

# Physical Nature of Information

G. Falkovich

January 25, 2022

How to receive, send and forget information

## Contents

<b>1</b>	<b>Thermodynamics and statistical physics</b>	<b>5</b>
1.1	Basics of thermodynamics . . . . .	5
1.2	Free energy . . . . .	10
1.3	Microcanonical distribution . . . . .	12
1.4	Canonical distribution . . . . .	14
<b>2</b>	<b>Appearance of irreversibility</b>	<b>16</b>
2.1	Evolution in the phase space . . . . .	17
2.2	Phase-space mixing and entropy growth . . . . .	19
2.3	Entropy decrease and non-equilibrium fractal measures . . . . .	26
<b>3</b>	<b>Physics of information</b>	<b>28</b>
3.1	Central limit theorem and large deviations . . . . .	29
3.2	Information as a choice . . . . .	32
3.3	Communication Theory . . . . .	36
3.4	Correlations in the signals . . . . .	39
3.5	Mutual information as a universal tool . . . . .	41
3.6	Hypothesis testing and relative entropy . . . . .	49

<b>4</b>	<b>Applications of Information Theory</b>	<b>54</b>
4.1	Distribution from information . . . . .	54
4.2	Exorcizing Maxwell demon . . . . .	58
4.3	Renormalization group and information loss . . . . .	62
4.4	Flies and spies . . . . .	64
4.5	Rate Distortion and Information Bottleneck . . . . .	70
4.6	Information is money . . . . .	74
<b>5</b>	<b>Stochastic processes</b>	<b>77</b>
5.1	Random walk and diffusion . . . . .	77
5.2	General fluctuation-dissipation relation . . . . .	79
5.3	Stochastic Web surfing and Google's PageRank . . . . .	83
<b>6</b>	<b>Conclusion</b>	<b>86</b>
6.1	Take-home lessons . . . . .	86
6.2	Epilogue . . . . .	90

## Preface

The book grew out of a one-semester course, 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 book) answers the most frequent question: *How much can we say and do about something we do not know?* Of course, the art of bluffing without blushing was perfected by people in many trades and walks of life. So when the course was taught in different institutions and countries, it was attended by a motley mix of students, post-docs and faculty from physics, mathematics, engineering, computer science, economics and biology. Eventually, it evolved into a meeting place where we learn from each other using the universal language of information theory, which is a statistical physics in disguise, albeit transparent.

The simplest way to answer the above question is called thermodynamics. It is a phenomenology that deals only with visible manifestations of the hidden, using symmetries and conservation laws to restrict possible outcomes and focusing on mean values ignoring fluctuations. More sophisticated approach derives the statistical laws by explicitly averaging over the hidden degrees of freedom. Those laws justify thermodynamics and describe the probabilities of fluctuations. More important, the basic notion of this approach (Gibbs entropy) turns out to be arguably the most important conceptual and technical tool of the modern science and technology.

The first Chapter recalls the 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 learn how irreversible entropy change appears from reversible flows in phase space via dynamical chaos. We understand that *entropy is not a property of a system, but of our knowledge of the system*. It is then natural to use the language of the information theory revealing the universality of the approach, which to a large extent is based on the simple trick of adding many random numbers. Building on that basis, one develops several versatile instruments, of which the mutual information and its quantum sibling, entanglement entropy, are presently most widely applied to the description of subjects ranging from bacteria and neurons to markets and quantum computers. We then discuss the so far most sophisticated way to forget information - renormalization group. Forgetting is a fascinating activity — one learns truly fundamental things this way. We end with the

stochastic thermodynamics and the generalizations of the second law.

Even though it is a graduate text, the book uses only elementary mathematical tools, but from all three fields — geometry, algebra and analysis — which correspond respectively to studying space, time and continuum in the physical world. We employ two complementary ways of thinking: continuous flows and discrete combinatorics (thus involving both brain hemispheres). Together, they produce a powerful and universal tool, applied everywhere, from computer science and machine learning to biophysics and economics. The book is panoramic, trying to combine into a reasonably coherent whole the subjects that are taught in much details in different departments: thermodynamics and statistical mechanics (as taught in physics and engineering), dynamical chaos (as taught in physics and applied mathematics), information and communication theories (as taught in computer science and engineering). At the end, recognizing the informational nature of physics and breaking the barriers of specialization is also of value for physicists. People working on quantum computers and the entropy of black holes use the same tools as those designing self-driving cars and market strategies, studying molecular biology, animal behavior and human languages, and figuring out how the brain works. Last, I felt compelled to tell the story worth telling: how we discover the limits imposed by uncertainty on engines, communications and computations.

Small-print parts can be omitted upon the first reading.

# 1 Thermodynamics and statistical physics

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

Yet even when we have a closed set of equations, they need initial or boundary conditions taken from measurements. Here again our knowledge is only partial because of a finite precision of measurements. This has dramatic consequences, when there is an instability, so that small variation of initial data leads to large deviation in evolution. In a sense, every new decimal in precision is a new degree of freedom for an unstable system.

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.

The first two sections of this Chapter present a phenomenological approach called thermodynamics. The last two sections serve as a brief reminder of statistical physics.

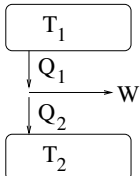
## 1.1 Basics of thermodynamics

One can teach monkey to differentiate, integration requires humans.

G Kotkin

It all started when practical needs to estimate the engine efficiency during the industrial revolution led to the development of the abstract concept of entropy. 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<sup>1</sup>. The work  $W$  is the difference between the heat given by the hot reservoir  $Q_1$  and the heat absorbed by the cold one  $Q_2$ . What is the maximal fraction of heat we can use for work? Carnot in 1824 stated that in all processes  $Q_1/T_1 \leq Q_2/T_2$ , so that the efficiency is bounded from above:

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


The diagram shows two rectangular boxes representing reservoirs. The top box is labeled  $T_1$  and the bottom box is labeled  $T_2$ . A downward arrow labeled  $Q_1$  points from the  $T_1$  box to a central point. From this point, a horizontal arrow labeled  $W$  points to the right. From the same central point, another downward arrow labeled  $Q_2$  points to the  $T_2$  box.

His elaborate arguments are of only historic interest now. Clausius in 1865 introduced the notion of entropy as a factor connecting temperature and heat, so we now interpret the Carnot criterium, saying that the entropy decrease of the hot reservoirs,  $\Delta S_1 = Q_1/T_1$ , must be less than the entropy increase of the cold one,  $\Delta S_2 = Q_2/T_2$ . Maximal work is achieved for minimal (zero) total entropy change,  $\Delta S_2 = \Delta S_1$ , which happens for reversible processes — if, for instance, a gas works by moving a piston then the pressure of the gas and the work are less for a fast-moving piston than in equilibrium. The efficiency is larger when the temperatures differ more.

Just like the progress from Carnot engine to a general thermodynamics, laws of nature appear usually by induction: from data and particular cases to a general law and from processes to state functions. The latter step requires integration (to pass, for instance, from the Newton equations of mechanics to the Hamiltonian or from thermodynamic equations of state to thermodynamic potentials). It is much easier to differentiate than to integrate, and so deduction (or postulation approach) is usually more simple and elegant. It also provides a good vantage point for generalizations and appeals to our brain, which likes to hypothesize 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 we present a deductive description of thermodynamics<sup>2</sup>. *Thermodynamics studies restrictions on the possible macroscopic properties that follow from the fundamental conservation laws.* Therefore, thermodynamics does

---

<sup>1</sup>Look under the hood of your car to appreciate the level of idealization achieved in that definition.

