Physical Nature of Information

G. Falkovich

March 20, 2020

How to get, send and forget information

Contents

The course was initially 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 statistical physics, because this discipline (and this course) answers the most frequent question: How much can we say about something we do not know? The simplest phenomenological approach (called thermodynamics) deals only with macroscopic manifestations of hidden degrees of freedom. It uses symmetries and conservation laws to restrict possible outcomes and focuses on mean values (averaged over many outcomes) ignoring fluctuations. More sophisticated approach is that of statistical physics, which derives the governing laws by explicitly averaging over the degrees of freedom. Those laws justify thermodynamic description of mean values and describe the probability of different fluctuations.

I shall start by briefly reminding basics of thermodynamics and statistical physics and 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. We first study how irreversible entropy change appears from reversible flows in phase space and learn the basics of dynamical chaos. We shall understand that *entropy is not a property of a system, but of our knowledge of the system.* It is then natural to re-tell the story using the language of information theory, which shows universal applicability of this framework: From bacteria and neurons to markets and quantum computers, one magic instrument appears over and over: mutual information and its quantum sibling, entanglement entropy. We then focus on the so far most sophisticated way to forget information - renormalization group. Forgetting is a fascinating activity — one learns truly fundamental things this way. At the end, I shall briefly describe stochastic thermodynamics and modern generalizations of the second law.

The course teaches two complementary ways of thinking: continuous flows and discrete combinatorics. Together, they produce a powerful and universal tool, applied everywhere, from computer science and machine learning to biophysics, economics and sociology. The course emphasis is less on giving wide knowledge and more on providing methods to acquire knowledge. At the end, recognizing the informational nature of physics and breaking the barriers of specialization is also of value for those who stay physicists (no matter what we are doing). 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 animal behavior and trying to figure out how the brain works. Small-print parts can be omitted upon the first reading.

1 Thermodynamics (brief reminder)

One can teach monkey to differentiate, integration requires humans.
Novosibirsk University saying

Physics is an experimental science, and laws appear usually 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 Newton equations of mechanics to Hamiltonian or Lagrangian functions 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 simple and elegant. It also provides a good vantage point for generalizations and appeals to our brain, which likes to theoretize even before receiving any data. 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 such a deduction is presented for thermodynamics following the spirit of the book H. B. Callen, *Thermodynamics* (1965).

1.1 Basic notions

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 believe that we found the way around the partial knowledge in mechanics, electricity and magnetism, where we have *closed description of the explicitly known degrees of freedom.* Even in those cases our knowledge is partial, but we restrict our description only to things that we can predict with full confidence. For example, planets are large complex bodies, and yet the motion of their centers of mass in the limit of large distances allows for a closed description of celestial mechanics. Already the next natural problem — how to describe a planet rotation — needs the account of many extra degrees of freedom, such as, for instance, oceanic flows (which slow down rotation by tidal forces).

In this course 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 that we know the symmetries and conservation laws of the system. *Thermodynamics studies restrictions* *on the possible properties of macroscopic matter that follow from the symmetries of the fundamental laws.* Therefore, thermodynamics does not predict numerical values but rather sets inequalities and establishes relations among different properties.

Thermodynamics started with physical systems, where the basic symmetry is invariance with respect to time shifts, which gives energy conservation[1]. That allows one to introduce the internal energy E. Energy change generally consists of two parts: the energy change of macroscopic 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. We wish to establish the energy of a given system in states independent of the way they are prepared. We call such states equilibrium, they are those that can be completely characterized by the *static* values of observable variables.

For a given system, any two equilibrium states A and B can be related by an adiabatic process either $A \to B$ or $B \to A$, which allows to measure the difference in the internal energy by the work W done by the system. Now, if we encounter a process where the energy decrease is not equal to the work done, we call the difference the heat flux into the system:

$$dE = \delta Q - \delta W \ . \tag{1}$$

This statement is known as the first law of thermodynamics. 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 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 any composite system. The values assumed by the parameters in the absence of an internal constraint maximize the entropy over the manifold of constrained equilibrium states.

[1] Be careful trying to build thermodynamic description for biological or social-economic systems, since generally they are not time-invariant. For instance, living beings age and the total amount of money generally grows (not necessarily in your pocket).

Thermodynamic limit. Traditionally, thermodynamics have dealt with extensive parameters whose value for a composite system is a direct sum of the values for the components. Of course, energy of a composite system is not generally the sum of the parts because there is an interaction energy. To treat energy as an extensive variable we therefore must 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 $V^{2/3}$ in comparison with the bulk terms that scale as V. Other extensive quantities are volume V, number of particles N, electric and magnetic moments, etc.

Thermodynamic entropy is an extensive variable[2], which is a homogeneous first-order function of all the extensive parameters:

$$S(\lambda E, \lambda V, \ldots) = \lambda S(E, V, \ldots) \ . \tag{2}$$

This function (called also fundamental relation) is *everything* one needs to know to solve the basic problem (and others) in thermodynamics.

To avoid misunderstanding, note that (2) 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 see in a moment that it is a convex function. The entropy is generally a monotonic function of energy[3], so that $S = S(E, V, \ldots)$ can be solved 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. Consider the case $(\partial E/\partial S)_X > 0$. An efficient way to treat partial derivatives is to use jacobians $\partial(u, v)/\partial(x, y) = (\partial u/\partial x)(\partial v/\partial y) - (\partial v/\partial x)(\partial u/\partial y)$ and the identity $(\partial u/\partial x)_y = \partial(u, y)/\partial(x, y)$. Then

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

Differentiating the last relation one more time we get

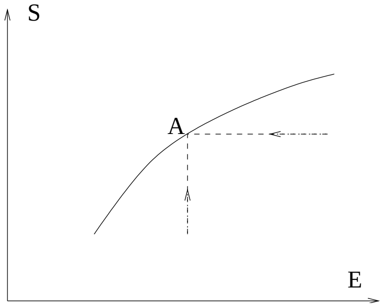
$$(\partial^2 E/\partial X^2)_S = -(\partial^2 S/\partial X^2)_E (\partial E/\partial S)_X \ ,$$

since the derivative of the second factor is zero as it is at constant X. We thus see that in the case $(\partial E/\partial S)_X > 0$ the equilibrium is defined by the

[2] We shall see later that the most interesting, fundamental and practical things are related to the non-extensive part of entropy.

[3] This is not always so, particularly for systems with a finite phase space, as shows a counter-example of the two-level system in Section 2.2.

energy minimum instead of the entropy maximum (very much like circle can be defined as the figure of either maximal area for a given perimeter or of minimal perimeter for a given area). 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 halfs with a bit different energies, the entropy must decrease: $2S(E) > S(E+\Delta) + S(E-\Delta) = 2S(E) + S''\Delta^2/2$, which requires $S'' < 0$. On the figure, unconstrained equilibrium states lie on the curve while all other states lie below. 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[4]. For example, for the energy one writes

$$\frac{\partial E}{\partial S} \equiv T(S,V,N)\,, \quad \frac{\partial E}{\partial V} \equiv -P(S,V,N) \quad \frac{\partial E}{\partial N} \equiv \mu(S,V,N)\ ,\ldots \tag{3}$$

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. From (3) we write

$$dE = \delta Q - \delta W = TdS - PdV + \mu dN\ . \tag{4}$$

Entropy is thus responsible for hidden degrees of freedom (i.e. heat) while other extensive parameters describe macroscopic degrees of freedom. We see

[4] In thermodynamics we have only extensive and intensive variables, because we take thermodynamic limit $N \to \infty$, $V \to \infty$ keeping N/V finite.

that in equilibrium something is maximal for hidden degrees of freedom but this "something" is not their energy.

Let us give an example 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,

$$dS = \frac{\partial S_1}{\partial E_1} dE_1 + \frac{\partial S_2}{\partial E_2} dE_2 = \frac{dE_1}{T_1} + \frac{dE_2}{T_2} = \left(\frac{1}{T_1} - \frac{1}{T_2}\right) dE_1 \,, \tag{5}$$

must be positive which means that energy flows from the hot subsystem to the cold one ($T_1 > T_2 \Rightarrow \Delta E_1 < 0$). We see that our definition (3) 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.

Both energy and entropy are homogeneous first-order functions of its 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

$$E = TS - PV + \mu N \ . \tag{6}$$

It may seem that a thermodynamic description of a one-component system requires operating functions of three variables. Let us show that there are only two independent parameters. For example, the chemical potential μ can be found as a function of T and P. Indeed, differentiating (6) and comparing with (4) 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 N variable, for instance, $E(S, V, N) = NE(s, v, 1) = Ne(s, v)$. In the entropy representation,

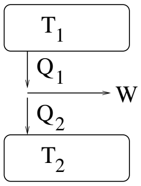
$$S = E\frac{1}{T} + V\frac{P}{T} - N\frac{\mu}{T} \,,$$

the Gibbs-Duhem relation is 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:

$$Ed(1/T) + Vd(P/T) - Nd(\mu/T) = 0 \ . \tag{7}$$

Processes. While thermodynamics is fundamentally about states it is also used for describing processes that connect states. Particularly important questions concern performance of engines and heaters/coolers. Heat engine works by delivering heat from a reservoir with some higher T_1 via some system to another reservoir with T_2 doing some work in the process[5]. If the entropy of the hot reservoir decreases by some ΔS_1 then the entropy of the cold one must increase by some $\Delta S_2 \geq \Delta S_1$. The work W is the difference between the heat given by the hot reservoir $Q_1 = T_1 \Delta S_1$ and the heat absorbed by the cold one $Q_2 = T_2 \Delta S_2$ (assuming both processes quasi-static). Engine efficiency is the fraction of heat used for work that is

$$\frac{W}{Q_1} = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{T_2 \Delta S_2}{T_1 \Delta S_1} \leq 1 - \frac{T_2}{T_1} \ . \tag{8}$$



One cannot overestimate the historical importance of this simple relation: empirical studies of the limits of engine efficiency lead to the distillation of the entropy concept. Indeed, maximal work is achieved for minimal entropy change $\Delta S_2 = \Delta S_1$, which happens for reversible (quasi-static) 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.

Similarly, refrigerator/heater is something that does work to transfer heat from cold to hot systems. The performance is characterized by the ratio of transferred heat to the work done. For the cooler, the efficiency is $Q_2/W \leq T_2/(T_1 - T_2)$, for the heater it is $Q_1/W \leq T_1/(T_1 - T_2)$. Now, the efficiency is large when the temperatures are close, as it requires almost no work to transfer heat.

Summary of formal structure: The fundamental relation (in energy representation) $E = E(S, V, N)$ is equivalent to the three equations of state

[5] Look under the hood of your car to appreciate the level of idealization achieved in that definition.

(3). 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 $R \simeq 8.3\,\mathrm{J/mole\,K} \simeq 2\,\mathrm{cal/mole\,K}$):

$$PV = NRT\ , \qquad E = 3NRT/2\ . \tag{9}$$

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 (6) in the form

$$S = E\frac{1}{T} + V\frac{P}{T} - N\frac{\mu}{T}\ . \tag{10}$$

Here we need to express intensive variables $1/T, P/T, \mu/T$ via extensive variables. The equations of state (9) give us two of them:

$$\frac{P}{T} = \frac{NR}{V} = \frac{R}{v}\ , \qquad \frac{1}{T} = \frac{3NR}{2E} = \frac{3R}{e}\ . \tag{11}$$

Now we need to find μ/T as a function of e, v using Gibbs-Duhem relation in the entropy representation (7). Using the expression of intensive via extensive variables in the equations of state (11), we compute $d(1/T) = -3Rde/2e^2$ and $d(P/T) = -Rdv/v^2$, and substitute into (7):

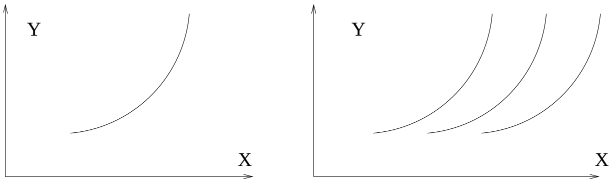
$$d\left(\frac{\mu}{T}\right) = -\frac{3}{2}\frac{R}{e}de - \frac{R}{v}dv\,, \quad \frac{\mu}{T} = C - \frac{3R}{2}\ln e - R\ln v\,,$$

$$s = \frac{1}{T}e + \frac{P}{T}v - \frac{\mu}{T} = s_0 + \frac{3R}{2}\ln\frac{e}{e_0} + R\ln\frac{v}{v_0}\ . \tag{12}$$

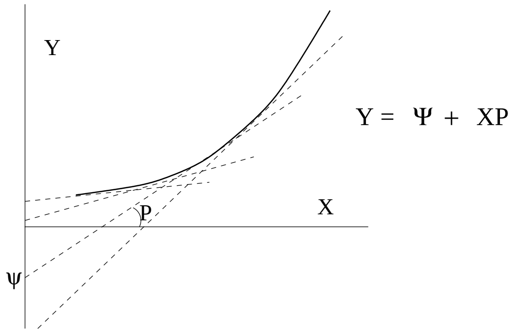
Here e_0, v_0 are parameters of the state of zero internal energy used to determine the temperature units, and s_0 is the constant of integration.

1.2 Legendre transform

Let us emphasize that 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:



For example, $Y = P^2/4$ correspond to the family of functions $Y = (X + C)^2$ for arbitrary C. To fix the shift, we specify for every P 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 characterized by the slope P *and* the intercept ψ. The function $\psi(P) = Y[X(P)] - PX(P)$ completely defines the curve; here one substitutes $X(P)$ found from $P = \partial Y(X)/\partial X$. The function $\psi(P)$ is the Legendre transform of $Y(X)$. From $d\psi = -PdX - XdP + dY = -XdP$ one gets $-X = \partial\psi/\partial P$ 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, $\partial P/\partial X = \partial^2 Y/\partial X^2 \neq 0$. Sign-definite second derivative means that the function is either concave or convex. This is the second time we meet convexity, which can be also related to stability. Indeed, for the function $E(S)$, one-to-one correspondence between S and $T = \partial E/\partial S$ guarantees uniformity of the temperature across the system. 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.

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: $dF(T, V, N, \ldots) = -SdT - PdV + \mu dN + \ldots$. It 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*. 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. Indeed, consider $F(T, X) = E[S(T, X), X] - TS(T, X)$, then $(\partial E/\partial X)_S = (\partial F/\partial X)_T$ that is they turn into zero simultaneously. Also, in the point of extremum, one gets $(\partial^2 E/\partial X^2)_S = (\partial^2 F/\partial X^2)_T$ i.e. both E and F are minimal in 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 $\Delta F = NRT \ln[V_2(V - V_2)/V_1(V - V_1)]$.

Enthalpy $H = E + PV$ is that partial Legendre transform of E which replaces the volume by the pressure $dH(S, P, N, \ldots) = TdS + VdP + \mu dN + \ldots$. It is particularly convenient for situation in which the pressure is maintained constant by a pressure reservoir (say, when the vessel is open into atmosphere). Just as the energy acts as a potential at constant entropy and the free energy as potential at constant temperature, so the enthalpy is a potential for the work done *by the system and the pressure reservoir* at constant pressure. Indeed, now the reservoir delivers pressure which can change the volume so that the differential of the total energy is

$$d(E + E_r) = dE - P_r dV_r = dE + P_r dV = d(E + P_r V) = d(E + PV) = dH \ .$$

Equilibrium minimizes H under the constant pressure. On the other hand, the heat received by the system at constant pressure (and N) is the enthalpy change:

$\delta Q = dQ = TdS = dH$. Compare it with the fact that the heat received by the system at constant volume (and N) is the energy change since the work is zero.

One can replace both entropy and volume obtaining (Gibbs) thermodynamics potential $G = E - TS + PV$ which has $dG(T, P, N, \ldots) = -SdT + VdP + \mu dN + \ldots$ and is minimal in equilibrium at constant temperature and pressure. From (6) we get (remember, they all are functions of different variables):

$$F = -P(T, V)V + \mu(T, V)N \, , \quad H = TS + \mu N \, , \quad G = \mu(T, P)N \ . \tag{13}$$

When there is a possibility of change in the number of particles (because our system is in contact with some particle source having a fixed chemical potential) then it is convenient to use the grand canonical potential $\Omega(T, V, \mu) = E - TS - \mu N$ which has $d\Omega = -SdT - PdV - Nd\mu$. The grand canonical potential reaches its minimum under the constant temperature and chemical potential.

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.

2 Statistical mechanics (brief reminder)

Here we introduce microscopic statistical description in the phase space and describe two principal ways (microcanonical and canonical) to derive thermodynamics from statistical mechanics.

2.1 Microcanonical distribution

Consider a *closed* system with the fixed number of particles N and the energy E_0. Boltzmann *assumed* that all microstates with the same energy have equal probability (ergodic hypothesis) which gives the *microcanonical distribution*:

$$\rho(p, q) = A\delta[E(p_1 \ldots p_N, q_1 \ldots q_N) - E_0] \ . \tag{14}$$

Usually one considers the energy fixed with the accuracy Δ so that the microcanonical distribution is

$$\rho = \begin{cases} 1/\Gamma & \text{for } E \in (E_0, E_0 + \Delta) \\ 0 & \text{for } E \notin (E_0, E_0 + \Delta) \, , \end{cases} \tag{15}$$

where Γ is the volume of the phase space occupied by the system

$$\Gamma(E,V,N,\Delta)=\int_{E<\mathcal{H}<E+\Delta} d^{3N}pd^{3N}q\ . \tag{16}$$

For example, for N noninteracting particles (ideal gas) the states with the energy $E=\sum p^2/2m$ are in the $\mathbf{p}$-space near the hyper-sphere with the radius $\sqrt{2mE}$. Remind that the surface area of the hyper-sphere with the radius R in $3N$-dimensional space is $2\pi^{3N/2}R^{3N-1}/(3N/2-1)!$ and we have

$$\Gamma(E,V,N,\Delta)\propto E^{3N/2-1}V^N\Delta/(3N/2-1)!\approx(E/N)^{3N/2}V^N\Delta\ . \tag{17}$$

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

$$S(E,V,N)=\ln\Gamma(E,V,N)\ . \tag{18}$$

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 (5). Assume that the indeterminacy in the energy of any subsystem, Δ, is much less than the total energy E. Then

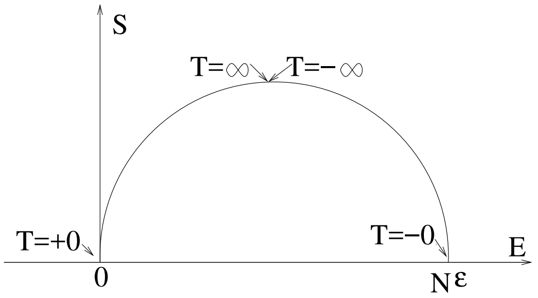
$$\Gamma(E)=\sum_{i=1}^{E/\Delta}\Gamma_1(E_i)\Gamma_2(E-E_i)\ . \tag{19}$$

We denote $\bar E_1,\bar E_2=E-\bar E_1$ the values that correspond to the maximal term in the sum (19). To find this maximum, we compute the derivative of it, which is proportional to $(\partial\Gamma_1/\partial E_i)\Gamma_2+(\partial\Gamma_2/\partial E_i)\Gamma_1=(\Gamma_1\Gamma_2)[(\partial S_1/\partial E_1)_{\bar E_1}-(\partial S_2/\partial E_2)_{\bar E_2}]$. Then the extremum condition is evidently $(\partial S_1/\partial E_1)_{\bar E_1}=(\partial S_2/\partial E_2)_{\bar E_2}$, that is the extremum corresponds to the thermal equilibrium

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

where the temperatures of the subsystems are equal. The equilibrium is thus where the maximum of probability is. It is obvious that $\Gamma(\bar{E}_1)\Gamma(\bar{E}_2) \leq \Gamma(E) \leq \Gamma(\bar{E}_1)\Gamma(\bar{E}_2)E/\Delta$. 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): $S(E, N) = \ln \Gamma \approx N \ln[N/(N-L)] + L \ln[(N-L)/L]$ at $N \gg 1$ and $L \gg 1$. The entropy as a function of energy is drawn in the Figure:



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.

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 part. 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/plus infinity to positive values and eventually reaches the temperature of the thermostat. That is negative temperatures are actually "hotter" than positive. 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 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.

2.2 Canonical distribution and fluctuations

Consider a small subsystem or a system in a contact 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 N, V, T. Let us derive the canonical distribution from the microcanonical. Here our system can have any energy and the question arises what is the probability $W(E)$. Let us find first the probability of the system to be in a given microstate a with the energy E. 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 $\Gamma_0(E_0 - E)$, where 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

$$w_a(E) = Z^{-1}\exp(-E/T)\ , \tag{20}$$

$$Z = \sum_a \exp(-E_a/T)\ . \tag{21}$$

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.

The probability to have a given energy is the probability of the state (20) times the number of states i.e. the statistical weight of the *subsystem*:

$$W(E) = \Gamma(E) w_a(E) = \Gamma(E) Z^{-1}\exp(-E/T)\ . \tag{22}$$

Here the weight $\Gamma(E)$ grows with E very fast for large N. 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 (20) tells that the probability of a given microstate exponentially decays with the energy of the state while (22) tells that the probability of a given energy has a peak.

An alternative and straightforward way to derive the canonical distribution is to use consistently the Gibbs idea of the canonical ensemble as a virtual set, of which the single member is the system under consideration and the energy of the total set is fixed. The probability to have our chosen system in the state a with the energy E_a is then given by the average number of systems $\bar{n}_a$ in this state divided by the total number of systems N. Any set of occupation numbers $\{n_a\} = (n_0, n_1, n_2 \ldots)$ satisfies obvious conditions

$$\sum_a n_a = N \ , \qquad \sum_a E_a n_a = E = \epsilon N \ . \tag{23}$$

Any given set is realized in $W\{n_a\} = N!/n_0!n_1!n_2!\ldots$ number of ways and the probability to realize the set is proportional to the respective W:

$$\bar{n}_a = \frac{\sum n_a W\{n_a\}}{\sum W\{n_a\}} \ , \tag{24}$$

where summation goes over all the sets that satisfy (23). We assume that in the limit when $N, n_a \to \infty$ the main contribution into (24) is given by the most probable distribution that is maximum of W (we actually look at the maximum of $\ln W$ which is the same yet technically simpler) under the constraints (23). Using the method of Lagrangian multipliers we look for the extremum of $\ln W - \alpha\sum_a n_a - \beta\sum_a E_a n_a$. Using the Stirling formula $\ln n! = n\ln n - n$ we write $\ln W = N\ln N - \sum_a n_a \ln n_a$. We thus need to find the value n_a^* which corresponds to the extremum of $\sum_a n_a \ln n_a - \alpha\sum_a n_a - \beta\sum_a E_a n_a$. Differentiating we obtain: $\ln n_a^* = -\alpha - 1 - \beta E_a$ which gives

$$\frac{n_a^*}{N} = \frac{\exp(-\beta E_a)}{\sum_a \exp(-\beta E_a)} \ . \tag{25}$$

The parameter β is given implicitly by the relation

$$\frac{E}{N} = \epsilon = \frac{\sum_a E_a \exp(-\beta E_a)}{\sum_a \exp(-\beta E_a)} \ . \tag{26}$$

Of course, physically $\epsilon(\beta)$ is usually more relevant than $\beta(\epsilon)$.

