Information Theory

a short course by a physicist

Gregory Falkovich

September 29, 2023

In memory of Leo Szilard and John Wheeler

Contents

1 Thermodynamics and statistical physics 7

1.1 Basics of thermodynamics 7

1.2 Thermodynamic potentials 14

1.3 Microcanonical distribution 16

1.4 Canonical distribution and fluctuations 19

1.5 Evolution in the phase space 21

2 Inevitability of irreversibility 23

2.1 Kinetic equation and H-theorem 24

2.2 Phase-space mixing and entropy growth 28

2.3 Entropy decrease and non-equilibrium fractal measures 36

3 Basics of Information Theory 40

3.1 Information as a choice 41

3.2 Communication Theory 46

3.3 Correlations in the signals 50

3.4 Mutual information as a universal tool 53

3.5 Channel capacity 57

3.6 Continuous case and Gaussian Channel 60

3.7 Hypothesis testing and Bayes' formula 62

3.8 Relative Entropy 66

4 Applications of Information Theory 72
4.1 The whole truth and nothing but the truth 72
4.2 Exorcizing Maxwell demon 76
4.3 Renormalization group and the art of forgetting 80
4.4 Information is life 88
4.5 Theory of mind 93
4.6 Rate Distortion and Information Bottleneck 97
4.7 Information is money 102

5 New Second Law of Thermodynamics 108
5.1 Stochastic Web surfing and Google's PageRank 108
5.2 Random walk and diffusion 111
5.3 Fluctuation relation and the new second law 113

6 Quantum information 117
6.1 Quantum mechanics and entropy 118
6.2 Quantum statistics and density matrix 121
6.3 Entanglement entropy 123
6.4 Quantum communications 127
6.5 Conditional entropy and teleportation 131
6.6 The way out is via a black hole 133

7 Conclusion 136
7.1 Take-home lessons 136
7.2 Epilogue 140

8 Appendix 143
8.1 Formal structure of thermodynamics 143
8.2 Central limit theorem and large deviations 144
8.3 Baker map 147
8.4 On Zipf law 151
8.5 Ising model of the brain 152
8.6 Unsupervised learning and infomax principle 153
8.7 Information loss by multi-dimensional RG 155
8.8 Brownian motion 156
8.9 Fluctuation relations in a multi-dimensional case 160
8.10 Quantum fluctuations and thermal noise 162
8.11 Quantum thermalization 164

Preface

This book answers the question: *How much can we say and do about something we do not know?* Of course, the art of bluffing without blushing was perfected by people in many trades and walks of life. This particular text grew out of a one-semester course, intended as a parting gift to those leaving physics for greener pastures and wondering what is worth taking with them. Statistically, most of the former physicists use statistics, because this discipline was first to develop quantitative tools to answer the above question. Yet when the course was taught in different institutions and countries, it attracted a motley mix of students, postdocs and faculty from physics, mathematics, engineering, computer science, economics and biology. Eventually, it evolved into a meeting place where we learn from each other using the universal language of information theory.

The simplest way to answer the above question is a phenomenology traditionally called thermodynamics. It deals only with visible manifestations of the hidden, using general principles (like symmetries and conservation laws) to restrict possible outcomes. The focus is on mean values, fluctuations are ignored. More sophisticated approach derives the statistical laws by explicitly averaging over the hidden degrees of freedom. Those laws justify thermodynamics and describe the probabilities of fluctuations. Two basic notions of this approach - entropy and free energy - turn out to be among the few most important conceptual and technical tools of the modern science and technology.

The book is an introduction, that is a prior knowledge of neither thermodynamics and statistics nor information theory is assumed. The first Chapter gives in a minimalist way the basics of thermodynamics and statistical physics and describes their double focus on what we have (energy) and what we don't (knowledge). When ignorance exceeds knowledge, the right strategy is to measure ignorance. Entropy does that. In the second Chapter, we learn how irreversible entropy change appears from reversible flows in phase space via partial description and dynamical chaos. We understand that *entropy is not a property of a system, but of our knowledge of the system.* It is then natural to start using in the third Chapter the language of the information theory revealing the universality of the approach. When viewed from the perspective of the information theory, the essence of statistics is essentially common sense, which could be compressed to the maxim: "the whole truth and nothing but the truth". That means that we must use all the available data and maximize missing information, that is look for the entropy maximum conditional on the data. This approach is valid not only for thermal equilibrium (where data are on the conserved quantities) but for any state. Mathematically, the approach is based to a large extend on the simple trick of adding many random numbers. Building on that basis, one develops several versatile instruments, like mutual information and its quantum sibling, entanglement entropy, which are widely applied to subjects ranging from bacteria and neurons to markets and quantum computers. The fourth Chapter describes several applications, elucidating different aspects of the approach and directions of its development. We also discuss the so far most sophisticated way to forget information - renormalization group. Forgetting is a fascinating activity — one learns truly fundamental things this way. In the fifth Chapter, we exploit a tireless random walker to rank webpages and obtain a more powerful form of the second law of thermodynamics. The last Chapter is a brief introduction into quantum information focused on the new uncertainty brought by the quantum nature of our world.

Even though it is a graduate text, which presents some advanced subjects in a relatively compact form, we use only elementary mathematical tools, but from all three fields — geometry, algebra and analysis — which correspond respectively to studying space, time and continuum in the physical world. We employ two complementary ways of thinking: continuous flows and discrete combinatorics (thus involving both brain hemispheres). Together, they will equip the reader with a powerful and universal tool, applied everywhere, from computer science and machine learning to biophysics and economics. The book is panoramic, trying to combine into a reasonably coherent whole the subjects that are taught in much details in different departments: thermodynamics and statistical mechanics (as taught in physics and engineering), dynamical chaos (as taught in physics and applied mathematics), information and communication theories (as taught in computer science and engineering). My desire is to reveal an essential unity between different fields and disciplines. In addition, I felt compelled to tell the story worth telling: how we discover the limits imposed by uncertainty on engines, communications, computations and perception. The protagonist of the story is the notion of entropy/information, which was born in the industrial revolution, matured during the digital revolution and leads the present revolution, which blurs the boundaries between physical, digital, and biological worlds.

At the end, recognizing the informational nature of physics and breaking the barriers of specialization is also of value for physicists. People working on quantum computers and the entropy of black holes use the same tools as those designing self-driving cars and market strategies, studying molecular biology, animal behavior and human languages, figuring out how the brain works and trying to quantify conscience. Many go out and apply the tools of physics to new domains. Few can come back enriched by the knowledge how the tools work in linguistics and brain research and look at the physical theories as an example of human language developed by human brain. It may open new perspectives.

The amount of material exceeds that for a standard one-semester course, so that lecturers can choose what is more appropriate for their audience. About 30 problems with detailed solutions will be provided for the problem-solving sessions and the exams. The book can also be used for independent study by senior undergraduate and graduate students, postdocs and faculty who want to see a bigger picture with connections between different disciplines and find new research opportunities. Readers familiar with thermodynamics and statistics can start from the second Chapter. Those who are also familiar with the basics of kinetics and dynamical chaos can directly go to the third Chapter, consulting the material from the first two when it is referred to. On the other hand, readers from computer science, engineering, mathematics or biology may benefit from reading the first two chapters as they provide some unifying framework for the rest of the book. Bear in mind that the book is written by a natural scientist focused more on "how it works" and "what it is like" and less on the rigor of proofs and definitions.

For a book with such a wide scope, it is probably inevitable not only that my limited expertise in engineering, computer science, biology, economics and linguistics caused some technical errors, but that the dilettante perspective distorted essential elements in the culture of these disciplines. As Schrodinger wrote, "some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them — and at the risk of making fools of ourselves." Fully accepting this risk, I shall maintain a website where objections and corrections will be gratefully received and discussed.

I wish to thank my friends and colleagues whose advices were helpful: Yakov Sinai, Alexander Zamolodchikov, Eric Siggia, Leonid Levitov, Massimo Vergassola, Alexei Kitaev, Jorge Kurchan, Boris Shraiman, William Bialek, Anton Kapustin and Krzysztof Gawedzki. I am grateful to students, postdocs and faculty who attended the lectures at Weizmann, Harvard, New York University, Simons Center, University of Arizona, Skoltech and High

School of Economics, and whose feedback was invaluable in writing this text. Special thanks to Michal Shavit at Weizmann and Saranesh Prembabu at Harvard for helping with tutorials, problems and solutions and for catching numerous errors (the remaining ones are my responsibility alone).

Small-print parts can be omitted upon the first few readings.

1 Thermodynamics and statistical physics

Our knowledge is always partial. If we study macroscopic systems, some degrees of freedom remain hidden. For small sets of atoms or sub-atomic particles, their quantum nature prevents us from knowing precise values of their momenta and coordinates simultaneously. We used to believe that we found the way around the partial knowledge in mechanics, electricity and magnetism, where we have *closed sets of equations describing explicitly known degrees of freedom.* In other words, we learned how to restrict our description only to things that can be considered independent of the unknown within given accuracy. For example, planets are large complex bodies, and yet the motion of their centers of mass in the limit of large distances satisfies closed equations of celestial mechanics[1].

Yet at some point we have realized how illusory was our belief in closed description, since we need to feed it with initial or boundary conditions taken from measurements. Here our knowledge is incomplete because of a finite precision of measurements. This has dramatic consequences, when there is an instability, so that small variation of initial data leads to large deviation in evolution. In a sense, every new decimal in precision is a new degree of freedom for unstable systems (including our Solar System).

In this Chapter we shall deal with *observable manifestations of the hidden degrees of freedom.* While we do not know their state, we do know their nature, whether those degrees of freedom are related to moving particles, spins, bacteria or market traders. That means, in particular, that we know the symmetries and conservation laws of the system.

The first two sections present a phenomenological approach called thermodynamics. The last two sections serve as a brief introduction into statistical physics.

1.1 Basics of thermodynamics

> One can teach monkey to differentiate, integration requires humans.
> G Kotkin

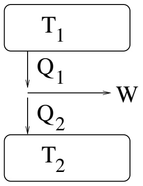
People are burning things to propel objects for at least a couple of thou-

[1] Already the next step — description of a planet rotation — needs the account of many extra degrees of freedom, for instance, oceanic flows (which slow down rotation by tidal forces).

sand years. A regular scientific inquiry on general principles governing conversion of heat into mechanical work was triggered by the practical needs to estimate the engine efficiency during the industrial revolution. That led to the development of the abstract concept of entropy.

A heat engine works by delivering heat from a reservoir with some temperature T_1 via some system to another reservoir with T_2 doing some work in the process. Look under the hood of your car to appreciate the level of abstraction achieved in that definition. The work W is the difference between the heat given by the hot reservoir Q_1 and the heat absorbed by the cold one Q_2. What is the maximal fraction of heat we can use for work? Carnot in 1824 stated that we cannot make Q_2 arbitrarily small: in all processes, $Q_2/T_2 \geq Q_1/T_1$, so that the efficiency is bounded from above:

$$\frac{W}{Q_1} = \frac{Q_1 - Q_2}{Q_1} \leq 1 - \frac{T_2}{T_1}\ . \qquad (1)$$



His elaborate arguments are of only historic interest now. Clausius in 1864 introduced the notion of entropy[2] as a factor connecting temperature and heat, so we now interpret the Carnot criterium, saying that the entropy decrease of the hot reservoirs, $\Delta S_1 = Q_1/T_1$, must be less than the entropy increase of the cold one, $\Delta S_2 = Q_2/T_2$. Maximal work is achieved for minimal (zero) total entropy change, $\Delta S_2 = \Delta S_1$, which happens for reversible processes — if, for instance, a gas works by moving a piston then the pressure of the gas and the work are less for a fast-moving piston than in equilibrium. The efficiency is larger when the temperatures differ more.

Just like the path from Carnot engine to a general thermodynamics, we discover the laws of nature by induction: from particular cases to a general law and from processes to state functions. The latter step requires integration (to pass, for instance, from the Newton equations of mechanics to the Hamiltonian or from thermodynamic equations of state to thermodynamic potentials). It is much easier to differentiate than to integrate, and so deduction (or postulation approach) is usually more pedagogical[3]. It also provides a good vantage point for generalizations and appeals to our brain, which likes

[2] ”Entropy” starts reminding energy and ends with tropos which means turn or way in Greek.

[3] In science, we strive to get the whole truth at any price. Then in teaching we sell its parts at affordable prices.

to hypothesize before receiving any data, as we shall see later. In such an approach, one starts from postulating a variational principle for some function of the state of the system. Then one deduces from that principle the laws that govern changes when one passes from state to state.

Here we present a deductive description of thermodynamics. *Thermodynamics studies restrictions on the possible macroscopic properties that follow from the fundamental conservation laws.* Therefore, thermodynamics does not predict numerical values but rather sets inequalities and establishes relations among different properties.

A traditional way to start building thermodynamics is to identify a conserved quantity, which can be exchanged but not created. It could be matter, money, energy, etc. For most physical systems, the basic symmetry is invariance of the fundamental laws with respect to time shifts[4]. Evolution of an isolated physical system is usually governed by the Hamiltonian (the energy written in canonical variables), whose time-independence means energy conservation. In what follows, the conserved quantity of thermodynamics is called energy and denoted E. We wish to ascribe to the states of the system the values of E. To start with, we focus on the states independent of the way they are prepared; such *equilibrium* states are completely characterized by the *static* values of observable variables.

Passing from state to state under external action involves the energy change, which generally consists of two parts: the energy change of visible degrees of freedom (which we shall call work) and the energy change of hidden degrees of freedom (which we shall call heat). To be able to measure energy changes in principle, we need adiabatic processes where there is no heat exchange, that is all the energy changes are visible. Ascribing to every state its energy (up to an additive constant common for all states) hinges on our ability to relate any two equilibrium states A and B by an adiabatic process either $A \to B$ or $B \to A$, which allows to measure the difference in their energies by the work W done by the system. Now, if we encounter a process where the energy change is not equal to the work, we call the difference the heat exchange δQ:

$$dE = \delta Q - \delta W \ . \tag{2}$$

[4] Be careful trying to build thermodynamics for biological or social-economic systems, since generally the laws that govern them are not time-invariant. For example, the metabolism of the living beings changes with age, and the number of market regulations generally increases (as well as the total money mass, albeit not necessarily in our pockets).

This statement is known as the first law of thermodynamics. It is nothing but declaration of our belief in energy conservation: if the visible energy balance does not hold then the energy of the hidden must change. The energy is a function of state so we use differential, but we use δ for heat and work, which aren't differentials of any function. Heat exchange and work depend on the path taken from A to B, that is they refer to particular forms of energy transfer (not energy content). The first law was experimentally discovered by Mayer in 1842; before that, heat was believed to be a separate fluid conserved by itself.

The basic problem of thermodynamics is the determination of the equilibrium state that eventually results after all internal constraints are removed in a closed composite system. The problem is solved with the help of extremum principle: there exists a quantity S called entropy which is a function of the parameters of the system. The values assumed by the parameters in the absence of an internal constraint maximize the entropy over the manifold of available states (Clausius 1865).

Thermodynamic limit. Traditionally, thermodynamics have dealt with extensive parameters whose value grows linearly with the number of degrees of freedom. Additive quantities like number of particles N, electric charge and magnetic moment are extensive. Energy generally is not additive, that is the energy of a composite system is not the sum of the parts because of an interaction energy: $E(N_1) + E(N_2) \neq E(N_1 + N_2)$. To treat energy as an additive variable we make two assumptions: i) assume that the forces of interaction are short-range and act only along the boundary, ii) take thermodynamic limit $V \to \infty$ where one can neglect surface terms that scale as $V^{2/3} \propto N^{2/3}$ in comparison with the bulk terms that scale as $V \propto N$.

In that limit, thermodynamic entropy is also an extensive variable[5], which is a homogeneous first-order function of all the extensive parameters:

$$S(\lambda E, \lambda V, \ldots) = \lambda S(E, V, \ldots) \ . \tag{3}$$

This function $S(E, V, \ldots)$, called also fundamental relation, is *everything* one needs to know to solve the basic problem (and others) in thermodynamics.

Of course, (3) does not mean that $S(E)$ is a linear function when other parameters fixed: $S(\lambda E, V, \ldots) \neq \lambda S(E, V, \ldots)$. On the contrary, we shall

[5] We shall see later that non-extensive parts of entropy are also important for studying interaction and correlations between subsystems.

see in a moment that it is a convex function. Nor is entropy necessarily a monotonic function of energy — an example of the two-level system in Section 1.4 shows that $S(E)$ could be non-monotonic for systems with a finite phase space. Yet for every interval of a definite derivative sign, say $(\partial E/\partial S)_X > 0$, we can solve $S = S(E, V, \ldots)$ uniquely for $E(S, V, \ldots)$ which is an equivalent fundamental relation. We assume the functions $S(E, X)$ and $E(S, X)$ to be continuous differentiable for any other parameter X. An efficient way to treat partial derivatives is to use jacobians

$$\frac{\partial(u,v)}{\partial(x,y)} \equiv \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial v}{\partial x}\frac{\partial u}{\partial y}\,, \qquad \left(\frac{\partial u}{\partial x}\right)_y = \frac{\partial(u,y)}{\partial(x,y)}\,.$$

Then

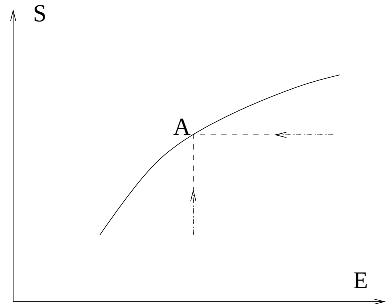
$$\left(\frac{\partial S}{\partial X}\right)_E = 0 \Rightarrow \left(\frac{\partial E}{\partial X}\right)_S = \frac{\partial(ES)}{\partial(XS)}\frac{\partial(EX)}{\partial(EX)} = -\left(\frac{\partial S}{\partial X}\right)_E \left(\frac{\partial E}{\partial S}\right)_X = 0\ .$$

That means that any entropy extremum is also an energy extremum. Differentiating the last relation one more time we differentiate only the first factor, since it turns into zero at equilibrium:

$$(\partial^2 E/\partial X^2)_S = -(\partial^2 S/\partial X^2)_E (\partial E/\partial S)_X\,.$$

The equilibrium is an entropy maximum, that is $-(\partial^2 S/\partial X^2)_E$ is negative. Which type of extremum energy has at equilibrium depends on the sign of $(\partial E/\partial S)_X$, which is called temperature, see (4) below. When the temperature is positive, the equilibrium is the energy minimum. Relation between entropy and energy at equilibrium is very much like circle can be defined as the figure of either maximal area for a given perimeter or minimal perimeter for a given area.

The figure shows the restriction imposed by thermodynamics on possible states: unconstrained equilibrium ones are on the curve while all other states lie below. It is important that the equilibrium curve $S(E)$ is convex, which guarantees stability of a homogeneous state. Indeed, if our system would break spontaneously into two halves with a bit different energies, the entropy must decrease: $2S(E) > S(E+\Delta) + S(E-\Delta) = 2S(E) + S''\Delta^2$, which requires $S'' < 0$ (that argument does not work for systems with long-range interaction where energy is non-additive). Convexity also guarantees that one can reach the state A either maximizing entropy at a given energy or minimizing energy at a given entropy:



One can work either in energy or entropy representation but ought to be careful not to mix the two.

Experimentally, one usually measures *changes* thus finding derivatives (called equations of state). The partial derivatives of an extensive variable with respect to its arguments (also extensive parameters) are intensive parameters. In thermodynamics we have only extensive and intensive variables, because we take thermodynamic limit $N \to \infty$, $V \to \infty$ keeping N/V finite. For example, for the energy one writes

$$\frac{\partial E}{\partial S} \equiv T(S,V,N)\,, \quad \frac{\partial E}{\partial V} \equiv -P(S,V,N) \quad \frac{\partial E}{\partial N} \equiv \mu(S,V,N)\ ,\ldots \tag{4}$$

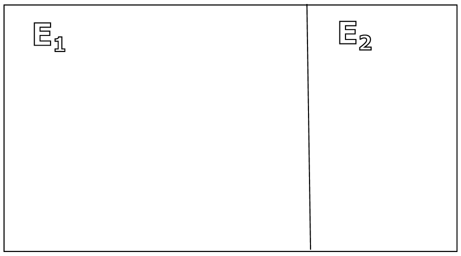
These relations are called the *equations of state* and they serve as *definitions* for temperature T, pressure P and chemical potential μ, corresponding to the respective extensive variables are S, V, N. We shall see later that entropy is the missing information, so that temperature is the energetic price of information. Our entropy is dimensionless, so that T is assumed to be multiplied by the Boltzmann constant $k = 1.3 \cdot 10^{-23} J/K$ and have the same dimensionality as the energy. From (4) we write

$$dE = \delta Q - \delta W = TdS - PdV + \mu dN\ . \tag{5}$$

Entropy is thus responsible for hidden degrees of freedom (i.e. heat) while other extensive parameters describe macroscopic degrees of freedom. We see that in equilibrium the missing information is maximal for hidden degrees of freedom.

The derivatives (4) are taken at equilibrium, where definite relation exists between variables, for instance, E and S. That means that (5) is true only for *quasi-static processes* i.e such that the system is close to equilibrium at every point of the process. A process can be considered quasi-static if its typical time of change is larger than the relaxation times (which for pressure can be estimates as L/c, for temperature as L^2/κ, where L is a system size, c - sound velocity and κ thermal conductivity). Finite deviations from equilibrium make $dS > \delta Q/T$ because entropy can increase without heat transfer. Only recently have we learnt how to measure equilibrium quantities in fast non-equilibrium processes, as will be described in Section 5.3.

Let us see how the entropy maximum principle solves the basic problem. Consider two simple systems separated by a rigid wall which is impermeable for anything but heat. The whole composite system is closed that is $E_1 + E_2 =$const.



The entropy change under the energy exchange,

$$dS = \frac{\partial S_1}{\partial E_1} dE_1 + \frac{\partial S_2}{\partial E_2} dE_2 = \frac{dE_1}{T_1} + \frac{dE_2}{T_2} = \left(\frac{1}{T_1} - \frac{1}{T_2}\right) dE_1 \,, \tag{6}$$

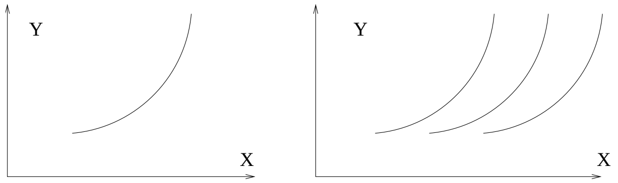
must be positive. For positive temperature, that means that energy flows from the hot subsystem to the cold one ($T_1 > T_2 \Rightarrow dE_1 < 0$). We see that our definition (4) is in agreement with our intuitive notion of temperature. When equilibrium is reached, $dS = 0$ which requires $T_1 = T_2$. If fundamental relation is known, then so is the function $T(E, V)$. Two equations, $T(E_1, V_1) = T(E_2, V_2)$ and $E_1 + E_2 =$const completely determine E_1 and E_2. In the same way one can consider movable wall and get $P_1 = P_2$ in equilibrium. If the wall allows for particle penetration we get $\mu_1 = \mu_2$ in equilibrium.

Example: Consider a system that is characterized solely by its energy, which can change between zero and E_{max}. The equation of state is the energy-temperature relation $E/E_{max} = (1 + e^{\epsilon/T})^{-1}$, which tends to 1/2 at $T \gg \epsilon$ and is exponentially small at $T \ll \epsilon$. In Section 1.3, we shall identify this with a set of elements with two energy levels, 0 and ϵ. To find the fundamental relation in the entropy representation, we integrate the equation of state:

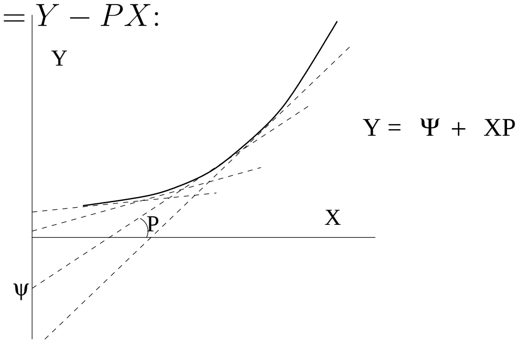
$$\frac{1}{T} = \frac{dS}{dE} = \frac{1}{\epsilon} \ln \frac{E_{max} - E}{E} \Rightarrow S(E) = \frac{E_{max}}{\epsilon} \ln \frac{E_{max}}{E_{max} - E} + \frac{E}{\epsilon} \ln \frac{E_{max} - E}{E} \,. \tag{7}$$

1.2 Thermodynamic potentials

The fundamental relation always relates extensive quantities. Therefore, even though it is always possible to eliminate, say, S from $E = E(S, V, N)$ and $T = T(S, V, N)$ getting $E = E(T, V, N)$, this *is not* a fundamental relation and it does not contain all the information. Indeed, $E = E(T, V, N)$ is actually a partial differential equation (because $T = \partial E/\partial S$) and even if it can be integrated the result would contain undetermined function of V, N. Still, it is easier to measure, say, temperature than entropy so it is convenient to have a complete formalism with an intensive parameter as operationally independent variable and an extensive parameter as a derived quantity. This is achieved by the Legendre transform: We want to pass from the relation $Y = Y(X)$ to that in terms of $P = \partial Y/\partial X$. Yet it is not enough to eliminate X and consider the function $Y = Y[X(P)] = Y(P)$, because such function determines the curve $Y = Y(X)$ only up to a shift along X, which changes neither Y nor P:



For example, consider the whole family of functions $Y = (X + C)^2$ for arbitrary C. We express $P = dY/dX = 2(X + C)^2$ and substitute $Y = (X + C)^2 = P^2/4$ — this single function corresponds to the whole family, that is does not allow to pick a single function we need. In other words, the family $Y = (X + C)^2$ solves the differential equation $Y = (dY/dX)^2/4$. To fix the shift, nail the curve and pick a single function, for every P we specify not Y but the position $\psi(P)$ where the straight line tangent to the curve intercepts the Y-axis: $\psi = Y - PX$:



In this way we consider the curve $Y(X)$ as the envelope of the family of the tangent lines, each characterized by the slope P *and* the intercept

ψ. The relation between them, $\psi(P) = Y[X(P)] - PX(P)$, completely defines the curve; here one substitutes $X(P)$ found from $P = dY(X)/dX$. The function $\psi(P)$ is called the Legendre transform of $Y(X)$. From $d\psi = -PdX - XdP + dY = -XdP$ one gets $-X = d\psi/dP$ i.e. the inverse transform is the same up to a sign: $Y = \psi + XP$.

The transform is possible when for every X there is one P, that is $P(X)$ is monotonic and $Y(X)$ is convex, $dP/dX = d^2Y/dX^2 \neq 0$. Sign-definite second derivative means that the function is either concave or convex. This is the second time we meet convexity, which we related above to the stability of a homogeneous state. Convexity and concavity will play an important role in this course.

Different thermodynamics potentials suitable for different physical situations are obtained replacing different extensive parameters by the respective intensive parameters. For example, free energy $F = E - TS$ (also called Helmholtz potential) is that partial Legendre transform of E which replaces the entropy by the temperature as an independent variable: $dF(T, V, N, \ldots) = -SdT - PdV + \mu dN + \ldots$. Counterpart to $(\partial E/\partial S)_{VN} = T$ is $(\partial F/\partial T)_{VN} = -S$. The free energy is particularly convenient for the description of a system in a thermal contact with a heat reservoir because then the temperature is fixed and we have one variable less to care about. The maximal work that can be done under a constant temperature (equal to that of the reservoir) is minus the differential of the free energy. Indeed, this is the work done *by the system and the thermal reservoir*. Is that work generally larger or smaller than the work done by the system alone? Let's see. That work is equal to the change of the total energy:

$$d(E + E_r) = dE + T_r dS_r = dE - T_r dS = d(E - T_r S) = d(E - TS) = dF \ .$$

In other words, the free energy $F = E - TS$ is that part of the internal energy which is *free* to turn into work, the rest of the energy TS we must keep to sustain a constant temperature. The equilibrium state minimizes F, not absolutely, but over the manifold of states with the temperature equal to that of the reservoir. Consider, for instance, minimization of $F(T, V) = E[S(T, V), V] - TS(T, V)$ with respect to volume:

$$\left(\frac{\partial F}{\partial V}\right)_T = \left(\frac{\partial E}{\partial V}\right)_S + \left(\frac{\partial E}{\partial S} - T\right)\frac{\partial S}{\partial V} = \left(\frac{\partial E}{\partial V}\right)_S ,$$

that is the derivatives turn into zero and E and F reach extrema simultaneously. Also, in the point of an extremum, one gets $(\partial^2 E/\partial V^2)_S =$

$(\partial^2 F/\partial V^2)_T$ i.e. both E and F have the same type of extremum (minimum in a positive-temperature equilibrium). Monatomic gas at fixed T, N has $F(V) = E - TS(V) = -NRT \ln V$+const. If a piston separates equal amounts N, then the work done in changing the volume of a subsystem from V_1 to V_2 is $\Delta F = NRT \ln[V_2(V - V_2)/V_1(V - V_1)]$.

System can reach the minimum of the free energy minimizing energy and maximizing entropy. The former often requires creating some order in the system, for instance, orienting all spins parallel in a ferromagnetic or anti-parallel in an anti-ferromagnetic. On the contrary, increasing entropy requires disorder. Which of these tendencies wins depends on temperature, setting their relative importance. In later sections, we shall see over and over again that looking for a minimum of some free energy is a universal approach, from finding an equilibrium state of a physical system to designing the most optimal algorithm of information processing.

Other thermodynamic potentials and the formal structure of thermodynamics are described in Appendix 8.1. Since the Legendre transform is invertible, all potentials are equivalent and contain the same information. The choice of the potential for a given physical situation is that of convenience: we usually take what is fixed as a variable to diminish the number of effective variables.

The next two sections present a brief recount of classical Boltzmann-Gibbs statistical approach: We introduce microscopic statistical description in the phase space and describe two principal ways (microcanonical and canonical) to derive thermodynamics from statistics.

1.3 Microcanonical distribution

Consider a *closed* system with the fixed energy E. Boltzmann *assumed* that all microstates with the same energy have equal probability (ergodic hypothesis). If the number of such states is $\Gamma(E)$ then the *microcanonical probability distribution* is as follows:

$$w_a(E) = 1/\Gamma(E), \tag{8}$$

To link statistical physics with thermodynamics one must define the fundamental relation i.e. a thermodynamic potential as a function of respective variables. For microcanonical distribution, Boltzmann introduced the entropy as

$$S(E, V, N) = \ln \Gamma(E, V, N) \ . \tag{9}$$

This is one of the most important formulas in physics[6] (on a par with $F = ma\,, E = mc^2$ and $E = \hbar\omega$).

Noninteracting subsystems are statistically independent. That means that the statistical weight of the composite system is a product - indeed, for every state of one subsystem we have all the states of another. If the weight is a product then the entropy is a sum. For interacting subsystems, this is true only for short-range forces in the thermodynamic limit $N \to \infty$.

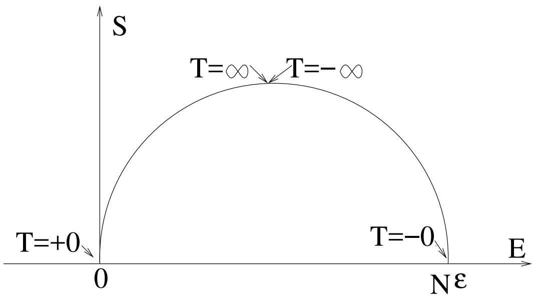
Consider two subsystems, 1 and 2, that can exchange energy. Let's see how statistics solves the basic problem of thermodynamics (to define equilibrium) that we treated above in (6). Assume that the indeterminacy in the energy of any subsystem, Δ, is much less than the total energy E. Then

$$\Gamma(E) = \sum_{i=1}^{E/\Delta} \Gamma_1(E_i)\Gamma_2(E - E_i) \ . \tag{10}$$

We denote $\bar{E}_1, \bar{E}_2 = E - \bar{E}_1$ the values that correspond to the maximal term in the sum (11). To find this maximum, we compute the derivative of it, which is proportional to $(\partial\Gamma_1/\partial E_i)\Gamma_2 + (\partial\Gamma_2/\partial E_i)\Gamma_1 = (\Gamma_1\Gamma_2)[(\partial S_1/\partial E_1)_{\bar{E}_1} - (\partial S_2/\partial E_2)_{\bar{E}_2}]$. The extremum condition, $(\partial S_1/\partial E_1)_{\bar{E}_1} = (\partial S_2/\partial E_2)_{\bar{E}_2}$, corresponds to the thermal equilibrium where the temperatures of the subsystems are equal. The equilibrium is thus where the maximum of probability is. It is obvious that $\Gamma(\bar{E}_1)\Gamma(\bar{E}_2) \le \Gamma(E) \le \Gamma(\bar{E}_1)\Gamma(\bar{E}_2)E/\Delta$. If the system consists of N particles and $N_1, N_2 \to \infty$ then $S(E) = S_1(\bar{E}_1) + S_2(\bar{E}_2) + O(logN)$ where the last term is negligible in the thermodynamic limit.

The same definition (entropy as a logarithm of the number of states) is true for any system with a discrete set of states. For example, consider the set of N particles (spins, neurons), each with two energy levels 0 and ϵ. If the energy of the set is E then there are $L = E/\epsilon$ upper levels occupied. The statistical weight is determined by the number of ways one can choose L out of N: $\Gamma(N, L) = C_N^L = N!/L!(N - L)!$. We can now define entropy (i.e. find the fundamental relation): $S(E, N) = \ln\Gamma$. At the thermodynamic limit $N \gg 1$ and $L \gg 1$, it gives $S(E, N) \approx N \ln[N/(N - L)] + L \ln[(N - L)/L]$, which coincides with (7). The entropy as a function of energy is drawn in the Figure:

[6] It is inscribed on the Boltzmann's gravestone.



The entropy is symmetric about $E = N\epsilon/2$ and is zero at $E = 0, N\epsilon$ when all the particles are in the same state.. The equation of state (temperature-energy relation) is $T^{-1} = \partial S/\partial E \approx \epsilon^{-1}\ln[(N-L)/L]$. We see that when $E > N\epsilon/2$ then the population of the higher level is larger than of the lower one (inverse population as in a laser) and the temperature is negative. Negative temperature may happen only in systems with the upper limit of energy levels and simply means that by adding energy beyond some level we actually decrease the entropy i.e. the number of accessible states. That example with negative temperature is to help you to disengage from the everyday notion of temperature and to get used to the physicist idea of temperature as the derivative of energy with respect to entropy [7] .

Available (non-equilibrium) states lie below the $S(E)$ plot. The entropy maximum corresponds to the energy minimum for positive temperatures and to the energy maximum for the negative temperatures. Imagine now that the system with a negative temperature is brought into contact with the thermostat (having positive temperature). To equilibrate with the thermostat, the system needs to acquire a positive temperature. A glance on the figure shows that our system must give away energy (a laser generates and emits light). If this is done adiabatically slow, that is along the equilibrium curve, the system first decreases the temperature further until it passes through minus infinity right into plus infinity and then down to positive values until it eventually reaches the temperature of the thermostat. That is negative temperatures are actually "hotter" than positive. If you put your hand on a negative temperature system, you feel heat flowing into you. By itself though the system is stable since $\partial^2 S/\partial E^2 = -N/L(N-L)\epsilon^2 < 0$ at any temperature. Stress that there is no volume in $S(E,N)$ that is we consider only subsystem or only part of the degrees of freedom. Indeed, real particles

[7] And yet deep within the deductive approach, it is worth remembering that in the inductive development of thermodynamics the unique role was played by the particular notion of temperature as a mean kinetic energy of the molecules of an ideal gas.

have kinetic energy unbounded from above and can correspond only to positive temperatures [negative temperature and infinite energy give infinite Gibbs factor $\exp(-E/T)$].

The derivation of thermodynamic fundamental relation $S(E,\ldots)$ in the microcanonical ensemble is thus via the number of states or phase volume.

1.4 Canonical distribution and fluctuations

Let us now describe the statistical description, which corresponds to the thermodynamic potential of free energy $F(T)$. Consider a system exchanging energy with a thermostat, which can be thought of as consisting of infinitely many copies of our system — this is so-called canonical ensemble, characterized by T. Here our system can have any energy and the question arises what is the probability to be in a given microstate a with the energy E. We derive that probability distribution (called canonical) from the microcanonical distribution of the whole system. Since all the states of the thermostat are equally likely to occur, then the probability should be directly proportional to the statistical weight of the thermostat $\Gamma_0(E_0-E)$. Here we assume $E \ll E_0$, expand (in the exponent!) $\Gamma_0(E_0-E) = \exp[S_0(E_0-E)] \approx \exp[S_0(E_0) - E/T)]$ and obtain

$$w_a(E) = Z^{-1}\exp(-E/T) \ , \tag{11}$$

$$Z = \sum_a \exp(-E_a/T) \ . \tag{12}$$

Note that there is no trace of the thermostat left except for the temperature. The normalization factor $Z(T,V,N)$ is a sum over all states accessible to the system and is called the partition function.

Our subsystem is macroscopic itself, so it has many ways to re-distribute the energy E among its degrees of freedom. In other words, it has many microscopic states that correspond to the same total energy of the subsystem. The probability for the subsystem to have a given energy is the probability of the state (12) times the number of states i.e. the statistical weight of the *subsystem*:

$$W(E) = \Gamma(E)w_a(E) = \Gamma(E)Z^{-1}\exp(-E/T) \ . \tag{13}$$

Here the weight $\Gamma(E)$ grows with E very fast for large E. But as $E \to \infty$ the exponent $\exp(-E/T)$ decays faster than any power. As a result, $W(E)$

is concentrated in a very narrow peak and the energy fluctuations around $\bar{E}$ are very small. For example, for an ideal gas $W(E) \propto E^{3N/2}\exp(-E/T)$. Let us stress again that the Gibbs canonical distribution (12) tells that the probability of a given microstate exponentially decays with the energy of the state while (14) tells that the probability of a given energy has a peak.

To get thermodynamics from the Gibbs distribution one needs to define the free energy because we are under a constant temperature. This is done via the partition function Z (which is of central importance since macroscopic quantities are generally expressed via the derivatives of it):

$$F(T) = -T \ln Z(T) \ . \tag{14}$$

To prove that, differentiate the identity $Z = \exp(-F/T) = \sum_a \exp(-E_a/T)$ with respect to temperature, which gives

$$Z\left(\frac{F}{T^2} - \frac{1}{T}\frac{\partial F}{\partial T}\right) = -\frac{1}{T^2}\sum_a E_a e^{-E_a/T} \ \Rightarrow \ F = \bar{E} + T\frac{\partial F}{\partial T} = E - TS \ ,$$

which we had in thermodynamics.

One can also relate statistics and thermodynamics by defining entropy. Remind that for a closed system Boltzmann defined $S = \ln \Gamma$ while the probability of state was $w_a = 1/\Gamma$. In other words, the entropy was minus the log of probability, $S = -\ln w_a$. For a subsystem at fixed temperature, different states have different probabilities, and both energy and entropy fluctuate. What should be the thermodynamic entropy: mean entropy $-\langle \ln w_a \rangle$ or entropy at a mean energy $\ln w_a(E)$? For a system that has a Gibbs distribution, $\ln w_a$ is linear in E_a, so that the entropy at a mean energy is the mean entropy, and we recover the standard thermodynamic relation. Indeed, the mean entropy,

$$\begin{aligned} \langle S \rangle &= - \ \ \langle \ln w_a \rangle = -\sum w_a \ln w_a = \sum w_a (E_a/T + \ln Z) \\ &= \ \ E/T + \ln Z = (E - F)/T \ , \end{aligned} \tag{15}$$

is the same as the logarithm of the probability at the mean energy:

$$S(E) = -\ln w_a(E) = -\ln\left[Z^{-1}\exp(-E/T)\right] = E/T + \ln Z = (E - F)/T \ . \tag{16}$$

Even though the Gibbs entropy, $S = -\sum w_a \ln w_a$ is derived here for equilibrium, this definition can be used for any set of probabilities w_a, since it provides a useful measure of our ignorance about the system, as we shall see later.

Are canonical and microcanonical statistical descriptions equivalent? Of course, not. The descriptions are equivalent only when fluctuations are neglected and consideration is restricted to mean values. That takes place in thermodynamics, where the distributions just produce different fundamental relations between the mean values: $S(E)$ for microcanonical, $F(T)$ for canonical. These functions are related by the Legendre transforms. How operationally does one check, for instance, the equivalence of canonical and microcanonical energies? One takes an isolated system at a given energy E, measures the derivative $\partial E/\partial S$, then puts it into the thermostat with the temperature equal to that $\partial E/\partial S$; the energy now fluctuates but the *mean* energy must be equal to E (as long as system is macroscopic and all the interactions are short-range).

As far as fluctuations are concerned, there is a natural hierarchy: microcanonical distribution neglects fluctuations in energy and number of particles, canonical distribution neglects fluctuations in N but accounts for fluctuations in E. The choice of description is dictated only by convenience in thermodynamics because it treats only mean values. But if we want to describe the whole statistics of the system in thermostat, we need to use canonical distribution, not the micro-canonical one.

1.5 Evolution in the phase space

So far we said precious little about how physical systems actually evolve to arrive at equilibrium. Let us start from a broad class of energy-conserving systems that can be described by a Hamiltonian evolution. Every such system is characterized by its momenta p and coordinates q, together comprising the phase space. Any state of a system is a point in the space. As time progresses, coordinates and moments change, and the point moves in the phase space. Since we cannot measure p, q exactly, we ought to consider finite regions. We define probability for a system to be in some $\Delta p \Delta q$-region of the phase space as the fraction Δt of the total observation time T it spends there: $w = \Delta t/T$. Assuming that the probability to find it within the volume $dpdq$ is proportional to this volume, we introduce the statistical distribution in the phase space as a density: $dw = \rho(p, q)dpdq$. By definition, the average with the statistical distribution is equivalent to the time average:

$$\bar{f} = \int f(p,q)\rho(p,q)dpdq = \frac{1}{T}\int_0^T f(t)dt \ . \tag{17}$$

We can now consider the evolution of the density $\rho(p,q)$ on timescales larger than T used to define it. In a flow with the velocity $\mathbf{v} = (\dot{p}, \dot{q})$, the density changes according to the continuity equation: $\partial\rho/\partial t + div\,(\rho\mathbf{v}) = 0$. For not very long time, we can neglect interaction between subsystems, so that the motion can be described by the Hamiltonian dynamics of the momenta and coordinates of the subsystem itself: $\dot{q}_i = \partial\mathcal{H}/\partial p_i$ and $\dot{p}_i = -\partial\mathcal{H}/\partial q_i$. The respective continuity equation for the probability density is called Liouville equation:

$$\frac{\partial\rho}{\partial t} = -div\,(\rho\mathbf{v}) = \sum_i \frac{\partial\mathcal{H}}{\partial p_i}\frac{\partial\rho}{\partial q_i} - \frac{\partial\mathcal{H}}{\partial q_i}\frac{\partial\rho}{\partial p_i} \equiv \{\rho, \mathcal{H}\}\,. \tag{18}$$

Here the Hamiltonian generally depends on the momenta and coordinates of the given subsystem and its neighbors.

The equation (19) describes the evolution of the density at a given point of the phase space since the time derivative at the left is partial, that is takes at fixed p_i, q_i. Any given physical system changes its momenta and coordinates, that is moves in the phase space. The density change for a system is then described by the full derivative taken along the flow: $d\rho/dt = \partial\rho/\partial t + (\mathbf{v}\nabla)\rho$. What is most important for us now is that any Hamiltonian flow in the phase space is incompressible: it conserves area in each plane p_i, q_i and the total volume: $div\,\mathbf{v} = \partial\dot{q}_i/\partial q_i + \partial\dot{p}_i/\partial p_i = 0$. That gives the Liouville theorem: $d\rho/dt = \partial\rho/\partial t + (\mathbf{v}\nabla)\rho = -\rho div\,\mathbf{v} = 0$. The statistical distribution is thus conserved along the phase trajectories of any system. As a result, ρ is an integral of motion.

At equilibrium, ρ must be expressed solely via the integrals of motion. We define statistical equilibrium as a state where macroscopic quantities are equal to the mean values. Assuming short-range forces we conclude that different macroscopic subsystems interact weakly and are statistically independent so that the distribution for a composite system ρ_{12} is factorized: $\rho_{12} = \rho_1\rho_2$. Since $\ln\rho$ is an additive quantity then in equilibrium it must be expressed linearly via the additive integrals of motions which for a general mechanical system are momentum $\mathbf{P}(p,q)$, the momentum of momentum $\mathbf{M}(p,q)$ and energy $E(p,q)$ (again, neglecting interaction energy of subsystems):

$$\ln\rho_a = \alpha_a + \beta E_a(p,q) + \mathbf{c}\cdot\mathbf{P}_a(p,q) + \mathbf{d}\cdot\mathbf{M}(p,q) \ . \tag{19}$$

Here α_a is the normalization constant for a given subsystem while the seven constants $\beta, \mathbf{c}, \mathbf{d}$ are the same for all subsystems (to ensure additivity of integrals) and are determined by the values of the seven integrals of motion for the whole system. We thus conclude that the additive integrals of motion is all we need to get the statistical distribution of a closed system (and any subsystem). These integrals replace all the enormous microscopic information. Considering subsystem which neither moves nor rotates we are down to the single integral, energy, which corresponds to the Gibbs' *canonical distribution*:

$$\rho(p,q) = A\exp[-\beta E(p,q)] \ . \tag{20}$$

It was obtained for any macroscopic subsystem of a very large system, which is the same as any system in contact with s thermostat. Note one subtlety: On the one hand, we considered subsystems weakly interacting to have their energies additive and distributions independent. On the other hand, precisely this weak interaction is expected to drive a complicated evolution visiting all regions of the phase space, thus making statistical description possible. Particular case of (21) is a distribution constant over all the phase space (kind of microcanonical), which is evidently invariant under the Hamiltonian evolution of an isolated system due to Liouville theorem. That distribution formally corresponds to $\beta = 1/T = 0$ — at infinite temperature canonical and microcanonical distributions coincide, since energy differences between different regions of the phase space do not matter.

Since the system spends largest time in equilibrium, it must be the most probable state, that is realize the entropy maximum. In particular, the canonical equilibrium distribution (21) corresponds to the maximum of the Gibbs entropy, $S = -\int \rho \ln \rho \, dpdq$, under the condition of the given mean energy $\bar{E} = \int \rho(p,q)E(p,q)\, dpdq$. Indeed, requiring zero variation $\delta(S + \beta\bar{E}) = 0$ we obtain (21). For an isolated system with a fixed energy, the entropy maximum corresponds to a uniform micro-canonical distribution.

2 Inevitability of irreversibility

> Où sont les neiges d'antan?
> François Villon

It is time for reflection. The most obvious contradiction we face is between distribution preservation by Hamiltonian evolution and the growth of its entropy. More generally, the puzzle here is how irreversible entropy growth appears out of reversible laws of mechanics. If we screen the movie of any evolution backwards, it will be a legitimate solution of the equations of motion. Will it have its entropy decreasing? Can we also decrease entropy by employing the Maxwell demon who can distinguish fast molecules from slow ones and selectively open a window between two boxes to increase the temperature difference between the boxes and thus decrease entropy?

These conceptual questions have been already posed in the 19 century. It took the better part of the 20 century to find answers, resolve the puzzles and make statistical physics conceptually trivial (and technically much more powerful). This required two things: i) better understanding dynamics and revealing the mechanism of randomization called dynamical chaos, ii) understanding entropy as a measure of uncertainty and developing the information theory. This Chapter is devoted to the first subject, the next Chapter — to the second one. Here we describe how irreversibility and relaxation to equilibrium essentially follows from necessity to consider ensembles (regions in phase space) due to incomplete knowledge. Initially small regions spread over the whole phase space under reversible Hamiltonian dynamics, very much like flows of an incompressible liquid are mixing. Such spreading and mixing in phase space correspond to the approach to equilibrium. On the contrary, to deviate a system from equilibrium, one adds external forcing and dissipation, which makes its phase flow compressible and distribution non-uniform. Difference between equilibrium and non-equilibrium distributions in phase space can then be expressed by the difference between incompressible and compressible flows.

