PHYS20352

Statistical Mechanics

Tobias Galla
The University of Manchester

May 6, 2019

$$S = k \ln \Omega$$

Contents

Chapter 1

Introduction: What is statistical mechanics?

> *"Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the same work, died similarly in 1933. Now it is our turn to study statistical mechanics.*
>
> *Perhaps it will be wise to approach the subject cautiously."*
>
> *(David Goodstein, 'States of Matter')*

1.1 A few practical things

I am this person here:

Dr Tobias Galla
Statistical Physics and Complex Systems Group
Email: tobias.galla@manchester.ac.uk
Web page: http://www.theory.physics.manchester.ac.uk/~galla
Office: 7.16 (Schuster)
Twitter: @tobiasgalla

I work on statistical mechanics (and applications), and I like this topic. I hope it will show! Hopefully I'll manage to convince you that statistical physics is a great topic — deep, fundamental, useful, and fun!

Example sheets and exam:
There will be a 90-minute exam at the end of the course, and one problem sheet per week. It is important that you do the weekly homework problems (before seeing the model answers), tutorial work and tutorial attendance contribute towards your mark.
Please use the tutorials to ask any questions you may have, discuss these with your tutors in the first instance. If you cannot resolve a question, please approach me! There is also the physics help desk service.

I welcome questions immediately before or after lectures. I am usually in the room a few minutes before and I hang around afterwards. So please approach me and ask any questions you may have. With ~ 320 students in the course I may not be able to always answer them all, but I will try my best! It is much easier to discuss physics face-to-face than by email.

What I hope to do as we go along:
In my view, teaching a core undergraduate course has several general objectives:

1. One is the obvious: to deliver the technical content, hopefully in an organised and clear manner; to help you understand the material, and to prepare you for the exam. This is the short-term goal for the time between now and the exam. This is an important part of the learning outcome, but perhaps not the most important.

2. To mention and introduce more general concepts as we go along. Ideas and approaches that you may use later in your life as a physicist. Perhaps in third or fourth year. Perhaps in your MPhys project, or your PhD if you do one. ('Did this guy Galla not say that there was the so-and-so approach to problems like this? – Let's look this up again.').

3. Most of all, to help you become professional physicists. Being a physicist is more than just knowing what time dilation is, or entropy or Maxwell's equations. It is a way of seeing and approaching the world. A way of thinking. One of my main aims is to help you develop this 'physics approach to problem solving'. This course is about statistical physics, but in some sense that is just a vehicle to build up these physics skills. For the next few months it will be statistical physics from 11-12 on Mondays, and from 2-3 on Thursdays. But it could have been any other area of physics.

You can help me to make this a good course and a good experience for everyone. If you have any (constructive) feedback then please let me know, either directly or through your student representatives. Keep in mind though that this is a large class, and naturally different students will have different personal tastes and preferences. So I will not be able to make everybody happy all the time. If you have criticism and feedback then try to be considerate, and do not only think about your personal requirements, but keep in mind the needs of the group as a whole (Google the 'veil of ignorance').

Even more important:
It is important that you go through all calculations, proofs etc in these notes step by step with a pencil and a piece of paper. Make absolutely sure that you understand all equalities, inequalities, and the logic behind the arguments.

Question everything.
Believe nothing.
Trust no one.
Not even your lecturer.

Never take anything for granted. Don't accept until you understand. Thinking 'that's just the way it is' and moving on is never acceptable. Go through all calculations in these notes. There is no other way of acquiring a good and thorough understanding of the material. But then going through those details, being pedantic, getting stuck and finally (hopefully) unstuck and making progress, that's what physics is about, that's where the fun is!

Most important of all
The most important thing is that we all **enjoy** ourselves while we explore this subtle, beautiful and incredibly powerful area of physics.

Recommended textbooks:

- Mandl, F., Statistical Physics, 2nd edition (Wiley)
- Bowley, R. & Sanchez, M., Introductory Statistical Mechanics, 2nd edition (Oxford)
- Zemansky, M.W. & Dittman, R.H., Heat and Thermodynamics, 7th edition (McGraw-Hill)
- Steane, A. M., A complete undergraduate course Thermodynamics (Oxford University Press)
- Blundell, S. J., Blundell, K. M., Concepts in Thermal Physics (Oxford University Press)
- Grimm, J. & Grimm, W. (Brothers Grimm), The Fairy Tales of the Brothers Grimm (Taschen GmbH)

Supplementary reading:

- Atkins, P., de Paula J., Physical Chemistry (Oxford University Press)
- Helrich, C., Modern Thermodynamics with Statistical Mechanics (Springer)
- Callen, H. B., Thermodynamics and an Introduction to Thermostatics (Wiley)
- Andersen, H. C., The Complete Fairy Tales (Wordsworth)

There are literally dozens of books on thermodynamics and statistical physics in the library. Just find the one(s) that you like most and use those.

Other sources:
Dr Judith McGovern has compiled extensive html-based material. It can be found here:

http://theory.physics.manchester.ac.uk/%7Ejudith/stat_therm/stat_therm.html

This material is very systematic and detailed. I highly recommend it. Some of my lecture notes rely on Dr McGovern's material, you will recognise some of the figures, and some of the text.

Acknowledgements:
In preparing this course I have used the notes and materials of lecturers who gave it before me. In sequence these are Prof. Alan McKane, Dr Judith McGovern, Prof. Ray Bishop and Dr Yang Xian and Prof. Jeff Forshaw for the parts on quantum gases. If you find these lecture notes useful, then credits should really go to these earlier lecturers. A lot of the text in these notes is taken directly from their materials. I have edited and re-organised the material to give it my personal 'spin', and I will continue to edit the notes. If you find typos or mistakes then they are likely to be mine. Please point them out to me, so that I can correct them.

1.2 The general objective of statistical physics: connecting the micro and the macro worlds

1.2.1 Microscopic and macroscopic properties of systems with many particles

The objective of statistical physics is to connect the behaviour of the macroscopic world with the rules that govern the behaviour of the microscopic world.

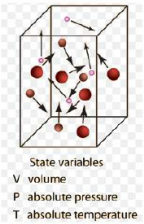
Macroscopic behaviour:
Statistical physics looks at systems made up of a large number of particles (typically $\sim 10^{23}$, Avogadro's number 6×10^{23}/ mole), e.g., a piece of metal, a cup of water, a box of gas, etc. When we talk about the 'macroscopic behaviour' of a system we refer to the collective behaviour coming about from the interaction of all constituents. A few examples follow below.

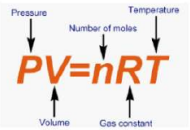
Microscopic perspective:
When we say we look at the 'microscopic' dynamics of a system we mean looking at the dynamics of individual constituents, including the interaction of that one constituent with other particles in the system. But the focus is on the individual particle, not the collective.

Examples:

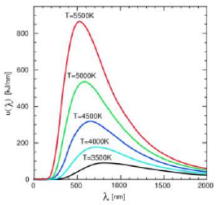
- In a gas a single molecule moves freely between 'bounces' off other gas molecules, and collisions are elastic (in the case of an ideal gas). Each gas molecule follows Newton's laws (if we consider a classical gas). In non-ideal gases there may be pairwise interactions between molecules. All this goes under microscopic behaviour, it concerns the motion of

individual fas molecules. But when one mole of such a gas undergoes a phase transition (say from gas to liquid) then this is a collective phenomenon, a macroscopic feature. Individual gas molecules do not undergo a phase transition – only the collective. On the macroscale we observe things such as the ideal gas law $PV = nRT$; this is a relation between macroscopic properties of the ideal gas. Individual gas molecules do not have a pressure or temperature – these quantities are only defined for the macroscopic system.

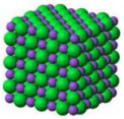




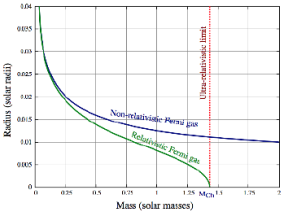
- We will look at quantum gases, for example a gas of photons. Individual photons are governed by laws that are different to those for the constituents of classical gases. For example we would have to know about the relation between energy and frequency of photons. These are the properties of the microscopic constituents. On the macro-level we would be interested in things such as the distribution of energies in a gas of photons contained in a cavity. This results for example in Planck's law – a property of the photon gas as a whole, not of individual photons.



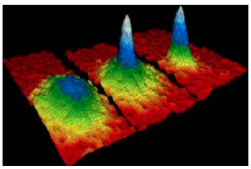
- Solid crystals are another example of systems with many constituents. These are the atoms of the crystal. In some cases they are ionised and we then also have to consider the electron gas in the crystal lattice. These microscopic constituents follow the laws of quantum mechanics. We can then ask what the macroscopic properties are of such a crystal, for example things such as its specific heat, Young's module (if you are interested in mechanical properties). The electron gas is a Fermi gas in the context of this course, we will investigate the macroscopic properties of such gases in detail.



- White dwarf stars are another example of a Fermi gas. The underlying constituents are electrons, governed by the laws of quantum mechanics (such as the Fermi exclusion principle). Macroscopically one finds a balance between gravitational pressure and outward pressure due to the exclusion principle. In order to study this, we need to investigate pressure in a degenerate Fermi gas – a collective (macroscopic) property arising from the interaction of the constituents.

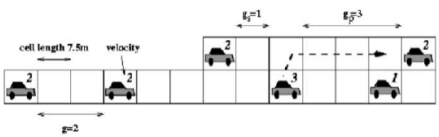


- Bose Einstein condensation is a collective phenomenon seen in gases of bosons at low temperatures. It is responsible for phenomena such as superfluidity and superconductivity. The theory of such gases was developed by Bose and Einstein, who predicted the condensation phenomenon in 1924/25. It was experimentally realised on 5th June 1995 by Cornell and Wiemann, and then a few months later by Ketterle. The three won the Nobel Prize in 2001. We will discuss both the theory and the experiment in this course.



- Pattern formation in biology. As an example I show a flock of birds. What are the rules governing the behaviour of individual birds? And how do they interact together to produce the beautiful macroscopic structure? Ultimately, this also falls within the remit of statistical mechanics.

