color centerfold), we observe mixing when we remove the partition, and we can calculate the change in entropy ($2k_BN \ln 2$). This is true for any two *different* particles A and B. the American chemist Josiah Willard Gibbs, a key contributor to the development of statistical mechanics, was the first to analyze this process of mixing. He was puzzled by the fact that the change in entropy in process of Figure 3.26a, is the same no matter what A and B are, for as long as they are different.

In fact, this is puzzling. However, Gibbs did not realize that this process of mixing is equivalent to two processes of expansion. Once you realize this equivalence you understand that in this process the change in SMI is one bit per particle, and therefore $2N$ bits for all the particles. This is independent of the types A and B. It only depends on the "loss of information," we initially knew that each particle is confined to a volume $V$, and at the final state it occupies a larger volume $2V$.

What happens when A and B are the same, i.e. identical particles? Figure 3.26b. Here, upon removing the partition we shall not *see* anything happening, and indeed the entropy change is 0. However, from the molecular point of view each particle does expand from $V$ to $2V$. Hence, the SMI associates with the *locational* distribution changes by $2N$ bits. So why is the entropy change 0? This apparent puzzle has something to do with the fact that the particles are indistinguishable. The analysis of this case is subtler and even Gibbs himself reached a wrong conclusion. The analysis involves some mathematics, and may be found elsewhere. [24]

### 3.4.4 Spontaneous heat transfer from a hot to a cold body

As a final example of a spontaneous process involving a positive change in entropy, consider the process shown in Figure 3.29.

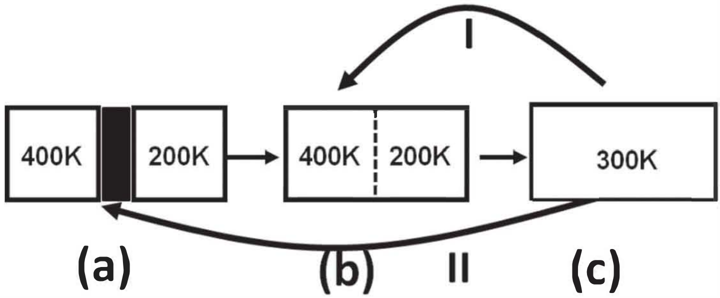


Figure 3.29. The initial, the final, and an intermediate state in the heat transfer process.

We start with two blocks of the same metal, each at different temperatures, say one at 400K, and the second at 200K. [25] The two bodies are separated by a partition which is an insulator, i.e. a wall that does not allow heat transfer between the two bodies. The entire combined system is isolated.

We now replace the insulator wall by a heat conducting wall. This is shown by a dashed line in Figure 3.29.

After some time, we shall see that the temperature of the hotter body will drop from 400K to 300K. The temperature of the cold body will rise from 200K to 300K. The process will ***always*** occur in this direction. We shall *never* observe the reversal (i.e. $c \rightarrow b$) of this process.

This process was the one in which Clausius first formulated the Second Law of Thermodynamics. It states that heat ***always*** flows from the hot to the cold body – *never* in the reverse direction.

This formulation of the Second Law is ***almost*** correct provided that first, we understand the word "always," and "never," in a probabilistic sense. "Always" means with very high probability, whereas "never" means "never" in many billions of years, but not in the absolute sense. Secondly, this formulation of the Second Law is valid only when the process occurs in an isolated system.

As we had done in the previous two examples, we can also ask the following three questions:

1. Can the process occur spontaneously in an isolated system in the reverse direction (i.e. $c \rightarrow b$)?
2. Can the SMI decrease from its value in state $c$ to its value in either state $a$, or $b$?
3. Can the entropy decrease from its value in state $c$ to its value in state $a$?

The answers to all these questions are the same as the ones we provided for the expansion, and mixing of ideal gases. The proof is not straightforward; it requires some mathematical arguments.[35]

We note that the system *can* go from the final equilibrium state $c$ to the initial state $b$ (but not to $a$). Regarding the change in the SMI in process $a$ to $c$, it is always positive. This statement is not trivial and requires proof. Basically, what happens here is that initially we have two different distributions of velocities (of all particles). The hotter body has a flatter distribution as we have seen in Section 3.3.5. The colder body has a sharper distribution of velocities. One can define the initial distribution of velocities in state *a* as an average distribution of the two bodies. After we remove the insulating wall the overall distribution of velocities will change. The new distribution of velocities corresponds to the final temperature at equilibrium. One can prove that the transition from the initial velocity-distribution to the final velocity-distribution always involves a *positive* change in SMI. Multiplying by the constant ($k_B \ln 2$) we can also calculate that the change in entropy in the process $a \rightarrow c$ will be positive. For more details, see Ben-Naim (2008, 2017b).

The interpretation of this positive entropy change is not simple. In both the expansion and the mixing processes discussed in Sections 3.3.1 and 3.3.3, the entropy change was due to the change in the *locational* distribution of the particles. In both processes each particle "expands" its accessible volume from $V$ to $2V$, and this involves a change of one bit per particle, hence, also positive change in entropy.

In the process of heat transfer, the situation is far more complicated. On one hand, the colder body is heated and its entropy *increases*. On the other hand, the hotter body cools down, and its entropy decreases. Though it is not trivial, one can prove that the *total* change in entropy must be positive.[27]

## 3.5 The Second Law of Thermodynamics

As we have noted in the preface, and as it is written in most textbooks, the two concepts of entropy and the Second Law are intertwined. You might find a statement of the Second Law as "the law of increasing entropy," or "the entropy of the universe always increases."

In this book I have defined and interpreted the concept of entropy without ever mentioning the Second Law. Likewise, one can introduce, discuss, and understand the Second Law without mentioning entropy. We shall see however that in some specific processes under specific conditions, the Second Law can be related to changes in entropy.

### 3.5.1 Thermodynamic and probabilistic formulation of the Second Law

There are many formulations of the Second Law. We have mentioned some in the previous sections. In this section, we shall formulate the Second Law in terms of thermodynamic quantities, as well as in terms of probability.

Consider the expansion process from $V$ to $2V$ described in Figure 3.18, in an isolated system. We can formulate the Second Law for this specific process in the following two, almost equivalent forms:

a. After the removal of the partition the system will “never” return spontaneously to the initial state.
b. The entropy of the system will never decrease and return to the initial value.

Note that in (a), I enclosed the word “never” in quotation marks, but not in (b). The reason is that in (a) we can observe the return to the initial state. As we saw, if *N* is small, then returning to the initial state is observable, however, for very large *N* we shall "never" observe such a return, here “never” is in practice, not absolute.

Regarding the entropy-formulation, we can say that the entropy will *never* decrease to its initial value. Here, we mean never in absolute sense. Since entropy is defined for equilibrium states only, the returning to its initial value means that the system must return to the *equilibrium* initial value, i.e., before the partition was removed. This will *never* occur spontaneously in an isolated system.

Thus, we see that the formulation (b) is *absolute*; no violations are allowed. On the other hand, formulation (a) is statistical, and "violations" are allowed. They are extremely rare for large *N*, and therefore any observation of returning to the initial state is not a violation of the Second Law but simply a rare event.

When people talk of violation of the Second Law in terms of (b), they actually mean SMI not entropy. As we have seen the SMI can change continuously after the removal of the partition. It also can fluctuate at equilibrium.

For 10 particles starting with all particles in the left compartment, both the number of particles in the left compartment, and in the SMI (associated with the distribution of particles in two compartments) fluctuates considerably as the system evolved after the removal of the partition between the two compartments. In fact, you can observe that the initial state (i.e. all particles in the left compartment) will be visited many times during the run of the simulation. When $N$ is 100, the change in the number of particles, as well as the SMI is smoother, but there are still noticeable fluctuations. When $N$ = 1000, we shall see an almost perfect monotonic change in SMI. Once the maximum value of the SMI is reached we do not observe any deviations.

When $N$ is of the order of $10^{23}$ we observe a *perfect* monotonic change in SMI and we will not observe any fluctuation in SMI once it reaches its maximum.

In the literature this behavior of the SMI is assigned to the entropy. In fact, even Boltzmann himself described this behavior of the entropy, i.e. that it increases monotonically towards its maximum, and once it reaches its maximum it can decrease, "it is not impossible but it is highly improbable." [28]

The apparent monotonic change of SMI with time (or the number of snapshots of the system) is the reason that such processes are referred to as irreversible, also as one directional change in entropy. This is unfortunately not exactly true. Every *process* is in principle reversible. However, when $N$ is very large, we *never* observe the reversal of this process, nor the reversal of the value of the SMI. This apparent irreversibility is a result of our extremely short lifetime compared with the time it will require for the system to return to its initial state, and the SMI to attain its initial value.

Note however that even when a system visits the initial system, the entropy does not decrease. The entropy by definition is related to the value of the SMI at *equilibrium*. In order for the entropy to decrease spontaneously in an isolated system, it must not only return to its initial state (see I in Figure 3.18) which is highly improbable, but it must return to the initial state and *stay there* which means it will return to the initial equilibrium state, Figure 3.18 process II. This will never occur; *never* in its absolute sense.[29]

In the following sections, we shall start with marbles in cells and explain the probabilistic nature of the Second Law. Next, we shall discuss systems of particles in cells or in compartments. Again, we shall see that understanding what happens, and why it happens in a preferred direction is based on probabilistic arguments.

Finally, we shall relate the probability interpretation of the Second Law to the entropy-formulation of the Second Law.

After reading this chapter you will see that the Second Law is not a mysterious almighty law, but rather as simple as the games we played in Section 1.1. Its understanding requires nothing more than plain common sense.

### 3.5.2 Let us play the 20Q game with marbles distributed in cells

In the following game we start with 10 marbles in 10 cells. The marbles are distinguishable, and they can have different colors, or different numbers. For any distribution of the marbles in the cells we can play the 20Q game as follows:

I show you the *distribution* of the *marbles*, such as the ones in Figure 3.30 (see color centerfold.

In general, the distribution of marbles is given by 10 numbers:

$$[N(1), N(2), N(3), \ldots, N(10)]$$

where $N(1)$ is the number of marbles in cell 1, $N(2)$ is the number of marbles in cell 2, and so on. In particular, the *distribution* of *marbles* in Figure 3.30 is (4, 0, 0, 1, 0, 2, 0, 0, 0, 3).

Now the rules of the game are as in the 20Q game. I think of a specific marble, and you have to find out in which cell the marble is by asking binary questions. You know the total number of cells and you also know the distribution of marbles in the cells. Your task is to find out where the marble is by asking minimum number of questions (if you like we can play the game by paying each other for each answer you get, and once you find out the marble I chose you get a prize).

How will you plan your strategy of asking questions? [30]

I hope you find this to be an easy game. Before we continue let me ask you a quick question: In Figure 3.31 (see color centerfold), I show you two different distributions of marbles.

In Figure 3.31a, the distribution of the marbles is:

$$[10, 0, 0, 0, 0, 0, 0, 0, 0, 0].$$

In Figure 3.31b, the distribution of the marbles is:

$$[1, 1, 1, 1, 1, 1, 1, 1, 1, 1].$$

Which of these two games is easier to play? By "easier," I mean that you can ask fewer questions in order to obtain the missing information.[31]

Now that you can distinguish between an easy and a difficult game, let me remind you that the SMI is simply a measure of the extent of difficulty of a specific game. It is related to the average number of questions you need to ask which guarantees that you will find the required information.

We next use the same system of 10 marbles in 10 cells, but now the game will evolve with time. We start with all marbles in cell 1, and we start shaking the whole box. If we shake vigorously we shall find out that some marbles will jump from one cell to the other. For instance, in Figure 3.32, we show a series of possible marble-distributions that we will observe after some shaking of the system.