<sup>2</sup>For a more detailed yet still compact presentation in this spirit, see the book H. B. Callen, *Thermodynamics* (1965).

not predict numerical values but rather sets inequalities and establishes relations among different properties.

You start building thermodynamics by identifying a conserved quantity, which can be exchanged but not created. It could be matter, money, energy, etc. For most physical systems, the basic symmetry is invariance of the fundamental laws with respect to time shifts<sup>3</sup>. Evolution of an isolated physical system is usually governed by the Hamiltonian (the energy written in canonical variables), whose time-independence means energy conservation. In what follows, the conserved quantity of thermodynamics is called energy and denoted  $E$ . We wish to ascribe to the states of the system the values of  $E$ . We focus on the states independent of the way they are prepared. We call such states equilibrium, they are completely characterized by the *static* values of observable variables.

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

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

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

---

<sup>3</sup>Be careful trying to build thermodynamics for biological or social-economic systems, since generally the laws that govern them are not time-invariant. For example, the metabolism of the living beings changes with age, and the number of market regulations generally increases (as well as the total money mass, albeit not necessarily in our pockets).

Mayer in 1842; before that, heat was believed to be a separate fluid conserved by itself.

**The basic problem** of thermodynamics is the determination of the equilibrium state that eventually results after all internal constraints are removed in a closed composite system. The problem is solved with the help of extremum principle: there exists a quantity  $S$  called entropy which is a function of the parameters of 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 (Clausius 1865).

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

In that limit, thermodynamic entropy is also an extensive variable<sup>4</sup>, which is a homogeneous first-order function of all the extensive parameters:

$$S(\lambda E, \lambda V, \dots) = \lambda S(E, V, \dots) . \quad (3)$$

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

Of course, (3) does not mean that  $S(E)$  is a linear function when other parameters fixed:  $S(\lambda E, V, \dots) \neq \lambda S(E, V, \dots)$ . On the contrary, we shall see in a moment that it is a convex function. Nor entropy is necessary a monotonic function of energy<sup>5</sup>. Yet for every interval of a definite derivative sign, say  $(\partial E / \partial S)_X > 0$ , we can solve  $S = S(E, V, \dots)$  uniquely for  $E(S, V, \dots)$  which is an equivalent fundamental relation. We assume the

---

<sup>4</sup>We shall see later that non-extensive parts of entropy are also important for studying interaction and correlations between subsystems.

<sup>5</sup>An example of the two-level system in Section 1.4 shows that  $S(E)$  could be non-monotonic for systems with a finite phase space.



functions  $S(E, X)$  and  $E(S, X)$  to be continuous differentiable. An efficient way to treat partial derivatives is to use jacobians

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

Then

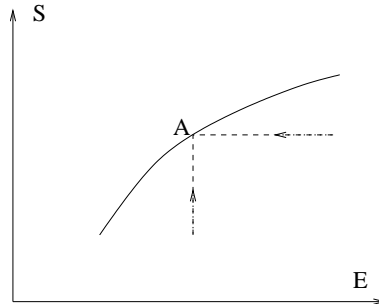
$$\left( \frac{\partial S}{\partial X} \right)_E = 0 \Rightarrow \left( \frac{\partial E}{\partial X} \right)_S = - \frac{\partial(ES)}{\partial(XS)} \frac{\partial(EX)}{\partial(EX)} = - \left( \frac{\partial S}{\partial X} \right)_E \left( \frac{\partial E}{\partial S} \right)_X = 0.$$

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 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 minimal perimeter for a given area).

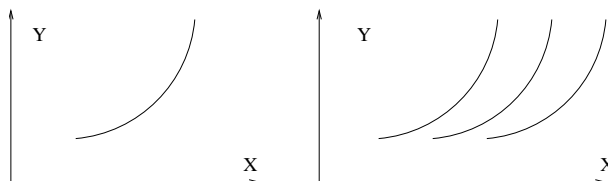
It is important that the equilibrium curve  $S(E)$  is convex, which guarantees stability of a homogeneous state. Indeed, if our system would break spontaneously into two halves with a bit different energies, the entropy must decrease:  $2S(E) > S(E + \Delta) + S(E - \Delta) = 2S(E) + S'' \Delta^2 / 2$ , which requires  $S'' < 0$  (that argument does not work for systems with long-range interaction where energy is non-additive). 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:



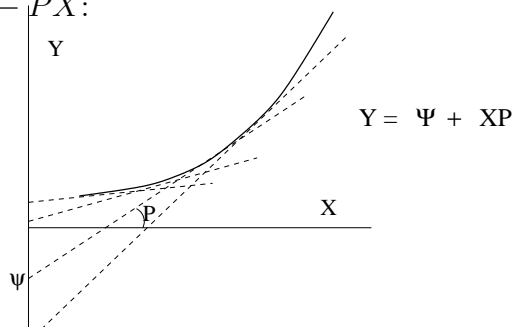
One can work either in energy or entropy representation but ought to be careful not to mix the two.

## 1.2 Free energy

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, the single function  $Y = P^2/4$  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  $\psi$ . 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:  $dF(T, V, N, \dots) = -SdT - PdV + \mu dN + \dots$ . 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.

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

The next two sections present a brief reminder of classical Boltzmann-Gibbs statistical mechanics. Here we introduce microscopic statistical description in the phase space and describe two principal ways (microcanonical and canonical) to derive thermodynamics from statistical mechanics.

### 1.3 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*:

$$\rho(p, q) = A\delta[E(p_1 \dots p_N, q_1 \dots q_N) - E_0] . \quad (4)$$

Usually one considers the energy fixed with the accuracy  $\Delta$  so that the microcanonical distribution is

$$\rho = \begin{cases} 1/\Gamma & \text{for } E \in (E_0, E_0 + \Delta) \\ 0 & \text{for } E \notin (E_0, E_0 + \Delta), \end{cases} \quad (5)$$

where  $\Gamma$  is the volume of the phase space occupied by the system

$$\Gamma(E, V, N, \Delta) = \int_{E < \mathcal{H} < E + \Delta} d^{3N}p d^{3N}q . \quad (6)$$

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

$$\Gamma(E, V, N, \Delta) \propto E^{3N/2-1}V^N \Delta / (3N/2 - 1)! \approx (E/N)^{3N/2}V^N \Delta . \quad (7)$$

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

$$S(E, V, N) = \ln \Gamma(E, V, N) . \quad (8)$$

This is one of the most important formulas in physics<sup>6</sup> (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  $N \rightarrow \infty$ .

---

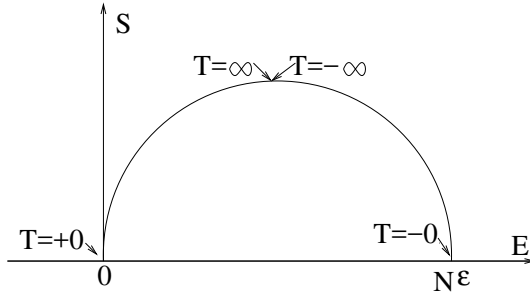
<sup>6</sup>It is inscribed on the Boltzmann's gravestone.