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):

$$F(T,V,N) = -T \ln Z(T,V,N) \ . \tag{27}$$

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

$$F = \bar{E} + T \left(\frac{\partial F}{\partial T} \right)_V \ ,$$

equivalent to $F = E - TS$ 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. For a subsystem at fixed temperature 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:

$$\begin{aligned} S &= - \ \langle \ln w_a \rangle = -\sum w_a \ln w_a = \sum w_a (E_a/T + \ln Z) \\ &= \ E/T + \ln Z = (E-F)/T = -\ln w_a(E) = S(E) \ . \end{aligned} \tag{28}$$

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.

Generally, 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, and eventually grand canonical distribution accounts for fluctuations both in E and N. The distributions are equivalent only when fluctuations are small. In describing thermodynamics, i.e. mean values, the distributions are equivalent, they just produce different fundamental relations between the mean values: $S(E,N)$ for microcanonical, $F(T,N)$ for canonical, $\Omega(T,\mu)$ for grand canonical, which are related by the Legendre transform. How operationally one checks, for instance, the equivalence of 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).

To describe fluctuations one needs to expand the respective thermodynamic potential around the mean value, using the second derivatives $\partial^2 S/\partial E^2$ and $\partial^2 S/\partial N^2$ (which must be negative for stability). That will give Gaussian distributions of $E - \bar{E}$ and $N - \bar{N}$. Of course, the probability distribution (22) is generally non-Gaussian, but in the thermodynamic limit it can be approximated by Gaussian not far from the (very sharp) maximum. A straightforward way to find the energy variance $\overline{(E-\bar{E})^2}$ is to differentiate with respect to β the identity $\overline{E-\bar{E}} = 0$. For this purpose one can use canonical distribution and get

$$\frac{\partial}{\partial\beta}\sum_a (E_a - \bar{E}) e^{\beta(F-E_a)} = \sum_a (E_a - \bar{E})\left(F + \beta\frac{\partial F}{\partial\beta} - E_a\right) e^{\beta(F-E_a)} - \frac{\partial \bar{E}}{\partial\beta} = 0\,,$$

$$\overline{(E-\bar{E})^2} = -\frac{\partial \bar{E}}{\partial\beta} = T^2 C_V\ . \tag{29}$$

Magnitude of fluctuations is determined by the *second* derivative of the respective thermodynamic potential:

$$\frac{\partial^2 S}{\partial E^2} = \frac{\partial}{\partial E}\frac{1}{T} = -\frac{1}{T^2}\frac{\partial T}{\partial E} = -\frac{1}{T^2 C_V}\ .$$

This is natural: the sharper the extremum (the higher the second derivative) the better system parameters are confined to the mean values. Since both $\bar{E}$ and C_V are proportional to N then the relative fluctuations are small indeed: $\overline{(E-\bar{E})^2}/\bar{E}^2 \propto N^{-1}$. Note that any extensive quantity $f = \sum_{i=1}^N f_i$ which is a sum over independent subsystems (i.e. $\overline{f_i f_k} = \bar{f}_i \bar{f}_k$) have a small relative fluctuation:

$$\frac{(\overline{f^2} - \bar{f}^2)}{\bar{f}^2} = \frac{\sum(\overline{f_i^2} - \bar{f}_i^2)}{(\sum f_i)^2} \propto \frac{1}{N}\ .$$

Let us repeat this important distinction: all thermodynamics potential are equivalent for description of mean values but respective statistical distributions are different. System that can exchange energy and particles with a thermostat has its extensive parameters E and N fluctuating and the grand canonical distribution describes those fluctuations. The choice of description is dictated only by convenience in thermodynamics because it treats only mean values. But in statistical physics, if we want to describe the whole statistics of the system in thermostat, we need to use canonical distribution, not the micro-canonical one. That does not mean that one cannot learn everything about the system by considering it isolated (micro-canonically). Indeed, we can determine C_V (and other second derivatives) for an isolated system and then will know the mean squared fluctuation of energy when we bring the system into a contact with a thermostat.

2.3 Central limit theorem and large deviations

The true logic of this world is to be found in the theory of probability.
Maxwell

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 not very large fluctuations around the mean have Gaussian probability distribution. Consideration of really large fluctuations requires 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

$$\mathcal{P}(X) \ \propto \ \mathrm{e}^{-NH(X/N)} \,. \tag{30}$$

To show this, let us characterize y by its generating function $\langle \mathrm{e}^{zy}\rangle \equiv \mathrm{e}^{G(z)}$ (assuming that the mean value exists for all complex 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 equal to the moments of

$(y - \langle y\rangle)$ called cumulants:

$$\begin{aligned}
&\langle \exp(zy)\rangle = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!}\langle y^n\rangle\,, \quad G(z) = \ln\langle e^{zy}\rangle = \ln\langle 1 + e^{zy} - 1\rangle \\
&= -\sum_{n=1} \frac{1}{n}\left(1 - \langle \exp(zy)\rangle\right)^n = -\sum_{n=1}\frac{1}{n}\left(-\sum_{m=1}^{\infty}\frac{z^m}{m!}\langle y^m\rangle\right)^n \qquad (31)\\
&= z\langle y\rangle + \left(\langle y^2\rangle - \langle y\rangle^2\right)\frac{z^2}{2!} + \ldots = \sum_{n=1}^{\infty}\frac{z^n}{n!}\langle (y - \langle y\rangle)^n\rangle = \sum_{n=1}^{\infty}\frac{z^n}{n!}\langle y^n\rangle_c\ .
\end{aligned}$$

An advantage in working with the cumulants is that for the sum of independent random variables their cumulants and the cumulant generating functions G sum up. For example, consider two random quantities A, B and the second cumulant of their sum: $\langle (A + B - \langle A\rangle - \langle B\rangle)^2\rangle = \langle (A - \langle A\rangle)^2\rangle + \langle (B - \langle B\rangle)^2\rangle)$, which is true as long as $\langle AB\rangle = \langle A\rangle\langle B\rangle$ i.e. A, B are independent. Generating functions are then multiplied. In our case, all y-s in the sum are independent and have identical distributions. Then the generating function of the moments of X has exponential dependence on N: $\langle \mathrm{e}^{zX}\rangle = \langle \exp\left(z\sum_{i=1}^{N} y_i\right)\rangle = \mathrm{e}^{NG(z)}$. The PDF $\mathcal{P}(X)$ is then given by the inverse Laplace transform $\frac{1}{2\pi i}\int \mathrm{e}^{-z\,X + NG(z)}\,dz$ with the integral over any axis parallel to the imaginary one. For large N, the integral is dominated by the saddle point z_0 such that $G'(z_0) = X/N$. This is similar to representing the sum (19) above by its largest term. If there are several saddle-points, the result is dominated by the one giving the largest probability. We now substitute $X = NG'(z_0)$ into $-zX + NG(z)$, and obtain the large deviation relation (30) with

$$H = -G(z_0) + z_0 G'(z_0)\ . \qquad (32)$$

We see that $-H$ and G are related by the ubiquitous Legendre transform. 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):

$$\begin{aligned} G''(z) &= \frac{d^2}{dz^2}\ln\int e^{zy}\mathcal{P}(y)\,dy \\ &= \frac{\int y^2 e^{zy}\mathcal{P}(y)\,dy\int e^{zy}\mathcal{P}(y)\,dy - \left[\int y e^{zy}\mathcal{P}(y)\,dy\right]^2}{\left[\int e^{zy}\mathcal{P}(y)\,dy\right]^2} \geq 0\ . \qquad (33) \end{aligned}$$

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 $\Delta/G'''(0)$.

A simple example is provided by the statistics of the kinetic energy, $E = \sum_1^N p_i^2/2$, of N classical identical unit-mass particles in 1d. The Maxwell distribution over momenta is Gaussian:

$$\rho(p_1,\ldots,p_N) = (2\pi T)^{-N/2}\exp\left(-\sum_1^N p_i^2/2T\right)\ .$$

The energy probability for any N is done by integration, using spherical coordinates in the momentum space:

$$\begin{aligned} \rho(E,N) &= \int \rho(p_1,\ldots,p_N)\delta\left(E - \sum_1^N p_i^2/2\right) dp_1\ldots dp_N \\ &= \left(\frac{E}{T}\right)^{N/2}\frac{\exp(-E/T)}{E\Gamma(N/2)}\ . \qquad (34) \end{aligned}$$

Plotting it for different N, one can appreciate how the thermodynamic limit appears. Taking the logarithm and using the Stirling formula one gets the large-deviation form for the energy $R = E/\bar{E}$, normalized by the mean energy $\bar{E} =$

$NT/2$:

$$\ln\rho(E,N)=\frac{N}{2}\ln\frac{RN}{2}-\ln\frac{N}{2}!-\frac{RN}{2}\approx\frac{N}{2}(1-R+\ln R)\ . \tag{35}$$

This expression has a maximum at $R=1$ i.e the most probable value is the mean energy. The probability of R is Gaussian near maximum when $R-1\le N^{-1/2}$ and non-Gaussian for larger deviations. Notice that this function is not symmetric with respect to the minimum, it has logarithmic asymptotic at zero and linear asymptotic at infinity.

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.

Asymptotic equipartition. The above law of large numbers state that the sum of iid random numbers $y_1+\ldots+y_N$ approaches $N\langle y\rangle$ as N grows. One can also look at the given sequence $y_1,\ldots,y_N$ and ask: how probable it is? This blatantly self-referential question is meaningful nevertheless. Since the numbers are independent, then the logarithm of the probability is the sum that satisfies the law of large numbers:

$$-\frac{1}{N}\ln p(y_1,\ldots,y_N)=-\frac{1}{N}\sum_{i=1}^{N}\ln\mathcal{P}(y_i)\ \to\ -\langle\ln\mathcal{P}(y)\rangle=S(Y)\,. \tag{36}$$

We see that the log of probability converges to N times the entropy of y. But how we find $S(Y)$? For a sufficiently long sequence, we assume that the frequencies of different values of y_i in our sequence give the probabilities of these values; we thus *estimate* $\mathcal{P}(y)$ and compute $S(Y)$. In other words, we assume that the sequence is typical. We then state that the probability of the typical sequence decreases with N exponentially: $p(y_1,\ldots,y_N)=exp[-NS(y)]$. Equivalently, the number of typical sequences grows with N exponentially with entropy setting 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".

3 Appearance of irreversibility

Où sont les neiges d'antan?
François Villon

After we recalled thermodynamics and statistical physics, it is time for reflection. The main 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 answer these questions, 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) consistent use of the information theory which turned out to be just another form of statistical physics. 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.

3.1 Evolution in the phase space

So far we said precious little about how physical systems actually evolve. Let us focus on a broad class of evergy-conserving systems that can be described by the Hamiltonian evolution. Every such system is characterized by its momenta p and coordinates q, together comprising the phase space. We define probability for a system to be in some $\Delta p \Delta q$ region of the phase space as the fraction of time it spends there: $w = \lim_{T\to\infty} \Delta t/T$. 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 density: $dw = \rho(p,q)dpdq$. By definition, the average with the statistical distribution is equivalent to the time average:

$$\bar{f} = \int f(p,q)\rho(p,q)dpdq = \lim_{T\to\infty} \frac{1}{T}\int_0^T f(t)dt \ . \tag{37}$$

The main idea is that $\rho(p,q)$ for a subsystem does not depend on the initial states of this and other subsystems so it can be found without actually solving equations 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 we usually do not know exactly the coordinates and momenta of all particles, we consider the ensemble of identical systems starting from different points in some domain of the phase space. 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, the motion can be considered conservative and described by the Hamiltonian dynamics: $\dot{q}_i = \partial\mathcal{H}/\partial p_i$ and $\dot{p}_i = -\partial\mathcal{H}/\partial q_i$, so that

$$\frac{\partial\rho}{\partial t} = \sum_i \frac{\partial\mathcal{H}}{\partial p_i}\frac{\partial\rho}{\partial q_i} - \frac{\partial\mathcal{H}}{\partial q_i}\frac{\partial\rho}{\partial p_i} \equiv \{\rho, \mathcal{H}\} \, .$$

Here the Hamiltonian generally depends on the momenta and coordinates of the given subsystem and its neighbors. 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 subsystem. As a result, ρ is an integral of motion and it must be expressed solely via the integrals of motion. Since in equilibrium $\ln\rho$ is an additive quantity then 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):

$$\ln\rho_a = \alpha_a + \beta E_a(p,q) + \mathbf{c}\cdot\mathbf{P}_a(p,q) + \mathbf{d}\cdot\mathbf{M}(p,q) \ . \tag{38}$$

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), those 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*:

$$\rho(p,q) = A\exp[-\beta E(p,q)]\ . \tag{39}$$

It was obtained for any macroscopic subsystem of a very large system, which is the same as any system in the contact with 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 of any subsystem, which makes it visiting all regions of the phase space, thus making statistical description possible. Particular case of (39) is a microcanonical (constant) distribution, which is evidently invariant under the Hamiltonian evolution of an isolated system due to Liouville theorem.

Assuming that the system spends comparable time in different available states (ergodic hypothesis) we conclude that since the equilibrium must be the most probable state, then it corresponds to the entropy maximum. In particular, the canonical equilibrium distribution (39) 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 (39). For an isolated system with a fixed energy, the entropy maximum corresponds to a uniform micro-canonical distribution.

3.2 Kinetic equation and H-theorem

How the system comes to the equilibrium and reaches the entropy maximum? What often causes confusion here is that the dynamics (classical and quantum) of any given system is time reversible. The Hamiltonian evolution described above is an incompressible flow in the phase space, $\operatorname{div}\mathbf{v} = 0$, so it conserves the total Gibbs entropy: $dS/dt = -\int d\mathbf{x} \ln \rho \frac{\partial \rho}{\partial t} = \int d\mathbf{x}\, \ln\rho \operatorname{div} \rho\mathbf{v} = -\int d\mathbf{x}\,(\mathbf{v}\nabla)\rho = -\int d\mathbf{x}\, \rho \operatorname{div}\mathbf{v} = 0$. How then the entropy 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 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 $\mathbf{P}=\{\mathbf{p}_1\dots\mathbf{p}_N\}$ 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:

$$\frac{\partial\rho(\mathbf{P},\mathbf{Q},t)}{\partial t}=\{\rho(\mathbf{P},\mathbf{Q},t),\mathcal{H}\}=\left[-\sum_i^N\frac{\mathbf{p}_i}{2m}\frac{\partial}{\partial\mathbf{q}_i}+\sum_{i<j}\theta_{ij}\right]\rho(\mathbf{P},\mathbf{Q},t)\ , \quad (40)$$

where

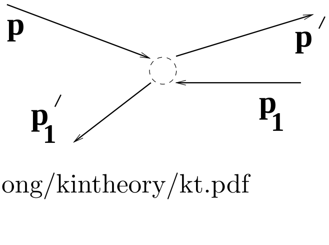
$$\theta_{ij}=\theta(\mathbf{q}_i,\mathbf{p}_i,\mathbf{q}_j,\mathbf{p}_j)=\frac{\partial U(\mathbf{q}_i-\mathbf{q}_j)}{\partial\mathbf{q}_i}\left(\frac{\partial}{\partial\mathbf{p}_i}-\frac{\partial}{\partial\mathbf{p}_j}\right)\ .$$

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 (40):

$$\frac{\partial\rho(\mathbf{p},t)}{\partial t}=\int\theta(\mathbf{q},\mathbf{p};\mathbf{q}',\mathbf{p}')\rho(\mathbf{q},\mathbf{p};\mathbf{q}',\mathbf{p}')\,d\mathbf{q}d\mathbf{q}'d\mathbf{p}'\ , \quad (41)$$

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,[7] 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 $\mathbf{p},\mathbf{p}_1$:



[7] see e.g. http://www.damtp.cam.ac.uk/user/tong/kintheory/kt.pdf

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 probabilities $\rho(\mathbf{p})\rho(\mathbf{p}_1)$ and depend both on initial momenta $\mathbf{p}$, $\mathbf{p}_1$ and the final ones $\mathbf{p}'$, $\mathbf{p}'_1$:

$$w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)\rho(\mathbf{p})\rho(\mathbf{p}_1)\,d\mathbf{p}d\mathbf{p}_1d\mathbf{p}'d\mathbf{p}'_1\ . \tag{42}$$

One may *believe* that (42) must work well when the distribution function evolves on a time scale much longer than that of a single collision. We assume that the medium is invariant with respect to inversion $\mathbf{r}\to-\mathbf{r}$ which gives the *detailed equilibrium*:

$$w\equiv w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)=w(\mathbf{p}',\mathbf{p}'_1;\mathbf{p},\mathbf{p}_1)\equiv w'\ . \tag{43}$$

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 $\mathbf{p}_1\mathbf{p}'\mathbf{p}'_1$:

$$\frac{\partial\rho}{\partial t}\ =\ \int(w'\rho'\rho'_1-w\rho\rho_1)\,d\mathbf{p}_1d\mathbf{p}'d\mathbf{p}'_1\,. \tag{44}$$

We now use the probability normalization which states the sum of transition probabilities over all possible states, either final or initial, is unity and so the sums are equal to each other:

$$\int w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}'_1)\,d\mathbf{p}'d\mathbf{p}'_1=\int w(\mathbf{p}',\mathbf{p}'_1;\mathbf{p},\mathbf{p}_1)\,d\mathbf{p}'d\mathbf{p}'_1\ . \tag{45}$$

Using (45) we transform the second term (44) and obtain the famous *Boltzmann kinetic equation* (1872):

$$\frac{\partial\rho}{\partial t}=\int w'(\rho'\rho'_1-\rho\rho_1)\,d\mathbf{p}_1d\mathbf{p}'d\mathbf{p}'_1\equiv I\,, \tag{46}$$

H-theorem. Let us look at the evolution of the entropy

$$\frac{dS}{dt}=-\int\frac{\partial\rho}{\partial t}\ln\rho\,d\mathbf{p}=-\int I\ln\rho\,d\mathbf{p}\ , \tag{47}$$

The integral (47) contains the integrations over all momenta so we may exploit two interchanges, $\mathbf{p}_1 \leftrightarrow \mathbf{p}$ and $\mathbf{p}, \mathbf{p}_1 \leftrightarrow \mathbf{p}', \mathbf{p}'_1$:

$$\begin{aligned} \frac{dS}{dt} &= \int w'(\rho\rho_1 - \rho'\rho'_1) \ln \rho \, d\mathbf{p} d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}'_1 \\ &= \frac{1}{2} \int w'(\rho\rho_1 - \rho'\rho'_1) \ln(\rho\rho_1) \, d\mathbf{p} d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}'_1 \\ &= \frac{1}{2} \int w' \rho\rho_1 \ln \frac{\rho\rho_1}{\rho'\rho'_1} \, d\mathbf{p} d\mathbf{p}_1 d\mathbf{p}' d\mathbf{p}'_1 \geq 0 \, , \end{aligned} \tag{48}$$

Here we may add 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 then use the inequality $x \ln x - x + 1 \geq 0$ with $x = \rho\rho_1/\rho'\rho'_1$.

Even if 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. 3.1. All this is true also for the inhomogeneous equilibrium in the presence of an external force.

One can look at the transition from (40) to (46) from a temporal viewpoint. N-particle distribution changes during every collision when particles exchange momenta. On the other hand, changing the single-particle distribution requires many collisions. In a dilute system with short-range interaction, the collision time is much shorter than the time between collisions, so the transition 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 the central notion of this course - mutual information - and will be given in Section 4.4 below. For now, a brief statement will suffice: If one starts from a set of uncorrelated particles and let them interact, then the interaction will build correlations and the total distribution will change, but the total entropy will not. However, the single-particle entropy will generally grow, since the Boltzmann equation 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, we broke time reversibility when we assumed particles uncorrelated before the collision.

One may think that for dilute systems there must be a regular expansions of different quantities in powers of density. In particular, molecular chaos factorization of two-particle distribution $\rho_{12} = \rho(\mathbf{q_1}, \mathbf{p_1}; \mathbf{q_2}, \mathbf{p_2})$ via one-particle distributions ρ_1 and ρ_2 is expected to be just the first term of such an expansion:

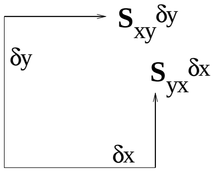
$$\rho_{12} = \rho_1\rho_2 + \int d\mathbf{q_3} d\mathbf{p_3} J_{123}\rho_1\rho_2\rho_3 + \dots \ .$$

In reality 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 3.5. 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.

3.3 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 see 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 it, 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 so that trajectories started from a small neighborhood are found in larger and larger regions of phase space 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 points, separated by $\mathbf{r}$, is determined by their velocity difference: $\delta v_i = r_j \partial v_i / \partial x_j = r_j \sigma_{ij}$. 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). We are interested here in deformation because it is the mechanism of the entropy growth. 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 volume change, must be zero: $\mathrm{Tr}\,\sigma_{ij} = \sum_i S_i = div\,\mathbf{v} = 0$ — 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:

$$r_i(t) = r_i(0) \exp\left[\int_0^t S_i(t')\, dt'\right] \ . \tag{49}$$

For a time-independent strain, the growth/decay is exponential in time. One

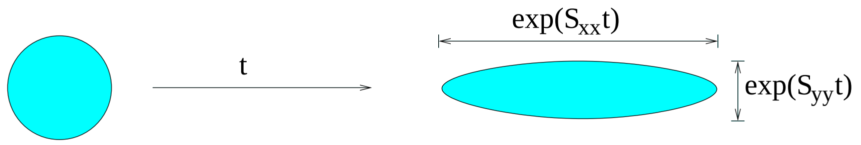


Figure 1: Deformation of a phase-space element by a permanent strain.

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

$$\begin{aligned}
&x(t) = e^{S_{11}t} x_0 \,, \\
&y(t) = e^{S_{22}t} y_0 = e^{S_{22}t}\sqrt{R^2 - x_0^2} = e^{S_{22}t}\sqrt{R^2 - x^2(t)e^{-2S_{11}t}} \,, \\
&x^2(t)e^{-2S_{11}t} + y^2(t)e^{-2S_{22}t} = R^2 \,. \qquad (50)
\end{aligned}$$

The equation (50) 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 $\sum_{i=1}^{6N} x_i^2(t)e^{-2S_i t}$ =const.

Of course, 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 how the system interacts with the environment, 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 result. It may seem counter-intuitive at first, but in a general case an exponential stretching persists on average and the majority of trajectories separate. Physicists think in two ways: one in space and another in time (unless they are relativistic and live in a space-time).