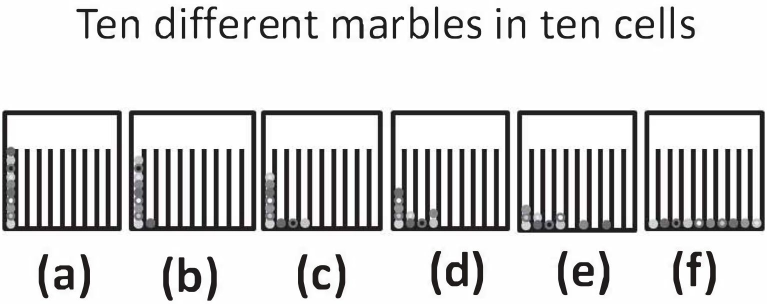


Figure 3.32. Starting with ten marbles in one cell (a), after some time we shall observe the evolution of the system, from (a), to (b), to (c), etc.

We can also *simulate* this "shaking" of the system on a computer. I start with all 10 marbles in cell 1, choose a specific marble and let is "jump" to another cell. While I do the "shaking" for a long time, I see that the distribution of marbles changes from Figure 3.32 (a), to 3.32 (b), to 3.32 (c), and so on.

For each *marble distribution* $[N(1), N(2), \ldots, N(10)]$. I can define a *probability distribution* $[p(1), p(2), \ldots, p(10)]$ where $p(i) = \frac{N(i)}{N}$. This is simply the fraction of the marbles in cell *i*.

For each marble distribution I can calculate the following three quantities, Figure 3.33:

1. The number of marbles left in cell 1.
2. The SMI of the 20Q game based on the marble distribution.
3. The probability of finding this particular distribution.

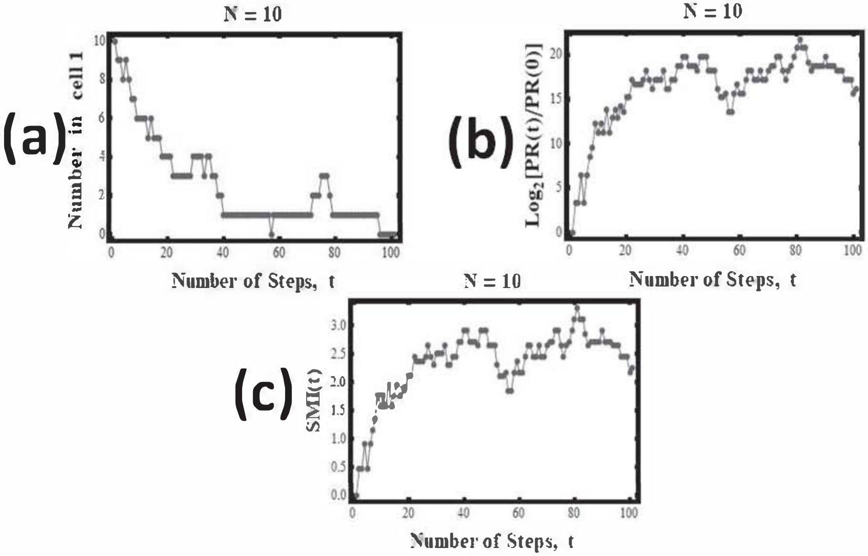


Figure 3.33. (a) Number of marbles in cell 1 after *t* steps.
(b) Logarithm of the probability ratio for the different configurations after *t* steps. (c) The change in the SMI as a function of the number of steps.

The first number is easy to calculate and easy to understand. We simply ***l●●k*** at how many marbles are left in cell 1. The second number is also easy to calculate. Given the marble distribution we can also calculate the probability distribution $[p(1), p(2), \dots, p(10)]$. On this probability distribution we can calculate the SMI if we know how to calculate the logarithm to base 2. [32]

If you do not know how to calculate logarithm you can do either of the following.

You can look at the graph of $\log_2$ (see Appendix A) for each of the probabilities $p(i)$, then calculate the average of the 10 quantities $-\log_2 p(i)$ with the probabilities $p(i)$.

The second way is just to estimate the average number of smart questions you will need to ask in order to get the required information. This will give you a qualitative idea of the value of the SMI. Since we started with SMI equal to zero, the SMI is the same as the change in the SMI for the process. The third way is to look at note 32, and tell yourself what those numbers mean.

The third quantity is a little more difficult to understand. If we shake the system for a long time, and take many snapshots of the marble-distributions we can calculate the number of times each of the possible distributions occurs. There is also a way to calculate this number theoretically.

Once we find out the number of times this distribution of marbles occurred we can also calculate the probability (Pr) of observing this distribution. Here, we have to be careful with the different probabilities I mentioned above. Read the next few sentences carefully and make sure you understand what is written therein.

For each *marble-distribution* we can construct a *probability distribution*. These two distributions are equivalent. On each marble (or probability) distribution we can also define the *probability* of occurrence of that marble (or probability) distribution.

For this reason, I refer to the second probability as the *super-probability* and denote it by Pr. This is the *probability* of obtaining a particular *probability distribution*.

In Figure 3.33 we plot the three quantities referred to above.

Details of the calculations are explained in Ben-Naim (2010), Chapter 3.[33] Here, I urge you to examine carefully these graphs to understand their meaning. The first as noted above is simple to understand. We started with all marbles in cell 1. Then we shook the system, and the number of marbles in cell 1 decreases almost monotonically to about 1. After a very long time you will see that the *average* number of marbles in each cell will be about 1. In this particular case we find that the number of marbles can also be 2 or 3, or more, but the most likely distribution is that shown in Figure 3.32f.

In Figure 3.33 (b), we also show the logarithm of ratio of super probabilities of the marble distributions. As we shake the system the configurations of the system changes, and the corresponding probability ratio also changes. We plot the logarithm of the ration because this ratio becomes very large number, and we want to draw the graph within reasonable range.

In Figure 3.33 (c), we show how the SMI changes as we shake the system. As you have already understood when we began this series of games, initially the value of the SMI was 0. This means that you know where the chosen marble is, therefore, you do not need to ask any questions in order to find out where the chosen marble is. As the system is being continuously shaken, the SMI increases to about 3. This means that after a long time the distribution of marbles will be almost uniform (i.e. about one marble in each cell). This is the most difficult game in this system. You will need on average, to ask three questions in order to find out in which cell the chosen marble is.

The most important conclusion from this experiment is that starting with any initial state, if we shake the system long enough the configuration of the system will tend to be uniform (about one marble per cell), and the SMI will increase to some maximum value. This means that the 20Q game will tend to be *more difficult* to play (i.e. more questions to ask on average).

I should also note here that the SMI increases almost monotonically to its maximum value. Once it reaches the maximum you will find some fluctuations in the SMI. When the number of marbles is very large the change in the SMI will be steadily upwards, and once it reaches the maximum it will stay there, and almost no fluctuations will be observed. [34]

### 3.5.3 Imagining playing the 20Q game with particles distributed in compartments

We now repeat almost the same experiment as in the previous section. This will lead us to formulate the Second Law for this particular experiment.

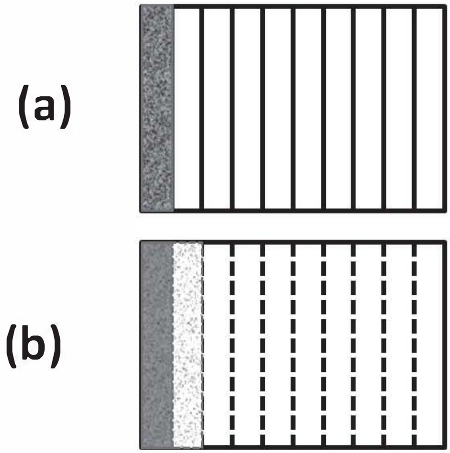


Figure 3.34. Ten moles of particles. (a) Initially, all in one compartment. (b) After a short period of time.

Look at Figure 3.34. Instead of marbles in cells we have *particles* in compartments. Instead of $N = 10$ marbles we have $N = 10$ moles of particles. (Each mole is an Avogadro number $6.023 \times 10^{23}$ of particles. This is a huge number, 10 followed by 23 zeros. For simplicity we shall refer to one mole as $10^{23}$ particles.

Remember, when we had 10 marbles in 10 cells we had a huge number of configurations. [35]

If you have any doubts about this number try to calculate the total number of configurations when you have two particles in two compartments, and three particles in two compartments. Repeat the same calculation for distinguishable and indistinguishable particles. [36]

As in the previous experiment, we also start with all $10 \times 10^{23}$ particles in the first compartment. We remove all the partitions between the compartments and watch what will happen.

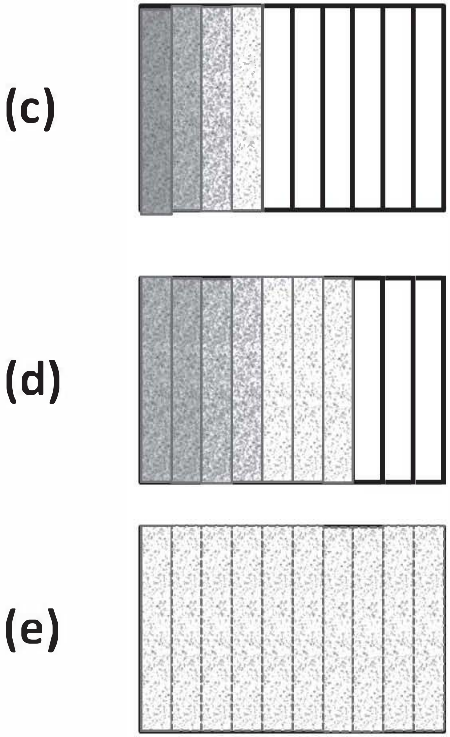


Figure 3.35. Ten moles of particles;
(c) and (d) after longer periods of time
(d) and (e) after very long time

In Figure 3.35, we show a few configurations obtained right after we removed the partitions. Note that we do not have to shake the system in order to achieve transitions of particles from one to another. The shaking is done from the "inside." The particles are moving at random velocities and directions.

These motions are enough to initiate the transitions of particles among the compartments.

Look again at the three quantities (1, 2, and 3) that we followed in the game of Section 3.5.2 and Figure 3.36 (note that *N=10* in this figure means ten moles of particles).

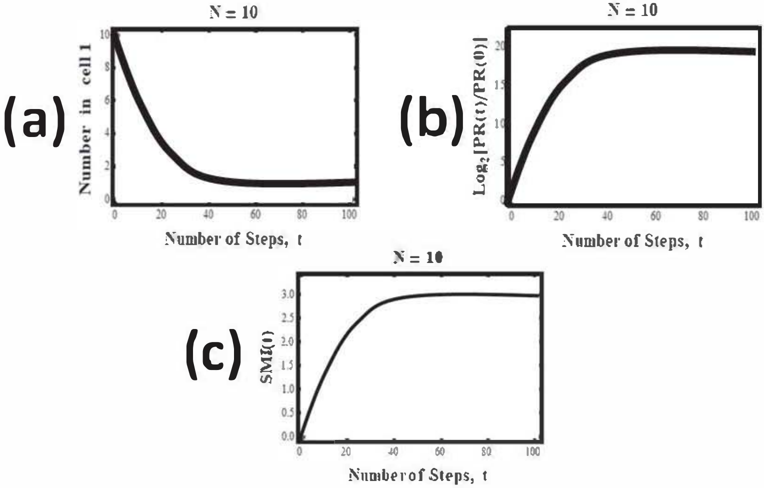


Figure 3.36. (a) Number of moles of particles in cell 1 after *t* steps. (b) Logarithm of the probability ratio for the different configuration after t steps. (c) The change in the SMI as a function of the number of steps (or time *t*). Note that *N*=10 means ten moles of particles.

Can you guess how these three quantities will change with time after we remove the partitions?

The answer is easy to guess. We shall observe the same trends as we recorded in Figure 3.33, but the curves will be smoother, so smooth that you get an impression that the process is strictly in one direction. Compare the graphs in Figure 3.33 with those in Figure 3.36.