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 (??). Assume that the indeterminacy in the energy of any subsystem,  $\Delta$ , is much less than the total energy  $E$ . Then

$$\Gamma(E) = \sum_{i=1}^{E/\Delta} \Gamma_1(E_i)\Gamma_2(E - E_i) . \quad (9)$$

We denote  $\bar{E}_1, \bar{E}_2 = E - \bar{E}_1$  the values that correspond to the maximal term in the sum (9). 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 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  $N_1, N_2 \rightarrow \infty$  then  $S(E) = S_1(\bar{E}_1) + S_2(\bar{E}_2) + O(\log N)$  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  $\epsilon$ . If the energy of the set is  $E$  then there are  $L = E/\epsilon$  upper levels occupied. The statistical weight is determined by the number of ways one can choose  $L$  out of  $N$ :  $\Gamma(N, L) = C_N^L = N!/L!(N - L)!$ . We can now define entropy (i.e. find the fundamental relation):  $S(E, N) = \ln \Gamma$ . At the thermodynamic limit  $N \gg 1$  and  $L \gg 1$ , it gives  $S(E, N) \approx N \ln[N/(N - L)] + L \ln[(N - L)/L]$ , which coincides with (??). 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.

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

## 1.4 Canonical distribution

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

$$w_a(E) = Z^{-1} \exp(-E/T) , \quad (10)$$

$$Z = \sum_a \exp(-E_a/T) . \quad (11)$$

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 (10) times the number of states i.e. the statistical weight of the *subsystem*:

$$W(E) = \Gamma(E)w_a(E) = \Gamma(E)Z^{-1} \exp(-E/T) . \quad (12)$$

Here the weight  $\Gamma(E)$  grows with  $E$  very fast for large  $N$ . But as  $E \rightarrow \infty$  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 (10) tells that the probability of a given microstate exponentially decays with the energy of the state while (12) tells that the probability of a given energy has a peak.

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

$$F(T, V, N) = -T \ln Z(T, V, N) . \quad (13)$$

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:

$$\begin{aligned} S &= - \langle \ln w_a \rangle = - \sum w_a \ln w_a = \sum w_a (E_a/T + \ln Z) & (14) \\ &= E/T + \ln Z = (E - F)/T = - \ln w_a(E) = S(E) . \end{aligned}$$

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

Are canonical and microcanonical descriptions equivalent? Of course, not. The descriptions are equivalent only when fluctuations are neglected and consideration is restricted to mean values. That takes place in thermodynamics,

where the distributions just produce different fundamental relations between the mean values:  $S(E, N)$  for microcanonical,  $F(T, N)$  for canonical,  $\Omega(T, \mu)$  for grand canonical. These relations are related by the Legendre transforms. How operationally one checks, for instance, the equivalence of canonical and microcanonical energies? One takes an isolated system at a given energy  $E$ , measures the derivative  $\partial E/\partial S$ , then puts it into the thermostat with the temperature equal to that  $\partial E/\partial S$ ; the energy now fluctuates but the *mean* energy must be equal to  $E$  (as long as system is macroscopic and all the interactions are short-range).

Let us repeat this important distinction: all thermodynamic potentials are equivalent for the 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 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 a weakly fluctuating system in thermal equilibrium 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 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.

## 2.1 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 energy-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:  $w = \lim_{T \rightarrow \infty} \Delta t / T$ . 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:

$$\bar{f} = \int f(p, q) \rho(p, q) dpdq = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(t) dt . \quad (15)$$

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  $\rho_{12}$  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:  $\partial\rho/\partial t + \text{div}(\rho\mathbf{v}) = 0$ . 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:  $\text{div}\mathbf{v} = \partial\dot{q}_i/\partial q_i + \partial\dot{p}_i/\partial p_i = 0$ . That gives the Liouville theorem:  $d\rho/dt = \partial\rho/\partial t + (\mathbf{v}\nabla)\rho = -\rho\text{div}\mathbf{v} = 0$ . The statistical distribution is thus conserved along the phase trajectories of any subsystem. As a result,  $\rho$  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):

$$\ln\rho_a = \alpha_a + \beta E_a(p, q) + \mathbf{c} \cdot \mathbf{P}_a(p, q) + \mathbf{d} \cdot \mathbf{M}(p, q). \quad (16)$$

Here  $\alpha_a$  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*:

$$\rho(p, q) = A \exp[-\beta E(p, q)]. \quad (17)$$

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 (17) 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 (17) 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 (17). For an isolated system with a fixed energy, the entropy maximum corresponds to a uniform micro-canonical distribution.

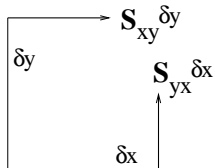
## 2.2 Phase-space mixing and entropy growth

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,  $\text{div } \mathbf{v} = 0$ , so it conserves the total Gibbs entropy:  $dS/dt = - \int d\mathbf{x} \ln \rho \frac{\partial \rho}{\partial t} = \int d\mathbf{x} \ln \rho \text{div } \rho \mathbf{v} = - \int d\mathbf{x} (\mathbf{v} \nabla) \rho = - \int d\mathbf{x} \rho \text{div } \mathbf{v} = 0$ . How then the entropy can grow?

To answer that question, 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:  $\text{Tr} \sigma_{ij} = \sum_i S_i = \text{div} \mathbf{v} = 0$  — 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:

$$r_i(t) = r_i(0) \exp \left[ \int_0^t S_i(t') dt' \right]. \quad (18)$$

For a time-independent strain, the growth/decay is exponential in time. One recognizes that a purely straining motion converts a spherical element into an ellipsoid with the principal diameters that grow (or decay) in time. Indeed,

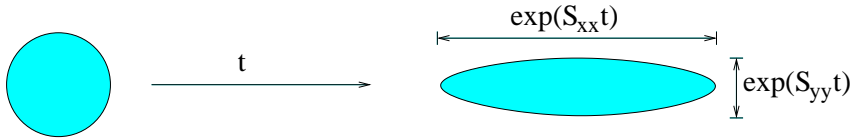


Figure 1: Deformation of a phase-space element by a permanent strain.

consider a two-dimensional projection of the initial spherical element i.e. a circle of the radius  $R$  at  $t = 0$ . The point that starts at  $x_0, y_0 = \sqrt{R^2 - x_0^2}$  goes into

$$\begin{aligned} x(t) &= e^{S_{11}t} x_0, \\ y(t) &= e^{S_{22}t} y_0 = e^{S_{22}t} \sqrt{R^2 - x_0^2} = e^{S_{22}t} \sqrt{R^2 - x^2(t) e^{-2S_{11}t}}, \\ x^2(t) e^{-2S_{11}t} + y^2(t) e^{-2S_{22}t} &= R^2. \end{aligned} \quad (19)$$

The equation (19) 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  $\sum_{i=1}^{6N} x_i^2(t) e^{-2S_i t} = \text{const}$ .

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)<sup>7</sup>.