- As another example consider a model of cars on a road (for example the celebrated Nagel-Schreckenberg cellular automaton). Each constituent in the model is a car, and they move forward according to very simple rules: at a given speed, keeping a certain distance from the car ahead and so on. These are the microscopic rules, the rules that you need to know to write a computer code for that model. But when you run the simulation you see things such as traffic jams. They are a collective phenomenon, seen at the macroscopic level (they involve many particles) These macroscopic phenomena 'emerge' from the interaction of the microscopic rules, but they cannot be understood from looking only at a few microscopic constituents.



1.2.2 The objective of statistical physics

Laws that govern the microscopic world are Newton's laws (classical), or Schrödinger's equation (quantum), etc. At the same time, we observe in experiments that macroscopic objects obey definitive laws: the ideal gas law, water boils at 100 degrees Celsius at standard atmospheric pressure, Planck's formula for black body radiation, etc.

General question: How to infer the laws that govern the macroscopic behaviour of systems with many constituents from the laws governing the microscopic behaviour of the individual constituents?

1.2.3 Deterministic versus statistical mechanics

Deterministic mechanics

In principle, the laws for the microscopic constituents apply to systems with a large number of such constituents. So one could take the view that it is now just a matter of dealing with this, for example solving Newton's equations for a gas of 10^{23} interacting classical particles. This would mean to solve 10^{23} coupled second-order differential equations. Tough, but well,

in principle the problem is well defined. Given the initial condition everything is fully deterministic. This is what I will call 'deterministic mechanics' – solving Newton's equation from a given starting point. There is nothing random here, nothing is statistical. Simply solving a (large) set of differential equations.
This view is related to the idea of Laplace's demon.

Laplace's demon:
At the beginning of the 19th century the belief was that studying systems with many constituents is mostly a matter of solving the required equations. We quote Pierre-Simon Laplace (1814):

> *"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."*

Clearly, this is not practical. Even with modern-day computers we cannot solve Newton's equations for all particles in the universe. Also, we do not know their positions and momenta (initial conditions). Some additional problems with quantum mechanics too (uncertainty).
Further, we are not really interested in the position and direction of motion of all particles in a macroscopic system. Instead what we often want to know are the bulk (macroscopic) properties. Or have you ever asked in a pub what the exact state of momentum of molecule number $1,424,743$ is in your pint of bitter? Instead you might ask about its temperature, or perhaps its colour (also a macroscopic property).

Statistical mechanics

In deterministic mechanics the state of a large system is mathematically described by a point in phase space. For a classical n-particle system, this would be a $6N$ dimensional space (three position and three momentum coordinates per particle). The dynamics of the system is then a trajectory in this space. If I know where the system is in this space at any one time, I can compute its past and its future, by solving Netwon's equations. That's Laplace's demon.

Statistical physics takes a very different view. It looks at ensembles – that is large groups of (virtual) copies of the system. Each of these systems is described by a point in phase space, and the ensemble defines a 'cloud' of points in phase space. We are not so much interested in how each and every phase space point in this cloud moves, instead we focus on the probability distribution defined by the cloud in phase space. We study mechanics in a statistical sense – what does the probability distribution of points in phase space look like?
In this course we will develop the principles that govern the shapes of these distributions, and study how the statistical behaviour of systems with many components defines the thermodynamic quantities that you already know – such as entropy for example.

What we will not do in this course is to analyse large data sets. This is not a course in statistics. This is statistical mechanics (or statistical physics).

Important Warning: Statistical Physics $\neq$ 'stats'
Please do not refer to statistical physics as statistics or 'stats'. What we do in this course is **not** statistics. We are not trying to analyse datasets. This course is about statistical physics – the theory which allows us to connect the microworld with the macroworld. If you email me and say *I have a question about 'stats'*, I will be in a bad mood straight away – TG.

1.3 How is statistical physics related to phenomenological thermodynamics?

1.3.1 Distinction between the two approaches

Broadly speaking, there are two approaches to understanding macroscopic physics. One is at the center of this course:

Statistical mechanics:
Starts from the knowledge of the microscopic nature of matter (atomic and molecular interactions etc) and aims to deduce the thermodynamic behaviour at the macro-level. Verifies thermodynamics but also provides a systematic formalism and method of calculating quantities at the macro-level.

You already know the other one from your 'Properties of Matter' course in first year. I will call it phenomenological thermodynamics.

Phenomenological thermodynamics:
Formulated without using knowledge of the microscopic nature of matter. Based on a small number of principles, the *laws of thermodynamics*. These are deduced from experiments.

Historically, thermodynamics of course comes before statistical physics. Most of thermodynamics was developed in the first half of the 19th century. The statistical physics perspective was first developed by Boltzmann and others, towards the end of the 1800s.

It is very important to understand the distinction between the two approaches. There are several:

1. Thermodynamics is 'phenomenological'. Thermodynamics describes phenomena observed in experiments, and to characterise these phenomena it uses quantities such as

heat, pressure or temperature. These quantities are all directly motivated by experiments. The main results of thermodynamics are not derived from mechanics, electrodynamics or quantum mechanics. Instead thermodynamics starts from empirically sensible relations or statements (the laws of thermodynamics).[1]

2. Statistical physics describes probability distributions in the space of states that a system can take. In such a probabilistic view there will be fluctuations, at least when the system is made up of a finite number of constituents. We will discuss this in more detail. Thermodynamics ignores these fluctuations, and studies the average behaviour resulting from the myriads of atomic transitions and degrees of freedom. It does not study these microscopic details themselves. Thermodynamics is only interested in the relations between macroscopic observables, after the microscopic details have been averaged out.

   Gibbs said this very nicely:

   > *"The laws of thermodynamics, as empirically determined, express the approximate and probable behavior of systems of a great number of particles, or, more precisely, they express the laws of mechanics for such systems as they appear to beings who have not the fineness of perception to enable them to appreciate quantities of the order of magnitude of those which relate to single particles, and who cannot repeat their experiments often enough to obtain any but the most probable results."*
   > *(J. Willard Gibbs)*

1.3.2 Brief summary of classical thermodynamics

Classical thermodynamics centres around the first and second law[2].

[1] I guess you could argue that classical mechanics or quantum mechanics also try do describe phenomena seen in experiments. That is undoubtedly true. But I think we would all agree that relations such as $F = ma$ or the Schrödinger equation are somehow less phenomenological than for example the first law of thermodynamics.

[2] The real first law of thermodynamics is of course that you do not talk about thermodynamics.

First law

The first law of thermodynamics states

$$\Delta E = Q + W \tag{1.1}$$

in any change of a thermodynamic system. In this equation

- Q is the amount of heat added to the system
- W is the work done on the system
- ΔE is the change in the internal energy of the system

Remark:
The first law is just a statement of the principle of conservation of energy, recognising the fact that heat is a form of energy.

Second law

The basic observations leading to the second law are remarkably simple:

- When two systems are placed in thermal contact they tend to come to equilibrium with each other - the reverse process, in which they revert to their initial states, never occurs in practice.
- Energy prefers to flow from hotter bodies to cooler bodies (and temperature is just a measure of the 'hotness'). It is everyday experience that heat tends to flow from hot bodies to cold ones when left to their own devices. There is a natural direction to spontaneous processes (e.g., the cooling of a cup of coffee)
- Physical systems, left to their own devices, tend to evolve towards disordered states, e.g. the mixing of milk in coffee. The reverse (spontaneous ordering) is not observed in closed systems.

Note: These observations can be used to define a 'direction of time'.

Related to this, it is important to realise that work and heat are simply different forms of energy transfer:

- Work is energy transfer via the macroscopically observable degrees of freedom. In this case the energy is transferred *coherently*.
- Heat is energy transfer between microscopic degrees of freedom. In this case the energy is transferred *incoherently*, the energy is stored in the thermal motion of the molecules of the substance.

The second law says that there is an intrinsic asymmetry in nature between heat and work: extracting ordered motion (i.e. work) from disordered motion (i.e. heat) is hard[3], whereas the converse is easy. In fact the second law says that it is impossible to convert heat completely into work.

There are two classic statements of the second law of thermodynamics.

Kelvin-Planck statement:
It is impossible to construct an engine which, operating in a cycle, produces no effect other than the extraction of heat from a reservoir and the performance of an equivalent amount of work.

Clausius statement:
It is impossible to construct a refrigerator which, operating in a cycle, produces no effect other than the transfer of heat from a cooler body to a hotter one.

Consequences of the second law

I label statements relating to Carnot engines as consequences of the second law. This is how the theory is best presented. Historically though, the work of Carnot (1824) preceded the Clausius and Kelvin statements (1850s). Planck contributed even later (he was only born in 1858).
First recall what a Carnot engine is:

Definition:
A Carnot engine is a reversible engine acting between only two heat reservoirs. That means that all processes are either isothermal (heat transfer at a constant temperature) or adiabatic (no heat transfer).

From the Kelvin-Planck or Clausius statements of the second law one can then derive Carnot's theorem.

[3]In his 1848 paper titled 'On an Absolute Thermometric Scale' Kelvin writes: 'In the present state of science no operation is known by which heat can be absorbed, without either elevating the temperature of matter, or becoming latent and producing some alteration in the physical condition of the body into which it is absorbed; and the conversion of heat (or caloric) into mechanical effect is probably impossible, certainly undiscovered.' In a footnote he then goes on to say: ' This opinion seems to be nearly universally held by those who have written on the subject. A contrary opinion however has been advocated by Mr Joule of Manchester; some very remarkable discoveries which he has made with reference to the generation of heat by the friction of fluids in motion, and some known experiments with magneto-electric machines, seeming to indicate an actual conversion of mechanical effect into caloric. No experiment however is adduced in which the converse operation is exhibited; but it must be confessed that as yet much is involved in mystery with reference to these fundamental questions of natural philosophy.' Beautiful – TG.

Carnot's theorem:
A reversible engine operating between two given reservoirs is the most efficient engine that can operate between those reservoirs.

You may remember the idea of the proof. You assume that there is a more efficient reversible engines operating between the two reservoirs. Then couple this with a Carnot engine run in reverse to construct a contradiction to the second law.
From Carnot's theorem (or directly from the second law) one can show, using a similar idea,

Corollary of Carnot's theorem:
Any reversible engine working between two heat reservoirs has the same efficiency as any other, irrespective of the details of the engine.