Let us first look at separation of trajectories from a temporal perspective, going with the flow: even when the average rate of separation along a given direction, $\Lambda_i(t) = \int_0^t S_i(t')dt'/t$, is zero, the average exponent of it is larger than unity (and generally growing with time):

$$\lim_{t\to\infty} \int_0^t S_i(t')dt' = 0 \,, \quad \lim_{T\to\infty} \frac{1}{T}\int_0^T dt \exp\left[\int_0^t S_i(t')dt'\right] \geq 1 \,. \qquad (51)$$

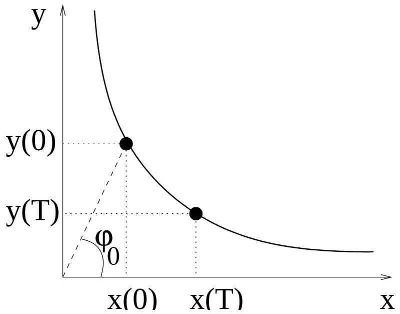


Figure 2: The distance of the point from the origin increases if the angle is less than $\varphi_0 = \arccos[1+\exp(2\lambda T)]^{-1/2} > \pi/4$. For $\varphi = \varphi_0$ the initial and final points are symmetric relative to the diagonal: $x(0) = y(T)$ and $y(0) = x(T)$.

This is because the intervals of time with positive $\Lambda(t)$ give more contribution 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 is $(1/2a)\int_a^{-a} e^\Lambda d\Lambda = (e^a - e^{-a})/2a > 1$.

Looking from a spatial perspective, consider the simplest flow field: two-dimensional[8] pure strain, which corresponds to an incompressible saddle-point flow: $v_x = \lambda x$, $v_y = -\lambda y$. Here we have one expanding direction and one contracting direction, their rates being equal. The vector $\mathbf{r} = (x, y)$ (the distance between two close trajectories) 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.

The net stretching and separation of trajectories is formally proved in mat-

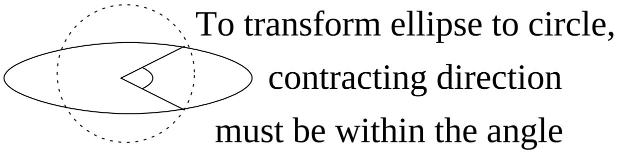
[8] Two-dimensional phase space corresponds to the trivial case of one particle moving along a line, yet it is great illustrative value. Also, remember that the Liouville theorem is true in $p_i - q_i$ plane projection.

hematics by considering 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

$$\lambda_i=\lim_{t\to\infty}t^{-1}\ln|\hat W\mathbf{f}_i| \tag{52}$$

define the so-called Lyapunov exponents, which can be thought of as the mean stretching rates. The sum of the exponents is zero due to the Liouville theorem so 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 $a=-d$, see home exercise.

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. Drop ink into a glass of water, gently stir (not shake) and enjoy the visualization of Furstenberg and Oseledets theorems.

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 box) is stretched in one (unstable) direction and contracted in another (stable) direction so that it turns into a long narrow strip (left and right boxes). Later in time, our resolution is still restricted - rectangles in the right box 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* in the phase space does not contradict the *time-irreversible filling of the phase 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 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 density spreads, entropy grows (as the logarithm of the volume occupied). 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.

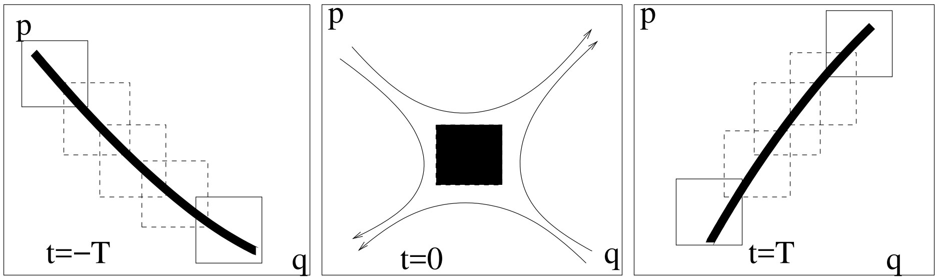
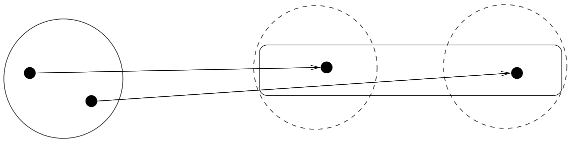


Figure 3: Increase of the phase volume upon stretching-contraction and coarse-graining. Central panel shows the initial state and the velocity field.

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 to briefly discuss our flow from this perspective. Apparently, if we have a finite resolution of a flow with positive Lyapunov exponents, with time we loose our ability to predict where the ensemble of the initially closed systems 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. The two points along a stretching direction that were hidden inside the resolution circle separate with time and can be distinguished:



Moreover, as time proceeds, we learn more and more about the initial locations of the points. The growth rate of such information about the past is again the sum of the positive Lyapunov exponents and is called 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 reminds me 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 make more precise by saying that the qualitative systems behavior is insensitive to small variations of its microscopic parameters).

At even larger time scales than the time of the velocity change for a trajectory, one can consider the motion as a series of uncorrelated random steps. That produces random walk considered in detail in Sect 7.1 below, where we will show that the spread of the probability density $\rho(\mathbf{r},t)$ is described by a simple diffusion: $\partial\rho/\partial t = \kappa\Delta\rho$. The total probability $\int \rho(\mathbf{r},t)\,d\mathbf{r}$ is conserved but the entropy increases monotonically under diffusion:

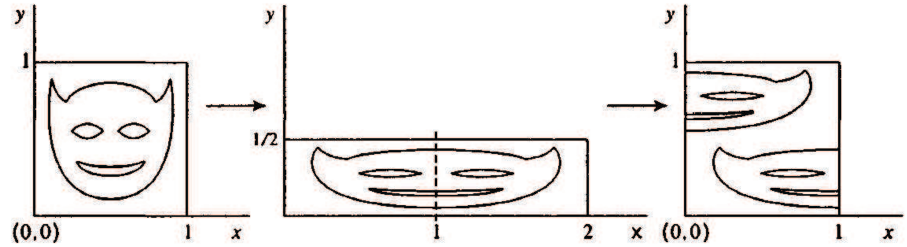
$$\frac{dS}{dt} = -\frac{d}{dt}\int \rho(\mathbf{r},t)\ln\rho(\mathbf{r},t)\,d\mathbf{r} = -\kappa\int \Delta\rho\ln\rho\,d\mathbf{r} = \kappa\int\frac{(\nabla\rho)^2}{\rho}\,d\mathbf{r} \geq 0\,. \qquad (53)$$

Asymptotically in time the solution of the diffusion equation takes the universal form $\rho(\mathbf{r},t) = (4\pi\kappa t)^{-d/2}\exp\left(-r^2/4\kappa t\right)$, see (148) below; substituting it into (53) we obtain a universal entropy production rate, $dS/dt = 1/2t$, independent of κ (which is clear from dimensional reasoning).

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]

3.4 Baker map

We consider a toy model of great illustrative value for the applications of chaos theory to statistical mechanics. Take the phase-space to be a unit square in the (x, y)-plane, with $0 < x, y < 1$. The measure-preserving transformation 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 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 pasta. This transformation is reversible except on the lines where the area was cut in two and glued back.



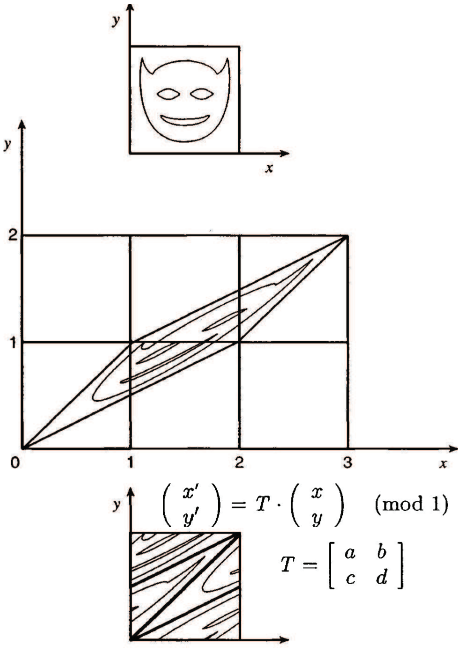
If we consider two initially closed points, then after n such steps the distance along x and y will be multiplied respectively by $2^n = e^{n \ln 2}$ and $2^{-n} = e^{-n \ln 2}$. It is then easy to see, without a lot of formalities, that there are two Lyapunov exponents corresponding to the discrete time n. One of them is connected to the expanding direction and has the value $\lambda_+ = \ln 2$. The

[9] As a student, I've participated (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, which is conceptually important, even though possible in idealized cases only.

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. Consider a small square with the side of the length 2^{-N} inside the unit square. The transformation will stretch this square horizontally and contract it vertically so that after N steps it will be roughly of horizontal dimension unity, and of vertical dimension 2^{-2N}. As the number of steps continues to increase, the original square 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 small set in the unit square will have the same fraction of its area occupied by these little strips of pasta as any other set on the square. This is the indicator of a mixing system. We conclude that a sufficiently smooth initial distribution function defined on the unit square will approach a uniform (microcanonical) distribution on the square.

To avoid impression that cutting and gluing of the baker map are necessary for mixing, consider a smooth model which has 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


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)
$T = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

in the x- and y-direction. The transformation matrix T (an analog of the transfer matrix $\hat{W}$ from the previous section) 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 $\ln \lambda_2$.

Baker map is area-preserving and does not change entropy, yet is we allow for coarse-graining along with the evolutions, then the entropy grows and eventually reaches the maximum, which is the logarithm of the phase volume, which befits the equilibrium microcanonical distribution.

Tutorial: Baker map as a reversible, deterministic, area-preserving transform with an irreversible transport equation for a reduced distribution function and an underlying stochastic-like microscopic dynamics. Dorfman 7.1-3, 8.3.

3.5 Entropy decrease and non-equilibrium fractal measures

As we have seen in the previous two sections, if we have indeterminacy in the data or consider an ensemble of systems, then Hamiltonian dynamics (an incompressible flow) effectively mixes and makes distribution uniform in the phase space. Since we have considered isolated systems, they conserve their integrals of motion, so that the distribution is uniform over the respective surface. In particular, dynamical chaos justifies micro-canonical distribution, uniform over the energy surface.

But what if the dynamics is non-Hamiltonian, that is Liouville theorem is not valid? The flow in the phase space is then generally compressible. For example, we accelerate particles by external forces f_i and damp their momenta with the dissipation rates γ_i, so that the equations of motion take the form: $\dot{p}_i = f_i - \gamma_i p_i - \partial H/\partial q_i$, which gives generally $div\,\mathbf{v} = \sum_i(\partial f_i/\partial p_i - \gamma_i) \neq 0$. Let us show that such flows create quite different distribution. Since $div\,\mathbf{v} \neq 0$, then the probability density generally changes along a flow: $d\rho/dt = -\rho div\,\mathbf{v}$. That produces entropy,

$$\frac{dS}{dt} = \int \rho(\mathbf{r},t) div\,\mathbf{v}(\mathbf{r},t)\, d\mathbf{r} = \langle \rho div\,\mathbf{v} \rangle \,. \tag{54}$$

with the rate equal to the Lagrangian mean of the phase-volume local expansion rate. If the system does not on average heats or cools (expands or contracts), then the whole phase volume does not change. That means that the global average (over the whole volume) of the local expansion rate is zero: $\langle div\,\mathbf{v} \rangle = \int div\,\mathbf{v}\, d\mathbf{r} = 0$. Yet for a non-uniform density, the entropy is not the log of the phase volume but the minus *mean* log of the phase density, $S = -\langle \rho \ln \rho \rangle$, whose derivative (54) is non-zero because of correlations between ρ and $div\,\mathbf{v}$. Since ρ is always smaller in the expanding regions where $div\,\mathbf{v} > 0$, then *the entropy production rate (54) is non-positive.* We conclude that the mean logarithm of the density (i.e. entropy) decreases. Since the uniform distribution has a maximal entropy under the condition of fixed normalization, then the entropy decrease means that the distribution is getting more non-uniform.

What happens then to the density? Of course, if we integrate density over all the phase space we obtain unity at any time: $\langle \rho \rangle = \int \rho(\mathbf{r},t)\, d\mathbf{r} = 1$. Let us now switch focus from space to time and consider the density of an arbitrary fluid element, which evolves as follows:

$$\rho(t)/\rho(0) = \exp\left[-\int_0^t div\,\mathbf{v}(t')\,dt'\right] = e^{C(t)}\,. \tag{55}$$

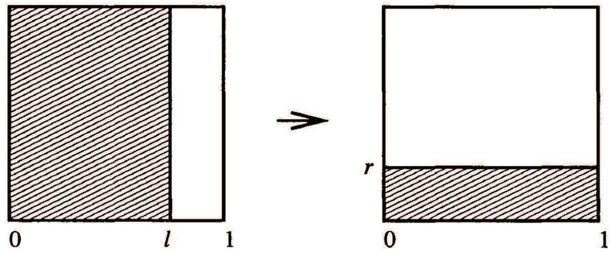
As we have seen in (51), if a mean is zero, the mean exponent generally exceeds unity because of concavity of the exponential function. Now the contraction factor averaged over the whole flow is zero at any time, $\langle C\rangle = 0$, and its average exponent is larger than unity: $\langle \rho(t)/\rho(0)\rangle = \langle e^C\rangle > 1$. That concavity simply means that the parts of the flow with positive C give more contribution into the exponent than the parts with negative C. Moreover, for a generic random flow the density of most fluid elements must grow non-stop as they move. Indeed, if the Lagrangian quantity (taken in the flow reference frame) $div\,\mathbf{v}(\mathbf{r},t)$ is random function with a finite correlation time, then at longer times its integral $\int_0^t div\,\mathbf{v}(t')\,dt'$ is Gaussian with zero mean and variance linearly growing with time (see section 2.3). 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 fractal set. That is opposite to the mixing by Hamiltonian incompressible flow.

In particular, for spatially smooth flow, the long-time Lagrangian average (along the flow)

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

is a sum of the Lyapunov exponents, which is then non-positive. 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.

This can be illustrated by a slight generalization of the baker map, expanding one region and contracting another, keeping the whole volume of the phase space unity:



The transformation has the form

$$x' = \begin{cases} x/l & \text{for } 0 < x < l \\ (x-l)/r & \text{for } l < x < 1\,, \end{cases}$$
$$y' = \begin{cases} ry & \text{for } 0 < x < l \\ r + ly & \text{for } l < x < 1 \end{cases}\,, \tag{56}$$

where $r + l = 1$. The Jacobian of the transformation is not equal to unity when $r \neq l$:

$$J = \left| \frac{\partial(x', y')}{\partial(x, y)} \right| = \begin{cases} r/l & \text{for } 0 < x < l \\ l/r & \text{for } l < x < 1\,, \end{cases}\,. \tag{57}$$

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 limit we obtain the Lyapunov exponents:

$$\lambda_+ \quad = \lim_{n\to\infty} \left[\tfrac{n_1}{n} \ln \tfrac{1}{l} + \tfrac{n_2}{n} \ln \tfrac{1}{r} \right] = -l \ln l - r \ln r\,, \tag{58}$$
$$\lambda_- \quad = \lim_{n\to\infty} \left[\tfrac{n_1}{n} \ln r + \tfrac{n_2}{n} \ln l \right] = r \ln r + l \ln l \tag{59}$$

The sum of the Lyapunov exponents, $\lambda_+ + \lambda_- = (l - r) \ln(r/l) = \overline{\ln J}$, is non-positive and is zero only for $l = r = 1/2$. Long-time average volume contraction of a fluid element and respective entropy production is the analog of the second law of thermodynamics. The volume contraction means that the expansion in the λ_+-direction proceeds slower than the contraction in the λ_--direction. Asymptotically our strips of pasta concentrate on a fractal set, that is one having non-integer dimensionality. Indeed, define the (box-counting) dimension of a set as follows

$$d = \lim_{\epsilon\to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)}\,, \tag{60}$$

where $N(\epsilon)$ is the number of boxes of length ϵ on a side needed to cover the set. After n iterations of the map, square having initial side $\delta \ll 1$ will be stretched into a long thin rectangle of length $\delta \exp(n\lambda_+)$ and width $\delta \exp(n\lambda_-)$. To cover contracting direction, we choose $\epsilon = \delta \exp(n\lambda_-)$, then $N(\epsilon) = \exp[n(\lambda_+ - \lambda_-)]$, so that the dimension is

$$d = 1 + \frac{\lambda_+}{|\lambda_-|} \ , \tag{61}$$

The dimension is between 1 and 2. The set is smooth in the x-direction and fractal in the y-direction, which respectively gives two terms in (61). We see the dramatic difference between equilibrium equipartition and non-equilibrium fractal distribution. Relation between compressibility and non-equilibrium is natural: to make system non-Hamiltonian one needs to act by some external forces, which pump energy into some degrees of freedom and, to keep a steady state, absorb it from other degrees of freedom — expansion and contraction of the momentum part of the phase-space volume. Thermal equilibrium requires fluctuation-dissipation theorem, which is thus violated. In a non-equilibrium steady state, the entropy extraction rate by the environment must be equal to the entropy production rate of the flow.

We thus see that the non-equilibrium steady state (NESS) distribution is singular, that is occupies zero-measure subset of the phase space. The singularity of the non-equilibrium measure is probably related to non-analyticity of kinetic coefficients and other quantities, mentioned at the end of the Section 3.2. In reality, approach to NESS is asymptotic; not much is known what distribution modifications provide for a permanent entropy decrease. Most likely, lower-order cumulants stabilize first, while farther and farther tails of the distribution continue to evolve. Of course, to have already the entropy of the uniform distribution finite we need a finite resolution or coarse-graining, which also stops the entropy decay in non-equilibrium at a finite value. Adding any noise to the system or treating it with a finite resolution makes the measure smooth.

To conclude this Chapter, let us stress the difference between the entropy growth described in the Sections 3.3-3.4 and the entropy decay described in the present Section. 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 rate of the volume growth was solely due to stretching directions and thus equal to the sum of the positive Lyapunov exponents, as described in Section 3.3.

On the contrary, the present section deals with compressible flows which decrease entropy by creating more inhomogeneous distributions, so that 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 the mean rate more than expanding regions.

Looking back at the previous Chapters, it is a good time to appreciate the complementarity of determinism and randomness expressed in terms "statistical mechanics" and "dynamical chaos".

4 Physics of information

This section presents an elementary introduction into the information theory from the viewpoint of a physicist. It re-tells the story of statistical physics using a different language, which lets us to see the Boltzmann and Gibbs entropies in a new light. An advantage of using different formulations is that it helps to understand things better and triggers different intuition in different people. 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 fast widens the region of applications of the universal tool of entropy (and related notion of mutual information): from physics, communications and computations to artificial intelligence and quantum computing.

4.1 Information as a choice

> "Nobody knows what entropy really is, so in a
> debate you will always have an advantage."
> von Neumann to Shannon

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 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

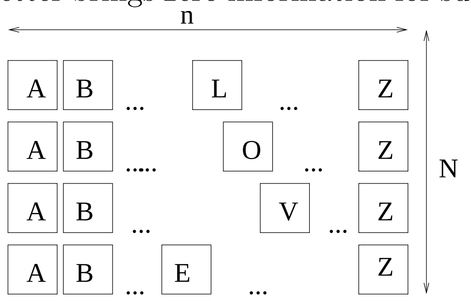
[10] The messages "in box 2 out of 2" and "in box 2 out of 22" bring the same candy but

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 (Hartley 1927, Shannon 1948)

$$I(n) = I(e)\ln n = k\ln n \tag{62}$$

If we measure information in binary choices or bits then $k^{-1} = \ln(2)$. 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. We can easily generalize the definition (62) 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 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 for such people).



not the same amount of information.

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 w_i assigned to each letter (or box) $i = 1, \dots, n$. It is then clear that different letters bring different information. When there is randomness, we evaluate the *average* information per symbol by repeating our choice, say, N times. As $N \to \infty$ we know that candy in the i-th box in Nw_i cases, that is we know that we receive the first alphabet symbol Nw_1 times, the second symbol Nw_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 is $N!/\Pi_i(Nw_i)!$ and the information that we obtained from N symbols is

$$I_N = k \ln\Big(N!/\Pi_i(Nw_i)!\Big) \approx -Nk\sum_i w_i \ln w_i + O(lnN)\ . \tag{63}$$

The mean missing information per symbol in the language coincides with the entropy (28):

$$I(w_1 \dots w_n) = \lim_{N\to\infty} I_N/N = -k\sum_{i=1}^n w_i \ln w_i\ . \tag{64}$$

Note that when $n \to \infty$ then (62) diverges while (64) may well be finite. Knowledge of w_i generally diminishes entropy.

Alternatively, one can derive (64) without any mention of randomness. Consider again n boxes and define $w_i = m_i/\sum_{i=1}^n m_i = m_i/M$, where m_i is the number of compartments in the box number i. The total information in which compartment $k \ln M$ must be a sum of the information about the box I plus the information about the compartment summed over the boxes: $k\sum_{i=1}^n w_i \ln m_i$. That gives the information about the box (letter) as the difference:

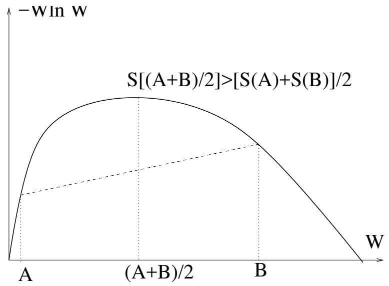
$$I = k\ln M - k\sum_{i=1}^n w_i \ln m_i = k\sum_{i=1}^n w_i \ln M - k\sum_{i=1}^n w_i \ln m_i = -k\sum_{i=1}^n w_i \ln w_i\ .$$

The mean information (64) is zero for delta-distribution $w_i = \delta_{ij}$; it is generally less than the information (62) and coincides with it only for equal probabilities, $w_i = 1/n$, 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, message brings maximum information per symbol; any prior knowledge can reduce the information. Mathematically, the property

$$I(1/n,\dots,1/n) \geq I(w_1 \dots w_n) \tag{65}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} s\left(w_i\right) \leq s\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) \ . \tag{66}$$



From here one gets the entropy inequality:

$$I(w_1 \dots w_n) = \sum_{i=1}^{n} s\left(w_i\right) \leq ns\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) = ns\left(\frac{1}{n}\right) = I\left(\frac{1}{n},\dots,\frac{1}{n}\right) \ . \tag{67}$$

The relation (66) 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 (66) for $n = 2$. To get from $n = 2$ to arbitrary n we use induction. For that end, we choose $\lambda = (n-1)/n$, $a = (n-1)^{-1}\sum_{i=1}^{n-1} w_i$ and $b = w_n$ to see that

$$\begin{aligned} s\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) &= s\left(\frac{n-1}{n}(n-1)^{-1}\sum_{i=1}^{n-1} w_i + \frac{w_n}{n}\right) \\ &\geq \frac{n-1}{n} s\left((n-1)^{-1}\sum_{i=1}^{n-1} w_i\right) + \frac{1}{n}s\left(w_n\right) \\ &\geq \frac{1}{n}\sum_{i=1}^{n-1} s\left(w_i\right) + \frac{1}{n}s\left(w_n\right) = \frac{1}{n}\sum_{i=1}^{n} s\left(w_i\right) \ . \end{aligned} \tag{68}$$

In the last line we used the truth of (66) for $n-1$ to prove it for n.