2.1 Kinetic equation and H-theorem

How a Hamiltonian evolution can bring a system to the equilibrium and reach the entropy maximum? Such evolution is an incompressible flow in the phase space, $\operatorname{div}\mathbf{v}=0$, so it conserves the *total* Gibbs entropy:

$$\frac{dS}{dt}=-\int d\mathbf{x}\frac{\partial\rho}{\partial t}\ln\rho=\int d\mathbf{x}\,\ln\rho\operatorname{div}\rho\mathbf{v}=-\int d\mathbf{x}\,(\mathbf{v}\nabla)\rho=\int d\mathbf{x}\,\rho\operatorname{div}\mathbf{v}=0\,.$$

Which entropy then can grow? Boltzmann answered this question by deriving the equation on the one-particle momentum probability distribution. Such equation must follow from integrating the N-particle Liouville equation (19) over all N coordinates and $N-1$ momenta. Consider the phase-space probability density $\rho(\mathbf{x},t)$ in the space $\mathbf{x}=(\mathbf{P},\mathbf{Q})$, where $\mathbf{P}=\{\mathbf{p}_1\ldots\mathbf{p}_N\}$

and $\mathbf{Q} = \{\mathbf{q}_1 \dots \mathbf{q}_N\}$. For the system with the Hamiltonian $\mathcal{H} = \sum_i \frac{p_i^2}{2m} + \sum_{i<j} U(\mathbf{q}_i - \mathbf{q}_j)$, the evolution of the density is described by the following Liouville equation:

$$\frac{\partial \rho(\mathbf{P}, \mathbf{Q}, t)}{\partial t} = \{\rho(\mathbf{P}, \mathbf{Q}, t), \mathcal{H}\} = \left[-\sum_i^N \frac{\mathbf{p}_i}{m} \frac{\partial}{\partial \mathbf{q}_i} + \sum_{i<j} \theta_{ij} \right] \rho(\mathbf{P}, \mathbf{Q}, t) \ , \quad (21)$$

where

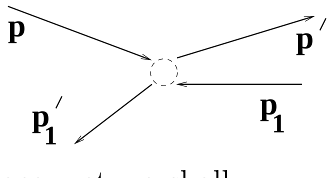
$$\theta_{ij} = \theta(\mathbf{q}_i, \mathbf{p}_i, \mathbf{q}_j, \mathbf{p}_j) = \frac{\partial U(\mathbf{q}_i - \mathbf{q}_j)}{\partial \mathbf{q}_i} \left(\frac{\partial}{\partial \mathbf{p}_i} - \frac{\partial}{\partial \mathbf{p}_j} \right)$$

has a meaning of an inverse typical time of the momentum change due to interaction. For a reduced description of the single-particle distribution over momenta $\rho(\mathbf{p}, t) = \int \rho(\mathbf{P}, \mathbf{Q}, t) \delta(\mathbf{p}_1 - \mathbf{p}) \, d\mathbf{p}_1 \dots d\mathbf{p}_N d\mathbf{q}_1 \dots d\mathbf{q}_N$, we integrate (22). The terms with $\partial/\partial \mathbf{q}_i$ do not contribute, and we get:

$$\frac{\partial \rho(\mathbf{p}, t)}{\partial t} = \int \delta(\mathbf{p}_1 - \mathbf{p}) \theta(\mathbf{q}_1, \mathbf{p}_1; \mathbf{q}_2, \mathbf{p}_2) \rho(\mathbf{q}_1, \mathbf{p}_1; \mathbf{q}_2, \mathbf{p}_2) \, d\mathbf{q}_1 d\mathbf{p}_1 d\mathbf{q}_2 d\mathbf{p}_2 \ . \ (22)$$

This equation is apparently not closed since the rhs contains two-particle probability distribution. If we write respective equation on that two-particle distribution integrating the Liouville equation over $N-2$ coordinates and momenta, the interaction θ-term brings three-particle distribution, etc. Consistent procedure is to assume a short-range interaction and a low density, so that the mean distance between particles much exceeds the radius of interaction. In this case we may assume for every binary collision that particles come from large distances and their momenta are not correlated. Statistical independence then allows one to replace the two-particle momenta distribution by the product of one-particle distributions.

Such derivation is cumbersome, but it is easy to write the general form that such a closed equation must have. For a dilute gas, only two-particle collisions need to be taken into account in describing the evolution of the single-particle distribution over moments $\rho(\mathbf{p}, t)$. Consider the collision of two particles having momenta $\mathbf{p}, \mathbf{p}_1$:



For that, they must come to the same place, yet we shall *assume* that the particle velocity is independent of the position and that the momenta of two particles are statistically independent, that is the probability is the product of single-particle probabilities: $\rho(\mathbf{p},\mathbf{p}_1)=\rho(\mathbf{p})\rho(\mathbf{p}_1)$. These very strong assumptions constitute what is called *the hypothesis of molecular chaos*. Under such assumptions, the number of such collisions (per unit time per unit volume) must be proportional to the probabilities $\rho(\mathbf{p})\rho(\mathbf{p}_1)$ and depend both on the initial momenta $\mathbf{p}$, $\mathbf{p}_1$ and the final ones $\mathbf{p}'$, $\mathbf{p}'_1$:

$$w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)\rho(\mathbf{p})\rho(\mathbf{p}_1)\,d\mathbf{p}d\mathbf{p}_1d\mathbf{p}'d\mathbf{p}'_1\ . \tag{23}$$

One may *believe* that (24) must work well when the one-particle distribution function evolves on a time scale much longer than that of a single collision.

We can now write the rate of the probability change as the difference between the number of particles coming and leaving the given region of phase space around $\mathbf{p}$ by integrating over all $\mathbf{p}_1\mathbf{p}'\mathbf{p}'_1$:

$$\frac{\partial\rho}{\partial t}\ =\ \int(w'\rho'\rho'_1-w\rho\rho_1)\,d\mathbf{p}_1d\mathbf{p}'d\mathbf{p}'_1\,. \tag{24}$$

The scattering probabilities $w\equiv w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)$ and $w'\equiv w(\mathbf{p}',\mathbf{p}'_1;\mathbf{p},\mathbf{p}_1)$ are nonzero only for quartets satisfying the conservation of energy and momentum. We assume that the scattering probabilities are invariant under time reversal which changes $\mathbf{p}\to-\mathbf{p}$ and interchange incoming and outgoing particles:

$$w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)=w(-\mathbf{p}',-\mathbf{p}'_1;-\mathbf{p},-\mathbf{p}_1)\ . \tag{25}$$

We also assume that the medium is invariant with respect to inversion $\mathbf{r},\mathbf{p}\to-\mathbf{r},-\mathbf{p}$, which gives $w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)=w(-\mathbf{p},-\mathbf{p}_1;-\mathbf{p}',-\mathbf{p}'_1)$. Translation invariance makes scattering the same at $\mathbf{r}$ and $-\mathbf{r}$. All three symmetries combined give:

$$w\equiv w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)=w(\mathbf{p}',\mathbf{p}'_1;\mathbf{p},\mathbf{p}_1)\equiv w'\ . \tag{26}$$

Using (27) we transform the second term in (25) and obtain the famous *Boltzmann kinetic equation* (1872):

$$\frac{\partial\rho}{\partial t}=\int w'(\rho'\rho'_1-\rho\rho_1)\,d\mathbf{p}_1d\mathbf{p}'d\mathbf{p}'_1\equiv I\,, \tag{27}$$

H-theorem. Let us now look at the evolution of the entropy of the one-particle distribution satisfying (28):

$$\frac{dS}{dt}=-\int\frac{\partial\rho}{\partial t}\ln\rho\,d\mathbf{p}=-\int I\ln\rho\,d\mathbf{p}\ , \tag{28}$$

The integral (29) contains the integrations over all momenta so we may exploit two interchanges, $\mathbf{p}_1 \leftrightarrow \mathbf{p}$ and $\mathbf{p}, \mathbf{p}_1 \leftrightarrow \mathbf{p}', \mathbf{p}'_1$:

$$\begin{aligned}\frac{dS}{dt} &= \int w'(\rho\rho_1 - \rho'\rho'_1)\ln\rho\, d\mathbf{p}d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}'_1\\ &= \frac{1}{2}\int w'(\rho\rho_1 - \rho'\rho'_1)\ln(\rho\rho_1)\, d\mathbf{p}d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}'_1\\ &= \frac{1}{2}\int w'\rho\rho_1 \ln\frac{\rho\rho_1}{\rho'\rho'_1}\, d\mathbf{p}d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}'_1 \geq 0\ , \qquad (29)\end{aligned}$$

Here we subtracted the integral $\int w'(\rho\rho_1 - \rho'\rho'_1)\, d\mathbf{p}d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}_1/2 = 0$ and used the inequality $x\ln x - x + 1 \geq 0$ with $x = \rho\rho_1/\rho'\rho'_1$.

Even though we use scattering probabilities obtained from mechanics reversible in time, $w(-\mathbf{p}, -\mathbf{p}_1; -\mathbf{p}', -\mathbf{p}'_1) = w(\mathbf{p}', \mathbf{p}'_1; \mathbf{p}, \mathbf{p}_1)$, our use of molecular chaos hypothesis made the kinetic equation irreversible. Equilibrium realizes the entropy maximum and so the distribution must be a steady solution of the Boltzmann equation. Indeed, the collision integral turns into zero by virtue of $\rho_0(\mathbf{p})\rho_0(\mathbf{p}_1) = \rho_0(\mathbf{p}')\rho_0(\mathbf{p}'_1)$, since $\ln\rho_0$ is the linear function of the integrals of motion as was explained in Sect. 1.5. All this is true also for the inhomogeneous equilibrium in the presence of an external force, see Section (5.3) below.

One can look at the transition from (22) to (28) from a temporal viewpoint. N-particle distribution changes during every collision when particles exchange momenta. On the other hand, the single-particle distribution is the average over $N-1$ particles, so changing it requires many collisions. Even though some of these collisions occur in parallel, in a dilute system with short-range interaction, the time between collisions is much longer than the collision time, so the single-particle distribution changes on a much longer scale. In other words, the transition from (22) to (28) is from a fast-changing function to a slow-changing one.

Let us summarize the present state of confusion. The full entropy of the N-particle distribution is conserved. Yet the one-particle entropy grows. Is there a contradiction here? Is not the full entropy a sum of one-particle entropies? The answer ("no" to both questions) requires introduction of a central notion of this course - mutual information - and will be given in Section 3.8 below. For now, a brief statement will suffice: We broke time reversibility and set the arrow of time when we assumed particles uncorrelated before the collision and not after. If one starts from a set of uncorrelated particles and lets them interact, then the interaction will build correlations and the total distribution will change, but the total entropy will not. Since correlations lower the entropy, that must be compensated by the growth of the single-particle entropies. That growth is described by the Boltzmann equation, which is valid for an uncorrelated initial state (and for some time after). Motivation for choosing such an initial state for computing one-particle evolution is that it is most likely in any generic ensemble. Yet that would make no sense to run the Boltzmann equation backwards from a correlated state, which is statistically a very unlikely initial state, since it requires momenta to be correlated in such a way that a definite state is produced after time t. In other words, Boltzmann equation describes at a macroscopic level (of one-particle distribution) not all but most of the microscopic (N-particle) evolutions.

Going a bit ahead of ourselves, we can say that neglecting inter-particles correlations by factorizing the two-particle distribution $\rho_{12} = \rho(\mathbf{q_1}, \mathbf{p_1}; \mathbf{q_2}, \mathbf{p_2}) = \rho_1 \rho_2$ means using incomplete information. This naturally leads to a further increase of uncertainty, that is of entropy. For dilute gases, such a factorization is just the first term of an expansion:

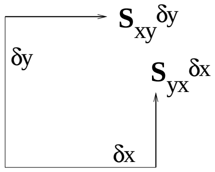
$$\rho_{12} = \rho_1 \rho_2 + \int d\mathbf{q_3} d\mathbf{p_3} J_{123} \rho_1 \rho_2 \rho_3 + \dots \ .$$

Is this a regular expansions? It turns out that such (so-called cluster) expansion is well-defined only for equilibrium distributions. For non-equilibrium distributions, starting from some term (depending on the space dimensionality), all higher terms diverge. The same divergencies take place if one tries to apply the expansion to kinetic coefficients like diffusivity, conductivity or viscosity, which are non-equilibrium properties by their nature. These divergencies can be related to the fact that non-equilibrium distributions do not fill the phase space, as described below in Section 2.3. Obtaining finite results requires re-summation and brings logarithmic terms. As a result, kinetic coefficients and other non-equilibrium properties are non-analytic functions of density. Boltzmann equation looks nice, but corrections to it are ugly, when one deviates from equilibrium. The corrections also violate H-theorem — indeed, dropping all the terms is a part of passing from the Liouville equation to the Boltzmann equation is what leads to the loss of information and entropy growth.

2.2 Phase-space mixing and entropy growth

We have seen that one-particle entropy can grow even when the full N-particle entropy is conserved. But thermodynamics requires the full entropy to grow. To accomplish that, let us return to the full N-particle distribution and recall that we have an incomplete knowledge of the system. That means that we always measure coordinates and momenta within some intervals, i.e. characterize the system not by a point in phase space but by a finite region there. We shall now show that quite general dynamics stretches this finite domain into a very thin convoluted strip whose parts can be found everywhere in the available phase space, say on a fixed-energy surface. The dynamics thus provides a stochastic-like element of mixing in phase space that is responsible for the approach to equilibrium, say to uniform microcanonical distribution. Yet by itself this stretching and mixing does not change the phase volume and entropy. Another ingredient needed is the necessity to continually treat our system with finite precision, which follows from the insufficiency of information. Such consideration is called *coarse graining* and, together with mixing, it is responsible for the irreversibility of statistical laws and for the entropy growth.

The dynamical mechanism of the entropy growth is the separation of trajectories in phase space: trajectories started from a small neighborhood are found farther and farther away as time proceeds. Denote again by $\mathbf{x} = (\mathbf{P}, \mathbf{Q})$ the $6N$-dimensional vector of the position and by $\mathbf{v} = (\dot{\mathbf{P}}, \dot{\mathbf{Q}})$ the velocity in the phase space. The relative motion of two close points, separated by $\mathbf{r}$, is determined by their velocity difference: $\delta v_i \approx r_j \partial v_i/\partial x_j = r_j \sigma_{ij}$. We can decompose the tensor of velocity derivatives into an antisymmetric part (which describes rotation) and a symmetric part $S_{ij} = (\partial v_i/\partial x_j + \partial v_j/\partial x_i)/2$ (which describes deformation). Separation of trajectories and entropy growth are due to deformation, so we focus on S_{ij}. The vector initially parallel to the axis j turns towards the axis i with the angular speed $\partial v_i/\partial x_j$, so that $2S_{ij}$ is the rate of variation of the angle between two initially mutually perpendicular small vectors along i and j axes. In other words, $2S_{ij}$ is the rate with which rectangle deforms into parallelograms:



Arrows in the Figure show the velocities of the endpoints. The symmetric tensor S_{ij} can be always transformed into a diagonal form by an orthogonal transformation (i.e. by the rotation of the axes), so that $S_{ij} = S_i \delta_{ij}$. According to the Liouville theorem, a Hamiltonian dynamics is an incompressible flow in the phase space, so that the trace of the tensor, which is the rate of the

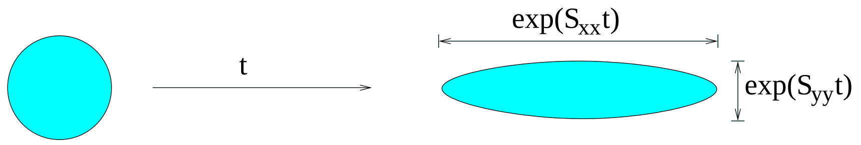


Figure 1: Deformation of a phase-space element by a permanent strain.

volume change, must be zero: $\mathrm{Tr}\,\sigma_{ij} = \sum_i S_i = div\,\mathbf{v} = 0$. That means that some components are positive, some are negative. Positive diagonal components are the rates of stretching and negative components are the rates of contraction in respective directions. Indeed, the equation for the distance between two points along a principal direction has a form: $\dot{r}_i = \delta v_i = r_i S_i$. The solution is as follows:

$$r_i(t) = r_i(0)\exp\left[\int_0^t S_i(t')\,dt'\right] \ . \tag{30}$$

For a time-independent strain, the growth/decay is exponential in time. One recognizes that a purely straining motion converts a spherical element into an ellipsoid with the principal diameters that grow (or decay) in time. Indeed, consider a two-dimensional projection of the initial spherical element i.e. a circle of the radius R at $t = 0$. The point that starts at $x_0, y_0 = \sqrt{R^2 - x_0^2}$ goes into

$$\begin{aligned} & x(t) = e^{S_{11}t}x_0\,, \\ & y(t) = e^{S_{22}t}y_0 = e^{S_{22}t}\sqrt{R^2 - x_0^2} = e^{S_{22}t}\sqrt{R^2 - x^2(t)e^{-2S_{11}t}}\,, \\ & x^2(t)e^{-2S_{11}t} + y^2(t)e^{-2S_{22}t} = R^2\,. \end{aligned} \tag{31}$$

The equation (32) describes how the initial circle turns into the ellipse whose eccentricity increases exponentially with the rate $|S_{11} - S_{22}|$. In a multi-dimensional space, any sphere of initial conditions turns into the ellipsoid defined by $\sum_{i=1}^{6N} x_i^2(t)e^{-2S_it}$ =const.

If our uncertainty about the initial state was confined within a sphere, then the uncertainty about the evolved state is within the ellipsoid. As the system moves in the phase space, both the strain values and the orientation of the principal directions change, so that expanding direction may turn into a contracting one and vice versa. Since we do not want to go into details of dynamics, then we consider such evolution as a kind of random process. The question is whether averaging over all values and orientations gives a zero net separation of trajectories. It may seem counter-intuitive, but an exponential stretching generally persists on average and the majority of trajectories separate. There are two ways to understand that: one in space and another in time[8].

Let us first go with the flow and see separation of trajectories with time. Denote the rate of separation along a given direction $\Lambda_i(t) = \int_0^t S_i(t')dt'/t$. Even when the time average is zero, $\lim_{t\to\infty} \int_0^t S_i(t')dt' = 0$, the average exponent of it is larger than unity (and generally grows with time):

$$\left\langle \frac{r_i(t)}{r_i(0)} \right\rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T dte^{\Lambda_i(t)} \geq 1 \ . \tag{32}$$

This is because the time intervals with positive $\Lambda(t)$ contribute more into the exponent than the intervals with negative $\Lambda(t)$. That follows from the *concavity* of the exponential function. In the simplest case, when Λ is uniformly distributed over the interval $-a < \Lambda < a$, the average Λ is zero, while the average exponent exceeds unity: $(1/2a) \int_a^{-a} e^\Lambda d\Lambda = (e^a - e^{-a})/2a > 1$.

Looking from a spatial perspective, consider the simplest case of a pure strain, which corresponds to an incompressible saddle-point flow in a plane: $v_x = \lambda x$, $v_y = -\lambda y$. We are in the reference frame of the trajectory corresponding to the center, so that $\mathbf{r} = (x, y)$ represents the separation between that trajectory and a close one. Even though $2d$ phase space corresponds to the trivial case of one particle moving along a line, it is of great illustrative value. Also, remember that the Liouville theorem is true in every $p_i - q_i$ plane projection. Here we have one expanding direction and one contracting direction, their rates being equal. The separation vector, $\mathbf{r} = (x, y)$, can look initially at any direction. The evolution of the vector components satisfies the equations $\dot{x} = v_x$ and $\dot{y} = v_y$. Whether the vector is stretched or contracted after some time T depends on its orientation and on T. Since $x(t) = x_0 \exp(\lambda t)$ and $y(t) = y_0 \exp(-\lambda t) = x_0 y_0 / x(t)$ then every trajectory is a hyperbole. A unit vector initially forming an angle φ with the x axis will have its length $[\cos^2\varphi \exp(2\lambda T) + \sin^2\varphi \exp(-2\lambda T)]^{1/2}$ after time T. The vector is stretched if $\cos\varphi \geq [1 + \exp(2\lambda T)]^{-1/2} < 1/\sqrt{2}$, i.e. the fraction of stretched directions is larger than half. When along the motion all orientations are equally probable, the net effect is stretching, increasing with the persistence time T.

[8] "Time and space are modes by which we think and not conditions in which we live" A. Einstein

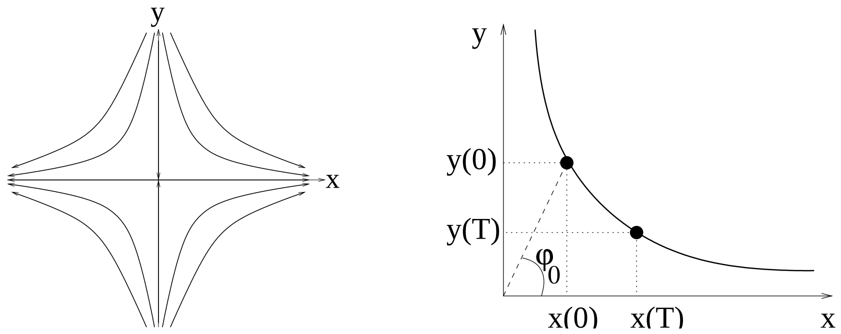


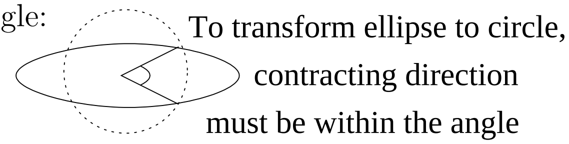
Figure 2: Left panel: streamlines of a saddle-point flow. Right panel: motion down a streamline. For $\varphi = \varphi_0$ the initial and final points are symmetric relative to the diagonal: $x(0) = y(T)$ and $y(0) = x(T)$. If $\varphi < \varphi_0 = \arccos[1 + \exp(2\lambda T)]^{-1/2} > \pi/4$, the distance from the origin increases.

The net stretching and separation of trajectories is formally proved in mathematics by considering a random strain matrix $\hat{\sigma}(t)$ and the transfer matrix $\hat{W}$ defined by $\mathbf{r}(t) = \hat{W}(t, t_1)\mathbf{r}(t_1)$. It satisfies the equation $d\hat{W}/dt = \hat{\sigma}\hat{W}$. The Liouville theorem $\mathrm{tr}\,\hat{\sigma} = 0$ means that $\det \hat{W} = 1$. The modulus $r(t)$ of the separation vector may be expressed via the positive symmetric matrix $\hat{W}^T\hat{W}$. The main result (Furstenberg and Kesten 1960; Oseledec, 1968) states that in almost every realization $\hat{\sigma}(t)$, the matrix $\frac{1}{t}\ln \hat{W}^T(t, 0)\hat{W}(t, 0)$ tends to a finite limit as $t \to \infty$. In particular, its eigenvectors tend to d fixed orthonormal eigenvectors $\mathbf{f}_i$. Geometrically, that precisely means than an initial sphere evolves into an elongated ellipsoid at later times. The limiting eigenvalues

$$\lambda_i = \lim_{t\to\infty} t^{-1} \ln |\hat{W}\mathbf{f}_i| \tag{33}$$

define the so-called Lyapunov exponents, which can be thought of as the mean stretching rates. The sum of the exponents is the mean volume growth rate, which is zero due to the Liouville theorem. As long as there is no special degeneracy , which makes all the exponents identically zero, there exists at least one positive exponent which gives stretching. Therefore, as time increases, the ellipsoid is more and more elongated and it is less and less likely that the hierarchy of the ellipsoid axes will change. Mathematical lesson to learn is that multiplying N random matrices with unit determinant (recall that determinant is the product of eigenvalues), one generally gets some eigenvalues growing and some decreasing exponentially with N. It is also worth remembering that in a random flow there is always a probability for two trajectories to come closer. That probability decreases with time but it is finite for any finite time. In other words, majority of trajectories separate but some approach. The separating ones provide for the exponential growth of positive moments of the distance: $E(a) = \lim_{t\to\infty} t^{-1} \ln\left[\langle r^a(t)/r^a(0)\rangle\right] > 0$ for $a > 0$. However, approaching trajectories have $r(t)$ decreasing, which guarantees that the moments with sufficiently negative a also grow. Mention without proof that $E(a)$ is a concave function, which evidently passes through zero, $E(0) = 0$. It must then have another zero which for isotropic random flow in d-dimensional space can be shown to be $a = -d$.

The probability to find a ball turning into an exponentially stretching ellipse thus goes to unity as time increases. The physical reason for it is that substantial deformation appears sooner or later. To reverse it, one needs to contract the long axis of the ellipse, that is the direction of contraction must be inside the narrow angle defined by the ellipse eccentricity, which is less likely than being outside the angle:



This is similar to the argument about the irreversibility of the Boltzmann equation in the previous subsection. Randomly oriented deformations on average continue to increase the eccentricity.

Armed with the understanding of the exponential stretching, we now return to the dynamical foundation of the second law of thermodynamics. We assume that our finite resolution does not allow us to distinguish between the states within some square in the phase space. That square is our ”grain” in coarse-graining. In the figure below, one can see how such black square of initial conditions (at the central panel) is stretched in one (unstable) direction and contracted in another (stable) direction so that it turns into a long narrow strip (left and right panels). Later in time, our resolution is still restricted - rectangles in the right panel show finite resolution (this is coarse-graining). Viewed with such resolution, our set of points occupies larger phase volume at $t = \pm T$ than at $t = 0$. Larger phase volume corresponds to larger entropy.

Time reversibility of any trajectory does not contradict time-irreversible filling of the space by the set of trajectories considered with a finite resolution. By reversing time we exchange stable and unstable directions (i.e. those of contraction and expansion), but the fact of space filling persists. We see from the figure that the volume and entropy increase both forward and backward in time. And yet our consideration does provide for time arrow: If we already

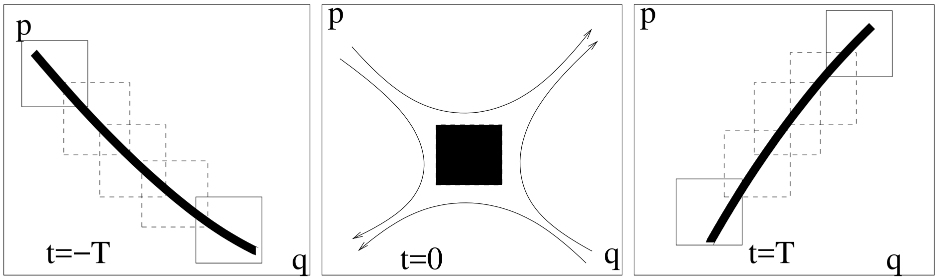
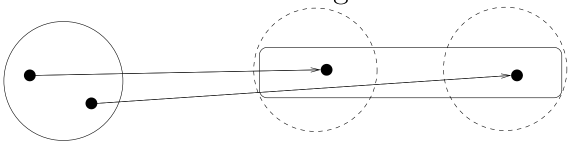


Figure 3: Increase of the phase volume upon stretching-contraction and coarse-graining. Central panel shows the initial state and the velocity field.

observed an evolution that produces a narrow strip, then its time reversal is the contraction into a ball; but if we consider a narrow strip as an initial condition, it is unlikely to observe a contraction because of the narrow angle mentioned above. Therefore, being shown two movies, one with stretching, another with contraction we conclude that with probability close (but not exactly equal!) to unity the first movie shows the true sequence of events, from the past to the future.

When the possible occupied region expands, the entropy grows (as the logarithms of the volume). If initially our system was within the phase-space volume ϵ^{6N}, then its density was $\rho_0 = \epsilon^{-6N}$ inside and zero outside. After stretching to some larger volume $e^{\lambda t}\epsilon^{6N}$ the entropy $S = -\int \rho \ln \rho d\mathbf{x}$ has increased by λt. The positive Lyapunov exponent λ determines the rate of the entropy growth. If in a d-dimensional space there are k stretching and $d - k$ contracting directions, then contractions eventually stabilize at the resolution scale, while expansions continue. Therefore, the volume growth rate is determined by the sum of the positive Lyapunov exponents $\sum_{i=1}^{k} \lambda_i$.

We shall formally define information later, here we use everyday intuition about it (as diminishing uncertainty) to briefly discuss our flow from this perspective. Consider an ensemble of systems having close initial positions within our finite resolution. In a flow with positive Lyapunov exponents, with time we loose our ability to predict where it goes. This loss of information is determined by the growth of the available phase volume, that is of the entropy. But we can look backwards in time and ask where the points come from. If we consider two points along a stretching direction, we can with confidence predict that they were closer before. During some time in the past, they were hidden inside the resolution circle, but they separate with time beyond the resolution and can now be distinguished:



Moreover, as time proceeds, we learn more and more about the initial locations of the points. The acquisition rate of such information about the past is again the sum of the positive Lyapunov exponents and is called the Kolmogorov-Sinai entropy. As time lag from the present moment increases, we can say less and less where we shall be and more and more where we came from. It illustrates the Kierkegaard's remark that the irony of life is that it is lived forward but understood backwards.

After the strip length reaches the scale of the velocity change (when one already cannot approximate the phase-space flow by a linear profile $\hat{\sigma}r$), strip starts to fold because rotation (which we can neglect for a ball but not for a long strip) is different at different parts of the strip. Still, however long, the strip continues locally the exponential stretching. Eventually, one can find the points from the initial ball everywhere which means that the flow is mixing, also called ergodic. Formal definition is that the flow is called ergodic in the domain if the trajectory of almost every point (except possibly a set of zero volume) passes arbitrarily close to every other point. An equivalent definition is that there are no finite-volume subsets of the domain invariant with respect to the flow except the domain itself. Ergodic flow on an energy surface in the phase space provides for a micro-canonical distribution (i.e. constant), since time averages are equivalent to the average over the surface. While we can prove ergodicity only for relatively simple systems, like the gas of hard spheres, we believe that it holds for most systems of sufficiently general nature (that vague notion can be made more precise by saying that the qualitative behavior is insensitive to small variations of the system's microscopic parameters).

One can think of any Hamiltonian dynamics as a map of phase space into itself. Appendix 8.3 describes a toy model of such a map, which is of great illustrative value for the applications of chaos theory to statistical mechanics.

Two concluding remarks are in order. First, the notion of an exponential separation of trajectories put an end to the old dream of Laplace to be able to predict the future if only all coordinates and momenta are given. Even if we were able to measure all relevant phase-space initial data, we can do it only with a finite precision ϵ. However small is the indeterminacy in the data, it is amplified exponentially with time so that eventually $\epsilon \exp(\lambda T)$ is large and we cannot predict the outcome. Mathematically speaking, limits $\epsilon \to 0$ and $T \to \infty$ do not commute. Second, the above arguments did not use the usual mantra of thermodynamic limit, which means that even the systems with a small number of degrees of freedom need statistics for their description at long times if their dynamics has a positive Lyapunov exponent (which is generic) - this is sometimes called *dynamical chaos*.[9]

Common lesson from the last two sections is that full knowledge persists while partial knowledge dissipates. If you know everything then this knowledge stays with you, which is a Liouville theorem. But if your knowledge is incomplete - either because you study only part of your degrees of freedom (Bolzmann) or because of finite precision (coarse-graining) - then your degree of uncertainty generally increases with time.

2.3 Entropy decrease and non-equilibrium fractal measures

As we have seen in the previous section, if we have indeterminacy in the data or consider an ensemble of systems, then an incompressible flow of a Hamiltonian dynamics effectively mixes and makes the distribution uniform in the phase space. Evolution is Hamiltonian for isolated systems, which conserve their integrals of motion, so that the distribution is uniform over the respective surface of constant integrals. In particular, dynamical chaos justifies micro-canonical distribution, uniform over the energy surface.

But what if the dynamics is non-Hamiltonian, that is the Liouville theorem is not valid? The flow in the phase space is then generally compressible. The simplest non-conservative effect is dissipation of kinetic energy, which shrinks all momenta and thus decreases the phase volume. We are interested, however, in a non-equilibrium steady state where we keep the total energy non-decreasing. For example, to compensate for the loss of momentum of the particles with the dissipation rates γ_i, we act on them by external forces

[9] As a student, I've participated (mostly as a messenger) in the discussion on irreversibility between Zeldovich and Sinai. I remember Zeldovich asking why coarse-graining alone (already introduced by Boltzmann) is not enough to explain irreversibility. Why one needs dynamical chaos to justify what one gets by molecular chaos? I believe that Sinai was right promoting separation of trajectories. It replaces arbitrary assumptions by clear demonstration from first principles and expands statistical approach to systems even with few degrees of freedom.

f_i, so that the equations of motion take the form: $\dot{p}_i = f_i - \gamma_i p_i - \partial H/\partial q_i$, $\dot{q}_i = \partial H/\partial p_i$, which gives $div\,\mathbf{v} = \sum_i(\partial f_i/\partial p_i - \gamma_i)$. When the system is in a thermostat, the forces f_i are due to random kicks, which are short-correlated comparing to times of order γ_i^{-1}. Such forces are in the detailed balance with the dissipation: after averaging over the short correlation time, $\langle \partial f_i/\partial p_i \rangle = \gamma_i$ for every i, so that $div\,\mathbf{v} \equiv 0$. For an example, see the consideration of a Brownian particle in Appendix 8.8, particularly (175).

Let us consider now a generic environment, where forces are correlated and $div\,\mathbf{v} \neq 0$ during finite intervals. Such phase-space flows create quite different distributions, since the probability density changes along a flow: $d\rho/dt = -\rho div\,\mathbf{v}$. That produces entropy,

$$\frac{dS}{dt} = \int \rho(\mathbf{r},t) div\,\mathbf{v}(\mathbf{r},t)\,d\mathbf{r} = \langle div\,\mathbf{v} \rangle\,. \tag{34}$$

The entropy-production rate is equal to the mean local expansion rate of the phase-volume. If the system does not on average heat or cool (expand or contract), then the whole phase volume does not change. That means that the volume integral of the local expansion rate is zero: $\int div\,\mathbf{v}\,d\mathbf{r} = 0$. Yet for a non-uniform density, the entropy is not the (Boltzmann) log of the phase volume but the (Gibbs) *mean* log of the inverse density, $S(t) = -\langle \ln \rho \rangle = -\int \rho(\mathbf{r},t) \ln \rho(\mathbf{r},t)\,d\mathbf{r}$, whose derivative (35) is non-zero because of correlations between ρ and $div\,\mathbf{v}$. Indeed, ρ is on average smaller in the expanding regions where $div\,\mathbf{v} > 0$. That means that the entropy production rate (35) is non-positive and the entropy) decreases. Under the only condition of normalization, a uniform distribution has a maximal entropy. Therefore, the entropy decrease means that the distribution is getting more and more non-uniform in the phase space.

Of course, integrating density over all the space gives unity at any time: $\int \rho(\mathbf{r},t)\,d\mathbf{r} = 1$. Let us now switch focus from space to time and consider the density of an arbitrary fluid element with the coordinate $\mathbf{r}(t)$, which satisfies $d\mathbf{r}/dt = \mathbf{v}$ and $\mathbf{r}(0) = \mathbf{r}_0$. The density then evolves as follows:

$$\frac{\rho(\mathbf{r}(t),t)}{\rho(\mathbf{r}_0,0)} = \exp\left[-\int_0^t div\,\mathbf{v}(\mathbf{r}(t'),t')\,dt'\right] = e^{C(t)}\,. \tag{35}$$

If the expansion rate in the flow reference frame, $s(t) = div\,\mathbf{v}(\mathbf{r}(t),t)$, is a random function with a finite correlation time τ, then its integral $C = \int_0^t div\,\mathbf{v}(t')\,dt'$ at $t/\tau = N \gg 1$ can be broken into a sum of many uncorrelated random numbers with a zero mean and some variance Δ. According to the Central Limit Theorem, see Section 8.2, the statistics of such a sum is Gaussian with a zero mean and the variance linearly growing with time: $\mathcal{P}(C) \propto e^{-C^2/2\Delta N}$. We then obtain for the average over all possible trajectories: $\overline{\exp(C)} = \int \mathcal{P}(C) e^C dC \propto e^{N\Delta/2}$. Therefore, for a generic random flow the density of most fluid elements must grow non-stop as they move. The reason is again the concavity of the exponential function, as in (33): if the mean is zero, the mean exponent generally exceeds unity.

Since the total measure is conserved, growth of density at some places must be compensated by its decrease in other places, so that the distribution is getting more and more non-uniform, which decreases the entropy. Looking at the phase space, one sees it more and more emptied with the density concentrated asymptotically in time on a small subset. That is opposite to the mixing by Hamiltonian incompressible flow.

If the density of any fluid element on average grows, then its volume decreases. In particular, for a spatially smooth flow, the long-time Lagrangian average (along the flow) of the volume change rate,

$$\overline{\frac{dS}{dt}} = \overline{div\,\mathbf{v}} = \lim_{t\to\infty} \frac{1}{t}\int_0^t div\,\mathbf{v}(t')\,dt' = \sum_i \lambda_i\,,$$

is a sum of the Lyapunov exponents, which is then non-positive (in distinction from an instantaneous average over space, which is zero at any time: $\int div\,\mathbf{v}\,d\mathbf{r} = 0$).

It is important that we allowed for a compressibility of a phase-space flow $\mathbf{v}(\mathbf{r}, t)$ but did not require its irreversibility. Indeed, even if the system is invariant with respect to $t \to -t$, $\mathbf{v} \to -\mathbf{v}$, the entropy production rate is generally non-negative and the sum of the Lyapunov exponents is non-positive for the same simple reason that contracting regions have more measure and give higher contributions. Backwards in time the measure also concentrates, only on a different set.

Let us show that the density could concentrate on a fractal set in a spatially smooth random compressible flow. One defines the (box-counting) dimension of a set as follows:

$$d_f = \lim_{\epsilon\to 0} \frac{\ln N(\epsilon)}{\ln(L/\epsilon)}\ , \tag{36}$$

Here $N(\epsilon)$ is the number of boxes of side ϵ needed to cover the set of the size L, see the Figure 4.

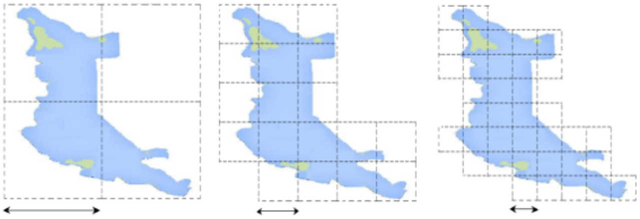


Figure 4: Number of covering squares depends on their size: $N(1/2) = 4$, $N(1/5) = 18$, $N(1/7) = 37$.

Consider a two-dimensional phase space with one positive Lyapunov exponent λ_+ and one negative Lyapunov exponent λ_-. After time t, a square having initial side $\delta \ll L$ will be stretched into a long thin strip of length $\delta \exp(t\lambda_+)$ and width $\delta \exp(t\lambda_-)$. To cover contracting direction, we choose $\epsilon = \delta \exp(t\lambda_-)$, then $N(\epsilon) = \exp[t(\lambda_+ - \lambda_-)]$, so that the dimension is

$$d_f = 1 + \frac{\lambda_+}{|\lambda_-|} \ , \tag{37}$$

Since $|\lambda_-| \geq \lambda_+$, then the dimension is between 1 and 2. The set is smooth in the expanding direction and fractal in the contracting direction, which respectively gives two terms in (38). How density concentrates on a fractal set in a random compressible flow is illustrated by a toy model presented in Appendix 8.3.

General (Kaplan-Yorke) conjecture is that $d_f = j + \sum_{i=1}^{j} \lambda_i / \lambda_{j+1}$, where j is the largest number for which $\sum_{i=1}^{j} \lambda_i \geq 0$ and $\sum_{i=1}^{j+1} \lambda_i < 0$. For incompressible flows, $j = d$.

Fractalization of the measure proceeds until the coarse-graining stops it. In distinction from the incompressible flow, coarse-graining at a small scale ϵ does not make the distribution uniform, but it makes the entropy finite: $S = \ln N(\epsilon) = d_f \ln(L/\epsilon)$. An equilibrium uniform (microcanonical) distribution in d-dimensional phase space has the entropy $S_0 = d \ln(L/\epsilon)$; a non-equilibrium steady state generally has a lower dimensionality $d_f < d$ with a lower entropy.

We thus see that for smooth dynamical systems, both temporal and spatial properties of the entropy are determined by the Lyapunov exponents. Entropy dependence on time (both forward and backward) is governed by the Kolmogorov-Sinai entropy, which is the sum of the positive Lyapunov exponents. Entropy dependence on spatial resolution is determined by the dimensionality.

To conclude this Chapter, let us appreciate the dramatic difference between the entropy growth described in Section 2.2 and the entropy decay described in the present Section (see also Appendix 8.3 for the examples of both). In the former, phase-space flows were area-preserving and the volume growth of an element was due to a finite resolution, which stabilized the size in the contracting direction, so that the mean volume growth rate was solely due to stretching directions and thus equal to the sum of the positive Lyapunov exponents, as described in Section 2.2. On the contrary, the present section deals with compressible flows. Relation between compressibility and non-equilibrium is natural: to make a system non-Hamiltonian one needs to pump energy into some degrees of freedom and absorb it from other degrees of freedom to keep a steady state, which corresponds to expansion and contraction of the momentum part of the phase-space. That decreases entropy by creating more inhomogeneous distributions. The mean rate of the entropy decay is the sum of all the Lyapunov exponents, which is non-positive since contracting regions contain more trajectories and contribute more than expanding regions. Long-time net contraction of a fluid element and respective entropy decay is the analog of the second law of thermodynamics: to deviate a system from equilibrium, one needs to lower its entropy until the resolution limit is reached.

Looking back at this Chapter, it is a good time to reflect on the complementarity of determinism and randomness expressed in terms "statistical mechanics" (19th century) and "dynamical chaos" (20th century). What shall we have in the 21st century: predictable uncertainty, multi-version reality?

3 Basics of Information Theory

This Chapter presents an elementary introduction into the information theory from the viewpoint of a natural scientist. It re-tells the story of statistical physics using a different language, which lets us see the Boltzmann and Gibbs entropies in a new light. Here we switch from continuous thinking in terms of phase-space flows to discrete combinatoric manipulations. What I personally like about the information viewpoint is that it erases paradoxes and makes the second law of thermodynamics trivial. It also allows us to see generality and commonality in the approaches (to partially known systems) of physicists, engineers, computer scientists, biologists, brain researchers, social scientists, market speculators, spies and flies. We shall see how the same tools used in setting limits on thermal engines are used in setting limits on communications, measurements and learning (which are essentially the same phenomena). The main mathematical tool exploits universality appearing upon summing many independent random numbers.

The central idea developed in this Chapter is that information lowers uncertainty. A convenient way to quantify it is by the number of questions whose answers together eliminate the uncertainty. If we are uncertain about the events with a priori equal probabilities, the number of such questions is a logarithm of the number n of possible outcomes, which is the Boltzmann entropy. To locate one out of n equally probable objects, one needs $\log_2 n$ yes-no questions. Alternatively, one can say that the information quantifies the degree of surprise: the larger the possible number of outcomes, the more surprising is any one of them. If we know the probabilities p_i of the events, then the surprise $\log_2 p_i$ is larger for the less probable ones, while the information rate per answer on average is equal to the Gibbs entropy, $S = -\sum_i p_i \log_2 p_i$ bits. That follows from the fact (shown in the next section) that the number of typical N-sequences of outcomes grows with N as 2^{NS}, so that any such sequence brings $\log_2 2^{NS} = NS$ bits, that is S bits per outcome on average (maybe the entropy is denoted S because it quantifies surprise).

But what if the answers are not completely reliable? In other words, we have an imperfect channel whose output A specifies the event (input) B_j not completely, but with some remaining uncertainty, which is characterized by the conditional entropy $S(B|A)$. The information received is then equal to $I(A,B) = S(B) - S(B|A)$ called the mutual information. Applications of the universal notions of entropy and mutual information widen fast: from physics, communications and computations to brain research, artificial intelligence and quantum computing, as will be described in the next Chapters.

3.1 Information as a choice

”Information is the resolution of uncertainty.”
C Shannon 1948

We want to know in which of n boxes a candy is hidden, that is we are faced with a choice among n equal possibilities. How much information do we need to get the candy? Let us denote the missing information by $I(n)$. Clearly, $I(1) = 0$, and we want the information to be a monotonically increasing[10] function of n. If we have several independent problems then information must be additive. For example, consider each box to have m compartments. To know in which from mn compartments is the candy, we need to know first in which box and then in which compartment inside the box: $I(nm) = I(n) + I(m)$. Now, we can write (Fisher 1925, Hartley 1927, Shannon 1948)

$$I(n) = I(e) \ln n = k \ln n \tag{38}$$

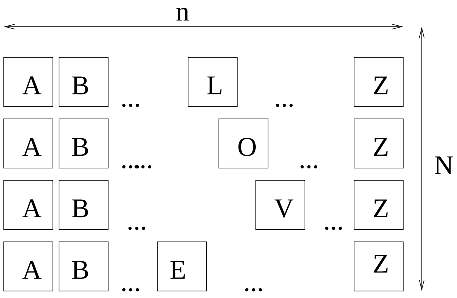
That information must be a logarithm is clear also from obtaining the missing information by asking the sequence of questions in which half we find the box with the candy, one then needs $\log_2 n$ of such questions and respective one-bit answers. If we measure information in binary choices or *bits* (abbreviation of "binary digits"), then $I(n) = \log_2 n$, that is $k^{-1} = \ln(2)$. To arrive at destination via the road with N forks one needs N bits, while via a street with M intersections $M \log_2 3$ bits, since there are three possible way at an intersection.

We can easily generalize the definition (39) for non-integer rational numbers by $I(n/l) = I(n) - I(l)$ and for all positive real numbers by considering limits of the series and using monotonicity. So the message carrying the single number of the lucky box with the candy brings the information $k \ln n$.

We used to think of information received through words and symbols. Essentially, it is always about in which box the candy is. Indeed, if we have an alphabet with n symbols then every symbol we receive is a choice out of n and brings the information $k \ln n$. That is n symbols are like n boxes. If symbols come independently then the message of the length N can potentially be one of n^N possibilities so that it brings the information $kN \ln n$. To convey the same information by smaller alphabet, one needs longer message. If all the 26 letters of the English alphabet were used with the same frequency then the word "love" would bring the information equal to $4 \log_2 26 \approx 4 \cdot 4.7 = 18.8$ bits. Here and below we assume that the receiver has no other prior knowledge on subjects like correlations between letters (for instance, everyone who knows English, can infer that there is only one four-letter word which starts with “lov...” so the last letter brings zero information

[10] The messages "in box 1 out of 2" and "in box 1 out of 22" bring the same candy but not the same amount of information.

for such people).



In reality, every letter brings on average even less information than $\log_2 26$ since we *know* that letters are used with different frequencies. Indeed, consider the situation when there is a probability p_i assigned to each letter (or box) $i = 1, \dots, n$. It is then clear that different letters bring different information. Let us evaluate the *average* information per symbol in a long message. To average, we consider the limit $N \to \infty$, then we know that the i-th letter appears Np_i times *in a typical sequence*, that is we know that we receive the first alphabet symbol Np_1 times, the second symbol Np_2 times, etc. What we didn't know and what any message of the length N brings is the order in which different symbols appear. Total number of orders (the number of different typical sequences) is equal to $N!/\Pi_i(Np_i)!$, and the information that we obtained from a string of N symbols is the logarithm of that number:

$$I_N = k \ln \frac{N!}{\Pi_i(Np_i)} \approx k\Big(N \ln N - \sum_i Np_i \ln Np_i\Big) = -Nk \sum_i p_i \ln p_i \,. \quad (39)$$

The mean information per symbol coincides with the Gibbs entropy (16):

$$S(p_1 \dots p_n) = \lim_{N\to\infty} I_N/N = -k \sum_{i=1}^{n} p_i \ln p_i \ . \quad (40)$$

Alternatively, one can derive (41) without any mention of randomness. Consider again n boxes and denote m_i the number of compartments in the box number i. When each compartment can be chosen independently of the box it is in, the i-th box is chosen with the frequency $p_i = m_i/\sum_{i=1}^n m_i = m_i/M$, that is a given box is chosen more frequently if it has more compartments. The information on a specific compartment is a choice out of M and brings information $k \ln M$. That information must be a sum of the information about the box I_n plus the information about the compartment,

$\ln m_i$, summed over the boxes: $k\sum_{i=1}^n p_i \ln m_i$. That gives the information I_n about the box (letter) as the difference:

$$I_n = k\ln M - k\sum_{i=1}^n p_i \ln m_i = k\sum_{i=1}^n p_i \ln M - k\sum_{i=1}^n p_i \ln m_i = -k\sum_{i=1}^n p_i \ln p_i = S\,.$$

A little more formally, one can prove that (41) is the only measure of uncertainty that is a continuous function of p_i, symmetric with respect to their permutations, and satisfies the inductive relation:

$$S(p_1, p_2, p_3 \dots p_n) = S(p_1 + p_2, p_3 \dots p_n) + (p_1 + p_2) S\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)\,.$$

That relation comes from considering a subdivision: first, receive the information whether one of the first two possibilities appeared, second, distinguish between 1 and 2.

Asymptotic equipartition. Let us look at a given sequence of symbols $y_1, \dots, y_N$ and ask: how probable it is? Can we answer this blatantly self-referential question without seeing other sequences?

Yes, we can, if the sequence is long enough and we know that the symbols are independently chosen. We use the law of large numbers which states that the sum of N random numbers fast approaches N times the mean value as N grows. To use the law we need to find numbers to sum. Since the symbols are independent, then the probability of any sequence is the product of probabilities and the logarithm of the probability is the sum: $N^{-1}\ln P(y_1, \dots, y_N) = N^{-1}\sum_{i=1}^N \ln P(y_i)$. For large N, it is the mean logarithm of the distribution, that is the entropy. Of which distribution? The probabilities of independent symbols depend not on the position i, but on which symbol from our alphabet, $y^1, y^2, \dots, y^n$ is there. Let us denote $p(y^j)$ the probabilities of different symbols. Then in the limit of large N we have $N^{-1}\sum_{i=1}^N \ln P(y_i) = \sum_{j=1}^n p(y^j)\ln p(y^j)$. But how to find $p(y)$? For a sufficiently long sequence, we assume that the frequencies of different symbols in our sequence give the true probabilities of these symbols. In other words, we assume that the sequence is typical. Then the log of probability converges to minus N times the entropy of y:

$$\frac{1}{N}\ln P(y_1, \dots, y_N) \to \sum_{j=1}^n p(y^j)\ln p(y^j) = \langle \ln p(y) \rangle = -S(y)\,. \qquad (41)$$

We then state that the probability of the typical sequence decreases with N exponentially: $P(y_1,\dots,y_N) = exp[-NS(y)]$. That probability is independent of the values $y_1,\dots,y_N$, that is the same for all typical sequences. We thus found that the best approximation for $P(y_1,\dots,y_N)$ is a uniform (microcanonical!) distribution. Equivalently, we can state that the number of typical sequences grows with N exponentially, and the entropy sets the rate of growths. That focus on typical sequences, which all have the same (maximal) probability, is known as asymptotic equipartition and formulated as *"almost all events are almost equally probable"*.

In physics, asymptotic equipartition is used, for instance, when we claim that the Boltzmann entropy is equivalent to the Gibbs entropy for systems whose energy is separable into independent parts in the thermodynamic limit (number of particles is an analog of a string length N). Like we argued for equivalence of energies in Section 1.4, we consider the microcanonical distribution taken at the energy equal to the mean energy of the canonical distribution (the typical set of the canonical ensemble). Then the Boltzmann entropy of such microcanonical distribution is equal to the entropy of the canonical distribution in the thermodynamic limit.