Thermodynamic definition of entropy

The second law of thermodynamics and the theory of Carnot engines can be used to introduce thermodynamic entropy. One first notes the Carnot relation for reversible engines operating between two heat baths (Q_H, Q_C are the amounts of heat exchanged with the hot and cold heat baths respectively).

Carnot relation

$$\frac{Q_C}{T_C} = \frac{Q_H}{T_H}, \tag{1.2}$$

applicable to Carnot engines.

This equality can be generalised to a broader set of reversible processes, see Eq. (1.4) below. Using the Carnot relation and the second law one can show the inequality in (1.3), known as Clausius' theorem.

Clausius' Theorem
If a system is taken through a cycle, the algebraic sum of the heat added weighted by the inverse of the temperature at which it is added, can never be greater than zero,

$$\oint \frac{đQ}{T} \leq 0\,. \tag{1.3}$$

The equality holds if the process is reversible, i.e.

$$\oint \frac{đQ^{\text{rev}}}{T} = 0\,. \tag{1.4}$$

The proof is somewhat subtle, a good account can be found in the book Blundell and Blundell, Concepts in Thermal Physics, Oxford University Press. The book by A. M. Steane, A complete undergraduate course Thermodynamics, Oxford University Press also has a very nice description. Go and look up these books! They are really good.

Based on Clausius' theorem one defines the function of state entropy as follows.

Definition:
For a given system, the change in entropy for an infinitesimal reversible change is defined by

$$dS = \frac{đQ}{T}. \tag{1.5}$$

Clausius' theorem (which in turn was derived from the second law) implies that S is a function of state.

Important: Please make sure you understand what this means – the 'function of state' property is not trivial!

Using this definition of entropy and Clausius' theorem one finds:

Important conclusion:
If an amount of heat $đQ$ is supplied to a system from a source at temperature T, then the change in the entropy of the system satisfied the inequality

$$dS \geq \frac{đQ}{T}. \tag{1.6}$$

The equality applies if the process is reversible.

Isolated systems:
If a system is thermally isolated $đQ = 0$. This means that

$$dS \geq 0 \quad \text{for thermally isolated systems} \tag{1.7}$$

$\Longrightarrow$ The entropy of an isolated system cannot decrease.

This leads to:

Maximum-entropy principle

An isolated system at equilibrium must be in the state of maximum entropy.

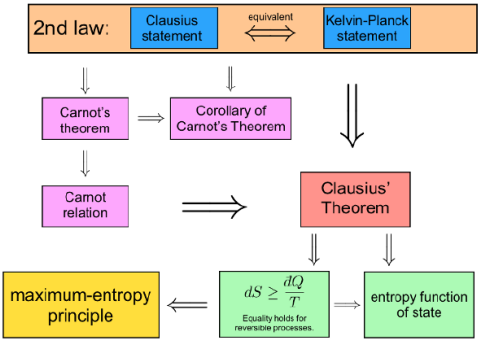


Figure 1.1: Summary of the logic in phenomenological thermodynamics.

This is an alternative statement of the second law. Note: Any system plus its surroundings (in thermodynamics deliberately called by the somewhat grandiose title 'the universe') forms an isolated system. Hence we have another statement:

> The entropy of the universe never decreases.

What this means is that any decrease in the entropy of a system must be at least compensated by an increase in the entropy of its surroundings.

1.4 Why is this course interesting?

I hope the text in Sec. 1.2 has already sparked your interest. Statistical mechanics is an incredibly powerful approach to studying systems made up of many constituents. In fact it is the only productive approach that we know. And the range of problems to which it applies is broad, both within physics and beyond.
That in itself should provide enough motivation, but I'd like to mention a few further aspects.

The place of thermodynamics in physics:
In the course on Properties of Matter you have studied thermodynamics in detail. A lot of this course was about things such as steam engines, their efficiency, etc. At first sight this may sound a little boring, and often don't realise the importance of thermodynamics and statistical physics (I had those difficulties when I was a student. Like most physics students I didn't

particularly like thermodynamics. We referred to it as 'thermodramatics'.)

Luckily, we are not going to study steam engines in this course.

But this course on statistical mechanics will hopefully help you to place thermodynamics in a wider context. We are now going to study 'what came after thermodynamics'. Hopefully this helps to convince you that thermodynamics is actually pretty fascinating. Classical thermodynamics, along with classical mechanics, electrodynamics and quantum mechanics forms one of the key pillars of physics. You cannot call yourself a trained physicist without knowing about these theories and areas. The development of thermodynamics was key for the industrial revolution in the early 19th century, and it continues to be relevant for an incredibly broad range of applications. These include all areas of engineering and chemistry, climate, as well as the thermodynamics of living systems.

History of thermodynamics:
Reading about the history of thermodynamics and now statistical physics you will get a feeling of the struggle the early heroes of thermodynamics went through in order to understand concepts such as heat, its conversion to mechanical energy and what we now call the first law. These things seem trivial to us now, but they were only properly understood about 150 years ago. As always in science you have to see these achievements in the context of what was known then and what wasn't. Up until that point people thought heat was some kind of fluid, the 'caloric', and they had no appreciation of the underlying microscopic processes. The idea of atoms wasn't to be established until the early 20th century. It is fascinating to see how a systematic understanding of these issues, now compactly formulated in the first, second and third law of thermodynamics, have enabled and continue to enable both industrial and technological advances. They have given us a deeper grasp of the world around us, including the living world. The construction of thermodynamics and the theory of statistical physics are a truly amazing success stories.

Stochastic thermodynamics:
The theories of classical thermodynamics and statistical physics apply to systems composed of *many modern* particles. Concepts such as heat, work or entropy really only make sense in this context. More recently the new field of 'stochastic thermodynamics' has emerged. Here, researchers apply the above concepts to systems composed of single particles or a few particles. Examples include the work done, entropy produced etc when single molecules are being stretched with optical tweezers (e.g. a single DNA molecule). Concepts from classical mechanics can be modified and adapted to describe such processes. This is very much in development, the corresponding experiments have only become possible in the last decade or so. We will (unfortunately) not have time to cover these aspects in this course. If you want to read about it, google 'stochastic thermodynamics' or 'Jarzynski inequality'.

Non-equilibrium statistical physics:
This course covers aspects of equilibrium thermodynamics and statistical physics. In most situations this means that we assume that all quantities of all systems we look at are time-

independent, and that all changes of e.g. external conditions we apply, are applied to slowly (quasi-statistcally) that the system is always in equilibrium. Occasionally we also look at irreversible processes. The system may then temporarily be out of equilibrium, but at the end of the process we let it 'equilibrate' again and look at the final equilibrium state that results. In this course we never really study the non-equilibrium states along irreversible trajectories, but only the equilibria at the beginning and end of the process. Most of the real world operates far from equilibrium though, and such systems *never* reach an equilibrium state. The best example is biology. Systems are driven by external energy, conditions change rapidly, currents flow and there is heat exchange and motion for example in cellular motors, the heart beat, etc. The theory of equilibrium statistical physics is now essentially complete (the final rosetta stone was the so-called renormalisation group, invented in the 1960s and 1970s). Most ongoing research is on off-equilibrium statistical physics, very interesting applications are found in biological systems and the 'physics of life'.

Complex systems:
In the last 20-30 years ideas from non-equilibrium statistical physics have been applied to questions outside physics, in particular applications in economics, the social sciences and in biology. One here studies so-called agent-based or individual-based systems, consisting of a larger number if interacting agents. These can be traders in the context of a stock market, cars if you model road traffic, or genes and proteins in biology. Such systems are often intrinsically off-equilbrium, and the tools from statistical physics can be used to derive a description of the phenomena shown by these systems at the macro-level (e.g. a traffic jam, stock market crash), starting from the interactions at the micro-level. Again, we do not have time to talk about these topics in this course, but they are discussed in the third-year course on nonlinear physics and in the fourth-year module on advanced statistical physics.

1.5 The distinction between macrostates and microstates

1.5.1 Macrostates and microstates

As stated, the main purpose of statistical physics is to connect the behaviour or state of the microscopic constituents of a system to its macroscopic properties. In order to conduct our business, we therefore have to be very clear with regards to what we mean by the state of the microscopic constituents, and the macroscopic state of the system as a whole. This is the distinction between macrostates and microstates.

The term macrostate pertains to the bulk features of a systems composed of many particles. Macrostates are described by a few thermodynamic variables, such as P, V, T and E, S for a gas for example, but makes no reference to the state of any individual constituent of that system. A macrostate is the current disposition of the system defined in terms of macroscopic variables.

A microstate, on the other hand, pertains to properties of the individual constituents of a system. For a single particle, we cannot speak of pressure, temperature of volume in a thermodynamic sense. But we can specify its physical state of motion. A microstate of a system is specified by the physical states of all of the system's constituents, not just by the bulk properties of the system.

This is best illustrated with a few examples.

1.5.2 Examples

Gas of classical particles

Microstates:
In classical mechanics, the state of a single particle is specified by its position, $\underline{r}$, and momentum $\underline{p}$. Suppose now we have a system of N such particles. What specifies the microstate? A microstate of this system is specified by the physical states of all constituents, i.e., by the positions and momenta of all N particles. Therefore microstates of the gas are specified by

$$(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_N, \underline{p}_1, \underline{p}_2, \ldots, \underline{p}_N)$$

Since each of these vectors has three components, the combined vector has $6N$ entries. It lives in a $6N$-dimensional space, the so-called phase space of the N-particle system.

Macrostates:
A macrostate of a gas is described by any combination of two variables out of P, V and T. For example P and V. Other representations are possible, e.g. E and V. You have discussed this in the context of thermodynamic potentials. The key thing to notice is that these quantities $(P, V, T, E, S, ...)$ have no particle index. Position vectors and momenta are defined for individidual particles, hence the index i in $\underline{r}_i$ and $\underline{p}_i$. Macroscopic quantities have no particle indices. Macrostates are specified by a combination of a suitable number of macro-quantities.