You probably noticed that (62) corresponds to the microcanonical description (18) giving information/entropy as a logarithm of the number of states, while (64) corresponds to the canonical description (28) giving it as an average. An advantage of Shannon entropy (64) is that is defined for arbitrary distributions, not necessarily equilibrium.

4.2 Communication Theory

After we learnt, what information messages bring on average, we are ready to discuss the best ways to transmit messages. 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 the next letter of the message, on the average. Entropy quantifies redundancy. We have seen that a communication channel on average transmits one unit of the information (64) per symbol. Receiving letter number i through a binary channel (transmitting ones and zeros)[11] brings information $\log_2(1/w_i) = \log_2 M - \log_2 m_i$ bits. Indeed, the remaining choice (missing information) is between m_i compartments. The entropy $-\sum_{i=a}^{z} w_i \log_2 w_i$ is the mean information content per letter. Note that less probable symbols have larger information content, but they happen more rarely. The mean information content for a given letter, $-w \log_2 w$, is maximal for $w = 1/2$.

So the entropy is the mean rate. What about the maximal rate? Following Shannon, we answer the question i) statistically, which makes sense in the limit of very long messages, when one can focus on typical sequences, as we did at the end of the Section 2.3. 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

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

typical strings — those that contain $N(1-p)$ 0's and Np 1'st. The number of such strings is given by the binomial C^N_{Np} which for large N is $2^{NI(p)}$, where $I(p) = -p\log_2 p - (1-p)\log_2(1-p)$. The strings differ by the order of appearance of 0 and 1. To distinguish between these $2^{NI(p)}$ messages, we specify any one using a binary string with lengthes starting from one and up to $NI(p)$. That maximal length is less than N, since $0 \le I(p) \le 1$ for $0 \le p \le 1$. We indeed achieve compression with the sole exception of the case of equal probability where $I(1/2) = 1$. True, the code must include a bit more to represent atypical messages, but in the limit of large N we may neglect the chance of their appearance and their contribution to the rate of transmission. Therefore, entropy sets both the mean and the maximal rate in the limit of long sequences. The idea 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.

But not any encoding guarantees the maximal rate of transmission. Designating sequences of the same length to letters with different probabilities is apparently sub-optimal. Signal compression is achieved by coding common letters by short sequences and infrequent letters 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 w_i listed in the table:

Symbol	w_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. Since there are exactly four two-bit words, we can use the Code 1.

An alternative is a variable-length Code 2. Here we want the least probable C and G to have the longest codewords of the same length differing by one bit that distinguishes between two of them. We then can combine C and G into a single source symbol with the probability 1/4, that is coinciding with T. We thus need to code T by one bit less, etc. Home exercise is to see which code, 1 or 2, uses less bits per base on average.

In English, the probability of "E" is 13% and of "Q" is 0.1%, so Morse[12] encodes "E" by a single dot and "Q" by "$- - \cdot -$" (first British telegraph managed to do without C,J,Q,U,X). One-letter probabilities give for the written English language the information per symbol as follows:

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

which is lower than $\log_2 26 = 4.7$ bits. Apart from one-letter probabilities, one can utilize more knowledge about the language by accounting for two-letter correlation (say, that "Q" is always followed by "U"). That will further lower the entropy, which now can be computed as the entropy of a gas with binary interactions. The simplest of such models is a 1d spin chain where we assume that each letter correlates only with two other letters. More elaborate Ising-type model is applied in the Section 5.1 below to a system of interacting neurons. We shall also discuss there how neurons encode information. Returning to English, long-range correlations lower the entropy down to approximately 1.4 bits per letter.

English messages bring 1.4 bits per letter *if no other information given.* If it was known that administrative messages related to this course contain restricted vocabulary, say, 9 nouns (professor, students, room, date, hour, lecture, tutorial, homework, exam), then the first word in "Students are smart" brings $\log_2 9$ bits. The whole message brings the same amount of information as the message "Professor is stupid".

Comparing 1.4 and 4.7, we conclude that the letters in an English text are about 70% redundant. This 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!"

We talked a lot about letters, but the human language encodes information not in separate letters but in words. An insight into the way we communicate is given by the frequency distribution of words and their meanings (Zipf 1949). It was found empirically that if one ranks words by the

[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.

frequency of their appearance in texts, then the frequency decreases as an inverse rank. For example, the first place with 7% takes *"the"*, followed by *"of"* with 3.5%, *"and"* with 1.7%, etc.

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

$$P(r) = (r+1)^{-1}\ . \tag{69}$$

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. 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 mean are balanced, In what follows we shall be minimizing a lot of two-term functionals and looking for a conditional entropy maximum (which is what most of statisticians do most of the time).

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, for instance, 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.

In distinction from lossless coding, jpeg, mpeg, mp3 and telephone use lossy compression which removes information presumed to be unimportant for humans. Moreover, any transcript of the spoken words has much lower entropy (by orders of magnitude) than the acoustic waveform, which means that it is possible to compress (much) further without losing any information about the words and their meaning, as will be discussed in the next subsection. 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. Incidentally, the great 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.

4.3 Mutual information as a universal tool

Answering the question i) in Sect. 4.1, we have found that the entropy of the set of symbols to be transferred determines the minimum mean number of bits per symbol, that is the maximal rate of information transfer. In this section, we turn to the question ii) and find out how this rate is lowered if the transmission channel can make errors. How much information then is lost on the way? In this context one can treat measurements as messages about the value of the quantity we measure. One can also view storing and retrieving information as sending a message through time rather than space.

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 introducing conditional and relative entropies and mutual information, which turn out to be 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 so-called conditional probability (of B_j in the presence of A_i), denoted $P(B_j|A_i)$. For every A_i, this is a usual 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*:

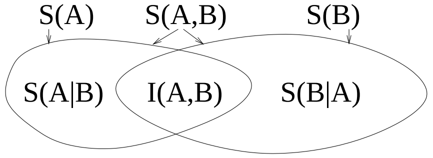
$$\begin{aligned} S(B|A) &= \sum_i P(A_i)S(B|A_i) = \sum_{ij} P(A_i)P(B_j|A_i)\log_2 P(B_j|A_i) \\ &= \sum_{ij} P(A_i, B_j)\log_2 P(B_j|A_i)\ . \end{aligned} \tag{70}$$

Here we related the conditional probability to the joint probability $P(A_i, B_j)$ by the evident formula $P(A_i, B_j) = P(B_j|A_i)P(A_i)$. The conditional entropy measures what on average remains unknown 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*:

$$I(A,B) = S(B) - S(B|A) = \sum_{ij} P(A_i, B_j)\log_2\left[\frac{P(B_j|A_i)}{P(B_j)}\right]\ . \tag{71}$$

Non-negativity of information means that on average measurements increase the conditional probability: $\langle\log_2[P(B_j|A_i)/P(B_j)]\rangle \geq 0$. For example, if B is a choice out of n equal possibilities ($P(B) = 1/n$ and $S(B) = \log_2 n$), but A can happen only in m cases out of those n ($P(B|A) = 1/m$ and $S(B|A) = \log_2 m$), then $I(A,B) = \log_2(n/m)$ bits. If $m = 1$ then A tells us all we need to know about B.

The formula $P(A_i, B_j) = P(B_j|A_i)P(A_i)$ gives the chain rule, $S(A,B) = S(A) + S(B|A) = S(B) + S(A|B)$, and $I(A,B)$ in a symmetric form:



$$I(A,B) = S(B) - S(B|A) = S(A) + S(B) - S(A,B) = S(A) - S(A|B) \tag{72}$$

When A and B are independent, the conditional probability is independent of A and the information is zero. When they are dependent, $P(B,A) >$

$P(A)P(B)$, so that that the information is indeed positive. 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. Non-negativity of the mutual information also gives the so-called sub-additivity of entropy:

$$S(A) + S(B) > S(A, B) \ . \tag{73}$$

One also uses $P(A, B) = P(B|A)P(A) = P(A|B)P(B)$ for estimating the conditional probability of the event B given the marginal probability of the measurements A:

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)} \ . \tag{74}$$

For example, experimentalists measure the sensory response of an animal to the stimulus, 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 stimulus $P(B)$ (see example at the beginning of Section 5.1). After that one obtains the conditional probability (74) that allows animal/robot to function in that habitat.

Mutual information sets the maximal rate of reliable communication thus answering the question ii) from the Section 4.2. Indeed, the number of possible outputs for a given input is an inverse of the conditional probability $P(A_i|B_J)$. For each typical input N-sequence, we have $[P(A|B)]^{-N} = 2^{NS(A|B)}$ possible output sequences, all of them equally likely. To identify the input without error, 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.

If one is just interested in the channel as specified by $P(B|A)$, then one maximizes $I(A, B)$ over all choices of the source statistics $P(B)$ and call it the Shannon's channel capacity, which quantifies the quality of communication systems or measurements: $\mathcal{C} = \max_{P(B)} I(A, B)$ in bits per symbol. 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 $P(B = 0) = P(B = 1) = 1/2$. 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 $P(B|A) = 1$, as in the case of $m = 1$ above. The capacity is lowered when the same outputs appear for different inputs, say, different groups of m inputs each gives the same output, so that $P(B|A) = 1/m$. In this case, one achieves error-free transition choosing 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. Lowered capacity means increased redundancy, that is a need to send more symbols to convey the same information.

In most cases, however, noise does not allow to separate inputs into groups with completely disjoint outputs, so errors always present. It was thought that in such cases it is impossible to make probability of error arbitrarily small when sending information with a finite rate R. Shannon have shown that it is possible, 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 one can 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).

In particular, if in a binary channel the probability of every single bit going wrong is q, then A is binary random variable, 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 is unity for a binary variable B. The rate of transmission is bounded from above by the capacity. That can be explained as follows: 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.

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 about 3. Number of ways to encode a given amino acids is approximately proportional to its frequency of appearance.

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 below in Section 4.6.

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.

When the measurement/transmission noise ξ is additive, that is the output is $A = g(B) + \xi$ with an invertible function g, we have $S(A|B) = S(\xi)$ and

$$I(A, B) = S(A) - S(\xi) \ . \tag{75}$$

The more choices of the output are recognizable despite the noise, the more is the capacity of the channel. In general, 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 $S(A)$.

Gaussian Channel. As an illustration, 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 $P(B) = (2\pi)^{-1/2}\exp(-B^2/2\mathcal{S})$, then $P(A) = [2\pi(\mathcal{N}+\mathcal{S})]^{-1/2}\exp[-A^2/2(\mathcal{N}+\mathcal{S})]$. Now, using (74) we can write

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

In particular, the estimate of B is linearly related to the measurement A:

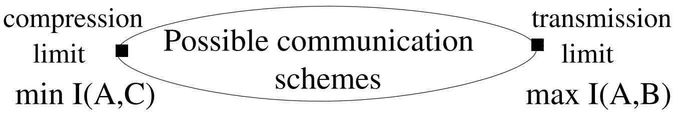
$$\bar{B} = \int BP(B|A)\,dB = \frac{A}{\mathcal{S}+\mathcal{N}} = A\frac{SNR}{1+SNR} , \tag{76}$$

where signal to noise ratio is $SNR = \mathcal{S}/\mathcal{N}$. The rule (76) 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. As is clear from this example, linear relation between the measurement and the best estimate requires two things: linearity of the input-output relation and Gaussianity of the statistics. Let us now find the mutual information (75):

$$\begin{aligned} I(A,B) &= S(A)-S(A|B) = S(A)-S(B+\xi|B) = S(A)-S(\xi|B) = S(A)-S(\xi) \\ &= \tfrac{1}{2}\left[\log_2 2\pi e(1+\mathcal{N}) - \log_2 2\pi e\mathcal{N}\right] = \tfrac{1}{2}\log_2(1+SNR) . \end{aligned} \tag{77}$$

Here we used the formula for the entropy of the Gaussian distribution. 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 $\rho(x) \propto \exp(-\lambda x^2)$.

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 former exceeds the latter, which requires positivity of the mutual information.

4.4 Hypothesis testing and Bayes' rule

All empirical sciences need a quantitative tool for confronting data with hypothesis. 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, which is the Bayes' rule:

$$P(h|e) = P(h)\frac{P(e|h)}{P(e)} \; . \tag{78}$$

That is the new (posterior) probability $P(h|e)$ that the hypothesis is correct after we receive the data 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 experimentalists in science and most wrong conclusions made by conspiracy theorists are connected to unfamiliarity with this simple formula. For example, your hypothesis is the existence of a massive covert conspiracy inside the US government and the evidence is September 11 terrorist attack. In this case $P(e|h)$ is high: a terrorist act 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. But of course, absent such an event, the prior probability $P(h)$ is vanishingly small. Only sequence of probability-increasing events may lead us to accept the hypothesis.

If choosing between two mutually exclusive hypotheses, h_1 and h_2, then

$$P(h_1|e) = P(h_1)\frac{P(e|h_1)}{P(h_1)P(e|h_1) + P(h_2)P(e|h_2)} \; . \tag{79}$$

Suppose that a test for using a particular drug 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 non-drug users (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 users of the drug, 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 than to sensitivity. In particular, checking an a priori improbable hypothesis, $P(h_1) \ll P(h_2)$, it is better to design experiment which minimizes $P(e|h_2)$ rather than maximizes $P(e|h_1)$, that is rules our alternative rather than supports the hypothesis. To see the combination of probabilities which defines the posterior probability of the hypothesis being true, it is instructive to present the result in the following form:

$$P(h_1|e) = \Big[1 + P(h_2)P(e|h_2)/P(h_1)P(e|h_1)\Big] \ . \tag{80}$$

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 and the recently established fact that signals between brain and sensory organs travel in both directions simultaneously. It is then likely that even our unconscious activity uses rational Bayes' rule, where e is sensory input. See e.g. "The Predictive Mind" by J. Hohwy.

Note a shift in interpretation of probability. Traditionally, mathematicians and gamblers treated 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. 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.

Relative Entropy. If the true distribution is p but our hypothetical distribution is q, what is the number N of trials we need to find out that our hypothesis is wrong? For that we need to estimate the probability of the stream of data observed. We shall observe the result i number of times which is $p_i N$ and *judge* the probability of it happening as $q_i^{p_i N}$ times the number of sequences with those frequencies:

$$\mathcal{P} = \prod_i q_i^{p_i N} \frac{N!}{\prod_j (p_j N)!} . \tag{81}$$

This is the probability of our hypothetical distribution being true. Considering limit of large N we obtain a large-deviation-type relation like (30):

$$\mathcal{P} \propto \exp\left[-N\sum_i p_i \ln(p_i/q_i)\right] . \tag{82}$$

The probability of not-exactly-correct hypothesis to approximate the data exponentially decreases with the number of trials. To measure the rate of that decrease, one introduces the *relative entropy* (also called Kullback-Liebler divergence):

$$D(p|q) = \sum_i p_i \ln(p_i/q_i) = \langle \ln(p/q)\rangle \ . \tag{83}$$

The relative entropy determines how many trials we need: we prove our hypothesis wrong when $ND(p|q)$ becomes large. The closer is our hypothesis to the true distribution, the larger is the number of trials needed. On the other hand, when $ND(p|q)$ is not large, our hypothetical distribution is just fine.

Relative entropy also quantifies how close to reality is the asymptotic equipartition estimate (36) 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) advices us 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 if the true probabilities of the letters are $\{p_k\}$, then the probability of the sequence is actually $\prod_k p_k^{Nq_k} = \exp(N\sum_k q_k \ln p_k) = \exp[N\sum_k(q_k \ln q_k + q_k \ln(p_k/q_k))] = \exp[-NS(q) + D(q|p)]$.

How many different probability distributions $\{q_k\}$ (called types in information theory) exist for an N-sequence? Since $q_k = n/N$, where n can take any of $N+1$ values $0, 1, \ldots, N$ then the number of possible K-vectors $\{q_k\}$ is at most $(N+1)^K$, which grows with N only polynomially. Since the number of sequences grows exponentially with N, then 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 $\mathcal{P}\{q_k\} \propto \exp[-ND(q|p)]$.

The relative entropy measures how different is the hypothetical distribution q from the true distribution p. Note that $D(p|q)$ is not the difference between entropies (which just measures difference in uncertainties). The relative entropy is not a true 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 our version of it (no matter how some philosophers want us to believe). Yet $D(p|q)$ has important properties of a distance. Since the probability does not exceed unity, the relative entropy is non-negative, it turns into zero only when distributions coincide, that is $p_i = q_i$ for all i.

Mutual information is the 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 $p(y_j) = \sum_i p(x_i, y_j)$: $D(p|q) = S(X) + S(Y) - S(X, Y) = I(X, Y) \geq 0$. It is also monotonic: $I(X, YZ) \geq I(X, Y)$.

If we observe less variables, then the relative entropy is less (property called monotonicity) $D[p(x_i, y_j)|q(x_i, y_j)] \geq D[p(x_i)|q(x_j)]$ where as usual $p(x_i) = \sum_j p(x_i, y_j)$ and $q(x_i) = \sum_j q(x_i, y_j)$. When we observe less 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

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

But when q is a product, D turns into I and we can use (72): $D[p(x_i, y_j, z_k)|q(x_i, y_j, z_k)] = S(X) + S(Y, Z) - S(X, Y, Z)$ so we obtain $S(X, Y) + S(Y, Z) \geq S(Y) + S(X, Y, Z)$, which is called strong sub-additivity.

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 $S(p) \le \bar{l} \le S(p) + 1$.

If i in p_i runs from 1 to M we can introduce $D(p|u) = \log_2 M - S(p)$, 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.

Remarks on the connections to Statistical Physics. The second law of thermodynamics is getting trivial from the perspective of mutual information. We have seen in Section 3.2 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 interparticle correlations, which are measured by the generalized (multi-particle) mutual information $\sum_i S(p_i, q_i) - S(p_1 \ldots p_n, q_1, \ldots q_n) = I(p_1, q_1; \ldots; p_n, q_n)$. Deriving the Boltzmann kinetic equation (44) in Section 3.2, 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 $S(p_1 \ldots p_n, q_1, \ldots q_n)$ 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 3.3: the full N-particle entropy grows because of phase-space mixing and continuous coarse-graining.

The nonnegativity of the relative entropy between an arbitrary distribution $\{p_i\}$ and the Gibbs distribution $q_i = \exp[\beta(F_\beta - E_i)]$ implies that the latter has the lowest possible free energy at a given temperature:

$$D(p|q) = \sum_i p_i \ln(p_i/q_i) = -S(p) + \beta E(p) - \beta F_\beta = \beta[F(p) - F_\beta] \geq 0 \ .$$

Relative entropy allows also to generalize the second law for non-equilibrium processes. In general, entropy can either increase upon evolution towards thermal equilibrium or decrease upon evolution towards a non-equilibrium state, as seen in Section 3.5. However, the relative entropy between the distribution and the steady-state distribution monotonously decreases with time. Also, the conditional entropy between values of any quantity taken at different times, $S(X_{t+\tau}|X_t)$, grows with τ when the latter exceeds the correlation time.

Continuous distributions. For the sake of completeness, we generalize (64) for a continuous distribution by dividing into cells (that is considering a limit of discrete points). Different choices of variables to define equal cells give different definitions of information. It is in such a choice that physics (or other specific knowledge) enters. Physics (quantum mechanics) requires that for Hamiltonian system the equal volumes in phase space contain equal number of states, so the measure is uniform in canonical coordinates; we then write the missing information in terms of the phase space density $\rho(P, Q, t)$, which may also depend on time:

$$I(t) = -\int \rho \ln \rho \, dP dQ \ . \tag{84}$$

It is maximal for the uniform distribution $\rho = 1/\Gamma$, $I = \ln \Gamma$. To avoid trivial errors, always consider $\Gamma \geq 0$.

If the density of the discrete points in the continuous limit is inhomogeneous, say $m(\mathbf{x})$, then the proper generalization is

$$I(t) = -\int \rho(\mathbf{x}) \ln[\rho(\mathbf{x})/m(\mathbf{x})] \, d\mathbf{x} \ .$$

It is invariant with respect to an arbitrary change of variables $\mathbf{x} \to \mathbf{y}(\mathbf{x})$ since $\rho(\mathbf{y})d\mathbf{y} = \rho(\mathbf{x})d\mathbf{x}$ and $m(\mathbf{y})d\mathbf{y} = m(\mathbf{x})d\mathbf{x}$ while (84) was invariant only with respect to canonical transformations (including a time evolution according to a Hamiltonian dynamics) that conserve the element of the phase-space volume. If we introduce the normalized distribution of points $\rho'(\mathbf{x}) = m(\mathbf{x})/\Gamma$, then

$$I(t) = \ln \Gamma - \int \rho(\mathbf{x}) \ln[\rho(\mathbf{x})/\rho'(\mathbf{x})] \, d\mathbf{x} \ . \tag{85}$$

The last term in (85) turns into zero when ρ and ρ' coincide and thus presents some measure of the difference between the distributions.

Since statistics and information is ultimately about counting, it must be discrete. Continuous treatment is just an approximation often convenient for analytic treatment; to avoid infinities and divergencies (like log of zero) it is convenient to work with differentials.

One can also write a continuous version of the mutual information:

$$I(Z,Y)=\int dzdyp(z,y)\ln\left[\frac{p(z|y)}{p(y)}\right]=\int dzdy\,p(z,y)\ln\left[\frac{p(z,y)}{p(z)p(y)}\right] . \qquad (86)$$

Here we used $p(z,y)=p(z|y)p(y)$ - the probability to get y,z is the probability to get y times the probability to get z for this y. Note that (86) is the particular case of multidimensional (85), where one takes $\mathbf{x}=(y,z)$, $\rho'=p(z)p(y)$, that is mutual information measures the difference between the true joint distribution and the distribution taken as if the quantities were statistically independent. It is straightforward to generalize it from the pair to many quantities.

4.5 Distribution from information

So far, we defined entropy and information via the distribution. Now, we want to use the idea of information to get the distribution. Statistical physics is a systematic way of guessing, making use of partial information. How to get the best guess for the probability distribution $\rho(x,t)$, based on the information given as $\langle R_j(x,t)\rangle = r_j$, i.e. as the expectation (mean) values of some dynamical quantities? Our distribution must contain *the whole truth* (i.e. all the given information) and *nothing but the truth* that is it must maximize the missing information, that 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 maximum of

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

we obtain the distribution

$$\rho(x,t)=Z^{-1}\exp\Big[-\sum_j\lambda_jR_j(x,t)\Big] , \qquad (87)$$

where the normalization factor

$$Z(\lambda_i)=\int\exp\Big[-\sum_j\lambda_jR_j(x,t)\Big]\,dx ,$$

can be expressed via the measured quantities by using

$$\frac{\partial \ln Z}{\partial \lambda_i} = -r_i \ . \tag{88}$$

For example, consider 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 (for impurity, think of X-ray with wavenumber k scattering on the lattice) but on average after many attempts we find, say, $\langle \cos(kj) \rangle = 0.3$. 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 $Z(\lambda) \approx n I_0(\lambda)$ where I_0 is the modified Bessel function. Equation $I_0'(\lambda) = 0.3 I_0(\lambda)$ has an approximate solution $\lambda \approx 0.63$.