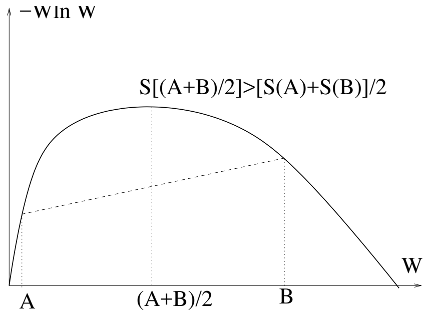
Now we recognize in (41) the asymptotic equipartition: N-string brings the information, which is the log of the number of typical strings: $I = NS$. Note that when $n \to \infty$ then (39) diverges while (41) may well be finite.

The mean information (41) is zero for delta-distribution $p_i = \delta_{ij}$; it is generally less than the information (39) and coincides with it only for equal probabilities, $p_i = 1/n$, when the entropy is maximum. Indeed, equal probabilities we ascribe when there is no extra information, i.e. in a state of maximum ignorance. In this state, a message brings maximum information per symbol; any prior knowledge can reduce the information. Mathematically, the property

$$S(1/n,\dots,1/n) \ge S(p_1 \dots p_n) \tag{42}$$

is called convexity. It follows from the fact that the function of a single variable $s(p) = -p \ln p$ is strictly **concave** since its second derivative, $-1/p$, is everywhere negative for positive p. For any concave function, the average over the set of points p_i is less or equal to the function at the average value (so-called Jensen inequality):

$$\frac{1}{n}\sum_{i=1}^{n} s\left(p_i\right) \le s\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) \ . \tag{43}$$



From here one gets the entropy inequality:

$$S(p_1 \dots p_n) = \sum_{i=1}^{n} s\left(p_i\right) \leq ns\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) = ns\left(\frac{1}{n}\right) = S\left(\frac{1}{n}, \dots, \frac{1}{n}\right) . \tag{44}$$

The relations (44-45) can be proven for any concave function. Indeed, the concavity condition states that the linear interpolation between two points a, b lies everywhere below the function graph: $s(\lambda a + b - \lambda b) \geq \lambda s(a) + (1 - \lambda)s(b)$ for any $\lambda \in [0, 1]$, see the Figure. For $\lambda = 1/2$ it corresponds to (44) for $n = 2$. To get from $n = 2$ to arbitrary n we use induction. For that end, we choose $\lambda = (n-1)/n$, $a = (n-1)^{-1}\sum_{i=1}^{n-1} p_i$ and $b = p_n$ to see that

$$\begin{aligned} s\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) &= s\left(\frac{n-1}{n}(n-1)^{-1}\sum_{i=1}^{n-1} p_i + \frac{p_n}{n}\right) \\ &\geq \frac{n-1}{n} s\left((n-1)^{-1}\sum_{i=1}^{n-1} p_i\right) + \frac{1}{n} s\left(p_n\right) \\ &\geq \frac{1}{n}\sum_{i=1}^{n-1} s\left(p_i\right) + \frac{1}{n} s\left(p_n\right) = \frac{1}{n}\sum_{i=1}^{n} s\left(p_i\right) . \end{aligned} \tag{45}$$

In the last line we used the truth of (44) for $n - 1$ to prove it for n.

You probably noticed that (39) corresponds to the microcanonical Boltzmann entropy (10) giving information/entropy as a logarithm of the number of states, while (41) corresponds to the canonical Gibbs entropy (16) giving it as an average. An advantage of Gibbs-Shannon entropy (41) is that it is defined for arbitrary distributions, not necessarily equilibrium.

3.2 Communication Theory

Here we start learning how to treat everything as a message. After we learnt, what information messages bring on average, we are ready to discuss the best ways to transmit them. That brings us to the Communication Theory, which is interested in two key issues, speed and reliability:

i) How much can a message be compressed; i.e., how redundant is the information? In other words, what is the maximal rate of transmission in bits per symbol?

ii) At what rate can we communicate reliably over a noisy channel; i.e., how much redundancy must be incorporated into a message to protect against errors?

Both questions concern redundancy – how unexpected is every letter of the message, on the average. Entropy quantifies redundancy. We have seen that a communication channel transmitting independent symbols on average transmits one unit of the information (41) per symbol. Receiving letter (box) number i through a binary channel (transmitting ones and zeros)[11] brings information $\log_2(1/p_i) = \log_2 M - \log_2 m_i$ bits. Indeed, the remaining choice (missing information) is between m_i compartments. In other words, we may say that the information measures the degree of surprise: less frequent events are more surprising. Less probable symbols bring larger information content, but they happen more rarely. The entropy $-\sum_{i=a}^{z} p_i \log_2 p_i$ is the mean information content per letter, that is the mean number of 0 or 1 needed to encode one letter of an alphabet.

So the entropy is the mean rate of the information transfer, since it is the mean growth rate of the number of typical sequences. What about the maximal rate of the information transfer? Following Shannon, we answer that question statistically, which makes sense in the limit of very long messages, when one can focus on typical sequences, as we did in the previous section deriving (40,42). Consider for simplicity a message of N bits, where 0 comes with probability $1-p$ and 1 with probability p. To compress the message to a shorter string of letters that conveys essentially the same information it suffices to choose a code that treats effectively the *typical* strings — those that contain $N(1-p)$ zeroes and Np ones. The number of such strings is given by the binomial C_{Np}^{N} which for large N is $2^{NS(p)}$, where $S(p) = -p\log_2 p - (1-p)\log_2(1-p)$. The strings differ by the order of appearance of 0 and 1. To distinguish between these $2^{NS(p)}$ messages, we encode any one using a binary string with lengths starting from one and ending at $NS(p)$ bits. For example, we encode by two one-bit words the two messages where all Np ones are together either at the beginning (followed by all $N(1-p)$ zeroes) or

[11] Binary code is natural both for signals (present-absent) and for logic (true-false).

at the end (preceded by all the zeroes). Then we encode by the four two-bit words the messages with one hole, etc. The maximal word length $NS(p)$ is less than N, since $0 \le S(p) \le 1$ for $0 \le p \le 1$. In other words, to encode all 2^N sequences we need words of N bits, but to encode all typical sequences, we need only words up to $NS(p)$ bits. We indeed achieve compression with the sole exception of the case of equal probability where $S(1/2) = 1$. True, the code must include a bit more of longer codewords to represent atypical messages. We then use longer and longer codewords for less and less probable sequences. In the limit of large N the chance of their appearance and their contribution to the rate of transmission gets negligible. Therefore, entropy sets both the mean and the maximal rate in the limit of long sequences. It gives the transfer rate of information when all the redundancy has been squeezed out.

The notion of typical messages in the limit $N \to \infty$ is an information-theory analog of ensemble equivalence in the thermodynamic limit. You may find it bizarre that one uses randomness in treating information communications, where one usually transfers non-random meaningful messages. One of the reasons is that encoding program does not bother to "understand" the message, and treats it as random. Draining the words of meaning is necessary for devising universal communication systems.

Maximal rate of transmission corresponds to the shortest mean length of the codeword. If we encode n equally probable objects by an alphabet with q symbols, the mean codeword cannot be shorter than $\log n/\log q = log_q n$. For example, to encode $n = 4$ bases of the genetic code by bits ($q = 2$) we need at least two-letter words. If we know that the source has the probability distribution $p(i)$, $i = 1, \ldots, n$, then we can use this information to shorten the mean codeword thus increasing the rate. Indeed, the entropy is now lower. Shannon proved that the shortest mean length of the codeword ℓ is bounded by

$$-\sum_i p(i) \log_q p(i) \le \ell < -\sum_i p(i) \log_q p(i) + 1 \,. \tag{46}$$

Of course, not any encoding guarantees the shortest mean codeword and the maximal rate of transmission. Designating sequences of the same length to objects with different probabilities is apparently sub-optimal. Inequality (45) quantifies that. To make the mean word length shorter and achieve signal compression in the limit of long messages, one codes frequent objects by short sequences and infrequent ones by more lengthy combinations - lossless compressions like zip, gz and gif work this way.

Consider a fictional creature whose DNA contains four bases A,T,C,G occurring with probabilities p_i listed in the table:

Symbol	p_i	Code 1	Code 2
A	1/2	00	0
T	1/4	01	10
C	1/8	10	110
G	1/8	11	111

We want a binary encoding for the four bases. As mentioned above, there are exactly four two-bit words, so that one can suggest the Code 1, which has exactly 4 words and uses 2 bits for every base. Here the word length is 2. However, it is straightforward to see that the entropy of the distribution $S = -\sum_{i=1}^{4} p_i \log_2 p_i = 7/4$ is lower that 2. One then may suggest a variable-length Code 2. It is built in the following way. We start from the least probable C and G, which we want to have the longest codewords of the same length differing by one (last) binary digit that distinguishes between the two of them. We then can combine C and G into a single source symbol with the probability 1/4, that is coinciding with the probability of T. To distinguish from C,G, we code T by two-bit word placing 0 in the second position. The combined C, G is now encoded 11, while T is encoded 10. We then can code A by one-bit word 0 to distinguish it from the combined T,C,G.

It is straightforward now to see that the Code 2 uses less bits per base on average, namely that its mean length of the codeword is exactly equal to the entropy: $(1/2)\cdot 1 + (1/4)\cdot 2 + (1/4)\cdot 3 = 7/4$. It is an example of the so-called Huffman code, which draws binary tree starting from its leaves: First, ascribe to the two least probable symbols two longest codewords differing in the last digit. Second, combine these two symbols into a single one and repeat. The procedure ends after $n-1$ steps where n is the size of the original alphabet. One may think that the variable-length code always requires an extra symbol (space or comma) to distinguish codewords in a continuous stream of 0 and 1. Actually, codes do not require a separating symbol, if they are prefix-free, that is no codeword can be mistaken for the beginning of another one. Such are, in particular, Huffman codes.

The most efficient code has the length of the mean codeword (the number of bits per base) equal to the entropy of the distribution, which determines the fastest mean transmission rate, that is the shortest mean codeword length.

To make yourself comfortable with the information brought by fractions of a bit, think about the decrease of uncertainty. One bit halves the uncertainty. For example, for a uniform distribution, receiving one bit shrinks its interval by the factor 2^{-1}. Receiving H bits shrinks the uncertainty interval to 2^{-H} fraction of its original length. Receiving half-bit shrinks the interval of possible values by the factor $2^{-1/2} \approx 0.7$.

The inequality (43) tells us, in particular, that using an alphabet is not optimal for the speech transmission rate as long as the probabilities of the letters are different. For example, if we use 26 letters, space and 5 punctuation marks (,.!?-), we need 5-bit words to encode these 32 symbols (actually used for teletype machines) We can use less symbols but variable codeword length to make the average codeword shorter than 5. Morse code uses just three symbols (dot, dash and space) to encode any language[12]. In English, the probability of "E" is 13% and of "Q" is 0.1%, so Morse encodes "E" by a single dot and "Q" by "$- \; - \cdot -$". One-letter probabilities give for the written English language the information per symbol as follows:

$$-\sum_{i=a}^{z} p_i \log_2 p_i \approx 4.11 \,\text{bits} \;,$$

which is lower than $\log_2 26 = 4.7$ bits.

3.3 Correlations in the signals

The first British telegraph managed to do without C,J,Q,U,X, which tells us that some letters can be guessed from their neighbors, and more generally that there is a correlation between letters. Apart from one-letter probabilities, one can utilize more knowledge about the language by accounting for two-letter correlation (say, that "Q" is almost always followed by "U", "H" often follows "T", etc). That will further lower the entropy.

A simple universal model with one-step correlations is a Markov chain. It is specified by the conditional probability $p(j|i)$ that the letter i is followed by j. For example $p(U|Q) = 1$. The probability is normalized for every i: $\sum_j p(j|i) = 1$. The matrix $p_{ij} = p(j|i)$, whose elements are positive and in every column sum to unity, is called stochastic. Do the vector of probabilities

[12] Great contributions of Morse were one-wire system and the simplest possible encoding (opening and closing the circuit), far more superior to multiple wires and magnetic needles of Ampere, Weber, Gauss and many others.

$p(i)$ and the transition matrix p_{ij} bring independent information? The answer is no, because the matrix p_{ij} and the vector p_i are not independent, but are related by the condition of stationarity: $p(i) = \sum p(j)p_{ji}$, that is $\mathbf{p} = \{p(a), \dots p(z)\}$ is an eigenvector with the unit eigenvalue of the matrix p_{ij}.

The probability of any N-string is then the product of $N-1$ transition probabilities times the probability of the initial letter. As in (42), minus the logarithm of the probability of a long N-string is a sum of uncorrelated numbers:

$$\log_2 p(i_1, \dots, i_N) = \log_2 p(i_1) + \sum_{k=2}^{N} \log_2 p(i_{k+1}|i_k) \,. \tag{47}$$

At large N the sum grows linearly with N with the rate, which is the mean value of the logarithm of conditional probability, $-\sum_j p(j|i)\log_2 p(j|i) = S_i$, called the conditional entropy S_i. Therefore, the number of typical sequences starting from i grows with N exponentially, as 2^{NS_i}. To get the mean rate of growth for all sequences, it must be averaged over different i with their probabilities $p(i)$. That way we express the language entropy via $p(i)$ and $p(j|i)$ by averaging over i the entropy of the transition probability distribution:

$$S = -\sum_i p_i \sum_j p(j|i)\log_2 p(j|i) \;. \tag{48}$$

That formula defines the information rate of the Markov source. We shall further discuss Markov chains describing index strategies in Section 4.7 and Google PageRank algorithm in Section 5.1 below.

One can go beyond two-letter correlations and statistically calculate the entropy of the next letter when the previous $L-1$ letters are known (Shannon 1950). As L increases, the entropy approaches the limit which can be called the entropy of the language. Long-range correlations and the fact that we cannot make up words further lower the entropy of English down to approximately 1.4 bits per letter, *if no other information given.* Comparing 1.4 and 4.7, we conclude that the letters in an English text are about 70% redundant. About the same value one finds asking people to guess the letters in a text one by one, which they do correctly 70% of the time. This redundancy makes possible data compression, error correction and crosswords. It is illustrated by the famous New York City subway poster of the 1970s:

"If u cn rd ths u cn gt a gd jb w hi pa!"

Triple redundancy of the alphabet encoding apparently serves the goal of protecting the message against errors of transmission. It could be that it also corresponds to the deeper need of our brain to obtain reinforcing of the prior guess[13] - "what I tell you three times is true".

So what is so special about alphabet? Redundant encodings are many. Moreover, the human language encodes meaning not in separate letters but in words. It was found empirically that if one ranks words by the frequency of their appearance in texts, then the frequency decreases as an inverse rank (Zipf 1949). For example, the first place with 7% takes *"the"*, followed by *"of"* with 3.5%, *"and"* with 1.7%, etc. More insight about the way we communicate from the frequency distribution of words and their meanings can be found in Appendix 8.4.

The oldest system of writing were logographic systems where every word or morpheme requires a separate symbol - logogram. Several independent such systems were developed: Egyptian hieroglyphics, cuneiform, Chinese characters, etc. Scribes and readers then learned thousands of symbols, which necessarily were restricted to a small part of society. The great democratizing invention of alphabetic writing, which dramatically improved handling of information (and irreversibly changed the ways we speak, hear and remember), was done only once in history. All known alphabets derive from that seminal (Semitic) script. The idea was to make writing not only conveying the meaning but also reproducing (extremely poorly!) the way the speech sounds. Of course, all known logographies have some phonetic component, generally based on the rebus principle (Putin=put+in). Alphabet makes a complete transition using phonograms instead of logograms. The way we hear is related to the notion of phonemes. Linguists define the phoneme as the smallest acoustic unit that makes a difference in meaning. Their numbers in different languages are subject to disagreements but generally are in tens. For example, most estimates for English give 45, that is comparable with the number of letters in the alphabet. Another interesting question is how we recognize words in a speech, which is essentially a running stream of sound, — apparently rhythm plays the leading role.

How redundant is the genetic code? There are four bases, which must encode twenty amino acids. There are 4^2 two-letter words, which is not enough. The designer then must use a triplet code with $4^3 = 64$ words, so that the redundancy factor is again about 3. Number of ways to encode a given amino acid is approximately proportional to its frequency of appearance.

Another example of redundancy for error-protection is the NATO phonetic alphabet used by the military and pilots. To communicate through a noisy acoustic channel, letters are encoded by full words: A is Alpha, B is Bravo, C is Charlie, etc.

How best to encode numbers? Using a separate symbol for every number stop

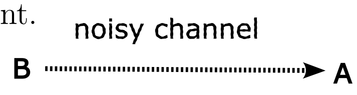
[13] See Section 4.5 below.

making sense when number N gets large. A simple way is to use one symbol and repeat it N times. It is immediately clear that one can encode better by dividing into groups, so that the number N can be encoded by $\log N$ symbols, which is much more efficient. One particular way of organizing numbers was another discovery of historical importance — a positional numeral system. Apart from $\log N$ economy, there is another profound consequence of encoding where the value depends on the position: it already implies algebraic operations. Indeed, reading (decoding) requires multiplying and adding: $2021 = 2 \times 1000 + 2 \times 10 + 1$. It then allowed simple automatic rules for computations (formulated by Persian al-Khwarizmi, from whose name the word algorithm appeared). Algebra, alcohol, etc, also have Arabic origin.

To conclude this subsection, recall that knowing the probability distribution one can compute entropy, which determines the most efficient rate of encoding. One can turn tables and estimate the entropy of the data stream looking for its most compact lossless encoding. It can be done in a one-pass (online) way, that is not looking at the whole string of data, but optimizing encoding as one processes the string from beginning to end. There are several such algorithms called adaptive codes (Lempel-Ziv, deep neural networks, etc). These codes are also called universal, since they do not require a priori knowledge of the distribution.

3.4 Mutual information as a universal tool

Answering the question i) in Sect. 3.2, we have found that the entropy of the set of objects determines the minimum mean number of bits per object (word length in the binary code), that is the maximal transfer rate of the information about the objects. In this section, we turn to the question ii) and find out how this rate is lowered if the transmission channel can make errors, so that one cannot unambiguously restore the input B from the output A. How much information then is lost on the way? In this context one can treat measurements A as messages about the value of the quantity B we measure. One can also view storing and retrieving information as sending a message through time rather than space. We can include into the same scheme forecast and observation, asking how much information about the experimental data B is contained in the theoretical predictions A. In all cases, A is what we have and B is what we want.



When the channel is noisy, the statistics of inputs $P(B)$ and outcomes $P(A)$ are generally different, that is we need to deal with two probability distributions and the relation between them. Treating inputs and outputs as taken out of distributions works for channels/measurements both with and without noise; in the limiting cases, the distribution can be uniform or peaked at a single value. Relating two distributions needs conditional probabilities, which we already introduced in Section 3.3. It will lead us to relative entropy and mutual information, which presently are the most powerful and universal tools of information theory.

The relation between the message (measurement) A_i and the event (quantity) B_j is characterized by the conditional probability (of B_j in the presence of A_i), denoted $P(B_j|A_i)$. For every A_i, this is a normalized probability distribution, and one can define its entropy $S(B|A_i) = -\sum_j P(B_j|A_i)\log_2 P(B_j|A_i)$. Since we are interested in the mean quality of transmission, we average this entropy over all values of A_j, which defines the so-called *conditional entropy*:

$$S(B|A) = \sum_i P(A_i)S(B|A_i) = -\sum_{ij} P(A_i)P(B_j|A_i)\log_2 P(B_j|A_i)\ . \quad (49)$$

We already encountered it in (49) considering correlations between subsequent terms in the sequence. If our sequence consisted of the related pairs A_i, B_j, like i_k, i_{k+1} in the previous section, it would bring the information (49,50).

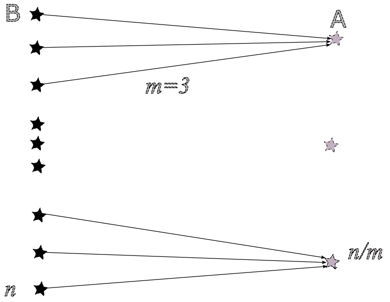
But we do not receive B_j. How much information about B brings the knowledge of A? The conditional entropy between input and output measures what on average remains unknown about B after the value of A is known. The missing information was $S(B)$ before the measurement and is equal to the conditional entropy $S(B|A)$ after it. Then what the measurements bring on average is their difference called *the mutual information*:

$$I(A,B) = S(B) - S(B|A) = \sum_{ij} P(A_i)P(B_j|A_i)\log_2\left[\frac{P(B_j|A_i)}{P(B_j)}\right]\ . \quad (50)$$

Information is a decrease in uncertainty, so that the mutual information must be non-negative. That means that measurements on average lower uncertainty by increasing the conditional probability relative to unconditional:

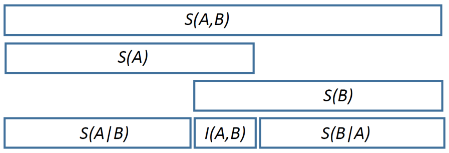
$$\langle \log_2[P(B_j|A_i)/P(B_j)]\rangle \geq 0\,.$$

For example, let B be a choice out of n equal possibilities: $P(B) = 1/n$ and $S(B) = \log_2 n$. Assume that for every A_i we can have m different values of B from disjoint sets, as shown in the Figure. Then $P(B|A) = 1/m$, $S(B|A) = \log_2 m$, and $I(A,B) = S(B) - S(B|A) = \log_2(n/m) \geq 0$, since evidently $m \leq n$. In this case, knowledge of B fixes A, so that $S(A|B) = 0$ and $I(A,B) = S(A)$. When there is one-to-one correspondence, $m = 1$, then A tells us all we need to know about B.



Probabilities are multiplied and entropies added for independent events. For correlated events, one uses conditional probabilities and entropies in what is called the chain rule:

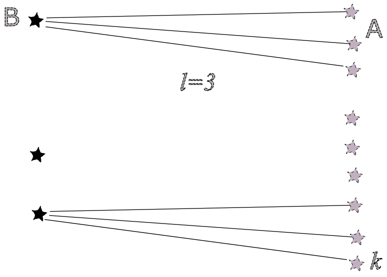
$$P(A_i, B_j) = P(B_j|A_i)P(A_i)\,, \tag{51}$$
$$S(A,B) = S(A) + S(B|A) = S(B) + S(A|B)\,.$$



It gives $I(A,B)$ in a symmetric form:

$$\begin{aligned} I(A,B) &= \sum_{ij} P(A_i,B_j)\log_2\left[\frac{P(A_i,B_j)}{P(A_i)P(B_j)}\right] \\ &= S(B) - S(B|A) = S(A) + S(B) - S(A,B) = S(A) - S(A|B)\,. \end{aligned} \tag{52}$$

To illustrate the symmetry, consider the case dual to the above $m-n$ example: For every equally probable input B, we have l equally probable values of A, whose total number is k. In this case, $P(A|B) = 1/l$ and $S(A|B) = \log_2 l$, so that $I(A,B) = S(A) - S(A|B) = \log_2(k/l) = S(B)$ bits, similar to the $m - n$ case.



To avoid confusion, let us state the obvious: there is no symmetry between A and B. They could be of very different nature - one is the position of an atom, another is the reading of the device, for instance. Neither their entropies, $S(A)$ and $S(B)$, nor the conditional entropies, $S(B|A)$ and $S(A|B)$, are generally equal or even comparable. Yet the degree of their correlation $I(A,B)$ is a symmetric function.

It is important to stress that measuring A decreases the entropy of B only on average over all values A_i: $S(B|A) \leq S(B)$. That follows from $P(B_j) = \sum_i P(B_j|A_i)P(A_i)$ and the convexity of the logarithm. Yet for any particular A_i the entropy $S(B|A_i)$ can be either smaller or larger than $S(B)$, depending on how this measurement changes the probability distribution (see the problem Conditional Entropy of Criminality). Note that $P(A_i, B_j)$ could be either larger or smaller than $P(A_i)P(B_j)$ when the pair A_i, B_j are respectively correlated or anti-correlated. Yet on average, the non-negativity of the mutual information gives the so-called sub-additivity of entropy:

$$S(A) + S(B) > S(A,B) \ . \tag{53}$$

When A and B are independent, the joint entropy is a sum, and the mutual information is zero. When A, B are related deterministically, $S(A) = S(B) = S(A,B) = I(A,B)$, where $S(A) = -\sum_i P(A_i)\log_2 P(A_i)$, etc. And finally, since $P(A|A) = 1$ then the mutual information of a random variable with itself is the entropy: $I(A,A) = S(A)$. So one can call entropy self-information. Another evident remark is that $I(A,B)$ exceeds neither $S(A)$ nor $S(B)$. Indeed, A cannot contain more information about B than about itself or than the information B contains about itself.

We have seen in the previous section that the mutual information between letters lowered the entropy of the language from the one-letter entropy, $-\sum_i p(i)\log p(i)$. That lowering is brought by the knowledge of the conditional probabilities $p(j|i)$, $p(j|i,k\ldots)$, which is more than knowledge of $p(i)$.

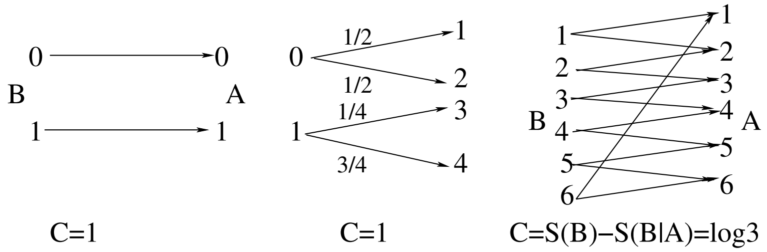
3.5 Channel capacity

If the mutual information is what on average brings an imperfect channel, how reliable is it? It is tempting to assume that the mutual information plays for noisy channels the same role the entropy plays for ideal channels, in particular, sets the maximal rate of reliable communication in the limit of long messages, thus answering the question ii) from the Section 3.2 Indeed, if there are different outputs for the same input, like in the above simple

$k - l$ example, the rate of information transfer is lower than for a one-to-one correspondence, since we need to divide our k outputs into groups of l, distinguishing only between the groups. More formally, for each typical N-sequence of independently chosen B-s, we have $[P(A|B)]^{-N} = 2^{NS(A|B)}$ possible output sequences, all of them equally likely. To get the rate of the useful information about distinguishing the inputs, we need to divide the total number of typical outputs $2^{NS(A)}$ into sets of size $2^{NS(A|B)}$ corresponding to different inputs. Therefore, we can distinguish at most $2^{NS(A)}/2^{NS(A|B)} = 2^{NI(A,B)}$ sequences of the length N, which sets $I(A, B)$ as the maximal rate of information transfer.

However, that was a rather trivial case when inputs can be distinguished from outputs without errors. But what if a single output can correspond to different inputs like in the above $m - n$ example? There is no way now to determine every input exactly. Can we still use this imperfect channel to convey information in a way where errors can be made arbitrarily small? Yes, we can if we avoid overlapping inputs, or in other words, choose correctly the statistics of inputs. Let us characterize the channel itself, maximizing $I(A, B)$ over all choices of the input statistics $P(B)$. That quantity is called the Shannon's channel capacity, which quantifies the quality of communication systems in bits per symbol:

$$\mathcal{C} = \max_{P(B)} I(A, B) .$$

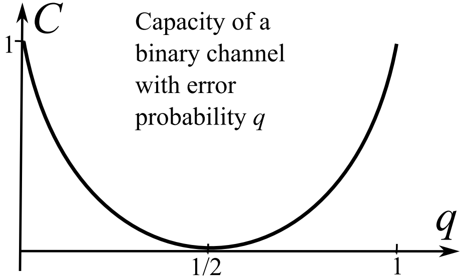


To put it simply, the channel capacity is the log of the maximal number of distinguishable inputs. For example, if our channel transmits the binary input exactly (zero to zero, one to one), then the capacity is 1 bit, which is achieved by choosing $P(B = 0) = P(B = 1) = 1/2$, see the left panel in the Figure. Let us stress that if $P(0) \neq P(1)$, then the average rate is less than the capacity (one bit per symbol) despite the channel being perfect. Even if the channel has many outputs for every input out of n, the capacity is still $\log_2 n$, if those outputs are non-overlapping for different inputs, so that the input can be determined without an error and $P(B|A) = 1$. Such case is presented in the middle panel in the Figure. In this case, the transfer rate is determined by the number of B-states; from the perspective of A-states, the rate is $S(A) - S(A|B) = 2 - 1 = 1$.

Like the mutual information, the capacity deviates down from $S(B)$ when the same outputs appear for different inputs, say, different groups of m inputs each give the same output, so that $P(B|A) = 1/m$. In this case, one cannot achieve error-free transition for uniform $P(B)$, one needs to choose only one input symbol from each of n/m groups, that is using $P(B) = m/n$ for the symbols chosen and $P(B) = 0$ for the rest; the capacity is then indeed $\mathcal{C} = \log_2(n/m)$ bits (in the right panel of the Figure $n = 6$, $m = 2$). Lowered capacity means increased redundancy, that is a need to send more symbols to convey the same information.

Let us treat at last the most generic case with random errors, when one cannot separate inputs/outputs into completely disjoint groups. Here, one may argue that taking the limit of large N does not help since the channel continues to make errors all the time. And yet Shannon showed (in the co-called noisy channel theorem) that one can keep a finite transmission rate and yet make the probability of error arbitrarily small at the limit $N \to \infty$. The idea is that to correct errors one needs to send extra bits, so to get the rate we need to compute how many bits are devoted to error correction and how many to transferring the information itself. Shannon showed that it is possible to make the probability of error arbitrarily small when sending information with a finite rate R, if there is any correlation between output A and input B, that is $\mathcal{C} > 0$. Then the probability of an error can be made $2^{-N(\mathcal{C}-R)}$, that is asymptotically small in the limit of $N \to \infty$, if the rate is lower than the channel capacity. This (arguably the most important) result of the communication theory is rather counter-intuitive: if the channel makes errors all the time, how can one decrease the error probability treating long messages? Shannon's argument is based on typical sequences and average equipartition, that is on the law of large numbers (by now familiar to you).

For example, if in a binary channel the probability of every single bit going wrong is q, then A is a binary random variable with equal probabilities of 0 and 1, so that $S(A) = \log_2 2 = 1$. Conditional probabilities are $P(1|0) = P(0|1) = q$ and $P(1|1) = P(0|0) = 1 - q$, so that $S(A|B) = S(B|A) = S(q) = -q \log_2 q - (1 - q) \log_2(1 - q)$. The mutual information $I(A, B) = S(A) - S(A|B) = 1 - S(q)$. This is actually the maximum, that is the channel capacity: $\mathcal{C} = \max_{P(B)}[S(B) - S(B|A)] = 1 - S(q)$, because the maximal entropy $P(B) = 1$, which corresponds to $P(0) = P(1) = 1/2$.



Let us now see how the rate of transmission is bounded from above by the capacity. To correct an error, we need to specify its place. In a message of length N, there are on average qN errors and there are $N!/(qN)!(N-qN)! \approx 2^{NS(q)}$ ways to distribute them. We then need to devote some m bits in the message not to data transmission but to error correction. Apparently, the number of possibilities provided by these extra bits, 2^m, must exceed $2^{NS(q)}$, which means that $m > NS(q)$, and the transmission rate $R = (N-m)/N < 1 - S(q)$. The channel capacity is zero for $q = 1/2$ and is equal to 0.988 bits per symbol for $q = 10^{-3}$. The probability of errors is binomial with the mean number of errors qN and the standard deviation $\sigma = \sqrt{Nq(1-q)}$. If we wish to bound the error probability from above, we must commit to correcting more than the mean number of errors, making the transmission rate smaller than the capacity.

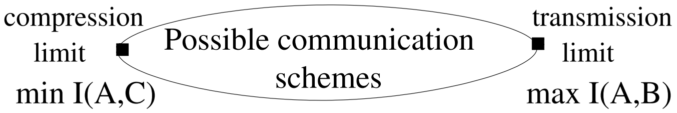
The conditional entropy $S(B|A)$ is often independent of the input statistics $P(B)$ like in the above example. Maximal mutual information, that is capacity, is then achieved for maximal $S(B)$. If no other restrictions imposed, that corresponds to the uniform distribution $P(B)$.

If the measurement/transmission noise ξ is additive, that is the output is $A = g(B) + \xi$ with an invertible function g, then $S(A|B) = S(\xi)$, so that

$$I(A, B) = S(A) - S(\xi) \ . \tag{54}$$

The more choices of the output are recognizable despite the noise, the more is the capacity of the channel. When the conditional entropy $S(A|B)$ is given, then to maximize the mutual information we need to choose the measurement/coding procedure, for instance, $g(B)$ above, that maximizes the entropy of the output $S(A)$.

Mutual information also sets the limit on the data compression $A \to C$, if coding has a random element so that its entropy $S(C)$ is nonzero. In this case, the maximal data compression, that is the minimal coding length in bits, is $\min I(A,C)$.



Take-home lesson: entropy of the symbol set is the ultimate data compression rate; channel capacity is the ultimate transmission rate. Since we cannot compress below the entropy of the alphabet and cannot transfer faster than the capacity, then transmission is possible only if the latter exceeds the former.

3.6 Continuous case and Gaussian Channel

Information theory is essentially discrete, since it is ultimately about counting. Moreover, the world of natural phenomena is described by digitized data both on a practical level because of finite resolution and on a fundamental level because of quantum bounds on maximal entropy in a given volume. Yet the analysis presents such a convenient mathematical tool with all the derivatives and integrals, that we generalize here the definition of the Gibbs entropy (41) for a continuous distribution.

In a continuous case, an indeterminacy is infinite, as for determining the position of a point on an interval L. If we agree to know the position with an accuracy ϵ, then the entropy is $S(B) = \log_2(L/\epsilon)$. How much information does a measurement A of the point position with a precision Δ bring? The indeterminacy in the point position after the measurement is $S(B|A) = \log_2(\Delta/\epsilon)$, so that the measurement brought the information independent of ϵ:

$$I(A,B) = S(B) - S(B|A) = \log \frac{L}{\Delta} \,. \tag{55}$$

We see that even though the entropies go to infinity in the continuous limit $\epsilon \to 0$, the mutual information stays finite. That property makes the mutual information and its quantum cousin, entanglement entropy, so important in physics, since they are insensitive to microscopic details and free from ultraviolet divergencies.

More generally, we define the entropy of a continuous distribution $\rho(x)$ by dividing into ϵ-intervals and denoting $p_i = \rho(x_i)\epsilon$. Such entropy in the limit $\epsilon \to 0$ consists of two parts:

$$-\sum_i p_i \log p_i \to -\int dx \rho(x) \log \rho(x) - \log \epsilon\,. \tag{56}$$

The second part is an additive constant depending on the resolution. When interested in the functional form of the distribution, we usually focus on the first term, which is called differential entropy $S(x)$. It is the difference between the entropies of the coarse-grained distribution and of the uniform distribution; when $\epsilon \to 0$ both diverge but their difference may stay finite. In distinction from a discrete case, it is invariant with respect to shifts but not re-scaling of the variables: $S(ax+b) = S(x) + \log a$. For example, the differential entropy of the Gaussian distribution $P(\xi) = (2\pi\mathcal{N})^{-1/2}\exp[-\xi^2/2\mathcal{N}]$ is as follows:

$$S(\xi) = -\int_{-\infty}^{\infty} P(\xi)\log_2 P(\xi) = \frac{1}{2}\log_2 2\pi e\mathcal{N}\,.$$

Consider a linear noisy channel: $A = B + \xi$, such that the noise is independent of B and Gaussian with $\langle\xi\rangle = 0$ and $\langle\xi^2\rangle = \mathcal{N}$. Then $P(A|B) = (2\pi\mathcal{N})^{-1/2}\exp[-(A-B)^2/2\mathcal{N}]$. If in addition we have a Gaussian input signal with $P(B) = (2\pi\mathcal{S})^{-1/2}\exp(-B^2/2\mathcal{S})$, then

$$P(A) = \int dBd\xi P(B)P(\xi)\delta(A-B-\xi) = [2\pi(\mathcal{N}+\mathcal{S})]^{-1/2}\exp[-A^2/2(\mathcal{N}+\mathcal{S})]\,.$$

Now, using the chain rule, we can write

$$P(B|A) = P(A|B)P(B)/P(A) = \sqrt{\frac{\mathcal{N}+\mathcal{S}}{2\mathcal{N}}}\exp\left[-\frac{\mathcal{S}+\mathcal{N}}{2\mathcal{N}}\left(B - \frac{A}{\mathcal{S}+\mathcal{N}}\right)^2\right]\,.$$

If we measure the value $A = a$, what is the best estimate for the value $B(a) = b$? It is computed using the conditional probability:

$$b = \int BP(B|a)\,dB = \frac{a\mathcal{S}}{\mathcal{S}+\mathcal{N}} = a\frac{SNR}{1+SNR}\,, \tag{57}$$

where signal to noise ratio is $SNR = \mathcal{S}/\mathcal{N}$. The rule (58) makes sense: To "decode" the output of a linear detector we use the unity factor at high SNR, while at low SNR we scale down the output since most of what we are seeing must be noise. Note that the estimate of b is linearly related to the measurement a, which requires two things: linearity of the input-output relation and Gaussianity of the statistics. Let us now find the mutual information (55):

$$\begin{aligned} I(A,B) &= S(A)-S(A|B)=S(A)-S(B+\xi|B)=S(A)-S(\xi|B)=S(A)-S(\xi) \\ &= \frac{1}{2}\left[\log_2 2\pi e(\mathcal{S}+\mathcal{N})-\log_2 2\pi e\mathcal{N}\right]=\frac{1}{2}\log_2(1+SNR)\ . \end{aligned} \tag{58}$$

The capacity of such a channel depends on the input statistics. One increases capacity by increasing the input signal variance, that is the dynamic range relative to the noise. For a given input variance, the maximal mutual information (channel capacity) is achieved by a Gaussian input, because the Gaussian distribution has maximal entropy for a given variance. Indeed, varying $\int dx\rho(x)(\lambda x^2-\ln\rho)$ with respect to ρ we obtain $\rho(x)\propto\exp(-\lambda x^2)$. Therefore (59) determines also the capacity of the Gaussian channel in bits per transmission: $\mathcal{C}=\log_2\sqrt{(\mathcal{N}+\mathcal{S})/\mathcal{N}}$. That means that receiving a value A allows to distinguish between $2^{\mathcal{C}}$ values, that is noise effectively makes a continuous channel discrete. We shall elaborate on that in section 4.6.

3.7 Hypothesis testing and Bayes' formula

...la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul

Laplace

All empirical sciences need a quantitative tool for confronting hypothesis with data. One (rational) way to do that is statistical: update prior beliefs in light of the evidence. It is done using conditional probability. Indeed, for any e and h, we have $P(e,h)=P(e|h)P(h)=P(h|e)P(e)$. If we now call h hypothesis and e evidence, we obtain the rule for updating the probability of hypothesis to be true:

$$P(h|e)=P(h)\frac{P(e|h)}{P(e)}\ . \tag{59}$$

This form of the chain rule is so important that it has been named after Bayes, who first introduced it (in 1763). That common-sense statement specifies how to update the probability that the hypothesis h is correct after we receive the data e: the new (posterior) probability $P(h|e)$ is the prior probability $P(h)$ times the quotient $P(e|h)/P(e)$, which presents the support e provides for h. Without exaggeration, one can say that most errors made by data analysis in science and most conspiracy theories are connected to neglect or abuse of this simple formula. For example, your hypothesis is the existence of a massive international conspiracy to increase the power of governments and the evidence is COVID pandemic. In this case $P(e|h)$ is high: a pandemic provoking increase of the state power is highly likely *given* such a conspiracy exists. This is presumably why some people stop thinking here and accept the hypothesis. They thus commit the error called inversion of the conditional, since we need to evaluate not $P(e|h)$, but $P(h|e)$. Even when the former is not small, the latter could be. Indeed, absent such an event, the prior probability $P(h)$ could be vanishingly small. To overcome that smallness by a large quotient support factor we need to evaluate total $P(e)$, that is the probability that pandemic happens with or without conspiracy.

If we choose between two mutually exclusive hypotheses, h_1 and h_2, then the total probability of the evidence consists of two terms: $P(e) = P(e, h_1) + P(e, h_2) = P(h_1)P(e|h_1) + P(h_2)P(e|h_2)$. Then the posterior probability of the hypothesis being true is as follows:

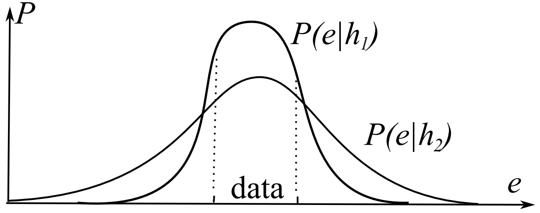
$$P(h_1|e) = P(h_1)\frac{P(e|h_1)}{P(e)} = P(h_1)\frac{P(e|h_1)}{P(h_1)P(e|h_1) + P(h_2)P(e|h_2)} \ . \qquad (60)$$

For example, when we want to check a priori improbable hypothesis, $P(h_1) \ll P(h_2)$, any data changing the probability of this hypothesis won't matter much because $P(e|h_1)$ in (61) is multiplied by a small number $P(h_1)$. It is better then to design experiment or look for the data which could minimize $P(e|h_2)$ rather than maximize $P(e|h_1)$, that is rule out alternatives rather than supports the hypothesis. This is why even good tests, with $P(e|h_1)$ close to unity and $P(e|h_2)$ small, are not very reliable at the beginning of a pandemic, when $P(h_1)$ is small. The same is true for drug test in a mostly drug-free population. Suppose that a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users (hypothesis h_1) and 99% true negative results for clean people (hypothesis h_2). If we denote e the positive test result, then $P(e|h_1) = 0.99$ and $P(e|h_2) = 1 - 0.99 = 0.01$. Suppose that 0.5% of people are drug users, that is $P(h_1) = 0.005$. The probability that a randomly selected individual with a positive test is a drug user is $0.005 \cdot 0.99/(0.99 \cdot 0.005 + 0.01 \cdot 0.995) \approx 0.332$ that is less that half. The result is more sensitive to specificity approaching unity, when $P(e|h_2) \to 0$, than to sensitivity.

The choice between two (not necessarily exclusive) hypotheses is determined by the ratio of their probabilities conditioned on the data:

$$\frac{P(h_1|e)}{P(h_2|e)} = \frac{P(h_1)}{P(h_2)} \frac{P(e|h_1)}{P(e|h_2)} \ . \tag{61}$$

Both factors here quantify Occam's razor, which is preference for simpler hypothesis. The second factor is applied to data and is mostly used by experimentalists. More complex hypothesis, say h_2, is capable of a wider variety of predictions, so it spreads its probability over the data space more thinly. If the evidence is compatible with both hypotheses (the data range is around their probability maxima as in the Figure), simpler hypothesis generally assigns more probability to the evidence.



In distinction from experimentalists, theoreticians apply Occam's razor to the first factor in (62) choosing prior beliefs on aesthetic grounds of mathematical beauty and simplicity.

Alternatively, one can always interpret higher probability as lower information brought by the choice. That interpretation of (62) is sometimes called minimum description length: one should prefer the hypothesis that communicates the data in the smaller number of bits. There are two subsequent messages communicated: first we choose the model and then communicate the data within this model. The length of the message is then $-\log_2 P(h) - \log_2 P(e|h) = -\log_2 P(e,h)$. This way we say that the choice of a simpler model is communicated in less bits and such model also communicates data prediction in less bits since a more narrow distribution has lower entropy. Technically, $P(e|h)$ is also evaluated in a two-step process, so the respective message has two parts: first we specify the choice parameters, then communicate the data in these terms. Increasing the number of parameters we are able to fit the data better which shortens the error list in the data message; optimization of the respective trade-off is briefly described at the end of Section 4.7.

Note the shift in the interpretation of probability brought by (60-62). Traditional sampling approach by mathematicians and gamblers treats probability as the *frequency of outcomes in repeating trials.* Bayesian approach defines probability as a *degree of belief*; that definition allows wider applications, particularly when we cannot have repeating identical trials, nor an ensemble of identical objects. For example, we have only one planet Earth and cannot yet restart it from the same or different initial conditions. Therefore, any estimate of the statistical significance of, say, global warming prediction must be based on the Bayesian approach. The approach may seem unscientific since it is dependent on the prior beliefs, which can be subjective. However, repeatedly subjecting our hypothesis to variable enough testing, we hope that the resulting flow in the space of probabilities will eventually come close to a fixed point independent of the starting position. Normally, only a sequence of data with a clear trend of increasing probability may lead us to accept the hypothesis.

Making prior assumptions explicit is important, both computationally and conceptually. There are neither inference nor prediction without assumptions, however uncomfortable some may feel about that. For example, given $5, 8, \ldots$ as two numbers of the sequence, one may put forward two hypotheses: h_1 predicts an arithmetic sequence $5, 8, 11, \ldots$, while h_2 predicts the Fibonacci sequence $5, 8, 13, \ldots$, where any number is the sum of two preceding ones. If the next number comes through the noisy channel as 12 ± 1, then $P(e|h_1) = P(e|h_2)$ and the choice in (62) is due to priors. Engineers and accountants would argue that arithmetic sequences are more frequently encountered, while natural scientists would point to pine cones, floral petals and seed heads to argue for Fibonacci.

Observing our own mental processes gives us both the idea of logic and of statistical inference. A Bayesian approach is used in brain research on multiple levels, from interpretation of neural spikes and functional brain imaging to modeling sensory processing and belief propagation. One such approach is described in Section 4.5.

One also uses the Bayes' formula for design. For example, experimentalists measure the sensory response A of an animal to the stimulus B, which gives $P(A|B)/P(A)$ or build a robot with the prescribed response. Then they go to the natural habitat of that animal/robot and measure the distribution of stimuli $P(B)$ (see the example at the beginning of Section 4.4). After that one obtains the conditional probability

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)} \ , \tag{62}$$

that allows animal/robot to perceive the environment and function effectively in that habitat.

3.8 Relative Entropy

The mutual information $I(A,B)$ measures the degree of correlation, which is essentially the difference between the true joint distribution $P(A,B)$ and the product distribution $P(A)P(B)$ of two independent quantities. As such, it is a particular case of a more general measure of difference between distributions. Let us ask the following question: How fast can a sequence of data invalidate an incorrect hypothesis? If the true distribution is p but our hypothetical distribution is q, what number N of trials is sufficient to decrease the probability $P(h|e)$ by some a priori set factor? For that we need to estimate how fast decreases with N the factor $\mathcal{P} = P(e|h)/P(e)$, that is to compute the probability of the stream of data observed given the distribution q. The result i is observed $p_i N$ times. We *judge* the probability of that happening as $q_i^{p_i N}$ times the number of sequences with those frequencies:

$$\mathcal{P} = \prod_i q_i^{p_i N} \frac{N!}{\prod_j (p_j N)!} . \tag{63}$$

This is how fast changes with N the probability of our hypothetical distribution being true given the set of data. Considering the limit of large N we obtain:

$$\mathcal{P} \approx \exp\{N[\ln N + \sum_i p_i(\ln q_i - \ln p_i N)]\} = \exp\left[-N\sum_i p_i \ln(p_i/q_i)\right] . \tag{64}$$

This a large-deviation-type relation like (155) in Appendix 8.2. The probability exponentially changes with the rate called the *relative entropy* (Kullback-Liebler divergence):

$$D(p|q) = \textstyle\sum_i p_i \ln(p_i/q_i) = \langle \ln(p/q) \rangle \ . \tag{65}$$

We need this quantity to be always non-negative so that the probability of not-exactly-correct hypothesis to approximate the data decreases with the number of trials. That can be shown using the simple inequality $\ln x \le x - 1$ (turning into equality only for $x = 1$):

$$-D(p|q) = \textstyle\sum_i p_i \ln(q_i/p_i) \le \sum_i (q_i - p_i) = 0 \, .$$

To prove our hypothesis wrong, we need the number N of trials making $ND(p|q)$ exceeding a threshold. The closer our hypothesis is to the true distribution, the larger the number of trials needed. On the other hand, when $ND(p|q)$ is below the threshold, our hypothetical distribution is just fine.

The relative entropy measures how different the hypothetical distribution q is from the true distribution p. Note that $D(p|q)$ is not the difference between entropies (which just measures difference in uncertainties). Nor the relative entropy is a geometrical distance since it does not satisfy the triangle inequality and is asymmetric, $D(p|q) \neq D(q|p)$. Indeed, there is no symmetry between reality and a hypothesis. Yet $D(p|q)$ has important properties of a distance: it is non-negative and turns into zero only when distributions coincide, that is $p_i = q_i$ for all i.

Nonnegativity and asymmetry are related. Indeed, if I believe that the distribution is p_i, then the entropy $-\sum_i p_i \ln p_i$ quantifies my average degree of surprise upon receiving the series of outcomes. But if somebody believes that the distribution is actually q, then her surprise upon the outcome i is $-\ln q_i$. I *judge* her average degree of surprise to be $-\sum_i p_i \ln q_i$. That must be larger than my own degree, since I naturally assume myself possessing the best distribution, otherwise I'd replaced it by a better option.

In particular, relative entropy quantifies how close to reality is the asymptotic equipartition estimate (42) of the probability of a given sequence. Assume that we have an N-sequence where the values/letters appear with the frequencies q_k, where $k = 1, \ldots, K$. Then the asymptotic equipartition (the law of large numbers) suggests that the probability of that sequence is $\prod_k q_k^{Nq_k} = \exp(N \sum_k q_k \ln q_k) = \exp[-NS(q)]$. But the frequencies we observe in a finite sequence are generally somewhat different from the true probabilities $\{p_k\}$. That difference has a price, so that the true probability is actually lower, which follows from the positivity of the relative entropy: $\prod_k p_k^{Nq_k} = \exp(N \sum_k q_k \ln p_k) = \exp[N \sum_k (q_k \ln q_k + q_k \ln(p_k/q_k))] = \exp\{-N[S(q) + D(q|p)]\}$. Asymptotic equipartition on average overestimates the probability of a given sequence. It is not suprising, since it disregards atypical sequences assuming the ensemble smaller than it really is.