First, it is clear that once we remove the partition the number of particles in cell 1 will be reduced. Unlike the curve in Figure 3.33, the change in the number of particles will be very smooth as shown schematically in Figure 3.36.

By smooth, we mean that we shall not observe any fluctuations as we observed in the experiment in Section 3.5.2. The number of particles in cell 1 will decrease from $10 \times 10^{23}$ to $10^{23}$ almost monotonically. In reality, we might see some very small fluctuations but these are so small that we cannot observe them. After some time (the length of time depends on the temperature, the higher the temperature the larger the average velocities of the particles, and the faster the process). We shall reach a point wherein there will be about $10^{23}$ particles in each compartment. Again, we note that fluctuations can occur. Small fluctuations (of a few hundreds or thousands of particles) are frequent, but they are so small that we can hardly notice them. On the other hand, fluctuations (say of $10^{20}$ particles, or returning to the initial state) are so rare that we shall "never" observe them. This is what the smooth curve means in Figure 3.36. This is also the meaning of the final equilibrium state. This means that once we reach this state we shall not be able to notice any deviation of the density of the particles in each of the compartment. This is also the meaning of the term *irreversible* introduced in thermodynamics. In principle the process is *reversible*. The system *can* return to its initial state. However, such a reversal is so rare that you will never observe it, not in your lifetime, not in the lifetime of the universe, and not in many billions of years. In this sense we can say that we shall *practically* never observe such a reversal, hence, the term *irreversible process*.

Next, we discuss the change in the SMI in this system. In Section 3.5.2, we saw that the system's SMI starts from zero and reached 3 after a lengthy shaking. We said that the game we played with marbles started as a very easy game, and became more difficult to play as we shook the system. The same will occur here but with two important differences.

First, the particles are indistinguishable. This affects the counting of the number of configurations. Second, the SMI depends on how we choose the probability distribution. If we are interested only in the locations of the particles in the different compartments, then for each particle we know that initially it was in cell 1, and at the end of the process it can be in any one of the 10 compartments with equal probability of $^1/_{10}$. Therefore, the change in SMI is from the initial value of 0 to the final value of $\log_2 10$ bits per particle, or $\mathrm{N} \log_2 10$ bits for the *N* particles. We see that the value of the SMI at the final state of equilibrium is the same as the change in SMI when we move from the initial to the final state.

$$\begin{aligned}&\text{Change of the SMI for the process}\\ &\qquad = \mathrm{SMI(final)} - \mathrm{SMI(initial)} = \mathrm{SMI(final)} - 0\\ &\qquad = \mathrm{SMI(final)} = \mathrm{N}\log_2 10 \text{ bits}\end{aligned} \tag{1}$$

Initially, we knew that each particle was in cell 1, hence, $\mathrm{SMI(initial)} = 0$. After the removal of all the partitions we did not know in which of the 10 cells the particle was. Therefore, we lost $\log_2 10$ bits per particles, and $\mathrm{N}\log_2 10$ bits for all the *N* particles. In Figure 3.36 (c), we draw the curve of SMI per mole (to compare with Figure 3.33 (c), per marble).

When we want to calculate the entropy change or the entropy difference the situation is different. The entropy in the initial state is not 0. Let us denote the entropy in the initial state by S(initial), and the entropy in the final state as S(final), then we obtain the change of entropy in the process:

$$\text{Change of entropy in the process} = \text{S(final)} - \text{S(initial)} \tag{2}$$

The entropy S (initial) may be computed from the SMI of the initial distribution of locations and velocities. This is not 0. The difference between the two SMI in equation (1) takes into account only the distribution with respect to the 10 compartments. Since we know that in the initial state all the particles are in compartment 1, we conclude that SMI(initial) = 0. On the other hand, the SMI based on the distribution of locations and velocities takes into account all possible molecular configurations (of locations and velocities). Thus, the initial entropy in equation (2) is related to the latter SMI by:

$$\text{S(initial)} = (k_B \ln 2)\text{SMI(initial, based on the molecular distribution)} \tag{3}$$

Once we know the S(initial), we can also calculate the change in entropy from equation (1) to (3) as:

$$\begin{aligned}&\text{change in entropy}\\ &= (k_B \ln 2) \times \text{ change in the SMI in the process}\\ &= k_B \ln 10 e.u. \qquad (4)\end{aligned}$$

Note that the change in the SMI in equation (1) is in bits. The change in the entropy is in *entropy* units (e.u.).[37]

Finally, we discuss the probability of the final state compared with the probability of the initial state. The general relationship for any process, initial→final, occurring in an isolated system is:

$$\frac{\Pr(final)}{\Pr(initial)} = \exp\left[\frac{S(final) - S(initial)}{k_B}\right] \qquad (5)$$

It should be stressed that the two states "initial" and "final" in this equation are not the same in the two sides of the equation.

On the right-hand side, $S(final)$ refers to the final *equilibrium* state. The quantity $S(initial)$ refers to the initial *equilibrium* state. Thus, the difference in entropies on the right-hand side of Equation (5), is between two *equilibrium* states.

On the other hand, the "final" in $\Pr(final)$ refers again to the final equilibrium state, but the "initial" in $\Pr(initial)$ *does* *not* refer to the initial state before removing the partition, but to the initial state just *after* the removal of the partition, this is state $b$ in Figure 3.18.

Note that the probability of state $a$, in Figure 3.18, is *one*. Also, the probability of state $c$ is *one*. These two are equilibrium states. However, state $b$ is not an equilibrium state. Here, we can ask about the probability of observing the state $b$ *after* we removed the partition, and after we arrived at the new equilibrium state $c$.

We can now conclude that the ratio of the probabilities on the left-hand side of the equation is extremely large (roughly of the order $10^N$). This is tantamount to saying that when we remove the partition the system will move from the "initial" state to the final state with probability nearly 1.

At the equilibrium state $c$, there is an extremely tiny probability to observe the state $b$. This event is so small that we can say that the process $b \rightarrow c$ is *irreversible*. This is a *practical* irreversibility, not an *absolute* one. The system will never spontaneously visit the initial state $a$. For this to occur we need to observe a system returning to state $a$, and *staying there*, as in the initial equilibrium state.

Before we continue we emphasize again that Equation (5) is valid for processes occurring in an isolated system.

### 3.5.4 The Entropy formulation of the Second Law for an isolated system

Traditionally, the Second Law for isolated systems is formulated in terms of entropy. You might find statements like:

1. “The entropy always increases and reaches a maximum at equilibrium.”
2. "The entropy of the universe always increases."
3. “The entropy of an isolated system increases and reaches maximum at equilibrium.”

These statements sound similar, but in fact they are quite different, and are all wrong.

The first is ambiguous. As we have seen, entropy is a quantity which is defined for a well-defined system. “Entropy increases” is meaningless as much as saying that “beauty increases,” “wisdom increases,” or “volume increases.” One must specify which entropy is being referred to.

The second is due to Clausius. Here, it seems that the system – the universe – is specified, but in fact it is not. The universe, is by definition all that exists. We have a long way to go as far knowing the size, the content, the total energy, etc. of the entire universe. Therefore, the universe is far from being a well-defined system. Besides, this formulation does not specify the parameter with respect to which the entropy increases. Most people believe that "entropy increases" means "increase with time." Unfortunately, the entropy in thermodynamics is defined for a well-defined system at equilibrium. As such, it is not a function of time.

Finally, the third formulation does specify the system, but does not specify the parameter with respect to which the entropy increases. In thermodynamics, entropy is said to be a *state function*. This means that the *macroscopic* state of the system is specified. An isolated system means a system having a constant energy ($E$), constant volume ($V$), and constant number of particles ($N$).[38]

For such a system, denoted $(E,V,N)$, the value of the entropy is fixed, and does not change with time. We denote this entropy by $S(E,V,N)$, i.e. the value of $S$ for a system characterized by the fixed value of $E$, $V$, and $N$.

We now state the thermodynamic formulation of the Second Law for an isolated system:

***For any isolated system $(E, V, N)$, at equilibrium, the entropy is maximum over all possible constrained equilibrium states of the same system.***

Note that this formulation uses only macroscopic quantities. Also, it applies only to equilibrium states. A constraint could be a partition which separates between one or more subsystems, or an inhibitor which prevents the occurrence of a chemical reaction.

The entropy formulation means that if we remove any of the constraints in any of the initial systems, the entropy will either increase or remain unchanged.

Therefore, an equivalent formulation of the Second Law is:

***Removing any constraint from a constrained equilibrium state of an isolated system will result in an increase of the entropy.***

Note carefully that we have *defined* entropy for equilibrium systems. The maximum entropy is also a maximum with respect to all *constrained equilibrium states*.

In many textbooks and popular science books you might come across a statement “that the entropy is the *cause* of the process (sometimes entropy is referred to as the "driving force" for the process). This is not true. The entropy change is a *result* of the process, and not its cause.

The ultimate cause of any process discussed above is the kinetic energy of the particles. If all the particles were at rest none of the processes described above would have occurred. The specific direction in which the process occurs (expansion not contraction, mixing not demixing, heat transfer from the hotter to the colder, not the other direction) is explained by the ratio of the probabilities. Therefore, the probability formulation of the Second Law as expressed by the ratio in Equation (5) is easier to understand. The fact that the system will move to a state of larger probability is simply a matter of common sense. The probability formulation of the Second Law is also more general; it applies to processes that do not necessarily occur in isolated systems.

In the next section we state very briefly two more formulations of the Second Law for processes at constant ($T$, $V$, $N$) and constant ($T$, $P$, $N$). Although it is not essential for the understanding of the entropy we discuss these two cases in order to demonstrate; first, that entropy formulation of the Second Law applies only for isolated systems, and second, that the probability formulation of the Second Law is much more general and applies to any process of removing constraints in a constrained equilibrium state.

### 3.5.5 Formulation of the Second Law for processes in ($T$, $V$, $N$) and ($T$, $P$, $N$) systems

In the previous section we formulated the Second Law for processes in an isolated system characterized by the constants $E$, $V$, $N$. In practice, a truly isolated system does not exist. All laboratory experiments are carried out on systems which are not isolated. The most frequent parameters used to describe a system are either ($T$, $V$, $N$) (i.e. constant temperature $T$, volume $V$, and number of particles N), or ($T$, $P$, $N$), i.e. at constant temperature $T$, pressure $P$, and number of particles $N$.

In this section we present, very briefly and without proofs, two other formulations of the Second Law. The reader can skip this section if he/she is interested only in entropy.

Before we formulate the Second Law for such a system we defined the Helmholtz energy (previously called *free energy*) $A$ by:

$$A = E - TS$$

Here, we have on the right-hand side of the equation the energy $E$, the temperature $T$, and the entropy $S$. Remember that

$S$ is *only* defined for equilibrium systems, and so is the Helmholtz energy.

This definition of the Helmholtz energy is valid for any system at equilibrium. In this definition we did not specify the independent variables with which we characterize the system. We have the liberty to choose any set of independent variables we choose, say $(E, V, N)$, $(T, V, N)$ or $(T, P, N)$. However, if we want to formulate the Second Law in terms of the Helmholtz energy we have no other choice but to choose the independent variables $(T, V, N)$. For a $(T, V, N)$ system the Helmholtz energy formulation of the Second Law is:

***For any $(T, V, N)$ system at equilibrium the Helmholtz energy has a minimum over all possible constrained equilibrium states of the same system.***

Recall that the entropy formulation of the Second Law was valid only for *isolated* systems, i.e. for $(E, V, N)$ systems. The Helmholtz energy formulation is valid only for systems that are isothermal (constant $T$), and isochoric (constant $V$), as well as closed (constant $N$).

An equivalent statement of the Helmholtz energy formulation is:

***Removing any constraint from a constrained equilibrium state in a $(T, V, N)$ system will result in a decrease in Helmholtz energy.***

Note carefully that it is only for a process at constant $(T, V, N)$ that this formulation is valid. It is not true that the Helmholtz energy decreases for any process occurring in any thermodynamic system.