System of quantum particles

In quantum mechanics, we use wavefunctions to describe the physical state of a single particle. Wavefunctions are usually specified by a set of quantum numbers. For example, a state of an electron in hydrogen atom is specified by a set of quantum numbers (n, l, m, σ), where n is the principle quantumnumber, l the angular momentum quantum number, m the z-component of angular momentum, and finally $\sigma = \pm 1/2$ is its spin quantum number. Notice that quantum states are discrete, i.e., the parameters (n, l, m, σ) are discrete numbers, in contrast to the classical micro-states which are described by continuous variables, $\underline{r}_i$ and $\underline{p}_i$.

For a quantum particle moving in a box, its state is a plane-wave specified by three discrete components of its momentum $\underline{p} = (p_x, p_y, p_z)$ together with its spin σ. We will discuss this

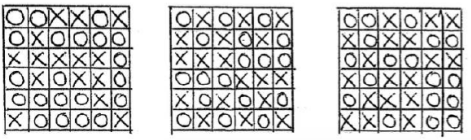


Figure 1.2: Illustration of microstates of a 6×6 arrangement of spin-1/2 particles (see text for details).

in details later. For a system of N quantum particles in a box, the collection of all quantum numbers

$$\left(\underline{p}_1, \underline{p}_2, \ldots, \underline{p}_N; \sigma_1, \sigma_2, ..., \sigma_N\right)$$

is used to specify the micro-states in the so-called independent particle approximation (ignoring the interactions between particles). Later, we will consider a system of N localized spins in a solid, for which a micro-state is specified by a given spin configuration $(\sigma_1, \sigma_2, ..., \sigma_N)$.

We notice immediately that the number of micro-states is huge when N is large, both in the classical case and in the quantum case.

Warning:
The distinction between microstates and macrostates occasionally seems to cause confusion among students. This is at least what I have noticed. When asked to explain the difference between microstates and macrostates, students sometimes give an answer along the lines of 'A microstate is a state of a small system, and a macrostate is the state of a large system'. While the second part of the sentence is not entirely inaccurate (but not very precise either), it is hopefully clear to you that the first part is just wrong.

Third example: system of distinguishable spin-1/2 particles

We'll now look at a system of spins on 6×6 checkerboard, see Fig. 1.2. Each square of the board represents a localised spin-1/2 particle with two possible spin orientations, one is spin-up (marked X), and the other spin-down (marked by O).

Microstates:
Any particular configuration of these up and down spins corresponds a microstate. Many different patterns are possible. In fact, it is easy to compute how many microstates there are. We have $N = 36$ spins, and each spin can be in one of two states. So the total number of microstates for this system is $\Omega = 2^N = 2^{36} \approx 7 \times 10^{10}$.

Macrostates:
To specify macrostates of the system, we are not interested in the state of each and every particle, but only in the bulk properties. The relevant bulk property here is the total number

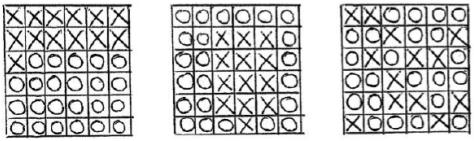


Figure 1.3: Illustration of 3 different microstates all belonging to the macro state '13 up-spins'.

of, say, up-spins $n_\uparrow$. The number of down-spins is then $N - n_\uparrow$, where $N = 36$. As a consequence there are 37 macrostates, namely those with $n_\uparrow = 0, 1, 2, \ldots, 36$ up-spins.

We immediately see that there are many microstates which belong to the same macrostate. For example, if the macrostate is "13 up-spins" ($n_\uparrow = 13$), then we can pick those 13 spins arbitrarily among the total of 36 spins. There are many possibilities to do that, some of these are illustrated in Fig. 1.3.

How many possible such microstates are there for a given macrostate? This is just the common combinatorial problem of splitting a group of N identical objects into two smaller of sizes $n_\uparrow$ and $(N - n_\uparrow)$. The number of ways of doing this is N-choose-$n_\uparrow$,

$$\left(\begin{array}{c} N \\ n_\uparrow \end{array} \right) := \frac{N!}{n_\uparrow!(N - n_\uparrow)!}. \tag{1.8}$$

In our example, $N = 36$ and $n_\uparrow = 13$, so the total is 2.31×10^9. Similarly for $n_\uparrow = 15$ there are only 5.57×10^9 ways, whereas for $n_\uparrow = 18$ there are 9.08×10^9 ways (this is the maximum for this case).

The numbers $N!/[n!(N - n)!]$ are the binomial coefficients. They appear in the binomial expansion

$$(a + b)^N = \sum_{n=0}^{N} \left(\begin{array}{c} N \\ n \end{array} \right) a^{N-n} b^n, \tag{1.9}$$

and in Pascal's triangle.

1.5.3 Distinguishable and indistinguishable particles

In the previous checkerboard example we have implicitly assumed that the different particles are *distinguishable*. Each one of them resides in a certain place on the checkerboad ('the crystal lattice'), and we can tell them apart. This is why there are 2^N microstates for this system, it makes a difference which one of the spins are up and which ones are down. If the spins had been *indistinguishable* our counting of microstates would have been different. If we can't tell the spins apart then all we can know is how many of them are up and how many are down. There are then only 37 microstates, namely the states with $n_\uparrow = 0, 1, \ldots, 36$ up spins. There are also 37 macrostates, each corresponding to one of the microstates.

This is further illustrated, for a different example, in Fig. 1.4. We consider a two-particle two-level system. Each particle can be in one of two energy states, the lower one has energy

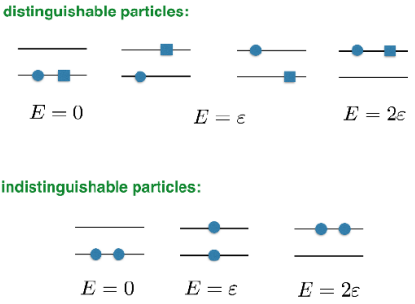


Figure 1.4: Microstates and macrostates for systems with distinguishable and indistinguishable particles.

0, the upper one energy ε. As a consequence there are three macrostates, with total energies $E = 0, \varepsilon, 2\varepsilon$. It does not matter whether particles are distinguishable or not.

However, looking at microstates, it makes a difference if particles can be distinguished or not. In the upper part of Fig. 1.4 the two particles are distinguishable. There are then four different microstates. One of these corresponds to $E = 0$, another to $E = 2\varepsilon$, and the macrostate $E = \varepsilon$ is made up of two microstates.

If the particles are indistinguishable, then there are only three microstates, as shown in the lower part of the figure.

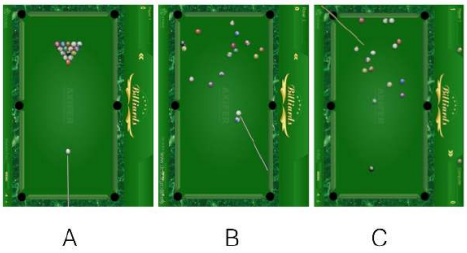
1.6 The arrow of time

One formulation of the second law states that the entropy of isolated systems never decreases with time. This defines a direction of time (the 'arrow of time'). Later in this course we will discuss a statistical interpretation of entropy, and of its increase. This provides a more fundamental interpretation of the second law.
This section contains a brief prelude to that.

Example 1:
Let us look at the following figure.

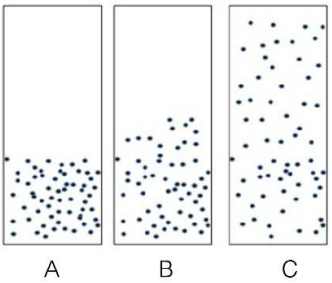
Clearly, there is no way to time-order the three images, all sequences A-B-C, B-A-C, C-B-A, etc. seem perfectly valid time orderings, and one can't say which one of these is more sensible than the other.

Example 2:
Next, we'll look at this example here:



It is hard to tell whether B comes before C, or C before B, but everybody will agree that A comes before B and C.

Example 3:
Finally, consider this example:



Here, there is only one valid time ordering, A-B-C.

Now, we try to analyse this in a bit more detail. Why exactly did we say that A comes before B and C in the second example? And why exactly is A-B-C the only time ordering that seems sensible to us in the third example? Well, because you will have never seen B or C turn into A in a match of pool (second example), and particles in a container never spontaneously assemble in one part of the container (sequence C-B-A in the third example). We just never see things like this.

Now here is the thing. Look at the third example. Suppose we start from A, and let the system go. It will reach B, and then finally C. Let's freeze the positions when C is reached. Then reverse all momenta (velocities), and let the system run again. It will go back to B, and then finally A. The sequence C-B-A is not actually forbidden by the laws of physics (Newton's equations in this example[4]). For the first and second examples (match of pool) we have to add a caveat[5]

Summary:

1. The natural ordering, commensurate with our experience, is A-B-C in example 3 (A before B or C in example 2).
2. However, the laws of physics do not forbid the sequence C-B-A.
3. Our experience comes from the fact that a situation like A is likely to develop into one like C, but the reverse, while not impossible, is very unlikely (it would require very special initial conditions).
4. In example 1, there is no natural ordering.

So then, why do we say (in the third example) that A-B-C is a valid time ordering, but C-B-A is not? Both are perfectly possible under the laws of physics.

The reason is as follows: if you start from a situation that looks like A in the third example, then it is very likely that you'll get to a situation that looks like B and then later C. The reverse is much harder. Yes, it is possible to construct an initial condition that looks like C and leading to A, but you have to choose the initial velocities of all particles very carefully. For most starting positions that 'look like C', you won't get to anything that 'looks like' A – that's why we never see this.

[4] Formally, Newton's equations are second-order differential equations, so invariance under time-reversal, $t \leftrightarrow -t$.

[5] The balls will slow down over time (due to friction). So if I show you a movie of the thing, you'd be able to tell if B comes before C or the other way round in the second example, just by looking at the speed of the balls. But for the sake of the argument let us say that there is no friction, and the total energy contained in the motion of the balls is conserved. Note that friction results in a term proportional to $\dot{r}$ in Newton's equations. This is a first-order derivative, and the invariance under time-reversal breaks down.