If our guess is the Gibbs distribution with a given temperature, $q_i = Z^{-1} \exp(-E_i/T)$, then the relative entropy is the difference of the free energies divided by that temperature:

$$D(p|q) = \ln Z + \sum_i p_i E_i/T - S(p) = -\frac{F(q)}{T} + \frac{E}{T} - S(p) = \frac{F(p) - F(q)}{T}\,.$$

Positivity of $D(p|q)$ corresponds to the known fact that the Gibbs distribution has the lowest free energy (which does not necessarily mean that it is a true distribution in every case). Therefore, one can also think of the relative entropy as a generalization of a free energy difference for non-Gibbs q-distribution.

How many different probability distributions $\{q_k\}$ (called types in information theory) exist for an N-sequence made out of an alphabet with K symbols? The distribution $\{q_k\}$ is a K-vector. Since q_k can take any of $N+1$ values $0, 1/N, \ldots, 1$, then the number of possible K-vectors is at most $(N+1)^K$, which grows with N only polynomially, where the alphabet size K sets the power. The number of sequences grows exponentially with N, so that there is an exponential number of possible sequences for each type. The probability to observe a given type (empirical distribution) is determined by the relative entropy $\mathcal{P}\{q_k\} \propto \exp[-ND(q|p)]$.

Mutual information is that particular case of the relative entropy when we compare the true joint probability $p(x_i, y_j)$ with the distribution made out of their separate measurements $q(x_i, y_j) = p(x_i)p(y_j)$, where $p(x_i) = \sum_j p(x_i, y_j)$ and $p(y_j) = \sum_i p(x_i, y_j)$:

$$D(p|q) = S(X) + S(Y) - S(X, Y) = I(X, Y) \geq 0\,.$$

If i in p_i runs from 1 to K we can introduce $D(p|u) = \log_2 K - S(p)$, where u is a uniform distribution. That allows one to show that both relative entropy and mutual information inherit from entropy convexity properties. You are welcome to prove that $D(p|q)$ is convex with respect to both p and q, while $I(X, Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$. In particular, convexity is important for making sure that the extremum we are looking for is unique and lies at the boundary of allowed states.

Relative entropy also measures the price of non-optimal coding. As we discussed before, a natural way to achieve an optimal coding would be to assign the length to the codeword according to the probability of the object encoded: $l_i = -\log_2 p_i$. Indeed, the information in bits about the object, $\log_2(1/p_i)$, must be exactly equal to the length of its binary encoding. For an alphabet with d

letters, $l_i = -\log_d p_i$. The more frequent objects are then coded by shorter words, and the mean length is the entropy. The problem is that l_i must all be integers, while $-\log_d p_i$ are generally not. A set of integer l_i effectively corresponds to another distribution with the probabilities $q_i = d^{-l_i}/\sum_i d^{-l_i}$. Assume for simplicity that we found encoding with $\sum_i d^{-l_i} = 1$ (unity can be proved to be an upper bound for the sum). Then $l_i = -\log_d q_i$ and the mean length is $\bar{l} = \sum_i p_i l_i = -\sum_i p_i \log_d q_i = -\sum_i p_i(\log_d p_i - \log_d p_i/q_i) = S(p) + D(p|q)$, that is larger than the optimal value $S(p)$, so that the transmission rate is lower. In particular, if one takes $l_i = \lceil \log_d(1/p_i) \rceil$, that is the integer part, then one can show that $S(p) \le \bar{l} \le S(p) + 1$, that is non-optimality is at most one bit.

Monotonicity and irreducible correlations. If we observe fewer variables, then the relative entropy is less, property called monotonicity:

$$D[p(x_i, y_j)|q(x_i, y_j)] \ge D[p(x_i)|q(x_j)] \,,$$

where as usual $p(x_i) = \sum_j p(x_i, y_j)$ and $q(x_i) = \sum_j q(x_i, y_j)$. With fewer variables, we need larger N to have the same confidence. In other words, information does not hurt (but only on average!). For three variables, one can define $q(x_i, y_j, z_k) = p(x_i)p(y_j, z_k)$, which neglects correlations between X and the rest. What happens to $D[p(x_i, y_j, z_k)|q(x_i, y_j, z_k)]$ if we do not observe Z at all? Integrating Z out turns q into a product. Monotonicity gives

$$D[p(x_i, y_j, z_k)|q(x_i, y_j, z_k)] \ge D[p(x_i, y_j)|q(x_i, y_j)] \,.$$

But when q is a product, D turns into I and we can use (53):

$$\begin{gathered} D[p(x_i, y_j, z_k)|q(x_i, y_j, z_k)] = \left\langle p(X, Y, Z) \log \tfrac{p(X,Y,Z)}{p(X)p(Y,Z)} \right\rangle = S(X) + S(Y, Z) \\ -S(X, Y, Z) \ge D[p(x_i, y_j)|q(x_i, y_j)] = S(X) + S(Y) - S(X, Y) \,, \end{gathered}$$

so we obtain $S(X, Y) + S(Y, Z) - S(Y) - S(X, Y, Z) \ge 0$, which is called strong sub-additivity. It can be presented as the positivity of the *conditional mutual information*

$$\begin{aligned} I(X, Z|Y) &= S(X|Y) + S(Z|Y) - S(X, Z|Y) = S(X, Y) - S(Y) + S(Z, Y) \\ &-S(Y) - S(X, Y, Z) + S(Y) = S(X, Y) + S(Y, Z) - S(Y) - S(X, Y, Z) \,. \end{aligned} \tag{66}$$

That allows one to make the next step in disentangling information encoding. The straightforward generalization of the mutual information for many objects, $I(X_1, \ldots, X_k) = \sum S(X_i) - S(X_1, \ldots, X_k)$, simply measures the total correlation. We can introduce a more sophisticated measure of correlations called the interaction (or multivariate) information, which measures the irreducible information in a set of variables, beyond that which is present in any subset of those variables. For three variables it measures the difference between the total correlation and that encoded in all pairs and is defined as follows (McGill 1954):

$$II = I(X,Z) - I(X,Z|Y) = S(X) + S(Y) + S(Z) - S(X,Y) - S(X,Z)$$
$$+S(X,Y,Z) - S(Y,Z) = I(X,Y) + I(X,Z) + I(Y,Z) - I(X,Y,Z). \quad (67)$$

Interaction information measures the influence of a third variable on the amount of information shared between the other two and could be of either sign. When positive, it indicates that the third variable accounts for some of the correlation between the other two, that is its knowledge diminishes the correlation. When negative, it indicates that the knowledge of the third variable facilitates the correlation between the other two. Alternatively, one may say that a positive $II(X,Y,Z)$ measures redundance in the information about the third variable contained in the other two separately, while negative one measures synergy which is the extra information about Y received by knowing X and Z together, instead of separately.

For example, a channel with input X, noise Z and output Y corresponds to $I(X,Z) = 0$ and $I(X,Z|Y) > 0$, that is $II(X,Y,Z) < 0$. Indeed, once you know the output, the unknown noise and input are related. Love triangles can be either redundant or synergetic (information-wise). If Y dates either X, both X, Z or none, then the dating states of X and Z are correlated. Knowing one tells us more about another (chooses from more possibilities) when the state of Y is not known than when it is: $I(X,Z) > I(X,Z|Y)$. On the contrary, if Y can date with equal probability one, another, both or none, the states of X and Z are uncorrelated, but the knowledge of Y induces correlation between X, Z: if we know that Y presently dates, then it is enough to know that X does not to conclude that Z does. Note that $II(X,Y,Z)$ is symmetric.

Capturing dependencies by using structured groupings can be generalized for arbitrary number of variables as follows:

$$I_n = \sum_{i=1}^n S(X_i) - \sum_{ij} S(X_i,X_j) + \sum_{ijk} S(X_i,X_j,X_k)$$
$$- \sum_{ijkl} S(X_i,X_j,X_k,X_l) + \ldots + (-1)^{n+1} S(X_1,\ldots,X_n)\,. \quad (68)$$

Entropy, mutual information and interaction information are the first three members of that hierarchy.

An important property of both relative entropy and all I_n for $n > 1$ is that they are independent of the additive constants in the entropies that is of the choice of units or bin sizes.

Connections to Statistical Physics. The second law of thermodynamics is getting trivial from the perspective of mutual information. We have seen in Section 2.1 that even when we follow the evolution with infinite precision, the full N-particle entropy is conserved, but one particle entropy grows. Now we see that there is no contradiction here: subsequent collisions impose more and more correlation between particles, so that mutual information growth compensates that of one-particle entropy. Indeed, the thermodynamic entropy of the gas is the sum of entropies of different particles $\sum S(p_i, q_i)$. In the thermodynamic limit we neglect inter-particle correlations, which are measured by the generalized (multi-particle) mutual information $\sum_i S(p_i, q_i) - S(p_1 \dots p_n, q_1, \dots q_n) = I(p_1, q_1; \dots; p_n, q_n)$. Deriving the Boltzmann kinetic equation (25) in Section 2.1, we replaced two-particle probability by the product of one-particle probabilities. That gave the H-theorem, that is the growth of the thermodynamic (uncorrelated) entropy. Since the Liouville theorem guarantees that the phase volume and the true entropy $S(p_1 \dots p_n, q_1, \dots q_n)$ do not change upon evolution, then the increase of the uncorrelated part must be compensated by the increase of the mutual information. In other words, one can replace the usual second law of thermodynamics by the law of conservation of the total entropy (or information): the increase in the thermodynamic (uncorrelated) entropy is exactly compensated by the increase in correlations between particles expressed by the mutual information. The usual second law then results simply from our renunciation of all correlation knowledge, and not from any intrinsic behavior of dynamical systems. Particular version of such renunciation has been presented in Section 2.2: the full N-particle entropy grows because of phase-space mixing and continuous coarse-graining.

But the mutual information can work in opposite direction too. Imagine that two systems were at respectively T_1 and T_2, heat dE_1 passed from 2 to 1, and also that the degree of their correlation changed by ΔI. The second law then generalizes (6) to

$$\left(\frac{1}{T_1} - \frac{1}{T_2}\right) dE_1 - \Delta I \geq 0 \ . \tag{69}$$

If correlations were absent before and appeared when the systems were brought into contact, then $\Delta I > 0$ and we still have heat flowing from hot to cold, its amount bounded from below: $dE_1(T_2 - T_1) \geq T_1 T_2 \Delta I > 0$. However, one can create a situation, where there was initial correlation between the systems and it was destroyed during the heat exchange, that is $\Delta I < 0$. In this case, the heat could flow from cold to hot system. An information-theoretic resource can be spent to perform refrigeration. That will be further discussed in connection with Maxwell demon in the next Chapter.

Relative entropy allows also to generalize the second law for non-equilibrium processes. Entropy itself can either increase upon evolution towards thermal equilibrium or decrease upon evolution towards a non-equilibrium state, as seen respectively in Sections 2.2,2.3. However, the relative entropy between the distribution and the steady-state distribution monotonously decreases with time.

4 Applications of Information Theory

My brothers are protons, my sisters are neurons
Gogol Bordello "Supertheory of Supereverything"

This Chapter puts some content into the general notions introduced above. Choosing out of enormous variety of applications, I tried to balance the desire to show beautiful original works and the need to touch diverse subjects to let you recognize the same ideas in different contexts. The Chapter is concerned with practicality no less than with optimality; we often sacrifice the latter for the former. The simplest and probably the most important lesson we learn here is that looking for a conditional entropy maximum is a universal approach, not restricted to thermal equilibrium.

4.1 The whole truth and nothing but the truth

So far, we defined entropy and information via the distribution. In practical applications, however, the distribution $\rho(x,t)$ is usually unknown and we need to guess it from some data. Information theory supplies a systematic way of guessing, making use of partial information. It is assumed to be given as $\langle R_j(x,t)\rangle = \int \rho(x,t) R_j(x,t)\, dx = r_j(t)$, i.e. as the ensemble averages of some dynamical quantities including normalization, $R_0 = \int \rho(x,t)\, dx = r_0 = 1$. How to get the best guess for $\rho(x,t)$, based on that information? Before, we used to find thermal equilibrium distribution looking for a conditional entropy maximum. But now we want to treat any distribution; among the parameters that we measure could be currents, gradients and other signs of non-equilibrium. Yet the approach is essentially the same. There are infinitely many distributions that contain *the whole truth* (i.e. are compatible with all the given information). Our distribution must also contain *nothing but the truth* that is it must maximize the missing information, which is the entropy $S = -\langle \ln \rho \rangle$. This is to provide for the widest set of possibilities for future use, compatible with the existing information. Looking for the extremum of

$$S + \sum_j \lambda_j \langle R_j(x,t) \rangle = \int \rho(x,t) \Big\{ -\ln[\rho(x,t)] + \sum_j \lambda_j R_j(x,t) \Big\} \, dx \ ,$$

we differentiate it with respect to $\rho(x,t)$ and obtain the equation $\ln[\rho(x,t)] = \sum_j \lambda_j R_j(x,t)$ which gives the distribution

$$\rho(x,t) = \exp\Big[\sum_j \lambda_j R_j(x,t)\Big] = Z^{-1} \exp\Big[\sum_{j\geq 1} \lambda_j R_j(x,t)\Big] \ . \qquad (70)$$

The normalization factor

$$Z(\lambda_i) = e^{-\lambda_0} = \int \exp\Big[\sum_{j=1} \lambda_j R_j(x,t)\Big] \, dx \ ,$$

can be expressed via the measured quantities by using

$$\frac{\partial \ln Z}{\partial \lambda_i} = r_i \ . \qquad (71)$$

The distribution (71) corresponds to the entropy extremum, but how we know that it is the maximum? Positivity of relative entropy proves that. Indeed, consider any other normalized distribution $g(x)$ which satisfies the constraints: $\int dx\, g(x) R_j(x) = r_j$. Then

$$\int dx\, g \ln \rho = \sum_j \lambda_i r_j - \ln Z = \int dx\, \rho \ln \rho = -S(\rho)$$

so that

$$S(\rho) - S(g) = -\int dx (g \ln \rho - g \ln g) = \int dx\, g \ln(g/\rho) = D(g|\rho) \geq 0 \ .$$

Gibbs distribution is (71) with R_1 being energy. When it is the kinetic energy of molecules, we have Maxwell distribution; when it is potential energy in an external field, we have Boltzmann distribution. For our initial "candy-in-the-box" problem (think of an impurity atom in a lattice if you prefer physics), let us denote the number of the box with the candy j. Different attempts give different j but on average after many attempts we find, say, the mean value $\langle j\rangle = r_1$. The distribution giving maximal entropy for a fixed mean is exponential, which in this case is the geometric distribution: $\rho(j) = (1-p)p^j$, where $p = r_1/(1+r_1)$ (home exercise). Similarly, if we scatter on the lattice X-ray with wavenumber k and find $\langle\cos(kj)\rangle = 0.3$, then

$$\rho(j) = Z^{-1}(\lambda)\exp[-\lambda\cos(kj)]$$

$$Z(\lambda) = \sum_{j=1}^{n}\exp[\lambda\cos(kj)]\,, \quad \langle\cos(kj)\rangle = d\log Z/d\lambda = 0.3\ .$$

We can explicitly solve this for $k \ll 1 \ll kn$ when one can approximate the sum by the integral so that $Z(\lambda) \approx nI_0(\lambda)$ where I_0 is the modified Bessel function. Equation $I_0'(\lambda) = 0.3I_0(\lambda)$ has an approximate solution $\lambda \approx 0.63$.

Note in passing that the set of equations (72) may be self-contradictory or insufficient so that the data do not allow to define the distribution or allow it non-uniquely. For example, consider $R_i = \int x^i\rho(x)\,dx$ for $i = 0,1,2,3$. Then (71) cannot be normalized if $\lambda_3 \neq 0$, but having only three constants $\lambda_0, \lambda_1, \lambda_2$ one generally cannot satisfy the four conditions. That means that we cannot reach the entropy maximum, yet one can prove that we can come arbitrarily close to the entropy of the Gaussian distribution $\ln[2\pi e(r_2-r_1^2)]^{1/2}$.

If, however, the extremum is attainable, then the information still missing after the measurements can be computed from (71): $S\{r_i\} = -\sum_j \rho(j)\ln\rho(j)$. It is analogous to thermodynamic entropy as a function of (measurable) macroscopic parameters. It is clear that S have a tendency to decrease whenever we add a constraint by measuring more quantities R_i.

If we know the given information at some time t_1 and want to make guesses about some other time t_2 then our information generally gets less relevant as the distance $|t_1 - t_2|$ increases. In the particular case of guessing the distribution in the phase space, the mechanism of loosing information is due to separation of trajectories described in Sect. 2.2. Indeed, if we know that at t_1 the system was in some region of the phase space, the set of trajectories started at t_1 from this region generally fills larger and larger regions as $|t_1 - t_2|$ increases. Therefore, missing information (i.e. entropy) increases with $|t_1 - t_2|$. Note that it works both into the future and into the past. Information approach allows one to see clearly that there is really no contradiction between the reversibility of equations of motion and the growth of entropy.

Yet there is one class of quantities where information does not age. They are integrals of motion. A situation in which only integrals of motion are known is called equilibrium. When we leave system alone, all currents dissipate and gradients diffuse. The distribution (71) then takes the equilibrium form, either canonical (21) if environment temperature is known, or microcanonical if only total energy is known.

From the information point of view, the statement that systems approach thermal equilibrium is equivalent to saying that all information is forgotten except the integrals of motion. If, however, we possess the information about averages of quantities that are not integrals of motion and those averages do not coincide with their equilibrium values then the distribution (71) deviates from equilibrium. Examples are fluxes and gradients.

Traditional way of thinking is operational: if we leave the system alone, it is in equilibrium; we need to act on it to deviate it from equilibrium. Informational interpretation lets us to see it in a new light: If we leave the system alone, our ignorance about it is maximal and so is the entropy, so that the system is in thermal equilibrium. When we act on a system in a way that gives us more knowledge of it, the entropy is lowered, and the system deviates from equilibrium.

We see that looking for the distribution that realizes the entropy extremum under given constraints is a universal powerful tool whose applicability goes far beyond equilibrium statistical physics. It is essentially a common sense expressed via the simple mathematics. A beautiful example of using this approach is obtaining the statistical distribution of the ensemble of neurons. In a small window of time, a single neuron either generates an action potential or remains silent, and thus the states of a network of neurons are described naturally by binary vectors $\sigma_i = \pm 1$. Most fundamental results of measurements are the mean spike probability for each cell $\langle \sigma_i \rangle$ and the matrix of pairwise correlations among cells $\langle \sigma_i \sigma_j \rangle$. One can successfully approximate the probability distribution of σ_i by maximum entropy distribution (71) that is consistent with the two results of the measurement. The probability distribution of the neuron signals that maximizes entropy is as follows:

$$\rho(\{\sigma\}) = Z^{-1} \exp\left[\sum_i h_i \sigma_i + \frac{1}{2}\sum_{i<j} J_{ij}\sigma_i\sigma_j\right] \;, \qquad (72)$$

where the Lagrange multipliers h_i, J_{ij} have to be chosen so that the averages $\langle\sigma_i\rangle$, $\langle\sigma_i\sigma_j\rangle$ in this distribution agree with the experiment. Such models bear the name Ising in physics, where they were first used for describing systems of spins (the model was formulated by Lenz in 1920 and solved in one dimension by his student Ising in 1925). The distribution (73) corresponds to the thermal equilibrium in the respective Ising model, yet it describes the brain activity, which is apparently far from thermal equilibrium (unless the person is brain dead). More in Appendix 8.5 .

4.2 Exorcizing Maxwell demon

> Demon died when a paper by Szilárd appeared, but it continues
> to haunt the castles of physics as a restless and lovable poltergeist.
> P Landsberg, quoted from Gleick "The Information"

Making a measurement R one changes the distribution from $\rho(x)$ to $\rho(x|R)$, which has its own *conditional* entropy

$$S(x|R) = -\int dxdR\,\rho(R)\rho(x|R)\ln\rho(x|R) = -\int dxdR\,\rho(x,R)\ln\rho(x|R)\,.$$

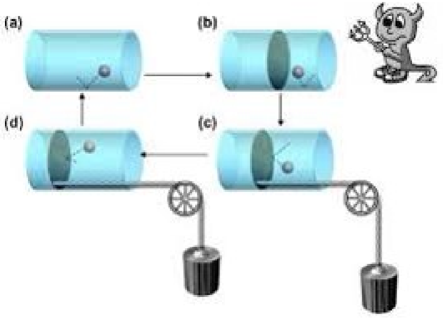
The conditional entropy quantifies my remaining ignorance about x once I know R. Measurement decreases the entropy of the system by the mutual information (51,53) — that how much information about x one gains:

$$\begin{aligned} S(x) - S(x|R) &= \int \rho(x|R)\ln\rho(x|R)\,dxdR - \int \rho(x)\ln\rho(x)\,dx \\ &= \int \rho(x,R)\ln\frac{\rho(x,R)}{\rho(x)\rho(R)}\,dxdR = S(x) + S(R) - S(x,R) = \Delta I\,. \quad (73) \end{aligned}$$

But all our measurements happen in a real world at a finite temperature. Does it matter? Yes, it determines the energy cost of measurements. Assume that our system is in contact with a thermostat having temperature T, which by itself does not mean that our system is in thermal equilibrium (as, for instance, a current-carrying conductor). We then can define a free energy $F(\rho) = E - TS(\rho)$. The Gibbs-Shannon entropy (41) and the mutual information (51,74) can be defined for arbitrary distributions. If the measurement does not change energy (like the knowledge in which half of the box the particles is), then the entropy decrease (74) increases the free energy that is the total work we are able to do. The first law of thermodynamics then requires that the minimal work to perform such a measurement is $F(\rho(x|R)) - F(\rho(x)) = T[S(x) - S(x|R)] = T\Delta I$.

Thermodynamics interprets F as the energy we are *free* to use keeping the temperature. Information theory reinterprets that in the following way: If we knew everything, we can possibly use the whole energy (to do work); the less we know about the system, the more is the missing information S and the less work we are able to extract. In other words, the decrease of $F = E - TS$ with the growth of S measures how available energy decreases with the loss of information about the system. Maxwell understood that already in 1878: "Suppose our senses sharpened to such a degree that we could trace molecules as we now trace large bodies, the distinction between work and heat would vanish."

The concept of entropy as missing information[14] (Brillouin 1949) allows one to understand that Maxwell demon or any other information-processing device do not really decrease entropy. Indeed, if at the beginning one has an information on position or velocity of any molecule, then the entropy was less by this amount from the start; after using and processing the information the entropy can only increase. Consider, for instance, a particle in the box at a temperature T. If we know in which half it is, then the entropy (the logarithm of *available* states) is $\ln(V/2)$. That teaches us that information has thermodynamic (energetic) value at a finite temperature: by placing a piston at the half of the box and allowing particle to hit and move it we can get the work $T\Delta S = T \ln 2$ out of thermal energy of the particle:



Energy conservation tells that to get such an information one must make a measurement whose minimum energetic cost at fixed temperature is $W_{meas} = T\Delta S = T \ln 2$ (that was realized by Szilard in 1929 who also introduced "bit" as a unit of information). Such work needs to be done for any entropy change by a measurement (74).

That guarantees that we cannot break the first law of thermodynamics. But our work of lifting the weight was done at the expense of the thermal

[14] that entropy is not a property of the system but of our knowledge about the system

energy of the system, that is we just turned heat into work. Indeed, hitting the moving piston, particle looses momentum and energy, which it replenishes back to T by hitting the walls with that temperature provided by the environment. We can then do the measurement using this work extracted from heat. Can we break the second law constructing a perpetuum mobile of the second kind, regularly using the thermal energy of the environment to do work and measuring particle position? To answer the question, we need to account for the fact that our demonic engine now includes both the working system A and the measuring device M. For ideal (or demonic) observer, which does not change its state upon measurements, the entropy change is the difference between the entropy of the system $S(A)$ and the entropy of the system interacting with the measuring device $S(A|M)$, that is *the mutual information* defined in the Section 3.4. When there is also a change in the free energy $-\Delta F_M$ of the measuring device, the measurement work could be less than the mutual information:

$$W_{meas} \geq T\Delta S - \Delta F_M = T[S(A) - S(A|M)] - \Delta F_M \ . \tag{74}$$

However, to make a full thermodynamic cycle, we need to return the demon's memory to the initial state. What is the energy price of *erasing* information? Such erasure involves compression of the phase space and is irreversible. For example, to erase information in which half of the box the particle is, we may compress the box to move the particle to one half irrespective of where it was. That compression decreases entropy and is accompanied by the heat $T\ln 2$ released from the system to the environment. If we want to keep the temperature of the system, we need to do exactly that amount of work compressing the box (Landauer 1961). In other words, demon cannot get more work from using the information than we must spend on erasing it to return the system to the initial state (to make a full cycle).

$$W_{eras} \geq \Delta F_M \ . \tag{75}$$

Together, the energy price of the cycle is again the mutual information:

$$W_{eras} + W_{meas} \geq T[S(A) - S(A|M)] = TI(A, M) \ , \tag{76}$$

Thermodynamic energy cost of measurement and information erasure depends neither on the information content nor on the free-energy difference; rather the bound depends only on the mutual correlation between the measured system and the memory. Inequality (77) expresses the trade off between the work required for erasure and that required for measurement: when one is smaller, the other one must be larger. Let us stress that information acquisition and processing have no intrinsic, irreducible thermodynamic cost whereas the seemingly trivial act of information destruction does have a cost. The relations (75,76,77) are versions of the second law of thermodynamics, in which information content and thermodynamic variables are treated on an equal footing.

Similarly, in the original Maxwell scheme, the demon observes the molecules as they approach the shutter, allowing fast ones to pass from A to B and slow ones from B to A. This is one way to use information to transfer heat from cold to hot, as described by (70). Creation of the temperature difference with a negligible expenditure of work lowers the entropy precisely by the amount of information that the demon collected. Erasing this information will also require work.

Landauer's principle not only exorcizes Maxwell's demon, but also imposes the fundamental physical limit on computations. Performing standard operations independent of their history requires irreversible acts (which do not have single-valued inverse). Any Boolean function that maps several input states onto the same output state, such as AND, NAND, OR and XOR, is logically irreversible. When a computer does logically irreversible operation at a finite temperature, the information is erased and heat must be generated. It is worth stressing that one cannot make this heat arbitrarily small making the process adiabatically slow: $T\ln 2$ per bit is the minimal amount of dissipation to erase a bit at a fixed temperature[15].

Take-home lesson: information is physical. We can get extra work out of it, for instance, improving the efficiency of thermal engines beyond the Carnot limit. Processing information without storing an ever-increasing amount of it must be accompanied by a finite heat release at a finite temperature. Of course, any real device dissipates heat just because it works at a finite rate. Lowering that rate one lowers the dissipation rate too. The message is that no matter how slowly we process information, we cannot make the dissipation rate lower than $T\ln 2$ per bit. This is in distinction from usual thermodynamic processes where there is no information processing involved and we can make heat release arbitrarily small making the process slower.

[15] In principle, any computation can be done using only reversible steps, thus eliminating the need to do work (Bennett 1973). That will require the computer to reverse all the steps after printing the answer.

4.3 Renormalization group and the art of forgetting

Erase the features Chance installed,
and you will see the world's great beauty.
A Blok

Erase the features Chance installed.
Watch by chance do not rub a hole.
V Nekrasov[16]

Not only economics and biology, but also physics deal with what are essentially large-scale low-resolution effective theories. Even what was once considered elementary particles is described now as low-energy large-scale excitations of fields whose microscopic behavior is generally unknown (say, at the Planck scale to be introduced in Section 6.6). The most fundamental question is which information about the microscopic properties determines the observable macroscopic behavior, and which is irrelevant and can be forgotten. One of the most fruitful ideas of the 20-th century is to look how one looses information step by step and what universal features appear in the process. We were loosing information about microscopic properties in Section 2.2 applying coarse-graining and treating only finite regions of phase-space. We can also do that explicitly by averaging over small-scale fluctuations or some other degrees of freedom. A general formalism which describes how to reduce description keeping only most salient features is called the renormalization group (RG). It consists in subsequently eliminating degrees of freedom, renormalizing remaining ones and looking for fixed points of such a procedure. There is a shift of paradigm brought by the renormalization group approach. Instead of being interested in this or that probability distribution, we are interested in different RG-flows in the space of distributions. Whole families (universality classes) of different systems described by different distributions flow under RG transformation to the same fixed point i.e. have the same asymptotic distribution.

As almost everything in this course, the simplest realization of RG refers to summing independent random numbers, the procedure described in detail in Appendix 8.2. Let us do summation step by step, summing two numbers at every step. Consider a set of random iid variables $\{x_1 \dots x_N\}$, each having the probability density $\rho(x)$ with zero mean and unit variance. The two-step RG reduces the number of random variables by replacing any two of them by their sum and re-scales the sum to keep the variance: $z_i = (x_{2i-1} + x_{2i})/\sqrt{2}$.

[16] Translated from Russian by A. Shafarenko

Since summing doubles the variance we divided by $\sqrt{2}$. The new random variables each has the following distribution:

$$\rho'(z) = \int dxdy\rho(x)\rho(y)\delta\left(z - \frac{x+y}{\sqrt{2}}\right) \ . \tag{77}$$

The distribution which does not change upon such procedure is called fixed point (even though it is not a point but rather a whole function) and satisfies the equation

$$\rho(x) = \sqrt{2}\int dy\rho(y)\rho(\sqrt{2}x - y) \ .$$

Since this is a convolution equation, the simplest is to solve it by the Fourier transform, $\rho(k) = \int \rho(x)e^{ikx}dx$. Multiplying by $e^{ikx\sqrt{2}}$ and integrating, we get

$$\rho(k\sqrt{2}) = \rho^2(k) \ . \tag{78}$$

In other words, $\rho(k)$ is the generating function, which is multiplied upon summation of independent variables. The solution of (79) is $\rho_0(k) \sim e^{-k^2}$ and $\rho_0(x) = (2\pi)^{-1/2}e^{-x^2/2}$. We thus have shown that the Gaussian distribution is a fixed point of repetitive summation and re-scaling of random variables, keeping variance fixed. This is not surprising, since it has a maximal entropy among the distributions with the same variance.

To turn that into the central limit theorem, we need also to show that this distribution is stable, that is RG indeed flows towards it. Let us analyze the flow near the fixed point, where we denote $\rho = \rho_0(1+h)$ and linearize the transform in h. The transformed distribution is then $h'(k) = 2h(k/\sqrt{2})$. The eigenfunctions of the linearized transform are $h_m(k) = k^m$ with eigenvalues $2^{1-m/2} = h'_m(k)/h_m(k)$. We see that the modes with $m = 0, 1$ grow while the mode with $m = 2$ do not decay. Fortunately, these three modes are forbidden by the three conservation laws of the transformation (78): the moments $\int x^n\rho(x)\,dx$ must be preserved for $n = 0$ (normalization), $n = 1$ (zero mean) and $n = 2$ (unit variance). The moments of $\rho(x)$ are derivatives of the generating function $\rho(k)$ at $k = 0$: $\int x^n\rho(x)\,dx = d^n\rho(k)/d(ik)^n_{k=0}$. Therefore, the three conservation laws mean that $h(0) = dh(0)/dk = d^2h(0)/dk^2 = 0$, so that only $m > 2$ are admissible. All the admissible perturbations decay upon RG flow, that is deviations from the fixed point decrease ,which means that the point is stable.

To conclude, the RG-flow eventually brings us to the distribution with the maximal entropy, forgetting all the information except the invariants - normalization, the mean and the variance.

Another natural transformation is replacing a pair by their mean $z_i = (x_{2i-1} + x_{2i})/2$. The fixed point of this distribution satisfies the equation

$$\rho(z) = \int \rho(x)\rho(y)\delta(z - x/2 - y/2)\, dxdy \;\Rightarrow \rho(k) = \rho^2(k/2)\,.$$

It has the solution $\rho(k) = \exp(-|k|)$ and $\rho(x) = (1 + x^2)^{-1}$, which is the Cauchy distribution mentioned in Section 8.2. In this case, the distribution has an infinite variance, and RG preserves only the mean (which is zero) and normalization. More generally, one can consider a family of re-scaling rules, $z_i = (x_{2i-1} + x_{2i})/2^\mu$, and obtain the family of universal distributions $\rho(k) = \exp(-|k|^\mu)$, characterized by the parameter.

When we look for limiting distributions in the real world, we often need to deal not with independent but with strongly correlated random variables. Let us consider the Ising model of interacting spins and describe the procedure of block spin transformation. The model was mentioned in Section 4.1: random variables are spins $\sigma_i = \pm 1$. To eliminate small-scale degrees of freedom, we divide all the spins into groups (blocks). It is natural to group into blocks the most strongly correlated spins. In neuron systems (73) correlation is not necessarily related to spatial proximity. Here we consider physical systems where strongest correlations are with the nearest neighbors. In this case, there are k^d spins in every block with the side k (d is space dimensionality). We then assign to any block a new variable σ' which is ± 1 when respectively the spins in the block are predominantly up or down. We *assume* that the system can be described equally well in terms of block spins with the distribution of the same form as original but with renormalized parameters.

Consider first a one-dimensional chain, where the Gibbs distribution is $\rho(\{\sigma_i\}) = Z^{-1} \exp\left(-K \sum_i \sigma_i\sigma_{i+1}\right)$, and $K = 1/T$ is the parameter which will be renormalized. The partition function is easy to compute by summing not over N spins but over the $N - 1$ bonds between them. A bond brings ether factor e^K when two spins have the same sign or e^{-K} when the signs are different. For a chain with open ends, we have also two possible values at the ends, which gives

$$Z(K) = \sum_{\{\sigma=\pm 1\}} \exp\Big[K \sum_i \sigma_i\sigma_{i+1}\Big] = 2(2\cosh K)^{N-1}\,.$$

Let us transform the partition function by the procedure (called decimation[17]) of eliminating degrees of freedom by ascribing (undemocratically) to

[17] the term initially meant putting to death every tenth soldier of a Roman army regiment

every block of $k = 3$ spins the value of the central spin. Consider two neighboring blocks $\sigma_1, \sigma_2, \sigma_3$ and $\sigma_4, \sigma_5, \sigma_6$ and sum over all values of σ_3, σ_4 keeping $\sigma'_1 = \sigma_2$ and $\sigma'_2 = \sigma_5$ fixed. The respective factors in the partition function can be written as follows: $\exp[K\sigma_3\sigma_4] = \cosh K + \sigma_3\sigma_4 \sinh K$, which is true for $\sigma_3\sigma_4 = \pm 1$. Denote $x = \tanh K$. Then only the terms with even powers of σ_3 and σ_4 contribute the factors in the partition function that involve these degrees of freedom:

$$\begin{aligned}
&\sum_{\sigma_3,\sigma_4=\pm 1} \exp[K(\sigma'_1\sigma_3 + \sigma_3\sigma_4 + \sigma_4\sigma'_2)] \\
&= \cosh^3 K \sum_{\sigma_3,\sigma_4=\pm 1} (1 + x\sigma'_1\sigma_3)(1 + x\sigma_4\sigma_3)(1 + x\sigma'_2\sigma_4) \\
&= 4\cosh^3 K(1 + x^3\sigma'_1\sigma'_2) = e^{-g(K)} \cosh K'(1 + x'\sigma'_1\sigma'_2)\,, \qquad (79) \\
&g(K) = \ln\left(\frac{\cosh K'}{4\cosh^3 K}\right)\,. \qquad (80)
\end{aligned}$$

The expression (80) has the form of the Boltzmann factor $\exp(K'\sigma'_1\sigma'_2)$ with the re-normalized constant $K' = \tanh^{-1}(\tanh^3 K)$ or $x' = x^3$ — this formula and (81) are called recursion relations. The partition function of the whole system in the new variables can be written as

$$\sum_{\{\sigma'\}} \exp\Big[-g(K)N/3 + K' \sum_i \sigma'_i\sigma'_{i+1}\Big]\,.$$

The term proportional to $g(K)$ represents the contribution into the free energy of the short-scale degrees of freedom which have been averaged out. This term does not affect the statistics of the remaining variables, which is determined by the renormalization of the constant, $K \to K'$. Let us discuss this renormalization. Since $K \propto 1/T$ then $T \to \infty$ correspond to $x \to 0+$ and $T \to 0$ to $x \to 1-$. One is interested in the set of the parameters which does not change under the RG, i.e. represents a fixed point of this transformation. Both $x = 0$ and $x = 1$ are fixed points of the transformation $x \to x^3$. The first one corresponds to the flat distribution with a zero mean, that is to a disordered state. The second one corresponds to the delta-function peaked either at $\sigma_i = 1$ or $\sigma_i = -1$ for all i. The first fixed point is stable and the second one unstable. Indeed, iterating the process for $0 < x < 1$, we see that x approaches zero and effective temperature infinity. That means

that run from a battlefield.

that large-scale degrees of freedom are described by the distribution with the effective temperature so high that the system is in a disordered (paramagnetic) state. It is in agreement with the general argument on impossibility of long-range order in one-dimensional systems with short-range interaction because any overturned spin breaks the correlation between left and right parts. For however small yet finite temperature, the distance to the next overturned spin is e^{-K}, that is finite. At this limit we have $K, K' \to 0$ so that the contribution of the small-scale degrees of freedom is getting independent of the temperature: $g(K) \to -\ln 4$. We see that spatial re-scaling leads to the renormalization of temperature: spin system looks hotter when viewed with less resolution.

Similarly, we may sum over every second spin which gives the recursive relation $\tanh K' = \tanh^2 K$. It corresponds to different steps, but the same flow and the same fixed points.

What entropic measure monotonically changes along that RG quantifying the irreversibility of forgetting? In other words, how to show that RG flow is gradient-like, that is irreversibly sliding down some potential slope? Eliminating some degrees of freedom necessary decreases the entropy of the whole system even when RG moves us towards more disordered state. Then it is more natural to be interested in the entropy per spin or in the mutual information between eliminated and remaining degrees of freedom. These two entropic measures are related. For RG, we can define the mutual information between two sub-lattices: eliminated and remaining. The positivity of the mutual information then implies the monotonic growth of the entropy per spin, $h(K) = \lim_{N\to\infty} S(K, N)/N$. Indeed, consider, for instance, the RG eliminating every second spin, $N \to N/2$, and renormalizing the coupling constant by $K \to K'$. Subtracting the entropy of the original lattice from the sum of the entropies of two identical sub-lattices gives the mutual information: $I = 2S(N/2, K') - S(N, K) = N[h(K') - h(K)] \geq 0$. That shows that in 1d the entropy per block spin growth with the block size upon RG at any distance from the fixed point.

Let us now consider a finite N, but come close to a fixed point. In a finite system with short-range correlations, the entropy for large N is generally as follows:

$$S(N) = hN + C\,, \qquad I = N[h(K') - h(K)] + 2C' - C\ . \tag{81}$$

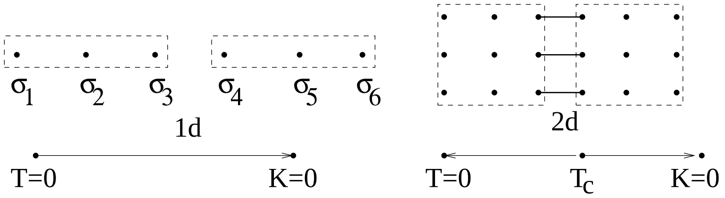
We now have two characteristics, h and C. In a fixed point, the extensive terms in I cancel and $I = C > 0$. This is why C is called *excess entropy*. One can explain positivity of C saying that a finite system appears more random than it is, since we haven't seen yet all the possible correlations.

Mutual information also naturally appears in the description of the information flows in the real space. Let us break the 1d N-chain into two parts, M and $N-M$. The mutual information between two parts of the chain (or between the past and the future of a message) is as follows: $I(M,N-M) = S(M)+S(N-M)-S(N)$. Here, the extensive parts (linear in M,N) cancel in the limit $N,M \to \infty$. Therefore, such mutual information is equal to C from (82).

After these general arguments, let us now compute h and C for the Ising model. Remind that the entropy is expressed via the partition function as follows:

$$S = \frac{E-F}{T} = T\frac{\partial \ln Z}{\partial T} + \ln Z \ .$$

For the 1d Ising chain, $Z = 2(2\cosh K)^{N-1}$ gives $h = \ln(2\cosh K) - K\tanh K$ and $C = K\tanh K - \ln(\cosh K)$. Upon RG flow, these quantities monotonously change from $h(K) \approx 3e^{-2K}$, $C \approx \ln 2$ at $K \to \infty$ to $h(K) \approx \ln 2$, $C \to 0$ at $K \to 0$. One can interpret this, saying that $C = \ln q$, where q is the degeneracy of the ground state. Indeed, $q = 2$ at the zero-temperature fixed point due to two ground states with opposite magnetization, while $q = 1$ in the fully disordered state. So this mutual information (and the excess entropy) indeed measures how much information per one degree of freedom one needs to specify (for non-integer q, obtained mid-way of the RG flow, one can think of it as viewing the system with finite resolution). Note that the past-future mutual information also serves as a measure of the message complexity (that is the difficulty of predicting the message). Without going into details, note also that C depends on the boundary conditions.



This is still rather trivial in 1d, where RG moves systems towards disorder, so that $K' < K$ and $h(K') > h(K)$. In higher dimensions, the next iteration of decimation cannot be performed, and the picture of RG flow is more interesting. There could exist fixed points (limiting distributions) which describe neither low-temperature fully ordered state nor high-temperature

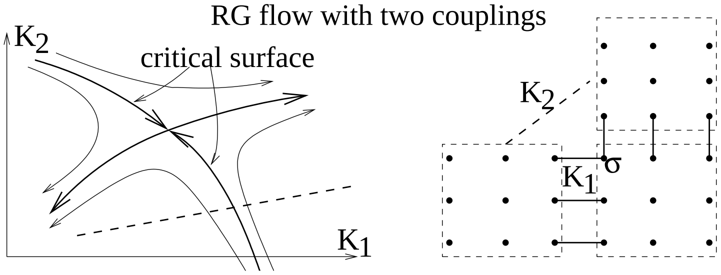


Figure 5: Left: Renormalization group flow with an unstable fixed point. Right: Next-to-nearest neighbor coupling K_2 due to a corner spin.

fully disordered state, but a critical state of the phase transition between the two. The zero-temperature fixed point is unstable in 1d, i.e. K decreases under RG transformation. Yet in the low-temperature region ($x \approx 1, K \to \infty$) it decreases very slow so that it does not change in the main order: $K' = K - const \approx K$. This can be readily interpreted: the interaction between k-blocks is mediated by their boundary spins which all look at the same direction, $K' \approx K\langle\sigma_3\rangle_{\sigma_2=1}\langle\sigma_4\rangle_{\sigma_5=1} \approx K$ (by the same token, at high temperatures $\langle\sigma\rangle \propto K$ so that $K' \propto K^3$). However, in d dimensions, there are k^{d-1} spins at the block side so that $K' \propto k^{d-1}K$ as $K \to \infty$ (in the case $k = 3$ and $d = 2$ we have $K' \approx 3K$, see the right panel of Figure 5). That means that $K' > K$ that is the low-temperature fixed point is stable at $d > 1$. On the other hand, the paramagnetic fixed point $K = 0$ is stable too, so that there must be an unstable fixed point in between at some K_c which corresponds to a critical temperature T_c. In distinction from summing random numbers, we are interested now in unstable fixed point, because it separates regions between two qualitatively different large-scale behavior - ordered and disordered. At a finite temperature, there are always ordered and disordered domains of different scales. At $T > T_c$, looking at larger and larger domains we find them less and less correlated with each other. Yet at $T < T_c$, the mean spins of larger and larger domains are more and more correlated with each other.

Yet we now need to consider RG flows not in the 1d space of K-values, but in multi-dimensional parameter spaces. Already in 2d, summing over corner spin σ produces diagonal coupling between blocks. In addition to K_1, which describes an interaction between neighbors, we need to introduce another parameter, K_2, to account for a next-nearest neighbor interaction.

In fact, RG generates all possible further couplings so that it is a flow in an infinite-dimensional $\mathbf{K}$-space. An unstable fixed point in this space determines critical behavior. The dimensionality of the attractor is determined by the Lyapunov exponents. Negative exponents correspond to the directions in which the flow is converging and erasing information about the microscopic distribution. Positive Lyapunov exponents correspond to unstable directions. To be at criticality, displacements in the unstable directions must be kept zero, which requires tuning respective parameter. We know, however, that we need to control a finite number of parameters to reach a phase transition; for Ising at zero external field and many other systems it is a single parameter, temperature. For all such systems, RG flow has only one unstable direction, all the rest must be contracting stable directions, like the projection on K_1, K_2 plane shown in the left panel of Figure 5. The line of points attracted to the fixed point is the projection of the critical surface, so called because the long-distance properties of each system corresponding to a point on this surface are controlled by the fixed point. The critical surface is a separatrix, dividing points that flow to high-T (paramagnetic) behavior from those that flow to low-T (ferromagnetic) behavior at large scales[18]. See also Appendix 8.7.

We can now understand why physicists are so interested in the critical surface, where the fixed point is actually stable and attractive. That picture of the RG flow explains universality of long-distance critical behavior: different physical systems (in different regions of the parameter $\mathbf{K}$-space) flow to the same fixed point, that is have the same statistics of large-scale correlations and fluctuations. Indeed, changing the temperature in a system with only nearest-neighbor coupling, we move along the line $K_2 = 0$. The point where this line meets critical surface defines K_{1c} and respective T_{c1}. At that temperature, the large-scale behavior of the system is determined by the RG flow i.e. by the fixed point. In another system with nonzero K_2, changing T we move along some other path in the parameter space, indicated by the broken line in the left panel of Figure 5. Intersection of this line with the critical surface defines some other critical temperature T_{c2}. But the long-distance properties of this system at that temperature are determined by the same fixed point.

[18] Mention in passing that in dimensions $d > 4$, the block-spin renormalization of the Ising-class models leads to asymptotic Gaussian distribution $\ln \rho(\eta) \propto -|\nabla \eta|^2$

4.4 Information is life

> What lies at the heart of every living thing is not a fire, not warm breath, not a 'spark of life.' It is information.
>
> Richard Dawkins

One may be excused thinking that living beings consume energy and matter to survive, unless one knows that energy and matter are conserved and cannot be consumed. All the energy, absorbed by plants from sunlight and by us from food, is emitted as heat. Life-sustaining substance is entropy: we consume information and generate entropy by intercepting flows from low-entropy energy sources to high-entropy body heat — just think how much information was processed to squeeze 500 kkal of chemical energy into 100 grams of a chocolate, and you enjoy it even more. For plants, the Sun is a low-entropy energy source due to its high temperature. The same is true for the whole Earth, which exports into space much more entropy than it receives from the Sun. Nor we consume matter, only make it more disordered: what we consume has much lower entropy than what comes out of our excretory system. In other words, we decrease entropy inside and increase it outside of our bodies. Consuming information is our way to resist (temporarily) the second law of thermodynamics and survive.

We have two separate systems for processing information, the genome and the brain.

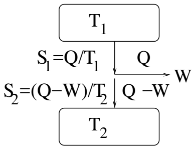
Genome way to stay out of the (most probable) state of thermal equilibrium is to use replication to generate highly ordered (and improbable) structures. The instructions for replication are encoded in genes. What are the error rates in the transmission of the genetic code? Typical energy cost of a mismatched DNA base pair is that of a hydrogen bond, which is about ten times the room temperature. If the DNA molecule was in thermal equilibrium with the environment, thermal noise would cause error probability $e^{-10} \simeq 10^{-4}$ per base. This is deadly. A typical protein has about 300 amino acids, that is encoded by about 1000 bases; we cannot have mutations in every tenth protein. Moreover, synthesis of RNA from DNA template and of proteins on the ribosome involve comparable energies and could cause comparable errors. That means that Nature operates a highly non-equilibrium state, so that bonding involves extra irreversible steps and burning more energy. This way of sorting molecules is called kinetic proofreading (Hopfield 1974, Ninio 1975) and is very much similar to the Maxwell demon discussed in Section 4.2.

Collectively, the evolution as a natural selection is an increasingly efficient encoding of information about the environment in the gene pool of its inhabitants. This process is accelerated by sex, which still provides one of the highest transfer rates of information (even though most of it is discarded). The survivors of natural selection are not the fittest individuals. The ultimate survivor is the information in the genes, which continues to exist long after many its former carriers, individual and species, went extinct.

On another level, nervous system maintains the body integrity consuming information by active inference, as described in Section 4.5. The genome way of information processing is clearly digital; what about the brain? Since neurons often either fire or not a standard pulse, it may seem that information is encoded in binary digits. Indeed, written language and many similar tasks are clearly handled by processing digital information. However, there are reasons to believe that the brain is also an analog device, for instance encoding information in the frequency of pulses, which could be varied continually.

If an elementary act of life as information processing (say, thought) generates ΔS, we can now ask about its energy price. Similar to our treatment of the thermal engine efficiency (1), we assume that one takes Q from the reservoir with T_1 and delivers $Q - W$ to the environment with T_2. Then $\Delta S = S_2 - S_1 = (Q - W)/T_2 - Q/T_1$ and the energy price is as follows:

$$Q = \frac{T_2 \Delta S + W}{1 - T_2/T_1} \ .$$



When $T_1 \to T_2$, the information processing is getting prohibitively ineffective, just like the thermal engine. In the other limit, $T_1 \gg T_2$, one can neglect the entropy change on the source, and we have $Q = T_2 \Delta S + W$. Hot Sun is indeed a low-entropy source.

So how many bits we consume per second? Let us now estimate our rate of information processing and entropy production. An average lazy human being dissipates about $W = 200$ watts of power at $T = 300\,K$. Since the Boltzmann constant is $k = 1.38 \times 10^{-23}$, that gives about $W/kT \simeq 10^{23}$ bits per second. The amount of information processed per unit of subjective time (per thought) is about the same, assuming that each moment of consciousness lasts about a second (Dyson, 1979).