The relationship between the change in the Helmholtz energy and the probability ratio is:

$$\frac{\Pr(final)}{\Pr(initial)} = \exp\left[-\frac{\mathrm{A}(final)-\mathrm{A}(initial)}{k_B T}\right] \qquad (6)$$

For a $(T, P, N)$ system the formulation of the Second Law is similar to the one for a $(T, V, N)$ system. Instead of the Helmholtz energy defined above, we need to define the Gibbs energy by:

$$G = E - TS + PV$$

We note again that the definition of the Gibbs energy applies to any thermodynamic system at equilibrium. We have the liberty to choose the independent variables characterizing the system. However, for the Gibbs energy formulation of the Second Law we must choose the specific independent variables

$(T, P, N)$. Here is the Gibbs energy formulation of the Second Law for the $T, P, N$ system:

***For any (T, P, N) system at equilibrium the Gibbs energy has a minimnm over all the possible constrained equilibrium states of the same system.***

It is important to emphasize that this formulation of the Second Law is valid for a system at constant temperature (isothermal), constant pressure (isobaric), and closed (i.e. impermeable to particles).

An equivalent statement of the Gibbs energy formulation is:

***Removing any constraint from a constrained equilibrium state of a (T, P, N) system will result in a decrease in the Gibbs energy.***

We emphasized again that it is not true that the Gibbs energy decreases in any process occurring in any system. The relationship between the change in the Gibbs energy and the probability ratio is:

$$\frac{\Pr(final)}{\Pr(initial)} = \exp\left[-\frac{G(final)-G(initial)}{k_B T}\right] \qquad (7)$$

### 3.5.6 The probability Formulation of the Second Law

As one can see, by comparing Equations (5), (6), and (7), the thermodynamic quantity that has an extremum (max $S$, min $A$, and min $G$) is different for different specifications of the system. The probability formulation (see the left-hand sides of Equations (5), (6), and (7) is the same. This formulation states that when we remove a constraint in a well-defined constrained equilibrium system, the system will move to a new equilibrium state with probability nearly one, or equivalently the ratio of the two probabilities is nearly "infinity." This is true for any well-defined thermodynamic system.

For this reason, the probability formulation of the Second Law is much more general than the entropy, Helmholtz energy, and Gibbs energy formulations. In practical applications the Gibbs energy formulation is the most useful, the second is the Helmholtz energy, and the last is the entropy formulation.

### 3.5.7 Conclusion

In this chapter, we described three different definitions of entropy. All three "converge" to the same value of entropy whenever entropy can be calculated either theoretically or experimentally.[38] My personal preference is in favor of the definition based on the SMI. The reasons for this preference are described below. Before doing this, I will briefly describe the definition of entropy based on the SMI, and the relationship between the SMI, the entropy and the Second Law.

We start by defining the SMI on the distribution of locations and velocities of a system of simple particles (i.e. having no internal degrees of freedom). We can define such a SMI for small or large systems, and for systems that are not at equilibrium.

The next step is to apply this SMI to a system of very large number of particles, then take the maximal value of this SMI, multiply by a constant and get the *entropy* of an ideal gas. More details in Note 39.

Thus, while entropy is a special case of SMI, the SMI is *not* entropy. Therefore, I suggest refraining from referring to SMI as either entropy, informational entropy, or Shannon's entropy.

In both the preface and the introduction to this chapter, I claimed that the definition based on the SMI is superior to the other definitions of entropy. Here is my explanation:

1. Since the entropy is a special case of SMI, it follows that whatever interpretation you choose for the SMI (average uncertainty, average unlikelihood, or a measure of

information), it will also apply for the entropy. This is the only valid, solid, and proven interpretation of entropy. You must also remember that entropy is not just “uncertainty,” as many authors write; it is the average uncertainty about the distribution of locations and velocities of all particles at equilibrium. This is the reason I have emphasized several times the restricted meaning of uncertainty when applied to either SMI or to entropy.

2. This definition leads to an exact relationship between the SMI and the super-probability (Pr), i.e. the probability of finding a specific distribution. The same relationship also applies for the entropy once we limit ourselves to equilibrium.[39]

3. This definition removes any mystery associated with entropy. One does not need to “invent” interpretations which are based on how the system of particles appears to us (order-disorder, chaos, spreading energy, etc.).

4. This definition also shows the limitations on the applicability of entropy and the Second Law. Hence, one should be careful not to apply entropy to systems for which it is inapplicable, such as living systems or the entire universe.

5. In the procedure of obtaining the entropy from the SMI, we saw how the maximum SMI leads to the *uniform* distribution of locations (in absence of an external field). Also, we saw why we get the Maxwell-Boltzmann distribution of velocities at equilibrium [for details, see Ben-Naim (2008, 2012)]. These two distributions are also the most probable distributions.

6. Finally, and most importantly, this definition shows clearly why entropy is *not* a function of time, and why it is incorrect to say that entropy tends to increase. It is the SMI of the system that can be a fimction of time, and can increase with time and reach a maximum value at equilibrium. Entropy (up to a multiplication constant) is the value of the SMI at equilibrium. As such, it is not a function of time, it does not increase with time, and it does not reach a maximum value at equilibrium. Entropy *is* the maximum value of the SMI of a thermodynamic system.

Having done with the advantages of the SMI-based definition of entropy, we tumed to the formulation of the *Second Law* based on *probability*, rather than on *entropy.* I am well aware of the fact that most scientists view the entropy as the core concept of the Second Law. Moreover, most scientists ascribe to entropy itself the power *to drive all* processes in the universe, not only processes in well-defined thermodynamic systems. Therefore, my suggestion is to replace entropy by probability in formulating the Second Law. This formulation is more general, on one hand, and makes the Second Law a matter of common sense, on the other hand.

# EPILOGUE

## MISINTERPRETATIONS AND OVER INTERPRETATIONS OF ENTROPY

Ever since the concept of entropy was introduced, people sought a simple and intuitive interpretation of its meaning. Many interpretations were suggested over the years such as: Arrow of time, disorder, mixing, chaos, spreading, ignorance, freedom, and many others. Unfortunately, none of these have been proven to be a correct interpretation of entropy.

If you open any dictionary you will find the following definitions of the word entropy.

In the Merriam-Webster Collegiate Dictionary (2003), "Entropy is defined as; "change," "literary turn," a measure of the "unavailable energy" in a closed thermodynamic system… a measure of the system's degree of order…"

In Yahoo's online dictionary, one finds:

1. The amount of thermal energy not available to work
2. A measure of the loss of information
3. A measure of disorder or randomness

In Merriam-Webster's online dictionary:

1. A measure of the unavailable energy in a closed thermodynamic system
2. A measure of the system's disorder
3. The degradation of matter and energy in the universe to an ultimate state of inert uniformity

$$\varepsilon\nu = in$$

$$\tau\rho o\pi\eta = trope = transformation$$

$$\varepsilon\nu\tau\rho o\pi\iota\alpha = entropy$$
$$= \text{transformation inwards}$$

In modern Greek, entropy means "turn into," or "turn to be," or "evolves into."

Unfortunately, all these "definitions" or "descriptions" of entropy are incorrect!

I believe that such misinterpretations contribute to the deepening mystery associated with entropy. The second contribution which enhances the mystery is the fact that many writers on entropy and the Second Law ascribe to entropy "super natural" powers, which is by far the most potent contributor to the mystery associated with it. It is not uncommon to find the expression “ravages of entropy,” ascribing to it the power of ravaging everything, from the spluttering of an egg, to decaying and death of living systems, to the eventual “thermal death” of the entire universe. Such exaggerated and overrated “powers” ascribed to entropy undoubtedly leaves the readers of popular science books with an awesome feeling. Not only does entropy possess physical powers, it also “controls,” and “drives” our thoughts, feelings, and creativity.

In my opinion, the reason why people talk about entropy of a living system, entropy of the universe, and the arrow of time, is simple: They simply reiterate what other people say, or have said about entropy. An erroneous statement by a famous and respected scientist will be propagated from generation to generation, until someone dares to challenge and question the validity of that statement.

The misunderstanding and misinterpretation of the entropy gives people a free hand in assigning to it all kinds of powers; “to ravaging everything,” "to driving the universe," the “be-all and end-all” of everything that happens in the universe, including our thinking, feelings, and creation of arts.

It is unfortunate that many readers of such popular science books believe that the author "knows" what he is writing about, but they do not have the tools to examine the truthfulness of any of the author’s claims.

Of course, if you believe that entropy is the be-all, and end-all of everything, how can you expect to ever understand entropy?

As a boy, I learned how God created the world, all its living creatures, and his best creation, humans. But God gave us the freedom of thought, feelings, and even the choice to believe, or not to believe in Him.

Now we learn that entropy controls everything. Perhaps, entropy also *decides* whether we can, or cannot understand entropy. Perhaps, the very understanding of the Second Law violates the Second Law.

All these nonsenses fill up many popular science books written by well-known scientists. One has to be very careful when reading any popular science book. If you do not have the tools to examine the veracity of the author's claims and statements, at least you can see whether the author provides any justification for such statements, or simply parrots what others have said or written. There are essentially three unjustified "applications" of entropy and the Second Law that have been propagated in the literature.

The most popular application of entropy is to living systems; you, and I, or any animal.

This application is based on the misconstrued (I would even say, perverted) interpretation of entropy as a measure of disorder, on one hand, and the view that life is a process towards more order, more structure, more organized, etc. on the other hand.

Combining these two erroneous views inevitably leads us to the association of life phenomena with a *decrease* in entropy. This in turn leads to the erroneous (perhaps meaningless) conclusion that life is a "struggle" against the Second Law. The fact is that entropy *cannot be defined for any living system*, and the Second Law, in its entropy formulation does not apply to living systems.

It is difficult to trace the origin of this misconstrued conclusion. I will be grateful for any information on this which the reader might be able to provide. There is no doubt however, that the most prominent and influential scientist who was responsible for much of the nonsensical writings in textbooks, as well as popular science books was Schrödinger himself.

Schrödinger, who contributed significantly to the development of quantum mechanics failed when he discussed entropy of living systems.

In his famous and widely praised book "What is Life," he expressed several times the erroneous idea that the Second Law is the "natural tendency" of things to go from *order* to *disorder*," and in addition: "Life seems to be orderly and lawful behavior of matter, not based exclusively on its tendency to go over from order to disorder."

From these two assertions, Schrödinger reached the most absurd conclusion to the question: "What then is the precious something contained in our food which keeps us from death?"

His answer: "What an organism feeds upon is *negative entropy* … the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive."

Thus, Schrödinger not only adopted the misinterpretation of entropy as a measure of disorder, and not only expressed the misconception regarding the role of entropy in living systems, but also "invented" a new concept of "negative entropy" to explain how a living system "keeps aloof of death."

We emphasize again that entropy is not definable for a living system. Any statement about the entropy change in a living system is therefore, meaningless. This is *a fortiori* true when we use the meaningless "negative entropy" in connection with living systems. I hope that the reader of my book knows by now that entropy is a positive number. It is defined for well-defined systems at equilibrium. Next time you read a statement about the entropy of a living system, ask the author to explain how he/she calculated, or measured that entropy.

The Second Application of entropy and the Second Law is to the *entire* universe.

In this case, I can easily pinpoint the culprit for the misuse – no other than Clausius himself. As is well known, Clausius formulated one version of the Second Law (heat flows from a hot body to a cold body). Clausius also defined the change in entropy for the transfer of a small quantity of heat into, or out from a system at a constant temperature. Clausius' ideas constituted the basis on which the whole science of thermodynamics was built upon, including the most general, and most powerful, and useful Second Law of Thermodynamics. For all these achievements Clausius deserved the highest scientific credit. Unfortunately, Clausius failed in *over generalizing* the Second Law. His well-known and well quoted statement:

***"The entropy of the universe always increases."***

I do not know how Clausius arrived at this formulation of the Second Law. I can only guess what has motivated him, as well as the many others who followed him to conclude that the entropy of the universe always increases. And "always increases," always means increases with time. This brings me to the final, and most common misconception about entropy and the Second Law.