To make the discussion more precise, we need to say with more precision what we mean by things such as 'situation that looks like C'. First of all, we note that the microstates of the system in example 3 is specified by the positions and the momenta of all particles. The momenta (velocities) are not drawn in the figure, but you can easily imagine them.
So, then what is a macrostate for our purposes? Imagine we could divide microstates into groups with an equal 'amount of disorder'. For example, there are many microstates that 'look like A' – let's group them all into one macrostate. 'Look like' in this context means 'has a similar degree of disorder'. We have not yet introduced a mathematical measure of disorder – doing this is one of the main purposes of this course. But let's imagine we had such a quantity. We'd then group microstates by their disorder. Saying that two microstates belong to the same macrostate is then the more precise version of saying 'looks like'. For example microstates B and C in example 2 have roughly the same degree of disorder ('B looks like C'), so we'd say that they both belong to the same macrostate.

Everything in the examples can then be explained based on the following two principles:

1. Macrostates with a high degree of disorder have more microstates than macrostates with low disorder.
2. If a system is started from given initial condition, then ultimately all possible microstates will be assumed with equal probability.

We can't verify statement 1 without a definition of 'disorder', but as we will see disorder (entropy) of a macrostate will be defined precisely through the number of microstates that belong to the macrostate. Statement 2 is a postulate at this point, but we will give very good information theoretic reasons for it later on.
Statment 2 means that ultimately any microstate is as likely as any other. But the vast majority of microstates will belong to the macrostate with highest disorder ('looks like C' in example 3). This is why, eventually, the system in example 3 will – with overwhelming probability – 'look like C' at long times. Systems tend from macrostates with low disorder (few microstates) to those with high disorder (many macrostates). And this is what our intuition tells us. In all cases above when we able to order states, it was based on the degree of disorder. Let's look at the three examples again.

Example 2:
In example 2 we said that A comes before B or C. Why? Because 'A looks more ordered'. We can't say whether B or C comes first, that's because they are equally disordered (roughly).

Example 3:
Our intuitive ordering is A-B-C, that's because C is more disordered than B, and B is more disordered than A.

Example 1:
In example 1 finally we could not order the three states. That's because they all look roughly equally disordered.

Chapter 2

Brief introduction to the mathematics of probability and information

> *"My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.'"*
>
> (Claude E. Shannon)

2.1 Probability and uncertainty

Probability theory is the mathematical tool used to describe uncertainty. In theoretical physics we describe the world we see in terms of simple (or not so simple) models. In the context of modelling people often distinguish between two types of uncertainty:

- **Aleatory uncertainty** (Latin alea=rolling of a dice): This refers to uncertainty which is intrinsic to the phenomenon observed. Even with perfect knowledge of the system this uncertainty would remain.
- **Epistemic uncertainty** (Greek episteme=knowledge): Uncertainty caused by the lack of knowledge. This type of uncertainty could, in principle, be eliminated by better knowledge of the system.

Tobias' view:
As a theoretical physicist I largely believe in a mechanistic world. One could argue that the only aleatory uncertainty known in physics (i.e. uncertainty intrinsic to the natural world) is quantum noise. Quantum mechanics is fundamentally a stochastic theory, experiments testing Bell's inequalities seem to suggest that 'hidden variable' theories are incorrect.

Leaving quantum mechanics aside, all other uncertainty is epistemic in my view – at least we are not aware of any other aleatory type of stochasticity other than quantum noise. Stochasticity we have in non-quantum models is randomness we put into our models to represent the fact that there are dynamics 'deeper down' that we do not know about in detail, or do not understand. It is 'effective randomness'.

2.2 Definition of probability

Definition (Sample space)
The sample space, Ω, of an experiment is a set of outcomes. Each element of Ω corresponds to one possible outcome of the experiment.

Examples

- Roll a dice once: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Roll a dice twice: $\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}$. The sample space has 36 elements.
- Roll two distinguishable dice: $\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}$. The sample space has 36 elements.
- Roll two indistinguishable dice:
  $\Omega = \{(1, 1), (1, 2), \ldots, (1, 6), (2, 2), (2, 3), \ldots, (2, 6), (3, 3), \ldots, (3, 6), \ldots, (6, 6)\}$. The sample space has 21 elements (according to my count).

Definition
An event A is a subset of Ω, $A \subset \Omega$.

Example
Flip a coin twice. $\Omega = \{HH, HT, TH, TT\}$. 'Get at least one H' is an event: $A = \{HH, HT, TH\} \subset S$.

Example
Flip two distinguishable coins. $\Omega = \{HH, HT, TH, TT\}$. If the coin is fair we have

$$p(HH) = p(HT) = p(TH) = p(TT) = \frac{1}{4}. \tag{2.1}$$

Say the event we are interested in is 'at least one H'. Then

$$A = \{HH, HT, TH\}, \tag{2.2}$$

and $P(A) = p(HH) + p(HT) + p(TH) = \frac{3}{4}$.

Remark
We note $P(\emptyset) = 0$ and $P(\Omega) = 1$.

2.3 Set theory and conditional probabilities

The following concepts are useful

- union of two sets, $A \cup B$
- intersection $A \cap B$.

Two events are *mutually exclusive* if $A \cap B = \emptyset$.

Theorem

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.3}$$

Proof
Draw a diagram.

Remark
If two events A and B are mutually exclusive then $P(A \cup B) = P(A) + P(B)$.

Definition (Conditional probability)
The conditional probability $P(A|B)$ ('A given B') is the probability that event A occurs given that we know that B occurs. It is defined as follows

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.4}$$

Remark

$$P(A) = P(A|B)P(B) + P(A|\text{not } B)P(\text{not } B) \tag{2.5}$$

Proof
Insert definition of conditional probabilities, and note that $P(A) = P(A \cap B) + P(A \cap (\text{not } B))$.

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.6}$$

Proof
Easy to prove, but hard to understand. For proof, insert definitions of $P(A|B)$ and $P(B|A)$.

Why is this useful? (Example)
You are presented with three coins, two are fair and one always lands on H. You pick one coin at random. What is the probability that it is the biased coin. Well, easy, that's $1/3$, one out of three.

Next you flip the coin 3 times, and you get HHH. This provides some additional information about the coin. But how does this observation of HHH change your assessment of the coin? Given that additional information, what is the probability that the coin you have picked is the fair one?

Answer
Define the following events: B=you have picked the biased coin; F=you have picked the fair coin; HHH you get three heads in a row.
We have $P(B) = 1/3$, $P(F) = 2/3$, $P(HHH|F) = 1/8$ and $P(HHH|B) = 1$.
What you would like to know is $P(B|HHH)$. Using Bayes

$$P(B|HHH) = \frac{P(HHH|B)P(B)}{P(HHH)}. \tag{2.7}$$

We also have $P(HHH) = P(HHH|B)P(B) + P(HHH|F)P(F)$ (note that F is the same as 'not B', the two are mutually exclusive). So $P(HHH) = 1 \times \frac{1}{3} + \frac{1}{8} \times \frac{2}{3} = \frac{10}{24}$. So we find

$$P(B|HHH) = \frac{1 \times \frac{1}{3}}{\frac{10}{24}} = \frac{4}{5}. \tag{2.8}$$

The probability that you have picked the biased coin, given the observation HHH is 4/5, so the probability that you have picked the fair coin is 1/5.

Remark
This is not the probability that you get HHH given that you have picked the fair coin.

Definition (Independent events)
Two events A and B are called independent if and only if

$$P(A \cap B) = P(A)P(B). \tag{2.9}$$

Remark

- This is not the same as being mutually exclusive ($P(A \cap B) = P(\emptyset) = 0$).
- **Important:** The independence of A and B is equivalent to $P(A|B) = P(A)$. (Knowledge that B has occurred tells you nothing about the probability that A occurs).

  Proof: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ by definition. The latter is equal to $P(A)$ if and only if $P(A \cap B) = P(A)P(B)$.
- Using the previous bullet the independence of A and B is equivalent to $P(B|A) = P(B)$ by symmetry. (You should understand why). This means in particular that $P(A|B) = P(A)$ is equivalent to $P(B|A) = P(B)$. Again you should understand this in detail.

2.4 Random variables and probability density functions

2.4.1 Random variables

Definition (Random variable)
A random variable X is a quantity whose value is determined by the outcome of a probabilistic experiment. Mathematically speaking X is a map from Ω to the real numbers $\mathbb{R}$:

$$X : \Omega \to \mathbb{R}. \tag{2.10}$$

Examples

1. A coin is flipped three times. $\Omega = \{HHH, HHT, \ldots, TTT\}$. Say X is the number of H's obtained in this experiment. So X takes values $0, 1, 2$ or 3:

$$\begin{aligned} X(HHH) &= 3, \\ X(HHT) = X(HTH) = X(THH) &= 2, \\ X(TTH) = X(THT) = X(HTT) &= 1, \\ X(TTT) &= 0. \end{aligned} \tag{2.11}$$

   The set $\{X = 1\}$ is an event, we have $\{X = 1\} = \{TTH, THT, HTT\}$. Similarly, $\{X = 0\}$, $\{X = 2\}$ and $\{X = 3\}$ are events.

2. Roll a die once. Define X as follows

$$X = \begin{cases} 1 & \text{outcome is even,} \\ 0 & \text{outcome is odd.} \end{cases} \tag{2.12}$$

   We have $X(1) = X(3) = X(5) = 0$ and $X(2) = X(4) = X(6) = 1$. If the die is fair, $P(X = 0) = P(X = 1) = 1/2$.

2.4.2 Discrete random variables

Basic properties

If X can take discrete (countably many[1]) values $x_1, x_2, \ldots$ then we write

$$p_i \equiv P(X = x_i) \tag{2.13}$$

for the probability of the event $\{X = x_i\}$. The quantity $P(x_i)$ is the probability that X takes the value x_i if the probabilistic experiment is carried out.