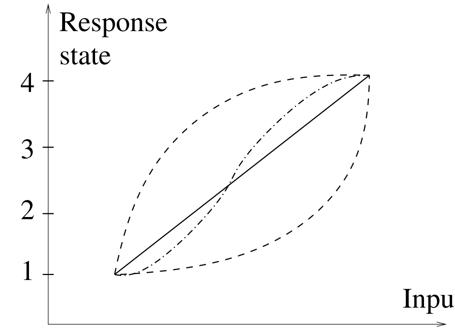
We now discuss how such beings actually process information. Do the Gibbs entropy and the mutual information have any quantitative relation to the way we react to the signals? Yes, they do! When one must react differently to different stimuli, the average choice-reaction time was found experimentally to be linearly proportional to the entropy of the statistical distribution of stimuli (Hick 1952, Hyman 1953). The more is the uncertainty, the longer it takes to recognize the event. For example, when one needs to name the letter or number that appear randomly on a screen, the average response time grows logarithmically with the size of the set. Logarithmic dependence on the set size means that the decision is made by subdividing strategy. Similarly, the time to find an item in an ordered menu grows logarithmically with the menu length, yet it grows linearly when the menu is disordered.

When the number of elements stays constant but the frequencies of their appearances are made unequal thus lowering entropy, the average response time decreases proportionally. Even more remarkably, when experimentalists introduce a correlation between subsequent stimuli, the response time goes down in proportion to the conditional entropy, which is less than unconditional.

One can turn the tables and prescribe the reaction time. As this time is getting shorter, we make more and more errors in naming the objects. How to quantify that? We need to compute the mutual information between the input (number i on a screen) and output (our name j for it). Making more errors means lower mutual information. Experimentally one measures the joint probability $p(i,j)$ from which one obtains the marginal probabilities $p(i) = \sum_j p(i,j)$, $p(j) = \sum_i p(i,j)$ and conditional probability $p(j|i) = p(i,j)/p(i)$. One then computes $S(j) = -\sum_i p(j)\log p(j)$, $S(j|i) = -\sum_{ij} p(i)p(j|i)\log p(j|i)$ and $I(i,j) = S(j) - S(j|i)$. The mutual information was found experimentally to be linearly proportional to the reaction time prescribed. We see how living beings use Boltzmann and Gibbs entropies, as well as the mutual information.

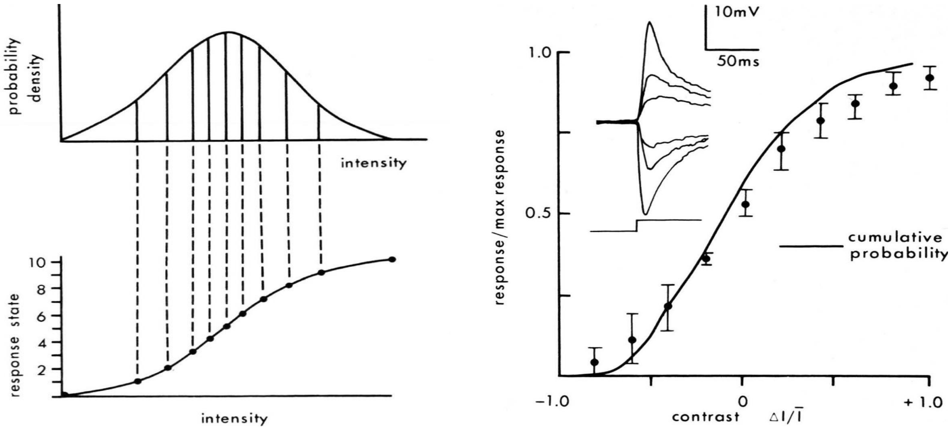
Since the time of processing is proportional to the amount of information, one can conclude that the system works to keep uniform an average amount of information processed per unit time, that is the rate. The next example presents more sophisticated strategy in processing stimuli, where the system maximizes information transfer rate by keeping it uniform through the dynamic range of the signal (such strategies are sometimes called infomax principle).

Maximizing capacity. Imagine yourself on the day five of Creation designing the response function for a sensory system of a living being. Technically, the problem is to choose thresholds for switching to the next level of response, or equivalently, to choose the function of the input for which we take equidistant thresholds. Suppose that we wish to divide the whole perceivable (finite) interval of signals into three regions, encoding them as weak (1,2), medium (2,3) and strong (3,4):



For given value intervals of input and response, should we take the solid line of linear proportionality between response and stimulus? Or choose the lowest curve that treats all low-intensity inputs as weak and amplifies difference in high-intensity signals? The choice depends on the significance of different intervals for survival. For example, the upper curve was actually chosen (on the day six) for the auditory system of animals and humans: our ear senses loudness as the logarithm of the intensity, which amplifies differences in weak sounds and damps strong ones. That way we better hear whisper of a close one and aren't that frightened by loud threats.

If, however, all the input amplitudes are of comparable significance, then the goal could be to maximize the mean information transfer rate (capacity) at the level of a single neuron/channel. In such a case, the response curve (encoding) must be designed by the Creator together with the probability distribution of stimuli. That it is indeed so was discovered in one of the first application of information theory to the real data in biology, namely to processing of visual signals (Laughlin 1981). It was conjectured that maximal-capacity encoding must use all response levels with the same frequency, which requires that the response function is an integral of the probability distribution of the input signals (see Figure). First-order interneurons of the insect eye were found to code contrast rather than absolute light intensity. Subjecting the fly in the lab to different contrasts x, the response function $y = g(x)$ was measured from the fly neurons; the probability density of inputs, $\rho(x)$, was measured across its natural habitat (woodlands and lakeside) using a detector which scanned horizontally, like a turning fly.



The coding strategy for maximizing information capacity by ensuring that all response levels are used with equal frequency. Upper left curve: probability density function for stimulus intensities. Lower left curve: the response function, which ensures that the interval between each response level encompasses an equal area under the distribution, so that each state is used with equal frequency. In the limit where the states are vanishingly small this response function corresponds to the cumulative probability function. Right panel: The contrast-response function of fly neuron compared to the cumulative probability function for natural contrasts. Simon Laughlin, *Naturforsch.* **36**, 910-912 (1981)

We can now explain it noting that the representation with the maximal capacity corresponds to the maximum of the mutual information between input and output: $I(x, y) = S(y) - S(y|x)$. Since we consider a one-to-one relation $y = g(x)$, that is an error-free transmission, then the conditional entropy $S(y|x)$ is zero. Therefore, according to Section 3.4, we need to maximize the entropy of the output assuming that the input statistics $\rho(x)$ is given. Absent any extra constraints except normalization, the entropy for a distribution on a finite interval is maximal when $\rho(y)$ is constant. Indeed, since $\rho(y)dy = \rho(x)dx = \rho(x)dydx/dy = \rho(x)dy/g'(x)$, then

$$\begin{aligned} S(y) &= -\int \rho(x)\ln[\rho(x)/g'(x)]\,dx = S(x) + \langle \ln[g'(x)] \rangle\,, \qquad (82)\\ \frac{\delta S}{\delta g} &= \frac{\partial}{\partial x}\frac{\rho}{g'(x)} = 0 \quad \Rightarrow \quad g'(x) = C\rho(x)\ , \end{aligned}$$

as in the Figure. In other words, we choose equal bins for the variable whose probability is flat. Since the probability $\rho(x)$ is positive, the response function $y = g(x)$ is always monotonic i.e. invertible. Note that our choice of response function is an exact analog of efficient encoding using longer codewords for less frequent letters. In that way, we utilized only the probability distribution of different signal levels, similar to language encoding which utilizes different frequencies of letters (and not, say, their mutual correlations).

We have also applied quasi-static approximation, neglecting dynamics and relating instantaneous values of x and y. Let yourself be impressed by the agreement of theory and experiment — there were no fitting parameters. The same approach works well also for biochemical and genetic input-output relations. For example, the dependence of a gene expression on the level of a transcription factor is dictated by the statistics of the latter. That also works when the conditional entropy $S(y|x)$ is independent of the form of the response function $y = g(x)$. See more details in the Appendix 8.6.

For particular types of signals, practicality may favor non-optimal but simple schemes like amplitude and frequency modulation (both are generally non-optimal but computationally feasible and practical). Even in such cases, the choice is dictated by the information-theory analysis of the efficiency. For example, neuron either fires a standard pulse (action potential) or stays silent, which makes it natural to assume that the information is encoded as binary digits (zero or one) in discrete equal time intervals (which would mean that the brain does digital rather than analog computations). Yet one can imagine that the information is encoded by the time delays between subsequent pulses (since time is continuous, this is more of an analog computation). On the engineer's language, the former method of encoding is a limiting case of amplitude modulation, while the latter case is that of frequency modulation. The maximal rate of information transmission in the former case is only dependent on the minimal time delay between the pulses determined by the neuron recovery time. On the other hand, in the latter case, the rate depends on both the minimal error of timing measurement and of admissible maximal time between pulses. In reality, brain activity "depends in one way or another on all the information-bearing parameters of an impulse — both on its presence or absence as a binary digit and on its precise timing" (MacKay and McCulloch 1952).

4.5 Theory of mind

Who are you going to believe, me, or your own eyes? Marx

And how the sensory information is processed and determines the behavior? An ambitious application of information theory is an attempt to understand and quantify sentient behavior. One idea going back to Helmholts is to view "perception as as unconscious inference". There is evidence that perception of our brain is inferential, that is based on the prediction and hypothesis testing. Among other things, this is manifested by the long known phenomenon of binocular rivalry which occurs when different pictures are presented to the two eyes. Rather than perceiving a stable, single amalgam of the two stimuli, one experiences alternations as the two stimuli compete for perceptual dominance, which can be influenced by priming. Another evidence is the recently established fact that signals between brain and sensory organs travel in both directions simultaneously.

Perception is thus treated not as a bottom-up encoding of sensory states Y into internal neuronal representation of the environmental states X, but as a combination of top-down prior expectation with bottom-up sensory signals. The combined bottom-up-top-down approach makes sense from evolutional and developmental perspectives. Indeed, the bottom-up approach assumes that there is some entity which processes the sensory inputs Y into a picture of the world $P(X|Y)$. Yet where that entity came from? Imagine a brain as a bunch of neurons in a black box receiving electrical signals, which do not carry with them labels "from the retina", "from the liver", "from your grandmother", etc. The best one can do is to send out signals which help you to survive. Since one have managed to survive up to this point, then the right survival strategy is continuation, which presumes receiving more or less the same signals as before.

In this spirit, we try to describe perception as hypothesis testing within the Bayes' framework, introduced in Section 3.7. The mechanics of the sensory system determines $P(Y|X)$, which is the conditional probability of sensory input for a given state of environment. In the example of the fly eye from the Section 4.4, x is a contrast in light intensity and y is the neuron signal. Upon receiving the particular input y, the simplest inference about the environment is that of maximal likelihood: taking the value x that maximizes $P(y|x)$. However, to make a decision or action based on the inference, we need a measure of confidence in the result. That means that our inference must be probabilistic, obtaining the whole posterior probability distribution $P(X|Y)$ — sharp distribution gives high and flat distribution low confidence. To obtain the posterior distribution, we need a prior distribution $P(X)$ and the Bayes' formula (60):

$$P(X|Y) = P(Y|X)P(X)/P(Y) \ . \tag{83}$$

Leaving aside for a while the normalizing factor $P(Y)$, we thus presume that the mind has so-called generative model, represented by the joint distribution $P(X,Y) = P(Y|X)P(X)$. Exact computation by (84) can be impossible or unpractical, for instance, due to necessity to average over many hidden states and variables. It is natural to assume that the brain uses variational approach based on optimizing some tractable proxy. The first thing to account is the degree of surprise or necessary change, characterized by the relative entropy between prior and posterior distributions. Averaged over all X and Y, it is nothing but the mutual information, that is the average information brought by sensory inputs:

$$D[P(X|Y)|P(X)] = \sum_Y P(Y) \sum_X P(X|Y) \log[P(X|Y)/P(X)] = I(X,Y)\,.$$

For perception, however, we shall need to evaluate the change at a given y. Changing beliefs and updating expectations entails a cognitive cost, as we know all too well. More important and probably related: expected states are preferred for survival (fish expects to stay in the water), while surprises are to be avoided. Generative model is strongly biased towards a narrow interval of parameters guaranteeing survival. Yet this natural tendency to minimize the change conflicts with the necessity to accommodate the data. Whenever we encounter a trade-off, free energy negotiates it. The working hypothesis is that for a given y the brain looks for the posterior distribution $Q(X)$ which minimizes the following free energy:

$$\begin{aligned} F[Q,y] &= \sum_x Q(x) \log \frac{Q(x)}{P(x,y)} = -\sum_x Q(x) \log P(x,y) - S(Q) \\ &= D[Q(x)|P(x)] - \sum_x Q \log P(y|x) = \sum_x Q(x) \left[\log \frac{Q(x)}{P(x)} - \log P(y|x)\right] \\ &= D[Q(x)|P(x|y)] - \log P(y)\,. \end{aligned} \tag{84}$$

As clear from the beginning of the first line, it measures the mismatch between the internal generative model $P(x,y)$ and current observation as a functional of $Q(x)$ and a function of y. The three lines suggest three different operational strategies according to the three different interpretations of the same quantity. The first line is written in the form of a free energy, $E/T - S(Q)$, where minus log of prior probability, $\log P(x,y) = -E(x,y)/T$, can be loosely interpreted as the "energy" of some Gibbs distribution (measured in bits, that is divided by the temperature). Minimization requires the trade-off between the energy-imposed "truth" of accounting for prior $P(x,y)$ and "nothing but the truth" maximization of the entropy $S(Q)$. We call energy what needs minimization, while entropy requires maximization. The second line describes the above trade-off between inertia and force of data: the first term on the right is the degree of change, while the second term quantifies the accuracy of data representation — $Q(x)$ must give more weight to those x which provide for higher probability to observe y according to $P(y|x)$, which is given. The third line does not describe any trade-off, but shows that the free energy is bounded from below by the sensory surprise $-\log P(y)$. Only when our variational $Q(x)$ is equal to the exact $P(x|y)$, the free energy reaches its global minimum.

The third line (85) suggests that perceptual inference, that is computing $Q(x)$, is not the only way to minimize $F(Q, y)$; another way is to change the sensory data y. That requires action: one can switch the channel or look the other way rather than change the beliefs. That brings us to the active inference approach, which puts action into perception (Parr, Pezzulo, Friston 2022). The picture is that living beings survive adapting action-perception loop with their environment. That means that every sensory input is not obtained passively, but is predicted by the brain and is solicited by an action intended for the predicted input. Mismatch between predicted and actually received sensory input leads to updating the predictive (generative) model, which then triggers new action leading to new sensory observations better corresponding to expectations. Perception and action are complementary ways to diminish the mismatch. Perception changes your mind replacing prior beliefs by posterior ones, while action changes the world to make it more compatible with the beliefs. To summarize: surprise minimization by active inference is our way to survive.

In particular, our perception of objects is very much determined by the generative model with its prediction of how actions change sensory input (encoded in conditional probability of *what could have happened*). Indeed, even with one eye closed we distinguish a three-dimensional object from its two-dimensional picture despite receiving identical visual signals. The reason is that our brain knows that moving our head will reveal the new parts of the image in the former case, but not in the later.

While still highly hypothetical, this theory finds some empirical support in measurements of the connectivity and activity of neural networks. For example, some connectivity patterns in a motor cortex support the idea of a motor command as a prediction, such that the prediction errors related to body position and motion can be resolved by reflexes without belief updating. Simply speaking, brain can infer the positions of body parts without receiving outside signals. The analysis of the experimental data on brain activity is facilitated by the asymmetry between descending signals carrying expectations and ascending signals bringing prediction errors — the latter involve nonlinear operations generating higher frequencies, which is measurable. And last but least - playing tennis would be impossible if brain just reacted on visual stimuli, since the time between light hitting retina and brain receiving a signal is in excess of 150 msec. The active inference approach is also useful in building models for analyzing data from behavioral experiments and disease processes, drawing inferences about inferences. When top-down signals totally dominate, one has hallucinations; what is considered normal perception could then be called "controlled hallucination".

I think that poetry and music appeal to our ever-predicting mind by creating expectations (using rhythm or melody) and then partially fulfilling and partially breaking them. An optimal mixture of expected and surprising is what makes for a great art, which still waits for its free-energy analysis. Another possible dramatic implication of the active inference approach is a treatment of emotions not solely as fixed universal patterns of brain and body inherited from animal ancestors and triggered by sensory inputs, but as constructed and learnt patterns of prediction and reaction amenable to significant variability and plasticity.

Mention in passing the suggestions to use relative information and mutual entropy for a more ambitious task of quantifying consciousness, understood as processing information from different channels in an integrated way, irreducible to processing information in the channels separately. Such approach is known as integrated information theory (Tononi 2008). Another recent example is the use of mutual information to quantify immersion of a person in an activity and the related rate of success (Melnikoff 2022).

4.6 Rate Distortion and Information Bottleneck

> There's no sense in being precise when you
> don't even know what are you talking about
> von Neumann

When we transfer information, we look for maximal transfer rate and thus define channel capacity as the maximal mutual information between input and output. But when we encode the information, we may be looking for the opposite: what is the *minimal* number of bits, sufficient to encode the data with a given accuracy.

For example, encoding a real number requires infinite number of bits. Representation of a continuous input B by a finite discrete output encoding A generally leads to some distortion, which we shall characterize by the real function $d(A,B)$. How large is the mean distortion, $\mathcal{D}=\sum_{ij}P(A_i,B_j)d(A_i,B_j)$, for a given statistics of B, encoding with R bits and 2^R values? It depends on the choice of the distortion function, which specifies what are the most important properties of the signal B. For Gaussian statistics (which is completely determined by the variance), one chooses the squared error function $d(A,B)=(A-B)^2$. We first learn to use it in the standard least squares approximations — now we can understand why squares and not other powers — because minimizing variance minimizes the entropy of a Gaussian distribution and thus the amount of information needed to characterize it.

Consider a Gaussian B with $\langle B\rangle=0$ and $\langle B^2\rangle=\sigma^2$. If we have one bit to represent it, apparently, the only information we can convey is the sign of B. The simplest is to encode positive/negative regions is by numbers $\pm A$. To minimize squared error, we choose $A=\pm\langle|B|\rangle=\pm\sigma\sqrt{2/\pi}$, which corresponds to

$$\mathcal{D}(1)=2(2\pi)^{-1/2}\int_0^\infty\Bigl(B-\sigma\sqrt{2/\pi}\Bigr)^2\exp[-B^2/2\sigma^2]\frac{dB}{\sigma}=\sigma^2(1-2/\pi)\ . \quad (85)$$

Let us now turn the tables and ask what is the minimal rate R sufficient to provide for distortion not exceeding $\mathcal{D}$. This is called *rate distortion function* $R(\mathcal{D})$. We know that the rate is the mutual information $I(A,B)$, but now we are looking not for its maximum (as in channel capacity) but for the minimum over all the encodings defined by the conditional probabilities $P(B|A)$, such that the distortion does not exceed $\mathcal{D}$. Since $I(A,B)=S(B)-S(B|A)$, then minima of $I(A,B)$ are maxima of $S(B|A)$. It is helpful to think of distortion as produced by the added noise ξ with the variance $\mathcal{D}$. For a fixed variance, maximal entropy $S(B|A)$ corresponds to the Gaussian distribution, so that we have an (imaginary) Gaussian channel with the variance $\langle(B-A)^2\rangle=\mathcal{D}$. Together with the Gaussian input having $\langle B^2\rangle=\sigma^2$, they provide for the minimal rate given by (59):

$$\begin{aligned}R(\mathcal{D})&=I(A,B)=S(B)-S(B|A)=S(B)-S(B-A|A)\\&\geq S(B)-S(B-A)=\tfrac{1}{2}\log_2(2\pi e\sigma^2)-\tfrac{1}{2}\log_2(2\pi e\mathcal{D})=\tfrac{1}{2}\log_2\tfrac{\sigma^2}{\mathcal{D}}\ . \quad (86)\end{aligned}$$

It goes to infinity for $\mathcal{D}\to 0$ and turns into zero for $\mathcal{D}=\sigma^2$. Indeed, for $\mathcal{D}\geq\sigma^2$ we can take $A=0$ with probability one making the mutual information zero - absolute minimum!

Often we need to represent by R bits m independent Gaussian signals with different variances σ_i, $i=1,\dots,m$ — for instance, signals from different spectral intervals. How to divide these bits between signals to minimize the total distortion? We look for the distortions $\mathcal{D}_i$ and minimize $\sum_i \mathcal{D}_i = \mathcal{D}$ under the condition that $\sum_i R(\mathcal{D}_i) = R$. Differentiating $\sum_i [\mathcal{D}_i + \lambda \log_2 \sigma_i^2/\mathcal{D}_i]$ with respect to $\mathcal{D}_i$, we find out that $\mathcal{D}_i$ are all equal. Therefore, all $\mathcal{D}_i = \mathcal{D}/m$, as long as this constant is less than all σ_i. Taking smaller R, we increase $\mathcal{D}$ and reach the moment when $\mathcal{D}/m$ exceeds some σ_j - then we need to take respective $R_j = 0$, that is allocate zero bites to this component. Alternatively, if we managed to decrease enough the variance of some component, it is not treated as fluctuating and does not deserve to be represented (except one bit for its mean if it is nonzero) — such is the logic of rate distortion theory.

One can show that the rate distortion function $R(\mathcal{D})$ is monotonous and convex for all systems. When the distortion is not a quadratic function, the conditional probability of encoding $P(A|B))$ is not Gaussian. In solving practical problems, it must be found solving the variational problem, where one finds a normalized $P(A|B))$, which minimizes the mutual information under the condition of a given mean distortion. For that one minimizes the functional

$$F = I + \beta\mathcal{D} = \sum_{ij} P(A_i|B_j)P(B_j) \left\{ \ln \frac{P(A_i|B_j)}{P(A_i)} + \beta d(A_i, B_j) \right\} \ . \tag{87}$$

After variation with respect to $P(A_i|B_j)$ we obtain

$$P(A_i|B_j) = \frac{P(A_i)}{Z(B_j,\beta)} e^{-\beta d(A_i,B_j)} \ , \tag{88}$$

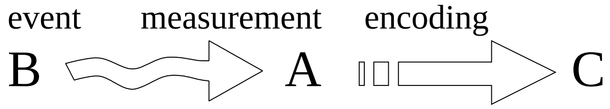
where the partition function $Z(B_j,\beta) = \sum_i P(A_j) e^{-\beta d(A_i,B_j)}$ is the normalization factor. Recall that what is given is $P(B)$, not $P(A)$. The latter must be expressed via the same conditional probability:

$$P(A_i) = \sum_i P(A_i|B_j)P(B_j) \ . \tag{89}$$

The system of equations (89,90) is usually solved by iterations.

Immediate physical analogy is that (89) is a Gibbs distribution with the ”energy” equal to the distortion function. Maximizing entropy for a given energy (Gibbs) is equivalent to minimizing mutual information for a given distortion function. As usual, what is given is in the exponent. Choice of the value of the inverse temperature β reflects our priorities: at small β the conditional probability is close to the unconditional one, that is we minimize information without much regard to the distortion. On the contrary, large β requires our conditional probability to be sharply peaked at the minima of the distortion function.

Similar, but more sophisticated optimization procedures are applied, in particular, in image processing. Images contain enormous amount of information. The rate at which visual data are collected by the photoreceptor mosaic of animals and humans is known to exceed 10^6 bits/sec. On the other hand, studies on the speed of visual perception and reading speeds give numbers around 40-50 bits/sec for the perceptual capacity of the visual pathway in humans. The brain then have to perform huge data compressions. This is possible because visual information is highly redundant due to strong correlations between pixels.



The measured quantity A thus contains too much data of low information value. We wish to compress A to C while keeping as much as possible information about B. Understanding the given signal A requires more than just predicting/inferring B, it also requires specifying which features of the set of possible signals $\{A\}$ play a role in the prediction. Here meaning seeps back into the information theory. Indeed, information is not knowledge (and knowledge is not wisdom). Non-surprisingly, the main tool in automated and AI-assisted pattern recognition in images and other data is the mutual information. We formalize this problem as that of finding a short code for $\{A\}$ that preserves the maximum information about the set $\{B\}$. That is, we squeeze the information that A provides about B through a "bottleneck" formed by a limited set of codewords $\{C\}$. This is reached via the method called Information Bottleneck (Tishby at al 2000) , targeted at characterizing the tradeoff between information preservation (accuracy of relevant predictions) and compression. Here one looks for the minimum of the functional

$$I(C,A) - \beta I(C,B) \ . \tag{90}$$

The coding $A \to C$ is also generally stochastic, characterized by $P(C|A)$. The quality of the coding is determined by the rate, that is by the average number of bits per message needed to specify an element in the codebook without confusion. This number per element A of the source space $\{A\}$ is bounded from below by the mutual information $I(C,A)$ which we thus want to minimize. Effective coding utilizes the fact that mutual information is usually sub-extensive in distinction from entropy which is extensive. Note the difference from the Section 3.4, where in characterizing the channel capacity (upper bound for the error-free rate) we *maximized* $I(A,B)$ over all choices of the source space $\{B\}$, while now we *minimize* $I(C,A)$ over all choices of coding. To put it differently, there we wanted to maximize the information transmitted, now we want to minimize the information processed. This minimization, however, must be restricted by the need to retain in C the relevant information about B which we denote $I(C,B)$.

Having chosen what properties of B we wish to stay correlated with the encoded signal C, we add the mutual information $I(C,B)$ with the Lagrange multiplier $-\beta$ to the functional (91). The sign is naturally chosen such that $\beta > 0$ (analog of inverse temperature), indeed, we want minimal coding $I(A,B)$ preserving maximal information $I(C,B)$, that is $I(C,B)$ is treated similarly to the channel capacity. The single parameter β again represents the tradeoff between the complexity of the representation measured by $I(C,A)$, and the accuracy of this representation, measured by $I(C,B)$. At $\beta = 0$ our quantization is the most sketchy possible — everything is assigned to a single point. At β grows, we are pushed toward detailed quantization. By varying β one can explore the tradeoff between the preserved meaningful information and compression at various resolutions. Comparing with the rate distortion theory functional (89), we recognize that we are looking for the conditional probability of the mapping $P(C|A)$, that is we explicitly want to treat some pixels A_i as more relevant than the others.

However, the constraint on the meaningful information is now nonlinear in $P(C|A)$, so this is a much harder variational problem. Indeed, (91) can be written as follows:

$$\begin{aligned} I(C,A) - \beta I(C,B) &= \sum_{ij} P(C_j|A_i)P(A_i)\ln\frac{P(C_j|A_i)}{P(C_j)} \\ &- \beta\sum_{jk} P(B_k|C_j)P(C_j)\left\{\ln\frac{P(B_k|C_j)}{P(B_k)}\right\} . \end{aligned} \tag{91}$$

The conditional probabilities of A, B under given C are related by the Bayes' rule

$$P(B_k|C_j) = \frac{1}{P(C_j)}\sum_i P(A_i)P(B_k|A_i)P(C_j|A_i) , \tag{92}$$

where the conditional probability of the measurements, $P(B_k|A_i)$, is assumed to be known. The variation of (92) with respect to the encoding conditional probability,

$P(C_j|A_i)$, now gives the equation (rather than an explicit expression):

$$
\begin{aligned}
P(C_j|A_i) = &\quad \tfrac{P(C_j)}{Z(A_i,\beta)} \exp\left[-\beta \textstyle\sum_k P(B_k|A_i) \log \tfrac{P(B_k|A_i)}{P(B_k|C_j)}\right] \\
= &\quad \tfrac{P(C_j)}{Z(A_i,\beta)} \exp\{-\beta D[P(B|A)||P(B|C)]\} \ , \qquad (93)
\end{aligned}
$$

We see that the relative entropy D between the two conditional probability distirbutions emerged as the effective distortion measure $\mathcal{D}$. The system of equations (93,94) is also solved by iterations. For example, one minimizes $I(A,C) + \beta D[P(B|A)||P(B|C)]$ in alternating iterations first over $P(C|A)$, then over $P(C)$, then over $P(B|A)$, then repeating the cycle. Doing compression procedure many times, $A \to C_1 \to C_2 \dots$ is used in multi-layered Deep Learning Algorithms. Here knowledge of statistical physics helps in several ways, particularly in identifying phase transitions (with respect to β) and the relation between processing from layer to layer and the renormalization group: features along the layers become more and more statistically decoupled as the layers gets closer to the fixed point.

Practical problems of iterations and machine learning are closely related to fundamental problems in understanding and describing the biological evolution. Here an important task is to identify classes of functions and mechanisms that are provably evolvable — can logically evolve into existence over realistic time periods and within realistic populations, without any need for combinatorially unlikely events to occur. Quantitative theories of evolution in particular aim to quantify the complexity of the mechanisms that evolved, which is done using information theory.

4.7 Information is money

This section is for those brave souls who decided to leave physics for gambling. If you have read till this point, you must be well prepared for that.

Let us start from the simplest game: you can bet on a coin, doubling your bet if you are right or loosing it if you are wrong. Surely, an intelligent person would not bet money hard saved during graduate studies on a totally random process with a zero gain. You bet only when you have an *information* that sides have unequal probabilities: $p > 1/2$ and $1-p$. To have a steady income and an average growth you want to play the game many times. Shall we look then for the maximal average return? The maximal mean arithmetic growth rate is $(2p)^N$ and corresponds to betting every time all your money on the more probably side. That mean, however, comes from a single all-win realization; the probability of that winning streak goes to zero with growing N as p^N. To avoid loosing it all with probability fast approaching unity, you bet only a fraction f of your money on the more probable p-side. What to do with the remaining money, keep it as an insurance or bet on a less probable side? The first option just diminishes the effective amount of money that works. Moreover, the other side also wins sometimes, so we put $1-f$ on the side with $1-p$ chance. If after N such bets the p-side came n times then your money is multiplied by the factor $(2f)^n[2(1-f)]^{N-n} = 2^{N\Lambda}$, where the rate is

$$\Lambda(f) = 1 + \frac{n}{N}\log_2 f + \left(1 - \frac{n}{N}\right)\log_2(1-f) \ . \tag{94}$$

As $N \to \infty$ we approach the mean geometric rate, which is $\lambda = 1 + p\log f + (1-p)\log(1-f)$. Note the similarity with the Lyapunov exponents from Sections 3.3–3.5 — we consider the logarithm of the exponentially growing factor since we know $\lim_{N\to\infty}(n/N) = p$ (it is called self-averaging quantity because it is again a sum of random numbers). Differentiating $\Lambda(f)$ with respect to f you find that the maximal growth rate corresponds to $f = p$ (proportional gambling) and equals to

$$\lambda(p) = 1 + p\log_2 p + (1-p)\log_2(1-p) = S(u) - S(p)\,, \tag{95}$$

where we denoted the entropy of the uniform distribution $S(u) = 1$ bit. We thus see that the maximal rate of money growth equals to the entropy decrease, that is to the information you have (Kelly 1950). What is beautiful here is that the proof of optimality is constructive and gives us the best betting strategy. An important lesson is that we maximize not the mean return but its mean logarithm, that is a geometric mean. Since it is a self-averaging quantity the probability to grow with this rate approaches unity as $N \to \infty$. Note, however, that the geometric mean is less than the arithmetic mean. Therefore, we may have a situation when the arithmetic growth rate is larger than unity while the geometric mean is smaller than unity. That would be unfortunate, since no matter the strategy, the probability to loose it all will tend to unity as $N \to \infty$, even though the mean returns grows unbounded.

It is straightforward to generalize (96) for gambling on horse races or investing, where many outcomes have different probabilities p_i and payoffs g_i. To maximizing $\sum p_i \log(f_i g_i)$ we look for the maximum of $\sum p_i \log f_i$. Since $\sum f_i = 1$ we can treat it as a distribution. The relative entropy $\sum p_i \log(p_i/f_i)$ is non-negative, so that $\sum p_i \log f_i$ reaches its maximum when all $f_i = p_i$ independent of g_i, that is our distribution coincide with the true distribution, which is proportional gambling. The rate is then

$$\lambda(p,g) = \sum_i p_i \ln(p_i g_i) \ . \tag{96}$$

Here you have a formidable opponent - the track operator, who actually sets the payoffs. Knowing the probabilities, a perfect operator would set the payoffs, $g_i = 1/p_i$, to make the game fair and your rate zero. Nobody's perfect, so it is more likely that the real operator has business sense to make the racecourse profitable by setting the payoffs a bit lower. That will make your λ negative[19]. Your only hope then is that your information is better. Indeed, if the operator assumes that the probabilities are q_i and sets payoffs as $g_i = 1/Zq_i$ with $Z > 1$, then

$$\lambda(p,q) = -\ln Z + \sum_i p_i \ln(p_i/q_i) = -\ln Z + D(p|q) \ . \tag{97}$$

That is if you know the true distribution but the operator uses the approximate one, the relative entropy $D(p|q)$ determines the rate with which your winnings can grow. Since you aren't perfect either, then it is likely that you use the distribution q', which is not the true one. In this case, you still have a chance if your distribution is closer to the true one: $\lambda(p,q,q') = -\ln Z + D(p|q) - D(p|q')$. Remind that the entropy determines the optimal rate of coding. Using incorrect distribution incurs the cost of non-optimal coding. Amazingly, (98) tells that if you can encode the data describing the sequence of track winners shorter than the operator, you get paid in proportion to that shortening.

In reality, people bet according to their whims rather than play a long game, while bookmakers set the rewards according to the statistics of betting rather than horse winnings. Average gambler losses and bookmaker income are independent of the outcome of racing, which is thus a pure sport.

To feel less smug, note that bacteria follow the same strategy without ever taking this or other course on information theory. Like in coin flipping, bacteria often face the choice between growing fast but being vulnerable to antibiotic or grow slow but being resistant. They use proportional gambling to allocate respective fractions of populations to different choices. There could be several lifestyle choices, which is analogous to horse racing (called phenotype switching in this case). The same strategy is used by many plants, where the fraction of the seeds do not germinate in the same year they were dispersed; the fraction increases together with environment variability.

More generally, the environment can be characterized by a set of parameters A, while the internal state of a gambler, plant or bacteria can be characterized by another set of parameters B. In the proportional gambling setting, A is the vector of probabilities $\{p_i\}$ and B is the vector of fractions $\{f_i\}$. In another setting, A could include the concentration of a nutrient and B - the amount of enzyme needed to metabolize the nutrient. The growth rate is then the function of these two parameters $r(A,B)$ and the mean growth rate is as follows:

$$\lambda = \int dA\, dB\, P(A,B) r(A,B) = \int dA\, P(A) \int dB\, P(B|A) r(A,B) \ . \tag{98}$$

[19] European roulette wheel has 18 red and 18 black pockets and a single green, so that even the highest-odds bets, on red or black, have a slightly less than half chance of success.

To maximize the growth, bacteria, plants and gamblers need to coordinate the internal state with that of environment. That coordination is determined by the conditional probability $P(B|A)$, which determines the mutual information between the external world and the internal state:

$$I(A,B) = \int dA\, P(A) \int dBP(B|A) \log_2 \frac{P(B|A)}{P(B)} \ . \tag{99}$$

But acquiring that information has its own cost aI. One then looks for a tradeoff between maximizing growth and minimizing information cost. Therefore, we look for the maximum of the functional $F = \lambda - aI$, which gives similarly to (88,89)

$$P(B|A) = \frac{P(B)}{Z(A,\beta)} e^{\beta r(A,B)} \ , \tag{100}$$

where $\beta = a^{-1} \ln 2$ and the partition function $Z(A,\beta) = \int dBP(B)e^{\beta r(A,B)}$ is the normalization factor. We now recognize the rate distortion theory from the previous subsection; the only difference is that the energy now is minus the growth rate. The choice of β reflects relative costs of the information and the metabolism. If information is hard to get, one chooses small β, which makes $P(B|A)$ weakly dependent of $r(A,B)$ and close to unconditional probability $P(B)$. If information is cheaper, (101) tells us that we need to peak our conditional probability around the maxima of the growth rate. All the possible states in the plane r, I are below some monotonic convex curve, much like in the energy-entropy plane in Section 1.1. One can reach optimal (Gibbs) state on the boundary either by increasing the growth rate at a fixed information of by decreasing the information at a fixed growth rate.

So far we assumed that the probabilities $\{p_i\}$ are known. But more often one needs to play the game to learn the chances. As one plays, incurs some gains and losses and collects some information, one needs to strike the right balance between exploitation of an existing information to maximize the gain and exploration for a new information. For example, there is a broad class of the so-called sequential allocation problems encompassing design of clinical trials, adaptive routing, job-scheduling, and military logistics. Optimal for all them is the remarkable index strategy, which we first illustrate using the simple problem of scheduling jobs: Job i takes time t_i and, on completion, gives reward r_i. It is important that later rewards are γ^t less valuable, where the discount factor $0 < \gamma < 1$. To maximize the total discounted reward, we do i before j, if $r_i\gamma^{t_i} + r_j\gamma^{t_i+t_j} > r_j\gamma^{t_j} + r_i\gamma^{t_i+t_j}$. Taking r_i-terms to the let and r_j-terms to the right, we can present this as inequality for the job indices:

$$\nu_i = \frac{\gamma^{t_i}}{1-\gamma^{t_i}} r_i > \nu_j = \frac{\gamma^{t_j}}{1-\gamma^{t_j}} r_j \ .$$

So we can compute this index ν_i for each job independently, and schedule them in decreasing order of the indices.

Let us now play more complicated multi-armed bandit game, where we can only make one bet at a time, choosing among several options (arms of slot-machines). Imagine that each arm gives the same reward $r = 1$ if you win and 0 if you loose. At the start we do not know the probabilities of winning s_i, so we assume uniform prior: $P(s_i) = 1$, $0 \leq s_i \leq 1$. We play each arm several times and compute the posterior distribution by Bayes' formula. If we encountered w_i wins and l_i losses, then for every value of s_i, the posterior probability is the binomial distribution of w_i, l_i happening:

$$P(s_i) = s_i^{w_i}(1-s_i)^{l_i}\frac{(w_i+l_i+1)!}{w_i!l_i!}\ . \tag{101}$$

Upon further trials with l' losses and w' wins, the distribution is multiplied by $s_i^{w'}(1-s_i)^{l'}$, that is preserves its form, renormalizing parameters. Bayesian update for every arm is equivalent to a random walk in a positive direction on a 2-dimensional lattice (w_i, l_i). Each of these lattice points is a state of a Markov process with one-step vector of transition probabilities $P = \{w_i/(w_i+l_i), l_i/(w_i+l_i)\}$.

We need a strategy which maximizes the sum of the discounted rewards: the expected value of the sum $r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots$. Even though the total number of steps is potentially infinite, the discount factor introduces an effective horizon $\simeq (1-\gamma)^{-1}$. The powerful statement that we give without a proof is that the optimal strategy is to play at each step the arm with the maximal Gittins index ν_i (Gittins, 1979). The index is the ratio of the expected sum of rewards to the discounted time, under the assumption that playing the arm will be terminated in the future after τ steps:

$$\nu(l_i, n_i, t) = \sup_{\tau>0}\frac{\sum_{k=0}^{\tau-1}\gamma^k\langle r_{t+k-1}\rangle}{\sum_{k=0}^{\tau-1}\gamma^k}\ . \tag{102}$$

Here

$$\langle r_{t+k-1}\rangle = \frac{w_i(t_i+k-1)}{w_i(t_i+k-1)+l_i(t_i+k-1)}$$

is the expected reward at the step k, and we sum the future rewards that one would obtain by choosing to play only the i-th arm up to the stopping time $t+\tau$. The brackets denote the averaging over all the lattice paths with expectations based on the distributions (102) at every lattice point $w_i(t_i+k-1), l_i(t_i+k-1)$. We take the maximum over the number of future steps, which is variable, since we admit the possibility of switches to another arm. That supremum can be shown to be achieved, that is the stopping time τ is finite, because the discounted time in the denominator of (103) grows with τ. Denote $L = \nu/(1-\gamma)$, then

$$L = (1-\gamma^\tau)\sum_{k=0}^{\tau-1}\gamma^k\langle r_{t+k-1}\rangle = \sum_{k=0}^{\tau-1}\gamma^k\langle r_{t+k-1}\rangle + \gamma^\tau L\ .$$

That formula means that the lump sum L now or after some optimal number of further rewards are equally good alternatives. One then obtains L (numerically) as a maximal reward which is a fixed point, that is does not change upon one step.

The game thus proceeds as follows: At the beginning, each index is equal to the prior probability of winning, which is an inverse number of arms. We start from an arbitrary arm and play it until the number of losses makes its index less than $1/2$, then we switch to another one, etc. After a while, all arms are played many times with switches occurring when enough losses encountered. In the limit $l_i + w_i \to \infty$, the probability shrinks to $P(s_i) = \delta(s_i - p_i)$, where $p_i = \lim_{l_i+w_i\to\infty} w_i/(w_i + l_i)$, the mean reward is $r_0 = p_i$ and the evident optimal strategy is to choose the arm with the highest p_i, that is $\nu_i = r_0 = p_i$. Generally, the finite-time index is larger than its infinite-time asymptotics, accounting for the possibility that the actual probability is larger than the observed one. As we play an arm, its distribution (102) is getting more and more narrow and the index decreases, which makes it possible to switch to another arm. Switching arms provides a possibility of exploration and obtaining new information.

Financial activity of people is not completely reducible to gambling and its essence understood much less. When you earn enough money, it may be a good time to start thinking about the nature of money itself. Money appeared first as a measure of value, it acquired probabilistic aspect with the development of credit. These days, when most of it is in bits, it is clear that this is not matter (coins, banknotes) but information. Moreover, the total amount of money grows on average, but could experience sudden drops when the crisis arrives. Yet in payments money behaves as energy, satisfying the conservation law. It seems that we need a new concept for describing money, which has properties of both entropy and energy. Free energy combines energy and entropy additively, describing, in particular, how an entropy increase (loss of information) diminished the amount of work one can do. Similarly, free energy can describe a decrease in purchasing power due to information loss. Yet we probably need a more sophisticated notion to describe money as a universal medium of exchange. Since money is essentially a social construct, the degree of universality varies. For example, cold hard cash is guaranteed by governments, but credit card payments are guaranteed by private banks, so these two kinds of money are not identical. Add to this non-bank money like cryptocurrencies, and we see that the value of money depends essentially on how many people agree to use it. It is a challenge to devise a conceptual framework able to handle both material and ephemeral sides of money, but it seems that the information theory is a right place to start.

The preceding sections gave a long list of representations of various efficiency. To finish with it, mention briefly the standard problem of choosing how many fitting parameters to use. While it is intuitively clear that one should not use too many parameters for too few data points, mutual information makes this choice precise. If we work with a given class of functions (say, Fourier harmonics) then it is clear that increasing the number K of functions we can approximate our N points better and better, so that the deviation $D(K)$ is a decreasing function of K. But we know that our data contain noise so it does not make much sense to approximate every fluctuation. In other words, every fitting parameter introduces its own entropy s_i (for a white noise, we have all Fourier harmonics having the same s_i). Technically, we need to minimize the mutual information of the representation, which would consist of two parts: $ND(K) + \sum_{i=1}^{K} s_i$. Here the first term comes from an imperfect data fitting, so it contains the relative entropy $D(K)$ between our hypothetical distribution and the true one, while the second term is the entropy related to our K degrees of freedom. Extremum comes from a competition between the two terms. When we obtain more data and N is getting large, the value of K, which gives a minimum, usually saturates.

5 New Second Law of Thermodynamics

So far, we quantified uncertainty mostly by combinatorics. Classifying and keeping count are among the most difficult mental processes (possibly, because it is not easy to be fair). It is best to hire somebody else to do the job. That tireless somebody, who never stops, is a random walker. In this Chapter, we explore and exploit random walks in different environments. We first use a random walk on a graph to describe the Google's PageRank algorithm designed to quantify not the amount of information but its perceived importance. We then consider a random walk on a lattice biased by an externally imposed time-dependent drift. That will lead us to the modern generalizations of the second law and fluctuation-dissipation relations. It is important both for fundamentals of science and for numerous modern applications related to fluctuations in nano-particles, macromolecules, stock market prices etc.

5.1 Stochastic Web surfing and Google's PageRank

> When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. As was natural, this was

> followed by an excessive depression. The certitude that some precious books were inaccessible seemed almost intolerable.
>
> J L Borges "The Library of Babel"

Let us try to find an objective and quantitative measure not only of the amount of information, but also of its importance. We need to know which are the most precious books in the Library. Indeed, for an efficient information retrieval from the Web Library, webpages need to be ranked by their importance to order search results. By this time, it should come as no surprise for the reader that such ranks can be found by a statistical approach: performing a random walk on the web.

For Internet with n pages, we organize their ranks into a vector $\mathbf{p} = \{p_1, \dots, p_n\}$ which we normalize: $\sum_{i=1}^{n} p_i = 1$. The idea is to *equate the rank p_i with the probability to arrive at this page* randomly clicking on links. A reasonable way to measure the probability to arrive at a page is to count the number of links that refer to it. Not all links are equal though — those from a more probable page must bring more probability. On the other hand, a link from a page with many links must bring to each link less probability. One then comes to the two rules: i) every page relays its rank to the pages it links to, dividing it equally between them, ii) the rank of a page is the sum of all ranks obtained by links. According to the above rules, $p_i = \sum_j p_j/n_j$ where n_j is the number of outgoing links on page j, which links to the page i. In other words, we are looking for the eigenvector of the hyperlink matrix, $\mathbf{p}\hat{A} = \mathbf{p}$, where the matrix elements $a_{ij} = 1/n_j$ if j links to i and $a_{ij} = 0$ otherwise. Does a unique eigenvector with all non-negative entries and a unit eigenvalue always exist? If yes, how to find it?

The iterative algorithm to find the rank eigenvector p_i is called PageRank[20] (Brin and Page 1998). It starts from ascribing equal probability to all pages, $p_i(0) = 1/n$, and generates the new probability distribution by applying the above rules of the rank relay:

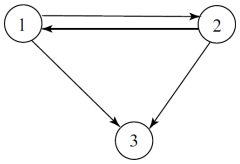
$$\mathbf{p}(t+1) = \mathbf{p}(t)\hat{A}\,. \tag{103}$$

We recognize that this stochastic process is a Markov chain, mentioned in Sections 3.3 and 4.7, which means that the future is determined by the present state, but not by the past. We thus interpret $\hat{A}$ as the matrix of transition probabilities between pages for our random surfer. In later modifications, one fills the elements of $\hat{A}$ not uniformly as $1/n_j$ but use information about actual frequencies of linking that can be obtained from access logs. Could our self-referential rules lead to a vicious circle or the iterations converge at $t \to \infty$? It better be convergent fast to be of any use for the instant-gratification generation. It is clear that if the largest eigenvalue λ_1 of $\hat{A}$ was larger than unity, than the iterations would diverge; if

[20] "Page" relates both to webpage and to Larry Page, who with Sergei Brin invented the algorithm and created Google.

$\lambda_1 < 1$, then the iterations would converge to zero. Both contradict normalization $\sum p_i = 1$. We need the largest eigenvalue to be unity and correspond to a single eigenvector, so that the iterations converge. How fast it converges then will be determined by the second largest eigenvalue λ_2 (which must be less than unity).

Moment reflection is enough to identify the problem: some pages do not link to any other page, which corresponds to rows of zeroes in $\hat{A}$. Such pages accumulate the score without sharing it. Another problem is caused by loops. The figure presents a simple example illustrating both problems:



If all transition probabilities are nonzero, the probability vector with time tends to $(0, 0, 1)$, that is the surfer is stuck at the page 3. When the probabilities a_{13}, a_{23} are very small, the surfer tend to be caught for long times in the loop $1 \longleftrightarrow 2$.

To release our random surfer from being stuck at a sink or caught in a loop, the original PageRank algorithm allowed it to jump randomly to any other page with equal probability. To be fair with pages that are not sinks, these random teleportations are added to all nodes in the Web: surfer either clicks on a link on the current page with probability d or opens up a random page with probability $1 - d$. To quote the original: "We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the damping factor is the probability at each page the "random surfer" will get bored and request another random page." This is equivalent to replacing $\hat{A}$ by $\hat{G} = d\hat{A} + (1 - d)\hat{E}$. Here the teleportation matrix $\hat{E}$ has all entries $1/n$, that is $\hat{E} = \mathbf{e}\mathbf{e}^T/n$, where $\mathbf{e}$ is the column vector with $e_i = 1$ for $i = 1, \ldots, n$. After that, all matrix entries g_{ij} are strictly positive and the graph is fully connected.

It is important that now our matrix has positive elements in every column whose sum is unity. Such matrices are called stochastic, since every column can be thought of as a probability distribution. Every stochastic matrix has unity as the largest eigenvalue. Indeed, since $\sum_j g_{ij} = 1$, then $\mathbf{e}$ is an eigenvector of the transposed matrix: $\hat{G}^T\mathbf{e} = \mathbf{e}$. Therefore, 1 is an eigenvalue for $\hat{G}^T$, and also for $\hat{G}$, which has the same eigenvalues. We can now use convexity to prove that this is the largest eigenvalue. For any vector $\mathbf{p}$, every element of $\mathbf{p}\hat{G}$ is a convex combination of the elements, $\sum_j p_j g_{ij}$, which cannot exceed the largest element of $\mathbf{p}$ since $\sum_j g_{ij} = 1$. For an eigenvector with an eigenvalue exceeding unity, at least one element of $\mathbf{p}\hat{G}$ must exceed the largest element of $\mathbf{p}$, therefore such eigenvector cannot exist. This is a particular case of the theorem: The eigenvalue with the largest absolute value of a positive square matrix is positive, and belongs to a positive eigenvector, where all of the vector's elements are positive. All other eigenvectors are smaller in absolute value (Markov 1906, Perron 1907).