The third miss-application is the Association of Entropy with the Arrow of Time.

The association of entropy with the Arrow of Time has been discussed in great detail in my book "The Briefest History of Time," Ben-Naim (2016a). Here, I will make only a few comments. The association of entropy with the so-called Arrow of Time is probably due to Eddington.

There are two very well-known quotations from Eddington's (1928) book, "The Nature of the Physical World." The first concerns the role of entropy and the Second Law, and the second introduces the idea of "time's arrow."

1. *"The law that entropy always increases, holds, I think, the supreme position among the laws of Nature.*

2. *" Let us draw an arrow arbitrarily. If as we follow the arrow we find more and more of the random element in the state of the world, then the arrow is pointing towards the future; if the random element decreases the arrow points towards the past. That is the only distinction known to physics. This follows at once if our fundamental contention is admitted that the introduction of randomness is the only thing which cannot be undone. I shall use the phrase 'time's arrow' to express this one-way property of time which has no analogue in space."*

In the first quotation Eddington reiterates the unfounded idea that "entropy always increases." Although I agree that the Second Law of thermodynamics is unique compared with other laws of physics [see also Ben-Naim (2008, 2015a)], I do not agree with the statement that "entropy always increases."

Although it is not explicitly stated, the second quotation alludes to the connection between the Second Law and the Arrow of Time. This is clear from the association of the "random element in the state of the world" with the "arrow pointing towards the future."

In my view it is far from clear that an Arrow of Time exists. It is also clear that entropy is not associated with randomness, and it is meaningless to say that entropy always increases. Therefore, my conclusion is that entropy has nothing to do with time! Unfortunately, many scientists even to this day still maintain that "entropy changes with time," "entropy explains time," and that entropy *is* the Arrow of time."[4●]

I hope that by following the steps in defining entropy you can judge for yourself why entropy cannot be used for any living system to the entire universe, and why entropy is not a function of time.

# NOTES

## Chapter 1

Note 1: Suppose we play with one die having six faces, numbered 1 to 6. We play the following game. I choose a number, say "1," and you choose any other number, say "4." I throw the die many times. Whenever it falls with its face "1" upward, you give me $1.00, and when it falls with "4" upwards, I will give you $ 1.00. If there is no reason to believe that any face has a preference to show up, then the game is *fair*, and the die is called a fair die. It is a *fair game* because both outcomes have the same chances of winning, or losing.

Now, suppose I know that the face with six dots is heavier than all the other faces, Figure 1.2b. Therefore, in each throw the face opposing the six dots, i.e. with one dot will show up more frequently. In this case, if I choose "1" and I let you choose any other number between "2" and "6," then this game is *unfair*. In this case the die will be described as an "unfair" die. Can you explain why this game is unfair? Remember, that I know that the face "6" is heavier, but you do not know this fact.

Note 2: In case *a*, the *probability* of the occurrence of the *blue* outcome is 1, or 100%. This means that in each toss, you are *certain* that the blue color will appear, therefore, you will win in each toss with *certainty*, or with probability one!

Note 3: Since there are five blue faces, and one red face, the relative chances of the occurrence of the blue and red outcomes are 5 to 1. In terms of probability, we can say that the *probability* of "blue" is five times larger than the *probability* of "red." This qualitative answer is enough for the moment. Later, you will say that the probability of the *event* "blue" is 5/6, and the probability of the *event* "red" is 1/6. Therefore, it is advantageous for you to choose *blue*.

Note 4: The average net earnings are calculated as follows:

When "blue" appears you get (+1) $ with probability 5/6. When "red" appears you pay one $ (-1) with probability 1/6. Therefore, the average expected earning on one toss is: $\frac{5}{6}(+1)+\frac{1}{6}(-1)$. The average expected earnings on 100 tosses is:$100\times\frac{4}{6}=\frac{400}{6}\approx 66.66\ldots$

This means that "on average" you are expected to earn about $ 67.00 after 100 tosses. We shall learn more about averages in Section 1.9.

Footnote: To Table 1.2.

In the fifth column in Table 1.2, we record the "weighted average." In Section 1.9 we will discuss the concept of an average. Here, if you are given two numbers, say 15 and 17, and you are asked to calculate the *average* of these two numbers, you are most likely to answer that the average is $(15 + 17)/2 = 32/2 = 16$. In this calculation, you implicitly assumed that the "weight" or the probability of each number is $\frac{1}{2}$. However, in general, if the number 15, which may be an outcome of an experiment occurs with probability (or with relative frequency) of say, 0.25, and the number 17 occurs with probability of 0.75, then the weighted average of the two numbers is calculated as:

$$0.25 \times 15 + 0.75 \times 17 = 3.75 + 12.75 = 16.5$$

This is slightly larger than your previous result of 16. The reason for this larger average is that we are giving larger "weight" to the number 17 than to the number 15.

Now, if I ask you to calculate the average of the two numbers $\frac{5}{6}$ and $\frac{1}{6}$, the "arithmetic" average would be:

$$\left(\frac{5}{6} + \frac{5}{16}\right)/2 = \frac{1}{2}$$

This average is not what we calculated in Table 1.2. Here, we calculate the "weighted" average of $\frac{5}{6}$ and $\frac{1}{6}$, using the weights $\frac{5}{6}$ and $\frac{1}{6}$, respectively. Thus, the average given in the fifth column is this weighted average of the two numbers $\Pr(blue)$ and $\Pr(red)$. As we shall see in Chapter 2, this special kind of average, is similar to the Shannon's measure of information.

Note 5: The probability of the "blue event" is $\frac{5}{6}$. Assuming that all the tosses are independent (see Section 1.6), we calculate the probability of earning $100.00, by multiplying $\frac{5}{6}$ by itself hundred times:

Probability of 100 blues

$$= \frac{5}{6} \times \frac{5}{6} \times \text{... hundred times ...} \times \frac{5}{6} = \left(\frac{5}{6}\right)^{100} \approx 0.000000012$$

This is a very small number. However, the probability of the outcome "red event" in 100 tosses is obtained by multiplying $\frac{1}{6}$ by itself hundred times:

Probability of 100 blues

$$= \frac{1}{6} \times \frac{1}{6} \times \text{... hundred times ...} \times \frac{1}{6} = \left(\frac{1}{6}\right)^{100} \approx 1.5 \times 10^{-87}$$

This is a far smaller number. The ratio between the two numbers is:

$$\frac{\text{Probability of 100 blues}}{\text{Probability of 100 reds}} = \left(\frac{5}{1}\right)^{100} \approx 8 \times 10^{69}$$

This is a very large number, about 1 followed by seventy zeroes!

Note 6: Qualitatively, it is clear that the probability of the occurrence of blue is larger than the probability of red. The relative chances are 4 to 2 (or 2 to 1). In terms of probabilities:

$$\text{The probability of blue} = \frac{4}{6} = \frac{2}{3}$$

$$\text{The probability of red} = \frac{2}{6} = \frac{1}{3}$$

Note 7: The average (or the expected) earning in 100 tosses in case *c*, is:

$$100\left[\frac{4}{6}(+1) + \frac{2}{6}(-1)\right] = 100\left(\frac{4}{6} - \frac{2}{6}\right) = \frac{100}{3} = 33.33$$

This means that on average, if you play 100 times with die *c*, your net earnings will be about $33.33. Not bad, yet not as good as in case *b*.

Note 8: The calculations of the probabilities is as follows: The probability of getting a blue in one toss is $\frac{4}{6} = \frac{2}{3}$, therefore, assuming that the tosses are independent (see Section 1.6), the probability of getting a "blue" in all of the 100 tosses is obtained by multiplying $\frac{2}{3}$ by itself hundred times:

$$\frac{2}{3} \times \frac{2}{3} \times \text{... hundred times ...} \times \frac{2}{3} = \left(\frac{2}{3}\right)^{100} \approx 2.4 \times 10^{-18}$$

This is a very small number, about 0.000··· 18 zeros··· 24

The probability of getting all "red" outcomes in all 100 tosses is:

$$\frac{1}{3} \times \frac{1}{3} \times \cdots \text{hundred times ...} \times \frac{1}{3} = \left(\frac{1}{3}\right)^{100} \approx 2 \times 10^{-48}$$

which is a far smaller number than the probability of getting all "blues."

The ratio of these two probabilities is:

$$\frac{\text{Probability of 100 blues}}{\text{Probability of 100 reds}} = 2^{100} \approx 10^{30}$$

This is one followed by 30 zeros. This is a huge number, yet it is far smaller than the number we obtained for this ratio in the case of die *b*, see note 4.

Note 9: There are three red, and three blue faces. Therefore, the chances of either the red or blue outcomes are equal. The probability of the outcome red is $\frac{3}{6}=\frac{1}{2}$, and the probability of the outcome blue is also $\frac{3}{6}=\frac{1}{2}$.

Note 10: The probability of occurrence of 100 blue outcomes in case ***d*** is:

$$\frac{1}{2}\times\frac{1}{2}\times\cdots\times\frac{1}{2}=\left(\frac{1}{2}\right)^{100}\approx 8\times 10^{-31}$$

This is the also probability for the occurrence of 100 red outcomes. The ratio of these two probabilities is:

$$\frac{\text{Probability of 100 blues}}{\text{Probability of 100 reds}}=\left(\frac{1}{1}\right)^{100}=1$$

Note 11: In table 1.11 we list all possible specific sequences of blues (B), and reds (R), and their probabilities. We group the sequence according to the number of Bs in each sequence, and calculate the corresponding probabilities of the generic sequence. Note that the probability of each specific sequence is the same 1/16.

Note 12: To calculate the probability of a specific sequence, say, RBRB, we assume that the tosses are independent (see Section 1.6). Therefore, the probability of this sequence is:

$$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$$

However, in order to calculate the probability of the generic sequence; i.e. a sequence of two Rs, and two Bs, we have to *sum* over all possible specific sequences of two Rs, and two Bs; these are:

BBRR

BRBR

RBBR

RBRB

RRBB

BRRB

Each of these have the same number of Rs, and Bs. Since these are disjoint events (see Section 1.3), the probability of the generic sequence is the sum over all the probabilities of these specific sequences, i.e.

$$6 \times \left(\frac{1}{2}\right)^{4} = \frac{6}{16} = \frac{3}{8}$$

Note 13: Note that in general, the average for two numbers $A_1$and $A_2$ is defined by:

$$Average\ of\ A = \Pr(1)A_1 + \Pr(2)A_2$$

See also Note 4.

We multiply each number by the probability of occurrence of that number, and then sum over all possible "events." Here, we have only two events "blue" and "red," and the average is over the probabilities themselves.

Average of probabilities

$$= \Pr(\text{blue})\Pr(\text{blue}) + \Pr(\text{red})\Pr(\text{red})$$

This can be referred to as the average probabilities of the die, or the average uncertainties, or the average likelihood over all possible outcomes.

Note that the fifth column in Table 1.2, we record the "weighted average." In Section 1.9 we will discuss the concept of an average. Here, if you are given two numbers, say 15 and

17, and you are asked to calculate the *average* of these two numbers, you are most likely to answer that the average is $(15 + 17)/2 = 32/2 = 16$. In this calculation, you implicitly assumed that the "weight" or the probability of each number is $\frac{1}{2}$. However, in general, if the number 15, which may be an outcome of an experiment, occurs with probability (or with relative frequency) of say, 0.25, and the number 17 occurs with probability of 0.75, then the weighted average of the two numbers is calculated as:

$$0.25 \times 15 + 0.75 \times 17 = 3.75 + 12.75 = 16.5$$

This is slightly larger than your previous result of 16. The reason for this larger average is that we are giving larger "weight" to the number 17 than to the number 15.