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):

$$\lim_{t \rightarrow \infty} \int_0^t S_i(t') dt' = 0, \quad \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T dt \exp \left[ \int_0^t S_i(t') dt' \right] \geq 1. \quad (20)$$

---

<sup>7</sup>"Time and space are modes by which we think and not conditions in which we live"  
A. Einstein

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  $\Lambda$  is uniformly distributed over the interval  $-a < \Lambda < a$ , the average  $\Lambda$  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<sup>8</sup> 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  $\varphi$  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 mathematics 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  $\text{tr } \hat{\sigma} = 0$  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  $t \rightarrow \infty$ . In particular, its eigenvectors tend to  $d$  fixed orthonormal eigenvectors  $\mathbf{f}_i$ . Geometrically, that precisely means that an initial sphere evolves into an elongated ellipsoid at later times. The limiting eigenvalues

$$\lambda_i = \lim_{t \rightarrow \infty} t^{-1} \ln |\hat{W} \mathbf{f}_i| \quad (21)$$

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,

---

<sup>8</sup>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 every  $p_i - q_i$  plane projection.

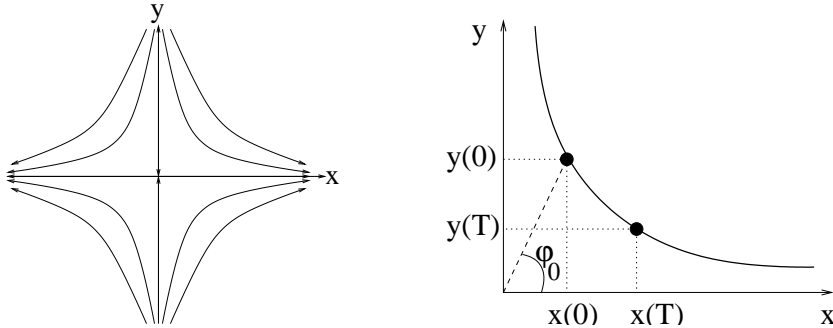
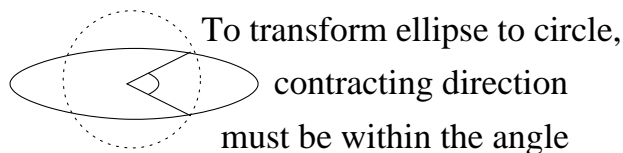


Figure 2: Left panel: streamlines of the saddle-point flow. Right panel: motion along a streamline. For  $\varphi = \varphi_0$  the initial and final points are symmetric relative to the diagonal:  $x(0) = y(T)$  and  $y(0) = x(T)$ . If  $\varphi < \varphi_0 = \arccos[1 + \exp(2\lambda T)]^{-1/2} > \pi/4$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

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:  $E(a) = \lim_{t \rightarrow \infty} t^{-1} \ln [\langle r^a(t) / r^a(0) \rangle] > 0$  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  $\epsilon^{6N}$ , 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  $\lambda t$ .



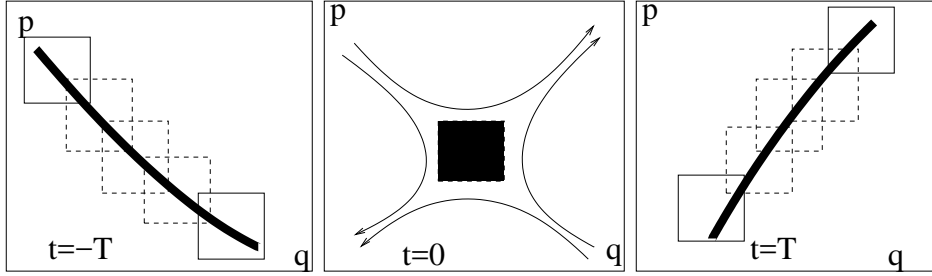
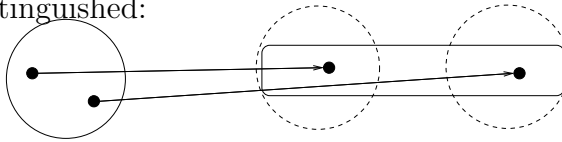


Figure 3: Increase of the phase volume upon stretching-contraction and coarse-graining. Central panel shows the initial state and the velocity field.

The positive Lyapunov exponent  $\lambda$  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. Consider an ensemble of systems having close initial positions within our finite resolution. In a flow with positive Lyapunov exponents, with time we loose our ability to predict where it goes. This loss of information is determined by the growth of the available phase volume, that is of the entropy. But we can look backwards in time and ask where the points come from. 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 acquisition rate of such information about the past is again the sum of the positive Lyapunov exponents and is called the Kolmogorov-Sinai entropy. As time lag from the present moment increases, we can say less and less where we shall be and more and more where we came from. It illustrates the Kierkegaard's remark that the irony of life is that it is lived forward but understood backwards.

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  $\epsilon$ . 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  $\epsilon \rightarrow 0$  and  $T \rightarrow \infty$  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*.<sup>9</sup>

### 2.3 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  $\gamma_i$ , so that the equations of motion take the form:  $\dot{p}_i = f_i - \gamma_i p_i - \partial H / \partial q_i$ ,  $\dot{q}_i = \partial H / \partial p_i$ , which gives 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,

$$\frac{dS}{dt} = \int \rho(\mathbf{r}, t) div \mathbf{v}(\mathbf{r}, t) d\mathbf{r} = \langle \rho div \mathbf{v} \rangle. \quad (22)$$

---

<sup>9</sup>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.

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 (22) is non-zero because of correlations between  $\rho$  and  $div \mathbf{v}$ . Since  $\rho$  is always smaller in the expanding regions where  $div \mathbf{v} > 0$ , then *the entropy production rate (22) 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:

$$\rho(t)/\rho(0) = \exp \left[ - \int_0^t div \mathbf{v}(t') dt' \right] = e^{C(t)}. \quad (23)$$

As we have seen in (20), 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:

$$\langle \rho(t)/\rho(0) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T dt \exp \left[ - \int_0^t div \mathbf{v}(t') dt' \right] = \langle e^C \rangle > 1 .$$

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 3.1). 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)

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

is a sum of the Lyapunov exponents, which is then non-positive (in distinction from an instantaneous average over space, which is zero at any time:  $\int \text{div } \mathbf{v} d\mathbf{r} = 0$ ). It is important that we allowed for a compressibility of a phase-space flow  $\mathbf{v}(\mathbf{r}, t)$  but did not require its irreversibility. Indeed, even if the system is invariant with respect to  $t \rightarrow -t$ ,  $\mathbf{v} \rightarrow -\mathbf{v}$ , 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.

To conclude this Chapter, let us stress the difference between the entropy growth described in the Sections 2.2-?? 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 2.2. 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" (19th century) and "dynamical chaos" (20th century). What shall we have in the 21st century?