Note in passing that the set of equations (88) may be self-contradictory or insufficient so that the data do not allow to define the distribution or allow it non-uniquely. If, however, the solution exists then (87) define the missing information $S\{r_i\} = -\sum_j \rho(j) \ln \rho(j)$, which is analogous to thermodynamic entropy as a function of (measurable) macroscopic parameters. It is clear that S have a tendency to increase whenever a constraint is removed (when we measure less quantities R_i).

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

$$S(x|R) = -\int dx dR \, \rho(R) \rho(x|R) \ln \rho(x|R) = -\int dx dR \rho(x,R) \ln \rho(x|R) \,. \tag{89}$$

The conditional entropy quantifies my remaining ignorance about x once I know R. Measurement decreases the entropy of the system by the mutual information (71,72) — that how much information about x one gains:

$$\begin{aligned} S(x|R) - S(x) &= -\int \rho(x|R) \ln \rho(x|R) \, dx dR + \int \rho(x) \ln \rho(x) \, dx \\ &= \int \rho(x,R) \ln[\rho(x,R)/\rho(x)\rho(R)] \, dx dR = S(x,R) - S(R) - S(x) \,. \end{aligned} \tag{90}$$

Assume that our system is in contact with a thermostat having temperature T, which by itself does not mean that it is in thermal equilibrium (as, for instance, a current-carrying conductor). We then can define a free energy $F(\rho) = E - TS(\rho)$. If the measurement does not change energy (like the knowledge in which half of the box the particles is), then the entropy decrease (90) increases the free energy, so that the minimal work to perform such a measurement is $F(\rho(x|R)) - F(\rho(x)) = T[S(x) - S(x|R)]$. We shall consider the energy price of information processing in more detail in Section 4.6.

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. 3.2. 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. The distribution (87) takes the canonical form (38,39) in equilibrium. On the other hand, taking micro-canonical as constant over the constant-energy surface corresponds to the same approach of not adding any additional information to what is known (energy).

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 (87) deviates from equilibrium. Examples are currents, velocity or temperature gradients like considered in kinetics.

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. A distribution that corresponds to the (conditional) maximum of entropy and yet does not describe thermal equilibrium is a universal tool. A beautiful example of using this approach is obtaining the statistical distribution of the ensemble of neurons (Schneidman, Berry, Segev and Bialek, 2006). 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 (87) that is consistent with the two results of the measurement. These models are thus Ising models , and the probability distribution of the neuron signals that maximizes entropy is as follows:

$$\rho(\{\sigma\}) = Z^{-1}\exp\left[\sum_i h_i\sigma_i + \frac{1}{2}\sum_{i<j} J_{ij}\sigma_i\sigma_j\right] \ , \qquad (91)$$

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. One can also measure some multi-cell correlations and check how well they agree with those computed from (91). 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. 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 brain actually develops and constantly modifies its own predictive model of probability needed in particular to accurately evaluate new events for their degree of surprise. The dominance of pairwise interactions means that learning rules based on pairwise correlations could be sufficient to generate nearly optimal internal models for the distribution of codewords in the brain vocabulary, thus allowing the brain to accurately evaluate probabilities.

It is interesting how 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 entropy is expected to turn into zero and it 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 custers 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.

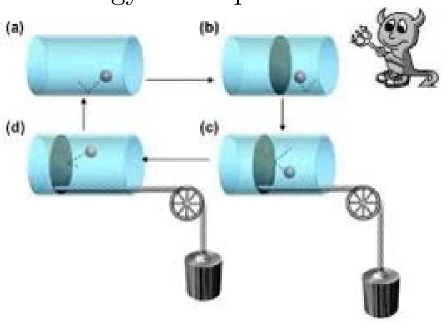
The distribution (91) corresponds to the thermal equilibrium in the auxiliary Ising model, yet it describes the brain activity, which is apparently far from thermal equilibrium (unless the person is brain dead).

4.6 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"

Here we want to adress the relation between information and energy, particularly, find out if information has any energy price. Since energy and entropy (information) have different dimensionalities, we need something to relate them. For example, this can be temperature, which is the derivative of the energy with respect to the entropy. That makes it natural to consider a system in contact with a thermostat. but not necessarily in thermal equilibrium. The Gibbs-Shannon entropy (64) and the mutual information (85,71,90) can be defined for arbitrary distributions. As we mentioned after (4.6), one can then define a free energy for any system in a contact with a thermostat having temperature T as $F(\rho) = E(\rho) - TS(\rho)$, even when the distribution of the system itself is not equilibrium. 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[13] (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: 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:



On the other hand, the law of 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). The same is true for any entropy change by a measurement (90). The entropy change generally is the difference between the entropy of the system $S(A)$ and the entropy of the system interacting with the measuring device $S(A, M)$. More generally, when there is a change in the free energy ΔF_M of the measuring device, the measurement work is

$$W_{meas} \geq T\Delta S + \Delta F_M = T[S(A) - S(A, M)] + \Delta F_M \ . \qquad (92)$$

If we cannot break the first law of thermodynamics, can we at least break the second one by constructing a perpetuum mobile of the second kind, regularly measuring particle position and using its thermal energy to do work? To make a full thermodynamic cycle of our demonic engine, we need to return the demon's

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

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 $S(M)$ than we must spend on erasing it to return the system to the initial state (to make a full cycle). More generally, we can lower the work at the price of cooling the measuring device:

$$W_{eras} \geq TS(M) - \Delta F_M \ . \tag{93}$$

Together, the energy price of the cycle is the temperature times what was defined in the Section 4.3 as *the mutual information*:

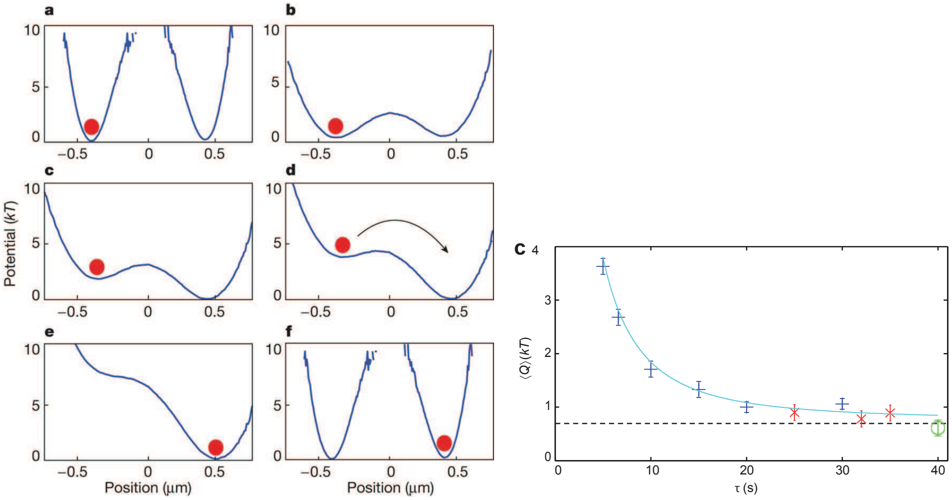
$$W_{eras} + W_{meas} \geq T[S(A) + S(M) - S(A, M)] = TI \ . \tag{94}$$

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 (94) 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. The relations (92,93,94) are versions of the second law of thermodynamics, in which information content and thermodynamic variables are treated on an equal footing.

Landauers principle not only exorcizes Maxwells demon, but also imposes the fundamental physical limit on irreversible computation. Indeed, when a computer does logically irreversible operation (which does not have a single-valued inverse) the information is erased and heat must be generated. Any Boolean function that maps several input states onto the same output state, such as AND, NAND, OR and XOR, is therefore logically irreversible. 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.

Take-home lesson: 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 we can make heat release arbitrarily small making the process slower, apparently because there is no information processing involved.

Experiment. Despite its fundamental importance for information theory and computer science, the erasure bound has not been verified experimentally until recently, the main obstacle being the difficulty of doing single-particle experiments in the low-dissipation regime (dissipation in present-day silicon-based computers still exceeds the Landauer limit by a factor $10^2 \div 10^3$ but goes down fast). The experiment realized erasure of a bit by treating colloidal particle in a double-well potential as a generic model of a one-bit memory (Berut et al, Nature 2012; Jun, Gavrilov, Bechhoefer, PRL 2014). The initial entropy of the system is thus $\ln 2$. The procedure is to put the particle into the right well irrespective of its initial position, see Figure below. It is done by first lowering the barrier height (Fig. b) and then applying a tilting force that brings the particle into the right well (Fig ce). Finally, the barrier is increased to its initial value (Fig f). At the end of this reset operation, the information initially contained in the memory has been erased and the entropy is zero.



The heat/work was determined by experimentally observing the particle trajectory $x(t)$ and computing the integral of the power using the known potential $U(x,t)$:

$$W = Q(\tau) = -\int_0^\tau \dot{x}(t)\frac{\partial U(x,t)}{\partial x}\, dt\ . \tag{95}$$

This heat was averaged over 600 realizations. According to the second law of thermodynamics,

$$\langle Q\rangle \geq -T\Delta S = T\ln 2\ . \tag{96}$$

One can see in the right panel of the figure above how the limit is approached as the duration of the process increases. We shall return to the Brownian particle in a potential in Section 7.3 where we present a generalization of (92,93).

5 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.

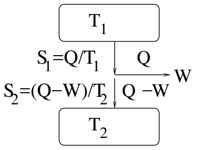
5.1 Flies and spies

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 to survive, unless one is a physicist and knows that energy is conserved and cannot be consumed. All the energy, absorbed by plants from sunlight and by us from food, is emitted as heat. Instead, we consume information and generate entropy by intercepting entropy flows to high-entropy body heat from low-entropy energy sources — just think how much information was processed to squeeze 500 kkal into 100 grams of a chocolate, and you enjoy it even more.

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 (8), 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:

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

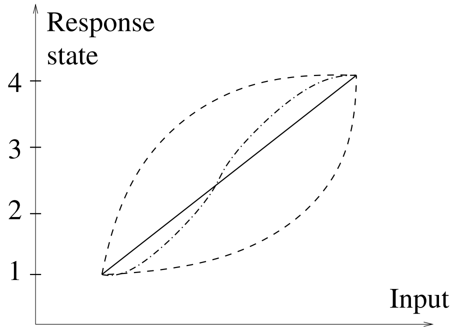


When $T_1 \to T_2$, the information processing is getting prohibitively ineffective, just like the thermal engine. In the other limit, $T_1 \gg T_2$, one can neglect the entropy change on the source, and we have $Q = T_2 \Delta S + W$. Hot Sun is indeed a low-entropy source.

Let us now estimate our rate of information processing and entropy production. A 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).

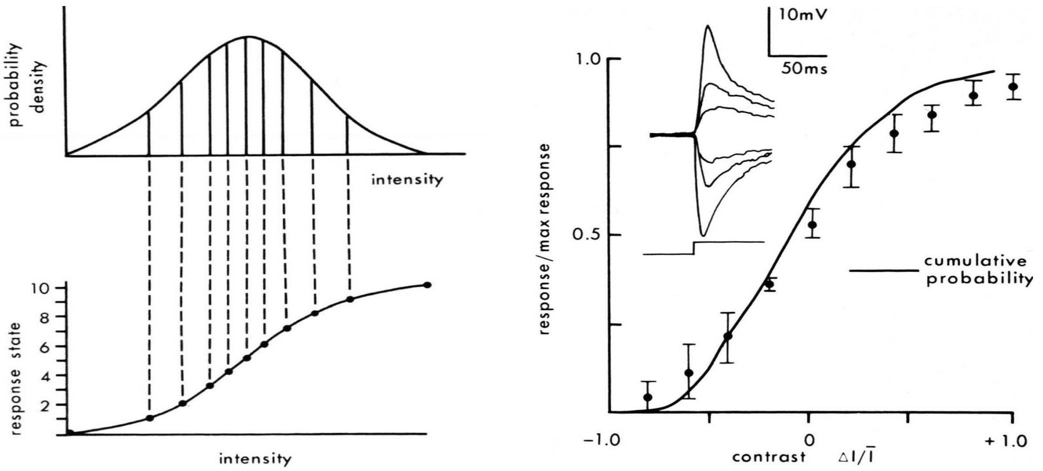
We now discuss how such beings actually process information.

Maximizing capacity. Imagine yourself on the day five of Creation designing the response function for a sensory system of a living being:



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 weak-intensity inputs the same and amplifies difference in strong signals? The choice depends on the goal. For example, the upper curve was actually chosen (on the day six) for the auditory system of animals and humans: our ear amplifies differences in weak sounds and damp strong ones, sensing loudness as the logarithm of the intensity. That way we better hear whisper of a close one and aren't that frightened by loud threats.

If, however, the goal is to maximize the mean information transfer rate (capacity) at the level of a single neuron/channel, then the response curve (encoding) must be designed by the Creator together with the probability distribution of visual stimuli. That it is indeed so was discovered in probably historically the first application of information theory to the real data in biology (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: $I(x,y) = S(y) - S(y|x)$. We assume that the measurement is noiseless and error-free, so that the conditional entropy $S(y|x)$ is zero. Therefore, according to Section 4.3, we need to maximize the entropy of the output assuming that the input entropy is given. Absent any extra constraints except normalization, the entropy is maximal when $\rho(y)$ is constant. Since $\rho(y)dy = \rho(x)dx = \rho(x)dydx/dy = \rho(x)dy/g'(x)$, then

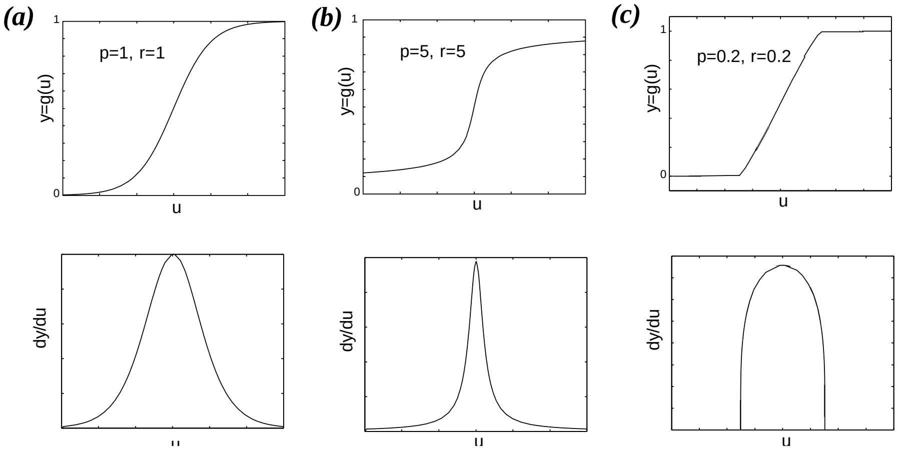
$$\begin{aligned} S(y) &= -\int \rho(x)\ln[\rho(x)/g'(x)]\,dx = S(x) + \langle \ln[g'(x)]\rangle\,, \qquad (97)\\ \frac{\delta S}{\delta g} &= \frac{\partial}{\partial x}\frac{\rho}{g'(x)} = 0 \quad\Rightarrow\quad g'(x) = \rho(x)\ , \end{aligned}$$

as in the Figure. Since the probability is positive, the response function $y = g(x)$ is always monotonic i.e. invertible. Note that 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. 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 nonzero but independent of the form of the response function $y = g(x)$.

This approach turns out quite useful in image and speech recognition by computers using unsupervised learning. That can be done by considering the "training set" of x-s to approximate the density $\rho(x)$. We then choose 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:

$$\Delta w \propto \frac{\partial}{\partial w} \ln \left(\frac{\partial g(x,w)}{\partial x} \right) \ . \tag{98}$$

For example, the form of the response function $y(u)$ popular for its flexibility is the two-parametric asymmetric generalized logistic function defined implicitly by $dy/du = y^p(1-y)^r$. Symmetric examples are given in the Figure, choosing $p \neq r$ one can work with skewed distributions as well.



After choosing the form of g, we need to properly center and normalize the output signal $u = wx + w_0$. Using (98) one trains computer to optimize the data processing with respect to all the parameters.

Let us now pass from a single channel to N inputs and outputs (neurons/channels). Consider a network with an input vector $\mathbf{x} = (x_1, \ldots, x_N)$ which is transformed into the output vector $\mathbf{y}(\mathbf{x})$ monotonically, that is $\det[\partial y_i/\partial x_k] \neq 0$. The multivariate probability density function of y is as follows:

$$\rho(\mathbf{y}) = \frac{\rho(\mathbf{x})}{\det[\partial y_i/\partial x_k]} \, , \tag{99}$$

Making it flat (distribute outputs uniformly) for maximal capacity is not straightforward now. 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 (97), which maximizes its entropy making the respective probability flat. For a good review and specific applications to visual sensory processing see Atick 1992.

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. Yet one can imagine that the information is encoded by the time delays between subsequent pulses. 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 is 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 the home exercise, you shall compare the two schemes. 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).

Minimizing correlation between components. Finding least correlated components can be a practical first step in maximizing capacity. Note how to *maximize* the mutual information between input and output, we *minimize* the mutual information between the components of the output. This is particularly true for natural signals where most redundancy comes from strong correlations (like that of the neighboring pixels in visuals). In addition, 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 spies: 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 also illustrated in the Figure below (from 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 (99) 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 that applies some variant of (98) to minimize mutual information. What was described is a single-layer processing. More powerful are multi-layer nonlinear schemes, where computing determinants and formulating learning rules is more complicated.

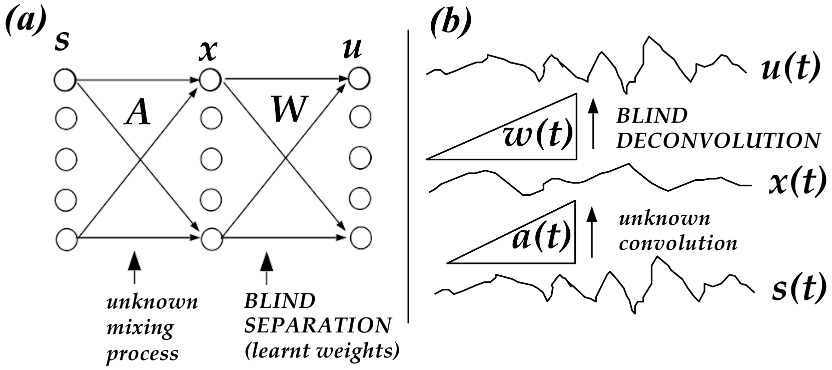


Figure 3: Network architectures for (a) blind separation of 5 mixed signals, and (b) blind deconvolution of a single signal.

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

$$E = \sum_i S(y_i) - \beta[S(\mathbf{y}) - S(\mathbf{x})] = \sum_i S(y_i) - \beta \ln \det[\partial y_i/\partial x_k] \ . \qquad (100)$$

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. Similar minimization procedures will be considered in the next Section.

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. Sometimes it is called *infomax* principle, the specific technique is called independent component analysis (ICA). 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.

5.2 Rate Distortion and Information Bottleneck

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, description of a real number requires infinite number of bits. Representation of a continuous input B by a finite discrete rate of the output encoding 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 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 — because minimizing variance minimizes the entropy of a Gaussian distribution. 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. To minimize squared error, we encode positive/negative values by $A = \pm\sigma\sqrt{2/\pi}$, which corresponds to

$$\mathcal{D}(1) = (2\pi)^{-1/2} \int_0^\infty \left(B - \sigma\sqrt{2/\pi}\right)^2 \exp[-B^2/2\sigma^2] \frac{dB}{\sigma} = \sigma^2/4 \ .$$

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 $P(B|A)$, such that the distortion does not exceed $\mathcal{D}$. 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 a Gaussian input with $\langle B^2\rangle = \sigma^2$ plus (imaginary) Gaussian channel with the variance $\langle (B-A)^2\rangle = \mathcal{D}$, and the minimal rate is given by (77):

$$\begin{aligned} R(\mathcal{D}) &= I(A,B) = S(B) - S(B|A) = S(B) - S(B-A|A) \geq S(B) - S(B-A) \\ &= \frac{1}{2}\log_2(2\pi e\sigma^2) - \frac{1}{2}\log_2(2\pi e\mathcal{D}) = \frac{1}{2}\log_2\frac{\sigma^2}{\mathcal{D}}\ . \end{aligned} \tag{101}$$

It turns into zero for $\mathcal{D} = \sigma^2$ and goes to infinity for $\mathcal{D} \to 0$. Presenting it as $\mathcal{D}(R) = \sigma^2 2^{-2R}$ we see that every bit of description reduces distortion by a factor of 4.

One can show that the rate distortion function $R(\mathcal{D})$ is monotonous and convex for all systems. In solving practical problems, it is usually found as a solution of the variational problem, where one needs to find normalized $P(B|A)$ which minimizes the functional

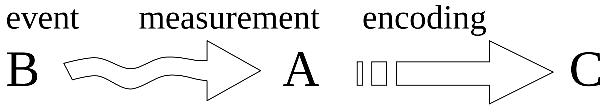
$$F = I + \beta\mathcal{D} = \sum_{ij} P(B_j|A_i)P(A_i)\left\{\ln\frac{P(B_j|A_i)}{P(B_j)} + \beta d(A_i,B_j)\right\}\ . \tag{102}$$

After variation with respect to $P(B_j|A_i)$ and accounting for the identity $P(B_j) = \sum_i P(B_j|A_i)P(A_i)$ we obtain

$$P(B_j|A_i) = \frac{P(B_j)}{Z(A_i,\beta)}e^{-\beta d(A_i,B_j)}\ , \tag{103}$$

where the partition function $Z(A_i,\beta) = \sum_j P(B_j)e^{-\beta d(A_i,B_j)}$ is the normalization factor. Immediate physical analogy is that this 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. 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 is 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. Mutual information is the main tool in the theory of (image, voice, pattern) recognition and AI.



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). 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, targeted at characterizing the tradeoff between information preservation (accuracy of relevant predictions) and compression. Here one looks for the minimum of the functional

$$I(C,A) - \beta I(C,B) \ . \tag{104}$$

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 4.3, 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 (104). 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 in the previous section). 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 (103), 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 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, variation of (104) with respect to the conditional probability now gives the equation (rather than an explicit expression):

$$P(C_j|A_i) = \frac{P(C_j)}{Z(A_i, \beta)} \exp\left[-\beta \sum_k P(B_k|A_i) \log \frac{P(B_k|A_i)}{P(B_k|C_j)}\right] , \qquad (105)$$

The conditional probability in the exponent is given by the Bayes' rule

$$P(B_k|C_j) = \frac{1}{P(C_j)} \sum_i P(A_i) P(B_k|A_i) P(C_j|A_i) , \qquad (106)$$

Technically, this system of equations is usually solved by iterations, for instance, via deep learning (which is one of the most successful paradigms for unsupervised learning to emerge over the last 15 years). Doing compression procedure many times, $A \to C_1 \to C_2 \ldots$ 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 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.