Now, if I ask you to calculate the average of the two numbers $\frac{5}{6}$ and $\frac{1}{6}$, the "arithmetic" average would be:

$$\left(\frac{5}{6} + \frac{5}{16}\right)/2 = \frac{1}{2}$$

This average is not what we calculated in Table 1.2. Here, we calculate the "weighted" average of $\frac{5}{6}$ and $\frac{1}{6}$, using the weights $\frac{5}{6}$ and $\frac{1}{6}$, respectively. Thus, the average given in the fifth column is this weighted average of the two numbers $\Pr(blue)$ and $\Pr(red)$. As we shall see in Chapter 2, this special kind of average is similar to the Shannon's measure of information.

Note 14: The eleven events and the corresponding probabilities are:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Event: | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Probability: | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

Note 15: The answers are:

For (a): $\frac{1}{6}$, For (b): $\frac{1}{2}$, For (c): $\frac{1}{3}$

Note 16: The probability is 1/3.

Note 17: The general "definition" is as follows: We make N experiments. We record the number of times the outcome *"B"* occurred. We define the probability of "B" as the limit of the ratio $N_B/N$ when N is very large (infinity).

Note 18: Note that the conditional probability is defined only for a condition, the probability of which is not zero. In the abovementioned example, we require that the event $B$ is not an impossible event.

Note 19: Younger children chose the urn on the right which is the *correct* choice, but for the *wrong* reasons. Furthermore, young children who chose the urn on the right and did not win switched to the urn on the left on the next game. When asked why they changed the urn they simply said: "The first choice was not good, it did not deliver the expected prize." Older children, aged nine years and above chose the urn on the right on the first game. They continued to choose the same urn on the second and third games even if they did not win on some of the games. Somehow they sensed that even if the urn on the right did not win, this urn was still the better choice. *Better*, does not guarantee that one wins every game. It means that on average you will win with higher probability. Therefore, if you choose the urn on the right-hand side, and do not win, you should not be discouraged. Be patient, continue to play and stick to the same urn. In the long run you will be winning.

Now for the probability of winning or losing:

For the urn on the left:

$$\Pr(win) = \frac{1}{2} \quad , \quad \Pr(lose) = \frac{1}{2}$$

For the urn on the right:

$$\Pr(win) = \frac{8}{12} = \frac{2}{3} \quad , \quad \Pr(lose) = \frac{4}{12} = \frac{1}{3}$$

---

Here, Pr is a shorthand for “Probability of.” Thus, $\Pr(win)$ means the Probability of winning.

You can see that the urn on the right has higher probability or likelihood of winning, and therefore it is advantageous to choose this urn in this “easy” game. It is advantageous not because this urn has *more* blue marbles, but because the *r●tio* of the number of blue marbles to red marbles is larger in the urn on the right than in the urn on the left.

Note 20: In game A, we assumed that the die is "fair," which means that all outcomes are equally likely to occur. There is no preferred outcome. Therefore, you cannot "explain" why you have chosen any particular number between 1 and 6. In this game, whatever outcome you choose, say 4 or 6, you can expect to "win," on average, one in six throws. If you play 1000 games, you can expect to "earn" $1000 \times \frac{1}{6} \approx 167$ dollars, independently of the number you have chosen.

In game B, again your earnings are independent of the choice of a particular number. On average, you will earn

$1000 \times \frac{1}{6} \approx 167$ dollars if the outcome coincides with the number you chose, but you will lose $1000 \times \frac{5}{6} \times \frac{21}{100} =$ 175 dollars. In this case, you will be losing, on average, about $175 - 167 = 8$ dollars after 1000 games.

Note 21: The probabilities are expected earnings are:

For "blue":

$\Pr = \frac{4}{19}$, with expected earnings of $\frac{4}{19} \times 2 = \frac{8}{19}$

For "red":

$\Pr = \frac{5}{19}$, with expected earnings of $\frac{5}{19} \times 3 = \frac{15}{19}$

For "green":

$\Pr = \frac{10}{19}$, with expected earnings of $\frac{10}{19} \times 1 = \frac{10}{19}$

Thus, although the probability of "red" has decreased (since you removed one "red" from the urn), your expected earnings are still larger when you choose red again. (These are the expected earnings per one draw. If you play the same game 1000 times you have to multiply these expected values of 1000).

Note 22: The probabilities and expected earnings are:

For "blue":

$\Pr = \frac{4}{18}$, with expected earnings of $\frac{4}{18} \times 2 = \frac{8}{18}$

For "red":

$\Pr = \frac{4}{18}$, with expected earnings of $\frac{4}{18} \times 3 = \frac{12}{18}$

For "green":

$\Pr = \frac{10}{18}$, with expected earnings of $\frac{10}{18} \times 1 = \frac{10}{18}$

Note 23: The last sentence might sound awkward. We are talking about the *probability* of finding a specific *probability distribution*. We shall further discuss this idea in Chapters 2 and 3.

Note 24: The generalization for a sequence of $N$ throws with $n$ Hs, and $(N - n)$ Ts is straightforward:

$$\Pr(\text{any sequence of } n \text{ Hs and } (N-n) \text{ Ts}) = \binom{N}{n} p^n q^{N-n}$$

where the symbol $\binom{N}{n}$ is $\frac{N!}{n!(N-n)!}$ and $N!$ means the product of all numbers from one to $N$, i.e. $1 \times 2 \times \cdots \times N$.

Note 25: Even without having a definition of an *average* quantity, it is clear that the statement made by ASU in 2000 is meaningless. The grades of *all* the students cannot be *above* the average. What the writer of this statement probably wanted to say is that the average in 2000 was above the averages of all students in the country. The *average* grade of the students in a given university is a number between the lowest and the highest grade.

Note 26: Yes, it is possible. It is easy to construct an example. Suppose HSU has seven professors with the following IQs: 100, 110, 120, 130, 140, 150 and 160; the average is 130. LSU has also seven professors with IQs: 50, 60, 70, 80, 90, 100 and 110; the average is 80.

Now, if the professor of the lowest IQ from HSU moves to LSU, the new situation is:

In HSU, there are six professors with IQs 110, 120, 130, 140, 150 and 160.

In LSU, there are eight professors with IQs 50, 60, 70, 80, 90, 100, 100 and 110.

Check the average IQ in HSU *increased* from 130 to 135, and also the average IQ in LSU *increased* from 80 to 82.5.

Of course, you cannot increase the average IQ of *all* of the professors in two universities. It was 105 before the move and 105 after the move. Explain why.

Note 27: Note that the average of the *two speeds* is $\frac{140}{20} =$ 70km/h. However, in order to calculate the average speed on the round trip, you take the total length and divide it by the total time of travel. In our examples, let the distance between Jerusalem to Tel Aviv be *a*. The travel time to Tel Aviv is $t_1 = \frac{a}{v_1} = \frac{a}{40}$, and the travel time from Tel Aviv to Jerusalem is $t_2 = \frac{a}{v_1} = \frac{a}{100}$. Therefore, the *average speed* in the round trip is $\frac{2a}{t_1+t_2} = \frac{2a}{\frac{a}{40}} = \frac{1800}{140} \approx 57$km/h.

In general, if the first speed is $v_1$ and the second is $v_2$, then the average of the two speeds is always *larger* than the average speed on the round trip, i.e.

$$\frac{v_1 + v_2}{2} = \frac{2v_1v_2}{v_1 + v_2} = \frac{(v_1 - v_2)^2}{2(v_1 - v_2)} \geq 0$$

Note 28: The average of the two speeds is approximately *half* the *speed of light*. The average speed of the entire trip is *half* the *donkey's speed*.

29. Rakoczy *et al* (2014)

## Notes to Chapter 2

Note 1: In die $a$, I am certain (or least uncertain) that I will earn more money than if I choose die $b$.

Note 2: Note that $0\log_2 0$ is equal to zero. One can take this as a definition of $x \log x$, when $x = 0$, or prove that in the limit of $x \to 0$, the function $x \log x \to 0$. The logarithm can be to any base.

It should be noted that the general behavior of the function SMI $(p)$ is similar (but not identical) to the one constructed in Figure 1.7. The reason is that the quantity SMI (p) can also be interpreted as average uncertainty with respect to the probability distribution. If you look at Figure A.1 (Appendix A), you will notice that for each value of $0 < p < 1$, $\log_2 p$ is a negative number. Therefore, $-\log_2 p$ is a positive number when $p$ changes between 0 and 1, $-\log_2 p$ changes between $\infty$ to zero, Figure A.3. Thus, the larger $p$ is, the larger the certainty, or the smaller the uncertainty of the occurrence of the event H. Also, $-\log_2 \mathrm{p}$ is larger, the larger the uncertainty. Therefore, SMI (p) may also be interpreted as a measure of the average uncertainty regarding the game with two possible outcomes.

Note 3: See simulated games at my site http://www.ariehbennaim.com.

Note 4: For more details, see Ben-Naim (2017b).

Note 5: It is about half a million questions! For details, see Ben-Naim (2017b).

Note 6: This follows from the property of the logarithm function. If we double $N$, we get $\log_2(2 \times \mathrm{N}) = \log_2 N + \log_2 2 = \log_2(N) + 1$.

Note 7: The number of questions is 5. You first divide the total objects or persons into two groups of 16. Ask whether the person or the object is on the right or the left. Then divide into two groups of 8, then into two groups of 4, then into two groups of 2, then you will have a choice between two possibilities. Therefore, in five questions you will find the required information.

Note 8: In this case you will need one more question, i.e. 6 questions. Explain why?

Note 9: For details, see Ben-Naim (2017b).

Note 10: We should add here that when $p_i = 0$, we are *certain* that the event *i will not occur*. It would be awkward to say in this case that the *uncertainty* in the occurrence of $i$ is zero. Fortunately, this awkwardness does not affect the value of SMI. Once we form the product $p_i \log p_i$, we get zero when either $p_i = 1$, or when $p_i = 0$.

Note 11: For specific examples, see Ben-Naim (2015, 2018).

## Notes to Chapter 3

Note 1: This is the same process of mixing that we observe when we add milk to coffee. However, in case of liquids or solids (that is, system with strong intramolecular reactions) the mixing is not always the spontaneous process that occurs.

Note 2: Note again that this is true for ideal gases and many liquids. But in some liquids or solids we might observe separation into two almost pure substances.

Note 3: This is true when the combined system of the two bodies is isolated.

Note 4: For more details, see Ben-Naim (2011a, 2015a, 2016a).

Note 5: Sometimes you see a slightly different notation, i.e. $dQ_{rev}/T$, where "rev" is short for reversible. We will not need this notion here. We require that the system is large enough at a specific temperature $T$, and that $dQ$ is small enough so that it does not change the temperature of the system. Some authors use the notation $\delta Q$ to emphasize that this quantity is not an exact differential. On the other hand, $dS$ is an exact differential. This means that there exists a function $S$, which is a state-function, i.e. a function of the parameters describing the system, say $T$, $P$, $N$, and is differentiable with respect to these variables.

Note 6: For more details, see Ben-Naim (2008, 2015a, 2016a).

Note 7: Higher temperature means larger kinetic energy of the particles. The relationship between the absolute temperature $T$ and the average kinetic energy $E_k$ of the particles is given by: $k_B T = \frac{2}{3}\left[\frac{m\langle v^2\rangle}{2}\right] = \frac{2}{3}E_k$.

Note 8: In Section 3.5 we shall see that the Second Law is indeed a law of probability. The system can go back to its initial state. However, the entropy can never decrease spontaneously in an isolated system.

Note 9: Basically, the paradoxes do not stem from Boltzmann's *definition* of entropy as stated above, but from another function Boltzmann had defined, and showed that it always decreases with time. This function was known as the H-function and the theorem Boltzmann proves was referred to as Boltzmann's H-theorem.

In fact, the H-function is a particular case of Shannon's measure of information. Indeed, it is true that H decreases with time until it reaches a minimum at equilibrium. The function $-H(t)$ looks like entropy but it is not the entropy of the system.