The great achievement of PageRank algorithm is the replacement of the iterative process (104) by

$$\mathbf{p}(t+1) = \hat{G}\mathbf{p}(t)\,. \tag{104}$$

That process cannot be caught into a loop and converges, which follows from the fact that $G_{ii} \neq 0$ for all i; that is there is always a probability to stay on the page breaking any loop. The eigenvalues of $\hat{G}$ are $1, d\lambda_2 \dots d\lambda_n$, where $\{\lambda_i\}$ are eigenvalues of $\hat{A}$ (prove it), so the choice of d affects convergence, the smaller the faster. On the other hand, it is somewhat artificial to use teleportation to an arbitrary page, so larger values of d give more weight to the true link structure of the Web. As in other optimization problems we encountered in this course, one needs a workable compromise. The standard Google choice $d = 0.85$ comes from estimating how often an average surfer uses bookmarks. As a result, the process usually converges after about 50 iterations.

One can design a personalized ranking by replacing the teleportation matrix by $\hat{E} = \mathbf{e}\mathbf{v}^T$, where the probability vector $\mathbf{v}$ has all nonzero entries and allows for personalization, that is can be chosen according to the individual user's history of searches and visits. That means that it is possible in principle to have our personal rankings of the webpages and make searches custom-made.

As mentioned, the sequence of the probability vectors defined by the relations of the type (104,105) is a Markov chain. In particular, the three random quantities $X \to Y \to Z$ is a Markov triplet if Y is completely determined by X, Z, while X, Z are independent conditional on Y, that is $I(X, Z|Y) = 0$. Such chains have an extremely wide domain of applications.

5.2 Random walk and diffusion

Let us now consider a particular yet fundamental stochastic process of a random walk on lattice, where the transition probability is nonzero only for neighboring cites. Our walker can hop randomly to any of $2d$ neighboring cites in the d-dimensional cubic lattice, starting from the origin at $t = 0$. We denote a the lattice spacing, τ the time between hops and $\mathbf{e}_i$ the orthogonal lattice vectors that satisfy $\mathbf{e}_i \cdot \mathbf{e}_j = a^2\delta_{ij}$. The probability to be in a given cite $\mathbf{x}$ evolves according to the equation

$$P(\mathbf{x}, t+\tau) = \frac{1}{2d}\sum_{i=1}^{d}\left[P(\mathbf{x}+\mathbf{e}_i, t) + P(\mathbf{x}-\mathbf{e}_i, t)\right] \ . \tag{105}$$

That can be rewritten in the form convenient for taking the continuous limit:

$$\frac{P(\mathbf{x},t+\tau)-P(\mathbf{x},t)}{\tau}=\frac{a^2}{2d\tau}\sum_{i=1}^{d}\frac{P(\mathbf{x}+\mathbf{e}_i,t)+P(\mathbf{x}-\mathbf{e}_i,t)-2P(\mathbf{x},t)}{a^2}\ . \tag{106}$$

This is a finite difference approximation to the diffusion equation, which appears when we take the continuous limit $a\to 0,\tau\to 0$ keeping finite the ratio $\kappa=a^2/2d\tau$: $(\partial_t-\kappa\Delta)P(\mathbf{x},t)=0$. The space density $\rho(x,t)=P(\mathbf{x},t)a^{-d}$ satisfies the same equation:

$$\partial_t\rho=\kappa\Delta\rho\ . \tag{107}$$

The solution with the initial condition $\rho(\mathbf{x},0)=\delta(\mathbf{x})$ is the Gaussian distribution:

$$\rho(\mathbf{x},t)=(4\pi\kappa t)^{-d/2}\mathrm{exp}\left(-\frac{x^2}{4\kappa t}\right)\ . \tag{108}$$

As well as (106,107), the diffusion equation conserves the total probability, $\int\rho(\mathbf{x},t)\,d\mathbf{x}$, because it has the form of a continuity equation, $\partial_t\rho(\mathbf{x},t)=-\mathrm{div}\,\mathbf{j}$ with the current $\mathbf{j}=-\kappa\nabla\rho$. Note that (108,109) are isotropic and translation invariant while the discrete version respected only cubic symmetries.

Another way to describe a random walk is to treat $\mathbf{e}_i$ as a random variable with $\langle\mathbf{e}_i\rangle=0$ and $\langle\mathbf{e}_i\mathbf{e}_j\rangle=a^2\delta_{ij}$, so that $\mathbf{x}=\sum_{i=1}^{t/\tau}\mathbf{e}_i$. The probability of the sum is (109), that is the product of Gaussian distributions of the components, with the variance growing linearly with t.

A path of a random walker behaves rather like a surface than a line. Two-dimensionality of the random walk is a reflection of the square-root diffusion law: $\langle x\rangle\propto\sqrt{t}$. Indeed, we defined in Section 2.3 the box-counting dimension (37) looking how the number of boxes $N(a)$ needed to cover the a geometric object grow as the box size a decreases. For a line, $N\propto 1/a$, generally $N\propto a^{-d}$. As we discussed above, diffusion requires the time step to shrink with the lattice spacing according to $\tau\propto a^2$. The number of elements is the number of steps and grows for a given t as $N(a)=t/\tau\propto a^{-2}$, so that $d=2$. One can also obtain dimension of a set as the relation between its size x and the number $N\propto x^d$ of standard-size elements needed to cover it. For a random walk, the number of elements is of order of the number of steps, $N\propto t\propto x^2$. Surfaces generally intersect along curves in 3d, they meet at isolated points in 4d and do not meet at $d>4$. That is reflected in special properties of critical phenomena in 2d (where random walker fills the surface) and 4d (where random walkers do not meet and hence do not interact). Is the mean time spent on a given site, $\sum_{t=0}^{\infty}P(\mathbf{x},t)$, finite or infinite? It follows from (109) that the answer depends on the space dimensionality: $\int\rho(\mathbf{x},t)\,dt\propto\int^{\infty}t^{-d/2}dt$ diverges for $d\leq 2$. In other words, the walker in 1d and 2d returns to any point infinite number of times.

One can see an example of how the properties of a random walk and diffusion appear in a physical system in Appendix 8.8, which describes the Brownian motion of a small particle in a fluid.

5.3 Fluctuation relation and the new second law

A significant generalization of equilibrium statistical physics can be achieved for systems with one or few degrees of freedom deviated arbitrary far from equilibrium. That can be done under the assumption that the rest of the degrees of freedom is in equilibrium and can be represented by a thermostat generating thermal noise. This new approach also allows one to treat non-thermodynamic fluctuations, such as a negative entropy change.

We illustrate these developments using the simplest example of a walker on a line x. Apart from a random walk with the diffusivity $\kappa = T = 1/\beta$, there is a drift with the velocity $v(x) = -\partial V(x)/\partial x$, that is down the gradient of the potential $V(x)$. One physical example is an the over-damped Brownian particle described in Appendix 8.8. According to the continuity equation, the drift adds ρv to the current j and $-\partial_x(\rho v)$ to $\partial\rho(x,t)/\partial t$. Combining this with (108), which describes diffusion from the random walk, we obtain the so-called Fokker-Planck equation for the probability $\rho(x,t)$:

$$\partial_t\rho = T\partial_x^2\rho + \partial_x(\rho\partial_x V) = -\hat{H}_{FP}\rho \ . \tag{109}$$

We have introduced the Fokker-Planck operator,

$$H_{FP} = -\frac{\partial}{\partial x}\left(\frac{\partial V}{\partial x} + T\frac{\partial}{\partial x}\right) \ ,$$

which allows one to exploit an analogy between quantum mechanics and statistical physics. We may say that the probability density is the ψ-function is the x-representation, $\rho(x,t) = \langle x|\psi(t)\rangle$. We then rewrite (110) as $d|\psi\rangle/dt = -\hat{H}_{FP}|\psi\rangle$, which has a formal solution $|\psi(t)\rangle = \exp(-tH_{FP})|\psi(0)\rangle$. The only difference with quantum mechanics is that their time is imaginary (of course, they think that our time is imaginary). In other terms, Schrodinger equation, $(\imath\hbar\Delta - 2mV)|\psi\rangle = 0$ corresponds to imaginary diffusivity. The transition (conditional) probability to come from x at t to x' at t' is given by the matrix element:

$$\rho(x',t';x,t) = \langle x'|\exp[(t-t')H_{FP})|x\rangle \ . \tag{110}$$

Without the coordinate-dependent field $V(x)$, the transition probability is symmetric, $\rho(x',t;x,0) = \rho(x,t;x',0)$, which is formally manifested by the fact that the respective Fokker-Planck operator ∂_x^2 is Hermitian. This property is called the detailed balance.

How the detailed balance is modified in an external field? If the potential V is time independent, then the stationary solution of (110) is the zero-current Gibbs state $\rho(x) = Z_0^{-1}\exp[-\beta V(x)]$, where $Z_0 = \int\exp[-\beta V(x,0)]\,dx$. That state satisfies a modified detailed balance: the probability current is the (Gibbs) probability density at the starting point times the transition probability; forward and backward currents must be equal in equilibrium:

$$\rho(x',t;x,0)e^{-V(x)/T} = \rho(x,t;x',0)e^{-V(x')/T} \ . \tag{111}$$
$$\langle x'|e^{-tH_{FP}}e^{-V/T}|x\rangle = \langle x|e^{-tH_{FP}}e^{-V/T}|x'\rangle = \langle x'|e^{-V/T}e^{-tH_{FP}^\dagger}|x\rangle \ .$$

Since this must be true for any x, x' then $e^{-tH_{FP}^\dagger} = e^{V/T}e^{-tH_{FP}}e^{-V/T}$ and

$$H_{FP}^\dagger \equiv \left(\frac{\partial V}{\partial x} - T\frac{\partial}{\partial x}\right)\frac{\partial}{\partial x} = e^{V/T}H_{FP}e^{-V/T} \ , \tag{112}$$

i.e. $e^{V/2T}H_{FP}e^{-V/2T}$ is hermitian, which can be checked directly. The quantum-mechanical notations thus allowed us to translate the detailed balance from the property of transition probabilities to that of the evolution operator (more on the analogy between thermal and quantum fluctuations can be found in Appendix 8.10).

If we now allow the potential to change in time, then the system goes away from equilibrium. Consider an ensemble of trajectories starting from the initial positions taken with the equilibrium Gibbs distribution corresponding to the initial potential: $\rho(x,0) = Z_0^{-1}\exp[-\beta V(x,0)]$. As time proceeds and the potential continuously changes, the system is never in equilibrium, so that $\rho(x,t)$ does not generally have a Gibbs form. Indeed, even though one can define a time-dependent Gibbs state $Z_t^{-1}\exp[-\beta V(x,t)]$ with $Z_t = \int\exp[-\beta V(x,t)]dx$, one can directly check that it is not any longer a solution of the Fokker-Planck equation (110) because of the extra term: $\partial_t\rho = -\beta\rho\partial_t V$. The distribution needs some time to adjust to the potential changes and is generally dependent on the history of these. For example, if we suddenly broaden the potential well, it will take diffusion (with diffusivity T) to broaden the distribution. Can we find some quantity which accounts for this history and lets us to generalize the detailed balance relation (112) we had in equilibrium? Such relation was found surprisingly recently despite its generality and relative technical simplicity of derivation.

To find the quantity that has a Gibbs form (i.e. have its probability determined by the instantaneous partition function Z_t), we need to find an equation which generalizes (110) by having an extra term that will cancel the time derivative of the potential. It is achieved by considering, apart from a position x, another random quantity defined as the potential energy change (or the external work done) along the particle trajectory during the time t:

$$W_t = \int_0^t dt' \frac{\partial V(x(t'), t')}{\partial t'} \ . \tag{113}$$

The time derivative is partial i.e. taken only with respect to the second argument, so that the integral is not equal by the difference between the start and the finish, but is determined by the whole history. The work is a fluctuating and even sign-changing quantity depending on the trajectory $x(t')$, which itself depends on the initial point and random walk realization.

Let us now run our random walker many times choosing different starting points $x(0)$ according to the Gibbs probability $\rho(x) = Z_0^{-1} \exp[-\beta V(x, 0)]$. It will give us many trajectories having different endpoints $x(t)$ and different energy changes W accumulated along the way. Now consider the joint probability $\rho(x, W, t)$ to come to x acquiring energy change W. This two-dimensional probability distribution satisfies the generalized Fokker-Planck equation, which can be derived as follows: Similar to the argument preceding (110), we note that the flow along W in $x - W$ space proceeds with the velocity $dW/dt = \partial_t V$ so that the respective component of the current is $\rho \partial_t V$ and the equation takes the form

$$\partial_t \rho = \beta^{-1} \partial_x^2 \rho + \partial_x (\rho \partial_x V) - \partial_{_W} \rho \partial_t V \ , \tag{114}$$

Since $W_0 = 0$ then the initial condition for (115) is

$$\rho(x, W, 0) = Z_0^{-1} \exp[-\beta V(x, 0)] \delta(W) \ . \tag{115}$$

While we cannot find $\rho(x, W, t)$ for arbitrary $V(t)$ we can multiply (115) by $\exp(-\beta W)$ and integrate over dW. Since $V(x, t)$ does not depend on W, we get the closed equation for $f(x, t) = \int dW \rho(x, W, t) \exp(-\beta W)$:

$$\partial_t f = \beta^{-1} \partial_x^2 f + \partial_x (f \partial_x V) - \beta f \partial_t V \ , \tag{116}$$

Now, *this* equation does have an exact time-dependent solution

$$f(x, t) = Z_0^{-1} \exp[-\beta V(x, t)] \,,$$

where the factor Z_0^{-1} is chosen to satisfy the initial condition (116). Note that $f(x, t)$ is instantaneously defined by $V(x, t)$ without any history dependence, in distinction from $\rho(x, t)$. In other words, the distribution weighted by $\exp(-\beta W_t)$ looks like Gibbs state, adjusted to the time-dependent potential at every moment of time. Remark that the phase volume defines probability only in equilibrium, yet the work divided by temperature is an analog of the entropy change (production), and the exponent of it is an analog of the phase volume change. Let us stress that $f(x,t)$ is not a probability distribution. In particular, its integral over x is not unity but the mean phase volume change, which remarkably is expressed via equilibrium partition functions at the ends (Jarzynski 1997):

$$\int f(x,t)dx = \int \rho(x,W,t)e^{-\beta W}dxdW = \left\langle e^{-\beta W}\right\rangle = \frac{Z_t}{Z_0} = \frac{\int e^{-\beta V(x,t)}dx}{\int e^{-\beta V(x,0)}dx} \,. \quad (117)$$

Here the bracket means double averaging: over the initial distribution $\rho(x,0)$ and over the different random walks during the time interval $(0,t)$. We can also obtain all weighted moments of x like $\langle x^n \exp(-\beta W_t)\rangle$ [21]. One can introduce the free energy $F_t = -T \ln Z_t$, so that $Z_t/Z_0 = \exp[\beta(F_0 - F_t)]$.

Let us reflect on where we have arrived following our random walker. We started from a Gibbs distribution but considered *arbitrary* temporal evolution of the potential. Therefore, our distribution was arbitrarily far from equilibrium during the evolution. And yet, we expressed the mean exponent of the work done via the partition functions of the equilibrium Gibbs distributions corresponding to the potential at the beginning and at the end. Even though the system is not in equilibrium at the end, the use of the respective Gibbs distribution is not that surprising, because the further relaxation to the equilibrium at the end value of the potential is not accompanied by doing any work W. What is surprising is that there is no dependence on the intermediate times. One can also look at it from the opposite perspective: no less remarkable is that one can determine a truly equilibrium property, the free energy difference, from non-equilibrium measurements (which could be arbitrary fast rather than adiabatically slow as we used to do in traditional thermodynamics).

The total heat release is the work minus the free energy change: $Q = W - F_t + F_0$. Divided by the temperature, it is minus the entropy change during the evolution. That allows to rewrite (118) as the following identity:

$$\langle e^{-\beta Q}\rangle = \langle e^{-\Delta S}\rangle = 1 \,, \quad (118)$$

which is a generalization of the second law of thermodynamics. Note that the entropy change ΔS is treated here as a fluctuating quantity, which could have either sign. Using the Jensen inequality $\langle e^A\rangle \geq e^{\langle A\rangle}$, one can obtain the usual second law of thermodynamics for the positivity of the mean entropy change:

$$\langle \beta W_d\rangle = \langle \Delta S\rangle \geq 0 \ .$$

When information processing is involved, it must be treated on equal footing, which allows one to decrease the work and the dissipation below the free energy difference:

$$\langle e^{-\beta Q - I}\rangle = \langle e^{-\Delta S}\rangle = 1 \ . \quad (119)$$

[21] I thank R. Chetrite for this derivation.

(Sagawa and Uedo, 2012; Sagawa 2012). We have considered such a case in Section 4.2, where we used $\langle Q\rangle \geq -IT = -T\Delta S$. The exponential equality (120) is a generalization of this inequality and (77).

So the modern form of the second law of thermodynamics is an equality rather than an inequality. The latter is just a partial consequence of the former. Compare it with the re-formulation of the second law in Section 3.3 as a conservation law rather than a law of increase. And yet (120) is not the most general form. The further generalization is achieved by relating the entropy production to irreversibility, stating that the probability to have a change $-\Delta S$ in a time-reversed process (marked by dagger) is as follows (Crooks 1999):

$$P^{\dagger}(-\Delta S) = P(\Delta S)e^{-\Delta S} \ . \tag{120}$$

Integrating (121) one obtains (120). That remarkable relation allows also one to express the mean entropy production via the relative entropy (66) between probabilities of the forward and backward evolution:

$$\langle \Delta S\rangle = \Big\langle \ln[P(\Delta S)/P^{\dagger}(-\Delta S)]\Big\rangle \ . \tag{121}$$

One can find derivation of the relation (121) for our toy model of the generalized baker map in the Appendix 8.3, and multi-dimensional versions in the Appendix 8.9.

6 Quantum information

Our world is fundamentally quantum mechanical. That adds some unavoidable uncertainty, which cannot be diminished by improving measurement precision or gathering more information. The unique source of the quantum uncertainty is superposition: a quantum system can be in many different states at the same time. The results of quantum mechanical measurement are then truly random (not because we did not bother to learn more on the system). Moreover, measurement dramatically differs in a quantum world since it irreversibly changes the system and this change cannot be made arbitrarily small. Account of quantum-mechanical uncertainty is thus of fundamental value for information theory.

Interest in quantum information is also pragmatic. Quantum superposition means that evolution proceeds in the space of factorially more dimensions than the respective classical system. This is a source of the parallelism of quantum computations. Moreover, classical systems, including computers, are limited by locality, that is operations have only local effects. Spatially separated quantum systems may be entangled with each other and operations may have non-local effects because of this. Those two basic facts motivate an interest in quantum computations.

Non-surprisingly, the quantum information theory is also based on the notion of entropy, which is similar to classical entropy yet differs in some important ways. Uncertainty and probability exist already in quantum mechanics where we consider an isolated system. On top of that we shall consider quantum statistics due to incomplete knowledge, which is caused by considering subsystems. This Chapter gives a very brief introduction to the subject, focusing on information and entropy and their most dramatic differences from the classical world. Recall that the entropy consideration by Planck is what started quantum physics in the first place. Looking at two asymptotics of a spectral curve, he decided to search for an analytic formula matching their *entropies*, simply adding them. The resulting formula is the logarithm of the number of ways to distribute a given energy in equal discrete portions — quantization was born.

6.1 Quantum mechanics and entropy

Quantum mechanics mathematically is quite elementary, since it is based on linear algebra, that is the study of vectors and linear operations on them. A quantum state of a physical system is a vector. We shall denote such (column) vectors either by ψ_i or use Dirac notation $|i\rangle$. The dual (row) vector then is denoted $\langle i|$ and the inner (scalar) product by $\langle i|j\rangle$. If in some orthonormal basis $\{|i\rangle\}$, two vectors are presented as $|v\rangle = \sum_i v_i\,|i\rangle$ and $|w\rangle = \sum_i w_i\,|i\rangle$, then $\langle v|w\rangle = \sum_i v_i^* w_i$. A property that can be measured is called an observable and is described a self-adjoint operator (matrix), say $\hat{O}$. The expectation value of an observable in a state ψ is an inner product $\langle\psi|\,\hat{O}\,|\psi\rangle$.

The fundamental statement is that any system can be in a single state ψ_i or in a superposition of states, $\psi = \sum_i a_i\psi_i$, where a_i are generally complex numbers. An example of a single state is a fixed-energy eigenstate of a Hamiltonian (which is an operator that is a matrix). The possibility of a superposition is the total breakdown from classical physics, where those states (say, with different energies) are mutually exclusive.

There are two things we can do with a quantum state: let it evolve without touching or measure it. Measurement is classical, it produces one and only state from the initial superposition; immediately repeated measurements will produce the same outcome. However repeated measurement of the identically prepared initial superposition, $\psi = \sum_i a_i\psi_i$, find different states: the state i appears with probability $p_i = |a_i|^2$.

There is an uncertainty already in a state of an isolated quantum system. If the operators of two observables are non-commuting, $[\hat{A},\hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A} \neq 0$, then the product of their variances is restricted from below:

$$|\langle\psi|[\hat{A},\hat{B}]|\psi\rangle|^2 = 4|\langle\psi|\hat{A}\hat{B}|\psi\rangle|^2 - |\langle\psi|\hat{A}\hat{B}+\hat{B}\hat{A}|\psi\rangle|^2$$
$$\leq 4|\langle\psi|\hat{A}\hat{B}|\psi\rangle|^2 \leq 4\langle\psi|\hat{A}^2|\psi\rangle\langle\psi|\hat{B}^2|\psi\rangle \ . \qquad (122)$$

Here the second step is the Cauchy-Schwarz inequality. In particular, momentum and coordinate are such pair, $\hat{A} = \hat{\mathbf{p}} - \langle\mathbf{p}\rangle$, $\hat{B} = \hat{\mathbf{q}} - \langle\mathbf{q}\rangle$. Since the momentum operator in the coordinate representation is $\hat{p}_x = \imath\hbar\partial_x$, then $[\hat{p}_x, x] = -\imath\hbar$, which gives the Heisenberg uncertainty principle: the variances of the coordinate and the momentum along the same direction, $\sigma_p = \langle p^2\rangle - \langle p\rangle^2$, $\sigma_q = \langle q^2\rangle - \langle q\rangle^2$, satisfy the inequality

$$\sqrt{\sigma_p\sigma_q} \geq \hbar/2\,. \qquad (123)$$

That means that we cannot describe quantum states as points in the phase space (p,q). Indeed, what we shall call below quantum entanglement is ultimately related to the fact that one cannot localize quantum states in a finite region — if coordinates are fixed somewhere, then the momenta are not.

Uncertainty relation (124) is an undergraduate version, let us describe now the graduate version, like we replaced the undergraduate version of the second law, $\langle\Delta S\rangle \geq 0$, by its graduate version $\langle e^{-\Delta S}\rangle = 1$. Indeed, variances are sufficient characteristics of uncertainty only for Gaussian distributions; the relation (124) was obtained for a free particle whose probability distributions in coordinate and momentum spaces are Gaussian. Generally, quantum probability distributions are non-Gaussian, so that their uncertainty must be quantified by an entropy. Taking log of the Heisenberg relation, we obtain $\log(2\sigma_p/\hbar)+\log(2\sigma_q/\hbar) = S(p)+S(q) \geq 0$, recasting it as requirements on the entropies of the Gaussian probability distributions of the momentum and the coordinate of a free particle. In d dimensions, different components commute, so that $\sqrt{\sigma_{\mathbf{p}}\sigma_{\mathbf{q}}} \geq d\hbar/2$ and

$$S(\mathbf{p}) + S(\mathbf{q}) \geq \log d\,.$$

When the probability distributions of non-commuting variables are not Gaussian, formulation in terms of variances does not make sense; yet the entropic uncertainty relation remains universally valid and is thus fundamental (Deutsch 1982).

More formally, if two operators do not commute, they cannot be diagonalized by a single orthonormal set. Assume that two non-commuting operators can be diagonalized by (project into) two different orthonormal bases $\{|x\rangle\}, \{|z\rangle\}$. If we measure a quantum state ψ by projecting onto the x-basis, the outcomes define a classical probability distribution $p(x)$. The Shannon entropy $S(X)$ quantifies how uncertain we are about the outcome before we perform the measurement. There is also a corresponding classical probability distribution of outcomes when we measure the same state ψ in the z-basis. The two bases are incompatible, so that there is a tradeoff between our uncertainty about X and about Z, captured by the inequality

$$S(X) + S(Z) \geq log(1/c)\,, \quad c = \max_{x,z} |\langle x|z\rangle|^2\,. \tag{124}$$

We see that the lower bound on the total uncertainty is given by the mean degree of mutual non-orthogonality of the two bases. We shall prove a more general form of this relation in the next subsection.

Two different bases $\{|x\rangle\}$, $\{|z\rangle\}$ for a d-dimensional space are called mutually unbiased if $|\langle x_i|z_k\rangle|^2 = 1/d$ for all i, k. That means that if we measure any x-basis state in the z-basis, all d outcomes are equally probable and give the same contribution into the total probability: $\sum_k |\langle x_i|z_k\rangle|^2 = \sum_i |\langle x_i|z_k\rangle|^2 = 1$. For measurements in two mutually unbiased bases performed on a pure state, the entropic uncertainty relation becomes

$$S(X) + S(Z) \geq \log d\,. \tag{125}$$

This inequality is saturated by x-basis states, for which $S(X) = 0$ and $S(Z) = \log d$. In one dimension $\log d = 0$.

Qubit. So far we dealt with the statistics of the measurement outcomes, and the entropy was the familiar classical Gibbs-Shannon entropy. Let us now deal with the *states* of quantum systems, rather than with the measurements. We have defined classical "bit" as a unit of information choosing between two states, so we can also call a bit a physical system, where we distinguish two states only. That could be a coin, a magnetic moment looking along or against an applied field, a photon with two polarizations, etc. Similarly, we define *qubit* — a quantum system having only two orthogonal states: $|0\rangle$ and $|1\rangle$. The most general state of a qubit A is a superposition of two states, $\psi_A = a\,|0\rangle + b\,|1\rangle$, where any observable is as follows:

$$\langle\psi_A|\hat{O}_A|\psi_A\rangle = |a|^2\langle 0|\hat{O}_A|0\rangle + |b|^2\langle 1|\hat{O}_A|1\rangle + (a^*b + ab^*)\langle 0|\hat{O}_A|1\rangle\ . \tag{126}$$

Normalization requires $|a|^2 + |b|^2 = 1$ and if the overall phase does not matter, then a qubit is characterized by two real numbers, say the amplitude $|a|$ and the relative phase between a and b. Alternatively, we may characterize it by a complex number. The qubit represents the unit of quantum information the same way the bit represents the unit of classical information. Here we see that indeed quantum systems operate with much more information - one needs many bits to record a complex number with a reasonable precision, and the difference grows exponentially when we compare the states of N classical bits with the possible states of N qubits. Moreover, qubit is not a classical bit because it can be in a superposition, nor can it be considered a random ensemble of classical bits with the probability $|a|^2$ in the state $|0\rangle$, because the phase difference of the complex numbers a, b matter, as seen from (127).

And yet quantum mechanics tells us that we cannot measure the complex numbers a, b, that is we cannot determine the quantum state of the qubit. This is in sharp contract with our ability to determine the state of a bit (say, when classical computer retrieves a memory). Measurements of a qubit bring either the result $|0\rangle$ with the probability $|a|^2$ or the result $|1\rangle$ with the probability $|b|^2 = 1 - |a|^2$. In other words, a quantum coin can defy gravity and stand on its edge at an arbitrary angle, but any measurements collapses it on one side, either heads or tails up. What use then in quantifying the quantum information if we cannot measure it? One should not despair though. While we cannot measure it directly, we can communicate it. Moreover, we shall describe below indirect ways to manipulate a quantum system so that a measurement gives a result, which depends distinctly on the state of the system. These ways involve entanglement between different subsystems.

6.2 Quantum statistics and density matrix

To consider subsystems, we need to pass from quantum mechanics to quantum statistics and introduce the fundamental notion of the **density matrix**. Consider a composite system AB, which is in a state ψ_{AB}. Denote by x the coordinates on A and by y on B. The expectation value of any $O(x)$ can be written as $\bar{O} = \sum_{x,y} \psi^*_{AB}(x,y)\hat{O}(x)\psi_{AB}(x,y)$. For independent sub-systems, one has $\psi_{AB}(x,y) = \psi_A(x)\psi_B(y)$ and $\bar{O} = \sum_x \psi^*_A(x)\hat{O}(x)\psi_A(x)$, so that one can forget about B and characterize A by the vector ψ_A. But generally, dependencies on x and y are not factorized, and the action of $\hat{O}(x)$ changes both x and y. We then ought to characterize A by the so-called density matrix

$$\rho(x,x') = \sum\nolimits_y \psi^*_{AB}(x',y)\psi_{AB}(x,y),$$

so that $\bar{O} = \sum_x [\hat{O}(x)\rho(x,x')]_{x'=x}$, where $\hat{O}(x)$ acts only on x and then we put $x' = x$.

More formally, if the state ψ_{AB} is a (tensor) product[22] of two states of A and B, $\psi_{AB} = \psi_A \otimes \psi_B$, then any operator $\hat{O}_A$ acting only in A has the expectation

[22] Multiplying every component of one N-vector by every component of another M-vector gives a MN-vector called tensor product. For example, $(a,b) \otimes (c,d,e) = (ac, ad, ae, bc, bd, be)$.

value

$$\langle\psi_{AB}|\hat{O}_A\otimes\hat{I}_B|\psi_{AB}\rangle = \langle\psi_A|\hat{O}_A|\psi_A\rangle\langle\psi_B|\hat{I}_B|\psi_B\rangle = \langle\psi_A|\hat{O}_A|\psi_A\rangle \ .$$

Here $\hat{I}$ is the identity operator. However, a general state ψ_{AB} could be not a single (tensor) product of states but a sum of tensor products:

$$\psi_{AB} = \sum_k \sqrt{p_k}\psi_A^k \otimes \psi_B^k \,, \tag{127}$$

Even more generally, for arbitrary orthonormal bases ϕ_A^i and ϕ_B^j, one generally has $\psi_{AB} = \sum_{ij}\alpha_{ij}\phi_A^i \otimes \phi_B^j$. A singular value decomposition allows one to represent the matrix of the coefficients as $\alpha_{ij} = u_{ik}d_{kk}v_{kj}$ where, $\hat{u}, \hat{v}$ are unitary and $\hat{d}$ is diagonal. We then define $\psi_A^k = u_{ki}\phi_A^i$ and $\psi_B^k = v_{jk}\phi_B^j$. That is called Schmidt decomposition by orthonormal vectors ψ_A^k, ψ_B^k, which allows us to present any state of AB as a sum of the products (128). The beauty of (128) is that there is only one sum: for each vector in A there is just one vector in B.

If there is more than one term in this sum, we call subsystems A and B *entangled*. There is no factorization of the dependencies in such a state. We can always make ψ_{AB} a unit vector, so that $\sum_i p_i = 1$ and these numbers can be treated as probabilities (to be in the state i). Now the operator acting only on A has the following expectation value

$$\langle\psi_{AB}|\hat{O}_A\otimes\hat{I}_B|\psi_{AB}\rangle = \sum_{i,j}\sqrt{p_ip_j}\langle\psi_A^i|\hat{O}_A|\psi_A^j\rangle\langle\psi_B^i|\hat{I}_B|\psi_B^j\rangle$$
$$= \sum_{i,j}\sqrt{p_ip_j}\langle\psi_A^i|\hat{O}_A|\psi_A^j\rangle\delta_{ij} = \sum_i p_i\langle\psi_A^i|\hat{O}_A|\psi_A^i\rangle = \mathrm{Tr}_A\rho_A\hat{O}_A \,,$$

where the density matrix in such notations is written as follows:

$$\rho_A = \sum_i p_i|\psi_A^i\rangle\langle\psi_A^i| \ . \tag{128}$$

It is all we need to describe A. From now on we shall distinguish pure states described by a vector and mixed states described by a density matrix. The matrix is hermitian, it has all non-negative eigenvalues and a unit trace. Every matrix with those properties can be "purified" that is presented (non-uniquely) as a density matrix of the subsystem A in the extended system AB, which as a whole is in a pure state ψ_{AB}. Possibility of purifications is quantum mechanical with no classical analog: the classical analog of a density matrix is a probability distribution which cannot be purified.

Statistical density matrix describes a mixed state or, in other words, an ensemble of states. Different ensembles can give the same density matrix, see home exercise. Mixed state described by a matrix must be distinguished from quantum-mechanical superposition described by a vector. The superposition is in both states simultaneously; the ensemble is in perhaps one or perhaps the other, characterized by probabilities - that uncertainty appears because we do not have any information of the state of the B-subsystem. Let us illustrate this in the simplest case of a two-qubit system A, B. Consider a pure quantum state of the form

$$\psi_{AB} = a\,|00\rangle + b\,|11\rangle \ . \tag{129}$$

A and B are correlated in that state, one can predict one by knowing another: the measurement of the second qubit always gives the same result as the measurement of the first one. Now any operator acting on A gives

$$\begin{aligned}&\langle\psi_{AB}|\hat{O}_A \otimes \hat{I}_B|\psi_{AB}\rangle = (a^*\langle 00| + b^*\langle 11|)\hat{O}_A \otimes \hat{I}_B|(a|00\rangle + b|11\rangle) \\ &= |a|^2\langle 0|\hat{O}_A|0\rangle + |b|^2\langle 1|\hat{O}_A|1\rangle \,,\end{aligned} \tag{130}$$

That corresponds to a diagonal density matrix:

$$\begin{aligned}\rho_A &= Tr_B\left(|a|^2\,|00\rangle\,\langle 00| + |b|^2\,|11\rangle\,\langle 11| + a^*b\,|00\rangle\,\langle 11| + ab^*\,|11\rangle\,\langle 00|\right) \\ &= |a|^2\,|0\rangle\,\langle 0| + |b|^2\,|1\rangle\,\langle 1| = \begin{bmatrix} |a|^2 & 0 \\ 0 & |b|^2 \end{bmatrix} .\end{aligned} \tag{131}$$

We can interpret this as saying that the system A is in a mixed state, that is with probability $|a|^2$ in the quantum state $|0\rangle$, and with probability $|b|^2$ it is in the state $|1\rangle$. Due to the orthogonality of B-states, the same results (131,132) are obtained if $\langle 0|\hat{O}_A|1\rangle \neq 0$ and for whatever relative phase between a and b, in distinction from (127). Being in a superposition is not the same as being in mixed state, where the relative phases of the states $|0\rangle\,, |1\rangle$ are experimentally inaccessible.

The system is called *entangled* [23] if its density matrix has more than one nonzero eigenvalue, so that there is more than one term in the sum (129).

We characterized the uncertainty in classical physics by a probability vector $\{p_i\}$ and in quantum mechanics by a state vector ψ_i. In quantum statistics, we need a matrix, generally non-diagonal, whose $i - j$ element quantifies how these two states of A are correlated via all possible states of B.

[23] Early idea of entanglement was conjured in 17 century: it was claimed that if two magnetic needles were magnetized at the same place and time, they would stay "in sympathy" forever at however large distances, and the motion of one is reflected on another. One con man tried to sell this communication device to Galileo, who didn't buy it.

6.3 Entanglement entropy

One can ascribe to any density matrix ρ_A the entropy by the formula analogous to the Gibbs-Shannon entropy of a probability distribution (von Neumann 1927, 1932):

$$S(\rho_A) = -\mathrm{Tr}\, \rho_A \log \rho_A \,. \tag{132}$$

Since we are dealing with diagonalizable matrices, a logarithm (or any other function) of the matrix is defined for a diagonal matrix: if $\rho = \sum_k p_k \left|k\right\rangle \left\langle k\right|$, then $\log \rho = \sum_k \log(p_k) \left|k\right\rangle \left\langle k\right|$. To avoid confusion, we shall always use Greek letters for the argument of von Neumann entropy and Latin letters for the argument of Shannon entropy.

The von Neumann entropy quantifies the sort of uncertainty, which exists only in a quantum world and is related to principal restriction of measurements to a finite volume. The classical entropy is the logarithm of the number of microstates compatible with the given macroscopic state. The quantum entropy $S(\rho_A)$ is, roughly speaking, the logarithm of the number of states of the inaccessible part B of the universe compatible with all measurements of A, together with a priori knowledge that A+B is in a pure state.

Evidently, $S(\rho_A)$ is invariant under a unitary transformation $\rho_A \to U\rho_A U^{-1}$, which is an analog of the Liouville theorem on the conservation of distribution by Hamiltonian evolution. Just like the classical entropy, it is non-negative, equals to zero only for a pure state and reaches its maximum $\log d$ for equipartition (when all d non-zero eigenvalues are equal), that is satisfies concavity (45). What does not have a classical analog is that the purifying system B has the same entropy as A (since the same p_i appears in its density matrix). Moreover, von Neumann entropy of a part $S(\rho_A)$ can be larger than that of the whole system $S(\rho_{AB})$. When AB is pure, $S(\rho_{AB}) = 0$, but $S(\rho_A)$ could be nonzero (Landau 1927). Information can be encoded in the correlations among the parts, yet be invisible when we look at one part of a quantum system. That purely quantum correlation between different parts is called entanglement, and the von Neumann entropy of a subsystem of pure state is called *entanglement entropy*.

Classically, we measured the nonlocality of information encoding by the mutual information $I(A, B) = S(A) + S(B) - S(A, B)$, which never exceeds the sum of two entropies. Quantum I is non-negative like classical, but generally is different. Nonlocality of information encoding is raised to the whole new level in the quantum world. For example, when AB is in an entangled pure state then $S(\rho_{AB}) = 0$, so that A and B together are perfectly correlated, but separately each one is in a mixed state with $S(\rho_A) = S(\rho_B) > 0$. Classically, the mutual information of perfectly correlated quantities is equal to each of their entropies, but quantum mutual information is their sum that is twice more: $I(\rho_{AB}) = S(\rho_A) + S(\rho_B) -$

$S(\rho_{AB}) = 2S(\rho_A)$. Quantum correlations are stronger than classical.

The von Neumann entropy of a density matrix is the Shannon entropy $S(p) = -\sum_i p_i \log p_i$ of its vector of eigenvalues, which is the probability distribution $\{p_i\}$ of its orthonormal eigenstates. In particular, for $\psi_{AB} = a\,|00\rangle + b\,|11\rangle$ we have $S(\rho_A) = -|a|^2 \log_2 |a|^2 - |b|^2 \log_2 |b|^2$. The maximum $S(\rho_A) = 1$ is reached when $|a|^2 = |b|^2 = 1/2$, which is called a state of maximal entanglement. In this case, when we trace out B (or A) we wipe out the information about the whole: any measurement on A or B cannot tell us anything about the state of the pair, since both outcomes are equally probable. On the contrary, when either $b \to 0$ or $a \to 1$, the entropy $S(\rho_A)$ goes to zero and measurements (of either A or B) give us definite information on the state of the system.

Original von Neumann argument involved a mixing process. Consider a gas of molecules, where pN are in a pure state $|a\rangle$ and $(1-p)N$ are in an orthogonal state $|b\rangle$. It is described by the N-the power of the density matrix $\rho = p\,|a\rangle\,\langle a| + (1-p)\,|b\rangle\,\langle b|$. Orthogonality of the states means that a-molecules can be separated from b-molecules, for instance, by a wall permeable only for one state. We then double our volume and moving a, b-walls from opposite ends to the center completely separate the molecules. We assume that the entropy is zero in such a state. The densities in the respective halves are different: their ratio is $p/(1-p)$. We can now squeeze one half by the factor p and another by $(1-p)$, making the densities equal and returning the whole volume to the original value. We can now restore the original mixture by removing the partition. Since the squeezing at a constant temperature T required the heat $-T[p \log p + (1-p)\log(1-p)] = TS(p)$, which increases the entropy by $S(p)$, then it is the mixing entropy of the original mixture.

Let $\rho = \sum_k p_k |\psi^k\rangle\langle\psi^k|$ be diagonal in the basis of eigenvectors $\{|\psi^k\rangle\}$, but we measure by projecting ρ on a different orthogonal set $\{|\phi^i\rangle\}$. In this case, the outcome i happens with the probability $q_i = \langle\phi^i|\rho|\phi^i\rangle = \sum_k p_k D_{ik}$, where $D_{ik} = |\langle\phi^i|\psi^k\rangle|^2$ is so-called double stochastic matrix, that is $\sum_i D_{ik} = \sum_k D_{ik} = 1$. The Shannon entropy of that probability distribution is larger than the von Neumann entropy,

$$S(q) = S(p) + \sum\nolimits_{ik} p_k D_{ik} \log\left(\sum\nolimits_n p_n D_{in}/p_k\right) = S(p) + D(q|p) \geq S(p) = S(\rho)\,,$$

that is such measurements are less predictable. Mathematically, the interpretation is that the diagonal elements are more random than the eigenvalues for a nonnegative Hermitian matrix.

We can now establish a *general uncertainty relation.* If we measure a *mixed* state ρ by projecting onto the orthonormal basis $\{|x\rangle\}$, the outcomes define the density matrix $\hat{M}_x\rho = \rho_x = \sum_x |x\rangle\,\langle x|\rho|x\rangle\,\langle x|$. The measurement operator $\hat{M}_z$ projecting onto another basis $\{|z\rangle\}$ defines $\hat{M}_z\rho = \rho_z = \sum_z |z\rangle\,\langle z|\rho|z\rangle\,\langle z|$. Both density matrices are diagonal, so that each von Neumann entropy is equal to the respective Shannon entropy: $S(\rho_x) = S(X)$ and $S(\rho_z) = S(Z)$. We now introduce the relative entropy for density matrices:

$$D(\rho|\rho_x) = Tr\,\rho(\log\rho - \log\rho_x) = Tr\,\rho\log\rho - Tr\,\rho_x\log\rho_x = S(X) - S(\rho)\,.$$

Here we have used the property of trace: $Tr\,\rho\log\rho_x = Tr\,\rho_x\log\rho_x$. As in the classical case, D is non-negative and quantifies the number of measurements needed to distinguish two density matrices. It also possesses an important property of monotonicity, that is non-increases upon any partial trace. This property is as intuitive as in the classical case — after all, it should be no easier to distinguish two density matrices looking only at subsystem, yet the proof is complicated and we do not give it here. We now use monotonicity of the relative entropy $D(\rho|\rho_x)$ under the action of the measurement in z-basis:

$$D(\rho|\rho_x) \geq D(\hat{M}_z\rho|\hat{M}_z\rho_x) = D(\rho_z|\hat{M}_z\rho_x) = -S(Z) - Tr\,\rho_z\log\hat{M}_z\rho_x\ . \qquad (133)$$

The new density matrix obtained by two measurements,

$$\hat{M}_z\rho_x = \hat{M}_z\hat{M}_x\rho = \sum\nolimits_z |z\rangle \sum\nolimits_x \langle z\,|x\rangle\,\langle x|\rho|x\rangle\,\langle x|\,z\rangle\,\langle z|\ ,$$

is diagonal, so that

$$\log\hat{M}_z\rho_x = \sum\nolimits_z |z\rangle\log\left(\sum\nolimits_x \langle z\,|x\rangle\,\langle x|\rho|x\rangle\,\langle x|\,z\rangle\right)\langle z|\ \ .$$

The logarithm is a monotonic function:

$$\log\left(\sum_x \langle z\,|x\rangle\,\langle x|\rho|x\rangle\,\langle x|\,z\rangle\right) \leq \log\left(\max_{x,z}|\langle x|z\rangle|^2\sum_x\langle x|\rho|x\rangle\right) = \log\left(\max_{x,z}|\langle x|z\rangle|^2\right)\ .$$

Substituting that into (134), we obtain the generalization of the uncertainty relation for a mixed state

$$S(X) + S(Z) \geq log(1/c) + S(\rho)\,, \quad c = \max_{x,z}|\langle x|z\rangle|^2\,.$$

Both sources of uncertainty in quantum statistics are here: non-orthogonality of states quantified by c and entanglement quantified by $S(\rho)$. Comparing that with the uncertainty relations (125,126) written for a pure state, we see that the von Neumann entropy quantifies the increase in uncertainty due to entanglement with the environment.

Coming to equilibrium. When a classical system is getting attached to a thermostat, it comes to thermal equilibrium with it, attaining maximum of entropy determined by the temperature of the thermostat. But what if a quantum system is getting attached to a large system with which they together form a pure quantum state with a zero entropy? Are thermalization and entropy growth possible for subsystems of a quantum system which as a whole remains in a pure quantum state? Yes they are! Thermalization takes place for any subsystem of a large system if the dynamics is ergodic and can be characterized by the growth of the entanglement entropy. Then the system as a whole acts as a thermal reservoir for its subsystems, provided those are small enough.

Consider a small quantum system which at some moment is getting attached to a large system. At this moment, the information is encoded locally, the entanglement entropy is zero and the subsystem is not in equilibrium with the whole system. As the small subsystem starts interacting with the large system and approaches equilibrium, the von Neumann entropy grows and reaches its maximum. Information, which was initially encoded locally in an out-of-equilibrium state, becomes encoded more and more nonlocally as the system evolves, eventually becoming invisible to an observer confined to the subsystem. Such thermalization can be quantified by a relative entropy. Denote the (evolving) density matrix of our subsystem ρ. If the evolution of the subsystem, when it is closed, is described by the Hamiltonian H, we can define the Gibbs density matrix as

$$\rho_0 = \frac{\exp(-\beta H)}{Tr\,\exp(-\beta H)} = Z^{-1}\exp(-\beta H)\ . \tag{134}$$

We now define the respective free energies via the energy and the von Neumann entropy:

$$F(\rho) = E - \beta^{-1}S(\rho) = Tr\,\rho H + \beta^{-1} Tr\,\rho\ln\rho = \beta^{-1}Tr\,\rho(\ln\rho + \beta H)\,,$$
$$F_0 = \beta^{-1}Tr\,\rho_0(\ln\rho_0 + \beta H) = -\beta^{-1}\ln Z = -\beta^{-1}\ln Tr\,\exp(-\beta H)\ .$$

The relative von Neumann entropy between ρ and ρ_0 can be expressed via the difference in the free energies:

$$\begin{aligned} D(\rho|\rho_0) &= Tr\,\rho\ln\rho - Tr\,\rho\ln\rho_0 \\ &= Tr\,\rho(\ln\rho + \beta H) + \beta^{-1}\ln Tr\,\exp(-\beta H) = \beta[F(\rho) - F_0] \geq 0\ , \end{aligned} \tag{135}$$

where the last inequality follows from positivity of the relative entropy. Therefore, the Gibbs state has the lowest free energy at a given temperature, which is determined by its environment treated as thermostat. Unitary evolution of the subsystem and its environment induces on a subsystem a decrease (by monotonicity) of $D(\rho, \rho_0)$, eventually bringing the subsystem to the Gibbs state.

6.4 Quantum communications

Without going into specifics of quantum processors and communication schemes, here we discuss how much information one can transfer by sending (or sharing) quantum objects. Let us first ask: How many bits of classical information can be recovered from a quantum system? Even though any qubit potentially contains a complex number, any measurement will only give one or another state, so that a pure state of a qubit can store one classical bit. The four orthogonal maximally entangled states of the qubit pair, $(|00\rangle \pm |11\rangle)/\sqrt{2}$ and $(|01\rangle \pm |10\rangle)/\sqrt{2}$ can store two bits. Generally, sending a quantum system whose state is determined by a d-dimensional complex vector, one can send at most $\log d$ bits of classical information (for instance, by sending one of the states from d basic vectors).

How this is related to the quantum mutual information can be realized by looking at a more symmetric variant of the same problem. Let a composite system AB be described by the density matrix ρ_{AB}. Alice has access to A, while Bob has access to B. The results of the measurements belong to classical information (can be written in the notebooks C_A and C_B). The maximal number of bits Alice can get from her measurements about those of Bob (and vice versa) is the classical mutual information between their notebooks, $I(C_A, C_B)$. Measurements correspond to tracing out some degrees of freedom, so that monotonicity guarantees that $I(C_A, C_B) \leq I(\rho_{AB}) \leq \log d$, where $I(\rho_{AB}) = S(\rho_A) + S(\rho_B) - S(\rho_{AB})$ is the mutual information of the initial density matrix ρ_{AB}.

Let us now turn to quantum information, that is to the information about quantum states themselves rather that to the measurements results. We pose the same natural question we asked for classical communications in Section 3.2: How much can a message be compressed, that is what is the maximum information one can transmit per quantum state? Is it given by von Neumann entropy or by Shannon entropy as in the classical case? Now the letters of our message are quantum states picked with their respective probabilities p_k, that is each letter is described by the density matrix and the message is a tensor product. Leaving aside how actual quantum communication devices handle information compression, we discuss here only the amount of quantum information, that is the number of combinations of states involved. If the states are mutually orthogonal and the density matrix is diagonal, it is essentially the classical case, that is the answer is given by the Shannon entropy $S(p) = -\sum_k p_k \log p_k$, which is the same as von Neumann entropy in this case. For example, the output of a qubit source which sends $|0\rangle$ with probability p and $|1\rangle$ with probability $1 - p$ can be compressed similarly to the classical source described in Section 3.2.

The new issue in quantum information theory is that nonorthogonal states cannot be perfectly distinguished, a feature with no classical analog. If a pure state AB was built from non-orthogonal states, taken with probabilities q_i, then the density matrix ρ_A is non-diagonal. There is then the difference between the Shannon entropy of the mixture and the von Neumann entropy of the matrix, $S\{q_i\} - S(\rho_A)$. It is non-negative and quantifies how much distinguishability is lost when we mix nonorthogonal pure states. Measuring ρ_A we can receive $S(\rho_A)$ bits, which is less than $S\{q_i\}$ bits that was encoded mixing the states with probabilities $\{q_i\}$.