We have

[1] This can be an infinite number of values. Countable means that they can be enumerated, but not that the set of possible values X can take is finite.

$$\begin{aligned} p_i &\geq 0 \ \forall i, \\ \sum_i p_i &= 1. \end{aligned} \tag{2.14}$$

Statistical physics context

Suppose we have a thermodynamic system which can be in one of Ω microstates at any one time. We label these by $i = 1, \ldots, \Omega$. In equilibrium the system will 'hop' from one microstate to the next on a very fast timescale. Say p_i is the probability to find the system in state i at any one time. Imagine for example, we took a photograph of the system (on a microscopic scale) at a random time, the quantity p_i tells us how likely it is that this photograph shows the system in microstate i.
We can then imagine physical properties associated with each of the microstates. These are typically macroscopic quantities, such as an energy, or a magnetisation. If microstate i is associated with an energy ε_i for example, then the mean energy of the system is

$$\langle \varepsilon \rangle = \sum_{i=1}^{\Omega} p_i \varepsilon_i. \tag{2.15}$$

More generally, if X is a function that assigns an observable value to each microstate i, and if the value of X in state i is written as x_i, then the expected value of x (also known as expectation value, average value or mean value) is defined as follows:

Expected value of x:

$$\langle x \rangle = \sum_i p_i x_i. \tag{2.16}$$

The normalisation condition, $\sum_i p_i = 1$ can be written as $\langle 1 \rangle = \sum_i p_i \times 1 = 1$.
We can also ask how much the value of X fluctuates from one micostate to the next. To this end we define the variance of X:

Variance of x:

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 = \sum_i p_i x_i^2 - \left(\sum_i p_i x_i \right)^2. \tag{2.17}$$

Exercise:
It should be clear to you

- that σ^2 can also be written as $\sigma^2 = \sum_i p_i (x_i - \overline{x})^2$, if we abbreviate $\overline{x} = \langle x \rangle$,
- why a vanishing variance, $\sigma^2 = 0$, means that $x_i = \langle x \rangle$ for all i, i.e. all microstates have the same associated value x_i, independent of i (and that value is then trivially the average $\langle x \rangle$).

These statements are important. You can (and should) derive these from the above properties and definitions.

Common discrete probability distributions over discrete states

- Geometric distribution, $p_i = (1-\lambda)\lambda^i$, where $0 < \lambda < 1$ is a parameter and $i = 0, 1, 2, \ldots,$
- Poisson distribution $p_i = e^{-\lambda} \frac{\lambda^i}{i!}$, where $\lambda > 0$ is a parameter and $i = 0, 1, 2, \ldots,$
- Binomial distribution, $p_i = \begin{pmatrix} m \\ i \end{pmatrix} \lambda^i (1-\lambda)^{m-i}$, where the integer m, and $\lambda > 0$ are parameters, and where $i = 0, 1, \ldots, m$.

Exercise:
Check the normalisation property for each of these, and compute $\langle i \rangle = \sum_i p_i i$.

2.4.3 Continuous random variables

Now consider the case in which X can take any real value[2].
For such a random variable it makes no sense to ask 'What is the probability that X takes a certain value x?'. This is a subtle point, but it becomes fairly obvious if you think about it. Make sure you understand this.

The meaningful question is here to ask 'What is the probability that X takes a value in the interval from a to b?'. We write $P(a \leq X \leq b)$ for this probability.

One introduces the probability density

$$p_X(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}. \tag{2.18}$$

(Note that this is a lower-case p.)

In other words, for small dx,

$$P(x \leq X \leq x + dx) = p(x) dx. \tag{2.19}$$

[2]The real numbers cannot be enumerated. The set $\mathbb{R}$ is not countable.

For finite intervals from a to b we have

$$P(a \leq X \leq b) = \int_a^b dx\ p(x). \tag{2.20}$$

Properties of $p(\cdot)$

- $p(x) \geq 0$ for all x.
- $\int_{-\infty}^{\infty} dx\ p(x) = 1$ (normalisation).

Statistical physics context:
Consider a single particle in a box, with position $\underline{r}$ and momentum $\underline{p}$. Now imagine that this (classical) particles bounces around in a container, a small box maybe, at high speed. If you take a picture at a random moment in time, you will essentially find the particle at a random position, and with momentum pointing in a random direction. How do we describe the statistics of such a system?

We now need to introduce a probability density in phase space. The phase space of this particle is 6-dimensional, and we write $\rho(\underline{r}, \underline{p})$ for the probability density in phase space. That is, if $d^3r d^3p$ is an infinitesimal 6-dimensional volume element in phase space, located at the point $(\underline{r}, \underline{p})$, then the probability to find the particle in that volume element in phase space is

$$\rho(\underline{r}, \underline{p}) d^3r d^3p. \tag{2.21}$$

So $\rho(\underline{r}, \underline{p})$ is a probability density per unit volume in phase space. It has the following properties

- $\rho(\underline{r}, \underline{p}) \geq 0$ (the density can never be negative any any point)
- $\int d^3r d^3p\ \rho(\underline{r}, \underline{p}) = 1$ (normalisation of the probability distribution).

We have written $\int d^3r d^3p \ldots$ for the 6-dimensional integral $\int dx \int dy \int dz \int dp_x \int dp_y \int dp_z \ldots$, where $\underline{r} = (x, y, z)$ and $\underline{p} = (p_x, p_y, p_z)$.

If $f(\underline{r}, \underline{p})$ is an observable for this single particle, then the definition of the expected value of f (or average value of f, or 'the mean of f') and variance are now as follows"

Expected value of f:

$$\langle f \rangle = \int d^3r d^3p\ \rho(\underline{r}, \underline{p}) f(\underline{r}, \underline{p}), \tag{2.22}$$

Variance of f:

$$\sigma^2 = \langle f^2 \rangle - \langle f \rangle^2 = \int d^3r d^3p\ \rho(\underline{r}, \underline{p}) \left(f(\underline{r}, \underline{p}) - \overline{f} \right)^2, \tag{2.23}$$

where we have abbreviated $\overline{f} = \langle f \rangle$.

Exercise:
If the second equality in Eq. (2.23) is not completely obvious to you, then please make sure you derive it.

For systems with N particles, phase space is $6N$-dimensional – we have three position coordinates and three momentum coordinates for each particle. Phase space functions are then of the form $f(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_N, \underline{p}_1, \underline{p}_2, \ldots, \underline{p}_N)$, and averages are defined as

$$\langle f \rangle = \int d^{3N}r d^{3N}p \; \rho(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_N, \underline{p}_1, \underline{p}_2, \ldots, \underline{p}_N) f(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_N, \underline{p}_1, \underline{p}_2, \ldots, \underline{p}_N). \quad (2.24)$$

Common continuous probability distributions

- the exponential distribution, $\rho(x) = \lambda e^{-\lambda x}$, where $\lambda > 0$ is a parameter, and where $x \geq 0$,
- the Gaussian distribution, $\rho(x) = (2\pi\sigma^2)^{-1/2} \exp\Big(-(x-\mu)^2/(2\sigma^2) \Big)$, where σ^2 and μ are parameters, and where $-\infty < x < \infty$.

Exercise:
Check the normalisation property for the exponential distribution, and compute $\langle x \rangle$.

2.5 Law of large numbers and the Central Limit Theorem

2.5.1 Independent identically distributed (iid) random variables

We now ask what the properties are of the average of measurements from multiple replications of the same experiment. More precisely, we consider N random variables, $X_1, X_2, \ldots, X_N$ all drawn independently from the same underlying distribution.

Definition
We say that $X_1, \ldots, X_N$ are *i.i.d.* (independent identically distributed) random variables if they are (pairwise) independent and if they all have the same underlying probability density.

We are interested in the properties of the average constructed from these measurements: $Z_N = \frac{1}{N} \sum_{i=1}^{N} X_i$.

Remark
Keep in mind that Z_N is a random variable itself, we are not talking about a specific realisation $\frac{1}{N} \sum_i x_i$ of observed values.

2.5.2 Law of large numbers

Law of large numbers (LLN)
Assume, $X_1, \ldots, X_N$ are N i.i.d random variables, with (finite) first moment $\langle X_i \rangle = \mu$. Their average

$$Z_N = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2.25}$$

then converges to a constant random variable as $N \to \infty$, which always takes the value μ, i.e.

$$Z_N \overset{N\to\infty}{\longrightarrow} \mu. \tag{2.26}$$

Remark
I am being very sloppy here, as I am not really telling you what exactly the notation

$$\text{“} \overset{N\to\infty}{\longrightarrow} \text{”} \tag{2.27}$$

actually means. Mathematicians have very precise definitions for this.

2.5.3 Central limit theorem

The law of large number says that the average of a large number of independent samples of a given distribution will be close to the mean of the distribution. In the limit of an infinite number of independent samples, the average of the samples will be exactly the mean of the distribution. This is intuitively what one would expect.

But how close will the average be, if we only draw a finite number of samples? The answer to this question is provided by the central limit theorem (CLT).

Central Limit Theorem
As before, suppose $X_1, \ldots, X_N$ are independent identically distributed random variables. We also assume that they have a finite first moment, μ, and a finite variance, σ^2. For large N, the random variable

$$Z_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu), \tag{2.28}$$

then follows a Gaussian distribution

$$p_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}, \tag{2.29}$$

of mean μ and with variance σ^2.

Corollary of Central Limit Theorem

As before, suppose $X_1, \ldots, X_N$ are independent identically distributed random variables. We also assume that they have a finite first moment, μ, and a finite variance, σ^2. The random variable

$$Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{2.30}$$

then follows a Gaussian distribution

$$p_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu)^2}{2\sigma_Y^2}}, \tag{2.31}$$

of mean μ and with variance $\sigma_Y^2 = \sigma^2/N$ for large N.

I do not include proofs of the LLN and CLT. To prove these one needs to know about characteristic functions (Fourier transforms) of probability distributions. Perhaps you'll do this 4th-year advanced statistical physics.

Exercise:

Derive the corollary from the CLT using the result of Qu. 1 on Example Sheet 2.

What does the CLT/corollary mean?

Broadly speaking the corollary to the CLT states the following:

Suppose you have an experiment, which generates a numerical outcome (a 'measurement'). Assume that the experiment can be repeated again and again independently. Alternatively you can think of a random number generator which generates a sequence of as many random numbers as you want, all from the same distribution, but each one independent of the previous random numbers. Say this underlying distribution (often unknown) has mean μ and variance σ^2.

1. You now carry out a batch of a large number N independent measurements (or let the random number generator produce N samples), and compute the numerical average of these N numbers. This is one realisation of the above random variable Z. It is a number.
2. Write this number down somewhere on a piece of paper.
3. Then carry out another batch of N repetitions of the experiment, each one independent of what you have done before. This gives another N random numbers ('readings'), and you average them. This average is another number (another realisation of Z).
4. Write this number down on the same piece of paper as before.
5. Then carry out another batch of N repetitions, compute the average, etc etc. I.e. repeat steps 3 and 4.
6. Do this many many times, i.e. do many many batches of N repetitions each. For each one compute the average and write it down. The piece of paper will now have 'many

many' numbers on it, each one obtained from a batch of N independent repetitions of the experiment.