### 3 Physics of information

This section presents an elementary introduction into the information theory from the viewpoint of a natural scientist. It re-tells the story of statistical

physics using a different language, which lets us to see the Boltzmann and Gibbs entropies in a new light. What I personally like about the information viewpoint is that it erases paradoxes and makes the second law of thermodynamics trivial. It also allows us to see generality and commonality in the approaches (to partially known systems) of physicists, engineers, computer scientists, biologists, brain researchers, social scientists, market speculators, spies and flies. We shall see how the same tools used in setting limits on thermal engines are used in setting limits on communications, measurements and learning (which are essentially the same phenomena). We shall see how fast widens the region of applications of the universal tool of entropy (and related notions of relative entropy and mutual information): from physics, communications and computations to artificial intelligence and quantum computing.

### 3.1 Central limit theorem and large deviations

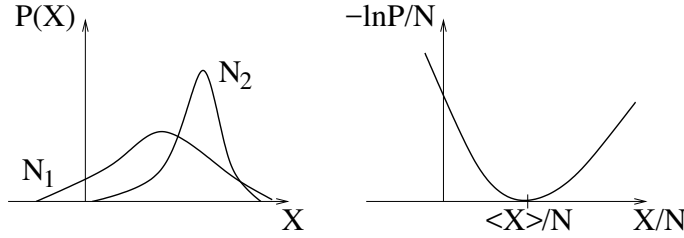
The true logic of this world is to be found in the theory of probability.  
Maxwell

Here we switch from continuous thinking in terms of phase-space flows to discrete combinatoric manipulations. As a bridge from statistical physics to information theory, we start from a simple technical tool used in both. Mathematics, underlying most of the statistical physics in the thermodynamic limit, comes from universality, which appears upon adding independent random numbers. The weakest statement is the law of large numbers: the sum approaches the mean value exponentially fast. The next level is the central limit theorem, which states that majority of fluctuations around the mean have Gaussian probability distribution. Consideration of large rare fluctuations requires the so-called large-deviation theory. Here we briefly present all three at the physical (not mathematical) level.

Consider the variable  $X$  which is a sum of many independent identically distributed (iid) random numbers  $X = \sum_1^N y_i$ . Its mean value  $\langle X \rangle = N\langle y \rangle$  grows linearly with  $N$ . Here we show that its fluctuations  $X - \langle X \rangle$  not exceeding  $\mathcal{O}(N^{1/2})$  are governed by the Central Limit Theorem:  $(X - \langle X \rangle)/N^{1/2}$  becomes for large  $N$  a Gaussian random variable with variance  $\langle y^2 \rangle - \langle y \rangle^2 \equiv \Delta$ . The quantities  $y_i$  that we sum can have quite arbitrary statistics, the only requirements are that the first two moments, the mean  $\langle y \rangle$  and the variance  $\Delta$ , 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

$$\mathcal{P}(X) \propto e^{-NH(X/N)}. \quad (24)$$



To show this, we write

$$\begin{aligned} \mathcal{P}(X) &= \int \delta\left(\sum_{i=1}^N y_i - X\right) \mathcal{P}(y_1) dy_1 \dots \mathcal{P}(y_N) dy_N \\ &= \int_{-\infty}^{\infty} dp \int \exp\left[ip\left(\sum_{i=1}^N y_i - X\right)\right] \mathcal{P}(y_1) dy_1 \dots \mathcal{P}(y_N) dy_N \\ &= \int_{-\infty}^{\infty} dp e^{-ipX} \prod_{i=1}^N \int e^{ipy_i} \mathcal{P}(y_i) dy_i = \int_{-\infty}^{\infty} dp e^{-ipX + NG(ip)}. \end{aligned} \quad (25)$$

Here we introduced the generating function  $\langle e^{zy} \rangle \equiv e^{G(z)}$ . The derivatives of the generating function with respect to  $z$  at zero are equal to the moments of  $y$ , while the derivatives of its logarithm  $G(z)$  are equal to the moments of  $(y - \langle y \rangle)$  called cumulants (see exercise).

For large  $N$ , the integral (25) is dominated by the saddle point  $z_0$  such that  $G'(z_0) = X/N$ . This is similar to representing the sum (9) above by its largest term. If there are several saddle-points, the result is dominated by the one giving the largest probability. We assume that contour of integration can be deformed in the complex plane  $z$  to pass through the saddle point without hitting any singularity of  $G(z)$ . We now substitute  $X = NG'(z_0)$  into  $-zX + NG(z)$ , and obtain the large deviation relation (24) with

$$H = -G(z_0) + z_0 G'(z_0). \quad (26)$$

We see that  $-H$  and  $G$  are related by the ubiquitous Legendre transform (which always appear in the saddle-point approximation of the integral Fourier or Laplace transformations). Note that  $NdH/dX = z_0(X)$  and

$$N^2 d^2 H/dX^2 = Nd z_0/dX = 1/G''(z_0).$$

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  $z$ ):

$$\begin{aligned} G''(z) &= \frac{d^2}{dz^2} \ln \int e^{zy} \mathcal{P}(y) dy \\ &= \frac{\int y^2 e^{zy} \mathcal{P}(y) dy \int e^{zy} \mathcal{P}(y) dy - [\int y e^{zy} \mathcal{P}(y) dy]^2}{[\int e^{zy} \mathcal{P}(y) dy]^2} \geq 0. \end{aligned} \quad (27)$$

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  $N \rightarrow \infty$ . Any smooth function is quadratic around its minimum with  $H''(0) = \Delta^{-1}$ , where  $\Delta = G''(0)$  is the variance of  $y$ . Quadratic entropy means Gaussian probability near the maximum — this statement is (loosely speaking) the essence of the central limit theorem. In the particular case of Gaussian  $\mathcal{P}(y)$ , the PDF  $\mathcal{P}(X)$  is Gaussian for any  $X$ . Non-Gaussianity of the  $y$ 's leads to a non-quadratic behavior of  $H$  when deviations of  $X/N$  from the mean are large, of the order of  $\Delta/G'''(0)$ .

One can generalize the central limit theorem and the large-deviation approach in two directions: i) for non-identical variables  $y_i$ , as long as all variances are finite and none dominates the limit  $N \rightarrow \infty$ , it still works with the mean and the variance of  $X$  being given by the average of means and variances of  $y_i$ ; ii) if  $y_i$  is correlated with a finite number of neighboring variables, one can group such "correlated sums" into new variables which can be considered independent.

The above figure and (24,26) show how distribution changes upon adding more numbers. Is there any distribution which does not change upon averaging, that is upon passing from  $y_i$  to  $\sum_{i=1}^N y_i/N$ ? That can be achieved for  $H \equiv 0$ , that is for  $G(z) = kz$ , which corresponds to the Cauchy distribution  $\mathcal{P}(y) \propto (y^2 + k^2)^{-1}$ . Since the averaging decreases the variance, it is no surprise that the invariant distribution has infinite variance. We shall return to distributions invariant under summation of variables considering Renormalization Group in Section 4.3.