Note 10: See Ben-Naim (2017, 2018).

Note 11: Note that $f(x)dx$ is the probability of finding the center of a particle between $x$ and $x + dx$. $f^*(x)dx$ has the same meaning but at equilibrium.

Note 12: In one dimension the SMI is $\log_2 N$ , and in 3D it is: $\log_2(N \times N \times N) = 3\log_2 N$

Note 13: For details see Ben-Naim (2008, 2012).

The 1D distribution of velocities is the normal distribution. It is also referred to as the Maxwell-Boltzmann distribution. This is an *exact* normal distribution. Sometimes, this 1D distribution is confused with the distribution of the speeds of particles, i.e. the absolute value of the velocity in 3D which is defined by $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$.

Note 14: A quick clarification. A distribution is a vector of numbers $p_1, \dots, p_N$ which are probabilities, i.e. $p_i$ is the probability of the event *i*. The *"super-probability"* mentioned in the text is $\Pr(p_1, \dots, p_n)$, i.e. the probability of obtaining the distribution $(p_1, \dots, p_N)$.

Note 15: In quantum mechanics, the microscopic states are defined as the solutions of the so-called stationary Schrödinger equation. The classical microscopic state is easier to visualize, i.e. specifying all the locations and velocities of the particles. The quantum mechanical microstates are, in general, not easy to visualize.

Note 16: A comment regarding the size of the system is in order. Remember that the SMI is defined for any distribution. Therefore, we can also define the SMI for a system of one, two, or three particles in a box. We can also proceed to define the corresponding entropy of such a system. However, if *N* is a small number, we shall find that the system will not obey the Second Law of Thermodynamics. In Section 3.5 we discuss the Second Law and relate it to the entropy. We shall see that if we expect the system to approach an equilibrium state, and stay there "forever," then we need to take the SMI of very large number of particles. Indeed, macroscopic systems contain a huge number of particles. This is the reason why we *never* observe large fluctuations from the equilibrium state.

Note 17: "Classical" means that we describe the microstates of the system by the locations and velocities (or momentum) of all its particles. "Ideal gas" means that the particles do not interact with each other. "Simple particles" means that the particles do not have any "internal degrees of freedom." This means that the microstate of each particle is fully described by its location and velocity. A non-simple atom or molecule has "internal structure." In addition to its location and velocity, one should also specify its *internal* state, i.e. the particle might be in different electronic, nuclear, vibrational or rotational states. All these are necessary for the description of the microstate of the system, and hence also contribute to the entropy.

Note 18: The reduction in the total number of microstates is obtained by first counting all the configurations when the particles are labeled (say, by numbers: $1, 2, \dots N$) then dividing by $N! = 1 \times 2 \times 3 \times \cdots \times N$.

Note 19: This is true for ideal gases. When there are interactions among the particles there could be situations of equilibrium with a non-uniform distribution of locations. For instance, water at 0°C can have two phases; liquid and solid (ice) at equilibrium. The locational distribution in such a system will not be uniform.

Note 20: If there are $W$ microstates and the distribution is uniform, then the probability of each microstate is $^1/_W$. The corresponding SMI is thus:

$$\text{SMI(for uniformly distributed microstates)}$$

$$= -\sum_{i=1}^{W} p_i \log_2 p_i$$

$$= W\frac{1}{w}\log_2 W = \log_2 W$$

multiply by $k_B$ ln 2, and get the entropy which is $k_B$ ln $W$. You have now discovered the Boltzmann formula for the entropy of a system having $W$ uniformly distributed states. One of the basic postulates of statistical mechanics is that all the quantum mechanical microstates of the system have equal probability. I should note here that $W$ is the *total* number of microstates. Sometimes one makes the distinction between equilibrium and not equilibrium microstates. Also you might read about the entropy associated with equilibrium and entropy associated with non-equilibrium microstates. The entropy is defined only for equilibrium macroscopic state, and it is related to the *total* number of microstate. The last statement is true for isolated systems.

Note 21: It can be shown that the maximal probability *decreases* as $N^{-1/2}$. In practice, we know that when the system reaches the state of equilibrium, it stays there *forever*. The reason is that the macroscopic state of equilibrium is not the same as the state for which $n^* = \frac{N}{2}$ , but it is this state along with a small neighborhood of $n^*$, say $n^* - \varepsilon N \leq n \leq n^* + \varepsilon N$, where $\varepsilon$ is a small number. For $N = 100$ and $\varepsilon = 0.01$, the probability of finding $n$ in the neighborhood of $n^*$ is about 0.235. For $N = 10^{10}$ particles, we can allow deviations of 0.001% of $N$ and the probability of staying in this neighborhood is nearly one. For more details, see Ben-Naim (2008, 2012).

Note 22: Most people will tell you that after you remove the partition, the entropy increases monotonically until it reaches its maximal value at equilibrium. This view cannot be justified either theoretically or experimentally. Think of the volume of the system. Before we remove the partition it is $V$, after reaching a final equilibrium the volume is $2V$. Can you define the "volume" of the gas at the intermediate states in Figure 3.23?

Note 23: More detailed calculations can be found in Ben-Naim (2008, 2010, and 2016).

Note 24: See Ben-Naim (2008, 2010, and 2015a).

Note 25: K represents the degrees in the Kelvin scale. We could also take two liquids or two gases. In the latter, it is easier to calculate the change in entropy for the process $b \rightarrow c$.

Note 26: For details, see Ben-Naim (2008, 2010, and 2015a).

Note 27: For a proof and simulation of this process, see Ben-Naim (2008, 2010, and 2016).

Note 28: This was essentially Boltzmann's answer to the critiques of his H-theorem. See also Ben-Naim (2018).

Note 29: This is different from Boltzmann's answer to his critics. Boltzmann's answer was correct regarding his $H(t)$ function. But Boltzmann was wrong in identifying the behavior of his $H(t)$ function with entropy.

Note 30: In this particular game it is advisable to ask the first question: "Is the chosen marble in the first five cells?" If the answer is either a Yes, or a No, with one more question you will know which marble I chose. Thus, with two questions you are guaranteed to find the missing information.

Note 31: In this particular case the game in Figure 3.31a is "much simpler." You do not need to ask questions.

Note 32: The SMI for any probability distribution $[p(1), p(2), \dots, p(10)]$ is given by the formula:

$$\text{SMI} = -\sum_{i=1}^{10} p(i)\log_2 p(i)$$

where the sum is over all the 10 cells.

Note 33: "Discover Entropy and the Second Law," Ben-Naim (2010).

Note 34: For more simulations with different numbers of marbles and cells, see Ben-Naim (2010).

Note 35: If the marbles are distinguishable, then the number of configurations is $10^{10}$. If the marbles are distinguishable then there are approximately $10^{10}/10!$ $(10! = 1 \times 2 \times 3 \times \cdots \times 10)$. Although this number is smaller than $10^{10}$, it is still a huge number.

Note 36: For two particles in two compartments there are $2^2 = 4$ configurations when the particles are distinguishable, and only two configurations when they are indistinguishable. For three particles there are $3^2 = 9$ configurations when the particles are distinguishable, and only *four* when the particles are indistinguishable. Note that for large numbers of particles the number of configurations is reduced from $N^N$ to $N^N/N$!

You can imagine that when $N$ is of the order of $10^{23}$, the number of configurations is about $10^{10^{23}} = 10^{100,000,000,000,000,000,000,000}$. This is ten followed by more than billions, and billions of zeros.

Note 37: Note the change in the base of the logarithm in equation (4).

$$\text{change in entropy} = k_B \ln 2 \times N\log_2 10$$

$$= k_B \ln 2 \times \frac{\ln 10}{\ln 2} = k_B \ln 10 e.u.$$

Remember the change in entropy for the expansion process from $V$ to 2 $V$, the entropy change was $k_B \ln 2$. Here, the entropy change is due to change of volume from $V$ to $10V$.

Note 38: This applies for a one-component system. If there are $c$-components, we denote them by $N_i$. The number of particles of species $i$, and we assume that all the $c$ numbers $N_i$ are constants.

Note 38: It should be noted here that today any quantity has the same mathematical form as the SMI is referred to as "entropy." These entropies have nothing to do with the entropy of a thermodynamic system. In fact, even Boltzmann himself defined a quantity which we denoted by H, and believed that this quantity exhibits the "behavior of entropy." This is not true. For more details, the reader is referred to Ben-Naim (2018).

Note 39: Here is a summary of the relationship between the SMI, entropy, and super-probability.

I will formulate it for classical systems, and for systems having a very large number of particles.

First, we need to distinguish between two "levels" or probabilities. The first is the probability of finding a specific configuration of all the particles. By configuration, I mean the locations and momenta of all the particles in the system. For simplicity, we assume that there is a finite number of configurations (i.e. we already took into account the uncertainty principle and the indistinguishability of the particles). A specific configuration is a vector $\boldsymbol{c} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N, \boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_N$, where $\boldsymbol{x}_i$ is the location vector and $\boldsymbol{p}_i$ is the momentum vector of particle *i*.

We define the probability $p(\boldsymbol{c})$ of finding a specific configuration $\boldsymbol{c}$. (In a continuous case, we shall refer to $p(\boldsymbol{c})$ as the probability density.)

For each macroscopic system we define the *probability* of finding the probability distribution $p(\boldsymbol{c})$. We denote this probability by $\Pr[p(\boldsymbol{c})]$. Thus, Pr is a function of the probability distribution (for the continuous case Pr is a *functional* of the probability density $p(\boldsymbol{c})$.)

Starting with a system with any arbitrary distribution $p(\boldsymbol{c})$, we can define both Pr and the SMI on this distribution. There is *one* distribution which maximizes both Pr and SMI. This distribution is referred to as the *equilibrium distribution* denoted $p^*(\boldsymbol{c})$.

At this point, we can formulate the Second Law as follows:

Starting from any constrained equilibrium state of well-defined macroscopic system, when removing the constraint, the system will move to a new equilibrium state having a new distribution $p^{**}(\boldsymbol{c})$. At this new equilibrium state, the probability $\Pr[p^{**}(\boldsymbol{c})]$ is overwhelmingly larger than the probability $\Pr[p^*(\boldsymbol{c})]$. Note that both probabilities: $\Pr[p^*(\boldsymbol{c})]$ and $\Pr[p^{**}(\boldsymbol{c})]$ pertain to the system *after* the removal of the constraint. But both $p^*(c)$ and $p^{**}(c)$ are the distributions at equilibrium; the latter and the former correspond to two equilibrium states *before*, and *after* the removal of the constraint.

So far we did not mention entropy. Entropy enters into the formulation only when the process is carried out in an *isolated system* $(E, V, N$ constants). In such a system, the ratio of $\Pr[p^{**}]/\Pr[p^*]$ is related to the *difference in the entropy* of the system in the final and the initial *equilibrium states*. In this case the change in entropy must be positive.

At this point, we recognize the advantage of the probabilistic formulation of the Second Law which is valid for processes in any well-defined system. The entropy formulation is valid only for an isolated system.

If the system is characterized by $T, V, N$ then whenever we remove a constraint from a constrained equilibrium state, the ratio of the probabilities $\Pr[p^{**}]/\Pr[p^*]$ is related to the *difference in the Helmholtz energy* of the system which must be negative.

If the system is characterized by $T, P, N$ then whenever we remove a constraint from a constrained equilibrium state, the ratio of the probabilities $\Pr[p^{**}]/\Pr[p^{*}]$ is related to the *difference in the Gibbs energy* of the system which must be negative.

Thus, we see that the probabilistic formulation of the Second Law is much more general and it applies to any thermodynamic system; $(E, V, N)$, $(T, V, N)$ or $(T, P, N)$. The "*driving force*" for moving from one equilibrium state to another may be ascribed to the probability ratio. As a *result* of this process, the entropy change will be positive for a $(E, V, N)$ system, the Helmholtz energy change will be negative for a $(T, V, N)$ system, and the Gibbs energy change will be negative for a $(T, P, N)$ system. Thus, we see that the probability Pr is the *central primary* concept in formulating the Second Law. The entropy, Helmholtz and Gibbs energy hold a secondary importance, and they are relevant to the specific systems characterized by specific thermodynamic variables.