5.3 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 $1 - p$. Betting all your money on the more probable side and retiring after one bet is not an option for two fundamental reasons: you haven't saved enough and you don't like a nonzero probability to loose it all. To have a steady income and an average growth you want to play the game many times. To avoid loosing it all, 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, so we choose the second option and 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 $(2f)^n[2(1-f)]^{N-n} = \exp(N\Lambda)$, where the rate is

$$\Lambda(f) = \ln 2 + \frac{n}{N} \ln f + \left(1 - \frac{n}{N}\right) \ln(1 - f) \ . \tag{107}$$

As $N \to \infty$ we approach the mean rate, which is $\lambda = \ln 2 + p \ln f + (1-p)\ln(1-f)$. 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 with respect to f you find that the maximal growth rate corresponds to $f = p$ (proportional gambling) and equals to

$$\lambda(p) = \ln 2 + p \ln p + (1-p)\ln(1-p) = S(u) - S(p) \,, \tag{108}$$

where we denoted the entropy of the uniform distribution $S(u) = \ln 2$. 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. It is straightforward to generalize that for gambling on horse races, where many outcomes have different probabilities p_i and payoffs g_i. Maximizing $\sum p_i \ln(f_i g_i)$ we find $f_i = p_i$ independent of g_i, so that

$$\lambda(p, g) = \sum_i p_i \ln(p_i g_i) \ . \tag{109}$$

Here you have a formidable opponent - the track operator, who actually sets the payoffs. Knowing the probabilities, the operator can set the payoffs, $g_i = 1/p_i$, to make the game fair and your rate zero. More likely is that the operator cannot resist the temptation to set the payoffs a bit lower to make your λ negative and relieve you of your money. 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

$$\lambda(p,q) = -\ln Z + \sum_i p_i \ln(p_i/q_i) = -\ln Z + D(p|q) \ . \tag{110}$$

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. Nobody's perfect so maybe 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, (110) 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.

To feel less smug, note that bacteria follow the same strategy without ever taking this or other course on statistical physics. Indeed, analogously to coin flipping, bacteria are often face the choice between growing fast but being vulnerable to antibiotic or grow slow but being resistant. They then use proportional gambling to allocate respective fractions of populations to different choices. There could be several lifestyle choices, which is analogous to horse racing problem (called phenotype switching in this case).

Bacteria, as well as gamblers, face the problem we haven't mentioned yet: acquiring information, needed for proportional gambling, has its own cost. One then looks for a tradeoff between maximizing growth and minimizing information cost. Assume that the environment is characterized by the parameter A, say, the concentration of a nutrient. The internal state of the bacteria is characterized by another parameter B, which can be the amount of enzyme needed to metabolize the nutrient. The growth rate is then the function of these two parameters $r(A,B)$. We are looking for the conditional probability $P(B|A)$, which determines the mutual information between the external world and the internal state:

$$I(A,B) = \int dA\, P(A) \int dB P(B|A) \log_2 \frac{P(B|A)}{P(B)} \ . \tag{111}$$

To decrease the cost aI of acquiring this information, we wish to let $P(B|A)$ closer to $P(B)$. Yet we also wish to maintain the average growth rate

$$\lambda = \int dA P(A) \int dB\, P(B|A) r(A,B) \ . \tag{112}$$

Therefore, we look for the maximum of the functional $F = \lambda - aI$, which gives similarly to (102,103)

$$P(B|A) = \frac{P(B)}{Z(A,\beta)} e^{\beta r(A,B)} \ , \tag{113}$$

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. If information is cheaper, (113) 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.

Economic activity of humans is not completely reducible to gambling and its essence understood much less. Therefore, when you earn enough money, it may be a good time to start thinking about the nature of money itself. These days, when most of it is in bits, it is clear to everyone 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 next crisis arrives. Yet in payments money behaves as energy, satisfying the conservation law. I have a feeling that we need a new concept for describing money, which has properties of both entropy and energy.

Concluding this Chapter, which is essentially a long list of representation examples of various efficiency, mention briefly the standard problem of choosing how many fitting parameters to choose. 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. But we know that our data contain noise so it does not make much sense to approximate every fluctuation. Technically, we need to minimize the mutual information of the representation, which would consist of two parts: $ND + Ks$. Here the first term comes from an imperfect data fitting, so it contains the relative entropy D between our hypothetical distribution and the true one, while the second term is the entropy related to our K degrees of freedom. For not very large N, increasing K decreases D, so it is a competition between two terms. When we obtain more data and N is getting large, the value of K, which gives a minimum, usually saturates.

6 Renormalization group and entanglement entropy

6.1 Analogy between quantum mechanics and statistical physics

Many aspects of quantum world are bizarre and have no classical analog. And yet there are certain technical similarities between descriptions of quantum randomness and thermal noise, related to the necessity of summing over different possibilities. One can see one such similarity using the formalism of the path integral, where one sums over different trajectories, — that will be briefly discussed in the Section 7.1 below. Here we describe the similarity in using the formalism of the transfer matrix for the systems with nearest neighbor interaction, which we shall also need in the next subsection. Indeed, in a nutshell, quantum mechanics is done by specifying two sets of states $|q\rangle$ and $\langle p|$, which has ortho-normality and completeness: $\langle p|q\rangle = \delta_{qp}$ and $\sum_q |q\rangle\langle q| = 1$. Physical quantities are represented by operators, and measurement corresponds to taking a trace of the operator over the set of states: trace $P{=}\sum_q \langle q|P|q\rangle$. One special operator, called Hamiltonian $\mathcal{H}$, determines the temporal evolution of all other operators according to $P(t) = \exp(i\mathcal{H}t)P(0)\exp(-i\mathcal{H}t)$. The operator $T(t) = \exp(i\mathcal{H}t)$ is called time translation operator or evolution operator. The quantum-mechanical average of any operator Q is calculated as a trace with the evolution operator normalized by the trace of the evolution operator:

$$\langle Q\rangle = \frac{\mathrm{trace}\, T(t)Q}{Z(t)}\,, \qquad Z(t) = \mathrm{trace}\, T(t) = \sum_a e^{-itE_a}\,. \tag{114}$$

The normalization factor is naturally to call the partition function, all the more if we formally consider it for an imaginary time $t = i\beta$

$$Z(\beta) = \mathrm{trace}\, T(i\beta) = \sum_a e^{-\beta E_a}\,. \tag{115}$$

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 (115) are just expectation values in the ground state.

That quantum mechanical description can be compared with the transfer-matrix description of the Ising model (which was formulated by Lenz in 1920 and solved in one dimension by his student Ising in 1925). It deals with the discrete spin variable $\sigma_i = \pm 1$ at every lattice site. The energy includes interaction with the external field and between nearest neighbors (n.n.):

$$\mathcal{H} = \frac{J}{2}\sum_{i=1}^{N-1}(1-\sigma_i\sigma_{i+1}) \ . \tag{116}$$

It is better to think not about spins but about the links between spins. Starting from the first spin, the state of the chain can be defined by saying whether the next one is parallel to the previous one or not. If the next spin is opposite it gives the energy J and if it is parallel the energy is zero. There are $N-1$ links. The partition function is that of the $N-1$ two-level systems:

$$Z = 2[1+\exp(-\beta J)]^{N-1} \ . \tag{117}$$

Here 2 because there are two possible orientations of the first spin.

To avoid considering the open ends of the chain (which must be irrelevant in the thermodynamic limit), we 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:

$$\begin{aligned} Z &= \sum_{\{\sigma_i\}} \exp\left[-\frac{\beta J}{2}\sum_{i=1}^{N-1}(1-\sigma_i\sigma_{i+1})\right] && (118)\\ &= \sum_{\{\sigma_i\}} \prod_{i=1}^{N-1} \exp\left[-\frac{\beta J}{2}(1-\sigma_i\sigma_{i+1})\right] && (119)\end{aligned}$$

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.

$$T = \begin{pmatrix} T_{1,1} & T_{1,-1} \\ T_{-1,1} & T_{-1,-1} \end{pmatrix} \tag{120}$$

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:

$$Z = \sum_{\{\sigma_i\}} T_{\sigma_1\sigma_2}T_{\sigma_2\sigma_3}\ldots T_{\sigma_N\sigma_1} = \sum_{\sigma_1=\pm 1}\langle\sigma_1|\hat{T}^{N-1}|\sigma_1\rangle = \mathrm{trace}\, T^{N-1} \ . \tag{121}$$

The eigenvalues λ_1, λ_2 of T are given by

$$\lambda_{1,2} = 1 \pm e^{-\beta J} \ . \tag{122}$$

The trace is the sum of the eigenvalues

$$Z = \lambda_1^{N-1} + \lambda_2^{N-1} \, . \tag{123}$$

Therefore

$$F = -T \log(\lambda_1^{N-1} + \lambda_2^{N-1}) = -T \Bigg[(N-1)\log(\lambda_1) \\ + \log\left(1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{N-1}\right)\Bigg] \to -NT\log\lambda_1 \quad \text{as} \quad N \to \infty \tag{124}$$

Note that the partition functions (123) and (117) give the same free energies only at the thermodynamics limit when a ring is indistinguishable from a chain with open ends.

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.

6.2 Renormalization group and information loss

Statistical physics in general is about lack of information. 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. Most often we loose information about microscopic properties. We can do that by averaging over small-scale fluctuations in a procedure called coarse-graining. A general formalism which describes how to make a coarse-graining to keep only most salient features in the description 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 dramatic 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 distribution. Whole families (universality classes) of different systems described by different distribution flow under RG transformation to the same fixed point i.e. have the same large-scale behaviour.

Renormalization group - brief reminder. Let us first describe the simplest case of RG, where we 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: $z_i = (x_{2i-1} + x_{2i})/\sqrt{2}$. Since summing doubles the variance we divided by $\sqrt{2}$. The new random variables each has the following distribution: $\rho'(z) = \sqrt{2} \int dxdy\rho(x)\rho(y)\delta(x+y-z\sqrt{2})$. The distribution which does not change upon such procedure is called fixed point:

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

Since this is a convolution equation, the simplest is to solve it by the Fourier transform, $\rho(k) = \int \rho(x)e^{ikx}dx$, which gives $\rho(k\sqrt{2}) = \rho^2(k)$. The solution is $\rho(k) \sim e^{-k^2}$ and $\rho(x) = (2\pi)^{-1/2}e^{-x^2/2}$. We thus proved the central limit theorem and have shown that the Gaussian distribution is fixed point of repetitive summation and re-scaling of random variables.

Let us now present a physical model in a discrete version - block spin transformation. For Ising model, to eliminate small-scale degrees of freedom, we divide all the spins into groups (blocks) with the side k so that there are k^d spins in every block (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 phenomena very near critical point can be described equally well in terms of block spins with the energy of the same form as original, $E' = -h' \sum_i \sigma'_i - J'/2 \sum_{ij} \sigma'_i \sigma'_j$, but with different parameters J' and h'.

Let us demonstrate how it works using 1d Ising model with $h = 0$ and $J/2T \equiv K$. The partition function is now $Z(K) = 2(2\cosh K)^{N-1}$ for the chain with open ends and $Z(K) = (2\cosh K)^{N-1} + (2\sinh K)^{N-1}$ for the ring, yet the renormalization group has the same action. Let us transform the partition function $\sum_{\{\sigma\}} \exp\Big[K \sum_i \sigma_i \sigma_{i+1}\Big]$ by the procedure (called decimation[14]) of eliminating degrees of freedom by ascribing (undemocratically) to 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 (assuming a chain of spins, i.e

[14] the term initially meant putting to death every tenth soldier of a Roman army regiment that run from a battlefield.

$\sigma_{i+N}=\sigma_i$):

$$\sum_{\sigma_3,\sigma_4=\pm 1}\exp[K(\sigma_1'\sigma_3+\sigma_3\sigma_4+\sigma_4\sigma_2')]$$

$$=\cosh^3 K\sum_{\sigma_3,\sigma_4=\pm 1}(1+x\sigma_1'\sigma_3)(1+x\sigma_4\sigma_3)(1+x\sigma_2'\sigma_4)$$

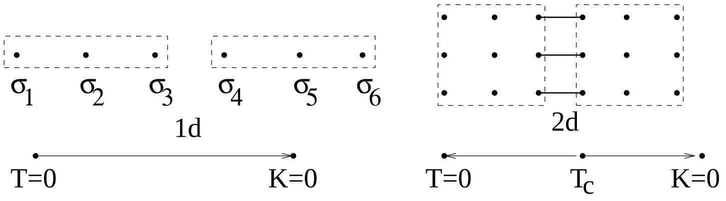
$$=4\cosh^3 K(1+x^3\sigma_1'\sigma_2')=e^{-g(K)}\cosh K'(1+x'\sigma_1'\sigma_2')\,, \tag{125}$$

$$g(K)=\ln\left(\frac{\cosh K'}{4\cosh^3 K}\right)\,. \tag{126}$$

The expression (125) 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 (126) 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 calculation of any spin correlation function. Yet the renormalization of the constant, $K\to K'$, influences the correlation functions. 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 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 large-scale degrees of freedom are described by the partition function where the effective temperature is high so the system is in a paramagnetic state in agreement with the general argument on impossibility of long-range order in one-dimensional systems with short-range interaction. 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.



Similarly, we may sum over every second spin:

$$\sum_{\sigma_{2k}=\pm 1}\exp[K(\sigma_{2k-1}\sigma_{2k}+\sigma_{2k}\sigma_{2k+1}]=2\cosh[K(\sigma_{2k-1}+\sigma_{2k+1})]\ ,\ . \tag{127}$$

which gives the recursive relation $e^{2K'} = \cosh 2K$.

It is done similarly in a continuous case. Every stage of the renormalization group consists of three steps: coarse-grain, re-scale, re-normalize. Thee first step is to decrease the resolution by increasing the minimum length scale from the microscopic scale a to ba where $b > 1$. This is achieved by integrating out small-scale fluctuations which we denote ϕ. The result is a renormalization of the probability distribution, which is now expressed in terms of a coarse-grained field η, whose profile smoother than the original and fluctuate less. In other words, the coarse-grained "picture" is grainier than the original and has less contrast. The original resolution can be restored by the second step called re-scaling: decreasing all length scales by a factor b. To restore the contrast, renormalizes the field, multiplying by b-dependent factor.

One starts from the Gibbs distribution that contains microscopic Hamiltonian: $\exp[\beta(F - \mathcal{H})]$. After averaging it over microscopic degrees of freedom *at a given large-scale field* $\eta(\mathbf{r})$, we obtain macroscopic probability distribution $\exp[\beta(F - \mathcal{F})]$. As we average over the small-scale degrees of freedom we incur an information loss $S(\eta, \phi) - S(\eta) > 0$, which was defined in (89) as the *conditional* entropy - the average entropy of the fast variables for a given configuration of slow variables: $S(\phi|\eta) = S(\eta, \phi) - S(\eta)$. One can think of ϕ as an input of a channel while η is the signal received. The entropy of the remaining variables decreases upon the renormalization group flow. That happens even when the RG flow is towards a high-temperature fixed point, where the system is getting disordered. That means that this measure of the information loss is not very informative about the flow in the space of coupling constants, it just measures the overall loss in the degrees of freedom. We expect, however, the entropy per degree of freedom (or lattice site) to increase, reflecting irreversibility of the transformation. Another extensive (that is also not very useful) information characteristic is the relative entropy between the distribution at any point in the RG flow and its fixed point value. Such relative entropy monotonously decreases upon the flow. The last natural step on this road is to subtract the information lost from the total information and obtain the mutual information $I(\eta, \phi) = S(\phi) - S(\phi|\eta) = S(\eta) + S(\phi) - S(\eta, \phi)$, compare with (90). It shows how much information about the eliminated variables is still stored in the renormalized effective action. Such quantity is also of much interest for computer modeling of systems with many degrees of freedom, which is usually done for a subset of modes. The mutual information is identically zero without interaction; multi-mode mutual information $I = \sum_k S_k - S\{\eta_k\}$ is the relative entropy $D(p|q)$ between the true distribution $p\{\eta_k\}$ and the distribution of non-interactive modes $q = \prod_k q(\eta_k)$. Eliminating short modes generally decreases I, but re-scaling and renormalization may increase it. This is apparently 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.

For the decimation procedure, it is useful to 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 site $h(K) = S(K,N)/N$. 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 MI: $I = 2S(N/2, K') - S(N,K) = N[h(K') - h(K)] \geq 0$. Unfortunately, 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, on the other hand, the next iteration of decimation cannot be performed.

In a finite system with short-range correlations, the entropy for large N is generally as follows:

$$S(N) = hN + C\,, \qquad I = N[h(K') - h(K)] + 2C' - C\ . \tag{128}$$

In the critical point the extensive terms in MI cancel and $I = C > 0$. One can explain positivity of C saying that a finite system appears more random than it is.

Mutual information also naturally appears in the description of the information flows in the real space. Points in space are generally correlated even when Fourier modes are not. Indeed, for a set of non-interacting waves, the total entropy is a direct sum: $S = \sum_k S_k = \sum_k \log \pi e n_k$. The pair correlation function between two points separated by r space is $f(0) = \sum_k n_k e^{\imath kr}$, which is delta-function only for k-independent n_k. By the same token, the mutual information between these points is $-\log[1 - f(r)/f(0)]$. For example, in thermal equilibrium, energy equipartition gives $n_k = T/\hbar\omega_k$, and non-constancy of ω_k (and thus n_k) encodes information about the interaction between spatial sites. For 1d N-chain, it pays to consider breaking it 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, also called *excess entropy*, is equal to C from (128).

For example, the partition function of the 1d Ising chain is $Z(K) = 2(2\cosh K)^{N-1}$. Remind that the entropy is expressed via the partition as follows:

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

So that we obtain for the chain: $h = \ln(2\cosh K) - K\tanh K$ 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, i.e. of the largest eigenvalue of the transfer matrix $T_{s,s'} = \exp(Kss')$. Indeed, $q = 2$ at 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 the how much information one needs to specify per one degree of freedom (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. For the ring, $C = K \tanh K - \ln(2\cosh K)$, but then the ring is not separated by one cut and I does not make sense.

It is important to stress the difference between the mutual information between points and regions in space and modes in Fourier space. For non-interacting modes, the MI is zero, but the MI between the points in physical space is not. Indeed, if the occupation numbers of the modes are n_k, then the total entropy of the respective Gaussian distribution is $\sum_k \ln(e\pi n_k)$. The variance in every space point is $N^{-1}\sum_{k=1}^N n_k$, so that the respective entropy is $\ln\Big(e\pi N^{-1}\sum_{k=1}^N n_k\Big)$. The sum of the entropies of the points is $N\ln\Big(e\pi N^{-1}\sum_{k=1}^N n_k\Big)$, while the total entropy of the system is the same for both representations. The difference between the sum of the point entropies and the total entropy is the N-point mutual information, $I(x_1,\dots,x_N) = N\ln\Big(e\pi N^{-1}\sum_{k=1}^N n_k\Big) - \sum_k \ln(e\pi n_k)$, which is positive because of convexity, whenever n_k are not all the same. For example, in thermal equilibrium, when $n_k = T/\omega_k$, the dependence of the frequency on the wave number encodes the information about interaction between points in space, which is measured by $I(x_1,\dots,x_N)$. Only when the spectrum of occupation numbers is flat, different spatial points are uncorrelated.)

In two dimensions, consideration is much more involved, and the reader deserves a warning: the discussion is getting quite superficial. Breaking a single bond now does not really separate, so in a plane H we 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 in a plane 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. As we shall describe in the next subsection, in a quantum case, we take the whole system in a ground state, trace out 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 $S(L) = -Tr\rho_L \log \rho_L$

For long lines in short-correlated systems, that information 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 not linear but logarithmic at criticality (when we have fluctuations of all scales and the correlation radius is infinite). Then one defines the function $c(r) = rdS_L(r)/dr$ (to cancel non-universal terms depending on the microscopic details) and shows that this is 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, which is again some measure of relevant degrees of freedom which respond to boundary perturbations. Moment reflection is enough to understand how difficult it must be to introduce proper intensive measure of information flow in dimensions higher than two, and yet recently, such function was also found for $d = 4$ (Komargodsky and Schwimmer 2011) and $Dd = 3$ (Klebanov).

6.3 Quantum information

this chapter should have been called
”quantum information theory for the impatient”
J Preskill

Since our world is fundamentally quantum mechanical, it is interesting what that reveals about the nature of information. Most important here are two aspects of a measurement in a quantum world: i) it irreversibly changes the system and this change cannot be made arbitrarily small, and ii) the results are truly random (not because we did not bother to learn more on the system). Apart from that fundamental aspect, interest in quantum information is also pragmatic. Classical systems, including computers, are limited by locality (operations have only local effects) and by the classical fact that systems can be in only one state at the time. However, a quantum system can be in a superposition of many different states at the same time, which means that quantum evolution exhibit interference effects and proceeds in the space of factorially more dimensions than the respective classical system. This is a source of the parallelism of quantum computations. Moreover, 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 computation and then in quantum information theory. Non-surprisingly, it is also based on the notion of entropy, which is similar to classical entropy yet differs in some important ways. Uncertainly and probability exist already in quantum mechanics where our knowledge of the state of the system is complete. On top of that we shall consider quantum statistics due to incomplete knowledge, which is caused by considering subsystems. In this subsection I'll give a mini-introduction to the subject, focusing on entropy. The presentation mostly follows Witten lectures with an occasional borrowing from the Preskill's book at

`https://arxiv.org/pdf/1805.11965.pdf`
`http://www.theory.caltech.edu/~preskill/ph219/chap10_6A.pdf`

Quantum mechanics describes a system by presenting its stationary states and possible jumps between those states. The fundamental statement is that any system can be in a *pure* state ψ_i (like, for instance, eigenstate of a Hamiltonian, which has fixed energy) or in a superposition of states: $\psi = \sum_i \sqrt{p_i}\psi_i$. This 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 (unitarily) without touching or measure it. Measurement is classical, it produces one and only pure state from the initial superposition; immediately repeated measurements will produce the same outcome. However repeated measurement of the identically prepared initial superposition find different states, the state i appears with probability p_i. A property that can be measured is called an observable and is described a self-adjoint operator.