**Asymptotic equipartition.** The above law of large numbers state that the sum of  $N$  iid random numbers approaches its mean value as  $N$  grows. One can also look at the given sequence  $y_1, \dots, y_N$  and ask: how probable it is? Can we answer this blatantly self-referential question without seeing other sequences? Yes, we can, if the sequence is long enough. To use the law of large numbers we need to find what to sum. Since the numbers are independent, then the probability of any sequence is the product of probabilities, and the logarithm of the probability is the sum that satisfies the law of large numbers:

$$-\frac{1}{N} \ln p(y_1, \dots, y_N) = -\frac{1}{N} \sum_{i=1}^N \ln \mathcal{P}(y_i) \rightarrow -\langle \ln \mathcal{P}(y) \rangle = S(Y). \quad (28)$$

We see that the log of probability converges to  $N$  times the entropy of  $y$ . But how we find  $S(Y)$  if we don't know  $\mathcal{P}(y)$ ? 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:  $p(y_1, \dots, y_N) = \exp[-NS(y)]$ . That probability is independent of the values  $y_1, \dots, y_N$ , that is the same for all typical sequences. 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".

### 3.2 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<sup>10</sup> function of  $n$ . If we have several independent problems then

---

<sup>10</sup>The messages "in box 2 out of 2" and "in box 2 out of 22" bring the same candy but not the same amount of information.

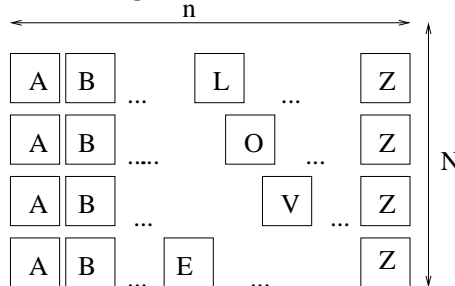


information must be additive. For example, consider each box to have  $m$  compartments. To know in which from  $mn$  compartments is the candy, we need to know first in which box and then in which compartment inside the box:  $I(nm) = I(n) + I(m)$ . Now, we can write (Fisher 1925, Hartley 1927, Shannon 1948)

$$I(n) = I(e) \ln n = k \ln n \tag{29}$$

That information must be a logarithm is clear also from obtaining the missing information by asking the sequence of questions in which half we find the box with the candy, one then needs  $\log_2 n$  of such questions and respective one-bit answers. If we measure information in binary choices or bits then  $I(n) = \log_2 n$ , that is  $k^{-1} = \ln(2)$ . We can easily generalize the definition (29) 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).



In reality, every letter brings on average even less information than  $\log_2 26$

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

$$I_N = k \ln(N! / \prod_i (Np_i)!) \approx -Nk \sum_i p_i \ln p_i + O(\ln N) . \quad (30)$$

The mean information per symbol coincides with the Gibbs entropy (14):

$$S(p_1 \dots p_n) = \lim_{N \rightarrow \infty} I_N / N = -k \sum_{i=1}^n p_i \ln p_i . \quad (31)$$

Here we recognize the asymptotic equipartition from the previous section:  $N$ -string brings the information, which is the log of the number of typical strings:  $I = S$ . Note that when  $n \rightarrow \infty$  then (29) diverges while (31) may well be finite.

Alternatively, one can derive (31) without any mention of randomness. Consider again  $n$  boxes and define  $p_i = m_i / \sum_{i=1}^n m_i = m_i / M$ , where  $m_i$  is the number of compartments in the box number  $i$ . When each compartment can be chosen independently of the box it is in, the  $i$ -th box is chosen with the frequency  $p_i$ , that is a given box is chosen more frequently if it has more compartments. The information on a specific compartment is a choice out of  $M$  and brings information  $k \ln M$ . That information must be a sum of the information about the box  $I_n$  plus the information about the compartment,  $\ln m_i$ , summed over the boxes:  $k \sum_{i=1}^n p_i \ln m_i$ . That gives the information  $I_n$  about the box (letter) as the difference:

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

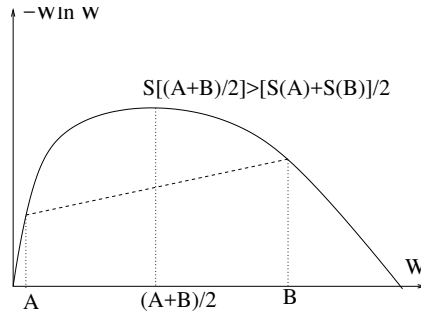
The mean information (31) is zero for delta-distribution  $p_i = \delta_{ij}$ ; it is generally less than the information (29) and coincides with it only for equal

probabilities,  $p_i = 1/n$ , when the entropy is maximum. Indeed, equal probabilities we ascribe when there is no extra information, i.e. in a state of maximum ignorance. In this state, a message brings maximum information per symbol; any prior knowledge can reduce the information. Mathematically, the property

$$S(1/n, \dots, 1/n) \geq S(p_1 \dots p_n) \quad (32)$$

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

$$\frac{1}{n} \sum_{i=1}^n s(p_i) \leq s\left(\frac{1}{n} \sum_{i=1}^n p_i\right). \quad (33)$$



From here one gets the entropy inequality:

$$S(p_1 \dots p_n) = \sum_{i=1}^n s(p_i) \leq n s\left(\frac{1}{n} \sum_{i=1}^n p_i\right) = n s\left(\frac{1}{n}\right) = S\left(\frac{1}{n}, \dots, \frac{1}{n}\right). \quad (34)$$

The relations (33-34) 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 (33) for  $n = 2$ . To get from  $n = 2$  to arbitrary  $n$  we use induction. For that end, we choose  $\lambda = (n - 1)/n$ ,  $a = (n - 1)^{-1} \sum_{i=1}^{n-1} p_i$  and  $b = p_n$  to see that

$$s\left(\frac{1}{n} \sum_{i=1}^n p_i\right) = s\left(\frac{n-1}{n}(n-1)^{-1} \sum_{i=1}^{n-1} p_i + \frac{p_n}{n}\right)$$

$$\begin{aligned}
&\geq \frac{n-1}{n} s \left( (n-1)^{-1} \sum_{i=1}^{n-1} p_i \right) + \frac{1}{n} s(p_n) \\
&\geq \frac{1}{n} \sum_{i=1}^{n-1} s(p_i) + \frac{1}{n} s(p_n) = \frac{1}{n} \sum_{i=1}^n s(p_i) .
\end{aligned} \tag{35}$$

In the last line we used the truth of (33) for  $n - 1$  to prove it for  $n$ .

You probably noticed that (29) corresponds to the microcanonical Boltzmann entropy (8) giving information/entropy as a logarithm of the number of states, while (31) corresponds to the canonical Gibbs entropy (14) giving it as an average. An advantage of Gibbs-Shannon entropy (31) is that it is defined for arbitrary distributions, not necessarily equilibrium.

### 3.3 Communication Theory

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

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

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

Both questions concern redundancy – how unexpected is every letter of the message, on the average. Entropy quantifies redundancy. We have seen that a communication channel transmitting independent symbols on average transmits one unit of the information (31) per symbol. Receiving letter (box) number  $i$  through a binary channel (transmitting ones and zeros)<sup>11</sup> brings information  $\log_2(1/p_i) = \log_2 M - \log_2 m_i$  bits. Indeed, the remaining choice (missing information) is between  $m_i$  compartments. The entropy  $-\sum_{i=1}^z p_i \log_2 p_i$  is the mean information content per letter. Less probable symbols have larger information content, but they happen more rarely.