7. Once you have collected those 'many many' numbers on your sheet of paper, plot a histogram of these numbers.

The CLT tells you that this histogram will (roughly) be a Gaussian curve with mean μ and variance σ^2/N where N is the size of each batch. Note that the variance gets smaller for larger batch sizes.

2.5.4 A few comments on statistical data analysis*

[For exam purposes, you can safely ignore this section.]

In my first year I was told: if you do an experiment (of whatever type) N times, and the have N readings of some quantity x, say $x_1, \ldots, x_N$, then do the following:

1. Compute the sample mean $\frac{1}{N}\sum_{i=1}^{N} x_i$. Call this $\overline{x}$.
2. Compute the sample standard deviation, $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$
3. Then quote the 'standard error of the mean', $s/\sqrt{N}$.

I understand the first part, computing the average like this makes sense. But do you *really* understand what items 2. and 3. are about?

Let's put this into more specific questions:

- In 2. above, why the strange denominator of $N-1$ if there are N measurements? Why not N? And if we have to subtract something then why 1, why not 17, 23 or 42?
- In 3. above, what exactly is the physical meaning of s and of $s/\sqrt{N}$? Why divide by $\sqrt{N}$, and when to use what, s or $s/\sqrt{N}$?

How to phrase this mathematically?

In the context of this problem we always assume that there is an underlying 'true' value of the quantity x, and that the measurements come with some sort of uncertainty, maybe because your instruments isn't fully precise or something. Mathematically this means that the N observations can be described as N random variables, $X_1, \ldots, X_N$, all independent of each other, and each with mean μ. We assume that there is no systematic bias in our measurements, so 'on average' the experimental observation agrees with the 'true' value of x. This means that μ, the average of the X_i is that 'true' value of the quantity we are measuring. The X_i have a variance σ^2. This variance comes from the uncertainty of your instrument for example.

The issue is though that μ (the true value of x) and the uncertainty σ^2 of your instrument will generally not be known to you. All you have is N repeat measurements of x, and the assumption that these are not biased in any way and that they are independent. So what do you do?

The $N-1$ issue

Assume for the moment that you knew μ (I know it isn't, but let's assume this for a moment), and that you want to estimate σ^2 (the unknown uncertainty of your instrument) from the data you have. That data consist of the N values $x_1, \ldots, x_N$ you have measured. This is all you have. If μ is known to you, then it is not unreasonable to make the guess

$$\sigma^2_{\text{guess}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2.32}$$

The quantity σ^2 as just defined is a realisation of a random variable, it depends on the readings you have obtained, and these are stochastic. If you take N readings, you can compute one value of σ_{guess}. If you then do another N measurements you get another value for σ_{guess} (lower case). They are both realisations of Σ_{guess} (upper case).
If we now compute the average of σ^2_{guess}, we get

$$\begin{aligned}\langle\sigma^2_{\text{guess}}\rangle &= \frac{1}{N}\langle(X_i-\mu)^2\rangle \\ &= \frac{1}{N}\sum_i \sigma^2 \\ &= \sigma^2.\end{aligned} \tag{2.33}$$

So this is fine. On average σ^2_{guess} will produce the correct value for the uncertainty of the instrument.
But, there is a catch. We have assumed that we know μ, the true value of the quantity we are measuring. But we don't. All we have are the $x_1, \ldots, x_N$. So let us replace μ in the above expression for σ^2_{guess} by our best estimate for μ, i.e. by $\frac{1}{N}\sum_j x_j$. Then we get an updated version

$$\sigma'^2_{\text{guess}} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_j x_j\right)^2 \tag{2.34}$$

Now what is the average of this? Well, let's see:

$$\begin{aligned}\langle\sigma'^2_{\text{guess}}\rangle &= \frac{1}{N}\sum_{i=1}^{N}\left\langle\left(x_i - \frac{1}{N}\sum_j x_j\right)^2\right\rangle \\ &= \frac{1}{N}\sum_i\langle x_i^2\rangle - \frac{2}{N^2}\sum_i\sum_j\langle x_i x_j\rangle + \frac{1}{N^2}\sum_{jk}\langle x_j x_k\rangle \\ &= \frac{1}{N}\sum_i\langle x_i^2\rangle - \frac{1}{N^2}\sum_i\sum_j\langle x_i x_j\rangle\end{aligned} \tag{2.35}$$

Now, how can we simply this further. Well, $\langle x_i^2\rangle - \mu^2 = \sigma^2$, so $\langle x_i^2\rangle = \sigma^2 + \mu^2$. If $i \neq j$ then we have $\langle x_i x_j\rangle = \langle x_i\rangle\langle x_j\rangle$ because of independence, and the latter is equal to μ^2. So $\langle x_i x_j\rangle = \mu^2$ if $i \neq j$. Keep in mind that we do not actually know μ and σ^2, we are computing averages

here, so mathematically it is all ok. Let's carry on:

$$\begin{aligned}\left\langle \sigma'^2_{\text{guess}} \right\rangle &= \frac{1}{N}\sum_i \left\langle x_i^2 \right\rangle - \frac{1}{N^2}\sum_i \sum_j \left\langle x_i x_j \right\rangle \\ &= \frac{1}{N} \times N \times (\mu^2+\sigma^2) - \frac{1}{N^2}\sum_i \left\langle x_i^2 \right\rangle - \frac{1}{N^2}\sum_{i\neq j} \left\langle x_i x_j \right\rangle . \end{aligned} \tag{2.36}$$

In the second term, $\sum_i \langle x_i^2 \rangle = N(\mu^2+\sigma^2)$, and in the third term $\sum_{i\neq j} \langle x_i x_j \rangle = N(N-1)\mu^2$ (the sum is over $N(N-1)$ combinations of i and j, and each term gives μ^2). So,

$$\begin{aligned}\left\langle \sigma'^2_{\text{guess}} \right\rangle &= (\mu^2+\sigma^2) - \frac{1}{N}(\mu^2+\sigma^2) - \frac{N(N-1)}{N^2}\mu^2 \\ &= (\mu^2+\sigma^2) - \frac{1}{N}(\mu^2+\sigma^2) - \left(1-\frac{1}{N}\right)\mu^2 \\ &= \left(1-\frac{1}{N}\right)\sigma^2 \\ &= \frac{N-1}{N}\sigma^2. \end{aligned} \tag{2.37}$$

So our estimator σ'^2_{guess} isn't quite right, we are getting things wrong by a factor of $(N-1)/N$. If we correct for this though, we have our answer. So we define $\sigma''^2_{\text{guess}} = \frac{N}{N-1}\sigma'^2_{\text{guess}}$, i.e.

$$\sigma''^2_{\text{guess}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_j x_j\right)^2 . \tag{2.38}$$

The above calculation then shows that

$$\left\langle \sigma''^2_{\text{guess}} \right\rangle = \sigma^2, \tag{2.39}$$

we we finally got it right. The pre-factor of $1/(N-1)$ is crucial, otherwise the estimator for σ^2 is biased and does not produce the correct experimental uncertainty on average.

This is kind of a long calculation, is perhaps not suitable for first-year data analysis. But you are in 2nd year now.

The use of s versus $s/\sqrt{N}$

After all this work it is now relatively easy to answer the second question. First of all we notice that $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2}$ (with $\overline{x} = N^{-1}\sum_{i=1}^{N} x_i$) is nothing else that σ''_{guess} as defined above.

So it is clear what s is. Given the N measurements, $x_1, \ldots, x_N$ it is a faithful (unbiased) estimator of the true standard deviation, σ, of the distribution from which the x_i are drawn. So, s is your best estimate (given the N observations $x_1, \ldots, x_N$) of the uncertainty associated with your instrument.

But what is the physical interpretation of $s/\sqrt{N}$? This is the uncertainty (standard deviation) of what?

Well, again we have already established this. If you follow the steps 1.-7. outlined in Sec. 2.5.3 you generate a histogram with variance σ^2/N, i.e. standard deviation $\sigma/\sqrt{N}$. Recall that this histogram is produced by obtaining a batch of N independent measurements, computing the average of these N numbers, and then making a histogram from many of such batches.

The quantity $s/\sqrt{N}$ (obtained from a single batch) is our best estimator of $\sigma/\sqrt{N}$. So $s/\sqrt{N}$ is in fact the best answer we can give, based on the measurements $x_1, \ldots, x_N$ to the following question:

"Assume I make N independent measurements of the quantity x, and compute the numerical average of these N numbers. How much will this average fluctuate if I do this many times? What is its standard deviation?"

Remark 1
Assume you have made N measurements, $x_1, \ldots, x_N$. You then ask the question: by how much will the average of M measurements fluctuate? Well, you can answer that easily, using the CLT. Your best estimator for σ, the standard deviation of a *single* measurement is given by

$$\sigma''_{\text{guess}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_j x_j\right)^2}, \tag{2.40}$$

as we have discussed.
A an average from a sample of M measurements (where M can be different from N) will then fluctuate with a standard deviation of $\sigma/\sqrt{M}$, and your best guess, based on $x_1, \ldots, x_N$ is

$$\frac{\sigma''_{\text{guess}}}{\sqrt{M}} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_j x_j\right)^2}}{\sqrt{M}}. \tag{2.41}$$

Remark 2
In the limit $M \to \infty$ this goes to zero. The average obtained from an infinite sample does not fluctuate, it is precisely μ. This reproduces the LLN.

Remark 3
You may now think that the LLN is just a special case of the CLT. Well, this is kind of true. But keep in mind that the CLT assumes a finite mean and finite variance of the X_i. For the LLN you only need finite means, but the variance need not exist.

If you have actually read this section, please email me a picture of a dinosaur – TG.

2.6 Shannon entropy: Quantifying uncertainty

Suppose we look at the outcomes of a recurring event, say your favourite football team plays once a week and you do not have the chance to see the matches yourself. After each match you learn the outcome. For simplicity we focus only on goal difference, i.e. a result of +3

would mean that your team won by three goals, 0 means a draw, and -2 means that your team lost by two goals.