There is an uncertainty already in a pure state of an isolated quantum system. *Quantum uncertainty principle* states that if the operators of two observables are non-commuting, then the values of the observables cannot fixed at the same time. In particular, momentum and coordinate are such pair, so that we cannot describe quantum states as points in the phase space. Indeed, quantum entanglement is ultimately related to the fact that one cannot localize quantum states in a finite region — if coordinates are zero somewhere, then the momenta are not. The uncertainty restriction was formulated initially by Heisenberg in terms of variances: for a free quantum particle, the variances of the coordinate and the momentum along the same direction satisfy the inequality $\sqrt{\sigma_p \sigma_q} \geq \hbar/2$. However, variances are sufficient characteristics of uncertainty only for Gaussian distributions. Indeed, 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 respective 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 we measure a quantum state ρ by projecting onto the orthonormal basis $\{|x\rangle\}$, the outcomes define a classical probability distribution $p(x) =$

$\langle x|\rho|x\rangle$, which is a probability vector whose entries are the diagonal elements of ρ in the x-basis. The Shannon entropy $S(X)$ quantifies how uncertain we are about the outcome before we perform the measurement. If $|z\rangle$ is another orthonormal basis, there is a corresponding classical probability distribution of outcomes when we measure the same state ρ in the z-basis. If the operators projecting on x, z do not commute, the two bases are incompatible, so that there is a tradeoff between our uncertainty about X and about Z, captured by the inequality

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

We say that two different bases $\{|x\rangle\}$, $\{|z\rangle\}$ for a d-dimensional Hilbert space are mutually unbiased if for all $|\langle x_i|z_k\rangle|^2 = 1/d$. 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 example, if a particle is in a definite point, all momentum directions are equally probable. For measurements in two mutually unbiased bases performed on a pure state, the entropic uncertainty relation becomes

$$S(X) + S(Z) \geq \log d\,.$$

Clearly this inequality is tight, as it is saturated by x-basis (or z-basis) states, for which $S(X) = 0$ and $S(Z) = \log d$. In particular, in one dimension $\log d = 0$. The uncertainty increases if ρ is not a pure state but a mixture, characterized by additional $S(\rho)$:

$$S(X) + S(Z) \geq log(1/c) + S(\rho)\,.$$

In reality we always study subsystems, which requires us to pass from quantum mechanics to quantum statistics and brings the fundamental notion of the **density matrix**. Consider a composite system AB, which is in a pure 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 $\bar{O} = \int dxdy\psi^*_{AB}(x,y)\hat{O}(x)\psi_{AB}(x,y)$. For non-interacting subsystems, one has $\psi_{AB}(x,y) = \psi_A(x)\psi_B(y)$ and $\bar{O} = \int dx\psi^*_A(x)\hat{O}(x)\psi_A(x)$, so that one can forget about B. But generally, dependencies on x and y are not factorized, so one ought to characterize A by the so-called density matrix $\rho(x,x') = \int dy\psi^*_{AB}(x',y)\psi_{AB}(x,y)$, so that $\bar{O} = \int dx[\hat{O}(x)\rho(x,x')]_{x'=x}$, where $\hat{O}(x)$ acts only on x and then we put $x' = x$.

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

$$\langle\psi_{AB}|\hat{O}_A \bigotimes 1_B|\psi_{AB}\rangle = \langle\psi_A|\hat{O}_A|\psi_A\rangle\langle\psi_B|1_B|\psi_B\rangle = \langle\psi_A|\hat{O}_A|\psi_A\rangle\ .$$

However, a general pure state ψ_{AB} is not a (tensor) product of pure states, it is what is called entangled: some generic $M \times N$ matrix where N, M dimensionalities of A, B respectively:

$$\left[\begin{array}{ccc} \psi_A^1\psi_B^1 & \dots & \psi_A^1\psi_B^M \\ \dots & & \\ \psi_A^N\psi_B^1 & \dots & \psi_A^N\psi_B^M \end{array}\right] \tag{129}$$

We can make it $N \times N$ diagonal with N positive eigenvalues $\sqrt{p_i}$ and extra $M-N$ rows of zeroes by applying unitary transformation in A and B, that is making $\psi_{AB} \to U\psi_{AB}V$ with $M \times M$ unitary V-matrix and $N \times N$ U-matrix. That is called Schmidt decomposition by orthonormal vectors ψ_A^i, ψ_B^i, which allows us to present the state of AB as a sum of the products:

$$\psi_{AB} = \sum_i \sqrt{p_i}\psi_A^i \bigotimes \psi_B^i \,, \tag{130}$$

If there is more than one term in this sum, we call subsystems A and B entangled. 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

$$\langle\psi_{AB}|\hat{O}_A \bigotimes 1_B|\psi_{AB}\rangle = \sum_{i,j} \sqrt{p_i p_j}\langle\psi_A^i|\hat{O}_A|\psi_A^j\rangle\langle\psi_B^i|1_B|\psi_B^j\rangle$$
$$= \sum_{i,j} \sqrt{p_i p_j}\langle\psi_A^i|\hat{O}_A|\psi_A^j\rangle\delta_{ij} = \sum_i p_i\langle\psi_A^i|\hat{O}_A|\psi_A^i\rangle = \mathrm{Tr}_A\rho_A\hat{O}_A \,,$$

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

$$\rho_A = \sum_i p_i|\psi_A^i\rangle\langle\psi_A^i| \;. \tag{131}$$

It is all we need to describe A. 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 some pure state ψ_{AB} in the extended system 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 the mixed state (or, in other words, an ensemble of states), which must be distinguished from quantum-mechanical superposition. 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 **qubit** — a quantum system having only two states: $|0\rangle$ and $|1\rangle$. The most general state of a qubit A is a superposition of two states, $\psi_A = a\,|0\rangle + b\,|1\rangle$, then any observable is as follows:

$$\langle\psi_A|\hat{O}_A|\psi_A\rangle = |a|^2\langle 0|\hat{O}_A|0\rangle + |b|^2\langle 1|\hat{O}_A|1\rangle + (a^*b + ab^*)\langle 0|\hat{O}_A|1\rangle \ . \tag{132}$$

Normalization requires $|a|^2 + |b|^2 = 1$ and the overall phase does not matter, so a qubit is really characterized by two real numbers. However, 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.

Consider now a pure quantum state of the two-qubit system A, B of the form

$$\psi_{AB} = a\,|00\rangle + b\,|11\rangle \ . \tag{133}$$

A and B are correlated in that state. Now any operator acting on A gives

$$\langle\psi_{AB}|\hat{O}_A \bigotimes 1_B|\psi_{AB}\rangle = (a^*\langle 00| + b^*\langle 11|)\hat{O}_A \bigotimes 1_B|(a|00\rangle + b|11\rangle)$$

$$= |a|^2\langle 0|\hat{O}_A|0\rangle + |b|^2\langle 1|\hat{O}_A|1\rangle\,, \tag{134}$$

$$\rho_A = |a|^2\,|0\rangle\,\langle 0| + |b|^2\,|1\rangle\,\langle 1| = \begin{bmatrix} |a|^2 & 0 \\ 0 & |b|^2 \end{bmatrix} . \tag{135}$$

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 $|1\rangle$. In distinction from (132), the same result (134) is obtained if $\langle 0|\hat{O}_A|1\rangle \neq 0$ due to the orthogonality of B-states. 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.

General 2×2 density matrix is characterized by three real numbers, $\mathbf{P} = (P_1, P_2, P_3)$:

$$\rho(\mathbf{P}) = \frac{1}{2}\begin{bmatrix} 1 - P_1 & P_2 + \imath P_3 \\ P_2 - \imath P_3 & 1 + P_1 \end{bmatrix} . \tag{136}$$

Non-negativity of the eigenvalues requires $P^2 = P_1^2 + P_2^2 + P_3^2 \leq 1$, that is the space of such density matrices is a ball. The boundary $P^2 = 1$ corresponds to zero determinant, that is to eigenvalues 0 and 1 — that means that ρ is a one-dimensional projector and the subsystem is in a pure state.

We call the system *entangled* if its density matrix has more than one nonzero eigenvalue, so that there is more than one term in the sum (131).

The **von Neumann entropy** of a density matrix ρ_A is defined by the formula analogous to the Gibbs-Shannon entropy of a probability distribution:

$$S(\rho_A) = S_A = -\mathrm{Tr}\,\rho_A \log \rho_A \,. \tag{137}$$

It is manifestly 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). It also satisfies concavity (67). 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 $S_A > 0$ can be larger than of the whole system S_{AB} (which is zero for a pure state, for instance). For a quantum system, but not for a classical one, information can be encoded in the correlations among the parts A, B of the system, yet be invisible when we look at one part. 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.

Are thermalization and entropy growth possible for a quantum system which as a whole remains in a pure quantum state? Yes they are! Thermalization and entropy growth take place for any subsystem of a large system if the dynamics is ergodic. Then the system as a whole acts as a thermal reservoir for its subsystems, provided those are small enough. As a quantum subsystem approaches equilibrium, the von Neumann entropy reaches its maximum. Information which is 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.

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. 3.3? 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 $x(t)$ and $p(0)$. Indeed, $\partial x(t)/\partial x(0) = \{x(t), p(0)\} \to \hbar^{-1}[x(t), p(0)]$. Of course, the mean value of this commutator is zero, but its square is not, which brings the concept of a so-called out-of-time-order correlator $\langle x(t)p(0)x(t)p(0)\rangle$ - 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). Respective Lyapunov exponent dimensionally must be temperature divided by $\hbar$ and indeed $2\pi T/\hbar$ was shown to be a universal limit (reached by black holes, for instance).

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, $S(\rho_A) = -|a|^2 \log_2 |a|^2 - |b|^2 \log_2 |b|^2$ for (133,134). The maximum $S(\rho_A) = 1$ is reached when $a = b = 1/\sqrt{2}$, which is called a state of maximal entanglement. Pure state of a qubit can store one bit. The four 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, yet when we trace out B (or A) we wipe out this information: 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.

If a pure state AB was built from non-orthogonal states, taken with probabilities q_i, then the density matrix 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 $\{q_i\}$.

On the other hand, we may measure by projecting ρ_A on the orthogonal set $\{|\psi_A^i\rangle\}$ which is different from the basis of eigenvectors $\{|\phi_A^k\rangle\}$, where $\rho_A = \sum_k p_k |\phi_A^k\rangle\langle\phi_A^k|$ is diagonal. In this case, the outcome i happens with the probability $q'(i) = \langle\psi_A^i|\rho_A|\psi_A^i\rangle = \sum_k p_k D_{ij}$, where $D_{ik} = |\langle\psi_A^i|\phi_A^k\rangle|^2$ 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, $S(q') = S(p) + \sum_{ik} D_{ik} \log D_i k \geq S(p) = S(\rho_A)$, that is such measurements are less predictable.

Similar to the classical conditional entropy (70) and the mutual information (71) we can define

$$S(A|B) = S_{AB} - S_B \,, \quad I(A,B) = S_A + S_B - S_{AB} \;. \tag{138}$$

Quantum I is non-negative as well as classical, but generally is different. For example, when AB is in an entangled pure state then $S_{AB} = 0$, A and B together are perfectly correlated, but separately each one is in a mixed state, so that $S_A = S_B > 0$. 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: $I(A,B) = S_A + S_B - S_{AB} = 2S_A$. Quantum correlations are stronger than classical.

Quantum $S(A|B)$ is even more bizarre; 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 $S(A|B) = -S_B < 0$. 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 $S(A|B)$ is negative?

That negativity allows for a quantum trick of *teleportation* of a quantum state. Imagine that Alice has in her possession a qubit A_0. Alice would like to help Bob create in his lab a qubit in a state identical to A_0. However, she wants to communicate by sending a classical message (sat, over the telephone). If Alice knows the state of her qubit, there is no problem: she tells Bob 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 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.

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

$$\psi_{A_1 B_1} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{A_1 B_1} \,. \tag{139}$$

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 $A_0 A_1$ 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 $A_0 A_1$ 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 $A_0 A_1$ satisfy that property:

$$\frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle)_{A_0 A_1} \,, \quad \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle)_{A_0 A_1} \,. \tag{140}$$

The states are chosen to be entangled, that is having A_0, 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, A_1.

Denote the unknown initial state of the qubit A_0 as $\alpha\,|0\rangle + \beta\,|1\rangle$, then the initial state of $A_0 A_1 B_1$ is

$$\frac{1}{\sqrt{2}}(\alpha\,|000\rangle + \alpha\,|011\rangle + \beta\,|100\rangle + \beta\,|111\rangle)_{A_0 A_1 B_1} \,. \tag{141}$$

Let's say that Alice's measurement, that is the projection on the states (140), reveals that $A_0 A_1$ is in the state

$$\frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)_{A_0 A_1} \,. \tag{142}$$

That means that only the first and the last terms in (141) 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 (138), we symmetrize and purify our problem. Notice that A_1 and B_1 are maximally entangled (come with the same weights), so that $S_B = \log_2 2 = 1$. On the other hand, A_1B_1 is in a pure state so $S_{A_1B_1} = 0$. Let us now add another system R which is maximally entangled with A_0 in a pure state, say

$$\psi_{A_0R} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{A_0R}\,. \tag{143}$$

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 $\psi_{RA_0} \bigotimes \psi_{A_1B_1}$. Since A_0 is maximally entangled with R then also $S_{A_0} = \log_2 2 = 1$ and the same is the entropy of the AB system $S(A_0A_1B_1) = S(A_0) = 1$ since A_1B_1 is a pure state. Therefore, $S(A|B) = S(A_0A_1|B) = S(A_0A_1B_1) - S(B_1) = 0$. One can show that teleportation only possible when $S(A|B)$ 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 $S(A|B) > 0$ 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. 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_{AB} is unchanged but S_B goes up by 1, so $S(A|B) = S_{AB} - S_B$ goes down by 1. So $S(A|B)$, 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.

Quantum communication poses the same natural question: i) 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? 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 of N matrices. If the states are mutually orthogonal then the density matrix is diagonal and 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. The central issue in quantum information theory is that nonorthogonal quantum states cannot be perfectly distinguished, a feature with no classical analog. For non-orthogonal states sending classical information about the probabilities of different states is not optimal since non-orthogonal states are not completely distinguished. For pure states the answer is given by $S(\rho)$ — von Neumann entropy is the number of qubits of quantum information carried per letter of a long message, and it is the best possible rate of quantum information transfer (the only difference from the classical case is that instead of typical sequences we consider typical subspaces). It is easy to see that $S(\rho)$ wont be the answer for mixed states. To give 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 $\rho_0 \bigotimes \rho_0 \bigotimes \ldots$ carries no information. If, however, 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. However, it is now different from von Neumann entropy because every mixed state ρ_k has a nonzero entropy:

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

Now Shannon entropy is less than von Neumann entropy:

$$
S(p) = S(\rho) - \sum_k p_k S(\rho_k) = S\left(\sum_k p_k\rho_k\right) - \sum_k p_k S(\rho_k) \equiv \chi(\rho_k, p_k)\,. \tag{144}
$$

It is believed that this quantity χ (called in quantum communications Holevo information) defines the limiting compression rate in all cases including non-orthogonal mixed states when it does not coincide with $S(p)$. The reason 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, (144) 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.

Early idea of entanglement was conjured in 17 century precisely for communications: 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 idea to Galileo, but the latter naturally wanted to see an experiment first.

Area Law for entanglement entropy and mutual information. Entropy is extensive that is proportional to the system volume or total number of degrees of freedom. We already noticed that extensive (volume) terms cancel in the entanglement entropy and mutual information, which are thus sub-extensive. This is a manifestation of the so-called *area law.* If we define, according to (131,137), the von Neumann entropy S_A of the region A or consider its classical analog, the mutual information between A and its complement H/A, they both are proportional to the boundary area rather than to the volume of A.

In particular, the entanglement entropy is thought to be responsible for the entropy of the black holes which is proportional to the boundary area and not to the volume. 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 represented by a theory which lives in the boundary of that region.

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. For some classes of systems, one can show (Kitaev, Preskill 2006) that in the combination $S_A + S_B + S_C + S_{ABC} - S_{AB} - S_{BC} - S_{AC}$, 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.

Problem-solving session: remind about the block-spin renormalization.

Gaussian distribution as a fix point of the renormalization group.

7 Stochastic processes

In this Section we further reveal connection between quantum fluctuations and thermal noise and present modern generalizations of the second law and fluctuation-dissipation relations. This is best done using the fundamental process of a random walk in different environments. It is interesting both for fundamentals of science and for numerous modern applications related to fluctuations in nano-particles, macro-molecules, stock market prices etc.

7.1 Random walk and diffusion

Consider a particle that can hop randomly to a neighboring cite of 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 $\mathbf{x}$ satisfies the equation

$$P(\mathbf{x}, t+\tau) = \frac{1}{2d}\sum_{i=1}^{d}\left[P(\mathbf{x}+\mathbf{e}_i, t) + P(\mathbf{x}-\mathbf{e}_i, t)\right] \ . \tag{145}$$

The first (painful) way to solve this equation is to turn it into averaging exponents as we always do in statistical physics. This is done using the Fourier transform, $P(\mathbf{x}) = (a/2\pi)^d \int e^{i\mathbf{kx}} P(\mathbf{k})\, d^d k$, which gives

$$P(\mathbf{k}, t+\tau) = \frac{1}{d}\sum_{i=1}^{d} \cos ak_i \, P(\mathbf{k}, t) \ . \tag{146}$$

The initial condition for (145) is $P(\mathbf{x}, 0) = \delta(\mathbf{x})$, which gives $P(\mathbf{k}, 0) = 1$ and $P(\mathbf{k}, t) = \left(d^{-1}\sum_{i=1}^{d}\cos ak_i\right)^{t/\tau}$. That gives the solution in space as an integral

$$P(\mathbf{x}, t) = (a/2\pi)^d \int e^{i\mathbf{kx}} \left(\frac{1}{d}\sum_{i=1}^{d}\cos ak_i\right)^{t/\tau} d^d k \ . \tag{147}$$

We are naturally interested in the continuous limit $a \to 0, \tau \to 0$. If we take $\tau \to 0$ first, the integral tends to zero and if we take $a \to 0$ first, the answer remains delta-function. A non-trivial evolution appears when the lattice constant and the jump time go to zero simultaneously. Consider the cosine expansion,

$$\left(\frac{1}{d}\sum_{i=1}^{d}\cos ak_i\right)^{t/\tau} = \left(1 - a^2k^2/2d + \ldots\right)^{t/\tau} \ ,$$

where $k^2 = \sum_{i=1}^d k_i^2$. The finite answer $\exp(-\kappa t k^2)$ appears only if one takes the limit keeping constant the ratio $\kappa = a^2/2d\tau$. In this limit, the space density of the probability stays finite and is given by the integral:

$$\rho(\mathbf{x},t) = P(\mathbf{x},t)a^{-d} \approx (2\pi)^{-d}\int e^{i\mathbf{kx}-t\kappa k^2}\, d^dk = (4\pi\kappa t)^{-d/2}\exp\left(-\frac{x^2}{4\kappa t}\right) . \quad (148)$$

The second (painless) way to get this answer is to pass to the continuum limit already in the equation (145):

$$P(\mathbf{x},t+\tau) - P(\mathbf{x},t) = \frac{1}{2d}\sum_{i=1}^{d}\left[P(\mathbf{x}+\mathbf{e}_i,t) + P(\mathbf{x}-\mathbf{e}_i,t) - 2P(\mathbf{x},t)\right] . \quad (149)$$

This is a finite difference approximation to the diffusion equation

$$(\partial_t - \kappa\Delta)P(\mathbf{x},t) = 0 . \quad (150)$$

Of course, ρ satisfies the same equation, and (148) is its solution. Note that (148,150) are isotropic and translation invariant while the discrete version respected only cubic symmetries. Also, 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 $\mathbf{j} = -\kappa\nabla\rho$.

Another way to describe it 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 (148), 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, one can define the dimensionality of a geometric object as a relation between its size R and the number N of standard-size elements(with fixed volume or area) needed to cover it. For a line, $N \propto R$, generally $N \propto R^d$. For a random walk, the number of elements is of order of the number of steps, $N \propto t$, while $R \propto x$ so that $d = 2$. 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).

The properties of random walks can be expressed alternatively in terms of sums over different paths. Let us write the transition probability indicating explicitly the origin: $\rho(\mathbf{x},t;0,0)$. Then we can write the convolution identity which simply states that the walker was certainly somewhere at an intermediate time t_1:

$$\rho(\mathbf{x},t;0,0) = \int \rho(\mathbf{x},t;\mathbf{x}_1,t_1)\rho(\mathbf{x}_1,t_1;0,0)\, d\mathbf{x}_1 . \quad (151)$$

We now divide the time interval t into an arbitrary large number of intervals and using (148) we write

$$\begin{aligned}\rho(\mathbf{x},t;0,0) &= \int \Pi_{i=0}^{n} \frac{d\mathbf{x}_{i+1}}{[4\pi\kappa(t_{i+1}-t_i)]^{d/2}} \exp\left[-\frac{(\mathbf{x}_{i+1}-\mathbf{x}_i)^2}{4\kappa(t_{i+1}-t_i)}\right] \\ &\to \int \mathcal{D}\mathbf{x}(t') \exp\left[-\frac{1}{4\kappa}\int_0^t dt'\dot{x}^2(t')\right] \ . \qquad (152)\end{aligned}$$

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 it gives the weight of every trajectory.

Looking at the transition probability (152), one can also see the analogy between the statistical physics of a random walk and quantum mechanics. According to Feynman, for a quantum non-relativistic particle with a mass M in the external potential $U(\mathbf{x})$, 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:

$$T(\mathbf{x},t;0,0) = \int \mathcal{D}\mathbf{x}(t') \exp\left[\frac{i}{\hbar}\int_0^t dt'\left(\frac{M\dot{x}^2}{2} - U(x)\right)\right] \ . \qquad (153)$$

Comparing with (152), 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. Note the action is an integral of the Lagrangian $M\dot{x}^2/2 - U(x)$ rather than Hamiltonian $M\dot{x}^2/2+U(x)$. In quantum theory, one averages over quantum rather than thermal fluctuations, yet the formal structure of the theory is similar.

7.2 Brownian motion

Let us see how the properties of the random walk and diffusion appear a physical system. 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. When the particle M is much larger than the molecular mass m then the particle velocity v is small comparing to the typical velocities of the molecules $v_T \simeq \sqrt{T/m}$. Then one can write the force $\mathbf{f}(\mathbf{p})$ acting on it 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 second term is much smaller than the first one; then one may ask 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. That resistance must be described by the second term, which then may be approximated as $\partial f_i/\partial p_j = -\lambda\delta_{ij}$. 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 $\eta \simeq mnv_T\ell$, where n is the concentration of the molecules. For a slow moving particle, $v \ll v_T\ell/R$, the resistance is given by the Stokes formula

$$\lambda = 6\pi\eta R/M \ . \tag{154}$$

We then obtain

$$\dot{\mathbf{p}} = \mathbf{f} - \lambda\mathbf{p} \ . \tag{155}$$

The solution of the linear equation (155) is

$$\mathbf{p}(t) = \int_{-\infty}^{t} \mathbf{f}(t')e^{\lambda(t'-t)}dt' \ . \tag{156}$$

We must treat the force $\mathbf{f}(t)$ as a random function since we do not track molecules hitting the particle, which makes (155) Langevin equation. 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 λ^{-1}. Since the integration time in (156) is of order λ^{-1} then the condition $\lambda\tau \ll 1$ 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

$$\begin{aligned}\sigma^2 &= \langle p_x^2\rangle = \langle p_y^2\rangle = \langle p_z^2\rangle = \int_0^\infty C(t_1-t_2)e^{-\lambda(t_1+t_2)}dt_1dt_2 \\ &\approx \int_0^\infty e^{-2\lambda t}\,dt \int_{-2t}^{2t} C(t')\,dt' \approx \frac{1}{2\lambda}\int_{-\infty}^{\infty} C(t')\,dt' \ .\end{aligned} \tag{157}$$

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):

$$\lambda = \frac{1}{2TM}\int_{-\infty}^{\infty} C(t')\,dt' \ . \tag{158}$$

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

$$\langle \mathbf{v}(t)\cdot\mathbf{v}(0)\rangle = (3T/M)\exp(-\lambda|t|) \tag{159}$$

which can be obtained from (156). Note that the friction makes velocity correlated on a longer timescale than the force. That gives

$$\langle|\Delta\mathbf{r}|^2(t')\rangle = \int_0^{t'} dt_1 \int_0^{t'} dt_2 \langle \mathbf{v}(t_1)\mathbf{v}(t_2)\rangle = \frac{6T}{M\lambda^2}(\lambda t' + e^{-\lambda t'} - 1)\ .$$

The mean squared distance initially grows quadratically (so-called ballistic regime at $\lambda t' \ll 1$). In the limit of a long time (comparing to the relaxation time λ^{-1} rather than to the force correlation time τ) we have the diffusive growth $\langle(\Delta\mathbf{r})^2\rangle \approx 6Tt'/M\lambda$. Generally $\langle(\Delta\mathbf{r})^2\rangle = 2d\kappa t$ where d is space dimensionality and κ - diffusivity. In our case $d = 3$ and then the diffusivity is as follows: $\kappa = T/M\lambda$ — the Einstein relation. Using (154) one can rewrite it as follows

$$\kappa = \frac{T}{M\lambda} = \frac{T}{6\pi\eta R}\ . \tag{160}$$

Note that the diffusivity depends on particle radius, but not 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[15].