So the entropy is the mean rate of the information transfer, since it is the mean growth rate of the number of typical sequences. What about the maximal rate of the information transfer? Following Shannon, we answer the question i) statistically, which makes sense in the limit of very long

---

<sup>11</sup>Binary code is natural both for signals (present-absent) and for logic (true-false).

messages, when one can focus on typical sequences, as we did at the end of the Section 3.1 and in deriving (30). Consider for simplicity a message of  $N$  bits, where 0 comes with probability  $1 - p$  and 1 with probability  $p$ . To compress the message to a shorter string of letters that conveys essentially the same information it suffices to choose a code that treats effectively the *typical* strings — those that contain  $N(1 - p)$  zeroes and  $Np$  ones. The number of such strings is given by the binomial  $C_{Np}^N$  which for large  $N$  is  $2^{NS(p)}$ , where  $S(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$ . The strings differ by the order of appearance of 0 and 1. To distinguish between these  $2^{NS(p)}$  messages, we encode any one using a binary string with lengths starting from one and ending at  $NS(p)$  bits. For example, we encode by a one-bit word the message where all  $Np$  ones are at the beginning followed by all  $N(1 - p)$  zeroes, then by two-bit word the message with one hole, etc. The maximal word length  $NS(p)$  is less than  $N$ , since  $0 \leq S(p) \leq 1$  for  $0 \leq p \leq 1$ . In other words, to encode all  $2^N$  sequences we need words of  $N$  bits, but to encode all typical sequences, we need only words up to  $NS(p)$  bits. We indeed achieve compression with the sole exception of the case of equal probability where  $S(1/2) = 1$ . True, the code must include a bit more to represent atypical messages, but in the limit of large  $N$  the chance of their appearance and their contribution to the rate of transmission are negligible. Therefore, entropy sets both the mean and the maximal rate in the limit of long sequences. The idea of typical messages in the limit  $N \rightarrow \infty$  is an information-theory analog of ensemble equivalence in the thermodynamic limit. You may find it bizarre that one uses randomness in treating information communications, where one usually transfers non-random meaningful messages. One of the reasons is that encoding program does not bother to "understand" the message, and treats it as random. Draining the words of meaning is necessary for devising universal communication systems.

Maximal rate of transmission correspond to the shortest mean length of the codeword. Consider an alphabet with  $q$  symbols and the source with the probability distribution  $p(i)$ ,  $i = 1, \dots, q$ . Then Shannon proved that the shortest mean length of the codeword  $\ell$  is bounded by

$$-\sum_i p(i) \log_q p(i) \leq \ell < -\sum_i p(i) \log_q p(i) + 1. \quad (36)$$

Of course, not any encoding guarantees the maximal rate of transmission. Designating sequences of the same length to objects with different probabilities is apparently sub-optimal. To make the mean word length shorter and

achieve signal compression, one codes frequent objects by short sequences and infrequent ones by more lengthy combinations - lossless compressions like zip, gz and gif work this way. Consider a fictional creature whose DNA contains four bases A,T,C,G occurring with probabilities  $p_i$  listed in the table:

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

We want a binary encoding for the four bases. Since there are exactly four two-bit words, one can suggest the Code 1, which has exactly 4 words and uses 2 bits for every base. Here the word length is 2. However, it is straightforward to see that the entropy of the distribution  $S = -\sum_{i=1}^4 p_i \log_2 p_i$  is lower than 2. One then may suggest a variable-length Code 2 (an example of the so-called Huffman code). It is built in the following way. We start from the least probable C and G, which we want to have the longest codewords of the same length differing by one (last) bit that distinguishes between the two of them. We then can combine C and G into a single source symbol with the probability 1/4, that is coinciding with the probability of T. To distinguish from C,G, we code T by two-bit word placing 0 in the second position. Now, we can code A by one-bit word 0 to distinguish it from T,C,G. Alternatively, one can start from the first bit ascribing 0 to A and 1 to T,C,G, then add the second bit to distinguish T from C,G and finish with adding the third bit to distinguish between C and G. Home exercise is to see which code, 1 or 2, uses less bits per base on average. The most efficient code has the length of the mean codeword (the number of bits per base) equal to the entropy of the distribution, which determines the fastest mean transmission rate, that is the shortest mean codeword length.

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

The inequality (32) tells us, in particular, that using an alphabet is not optimal for the information transmission rate as long as the probabilities of the letters are different. We can use less symbols but variable codeword

length. Morse code uses just three symbols (dot, dash and space) to encode any language<sup>12</sup>. In English, the probability of "E" is 13% and of "Q" is 0.1%, so Morse encodes "E" by a single dot and "Q" by "- - · -" (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:

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

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

### 3.4 Correlations in the signals

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", "H" often follows "T", etc). That will further lower the entropy.

A simple universal model with neighboring correlations is a Markov chain. It is specified by the conditional probability  $p(j|i)$  that the letter  $i$  is followed by  $j$ . For example  $p(U|Q) = 1$ . The probability is normalized for every  $i$ :  $\sum_j p(j|i) = 1$ . The matrix  $p_{ij} = p(j|i)$ , whose elements are positive and in every column sum to unity, is called stochastic. The vector of probabilities  $p(i)$  and the transition matrix  $p_{ij}$  are not independent but are related by the detailed balance:  $p(i)p_{ij} = p(j)p_{ji}$ . Summing over  $j$ , we obtain  $p(i) = \sum_j p(j)p_{ji}$ , that is  $\mathbf{p} = \{p(a), \dots, p(z)\}$  is an eigenvector with the unit eigenvalue of the matrix  $p_{ij}$ .

The probability of any  $N$ -string is then the product of  $N - 1$  transition probabilities times the probability of the initial letter. As in (28), minus the logarithm of the probability of a long  $N$ -string grows linearly with  $N$ :

$$\log_2 p(i_1, \dots, i_N) = \log_2 p(i_1) + \sum_{k=2}^N \log_2 p(i_{k+1}|i_k). \quad (37)$$

Therefore, the number of typical sequences starting from  $i$  grows with  $N$  exponentially, as  $2^{NS}$ , where  $S$  is the conditional entropy  $-\sum_j p(j|i) \log_2 p(j|i)$ ,

---

<sup>12</sup>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.

which must be averaged over different  $i$  with their probabilities  $p(i)$ . That way we express the language entropy via  $p(i)$  and  $p(j|i)$  by averaging over  $i$  the entropy of the transition probability distribution:

$$S = - \sum_i p_i \sum_j p(j|i) \log_2 p(j|i) . \quad (38)$$

That formula defines the information rate of the Markov source. We shall further discuss Markov chains describing Google PageRank algorithm in Section 5.3 below.

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

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

Note in passing that the human language encodes meaning 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 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.

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



So what is so special about alphabet? Redundant encodings are many. Note first that the human language encodes meaning 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 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.

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

Another encoding of profound importance is a positional numeral system, based on the fundamental discovery that number  $N$  can be encoded by  $\log N$  symbols instead of  $N$  times repeating the same mark. One cannot overestimate the importance of encoding where the value depends on the position, since it already implies algebraic operations. Indeed, reading (decoding) requires multiplying and adding:  $2021 = 2 \times 1000 + 2 \times 10 + 1$ . It then allowed simple automatic rules for computations (formulated by Persian al-Khwarizmi, from whose name the word algorithm appeared).