For example, non-orthogonal states $|0\rangle$ and the superposition $|s\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$ cannot be distinguished if a measurement in the basis $|0\rangle, |1\rangle$ brings $|0\rangle$, only when it brings $|1\rangle$. Consider an output producing non-orthogonal states $|0\rangle$ with the probability p and $|s\rangle$ with the probability $1-p$. Sending classical information about these states bring the information $S(p)$. Could we use different encoding corresponding to a shorter mean length of a codeword? The letters of our alphabet, $|0\rangle$ and $|s\rangle$, both contain the state $|0\rangle$, which means redundancy. The redundancy must allow for tighter compression than $S(p)$. That can be demonstrated using essentially the argument from Section 3.2 with the only difference that instead of typical sequences we consider typical subspaces. Could we decrease the entropy using the orthogonal states $|0\rangle, |1\rangle$? Indeed, a long N-string emitted by the source will look like a superposition of the terms having up to reordering the following form:

$$|0\rangle^{\otimes Np} |s\rangle^{\otimes N(1-p)} \approx |0\rangle^{\otimes N(1+p)/2} |1\rangle^{\otimes N(1-p)/2} \ . \tag{136}$$

This is because in the limit $N(1-p) \gg 1$ the product $|s\rangle^{\otimes N(1-p)}$ can be approximated by the superposition of the states with equal probability of $|0\rangle$ and $|1\rangle$. The number of the states of the form (137) is given by the number of $N(1+p)/2$ choices out of N; the logarithm of the number of states by the Stirling formula is $NS(1/2+p/2)$. If $(1+p)/2 > 1-p$, that is $p > 1/3$, we can choose a coding scheme, where we use the states of the form (137) as the new alphabet letters and neglect atypical states. We then achieve compression for sending classical information about these states, since $S(p) - S(1/2+p/2) = (1/2+p/2)\log(1/2+p/2) + (1/2-p/2)\log(1/2-p/2) - p\log p - (1-p)\log(1-p) > 0$ for $p > 1/3$. That bound makes sense since at $p < 1/3$ the chosen way of encoding actually increases redundancy, and one needs to use different encoding.

Yet the most efficient encoding uses the states, where the density matrix of our source is diagonal, instead of the states $|0\rangle, |1\rangle$, where it is not. Since the probabilities of the non-orthogonal states $|0\rangle$ and $(|0\rangle + |1\rangle)/\sqrt{2}$ are respectively p and $1-p$, the density matrix in the orthogonal basis $|0\rangle, |1\rangle$ is as follows:

$$\rho = p\,|0\rangle\langle 0| + \tfrac{1-p}{2}\,|0+1\rangle\langle 0+1| = \tfrac{1}{2}\begin{bmatrix} 1+p & 1-p \\ 1-p & 1-p \end{bmatrix} \ . \tag{137}$$

The eigenvalues are $q = (1 \pm \sqrt{2p^2 - 2p + 1})/2$ and $1 - q$ with the respective eigenstates $\sqrt{q}\,|1\rangle + \sqrt{1-q}\,|0\rangle$ and $-\sqrt{q}\,|0\rangle + \sqrt{1-q}\,|1\rangle$. The von Neumann entropy is the Shannon entropy in this orthonormal representation: $S(\rho) = S(q)$. One can show that $S(q) \leq S(p)$ for any p. In particular, for $p = 1/2$, we have $q = (2 + \sqrt{2})/2 = \sin^2(\pi/8)$ and

$$S(\rho) = S(q) = -q \log q - (1-q)\log(1-q) = 1 + \frac{1}{2} - \frac{1}{2\sqrt{2}} \log \frac{2+\sqrt{2}}{2-\sqrt{2}} \approx 0.6 \text{ bits} ,$$

that is indeed less than $S(p) = 1$.

The number of typical strings is then given by the Shannon entropy of this representation, which is now equal to the von Neumann entropy and is strictly lower than the Shannon entropy for any representation where the density matrix is not diagonal, as we discussed above.

We thus conclude that the best possible rate of quantum information transfer is given by the von Neumann entropy of the density matrix of the source. Yet the von Neumann entropy $S(\rho)$ gives the number of qubits of quantum information carried per letter of a long message only when we deal with a mixture of pure states. This is not true when $\rho = \sum_k p_k \rho_k$ and ρ_k are mixed states. It is easy to see from a trivial example: Suppose that a particular mixed state ρ_0 with $S(\rho_0) > 0$ is chosen with probability $p_0 = 1$. Then the message $\rho_0 \otimes \rho_0 \otimes \ldots$ carries no information.

When our alphabet is made of mixed yet mutually orthogonal states, then the states are distinguishable and the problem is classical, since we can just send the probabilities of the states, so the maximal rate is the Shannon entropy $S(p)$. However, it is *less than* the von Neumann entropy, which now includes also a nonzero entropy of every mixed state ρ_k. Indeed, because all ρ_k are orthogonal, they could be made diagonal simultaneously, and we obtain:

$$\begin{aligned} S(\rho) &= -\sum_k \mathrm{Tr}(p_k\rho_k) \log(p_k\rho_k) \\ &= -\sum_k (p_k \log p_k + p_k \mathrm{Tr}\, \rho_k \log \rho_k) = S(p) + \sum_k p_k S(\rho_k) \ . \end{aligned}$$

That shows that when ρ_k are mixed states, $S(\rho)$ is no longer a good measure of quantum entanglement since it clearly mixes quantum and classical correlations. In this case, von Neumann entropy exceeds the Shannon entropy:

$$S(p) = S(\rho) - \sum_k p_k S(\rho_k) = S\left(\sum_k p_k \rho_k\right) - \sum_k p_k S(\rho_k) \,. \tag{138}$$

To conclude, the information transfer rate
i) by orthogonal pure states is equal to $S(p) = S(\rho)$,

ii) by non-orthogonal pure states is equal to $S(\rho)$, which is less than $S(p)$,
iii) by orthogonal mixed states is equal to $S(p)$, which is less than $S(\rho)$.

For non-orthogonal mixed states, it is believed that

$$\chi(\rho_k, p_k) = S\left(\sum_k p_k \rho_k\right) - \sum_k p_k S(\rho_k)$$

(called in quantum communications Holevo information) defines the limiting compression rate in all cases including when it does not coincide with $S(p)$. The reason for the belief is that χ is monotonic (i.e. decreases when we take partial traces), but $S(\rho)$ is not - indeed one can increase von Neumann entropy by going from a pure to a mixed state. It follows from concavity that χ is always non-negative. We see that it depends on the probabilities p_k that is on the way we prepare the states. Of course, (139) is a kind of mutual information, it tells us how much, on the average, the von Neumann entropy of an ensemble is reduced when we know which preparation was chosen, exactly like classical mutual information $I(A, B) = S(A) - S(A|B)$ tells us how much the Shannon entropy of A is reduced once we get the value of B. So we see that classical Shannon information is a mutual von Neumann information. One also calls χ the accessible information of an ensemble of quantum states, that is the maximal number of bits of information that can be acquired about the preparation of the state on the average.

6.5 Conditional entropy and teleportation

Similar to the classical conditional entropy (50), one defines for von Neumann entropy

$$S(\rho_{AB}|\rho_B) = S(\rho_{AB}) - S(\rho_B) \ . \tag{139}$$

However, this is not an entropy conditional on something known, moreover it is not zero for correlated quantities but negative! Indeed, for pure AB, one has $S(\rho_{AB}|\rho_B) = -S(\rho_B) < 0$. Classical conditional entropy measures how many classical bits we need to add to B to fully determine A. Similarly, we would expect quantum conditional entropy to measure how many qubits Alice needs to send to Bob to reveal herself. But what does it mean when $S(\rho_{AB}|\rho_B)$ is negative?

That negativity is due to entanglement between A and B, which allows the trick of *teleportation*. Teleportation moves quantum states around without a quantum channel, and we shall see below that negative von Neumann conditional entropy counts the number of possible future teleportations. Imagine that Alice has in her possession a qubit A_0. Alice want Bob to create in his lab a qubit in a state identical to A_0. However, she is only able to communicate by sending a classical message. If Alice knows the state of her qubit, there is no problem (except that communicating a complex number exactly requires infinite number of classical bits): she tells Bob (say, over the telephone) the state of her qubit and he creates one like it in his lab. If, however, Alice does not know the state of her qubit, all she can do is to make a measurement, which will give some information about the prior state of qubit A_0. She can tell Bob what she learns, but the measurement will destroy the remaining information about A_0 and it will never be possible for Bob to recreate it. So she need to make a measurement revealing no information about A_0. Then what information that measurement reveals? It must be about something else which Alice and Bob share.

Suppose then that Alice and Bob have previously shared a qubit pair A_1, B_1 in a known entangled state, for example,

$$\psi_{A_1B_1} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{A_1B_1} . \tag{140}$$

Bob then took B_1 with him, leaving A_1 with Alice. In this case, Alice can solve the problem by making a joint measurement of her system A_0A_1 in a basis, that is chosen so that no matter what the answer is, Alice learns nothing about the prior state of A_0. In that case, she also loses no information about A_0. But after getting her measurement outcome, she knows the full state of the system and she can tell Bob what to do to recreate A_0. To see how this works, let us describe a specific measurement that Alice can make on A_0A_1 that will shed no light on the state of A_0. The measurement must be a projection on a state where the probability of A_0 to be in the state $|0\rangle$ is exactly equal to the probability to be in the state $|1\rangle$. The following four states of A_0A_1 satisfy that property:

$$\frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle)_{A_0A_1} , \quad \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle)_{A_0A_1} . \tag{141}$$

The states are chosen to be entangled, that is having A_0 and A_1 correlated. We don't use the state with $|00\rangle \pm |10\rangle$, which has equal probability of zero and one for A, but no correlation between the values of A_0 and A_1.

Denote the unknown initial state of the qubit A_0 as $\alpha\,|0\rangle + \beta\,|1\rangle$, then the initial state of $A_0A_1B_1$ is

$$\frac{1}{\sqrt{2}}(\alpha\,|000\rangle + \alpha\,|011\rangle + \beta\,|100\rangle + \beta\,|111\rangle)_{A_0A_1B_1} . \tag{142}$$

Let's say that Alice's measurement, that is the projection on the states (142), reveals that A_0A_1 is in the state

$$\frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)_{A_0A_1} . \tag{143}$$

That means that only the first and the last terms in (143) contribute (with equal weights but opposite signs). After that measurement, B_1 will be in the state $(\alpha\,|0\rangle - \beta\,|1\rangle)_{B_1}$, whatever the (unknown) values of α, β. Appreciate the weirdness of the fact that B_1 was uncorrelated with A_0 initially, but instantaneously acquired correlation after Alice performed her measurement thousand miles away. Knowing the state of B_1, Alice can send two bits of classical information, telling Bob that he can recreate the initial state $\alpha\,|0\rangle + \beta\,|1\rangle$ of A_0 by multiplying the vector of his qubit B_1 by the matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, that switches the sign of the second vector of the basis. The beauty of it is that Alice learnt and communicated not what was the state A_0, but how to recreate it.

To understand the role of the quantum conditional entropy (140) in teleportation, we symmetrize and purify our problem. Generally, weirdness of quantum entropies can be traced to the purely quantum possibility of purification. Notice that A_1 and B_1 are maximally entangled (come with the same weights), so that $S(\rho_B) = \log_2 2 = 1$. On the other hand, A_1B_1 is in a pure state so its von Neumann entropy is zero. Let us now add another system R which is maximally entangled with A_0 in a pure state A_0R, say

$$\psi_{A_0R} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{A_0R}\,. \tag{144}$$

Neither Alice nor Bob have access to R. From this viewpoint, the combined system $RAB = RA_0A_1B_1$ starts in a pure state which is a direct product $\psi_{RA_0} \otimes \psi_{A_1B_1}$. Since A_0 is maximally entangled with R then also $S(\rho_{A_0}) = \log_2 2 = 1$ and the same is the entropy of the AB system $S(\rho_{A_0A_1B_1}) = S(\rho_{A_0}) = 1$ since A_1B_1 is a pure state. Therefore, $S(\rho_{AB}|\rho_B) = S(\rho_{A_0A_1})|\rho_B) = S(\rho_{A_0A_1B_1}) - S(\rho_{B_1}) = 0$. One can show that teleportation only possible when $S(\rho_{AB}|\rho_B)$ is non-positive.

Recall that classically $S(A|B)$ measures how many additional bits of information Alice has to send to Bob after he has already received B, so that he will have full knowledge of A. Quantum analog of this involves qubits rather than classical bits. Suppose that $S(\rho_{AB}|\rho_B) > 0$ and Alice nevertheless wants Bob to recreate her states. She can simply send her states. Alternative is to do teleportation, which requires sharing with Bob an entangled pair for every qubit of her state to be teleported. Either way, Alice must be capable of *quantum communication*, that is of sending a quantum system while maintaining its quantum state. For teleportation, she first creates some maximally entangled qubit pairs and sends half of each pair to Bob. Each time she sends Bob half of a pair, $S(\rho_{AB})$ is unchanged but $S(\rho_B)$ goes up by 1, so $S(\rho_{AB}|\rho_B) = S(\rho_{AB}) - S(\rho_B)$ goes down by 1. So $S(\rho_{AB}|\rho_B)$, if positive, is the number of such qubits that Alice must send to Bob to make $S(A|B)$ non-positive and so make teleportation possible without any further quantum communication. Negative quantum conditional entropy measures the number of possible future qubit teleportations. We thus see that entanglement is an important resource in quantum communications.

6.6 The way out is via a black hole

All things physical are information-theoretic in origin.
J Wheeler, 1990

A black hole presents a way to eliminate all uncertainty about a system by swallowing it, thus forever eliminating from our world. No body, no uncertainty. Our religious belief that uncertainty can only increase leads us to the entropy of a black hole and to the ultimate restriction on the amount of information which can be encoded in a physical system.

Area law. Black hole is an object, whose size is smaller than its horizon $r_h = 2GM/c^2$, where M is the mass and G is the gravitational constant. One cannot escape from within r_h, since the speed needed for that exceeds c. The quantum entanglement entropy (between interior and exterior) is thought to be responsible for the entropy of black holes. To estimate it, we need an equation of state, that is the relation between energy and temperature. The energy of the hole is simply $E_{BH} = Mc^2 = c^4 r_h/2G$. The temperature of the hole is determined by its radiation, which is due to a purely quantum phenomenon of particle-antiparticle pairs appearing from vacuum fluctuations. Such pairs usually stay together and soon annihilate. If, however, such a pair straddles the horizon, then the inside part is absorbed by the hole, while the outside part can escape and be registered as radiation (this is how the entanglement appears). The typical wavelength of such radiation is r_h and its energy/temperature is then $T = \hbar c/4\pi r_h$. Now we can obtain entropy integrating the equation of state $T = dE/dS$:

$$T = \frac{\hbar c}{4\pi r_h} = \frac{dE_{BH}}{dS_{BH}} = \frac{c^4}{2G}\frac{dr_h}{dS_{BH}} \Rightarrow S_{BH} = \frac{\pi r_h^2 c^3}{G\hbar} \, .$$

Since any entropy is dimensionless, then $\hbar G/c^3$ must be a square of some fundamental length. It is called the Planck length, $l_p = \sqrt{\hbar G/c^3} \simeq 10^{-17}\, cm$, and it is the only combination with that dimensionality of the three fundamental physical constants, c, $\hbar$, G; it is the scale where quantization of gravity is expected to be important[24]. The entanglement entropy of a black hole can thus be written as

[24] One-parameter theories: G-theory, 17th century; c-theory, 19-20 centuries; $\hbar$-theory, early 20th century. Two-parameter theories: c, G - general relativity, $c, \hbar$ -quantum electrodynamics, 20th century. Hopefully, c, G, h-theory will appear in the 21st century.

follows:

$$S_{BH} = \frac{\pi r_h^2 c^3}{G\hbar} = \frac{\pi r_h^2}{l_p^2} = \frac{4\pi G M^2}{\hbar c} \ . \tag{145}$$

The area law behavior of the entanglement entropy in microscopic theories could be related to the holographic principle — the conjecture that the information contained in a volume of space can be completely encoded by the degrees of freedom which live on the boundary of that region.

Bekenstein bound. We can now estimate the information capacity (not a channel capacity!) defined as the maximal amount of information that can be encoded in a system by exploiting all of its degrees of freedom down to the quantum level. Is there a universal limit on how large the entropy of a physical system can be? The answer is given by the so-called Bekenstein bound (and its generalizations). On a dimensional ground it can be guessed as follows. The entropy must be the total energy E (including any rest masses) divided by a temperature (in energy units). The temperature must be determined by the system size R — the smaller the size the higher the temperature. Indeed, confining a system to a smaller region by quantum indeterminacy one increases the kinetic energy. The only combination with the dimensionality of energy one can make out of R and the world constants $\hbar, c$ is $\hbar c/R$. That suggests the bound in the following form: $S \le RE/\hbar c$ (Bekenstein 1981, 2004; Casini 2008). That bound was argued by exploiting the only known way to eliminate entropy from the observable world — to drop it into a black hole. If we drop a body of the energy E and entropy S into a black hole of large mass $M \gg E/c^2$, then the black hole's mass will grow by E/c^2. According to (146), the entropy of the hole will then grow by $8\pi GME/\hbar c^3$ plus a negligible term of order E^2. Meanwhile the entropy S has gone forever out of this world. The second law then requires that $S < 8\pi GME/\hbar c^3 = 4\pi r_h E/\hbar c$. A black hole, that can absorb the body, must have the horizon exceeding the body size, which gives the estimate for the bound (extra numerical numerical factor 1/2 comes from the actual consideration of the process of adiabatic lowering of the body into the hole):

$$S \le \frac{2\pi RE}{\hbar c} \ . \tag{146}$$

We assumed that the body is itself not a black hole, that is its size exceeds its horizon, $R > r_h(E) = 2GE/c^4$, so that the entropy restriction can be formulated solely in terms of the radius:

$$S \le \frac{\pi R^2 c^3}{G\hbar} \ . \tag{147}$$

Comparing (148) and (146), we conclude that a system must be a black hole to realize the capacity limit. Note without elaboration that (147,148) actually refer to the difference between the entropy of the system with the energy E and the entropy of the quantum vacuum in the region of the size R.

In a thermodynamic limit, the classical total entropy is extensive, that is proportional to the system volume or total number of degrees of freedom. In other words, the entropy is proportional to a volume as long as one can squeeze more and more distinguishable matter into it. When there is so much matter or so little space that the system turns into a black hole, we can see only the horizon and the entropy is proportional to the area (like a hologram where 3d image is encoded on 2d surface).

While the gravitational constant G does not enter (147), the appearance of gravity in the argument and in (148) deserves reflection. Via black holes, gravity provides the gates out of the observable world, which is a source of the bound. A counterpart to it is a Big Bang which provided a gate into this world — how something comes out nothing could probably teach us important lessons about the nature of information as well.

7 Conclusion

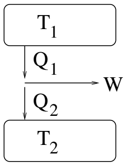
This Chapter attempts to compress the book to its most essential elements.

7.1 Take-home lessons

1. Thermodynamics studies restrictions imposed by hidden on observable. It deals with two extensive quantities. The first one (energy) E is conserved for a closed system, and its changes are divided into work (due to observable degrees of freedom) and heat (due to hidden ones). The second quantity (entropy) S can only increase for a closed system and reaches its maximum in thermal equilibrium, where the system entropy is a convex function of the energy. All available states lie below this convex curve in $S-E$ plane.

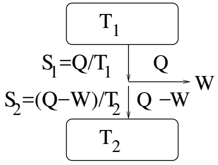
2. Convexity of the dependence $E(S)$ allows us to introduce temperature as the derivative of the energy with respect to the entropy. Extremum of the entropy means that the temperatures of the connected subsystems are equal in equilibrium. The same is true for the energy derivatives with respect to volume and other extensive variables. The entropy increase (called the second law of thermodynamics) imposes restrictions on thermal engine efficiency, that is the fraction of heat used for work:

$$\frac{W}{Q_1}=\frac{Q_1-Q_2}{Q_1}=1-\frac{T_2\Delta S_2}{T_1\Delta S_1}\leq 1-\frac{T_2}{T_1}\ .$$



If information processing generates $\Delta S = S_2 - S_1 = (Q - W)/T_2 - Q/T_1$, its energy price is as follows:

$$Q = \frac{T_2 \Delta S + W}{1 - T_2/T_1} \ .$$



3. Need in statistics appear due to incomplete knowledge: we are able to follow only part of the degrees of freedom and only with a finite precision. Statistical physics defines the (Boltzmann) entropy of a closed system as the log of the phase volume, $S = \log \Gamma$ and assumes (for the lack of any knowledge) the uniform distribution $w = 1/\Gamma$ called microcanonical. For a subsystem, the (Gibbs) entropy is defined as the mean phase volume: $S = -\sum_i w_i \log w_i$; the probability distribution is then obtained requiring maximal entropy for a given mean energy: $\log w_i \propto -E_i$. Information theory generalizes this approach, see 13 below.

4. Irreversible entropy growth may seem to contradict Hamiltonian dynamics, which is time-reversible and preserves the N-particle phase-space density. However, one can obtain the equation on a one-particle density for a dilute gas. Assuming that before every collision particles were independent, one obtains the Boltzmann kinetic equation, which, in particular, describes the irreversible growth of the one-particle entropy. Therefore, the difference must grow between the growing sum of one-particle entropies and the constant total entropy. That difference describes correlations and is called the mutual information. The lesson is that if we follow precisely all the degrees of freedom, the entropy is conserved and no information is lost. But if we follow only part of them, the entropy of that part will generally grow as it interacts with the rest — whatever information we had is getting less relevant with time. Similarly, the thermalization of a quantum subsystem increases the entanglement entropy since the information is getting encoded in interaction with the environment and inaccessible locally.

5. Total entropy growth can appear even if we follow all the degrees of freedom, but do it with a finite precision, that is consider evolution of finite phase-space regions. Instability leads to separation of trajectories, which spread over the whole phase space under a generic reversible Hamiltonian dynamics, very much like flows of an incompressible liquid are mixing (metaphorically, extra digits in precision add new degrees of freedom for unstable systems). Spreading and mixing in phase space correspond to the approach to equilibrium and entropy growth. On the contrary, to deviate a system from equilibrium, one adds external forcing and dissipation, which makes its phase flow compressible and distribution non-uniform.

6. Since uncertainty due to lack of knowledge plays such a prominent role, we wish to quantify it. The measure of uncertainly is the amount of information needed to remove it. We start in a discrete approach by receiving information

as answers to "yes-no" questions (called bits). The amount of information is the number of such answers, that is $\log_2 n$, where n is the number of equally probable possibilities (Boltzmann entropy) or the mean logarithm $-\sum_i p_i \log_2 p_i$ if the probabilities p_i are different from $1/n$ (Shannon-Gibbs entropy). Convexity of the function $-p\log p$ helps us to prove that the information/entropy has its maximum for equal probabilities (when our ignorance is maximal).

7. A simple mathematical notion of convexity is a powerful tool. We first use it in thermodynamics to make sure that the extremum is on the boundary of the region and to make Legendre transform of thermodynamic potentials. We then use convexity of the exponential function to show that even when the mean of a random quantity is zero, its mean exponent exceeds unity. That provides for an exponential separation of trajectories in an incompressible flow and exponential growth of the density of an element in a compressible flow. On the other hand, if the mean exponent is unity, $\langle e^{-\Delta S}\rangle = 1$, then the mean itself is negative: $-\langle \Delta S\rangle \leq 0$. Convexity of the entropic measures (including relative and von Neumann entropies) plays a central role in classical and quantum statistics. It is used to establish hierarchies and find the extremum.

8. In our discrete thinking, we use another basic mathematical object — the sum of independent random numbers $X = \sum_{i=1}^{N} y_i$. Three concentric statements were made. The weakest one is that X approaches its mean value $\bar{X} = N\langle y\rangle$ exponentially fast in N. The next statement is that the distribution $\mathcal{P}(X)$ is Gaussian in the vicinity of the width $\simeq N^{-1/2}$ around the maximum. The whole distribution is also very sharp, which is described by the large deviation form: $\mathcal{P}(X) \propto \mathrm{e}^{-NH(X/N)}$ where $H \geq 0$ and $H(\langle y\rangle) = 0$. Applying this to the log of the probability of a given sequence, $\lim_{N\to\infty} \log p(y_1 \dots y_N) = -NS(Y)$, we learn two lessons: i) the probability is independent of a sequence for most of them (almost all events are almost equally probable), ii) the number of typical sequences **grows exponentially and the entropy is the rate**.

9. The number of typical binary sequences of length N is then 2^{NS}, which cannot exceed 2^N. The efficient encoding of the typical sequences thus involves words with lengths from unity to NS, which is less than N if the probabilities of 0 and 1 are not equal. That means that the entropy is both the mean and the fastest rate of the reception of information brought by long messages/measurements. To squeeze out all the unnecessary bits, encoding is used both in industry and in nature, where sources often bring highly redundant information, like in visual signals.

10. If the transmission channel $B \to A$ makes errors, then the message does not completely eliminate uncertainty; what remains is the conditional entropy $S(B|A) = S(A,B) - S(A)$, which is the mean rate of growth of the number of possible errors. Sending extra bits to correct these errors lowers the transmission

rate from $S(B)$ to the mutual information $I(A,B) = S(B) - S(B|A) = S(A) + S(B) - S(A,B)$, which is the mean difference of the uncertainties before and after the message. The great news is that one can still achieve an asymptotically error-free transmission if the transmission rate is lower than I. The maximum of I over all source statistics is the channel capacity, which is the maximal rate of asymptotically error-free transmission. In particular, to maximize the capacity of sensory processing, the response function of a living beings or a robot must be a cumulative probability of stimuli.

11. Very often our goal is not to transmit as much information as possible, but to compress it and process as little as possible, looking for an encoding with a minimum of the mutual information. For example, the rate distortion theory looks for the minimal rate I of information transfer under the restriction that the signal distortion does not exceed the threshold $\mathcal{D}$. This is done by minimizing the functional $I + \beta\mathcal{D}$. Another minimization task could be to separate the signal into independent components with as little as possible (ideally zero) mutual information between them.

12. The conditional probability allows for hypothesis testing by the Bayes' rule: $P(h|e) = P(h)P(e|h)/P(e)$. That is the probability $P(h|e)$ that the hypothesis is correct after we receive the data e is the prior probability $P(h)$ times the support $P(e|h)/P(e)$ that e provide for h. Taking a log and averaging we obtain familiar $S(h|e) = S(h) - I(e,h)$. Bayes' approach demonstrates that there is no inference without prior assumption. If our hypothesis concerns the probability distribution itself, then the difference between the true distribution p and the hypothetical distribution q is measured by the relative entropy $D(p|q) = \langle \log_2(p/q) \rangle$. This is yet another rate — with which the error probability grows with the number of trials. D also measures the decrease of the transmission rate due to non-optimal encoding: the mean length of the codeword is not $S(p)$ but bounded by $S(p) + D(p|q)$. Mutual information is a particular case of relative entropy, they are both invariant with respect to arbitrary transformations of variables in a continuous case, which facilitates their ever-widening area of applications.

13. Since so much hangs on getting the right distribution, how best to guess it from the data? This is achieved by maximizing the entropy under the given data — "the truth and nothing but the truth". That explains and makes universal the entropy maximization from the point 3. What was thought to be a unique property of thermal equilibrium is now understood as a universally applicable common sense. It also sheds new light on physics, telling us that on some basic level all states are constrained equilibria. Whenever we encounter a trade-off, free energy appears, whose two terms quantify the opposite tendencies. Not only its (conditional) minima describe physical systems, but are presently the most powerful technical tools of optimization, from our Bayesian brain to machine-learning algorithms.

14. Information is physical. At a finite temperature both learning and erasing information requires work. The energetic price of a cycle is T times the mutual information between the system and the measuring device. Another side of the physical nature of information is that there is the (Bekenstein) limit on how much entropy one can squeeze inside a given radius; surprisingly, the limit is proportional to the area rather than the volume and is realized by black holes — our gates outside of this world.

15. The Renormalization Group (RG) is a best so far known way to forget information. As always with forgetting, the trick is to choose what to keep, which is decided by the renormalization. For example, we can either divide the sum of two random numbers by 2 keeping the mean or by $\sqrt{2}$ keeping the variance. That leads to different asymptotic distributions, which is the main focus of RG. We find that the entropy of the partially averaged and renormalized distribution is the proper measure of forgetting in simple cases, like adding random numbers on the way to the central limit theorem. In physical systems with many degrees of freedom, the quantity that changes monotonically upon RG can be the mutual information defined in two ways: either between remaining and eliminated degrees of freedom or between different parts of the same system.

16. Two central themes of quantum information and the two respective sources of quantum uncertainty are non-orthogonality and entanglement. The first theme appears already in quantum mechanics, respective uncertainty can be characterized by classical entropy. Quantum statistics appears when we treat subsystems and must deal with the density matrix and its von Neumann entropy. The quantum entropy of the whole can be less than the entropy of a part. In particular, the whole system can be in a pure state with zero entropy, then all the entropy of a subsystem comes from entanglement.

17. The last lesson is two progressively more powerful forms of the second law of thermodynamics, which originally was $\langle \Delta S \rangle \geq 0$. The first new form, $\langle e^{-\Delta S} \rangle = 1$, is the analog of a Liouville theorem. The second form relates the probabilities of forward and backward process: $\rho^\dagger(-\Delta S) = \rho(\Delta S) e^{-\Delta S}$.

7.2 Epilogue

The central idea of this course is that learning about the world means building a model, which is essentially finding an efficient representation of the data. Optimizing information transmission or encoding may seem like a technical problem, but it is actually the most important task of science, engineering and survival. Science works on more and more compact encoding of the strings of data, which culminates in formulating a law of nature, potentially describing infinity of phenomena. The mathematical tool we learnt here is an ensemble equivalence in the thermodynamic limit, its analog is the use of typical sequences in communication theory. The result is two roles of entropy: it defines maximum transmission and minimum compression.

Another central idea is that entropy is not a property of the physical world, but is an information we lack about it. And yet the information is physical — it has an energetic value and a monetary price. Indeed, the difference between work and heat is that we have information about the former but not the later. That means that one can turn information into work and one needs to release heat to erase information. We also have learnt that one not only pays for information but can turn information into money as well. The physical nature of information is manifested in the universal limit on how much of it we can squeeze into a space restricted by a given area.

The panoramic view accepted here works on different levels. Natural scientists see analogies between phenomena. One analogy discussed above is that measurements, predictions, recording retrievals, etc can all be treated and described uniformly as different forms of communication. Another analogy is between finding optimal strategy in economics (proportional gambling), biology (phenotype switching), engineering, data processing, perceptual inference, etc. On a higher level, mathematicians see analogies between analogies. For the above two analogies, the unifying mathematical notions are the relative entropy and free energy. Convexity is another example of a recurring mathematical notion unifying different approaches to the classes of phenomena, rather than phenomena themselves.

No rigorous proofs were given in this book, replaced instead by hand-waving arguments of varying plausibility or a particular example. More rigorous and detailed yet still compact deductive presentation of thermodynamics can be found in Callen, "Thermodynamics" (1965). Those interested in proofs for Chapter 2 can find them in Dorfman "An Introduction to Chaos in Nonequilibrium Statistical Mechanics". Detailed information theory with proofs can be found in Cowen & Thomas "Elements of Information Theory", whose Chapter 1 gives a concise overview. More practical and problem-oriented approach with numerous exercises can be found in MacKay "Information Theory, Inference and Learning algorithms". I wish also to stress that the examples given in this book represent a small slice of the ever-widening avalanche of applications; more biological applications can be found in "Biophysics" by Bialek, others in original articles and reviews. On quantum information the two comprehensive books are those by Preskill, and Nielsen & Chuang. Numerous references scattered through the text, like (Zipf 1949), give you the most compact encoding for a search.

Mention briefly several important subjects left out of this course. Our focus was largely (though not entirely) on finding a data description that is good on average. Yet there exists a closely related approach that focuses on finding the shortest description and ultimate data compression for a given string of data. The Kolmogorov complexity is defined as the shortest binary computer program able to compute the string. It allows us to quantify how much order and randomness is in a given sequence — truly random sequence cannot be described by an algorithm shorter than itself, while any order allows for compression. Complexity is (approximately) equal to the entropy if the string is drawn from a random distribution, but is actually a more general concept, treated in courses on Computer Science. Another fundamental issue not treated here is the dramatic difference between the classical and quantum classifications of computational complexity. Entropy as a measure of irreversibility also finds beautiful applications in geometry (see e.g. Perelman 2002).

Taking a wider view, I invite you to reflect on the history of our attempts to realize limits of possible, from heat engines through communication channels to computations. Will the next step be to study the natural limits of thinking and feeling?

Looking back one may wonder why accepting the natural language of information took so much time and was so difficult for physicists and engineers. Generations of students (myself including) were tortured by "paradoxes" in the statistical physics, which disappear when information language is used. I suspect that the resistance was to a large extent caused by the misplaced desire to keep scientist out of science. A dogma that science must be something "objective" and only related to the things independent of our interest in them obscures the simple fact that science is a form of human language. True, we expect it to be objectively independent of personality of this or that scientist as opposite, say, to literature, where we celebrate the difference between languages (and worlds) of Shakespear and Tolstoy. However, science is the language designed by and for humans, so that it necessarily reflects both the way body and mind operate and the restrictions on our ability to obtain and process the data. Presumably, omnipresent and omniscient being would have no need in the statistical information approach described here. As far as physics is concerned, I do not share the belief, widely held inside and outside of the discipline, that physicist's notions are truly objective and fundamental, as opposite even to biology (where distinction between organic and inorganic is due to our distinctively human interest in life), not speaking on linguistics or economics. I believe that we, physicists, can benefit from better appreciating the essential presence of scientist in science (for instance, to understand the special status of measurement in quantum mechanics).

As we learnt here, better understanding must lead to a more compact presentation; hopefully, the next edition will be shorter.

8 Appendix

8.1 Formal structure of thermodynamics

Both energy and entropy are homogeneous first-order functions of their variables: $S(\lambda E, \lambda V, \lambda N) = \lambda S(E, V, N)$ and $E(\lambda S, \lambda V, \lambda N) = \lambda E(S, V, N)$ (here V and N stand for the whole set of extensive macroscopic parameters). Differentiating the second identity with respect to λ and taking it at $\lambda = 1$ one gets the Euler equation

$$E = TS - PV + \mu N \ . \tag{148}$$

The equations of state are homogeneous of zero order, for instance,

$$T(\lambda E, \lambda V, \lambda N) = T(E, V, N) \, .$$

That confirms that the temperature, pressure and chemical potential are the same for a portion of an equilibrium system as for the whole system.

Generally, thermodynamics can be developed for as many quantities as we observe. But what is the minimal number of observables for a meaningful description? It may seem that a thermodynamic description of a one-component mechanical system requires operating functions of three intensive variables. Let us show that the homogeneity leaves only two independent parameters. For example, the chemical potential μ can be found as a function of T and P. Indeed, differentiating (149) and comparing with (5) one gets the so-called Gibbs-Duhem relation (in the energy representation) $Nd\mu = -SdT + VdP$ or for quantities per mole, $s = S/N$ and $v = V/N$: $d\mu = -sdT + vdP$. In other words, one can choose $\lambda = 1/N$ and use first-order homogeneity to get rid of the variable N, for instance: $E(S, V, N) = NE(s, v, 1) = Ne(s, v)$. In the entropy representation,

$$S = E\frac{1}{T} + V\frac{P}{T} - N\frac{\mu}{T} \, ,$$

the Gibbs-Duhem relation again states that because $dS = (dE + PdV - \mu dN)/T$ then the sum of products of the extensive parameters and the differentials of the corresponding intensive parameters vanish:

$$Ed(1/T) + Vd(P/T) - Nd(\mu/T) = 0 \ . \tag{149}$$

Let us summarize the formal structure: The fundamental relation is equivalent to the three equations of state (4). If only two equations of state are given then Gibbs-Duhem relation may be integrated to obtain the third relation up to an integration constant; alternatively one may integrate molar relation $de = Tds - Pdv$ to get $e(s, v)$, again with an undetermined constant of integration.

Example: Consider an ideal monatomic gas characterized by two equations of state (found, say, experimentally with $R \simeq 8.3\,\mathrm{J/mole\,K} \simeq 2\,\mathrm{cal/mole\,K}$):

$$PV = NRT \ , \qquad E = 3NRT/2 \ . \tag{150}$$

The extensive parameters here are E, V, N so we want to find the fundamental equation in the entropy representation, $S(E, V, N)$. We write (149) in the form

$$S = E\frac{1}{T} + V\frac{P}{T} - N\frac{\mu}{T} \ . \tag{151}$$

Here we need to express intensive variables $1/T, P/T, \mu/T$ via extensive variables. The equations of state (151) give us two of them:

$$\frac{P}{T} = \frac{NR}{V} = \frac{R}{v} \ , \qquad \frac{1}{T} = \frac{3NR}{2E} = \frac{3R}{2e} \ . \tag{152}$$

Now we need to find μ/T as a function of e, v using Gibbs-Duhem relation in the entropy representation (150). Using the expression of intensive via extensive variables in the equations of state (153), we compute $d(1/T) = -3Rde/2e^2$ and $d(P/T) = -Rdv/v^2$, and substitute into (150):

$$\begin{aligned} d\left(\frac{\mu}{T}\right) &= -\frac{3}{2}\frac{R}{e}de - \frac{R}{v}dv \, , \quad \frac{\mu}{T} = C - \frac{3R}{2}\ln e - R\ln v \, , \\ s &= \frac{1}{T}e + \frac{P}{T}v - \frac{\mu}{T} = s_0 + \frac{3R}{2}\ln\frac{e}{e_0} + R\ln\frac{v}{v_0} \ . \end{aligned} \tag{153}$$

Here we assumed that the system has the entropy s_0 in the state with the parameters e_0, v_0.

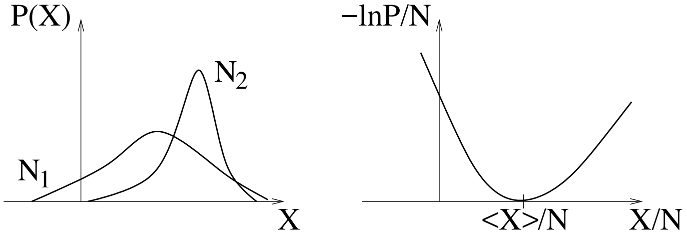
8.2 Central limit theorem and large deviations

The true logic of this world is to be found in the theory of probability.
Maxwell

A bridge from statistical physics to information theory is a simple technical tool used in both. Mathematics, underlying most of the statistical physics in the thermodynamic limit, comes from universality, which appears upon adding independent random numbers. The weakest statement is the law of large numbers: the sum approaches the mean value exponentially fast. The next level is the central limit theorem, which states that majority of fluctuations around the mean have Gaussian probability distribution. Consideration of large rare fluctuations requires the so-called large-deviation theory. Here we briefly present all three at the physical (not mathematical) level.

Consider the variable X which is a sum of many independent identically distributed (iid) random numbers $X=\sum_1^N y_i$. Its mean value $\langle X\rangle = N\langle y\rangle$ grows linearly with N. Here we show that its fluctuations $X-\langle X\rangle$ not exceeding $\mathcal{O}(N^{1/2})$ are governed by the Central Limit Theorem: $(X-\langle X\rangle)/N^{1/2}$ becomes for large N a Gaussian random variable with variance $\langle y^2\rangle - \langle y\rangle^2 \equiv \Delta$. The quantities y_i that we sum can have quite arbitrary statistics, the only requirements are that the first two moments, the mean $\langle y\rangle$ and the variance Δ, are finite. Finally, the fluctuations $X-\langle X\rangle$ on the larger scale $\mathcal{O}(N)$ are governed by the Large Deviation Theorem that states that the PDF of X has asymptotically the form

$$\mathcal{P}(X) \;\propto\; \mathrm{e}^{-NH(X/N)}\,. \tag{154}$$



To show this, we write

$$\begin{aligned}\mathcal{P}(X) &= \int \delta\left(\sum_{i=1}^N y_i - X\right)\mathcal{P}(y_1)dy_1\ldots\mathcal{P}(y_N)\,dy_N\\ &= \int_{-\infty}^{\infty} dp \int \exp\left[\imath p\left(\sum_{i=1}^N y_i - X\right)\right]\mathcal{P}(y_1)dy_1\ldots\mathcal{P}(y_N)\,dy_N\\ &= \int_{-\infty}^{\infty} dp e^{-\imath pX}\prod_{i=1}^N \int e^{\imath p y_i}\mathcal{P}(y_i)dy_i = \int_{-\infty}^{\infty} dp e^{-\imath pX+NG(\imath p)}\,.\end{aligned} \tag{155}$$

Here we introduced the generating function $\langle \mathrm{e}^{zy}\rangle \equiv \mathrm{e}^{G(z)}$. The derivatives of the generating function with respect to z at zero are equal to the moments of y, while the derivatives of its logarithm $G(z)$ are called cumulants (see exercise).

For large N, the integral (156) is dominated by the saddle point z_0 such that $G'(z_0)=X/N$. This is similar to representing the sum (11) above by its largest term. If there are several saddle-points, the result is dominated by the one giving the largest probability. We assume that contour of integration can be deformed in the complex plane z to pass through the saddle pint without hitting any singularity of $G(z)$. We now substitute $X=NG'(z_0)$ into $-zX+NG(z)$, and obtain the large deviation relation (155) with

$$H = -G(z_0) + z_0 G'(z_0)\ . \tag{156}$$

We see that $-H$ and G are related by the ubiquitous Legendre transform (which always appear in the saddle-point approximation of the Fourier or Laplace transformations). Note that $NdH/dX = z_0(X)$ and

$$N^2 d^2H/dX^2 = Ndz_0/dX = 1/G''(z_0)\,.$$

The function H of the variable $X/N - \langle y\rangle$ is called Cramér or rate function since it measures the rate of probability decay with the growth of N for every X/N. It is also sometimes called entropy function since it is a logarithm of probability.

Several important properties of H can be established independently of the distribution $\mathcal{P}(y)$ or $G(z)$. It is a convex function as long as $G(z)$ is a convex function since their second derivatives have the same sign. It is straightforward to see that the logarithm of the generating function has a positive second derivative (at least for real z):

$$\begin{aligned} G''(z) &= \frac{d^2}{dz^2}\ln\int e^{zy}\mathcal{P}(y)\,dy \\ &= \frac{\int y^2 e^{zy}\mathcal{P}(y)\,dy \int e^{zy}\mathcal{P}(y)\,dy - \left[\int y e^{zy}\mathcal{P}(y)\,dy\right]^2}{\left[\int e^{zy}\mathcal{P}(y)\,dy\right]^2} \geq 0\ . \qquad (157)\end{aligned}$$

This uses the Cauchy-Bunyakovsky-Schwarz inequality which is a generalization of $\langle y^2\rangle \geq \langle y\rangle^2$. Also, $H(z_0)$ takes its minimum at $z_0 = 0$, i.e. for X taking its mean value $\langle X\rangle = N\langle y\rangle = NG'(0)$. The maximum of probability does not necessarily coincides with the mean value, but they approach each other when N grows and maximum is getting very sharp — this is called the law of large numbers. Since $G(0) = 0$ then the minimal value of H is zero, that is the probability maximum saturates to a finite value when $N \to \infty$. Any smooth function is quadratic around its minimum with $H''(0) = \Delta^{-1}$, where $\Delta = G''(0)$ is the variance of y. Quadratic entropy means Gaussian probability near the maximum — this statement is (loosely speaking) the essence of the central limit theorem. In the particular case of Gaussian $\mathcal{P}(y)$, the PDF $\mathcal{P}(X)$ is Gaussian for any X. Non-Gaussianity of the y's leads to a non-quadratic behavior of H when deviations of X/N from the mean are large, of the order of $\Delta/G'''(0)$.

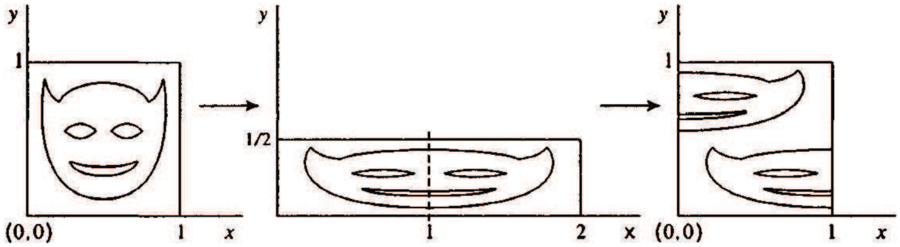
One can generalize the central limit theorem and the large-deviation approach in two directions: i) for non-identical variables y_i, as long as all variances are finite and none dominates the limit $N \to \infty$, it still works with the mean and the variance of X being given by the average of means and variances of y_i; ii) if y_i is correlated with a finite number of neighboring variables, one can group such "correlated sums" into new variables which can be considered independent.

The above figure and (155,157) show how distribution changes upon adding more numbers. Is there any distribution which does not change upon averaging, that is upon passing from y_i to $\sum_{i=1}^N y_i/N$? That can be achieved for $H \equiv 0$, that is for $G(z) = kz$, which corresponds to the Cauchy distribution $\mathcal{P}(y) \propto (y^2 + k^2)^{-1}$. Since the averaging decreases the variance, it is no surprise that the invariant distribution has infinite variance. Distributions invariant under summation of variables are treated in considering Renormalization Group in Section 4.3.

8.3 Baker map

Here we present a toy model, which is able to describe both mixing of area-preserving flows and fractalization of compressible flows. The phase-space ia a unit square in the (x, y)-plane, with $0 < x, y < 1$.

Area-preserving map and mixing. Consider first the area-preserving transformation, which is an expansion in the x-direction and a contraction in the y-direction, arranged in such a way that the unit square is mapped onto itself at each step. The transformation consists of two steps: First, the unit square is contracted in the y-direction and stretched in the x-direction by a factor of 2. This doesn't change the volume of any initial region. The unit square becomes a rectangle occupying the region $0 < x < 2$; $0 < y < 1/2$. Next, the rectangle is cut vertically in the middle and the right half is put on top of the left half to recover a square. This doesn't change volume either. That way bakers prepare long thin strips of pasta. This transformation is reversible except on the lines where the area was cut in two and glued back.



If we consider two initially closed points, then after n such steps the distance along x and y will be multiplied respectively by $2^n = e^{n \ln 2}$ and $2^{-n} = e^{-n \ln 2}$. It is then easy to see, without a lot of formalities, that there are two Lyapunov exponents corresponding to the discrete time n. One of them is connected to the expanding direction and has the value $\lambda_+ = \ln 2$. The other Lyapunov exponent is connected to the contracting direction and has the value $\lambda_- = -\ln 2$. For the forward time operation of the baker's transformation, the expanding direction is along the x-axis, and the contracting direction is along the y-axis. If one considers the time-reversed motion, the expanding and contracting directions change places. Therefore, for the forward motion nearby points separated only in the y-direction approach each other exponentially rapidly with the rate $\lambda_- = -\ln 2$. In the x-direction, points separate exponentially with $\lambda_+ = \ln 2$. The sum of the Lyapunov exponents is zero, which reflects the fact that the baker's transformation is area-preserving.

Let us argue now that the baker transformation is mixing, that is spreading

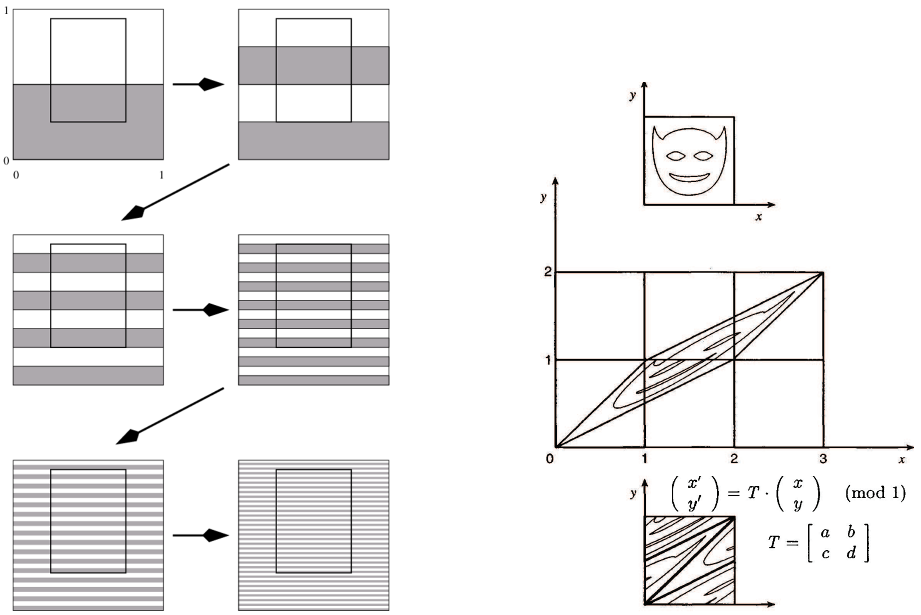

1
0
0
1
y
x
y
2
1
0
1
2
3
x
y
$\begin{pmatrix} x' \\ y' \end{pmatrix} = T \cdot \begin{pmatrix} x \\ y \end{pmatrix}$ (mod 1)
$T = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

Figure 6: Left panel: iterations of the baker map. Right panel: toral map.

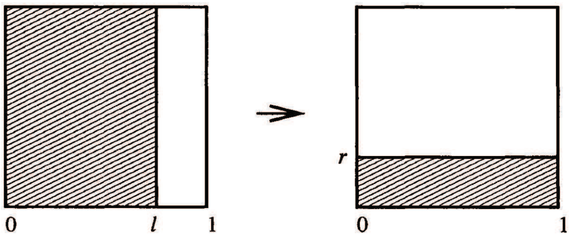
the measure uniformly over the whole phase space. Indeed, if a measure is initially concentrated in any domain, as in the grey area in the left panel of the Figure 6, after sufficiently long time the domain is transformed into a large number of very thin horizontal strips of length unity, distributed more and more uniformly in the vertical direction. Eventually any set in the unit square will have the same fraction of its area occupied by these little strips of pasta as any other set. This is the indicator of a mixing system. If we add to that however small coarse-graining, at sufficiently long time it blurs our measure to a constant one. We conclude that a sufficiently smooth initial distribution function defined on the unit square will approach a uniform (microcanonical) distribution.