The probability distribution of displacement at $\lambda t' \gg 1$,

$$\rho(\Delta\mathbf{r}, t') = (4\pi\kappa t')^{-3/2}\exp[-|\Delta\mathbf{r}|^2/4\kappa t']\,,$$

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: $\partial n/\partial t' = \kappa\nabla^2 n$.

In the external field $V(\mathbf{q})$, the particle satisfies the equations

$$\dot{\mathbf{p}} = -\lambda\mathbf{p} + \mathbf{f} - \partial_q V\,, \quad \dot{\mathbf{q}} = \mathbf{p}/M\ . \tag{161}$$

Note that these equations characterize the system with the Hamiltonian $\mathcal{H} = p^2/2M + V(\mathbf{q})$, that interact with the thermostat, which provides friction $-\lambda\mathbf{p}$ and agitation $\mathbf{f}$ - the balance between these two terms expressed by (158) means that the thermostat is in equilibrium.

[15] With temperature in degrees, (160) contains the Boltzmann constant, $k = \kappa M\lambda/T$, 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.

We now pass from considering individual trajectories to the description of the "cloud" of trajectories. Consider the over-damped limit $\lambda^2 M \gg \partial^2_{qq} V$, where we can neglect the acceleration term on timescales exceeding the force correlation time τ: $\lambda \mathbf{p} \gg \dot{\mathbf{p}}$. For example, if we apply to a charged particle an electric field $\mathbf{E} = -\partial_{\mathbf{q}} V$ constant in space, then $\partial^2_{qq} V = 0$; averaging (coarse-graining) over times exceeding τ, we can neglect acceleration, since the particle move on average with a constant velocity $\mathbf{E}/\lambda M$. In this limit our second-order equation (161) on $\mathbf{q}$ is reduced to the first-order equation:

$$\lambda \mathbf{p} = \lambda M \dot{\mathbf{q}} = \mathbf{f} - \partial_q V \ . \tag{162}$$

We can now derive the equation on the probability distribution $\rho(q,t)$, which changes with time due to random noise and evolution in the potential, the two mechanisms can be considered additively. We know that without V,

$$\mathbf{q}(t) - \mathbf{q}(0) = (\lambda M)^{-1} \int_0^t \mathbf{f}(t')dt' \,, \quad \langle |\mathbf{q}(t) - \mathbf{q}(0)|^2 \rangle = 2Dt \,,$$

and the density $\rho(q,t)$ satisfies the diffusion equation. The dynamical equation without any randomness, $\lambda M \dot{\mathbf{q}} = -\partial_q V$, corresponds to a flow in $\mathbf{q}$-space with the velocity $\mathbf{w} = -\partial_{\mathbf{q}} V/\lambda M$. In that flow, density satisfies the continuity equation $\partial_t \rho = -div\, \rho \mathbf{w} = -\partial_{q_i} w_i \rho$. Together, diffusion and advection give the so-called Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \kappa \nabla^2 \rho + \frac{1}{\lambda M} \frac{\partial}{\partial q_i} \rho \frac{\partial V}{\partial q_i} = -\mathrm{div}\, \mathbf{J} \ . \tag{163}$$

More formally, one can derive this equation by writing the Langevin equation (162) as $\dot{q}_i - w_i = \eta_i$ and taking the random force Gaussian delta-correlated: $\langle \eta_i(0) \eta_j(t) \rangle = 2\kappa \delta_{ij} \delta(t)$. Since it is the quantity $\dot{\mathbf{q}} - \mathbf{w}$ which is Gaussian now, then the path integral representation (152) changes into

$$\rho(\mathbf{q}, t; 0, 0) \quad = \quad \int \mathcal{D}\mathbf{q}(t') \exp\left[-\frac{1}{4\kappa} \int_0^t dt' |\dot{\mathbf{q}} - \mathbf{w}|^2 \right] \ , \tag{164}$$

To describe the time change, consider the convolution identity (151) 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 (164) and obtain

$$\rho(\mathbf{q}, t) = \int d\mathbf{q}' (4\pi\kappa\epsilon)^{-d/2} \exp\left[-\frac{[\mathbf{q} - \mathbf{q}' - \epsilon \mathbf{w}(\mathbf{q}')]^2}{4\kappa\epsilon} \right] \rho(\mathbf{q}', t - \epsilon) \ . \tag{165}$$

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 $\mathbf{q}$. 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 $\rho(\mathbf{q}', t - \epsilon)$:

$$\begin{aligned}\rho(\mathbf{q},t) &\approx (1-\epsilon\partial_\mathbf{q}\cdot\mathbf{w})\int d\mathbf{y}(4\pi\kappa\epsilon)^{-d/2}e^{-y^2/4\kappa\epsilon}\rho(\mathbf{q}+\mathbf{y}-\epsilon\mathbf{w},t-\epsilon)\\ &\approx (1-\epsilon\partial_\mathbf{q}\cdot\mathbf{w})\int d\mathbf{y}(4\pi\kappa\epsilon)^{-d/2}e^{-y^2/4\kappa\epsilon}\Big[\rho(\mathbf{q},t)+(\mathbf{y}-\epsilon\mathbf{w})\cdot\partial_\mathbf{q}\rho(\mathbf{q},t)\\ &+(y_iy_j-2\epsilon y_iw_j+\epsilon^2w_iw_j)\partial_i\partial_j\rho(\mathbf{q},t)/2-\epsilon\partial_t\rho(\mathbf{q},t)\Big]\\ &=(1-\epsilon\partial_\mathbf{q}\cdot\mathbf{w})[\rho-\epsilon\mathbf{w}\cdot\partial_\mathbf{q}\rho+\epsilon\kappa\Delta\rho-\epsilon\partial_t\rho+O(\epsilon^2)]\ , \qquad (166)\end{aligned}$$

and obtain (163) 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 solution which corresponds to the particle in an external field and in thermal equilibrium with the surrounding molecules:

$$\rho(q) \propto \exp[-V(q)/\lambda M\kappa] = \exp[-V(q)/T]\ . \qquad (167)$$

Apparently it has a Boltzmann-Gibbs form, and it turns into zero the probability current: $\mathbf{J} = -\rho\partial V/\partial\mathbf{q} - \kappa\partial\rho/\partial\mathbf{q} = e^{-V/T}\partial(\rho e^{V/T})/\partial\mathbf{q} = 0$. We shall use the Fokker-Planck equation in the next section for the consideration of the detailed balance and fluctuation-dissipation relations.

7.3 General fluctuation-dissipation relation

Fluctuation-dissipation theorem and Onsager reciprocity relations treated small deviations from equilibrium. Recently, a significant generalization of equilibrium statistical physics appeared for systems with one or few degrees of freedom deviated arbitrary far from equilibrium. This is 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, like the negative entropy change.

Consider again the over-damped Brownian particle with the coordinate $x(t)$ in a time-dependent potential $V(x,t)$:

$$\dot{x} = -\partial_x V + \eta\ . \qquad (168)$$

Here the random function $\eta(t)$ can be thought of as representing interaction with a thermostat with the temperature T so that $\langle\eta(0)\eta(t)\rangle = 2T\delta(t)$. This equation (used very often in different applications) can be applied to any macroscopic observable, where one can distinguish a systematic and random part of the evolution.

The Fokker-Planck equation for the probability $\rho(x,t)$ has the form (163):

$$\partial_t\rho = T\partial_x^2\rho + \partial_x(\rho\partial_x V) = -\hat{H}_{FP}\rho \ . \tag{169}$$

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 another instance of the analogy between quantum mechanics and statistical physics. We may say that the probability density is the ψ-function is the x-representation, $\rho(x,t) = \langle x|\psi(t)\rangle$. In other words, we consider evolution in the Hilbert space of functions so that we may rewrite (169) in a Schrödinger representation 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). The transition probability is given by the matrix element:

$$\rho(x',t';x,t) = \langle x'|\exp[(t-t')H_{FP})|x\rangle \ . \tag{170}$$

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 we have a Gibbs steady 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:

$$\rho(x',t;x,0)e^{-V(x)/T} = \rho(x,t;x',0)e^{-V(x')/T} \ . \tag{171}$$
$$\langle x'|e^{-tH_{FP}-V/T}|x\rangle = \langle x|e^{-tH_{FP}-V/T}|x'\rangle = \langle x'|e^{-V/T-tH_{FP}^\dagger}|x\rangle \ .$$

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

$$H_{FP}^\dagger \equiv \left(\frac{\partial V}{\partial x} - T\frac{\partial}{\partial x}\right)\frac{\partial}{\partial x} = e^{V/T}H_{FP}e^{-V/T} \ , \tag{172}$$

i.e. $e^{V/2T}H_{FP}e^{-V/2T}$ is hermitian, which can be checked directly.

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 (169) because of the extra term: $\partial_t\rho = -\beta\rho\partial_t V$. Indeed, 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. Still, can we find some use of the Gibbs factor and also have anything generalizing the detailed balance relation (171) 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 (169) 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) during the time t:

$$W_t = \int_0^t dt' \frac{\partial V(x(t'),t')}{\partial t'} \ . \tag{173}$$

The time derivative is partial i.e. taken only with respect to the second argument. The work is a fluctuating quantity depending on the trajectory $x(t')$, which depends on the initial point and noise.

Let us now take many different realizations of the noise $\eta(t)$, choose initial $x(0)$ with the Gibbs probability ρ_0 and run (168) many times with every initial data and every noise realization. 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 (163), 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

$$\partial_t\rho = \beta^{-1}\partial_x^2\rho + \partial_x(\rho\partial_x V) - \partial_W \rho\partial_t V \ , \tag{174}$$

Since $W_0 = 0$ then the initial condition for (174) is

$$\rho(x,W,0) = Z_0^{-1}\exp[-\beta V(x,0)]\delta(W) \ . \tag{175}$$

While we cannot find $\rho(x, W, t)$ for arbitrary $V(t)$ we can multiply (174) by $\exp(-\beta W)$ and integrate over dW. Since $V(x,t)$ does not depend on W, we get the closed equation for $f(x,t) = \int dW \rho(x,W,t)\exp(-\beta W)$:

$$\partial_t f = \beta^{-1}\partial_x^2 f + \partial_x(f\partial_x V) - \beta f \partial_t V \ , \tag{176}$$

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 (175). Note that $f(x,t)$ is instantaneously defined by $V(x,t)$, no history dependence as we have generally in $\rho(x,t)$. 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 entropy is defined 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):

$$\int f(x,t)dx = \int \rho(x,W,t)e^{-\beta W}dxdW = \left\langle e^{-\beta W}\right\rangle = \frac{Z_t}{Z_0} = \frac{\int e^{-\beta V(x,t)}dx}{\int e^{-\beta V(x,0)}dx} \, . \tag{177}$$

Here the bracket means double averaging, over the initial distribution $\rho(x,0)$ and over the different realizations of the Gaussian noise $\eta(t)$ during the time interval $(0,t)$. We can also obtain all weighted moments of x like $\langle x^n \exp(-\beta W_t)\rangle$ [16]. 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. 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, to obtain the mean exponent of the work done, it is enough to know 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). This is, of course, because the further relaxation to the equilibrium at the end value of the potential is not accompanied by doing any work. Remarkable that there is no dependence on the intermediate times. One can also look from it from the opposite perspective: no less remarkable is that one can determine the 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).

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

We can write for the dissipation $W_d = W - F_t + F_0$ (the work minus the free energy change) the following identity:

$$\langle e^{-\beta W_d} \rangle = \int dW_d \rho(W_d) \exp(-\beta W_d) = 1 \,, \tag{178}$$

which is a generalization of the second law of thermodynamics. Indeed, the mean dissipation divided by temperature is the thermodynamic entropy change. Using the Jensen inequality $\langle e^A \rangle \geq e^{\langle A \rangle}$, one can obtain the usual second law of thermodynamics in the following form:

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

Moreover, the Jarzynski relation is a generalization of the fluctuation-dissipation theorem, which can be derived from it for small deviations from equilibrium. Namely, we can consider $V(x,t) = V_0(x) - f(t)x$, consider limit of $f \to 0$, expand (177) up to the second-order terms in f and express the response to the field as the time derivative of the second moment.

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:

$$\langle e^{-\beta W_d - I} \rangle = \langle e^{-\Delta S} \rangle = 1 \ . \tag{179}$$

(Sagawa and Uedo, 2012; Sagawa 2012). We have considered such a case in Section 4.6, where we denoted $W_d = Q$ and used $\langle W_d \rangle \geq -IT = -T\Delta S$. The exponential equality (179) is a generalization of this inequality and (94).

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 (179) 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 is as follows (Crooks 1999):

$$\rho^\dagger(-\Delta S) = \rho(\Delta S) e^{-\Delta S} \ . \tag{180}$$

Integrating this relation one obtains (177,178,179).

Let us prove this relation for our toy model of the generalized baker map from the Section 3.5. Remind that we derived there a long-time average rate of the entropy production, which corresponded to the volume contraction 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 for such n-sequence is $\Delta S = n \ln J = n_1 \ln \frac{r}{l} + n_2 \ln \frac{l}{r}$ and its probability 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:

$$\frac{P(\Delta S)}{P(-\Delta S)} = \frac{P(\sigma)}{P(-\sigma)} = \left(\frac{l}{r}\right)^{n_2-n_1} = e^{n\sigma} = e^{\Delta S} \ . \tag{181}$$

In a multi-dimensional case, apart from making the potential time-dependent, there is another way to deviate the system from equilibrium: to apply in addition to the random thermal force $\mathbf{f}(t)$ a coordinate-dependent force $\mathbf{F}(\mathbf{q}, t)$ which is non-potential (not a gradient of any scalar). Again, when the dynamical equation are non-Hamiltonian, contact with the thermostat does not lead to thermal equilibrium. Indeed, there is no Gibbs steady solution to the Fokker-Planck equation and the detailed balance is now violated in the following way:

$$H_K^\dagger = e^{\beta\mathcal{H}} H_{FP} e^{-\beta\mathcal{H}} + \beta(\mathbf{F}\cdot\mathbf{p}) \ , \tag{182}$$

The last term is again the power $(\mathbf{F}\cdot\mathbf{p}) = (\mathbf{F}\cdot\dot{\mathbf{q}})$ divided by temperature i.e. the entropy production rate. A close analog of the Jarzynski relation can be formulated for the production rate averaged during the time t:

$$\sigma_t = \frac{1}{tT}\int_0^t (\mathbf{F}\cdot\dot{\mathbf{q}})\, dt \ . \tag{183}$$

Would $\mathbf{F} = dV/d\mathbf{q}$, that is a gradient of a scalar, then $(\mathbf{F}\cdot\dot{\mathbf{q}}) = dV(\mathbf{q}(t))/dt$. The quantity (183) fluctuates from realization to realization. The probabilities $P(\sigma_t)$ satisfy the relation, analogous to (180), which we give without general derivation

$$\frac{P(\sigma_t)}{P(-\sigma_t)} \propto e^{t\sigma_t} \ . \tag{184}$$

The second law of thermodynamics states that to keep the system away from equilibrium, the external force $\mathbf{F}$ must on average do a positive work. Over a long time we thus expect σ_t to be overwhelmingly positive, yet fluctuations do happen. The relations (184,181) show 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 (183) must have a large-deviation form, $P(\sigma_t) \propto \exp[-tH(\sigma_t)]$, so that (184) means that $H(\sigma_t) - H(-\sigma_t) = -\sigma_t$, as if $P(\sigma_t)$ was Gaussian.

One calls (180,184) detailed fluctuation-dissipation relations since they are stronger than integral relations of the type (177,178). Indeed, it is straightforward to derive $\langle \exp(-t\sigma_t)\rangle = 1$ from (184).

The relation similar to (184) 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 $s_1, \dots, s_N$ 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 (184), 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 Conclusion

The Chapter is an attempt in compressing the course to its most essential elements.

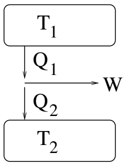
8.1 Take-home lessons

1. Thermodynamics studies restrictions imposed on observables by hidden. It deals with two fundamental extensive quantities. The first one (energy) 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) 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 lies below this convex curve in $S - E$ plane.

2. Temperature is the derivative of the energy with respect to the entropy. Extremum of entropy means that temperatures of the connected subsystems are equal in equilibrium. The same is true for the energy derivative with respect to volume (pressure). The entropy increase (called the second law of thermodynamics) imposes restrictions on thermal engine efficiency, that is the fraction of heat

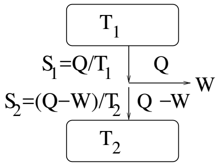
used for work:

$$\frac{W}{Q_1} = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{T_2 \Delta S_2}{T_1 \Delta S_1} \le 1 - \frac{T_2}{T_1} \; .$$



Similarly, if information processing generates ΔS, its energy price is as follows:

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



3. 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 better options) the uniform (microcanonical) distribution $w = 1/\Gamma$. 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: $\log w_i \propto -E_i$. We generalize this approach in 13 below within the information theory.

4. Basic mathematical object is the sum of independent random numbers $X = \sum_{i=1}^{N} y_i$. Three concentric statements are made. The weakest statement 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 $N^{-1/2}$ of the maximum. For even larger deviations, the distribution is very sharp: $\mathcal{P}(X) \propto \mathrm{e}^{-NH(X/N)}$ where $H \ge 0$ and $H(\langle y \rangle) = 0$. Applying this to the log of the probability of a given sequence, $\lim_{N\to\infty} p(y_1 \ldots y_N) = -NS(Y)$, 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**.

5. Irreversibility of the entropy growth seems to contradict Hamiltonian dynamics, which is time-reversible and preserves the N-particle phase-space probability density. However, one can obtain the equation on a one-particle density for a dilute gas, where only pair collisions matter. If we then assume that before every collision the two particles were independent, then we obtain the Boltzmann kinetic equation, which, in particular, describes the irreversible growth of the one-particle entropy. Since the total entropy is concerned, while the sum of one-particle entropies grow, we conclude that their difference must grow too. Later, we call it mutual information.

6. The N-particle entropy growth comes due to incomplete knowledge of initial conditions, which requires one to consider finite regions in phase space. These regions spread over the whole phase space under a generic 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.

7. Another simple mathematical property we use throughout is convexity. We first use it in the 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 in establishing that even when the mean of a random quantity is zero, its mean exponent is positive, which provides for an exponential separation of trajectories in an incompressible flow and exponential growth of the density of an element in a compressible flow.

8. Since the lack of knowledge or uncertainty play so prominent role, we wish to quantify it. That will measure simultaneously the amount of information needed to remove that uncertainty. This is consistently done in a discrete case, for instance, by counting the number of bits, that is answers to "yes-no" questions. That way we realize that the information is the logarithm of the number of equally probable possibilities (Boltzmann entropy) or the mean logarithm if the probabilities are different (Shannon-Gibbs entropy). Here convexity of the function $-w \log w$ helps us to prove that the information/entropy has its maximum for equal probabilities (when our ignorance is maximal).

9. The point 4 above states that the number of typical sequences grows with the rate equal to entropy. That means that the entropy is both the mean and the fastest rate of the reception of information brought by long messages/measurements. If the natural source brings highly redundant information (like in visual signals), to achieve the fastest rate we need to compress it, squeezing all the unnecessary bits out. That can be also accomplished by encoding.

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 rate of growth of possible number of errors. Sending extra bits to correct these errors lowers the transmission rate from $S(B)$ to the mutual information $I(A, B) = S(B) - S(B|A)$, 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 to guarantee 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 $S(h|e) = S(h) - I(e,h)$. 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 $S(p) + D(p|q)$.

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 approach from the point 3. It also sheds new light on physics, telling us that on some basic level all states are constrained equilibria.

14. Information is physical: to learn $\Delta S = S(A) - S(A,M)$ one does the work $T\Delta S$, where A is the system and M is the measuring device. To erase information, one needs to convert $TS(M)$ into heat. Both acts require a finite temperature. The energetic price of a cycle is T times the mutual information: $TI(A,M)$.

15. Looking at Renormalization Group as a way to forget information we find that the proper measure is again the mutual information, which can defined in two ways: either between remaining and eliminated degrees of freedom or between different parts of the same system. In particular, it shows us examples of the area law, when I is sub-extensive.

16. That area law was the initial motivation for the quantum information theory, since the entropy of a black hole is proportional to its area rather than volume. Nowadays, of course, it is driven by the quest for a quantum computer. Already quantum mechanics has a natural entropic formulation. Quantum statistics appears when we treat subsystems and must deal with the density matrix and its entanglement entropy, which is a sibling of the mutual information, since it also measures the degree of correlation. We learn that entropy of the whole can be less than the entropy of a part.

17. The last lesson is two progressively more powerful forms of the second law of thermodynamics, which originally was $\Delta S \geq 0$. The first new one, $\langle e^{-\Delta S} \rangle = 1$, is the analog of a Liouville theorem. The second one relates probabilities of forward

and backward process: $\rho^\dagger(-\Delta S) = \rho(\Delta S)e^{-\Delta S}$.

8.2 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. The mathematical tool 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.

Our focus was mainly 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 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 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.

Reader surely recognized that no rigorous proofs were given, replaced instead by plausible hand-waving argument or even a particular example. Those interested in proofs for Chapter 3 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. Nor the examples given are representative of the ever-widening avalanche of applications; more biological applications can be found in "Biophysics" by Bialek, others in original articles and reviews. Numerous references scattered through the text, like (Zipf 1949), give you the most compact encoding of what is to google to find details.

I invite you to reflect on the history of our attempts to realize limits of possible, from heat engines through communication channels to computations.

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 languages (and worlds) of Tolstoy and Dostoevsky differ so much. 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 supreme being would have no need in the statistical information approach described here. One may also wonder to what extent essential presence of scientist in science may help us understand the special status of measurement in quantum mechanics.

I wish to thank my colleagues whose advices were invaluable in preparing the course: Yakov Sinai, Alexander Zamolodchikov, Massimo Vergassola, Eric Siggia, Leonid Levitov. Special thanks to Michal Shavit for preparing tutorials and catching numerous errors (the remaining ones are my responsibility alone).

As we learnt here, better understanding must lead to a more compact presentation; hopefully, the next version of these lecture notes will be shorter.