On a particular match day you learn the result of the latest game. How surprised you will be will depend on the result. Say your team is one of the top teams and usually wins. You'd then probably be very surprised if you found an event of type -3, but you'd be less surprised if the outcome is $+2$. The -3 outcome is very unusual, and unlikely events are a source of surprise.

How can we quantify the amount of surprise mathematically? We run a probabilistic experiment, which can have different results, say $R_1, R_2, R_3, \ldots$. Result R_i occurs with probability p_i, and we assume that you know these probabilities in advance. You now run one instance of the experiment, and find a particular result, say R_6. How much surprise should you assign to this? We'd like to define a function $S(p)$ quantifying the amount of surprise resulting from the occurrence of an event, of which you know that it comes up with probability p.

What properties do we want this function $S(p)$ to have?

We want this to be positive or at least not negative, $S(p) \geq 0$.

Secondly if the outcome of the experiment is absolutely certain ($p = 1$) then there is no element of suprise at all. If your team *always* wins $3:0$, then there is no point even looking at the results. So we want $S(p = 1) = 0$.

Next assume now we do two *independent* iterations of the experiment, and you know that they are independent. We get results R_{11} and R_4. The probability for this to happen is $p_{11}p_4$, due to the independence. How surprised should you be to observe R_{11} followed by R_4? Well, if the two events are independent, you'd say that the total surprise ought to be $S(R_{11}) + S(R_4)$. The fact that you find R_{11} in the first instance of the experiment does not change the information you get from the R_{11}-outcome of the second experiment. So we want the function S to have the property $S(pq) = S(p) + S(q)$.

Let's summarise:

Properties of the 'Surprise' function $S(p)$:

1. $S(p) \geq 0$ for all $0 < p \leq 1$;
2. $S(1) = 0$;
3. $S(pq) = S(p) + S(q)$ for $0 < p, q \leq 1$
4. We also want S to be continuous and differentiable.

We have excluded the values $p = 0$ and $q = 0$. That's not a problem, if a particular event never occurs, then we do not need to assign an amount of surprise.

Claim: The only functions with these properties are of the form $S(p) = c \ln p$, with $c < 0$ constant.

Proof:
We want $S(pq) = S(p) + S(q)$. Differentiating both sides with respect to p and using the chain rule on the left-hand side we find

$$qS'(pq) = S'(p), \tag{2.42}$$

where we have written S' for the derivative of S. We now differentiate both sides of Eq. (2.42) again, now with respect to q. We have

$$S'(pq) + pqS''(pq) = 0. \tag{2.43}$$

Replacing pq by u we have found the differential equation

$$u\frac{d^2}{du^2}S(u) + \frac{d}{du}S(u) = 0, \tag{2.44}$$

which can be written as $\frac{d}{du}\left[u\frac{d}{du}S(u)\right] = 0$. This means $u\frac{d}{du}S(u) = \text{const.}$, i.e.

$$\frac{d}{du}S(u) = \frac{c}{u}, \tag{2.45}$$

so

$$S(u) = c \ln u. \tag{2.46}$$

Now, let's check the other properties. We want S to be non-negative, this means that the constant c must be negative. Further, we want $S(u = 1) = 0$. That's always fine, no matter what value we choose for c. Finally $S(pq) = S(p) + S(q)$ is ok for $S(p) = c \ln(p)$. Finally, S is continuous and differentiable, so we are done, end of proof.

So far we have not made a choice for the constant c. One common choice is to choose c such that

$$S(p) = -\log_2 p. \tag{2.47}$$

This can be achieved by setting $c = -1/\ln(2)$. The reason for this choice will become clear below.

Average surprise:
Returning to our probabilistic experiment above with outcomes $R_1, R_2, \ldots$ happening with probabilities $p_1, p_2, \ldots$ we can also introduce the average surprise

$$S = -\sum_i p_i \log_2 p_i \tag{2.48}$$

We now have everything we need to introduce Shannon entropy.

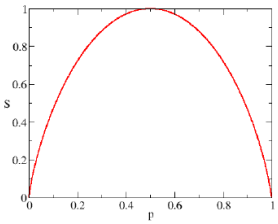


Figure 2.1: Graph of the function $S = -p \log_2 p - (1-p) \log_2 (1-p)$.

Definition of Shannon entropy:
The Shannon entropy of a probability distribution p_i is defined as

$$S = -\sum_i p_i \log_2 p_i. \tag{2.49}$$

For continuous distributions with density $p(x)$ one defines

$$S = -\int dx\, p(x) \ln p(x). \tag{2.50}$$

These expressions do not look particularly transparent, but we will give a more intuitive interpretation in the next section. Before we do this, we briefly look at a very simple example, a Bernoulli experiment with only two possible outcomes. The first occurs with probability p, and the second with probability $1-p$. The 'suprise' associated with the first outcome is then $-\log_2 p$ and that of the other event is $\log_2 (1-p)$. The Shannon entropy is the average surprise

$$S = -p \log_2 p - (1-p) \log_2 (1-p). \tag{2.51}$$

This is a relatively simple function of p, shown in Fig. 2.1. As can be seen from the figure the entropy is zero if either $p = 0$ or $p = 1$. This makes sense, in either of these two cases the outcome of the experiment is deterministic, there is the nothing random in the experiment and so no surprise is produced. Shannon entropy is maximal for $p = 1/2$, we then have $S = 1$. What this value means will become clear in the next section. It is natural though that the value $p = 1/2$ maximises Shannon entropy for the Bernoulli experiment. For $p = 1/2$ both outcomes are equally likely, and hence the uncertainty is biggest.

2.7 Shannon entropy and 'information'

2.7.1 Basic example

Suppose you have a probabilistic experiment with four different outcomes, we label them A, B, C and D. They come up with the following probabilities:

$$A : p_A = \frac{1}{2}, \qquad B : p_B = \frac{1}{4},$$
$$C : p_C = \frac{1}{8}, \qquad D : p_D = \frac{1}{8}. \tag{2.52}$$

Suppose now a friend of yours runs the experiment once. They will get result A, B, C or D, so they have 'information' about the outcome. You on the other hand, only know the general setup, i.e., the fact that there are four possible outcomes, and the probabilities with which they occur. But you do not know what exactly the outcome was this time. So your friend has more information than you do, because they know what specific result came up. How much information does your friend have?

This can be quantified by asking how many 'yes/no' questions you'd have to ask your friend to find out whether the outcome was A, B, C or D. (We assume that your friend gives truthful answers). By 'yes/no' questions we mean questions which have a binary outcome. You can always achieve this with two such questions. The first could be 'Was the outcome either A or B?'. If your friend says 'yes', your next question is: 'OK, was it A?'. Then you know. If your friend answers 'no' to the first question, you know the result must either be C or D, and so your second question is for example 'Was it C?'. Either way, with two questions per iteration of the experiment you can guarantee to know the outcome each time the experiment is run.

However, this is not the most efficient way of doing this. Result A is the most likely one, and importantly you know that this is so (remember that we assume that you are told the values of p_A, p_B, p_C and p_D before the experiment is run). So by asking 'Was the outcome either A or B?' you are 'wasting' resources, you are treating A and B equally, when A is more likely than B.

It is more efficient to do this as follows:

1. First question: 'Is it an A?'. If 'yes', you know the result for sure. End. Only one question needed.
2. If no, second question: 'Is it a B?', If 'yes', you know the result. Two questions were needed.
3. If no again, third question: 'Is it a C?'. The answer will either be yes or no, and either way you know the outcome of the experiment (if 'yes' it is a C, if 'no' it is a D). Three questions were needed.

Now, at first it may seem that this is worse than the two-question algorithm above. If the outcome is a C or a D we need three questions after all. But these outcomes are much less likely than the others. On average you have save questions, compared to the more inflexible 'always-ask-two-questions' algorithm. More precisely, with the second algorithm we only need

$$\frac{1}{2} \times \text{one question} + \frac{1}{4} \times \text{two questions} + \frac{1}{4} \times 3 \text{ questions} = 1.75 \text{ questions} \tag{2.53}$$

on average. I.e., if you use the more flexible sequence of questions you'll be able to save questions. You may need three questions for some of the iterations of the experiment, but if a large number of repeats is run, you'll be more efficient in the long run.

Now, let's compute the Shannon entropy of the distribution

$$\begin{aligned} S &= -p_A \log_2 p_A - p_B \log_2 p_B - p_C \log_2 p_C - p_D \log_2 p_D \\ &= \frac{1}{2} \times 1 + \frac{1}{2} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 \\ &= 1.75 \end{aligned} \tag{2.54}$$

So the Shannon entropy – at least for this example – is the average number of binary questions you need to ask to find out what the outcome of a probabilistic experiment is. More precisely, the values of $-\log p_i$ ($i = A, B, C, D$) corresponds to the number of questions you need to ask in each of the cases above, and this is then averaged over the different outcomes.

Admittedly, this is a very simple example, and the p_i above were chosen to keep things easy. They are all of the form $(1/2)^n$, so that the logs return integer values. We give a slightly more complicated example below, but we cannot show here in all generality that Shannon entropy has this interpretation. This in fact the statement of 'Shannon's source coding theorem', and we'd have to go deeper into the mathematics of information theory to actually prove this.

2.7.2 Shannon entropy and coding

We can briefly illustrate how Shannon entropy relates to coding theory. Assume we run the above experiment many many times, this generates an output string of the form $'AABACD...'$, and say we want to transmit this string along a communication line. Digital communication means sending bits, i.e. '0' or '1'. What is the most efficient way of doing this, what is the minimum number of bits (0 or 1s) that we need *on average* per iteration of the experiment? It is clear that sending bits is essentially the same as asking yes/no questions, so we expect that we'd need 1.75 bits per iteration of the experiment on average. Here is how one could do this:

outcome A:	send codeword '0'
outcome B:	send codeword '10'
outcome C:	send codeword '110'
outcome D:	send codeword '111'

Taking into account the probabilities with which each of the four cases comes up, the average codeword length is 1.75. At the same time these code words can uniquely be converted back into a string of A, B, C and D. Start reading the string from the beginning. Keep going until you either hit a 0 or have read three bits. By that point you have read in one of the four code words above, and have reached the end of one symbol (A, B, C or D). For example $'00100110111'$ corresponds to $AABACD$.