Note 40: For more details, see Ben-Naim (2018): Time's Arrow, and the Timeless Nature of Thermodynamics.

# APPENDIX A

# A QUALITATIVE PRESENTATION OF THE CONCEPT OF LOGARITHM

The concept of logarithm is the *inverse* of the concept of exponential. Therefore, we start with the idea of exponentiation. The simplest case of exponentiation is to look at the series of numbers.

| Explicit numbers | Short hand notation for the same number |
|---|---|
| $10 = 10$ | $10^1$ |
| $100 = 10 \times 10$ | $10^2$ |
| $1000 = 10 \times 10 \times 10$ | $10^3$ |
| $10000 = 10 \times 10 \times 10 \times 10$ | $10^4$ |
| $100000 = 10 \times 10 \times 10 \times 10 \times 10$ | $10^5$ |

Instead of writing such long numbers explicitly, we use a shorthand notation; the *n*th *power* of 10 is written as $10^n$ which means multiply 10 by itself *n*-times. Thus, $10^2$ means simply the number $10 \times 10$, and $10^4$ means multiply 10 by itself four times. One can generalize this notation for any number *a* (not necessarily 10) *raised* to the power *y* (not necessarily an integer). We write this as:

$$x = a^y \qquad \text{(A.1)}$$

When *y* is not an integer, say $y = 2.7$, we cannot say that we multiply *a* by itself 2.7 *times*. Instead, one can imagine that $x = a^y$ is a number in between the two numbers:

$$a^2 \leq a^{2.7} \leq a^3 \qquad \text{(A.2)}$$

This means that $a^{2.7}$ is somewhere between the two numbers $a^2$ and $a^3$, the latter can be calculated by multiplying *a* by itself 2 and 3 times, respectively. Thus, for any pair of numbers *a* and *y*, we define the number:

$$x = a^y \qquad \text{(A.3)}$$

The logarithm is an inverse of the exponentiation operation in the sense that given *x*, and a *base* *a*, we define the logarithm of *x* to the *base* *a* as:

$$y = \log_a x \qquad \text{(A.4)}$$

This is easy to understand when $a$ is 10, and $y$ is an integer. In this case, the logarithm of $x = 10000$ is simply the number of times you have to multiply 10 by itself to get 10000. In this case:

$$y = \log_{10} x = \log_{10} 10000 = 4 \qquad \text{(A.5)}$$

Similarly, for $8 = 2^3 = 2 \times 2 \times 2$, the logarithm of 8 to the base 2 is $\log_2 8 = 3$ . this is the number of times you have to multiply 2 to get the number 8.

Figure A.1 shows the graph of $\log_{10} x$ for the different values of $x$. Figure A.1b shows the graph of $\log_2 x$, and Figure A.1c shows the graph of $\log_e x$.

The last one is called the *natural* logarithm which is a logarithm to the base of the number $e$ which is approximately:

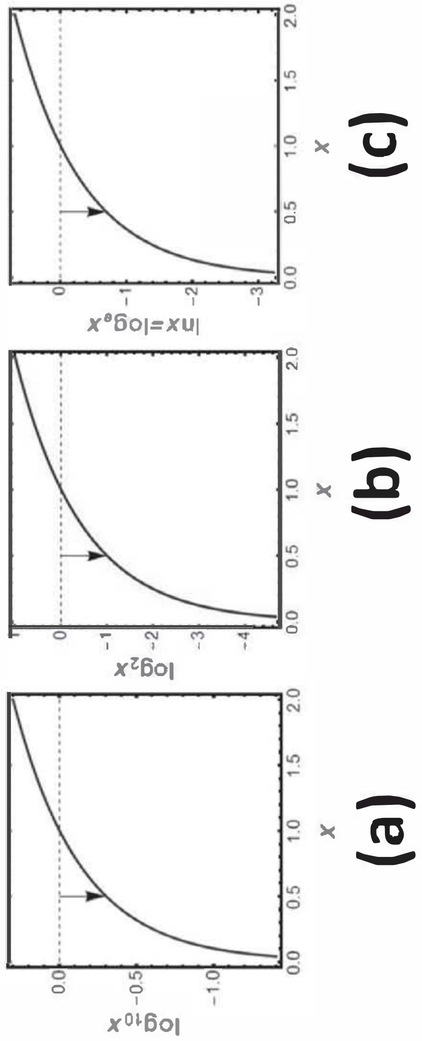


Figure A.1. Logarithm with base 10, 2, and *e*.

$$e \simeq 2.718 \ldots \tag{A.6}$$

This number is called Euler number, and denoted by *e*. It has many interesting properties, and it also arises "naturally" in the sciences. It is the limit of the quantity $\left(1+\frac{1}{n}\right)^n$, when $n$ goes to infinity. Figure A.2 shows this function. It is clear that as $n$ increases this quantity tends to a constant which has the value show in (A.6).

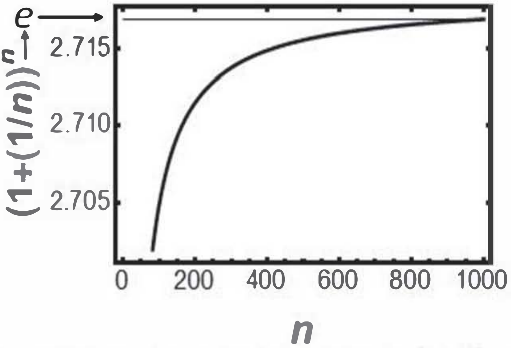


Figure A.2. Definition of $e$ as the limit of the $(1+(1/n))^n$. as $n$ tends to infinity (see arrows).

The natural logarithm is usually written as $\ln x = \log_e x$. This logarithm is useful in physics and mathematics, and in particular in thermodynamics. In information theory the base 2 is the more useful one.

If you still have difficulties in grasping the meaning of the logarithm, you can just look at the value of $\log x$ for any $x$, in either Figures A.1; (a), (b) or (c). For instance, for $x=\frac{1}{2}$ look at the value of $\log_{10}\left(\frac{1}{2}\right)$, or $\log_2\left(\frac{1}{2}\right)$, or $\log_e\left(\frac{1}{2}\right)$ in the respective graphs. Note that $\log 1 = 0$ for the three bases discussed above, and $\log x$ tends to minus infinity $(-\infty)$ when $x$ tends to zero.

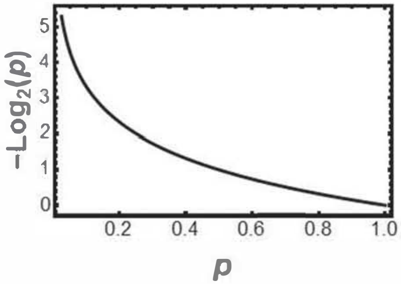


Figure A.3. Plot of $-\mathrm{Log}_2(p)$

In all our applications in this book we use the logarithm of $p$, where $p$ is number between zero and one. Therefore, $\log p$ will be a negative number (or zero when $p = 1$). In Figure A.3 we also draw $-\log(p)$ as a function of $p$. This relationship will be useful in our interpretation of $-\log(p_i)$ as a measure of the extent of uncertainty with respect to the occurrence of the event $i$, in Section 2.7.

# REFERENCES AND RECOMMENDED LITERATURE

Ben-Naim, A. (1987), *Is Mixing a Thermodynamic Process?* Am. J. Phys. **55**, 725.

Ben-Naim, A. (1992), *Statistical Thermodynamics for Chemists and Biochemists*. Plenum Press, New York.

Ben-Naim, A. (2006), *Entropy of Mixing and Entropy of Assimilation, an Information-Theoretical Prospective,* American Journal of Physics, **74**,1126.

Ben-Naim, A. (2007), *Entropy Demystified. The Second Law of Thermodynamics Reduced to Plain Common Sense*. World Scientific, Singapore.

Ben-Naim, A. (2008), *A Farewell to Entropy: Statistical Thermodynamics Based on Information*. World Scientific, Singapore.

Ben-Naim, A. (2009), *An Informational-Theoretical Formulation of the Second Law of Thermodynamics*. J. Chem. Education, **86**, 99.

Ben-Naim, A. (2010), *Discover Entropy and the Second Law of Thermodynamics. A Playful Way of Discovering a Law of Nature*. World Scientific, Singapore.

Ben-Naim, A. (2011), *Entropy: Order or Information*. J. Chem. Education, **88**, 594.

Ben-Naim, A. (2012), *Entropy and the Second Law. Interpretation and Misss-Interpretationsss*. World Scientific, Singapore.

Ben-Naim, A. (2015a), *Information, Entropy, Life and the Universe. What we know and what we do not know*. World Scientific, Singapore.

Ben-Naim, A. (2015b), *Discover Probability. How to Use It, how to Avoid Misusing It, and How It Affects Every Aspect of Your Life*. World Scientific, Singapore.

Ben-Naim, A. (2016a), *The Briefest History of Time*. World Scientific, Singapore.

Ben-Naim, A. (2016b), *Entropy the Truth the Whole Truth and Nothing but the Truth*, World Scientific Publishing, Singapore.

Ben-Naim, A. and Casadei, D. (2017a), *Modern Thermodynamics*, World Scientific Publishing, Singapore.

Ben-Naim, A. (2017b), *Information Theory*, World Scientific Publishing, Singapore.

Ben-Naim, A. (2017c), *The Four Laws that do not drive the Universe*. World Scientific Publishing, Singapore.

Ben-Naim, A. (2017d), *Entropy, Shannon's Measure of Information and Boltzmann's H-Theorem, in Entropy*, **19**, 48-66, (2017).

Boltzmann, L. (1877), *Vienna Academy*. **42**, *"Gesammelte Werke"* p. 193.

Boltzmann, L. (1896), *Lectures on Gas Theory*. Translated by S.G. Brush, Dover, New York (1995).

Brillouin, L. (1962), *Science and Information Theory*. Academy Press, New York.

Brush, S. G. (1976), *The Kind Of Motion We Call Heat. A History Of The Kinetic Theory of Gases In The 19th Century, Book 2: Statistical Physics and Irreversible Processes*. North-Holland Publishing Company.

Brush, S. G. (1983), *Statistical Physics and the Atomic Theory of Matter, from Boyle and Newton to Landau and Onsager*. Princeton University Press, Princeton.

Callen, H.B. (1960), *Thermodynamics*. John Wiley and Sons, New York.

Callen, H.B. (1985), *Thermodynamics and an Introduction to Thermostatics*. 2nd edition. Wiley, New York.

Cercignani, C. (2003), *Ludwig Boltzmann. The Man Who Trusted Atoms*, Oxford University Press.

Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory*, John Wiley and Sons, New York.

Denbigh, K. (1981), *The Principles of Chemical Equilibrium with Applications in Chemistry and Chemical Engineering*, Cambridge Univ. Press, Cambridge, New York.

Feller, W. (1957), *An Introduction to Probability Theory and its Applications*, Vol.I, Wiley, New York.

Feynman, R. (1996), *Feynman lectures on computation*, Addison-Wesley, Reading.

Rakoczy, H., Cluver, A., Saucke, L., Stoffregen, N, Grabener, A., Migura, J. and Call, J., (2014), Apes *Are Intuitive Statisticians,* Cognition, **131**, 60

Reif, F. (1965), *Fundamentals of Statistical and Thermal Physics*, Mc-Graw Hill Book Company, N.Y.

Shannon, C.E., (1948), *A Mathematical Theory of Communication*, Bell Syst. Tech. J., **27**, 379, 623

Tribus, M. and McIrvine, E.C. (1971) *Energy and Information*, Scientific American, 179-188

# INDEX