The baker map is area-preserving and does not change entropy, yet when we allow for repeating coarse-graining along with the evolution, then the entropy grows and eventually reaches the maximum, which is the logarithm of the phase volume, which corresponds to the equilibrium microcanonical distribution.

One can encode any point on the square using the binary code of the expansion $x = \sum_{k=0}^{\infty} a_k 2^{-k-1}$, $y = \sum_{i=1}^{\infty} b_{-i} 2^{-i}$, where all a_k, b_{-i} are either 1 or 0. Simply, a_1 encodes in which half, a_2 encodes in which quarter within the half, etc. Encoding any point as a bi-infinite string, $(x, y) = \ldots b_{-3} b_{-2} b_{-1} . a_0 a_1 a_2 \ldots$, one finds out that the Baker map shifts the point . one step to the right. Using such so-called symbolic dynamics one can analyze mixing properties of maps.

To avoid impression that cutting and gluing of the baker map are necessary for mixing, consider a smooth model with a similar behavior. Namely, consider a unit two-dimensional torus, that is unit square with periodic boundary conditions, so that all distances are measured modulo 1 in the x- and y-direction. The action of such toral map is shown in the right panel of Figure 6. The transformation matrix T (an analog of the transfer matrix $\hat{W}$ from the Section 2.2) maps unit torus into itself if a, b, c, d are all integers. The eigenvalues $\lambda_{1,2} = (a+d)/2 \pm \sqrt{(a-d)^2/4 + bc}$ are real when $(a-d)^2/4 + bc \geq 0$, in particular, when the matrix is symmetric. For the transform to be area-preserving, the determinant of the matrix T, that is the product of the eigenvalues must be unity: $\lambda_1\lambda_2 = ad - bc = 1$. In a general case, one eigenvalue is larger than unity and one is smaller, which corresponds respectively to positive and negative Lyapunov exponents $\ln\lambda_1$ and $\ln\lambda_2$.

Compressible map and fractalization. To illustrate the entropy decay and measure fractalization in compressible flows, we consider a slight generalization of the baker map, expanding one region and contracting another, keeping the whole area unity:



The transformation has the form

$$x' = \begin{cases} x/l & \text{for } 0 < x < l \\ (x-l)/r & \text{for } l < x < 1\,, \end{cases}$$

$$y' = \begin{cases} ry & \text{for } 0 < x < l \\ r + ly & \text{for } l < x < 1 \end{cases}, \tag{158}$$

where $r + l = 1$. The Jacobian of the transformation is not identically equal to unity when $r \neq l$:

$$J = \left|\frac{\partial(x', y')}{\partial(x, y)}\right| = \begin{cases} r/l & \text{for } 0 < x < l \\ l/r & \text{for } l < x < 1\,, \end{cases}. \tag{159}$$

If $l > 1/2$, then $r = 1 - l < l$, so that $J < 1$ in the shadowed region where $x < L$, and $J > 1$ in the white region where $x > L$. Of course, the mean Jacobian $\overline{J} = r + l$ is unity, since we always occupy the same unit square. Like in the treatment of the incompressible baker map in the previous section, consider two initially closed points. If during n steps the points find themselves n_1 times in the region $0 < x < l$ and $n_2 = n - n_1$ times inside $l < x < 1$ then the distances along x

and y will be multiplied respectively by $l^{-n_1}r^{-n_2}$ and $r^{n_1}l^{n_2}$. Taking the log and the limit we obtain the Lyapunov exponents:

$$\lambda_+ = \lim_{n\to\infty} \tfrac{1}{n} \ln \tfrac{\delta x(n)}{\delta x(0)} = \lim_{n\to\infty} \left[\tfrac{n_1}{n}\ln\tfrac{1}{l} + \tfrac{n_2}{n}\ln\tfrac{1}{r}\right] = -l\ln l - r\ln r\,, \quad (160)$$

$$\lambda_- = \lim_{n\to\infty} \tfrac{1}{n} \ln \tfrac{\delta y(n)}{\delta y(0)} = \lim_{n\to\infty} \left[\tfrac{n_1}{n}\ln r + \tfrac{n_2}{n}\ln l\right] = r\ln l + l\ln r\,. \quad (161)$$

The sum of the Lyapunov exponents, $\lambda_+ + \lambda_- = (l-r)\ln(r/l) = \overline{\ln J}$, is non-positive and is zero only for $l = r = 1/2$. Again, convexity of the logarithmic function means that $\overline{\ln J} \le \ln \overline{J} = 0$. The volume contraction means that the expansion in the λ_+-direction proceeds slower than the contraction in the λ_--direction. After n iterations of the map, a square having initial side $\delta \ll L$ will be stretched into a long thin rectangle of length $\delta\exp(n\lambda_+)$ and width $\delta\exp(n\lambda_-)$. Asymptotically our strips of pasta concentrate on a fractal set, which is smooth in the x-direction and fractal in the y-direction, which respectively gives two terms in the non-integer dimensionality $d_f = 1 + \lambda_+/|\lambda_-|$, see (38).

The singularity of the non-equilibrium measure in the limit $\epsilon \to 0$ is probably related to non-analyticity of kinetic coefficients mentioned at the end of the Section 2.1.

Let us now use our model to derive the relation between the probabilities of entropy increase and decrease:

$$P^\dagger(-\Delta S) = P(\Delta S)e^{-\Delta S}\,. \quad (162)$$

Here $P^\dagger$ refers to a time-reversed process. At every step the volume contraction factor is the Jacobian of the transformation: $J = r/l$ for $x \in (0,l)$ and $J = l/r$ for $x \in (l,1)$. A long-time average rate of the entropy production, $\overline{\ln J} = (l-r)\ln(r/l)$, is the volume contraction rate of a fluid element. However, during a finite time n, there is always a finite probability to observe an expansion of an element. This probability must decay exponentially with n, and there is a universal law relating relative probabilities of the extraction and contraction. If during n steps a small rectangular element finds themselves n_1 times in the region $0 < x < l$ and $n_2 = n - n_1$ times inside $l < x < 1$ then its sides along x and y will be multiplied respectively by $l^{-n_1}r^{-n_2}$ and $r^{n_1}l^{n_2}$. The volume contraction factor for such n-sequence is $(l/r)^{n_2-n_1}$ and its log is $\Delta S = n\ln J = n_1\ln\frac{r}{l} + n_2\ln\frac{l}{r}$. The probability of the sequence is $P(\ln J) = l^{n_1}r^{n_2}$. Opposite sign of $\ln J$ will takes place, for instance, in a time-reversed sequence. Time reversal corresponds to the replacement $x \to 1-y, y \to 1-x$, that is the probability of such sequence is $P(-\ln J) = r^{n_1}l^{n_2}$. Therefore, denoting the entropy production rate $\sigma = -\ln J$, we obtain the universal probability independent of r, l:

$$\frac{P(\Delta S)}{P(-\Delta S)} = \frac{P(\sigma)}{P(-\sigma)} = \left(\frac{l}{r}\right)^{n_2-n_1} = e^{n\sigma} = e^{\Delta S}\,. \quad (163)$$

8.4 On Zipf law

Empirical law for the frequency of the r-th most frequent word in English is well approximated by the Zipf law:

$$p_r = \begin{cases} (10r)^{-1} & \text{for r=1,...,12367} \\ 0 & \text{for r >12367} \end{cases} \; . \tag{164}$$

Incidentally, if words were independent, then the entropy per word can be computed from (165) to be 9.7 bits. With the average word length 4.7 letters, that would give approximately 2 bits per letter, which compares well with the known value of 1.4 bits per letter.

Probably the simplest model that gives such a distribution is random typing: all letters plus the space are taken with equal probability (Wentian Li 1992). Then any word with the length L is flanked by two spaces and has the probability $P_i(L) = (M+1)^{-L-2}/Z$, where $i = 1, 2, \ldots, M^L$ and M is the alphabet size. The normalization factor is $Z = \sum_L M^L(M+1)^{-L-2} = (M+1)^2/M$. On the other hand, the rank $r(L)$ of any L-word satisfies the inequality

$$M(M^{L-1}-1)/(M-1) = \sum_{i=1}^{L-1} M^i < r(L) \le \sum_{i=1}^{L} M^i = M(M^L-1)/(M-1)\,,$$

which can be written as $P_i(L) < C[r(L)+B]^{-\alpha} \le P_i(L-1)$ with $\alpha = \log_M(M+1)$, $B = M/(M-1)$ and $C = B^\alpha/M$. In the limit of large alphabet size, $M \gg 1$, we obtain

$$P(r) = (r+1)^{-1}\ . \tag{165}$$

This asymptotic actually takes place for wide classes of letter distributions, not necessarily equiprobable. Closely related way of *interpreting* statistical distributions is to look for variational principle it satisfies. What we mostly did in this course and what most of statisticians do most of the time is looking for a conditional entropy maximum and minimizing a two-term functional. In this case, one may require maximal information transferred with the least effort. The rate of information transfer is $S = -\sum_r P(r)\log P(r)$. The effort must be higher for less common words, that is to grow with the rank. Such growth can be logarithmic (for instance, when the effort is proportional to the word length). The mean effort is then $W = \sum_r P(r)\log r$. Looking for the minimum of $S - \lambda W$, we obtain $P(r) \propto r^{-\lambda}$. Zipf law corresponds to $\lambda = 1$, when goals and efforts are balanced.

Does then the Zipf law trivially appear because both number of words (inverse probability) and rank increase exponentially with the word size? The answer is negative because the number of distinct words of the same length in real language is not exponential in length and is not even monotonic. It is reassuring that our texts are statistically distinguishable from those produced by an imaginary monkey with a typewriter. Moreover, words have meaning. The number of meanings (counted from the number of dictionary entries for a word) grows approximately as the square root of the word frequency: $m_i \propto \sqrt{P_i}$. Meanings correspond to objects of reference having their own probabilities, and it seems that the language combines these objects into groups whose sizes are proportional to the mean probability of the group p_i, so that $P_i = m_i p_i \propto m_i^2$. It is tempting to suggest that the distributions appeared due to the balance between minimizing efforts of writers and readers, speakers and listeners. Writers and speakers would minimize their effort by having one word meaning everything and appearing with the probability one. On the other end, difficulty of perception is proportional to the depth of the memory keeping the context, needed, in particular, for choosing the right meaning. Readers and listeners then prefer a lot of single-meaning words. So far, no convincing optimization scheme giving different features of word statistics was found[25].

8.5 Ising model of the brain

As was mentioned at the end of Section 4.1, the activity of a network of neurons is described by one-bit variables $\sigma_i = \pm 1$ (active or inactive). The probability distribution that maximizes entropy under the given set of the mean activities $\langle \sigma_i \rangle$ and the pairwise correlations $\langle \sigma_i \sigma_j \rangle$ is that of an Ising model (73):

$$\rho(\{\sigma\}) = Z^{-1} \exp\left[\sum_i h_i \sigma_i + \frac{1}{2}\sum_{i<j} J_{ij}\sigma_i\sigma_j\right] . \tag{166}$$

One can also measure some multi-cell correlations and check how well they agree with those computed from (167). Despite apparent patterns of collective behavior, that involve many neurons, it turns out to be enough to account for pairwise correlations to describe the statistical distribution remarkably well (Schneidman, Berry, Segev and Bialek, 2006). This is also manifested by the entropy changes: measuring triple and multi-cell correlations imposes more restrictions and lowers the entropy maximum. One then checks that accounting for pairwise correlations changes entropy significantly while account for further correlation changes entropy relatively little. The sufficiency of pairwise interactions provides an enormous simplification, which may be important not only for our description, but also for the brain. The reason is that our mind actually develops and constantly

[25] Maybe, some statistical laws of language can be better understood viewing conversations which are not exchanges of information but rather mating games with multiple synonyms and meanings as a verbal plumage.

modifies its own predictive model of probability needed in particular to accurately evaluate new events for their degree of surprise, as described in Section 4.5. The dominance of pairwise interactions means that learning rules based on pairwise correlations could be sufficient to generate nearly optimal internal models to accurately evaluate probabilities. Side remark: we should not think that what is encoded from sensors into an electrical neuron activity is then "decoded" inside the brain. Whatever it is, brain is not computer.

It is interesting how the entropy scales with the number of interacting neurons N. The entropy of non-interacting (or nearest-neighbor interacting) neurons is extensive that is proportional to N. The data show that J_{ij} are non-zero for distant neurons as well. That means that the entropy of an interacting set is lower at least by the sum of the mutual information terms between all pairs of cells. The negative contribution is thus proportional to the number of interacting pairs, $N(N-1)/2$, that is grows faster with N, at least when it is not too large. One can estimate from low-N data a "critical" N, when the quadratic term is expected to turn entropy into zero. That critical N corresponds well to the empirically observed sizes of the clusters of strongly correlated cells. The lesson is: even when pairwise correlations are weak, sufficiently large clusters can be strongly correlated. It is also important that the interactions J_{ij} have different signs, so that frustration can prevent the freezing of the system into a single state (like ferromagnetic or anti-ferromagnetic). Instead there are multiple states that are local minima of the effective energy function, as in spin glasses.

8.6 Unsupervised learning and infomax principle

The maximal-capacity approach described in Section 4.4 turns out quite useful in image and speech recognition by iterative algorithms. One chooses some form of the response function $y = g(x, w)$ characterized by the parameter w and find optimal value of w using an "online" stochastic gradient ascent learning rule giving the change of the parameter:

$$\Delta w \propto \frac{\partial}{\partial w} \ln \left(\frac{\partial g(x, w)}{\partial x} \right) . \tag{167}$$

Of course, eye or camera provide not a single input signal, but the whole picture. Let us consider N inputs and outputs (neurons/channels). Consider a network with an input vector $\mathbf{x} = (x_1, \dots, x_N)$ which is transformed into the output vector $\mathbf{y}(\mathbf{x})$ one to one, that is $\det[\partial y_i/\partial x_k] \neq 0$. The multivariate probability density function of y is as follows:

$$\rho(\mathbf{y}) = \frac{\rho(\mathbf{x})}{\det[\partial y_i/\partial x_k]} , \tag{168}$$

Making it flat (distribute outputs uniformly) for maximal capacity is not straightforward now. In one dimension, it is enough to follow the gradient to arrive at an extremum, but there are many possible paths to the mountain summit. Maximizing the total mutual information between input and output, which requires maximizing the output entropy, is often (but not always) achieved by minimizing first the mutual information between the output components. For two outputs we may start by maximizing $S(y_1, y_2) = S(y_1) + S(y_2) - I(y_1, y_2)$, that is minimize $I(y_1, y_2)$. If we are lucky and find encoding in terms of independent components, then we choose for each component the transformation (83), which maximizes its entropy making the respective probability flat. For a review and specific applications to visual sensory processing, see Atick 1992.

Finding least correlated components can be a practical first step in maximizing capacity. To *maximize* the mutual information between input and output, we *minimize* the mutual information between the components of the output. This is particularly useful for natural signals where most redundancy comes from strong correlations (like that of the neighboring pixels in visuals). Also, finding an encoding in terms of least dependent components is important by itself for its cognitive advantages. For example, such encoding generally facilitates pattern recognition. In addition, presenting and storing information in the form of independent (or minimally dependent) components is important for associative learning done by brains and computers. Indeed, for an animal or computer to learn a new association between two events, A and B, the brain should have knowledge of the prior joint probability $P(A, B)$. For correlated N-dimensional A and B one needs to store $N \times N$ numbers, while only $2N$ numbers for quantities uncorrelated (until the association occurs).

Another cognitive task is the famous ”cocktail-party problem” posed by security services: N microphones (flies on the wall) record N people speaking simultaneously, and we need the program to separate them — so-called *blind separation* problem. Here we assume that uncorrelated sources $s_1, \dots, s_N$ are mixed linearly by an unknown matrix $\hat{A}$. All we receive are the N superpositions of them $x_1, \dots, x_N$. The task is to recover the original sources by finding a square matrix $\hat{W}$ which is the inverse of the unknown $\hat{A}$, up to permutations and re-scaling. Closely related is the *blind de-convolution* problem (see e.g. Bell and Sejnowski, 1995): a single unknown signal $s(t)$ is convolved with an unknown filter giving a corrupted signal $x(t) = \int a(t - t')s(t')\, dt'$, where $a(t)$ is the impulse response of the filter. The task is to recover $s(t)$ by integrating $x(t)$ with the inverse filter $w(t)$, which we need to find by learning procedure. Upon discretization, s, x are turned into N-vectors and w into $N \times N$ matrix, which is lower triangular because of causality: $w_{ij} = 0$ for $j > i$ and the diagonal values are all the same $w_{ii} = \bar{w}$. The determinant in (169) is simplified in this case. For $\mathbf{y} = g(\hat{w}\mathbf{x})$ we have $\det[\partial y(t_i)/\partial x(t_j)] = \det \hat{w} \prod_i^N y'(t_i) = \bar{w}^N \prod_i^N y'(t_i)$. One then applies some variant of (168) to minimize mutual information.

Ideally, we wish to find the (generally stochastic) encoding $\mathbf{y}(\mathbf{x})$ that achieves the absolute minimum of the mutual information $\sum_i S(y_i) - S(\mathbf{y})$. One way to do that is to minimize the first term while keeping the second one, that is under condition of the fixed entropy $S(\mathbf{y}) = S(\mathbf{x})$. In general, one may not be able to find such encoding without any entropy change $S(\mathbf{y}) - S(\mathbf{x})$. In such cases, one defines a functional that grades different codings according to how well they minimize *both* the sum of the entropies of the output components and the entropy change. The simplest energy functional for statistical independence is then

$$E = \sum_i S(y_i) - \beta[S(\mathbf{y}) - S(\mathbf{x})] = \sum_i S(y_i) - \beta \ln \det[\partial y_i/\partial x_k] \ . \qquad (169)$$

A coding is considered to yield an improved representation if it possesses a smaller value of E. The choice of the parameter β reflects our priorities — whether statistical independence or increase in indeterminacy is more important.

Maximizing information transfer and reducing the redundancy between the units in the output is applied practically in all disciplines that analyze and process data, from physics and engineering to biology, psychology and economics. Within the general infomax domain, this specific technique is called independent component analysis. More sophisticated schemes employs not only mutual information, but also interaction information (68). Note that the redundancy reduction is usually applied after some procedure of eliminating noise. Indeed, our gain function provides equal responses for probable and improbable events, but the latter can be mostly due to noise, which thus needs to be suppressed. Moreover, if input noises were uncorrelated, they can get correlated after coding. And more generally, it is better to keep some redundancy for corrections and checks when dealing with noisy data.

8.7 Information loss by multi-dimensional RG

It seems reasonable to expect irreversibility of renormalization group since it is a way of forgetting. Yet it is far from trivial to find an entropic characteristics which changes monotonically upon RG in a multi-dimensional space. Eliminating modes step by step generally decreases the mutual information between degrees of freedom I. However, re-scaling and renormalization may increase it, because some of the information about eliminated degrees of freedom is stored in the renormalized values of the parameters of the distribution. Increase or decrease of I upon RG thus shows whether the large-scale behavior is respectively ordered or disordered.

In Section 4.3 we characterized information exchange in one dimension looking at a single bond which separates two parts of a spin chain. Breaking a single bond in more than one dimension does not separate. In 2d plane, one can consider a (finite) line L and break the direct interactions between the degrees of freedom on the different sides of it. That is we make a cut and ascribe to its every point two (generally different) values on the opposite sides. The statistics of such set is now characterized not by a scalar function - probability on the line - but by a matrix, similar to the density matrix in quantum statistics, described in Section 6.2: One takes the whole quantum system in a ground state, traces out (average over) all the degrees of freedom outside the line and obtain the density matrix ρ_L of the degrees of freedom on the line. For that density matrix one defines von Neumann entropy $S_L = -\mathrm{Tr}\rho_L \log \rho_L$

For long lines in short-correlated systems, that quantity can be shown to depend only on the distance r between the end points (and not on the form of a line connecting them, that is information flows like an incompressible fluid). Moreover, this dependence is logarithmic at criticality (when we have fluctuations of all scales and the correlation radius is infinite). To cancel non-universal terms depending on the microscopic detail, one defines the function $c(r) = r dS_L(r)/dr$. which is shown to be a monotonic zero degree function, using positivity of the mutual information (sub-additivity of the entropy) between lines with r and $r + dr$ (Zamolodchikov 1986, Casini and Huerta 2006). The same function changes monotonically under RG flow and in a fixed point takes a finite value equal to the so-called zero charge of the respective conformal field theory. The zero charge is a measure of relevant degrees of freedom that respond to boundary perturbations. It is even more difficult to introduce proper intensive measure of information flow in dimensions higher than two, so far it is done in a quite model-specific way (see e.g. Komargodsky and Schwimmer 2011, Klebanov 2011) .

In looking for fundamental characteristics of order in fluctuating systems in higher dimensions, one can go even deeper. For instance, one can consider for quantum system in 2+1 dimensions the relative entanglement of three finite planar regions, A, B, C, all having common boundaries. As a quantum analog of the interaction information (68), one can introduce so-called topological entanglement entropy $S_A + S_B + S_C + S_{ABC} - S_{AB} - S_{BC} - S_{AC}$. For some classes of systems, one can show that in the combination, the terms depending on the boundary lengthes cancel out and what remains (if any) can be thus independent of the deformations of the boundaries, that is characterizing the topological order, if it exists in the system (Kitaev, Preskill 2006).

8.8 Brownian motion

We consider the motion of a small particle in a fluid. The momentum of the particle, $\mathbf{p} = M\mathbf{v}$, changes because of collisions with the molecules. Thermal equipartition guarantees that the mean kinetic energy of the particle is the same as of any molecule and equal to $T/2$. When the particle M is much larger than the molecular mass m then the rms particle velocity $v = \sqrt{T/M}$ is small comparing to the typical velocities of the molecules $v_T = \sqrt{T/m}$. That allows one to write the force $\mathbf{f}(\mathbf{p})$ acting on the particle as Taylor expansion in $\mathbf{p}$, keeping the first two terms, independent of $\mathbf{p}$ and linear in $\mathbf{p}$: $f_i(\mathbf{p}, t) = f_i(0, t) + p_j(t)\partial f_i(0, t)/\partial p_j(t)$ (note that we neglected the dependence of the force of the momentum at earlier times). Such expansion makes sense if the third term is much less than the second one, but then the second term must be much smaller than the first one — what is the reason to keep both? The answer is that molecules hitting standing particle produce force whose average is zero. The mean momentum of the particle is zero as well. However, random force by itself would make the squared momentum grow with time exactly like the squared displacement of a random walker in the previous section. To describe the particle in equilibrium with the medium, the force must be balanced by resistance which is also provided by the medium: the particle meets more molecules in the direction it moves and looses its momentum to them. That resistance has non-zero mean and must be described by the second term, which then may be approximated as $\partial f_i/\partial p_j = -\gamma\delta_{ij}$. If the particle radius R is larger than the mean free path ℓ, in calculating resistance, we can consider fluid as a continuous medium and characterize it by the viscosity η. For a slow moving particle, $v \ll v_T\ell/R$, the resistance is given by the Stokes formula

$$\gamma = 6\pi\eta R/M \ . \tag{170}$$

We then obtain

$$\dot{\mathbf{p}} = \mathbf{f} - \gamma\mathbf{p} \ . \tag{171}$$

The solution of the linear equation (172) is

$$\mathbf{p}(t) = \int_{-\infty}^{t} \mathbf{f}(t')e^{\gamma(t'-t)}dt' \ . \tag{172}$$

We must treat the force $\mathbf{f}(t)$ as a random function since we do not track molecules hitting the particle. We assume that $\langle\mathbf{f}\rangle = 0$ and that $\langle\mathbf{f}(t')\cdot\mathbf{f}(t'+t)\rangle = 3C(t)$ decays with t during the correlation time τ which is much smaller than γ^{-1}. Since the integration time in (173) is of order γ^{-1}, then the condition $\gamma\tau \ll 1$ means that the momentum of a Brownian particle can be considered as a sum of many independent random numbers (integrals over intervals of order τ) and so it must have a Gaussian statistics $\rho(\mathbf{p}) = (2\pi\sigma^2)^{-3/2}\exp(-p^2/2\sigma^2)$ where

$$\begin{aligned}\sigma^2 &= \langle p_x^2\rangle = \langle p_y^2\rangle = \langle p_z^2\rangle = \int_0^\infty C(t_1 - t_2)e^{-\gamma(t_1+t_2)}dt_1dt_2 \\ &\approx \int_0^\infty e^{-2\gamma t}\,dt \int_{-2t}^{2t} C(t')\,dt' \approx \frac{1}{2\gamma}\int_{-\infty}^{\infty} C(t')\,dt' \ . \end{aligned} \tag{173}$$

On the other hand, equipartition guarantees that $\langle p_x^2\rangle = MT$ so that we can express the friction coefficient via the correlation function of the force fluctuations (a particular case of the fluctuation-dissipation theorem, which follows from the detailed balance):

$$\gamma = \frac{1}{2TM}\int_{-\infty}^{\infty} C(t')\,dt' \ . \tag{174}$$

Displacement

$$\Delta\mathbf{r}(t') = \mathbf{r}(t+t') - \mathbf{r}(t) = \int_0^{t'} \mathbf{v}(t'')\,dt''$$

is also Gaussian with a zero mean. To get its second moment we need the different-time correlation function of the velocities

$$\langle \mathbf{v}(t)\cdot\mathbf{v}(0)\rangle = (3T/M)\exp(-\gamma|t|) \tag{175}$$

which can be obtained from (173). Note that the friction makes velocity correlated on a longer timescale than the force. That gives

$$\langle|\Delta\mathbf{r}|^2(t')\rangle = \int_0^{t'} dt_1 \int_0^{t'} dt_2 \langle\mathbf{v}(t_1)\mathbf{v}(t_2)\rangle = \frac{6T}{M\gamma^2}(\gamma t' + e^{-\gamma t'} - 1) \ .$$

The mean squared distance initially grows quadratically (so-called ballistic regime at $\gamma t' \ll 1$). In the limit of a long time (comparing to the relaxation time γ^{-1} rather than to the force correlation time τ) we have the diffusive growth $\langle(\Delta\mathbf{r})^2\rangle \approx 6Tt'/M\gamma$. Generally $\langle(\Delta\mathbf{r})^2\rangle = 6\kappa t$ where κ is the diffusivity, which thus satisfies the Einstein relation:

$$\kappa = \frac{T}{M\gamma} = \frac{T}{6\pi\eta R} \ . \tag{176}$$

The diffusivity depends on the particle radius but not the mass. Heavier particles are slower both to start and to stop moving. Measuring diffusion of particles with a known size one can determine the temperature[26].

The probability distribution of displacement at $\gamma t' \gg 1$,

$$\rho(\Delta\mathbf{r}, t') = (4\pi\kappa t')^{-3/2}\exp[-|\Delta\mathbf{r}|^2/4\kappa t'] \,,$$

satisfies the diffusion equation $\partial\rho/\partial t' = \kappa\nabla^2\rho$. If we have many particles initially distributed according to $n(\mathbf{r},0)$ then their distribution $n(\mathbf{r},t) = \int \rho(\mathbf{r}-\mathbf{r}',t)n(\mathbf{r}',0)\,d\mathbf{r}'$, also satisfies the diffusion equation: $\partial n/\partial t' = \kappa\nabla^2 n$.

[26] With temperature in degrees, (177) contains the Boltzmann constant, $k = \kappa M\gamma/T$, which was actually determined by this relation and found constant indeed, i.e. independent of the medium and the type of particle. That proved the reality of atoms - after all, kT is the kinetic energy of a single atom.

An external field $V(\mathbf{q})$ adds the force:

$$\dot{\mathbf{p}} = -\gamma\mathbf{p} + \mathbf{f} - \partial_q V\,, \quad \dot{\mathbf{q}} = \mathbf{p}/M\ . \tag{177}$$

These equations characterize the system with the Hamiltonian $\mathcal{H} = p^2/2M + V(\mathbf{q})$. The system interacts with the thermostat, which provides friction $-\gamma\mathbf{p}$ and agitation $\mathbf{f}$ - the balance between these two terms expressed by (175) means that the thermostat is in equilibrium.

We now pass from considering individual trajectories to the description of the "cloud" of trajectories and its statistics. Remind that our particle is macroscopic, that is we consider the so-called over-damped limit $\gamma\tau \gg 1$, where τ is the random force correlation time. Since we shall not be interested in small irregular changes of the velocity, but only in the statistics of displacement, we average (coarse-grain) over moving time window, $p(t) \to p(t) = \int_{t-\tau}^{t+\tau} p(t')dt'$. After the average, we can neglect acceleration. In this limit our second-order equation (178) on $\mathbf{q}$ is reduced to the first-order equation (we keep the same notations for coarse-grained quantities):

$$\gamma\mathbf{p} = \gamma M\dot{\mathbf{q}} = \mathbf{f} - \partial_q V\ . \tag{178}$$

We can now derive the equation on the probability distribution $\rho(\mathbf{q}, t)$. which changes with time due to random noise and evolution in the potential, the two mechanisms can be considered additively. Together, diffusion and advection give the Fokker-Planck equation, which is a multi-dimensional generalization of (110):

$$\frac{\partial\rho}{\partial t} = \kappa\nabla^2\rho + \frac{1}{\gamma M}\frac{\partial}{\partial q_i}\rho\frac{\partial V}{\partial q_i} = -\mathrm{div}\,\mathbf{J}\ . \tag{179}$$

More formally, one can derive (180) by writing (179) as $\dot{q}_i - w_i = \eta_i$ and taking the random force Gaussian delta-correlated: $\langle\eta_i(0)\eta_j(t)\rangle = 2\kappa\delta_{ij}\delta(t)$. One can write the conditional probability $\rho(\mathbf{q}, t; 0, 0)$ as an average over all possible paths each with its own weight determined by the Gaussian statistics of η_i. Such path integral representation is presented in Section 8.10 below. Since it is the quantity $\dot{\mathbf{q}} - \mathbf{w}$ which is Gaussian now, then(191) changes into

$$\rho(\mathbf{q}, t; 0, 0) \;\; = \;\; \int \mathcal{D}\mathbf{q}(t') \exp\left[-\frac{1}{4\kappa}\int_0^t dt' |\dot{\mathbf{q}} - \mathbf{w}|^2\right]\ , \tag{180}$$

To describe the time change, consider the convolution identity (190) for an infinitesimal time shift ϵ, then instead of the path integral we get simply the integral over the initial position $\mathbf{q}'$. We substitute $\dot{\mathbf{q}} = (\mathbf{q} - \mathbf{q}')/\epsilon$ into (181) and obtain

$$\rho(\mathbf{q}, t) = \int d\mathbf{q}' (4\pi\kappa\epsilon)^{-d/2} \exp\left[-\frac{[\mathbf{q} - \mathbf{q}' - \epsilon\mathbf{w}(\mathbf{q}')]^2}{4\kappa\epsilon}\right]\rho(\mathbf{q}', t - \epsilon)\ . \tag{181}$$

What is written here is simply that the transition probability is the Gaussian probability of finding the noise η with the right magnitude to provide for the transition from $\mathbf{q}'$ to $\mathbf{q}$. It is a coarse-grained continuous version of (106). We now change integration variable, $\mathbf{y} = \mathbf{q}' + \epsilon \mathbf{w}(\mathbf{q}') - \mathbf{q}$, and keep only the first term in ϵ: $d\mathbf{q}' = d\mathbf{y}[1 - \epsilon \partial_{\mathbf{q}} \cdot \mathbf{w}(\mathbf{q})]$. Here $\partial_{\mathbf{q}} \cdot \mathbf{w} = \partial_i w_i = div\, \mathbf{w}$. In the resulting expression, we expand the last factor $\rho(\mathbf{q}', t - \epsilon)$:

$$\begin{aligned}
\rho(\mathbf{q}, t) &\approx (1 - \epsilon \partial_{\mathbf{q}} \cdot \mathbf{w}) \int d\mathbf{y} (4\pi\kappa\epsilon)^{-d/2} e^{-y^2/4\kappa\epsilon} \rho(\mathbf{q} + \mathbf{y} - \epsilon \mathbf{w}, t - \epsilon) \\
&\approx (1 - \epsilon \partial_{\mathbf{q}} \cdot \mathbf{w}) \int d\mathbf{y} (4\pi\kappa\epsilon)^{-d/2} e^{-y^2/4\kappa\epsilon} \Big[\rho(\mathbf{q}, t) + (\mathbf{y} - \epsilon \mathbf{w}) \cdot \partial_{\mathbf{q}} \rho(\mathbf{q}, t) \\
&+ (y_i y_j - 2\epsilon y_i w_j + \epsilon^2 w_i w_j) \partial_i \partial_j \rho(\mathbf{q}, t)/2 - \epsilon \partial_t \rho(\mathbf{q}, t) \Big] \\
&= (1 - \epsilon \partial_{\mathbf{q}} \cdot \mathbf{w})[\rho - \epsilon \mathbf{w} \cdot \partial_{\mathbf{q}} \rho + \epsilon \kappa \Delta \rho - \epsilon \partial_t \rho + O(\epsilon^2)] \ , \qquad (182)
\end{aligned}$$

and obtain (180) collecting terms linear in ϵ. Note that it was necessary to expand until the quadratic terms in y, which gave the contribution linear in ϵ, namely the Laplacian, i.e. the diffusion operator.

The Fokker-Planck equation has a stationary zero-current Boltzmann-Gibbs solution which corresponds to the particle in an external field and in thermal equilibrium with the surrounding molecules:

$$\rho(\mathbf{q}) \propto \exp[-V(\mathbf{q})/\gamma M \kappa] = \exp[-V(\mathbf{q})/T] \ . \qquad (183)$$

From the perspective of the Information Theory, if the only thing we know is that a particle at $\mathbf{q}$ has the mean energy $V(\mathbf{q})$ then the probability distribution is an exponent of the energy.

8.9 Fluctuation relations in a multi-dimensional case

Apart from making the potential time-dependent, there is another way to deviate the system from equilibrium in more than one dimension: to add to the random thermal force $\mathbf{f}(t)$ and the potential force $-\partial_{\mathbf{q}} V(\mathbf{q})$ another external coordinate-dependent force $\mathbf{F}(\mathbf{q})$ which is non-potential (not a gradient of any scalar):

$$\dot{\mathbf{p}} = -\gamma \mathbf{p} + \mathbf{f} - \partial_{\mathbf{q}} V + \mathbf{F} \, , \quad \dot{\mathbf{q}} = p/M \ .$$

The non-potential force makes the system non-Hamiltonian even without any contact with a thermostat, that is when $\gamma = 0$ and $\mathbf{f} = 0$. Bringing such a system into a contact with a thermostat generally does not lead to thermal equilibrium, as we discussed in Section 2.3. The equation on the full phase-space distribution $\rho(\mathbf{p}, \mathbf{q}, t)$ has the form

$$\partial_t \rho = \{\mathcal{H}, \rho\} + T \Delta_p \rho + \partial_{\mathbf{p}} \rho [\mathbf{F} - \gamma \mathbf{p}] = H_K \rho \, . \qquad (184)$$

It is called the Kramers equation. Fokker-Planck equation follows from it in the overdamped limit. Only without $\mathbf{F}$, the Gibbs distribution $\exp(-\mathcal{H}/T)$ is a steady solution of (185) and one can formulate the detailed balance

$$H_K^\dagger = \Pi e^{\beta\mathcal{H}} H_K e^{-\beta\mathcal{H}} \Pi^{-1} \,, \tag{185}$$

where we added the operator inverting momenta: $\Pi \mathbf{p} \Pi^{-1} = -\mathbf{p}$. A non-potential force violates the detailed balance in the following way:

$$H_K^\dagger = \Pi e^{\beta\mathcal{H}} H_K e^{-\beta\mathcal{H}} \Pi^{-1} + \beta(\mathbf{F} \cdot \dot{\mathbf{q}}) \ . \tag{186}$$

The last (symmetry-breaking) term is again the power $(\mathbf{F} \cdot \dot{\mathbf{q}})$ divided by temperature i.e. the entropy production rate. The work done by that force depends on the trajectory in distinction from the case of a time-independent potential force. That dependence of the work on the trajectory precludes thermal equilibrium and is common for non-potential forces and for time-dependent potential forces. A close analog of the Jarzynski relation can be formulated for the production rate averaged during the time t:

$$\sigma_t = \frac{1}{tT} \int_0^t (\mathbf{F} \cdot \dot{\mathbf{q}})\, dt \ . \tag{187}$$

The power $(\mathbf{F} \cdot \dot{\mathbf{q}})$ is identically zero for a magnetic Lorentz force, which is perpendicular to the velocity. For a potential force, $\mathbf{F} = dU/d\mathbf{q}$, we have $(\mathbf{F} \cdot \dot{\mathbf{q}}) = dU(\mathbf{q}(t))/dt$, and the integral turns into zero on average. Non-potential external force $\mathbf{F}$ must on average do a positive work to keep the system away from equilibrium. Over a long time we thus expect σ_t to be overwhelmingly positive, yet fluctuations do happen. The probabilities $P(\sigma_t)$ satisfy the relation, analogous to (121), which we give without general derivation

$$\frac{P(\sigma_t)}{P(-\sigma_t)} \propto e^{t\sigma_t} \ . \tag{188}$$

This relation shows how low is the probability to observe a negative entropy production rate - this probability decays exponentially with the time of observation. Such fluctuations were unobservable in classical macroscopic thermodynamics, but they are often very important in modern applications to nano and bio objects. In the limit $t \to \infty$, when the probability of the integral (188) must have a large-deviation form, $P(\sigma_t) \propto \exp[-tH(\sigma_t)]$, so that (189) means that $H(\sigma_t) - H(-\sigma_t) = -\sigma_t$, as if $P(\sigma_t)$ was Gaussian with $H(\sigma_t) = (\sigma_t - 1)^2/2$.

One calls (121,189) detailed fluctuation-dissipation relations since they are stronger than integral relations of the type (118,119). Indeed, it is straightforward to derive $\langle \exp(-t\sigma_t) \rangle = 1$ from (189).

The relation similar to (189) can be derived for any system symmetric with respect to some transformation, to which we add perturbation anti-symmetric with respect to that transformation. Consider a system with the variables $s_1, \ldots, s_N$ and the even energy: $E_0(\mathbf{s}) = E_0(-\mathbf{s})$. Consider the energy perturbed by an odd term, $E = E_0 - hM/2$, where $M(\mathbf{s}) = \sum s_i = -M(-\mathbf{s})$. The probability of the perturbation $P[M(\mathbf{s})]$ satisfies the direct analog of (189), which is obtained by changing the integration variable $\mathbf{s} \to -\mathbf{s}$:

$$P(a) = \int d\mathbf{s}\delta[M(\mathbf{s}) - a]e^{\beta(ha-E_0)} = \int d\mathbf{s}\delta[M(\mathbf{s}) + a]e^{-\beta(ha+E_0)} = P(-a)e^{-2\beta ha} \ .$$

The validity condition for the results in this Section is that the interaction with the thermostat is represented by noise independent of the the evolution of the degrees of freedom under consideration.

8.10 Quantum fluctuations and thermal noise

Many aspects of quantum world are bizarre and have no classical analog. And yet there are certain technical similarities between descriptions of quantum and thermal fluctuations due to the necessity of summing over different possibilities. One analogy was already exploited in Section 5.3, where Schrodinger equation was treated as a diffusion equation with an imaginary diffusivity. One can also treat propagation of a quantum particle as a random walk in an imaginary time using the formalism of the path integral, where one sums over different trajectories. Let us write the transition probability for an unbiased random walk indicating explicitly the origin: $\rho(\mathbf{x}, t; 0, 0)$. It is the probability to come after time t to x conditional on starting at 0. We can write the convolution identity which simply states that the walker was certainly somewhere at an intermediate time t_1:

$$\rho(\mathbf{x}, t; 0, 0) = \int \rho(\mathbf{x}, t; \mathbf{x}_1, t_1)\rho(\mathbf{x}_1, t_1; 0, 0)\, d\mathbf{x}_1 \ . \tag{189}$$

We now divide the time interval t into an arbitrary large number of intervals and using (109) we write

$$\begin{aligned}\rho(\mathbf{x}, t; 0, 0) &= \int \Pi_{i=0}^{n} \frac{d\mathbf{x}_{i+1}}{[4\pi\kappa(t_{i+1} - t_i)]^{d/2}} \exp\left[-\frac{(\mathbf{x}_{i+1} - \mathbf{x}_i)^2}{4\kappa(t_{i+1} - t_i)}\right] \\ &\to \int \mathcal{D}\mathbf{x}(t') \exp\left[-\frac{1}{4\kappa}\int_0^t dt' \dot{x}^2(t')\right] \ . \end{aligned} \tag{190}$$

The last expression is an integral over paths that start at zero and end up at $\mathbf{x}$ at t. Notation $\mathcal{D}\mathbf{x}(t')$ implies integration over the positions at intermediate times normalized by square roots of the time differences. The exponential gives the weight of every trajectory.

Looking at the transition probability (191), one can see the analogy between the statistical mechanics of a random walk and quantum mechanics. According to Feynman, for a quantum non-relativistic particle with a mass M, the transition amplitude $T(\mathbf{x},t;0,0)$ from zero to $\mathbf{x}$ during t is given by the sum over all possible paths connecting the points. Every path is weighted by the factor $\exp(iS/\hbar)$ where S is the classical action:

$$T(\mathbf{x},t;0,0) \quad = \quad \int \mathcal{D}\mathbf{x}(t') \exp\left[\frac{i}{\hbar}\int_0^t dt' \frac{M\dot{x}^2}{2}\right] \ . \tag{191}$$

Comparing with (191), we see that the transition probability of a random walk is given by the transition amplitude of a free quantum particle during an imaginary time. In quantum theory, one averages over quantum rather than thermal fluctuations, yet the formal structure of the theory is similar.

Another similarity is revealed using the Heisenberg representation of the operators evolving in time. Remind that one special operator, called Hamiltonian $\hat{\mathcal{H}}$, determines the temporal evolution of any other operator $\hat{P}$ according to $\hat{P}(t) = \exp(i\mathcal{H}t)P(0)\exp(-i\mathcal{H}t)$. The evolution operator $\hat{T}(t) = \exp(i\hat{\mathcal{H}}t)$ was introduced in Section 5.3. The quantum-mechanical average of $\hat{P}(t)$ is calculated as a trace with the evolution operator normalized by the trace of the evolution operator:

$$\langle \hat{P} \rangle = \frac{\mathrm{Tr}\,\hat{T}(t)\hat{P}}{Z(t)} \, , \qquad Z(t) = \mathrm{Tr}\,\hat{T}(t) = \sum_a e^{-itE_a} \, . \tag{192}$$

The normalization factor is naturally to call the partition function, all the more if we formally consider it for an imaginary time $t = i\beta$, now related to the inverse temperature:

$$Z(\beta) = \mathrm{Tr}\,\hat{T}(i\beta) = \sum_a e^{-\beta E_a} \, . \tag{193}$$

If the inverse "temperature" β goes to infinity then all the sums are dominated by the ground state, $Z(\beta) \approx \exp(-\beta E_0)$ and the average in (194) are just expectation values in the ground state.

That quantum mechanical description can be compared with the transfer-matrix description for the systems with nearest neighbor interaction. Take for simplicity the Ising model whose Gibbs probability distribution, $\exp(-\beta\mathcal{H})$, is expressed via the classical Hamiltonian,

$$\mathcal{H} = \frac{J}{2}\sum_{i=1}^{N-1}(1 - \sigma_i\sigma_{i+1}) \, , \quad \sigma_i = \pm 1 \ . \tag{194}$$

Consider it on a ring so that $\sigma_{N+1} = \sigma_1$ and write the partition function as a simple sum over spin value at every cite:

$$Z = \sum_{\{\sigma_i\}} \exp\left[-\frac{\beta J}{2}\sum_{i=1}^{N-1}(1-\sigma_i\sigma_{i+1})\right] \tag{195}$$

$$= \sum_{\{\sigma_i\}} \prod_{i=1}^{N-1} \exp\left[-\frac{\beta J}{2}(1-\sigma_i\sigma_{i+1})\right] \tag{196}$$

Every factor in the product can have four values, which correspond to four different choices of $\sigma_i = \pm 1, \sigma_{i+1} = \pm 1$. Therefore, every factor can be written as a matrix element of 2×2 matrix: $\langle \sigma_j | \hat{T} | \sigma_{j+1} \rangle = T_{\sigma_j \sigma_{j+1}} = \exp[-\beta J(1 - \sigma_i \sigma_{i+1})/2]$. It is called the transfer matrix because it *transfers* us from one cite to the next.

$$T = \begin{pmatrix} T_{1,1} & T_{1,-1} \\ T_{-1,1} & T_{-1,-1} \end{pmatrix} \tag{197}$$

where $T_{11} = T_{-1,-1} = 1$, $T_{-1,1} = T_{1,-1} = e^{-\beta J}$. For any matrices $\hat{A}, \hat{B}$ the matrix elements of the product are $[AB]_{ik} = A_{ij}B_{jk}$. Therefore, when we sum over the values of the intermediate spin, we obtain the matrix elements of the matrix squared: $\sum_{\sigma_i} T_{\sigma_{i-1}\sigma_i} T_{\sigma_i \sigma_{i+1}} = [T^2]_{\sigma_{i-1}\sigma_{i+1}}$. The sum over $N-1$ spins gives T^{N-1}. Because of periodicity we end up with summing over a single spin which corresponds to taking trace of the matrix:

$$Z = \sum_{\{\sigma_i\}} T_{\sigma_1\sigma_2} T_{\sigma_2\sigma_3} \ldots T_{\sigma_N\sigma_1} = \sum_{\sigma_1 = \pm 1} \langle \sigma_1 | \hat{T}^{N-1} | \sigma_1 \rangle = \mathrm{Tr}\, T^{N-1} \ . \tag{198}$$

We thus see that taking the sum over two values of σ at every cite in the Ising model is the analog of taking trace in quantum-mechanical average. If there are m values on the cite, then T is $m \times m$ matrix. For a spin in n-dimensional space (described by so-called $O(n)$ model), trace means integrating over orientations. The translations along the chain are analogous to quantum-mechanical translations in (imaginary) time. This analogy is not restricted to 1d systems, one can consider 2d strips that way too.

8.11 Quantum thermalization

Is there any quantum analog of chaos which underlies thermalization the same way that dynamical chaos underlies mixing and thermalization in the classical statistics, as described in Sect. 2.2? Writing the classical formula of exponential separation, $\delta x(t) = \delta x(0) e^{\lambda t}$ as $\partial x(t)/\partial x(0) = e^{\lambda t}$ and replacing quantum-mechanically the space derivative by the momentum operator, one naturally comes to consider the commutator of $\hat{x}(t)$ and $\hat{p}(0)$. Indeed,

$$\frac{\partial x(t)}{\partial x(0)} = \frac{\partial x(t)}{\partial x(0)}\frac{\partial p(0)}{\partial p(0)} - \frac{\partial x(t)}{\partial p(0)}\frac{\partial p(0)}{\partial x(0)} = \{x(t), p(0)\} \to \hbar^{-1}[\hat{x}(t), \hat{p}(0)]\,.$$

That corresponds to the Heisenberg representation, where operators are time-dependent. The commutator measures the effect of having at $t = 0$ the value of $\hat{p}$ on the later measurement of $\hat{x}(t)$. The average value of this commutator over Gibbs distribution with a finite temperature T is zero. Averaging the square, $C(t) = \langle [x(t)p(0)]^2)\rangle$, brings the concept of a so-called out-of-time-order correlation function like $\langle x(t)p(0)x(t)p(0)\rangle$. Such quantities are found to grow exponentially in time in some quantum systems (complicated enough to allow chaos and simple enough to allow for analytic solvability): $C(t) = \hbar^2 e^{2\lambda t}$, where uncertainty relation gives the starting value at $t = 0$. The commutator squared is bounded, so that the exponential growth saturates when $C(t)$ is getting comparable with $\langle p^2\rangle\langle x^2\rangle$ - that value is supposed to be much larger than $\hbar^2$, which requires quasi-classical limit. Respective Lyapunov exponent dimensionally must be energy (temperature) divided by $\hbar$ and indeed $\lambda = 2\pi T/\hbar$ was shown to be a universal upper limit. To appreciate this, note that for a particle with the mass m, the time of effective scattering λ^{-1} could not be less than the de Broglie wavelength $\hbar/\sqrt{mT}$ divided by the thermal velocity $\sqrt{T/m}$ (and the mass drops out!). The limit is reached, for instance, by black holes, which scramble quantum information at the greatest possible rate.

When there are many interacting particles, then the growth of the many-particle version, $C_{ij} = \langle [x_i(t)p_j(0)]^2)\rangle$ describes how entanglement of more and more distant particles appears on the way to thermalization. Indeed, the evolution of the operators in the Heisenberg representation is governed by the Hamiltonian $\mathcal{H}\{\hat{x}_i, \hat{p}_i\}$:

$$\hat{x}_i(t) = e^{\imath\mathcal{H}t}\hat{x}_i(0)e^{-\imath\mathcal{H}t} = \sum_{j=0}^{\infty}\frac{(\imath t)^j}{j!}[\mathcal{H}\ldots[\mathcal{H}\hat{x}_i(0)]\ldots]\,. \tag{199}$$

Since the Hamiltonian describes interaction between particles, then the subsequent terms of the expansion will involve more and more particles, which describes the growth of entanglement with time.

Take for example, the chain with harmonic interaction described by the Hamiltonian $\mathcal{H}\{\hat{x}_i, \hat{p}_i\} = K + U = \sum_i \hat{p}_i^2/2m + (\hat{x}_i - \hat{x}_{i+1})^2$ and consider the one-particle creation operator at some point i. Commutator with the kinetic energy just moves it to another point without increasing the number of particles involved: $[K, c_i^\dagger] = c_{i-1}^\dagger + c_{i+1}^\dagger$. But each commutator with the potential energy increases the number of the operators involved: $[U, c_i^\dagger] \propto c_{i-1}^\dagger c_i^\dagger c_{i+1}^\dagger$.