### 3.5 Mutual information as a universal tool

Answering the question i) in Sect. 3.3, 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 presently are the most powerful and universal tools of information theory.

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

$$\begin{aligned} S(B|A) &= \sum_i P(A_i) S(B|A_i) = -\sum_{ij} P(A_i) P(B_j|A_i) \log_2 P(B_j|A_i) \\ &= -\sum_{ij} P(A_i, B_j) \log_2 P(B_j|A_i) . \end{aligned} \quad (39)$$

We already encountered it in (38) considering correlations between subsequent terms in the sequence. In this subsection we use conditional probability between input and output. 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*:

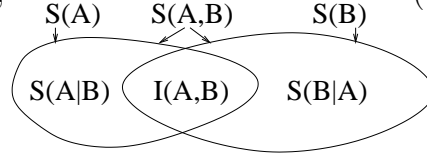
$$I(A, B) = S(B) - S(B|A) = \sum_{ij} P(A_i, B_j) \log_2 \left[ \frac{P(B_j|A_i)}{P(B_j)} \right] . \quad (40)$$

Indeed, information is a decrease in uncertainty. Non-negativity of the mutual information means that on average measurements lower uncertainty by

increasing the conditional probability relative to unconditional :  $\langle \log_2[P(B_j|A_i)/P(B_j)] \rangle \geq 0$ . For example, let  $B$  be a choice out of  $n$  equal possibilities:  $P(B) = 1/n$  and  $S(B) = \log_2 n$ . If for every  $A_i$  we can have  $m$  different values of  $B$ , then  $S(B|A) = \log_2 m$  and  $I(A, B) = \log_2(n/m)$  bits. When there is one-to-one correspondence,  $m = 1$  and  $A$  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,

$$S(A, B) = S(A) + S(B|A) = S(B) + S(A|B), \quad (41)$$



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

$$I(A, B) = S(B) - S(B|A) = S(A) + S(B) - S(A, B) = S(A) - S(A|B). \quad (42)$$

Exactly like in the above  $m - n$  example, when  $A$  is a choice out of  $k$  equal possibilities and for every input  $B_i$  we can have  $l$  different equally probable values of  $A$ , then  $S(A|B) = \log_2 l$  and  $I(A, B) = S(A) - S(A|B) = \log_2(k/l)$  bits.

To avoid confusion, let us state the obvious: there is no symmetry between  $A$  and  $B$ . They could be of very different nature - one is the position of an atom, another is the reading of the device, for instance. Neither their entropies,  $S(A)$  and  $S(B)$ , nor the conditional entropies,  $S(B|A)$  and  $S(A|B)$ , are generally equal or even comparable. Yet the degree of their correlation  $I(A, B)$  is a symmetric function. It is important that the mutual information  $I(A, B)$  is a universal measure of correlation, insensitive to the nature of the relationship between  $A$  and  $B$ , whether it is linear or nonlinear, direct or inverse, etc. On the contrary, correlation functions cannot serve as a universal measure: for instance,  $\langle AB \rangle$  is a proper measure of correlation for a Gaussian distribution only.

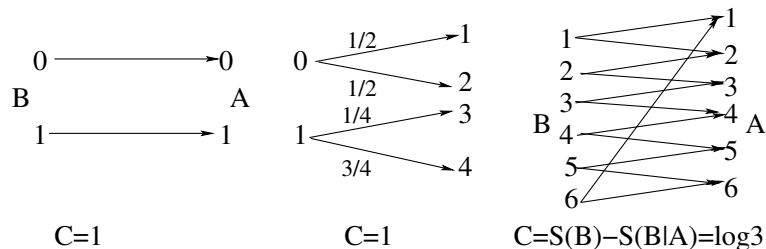
When  $A$  and  $B$  are independent, the joint entropy is a sum, and the information is zero. When they are dependent,  $P(B, A) > P(A)P(B)$ , so that that the information is indeed positive. When  $A, B$  are related deterministically,  $S(A) = S(B) = S(A, B) = I(A, B)$ , where  $S(A) = -\sum_i P(A_i) \log_2 P(A_i)$ , 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:

$$S(A) + S(B) > S(A, B). \quad (43)$$

If the mutual information is what on average brings an imperfect channel, how reliable it is? It is tempting to assume that the mutual information plays for noisy channels the same role the entropy plays for ideal channels, in particular, sets the maximal rate of reliable communication in the limit of long messages, thus answering the question ii) from the Section 3.3. Indeed, if there are different outputs for the same input, like in the above simple  $k - l$  example, the rate of information transfer is lower than for a one-to-one correspondence, since we need to divide our  $k$  outputs into groups of  $l$ , distinguishing only between the groups. More formally, for each typical  $N$ -sequence of independently chosen  $B$ -s, we have  $[P(A|B)]^{-N} = 2^{NS(A|B)}$  possible output sequences, all of them equally likely. To get the rate of the useful information about distinguishing the inputs, we need to divide the total number of typical outputs  $2^{NS(A)}$  into sets of size  $2^{NS(A|B)}$  corresponding to different inputs. Therefore, we can distinguish at most  $2^{NS(A)}/2^{NS(A|B)} = 2^{NI(A,B)}$  sequences of the length  $N$ , which sets  $I(A, B)$  as the maximal rate of information transfer.

However, that was a rather trivial case when inputs can be distinguished from outputs without errors. Real problem start when errors are made, for instance, when a single output can correspond to different inputs. Here, one may argue that taking the limit of large  $N$  does not help since the channel continues to make errors all the time. And yet Shannon have shown (in the co-called noisy channel theorem) that one can keep a finite transmission rate and yet make the probability of error arbitrary small at the limit  $N \rightarrow \infty$ . The idea is that to correct errors one needs to send extra bits, so to get the rate we need to compute how many bits are devoted to error correction and how many to transferring the information itself. To do that, we need to characterize the channel itself as specified by  $P(B|A)$ . Let us maximize  $I(A, B)$  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 in bits per symbol:

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



The channel capacity is the log of the maximal number of distinguishable signals. For example, if our channel transmits the binary input exactly (zero to zero, one to one), then the capacity is 1 bit, which is achieved by choosing  $P(B = 0) = P(B = 1) = 1/2$ , see the left panel in the Figure. Let us stress that if  $P(B) \neq P(A)$ , then the average rate is less than the capacity (one bit per symbol) despite the channel being perfect. Even if the channel has many outputs for every input out of  $n$ , the capacity is still  $\log_2 n$ , if those outputs are non-overlapping for different inputs, so that the input can be determined without an error and  $P(B|A) = 1$ . Such case is presented in the middle panel in the Figure. In this case, the transfer rate is determined by the number of  $B$ -states; from the perspective of  $A$ -states, the rate is  $S(A) - S(A|B) = 2 - 1 = 1$ . Like the mutual information, the capacity 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  $C = \log_2(n/m)$  bits (in the right panel of the Figure  $n = 6$ ,  $m = 2$ ). Lowered capacity means increased redundancy, that is a need to send more symbols to convey the same information. As mentioned, shorter alphabet requires longer messages.

Let us treat at last to the most generic case, when noise does not allow to separate inputs into groups with completely disjoint outputs, so errors are 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 has shown that it is possible, if there is any correlation between output  $A$  and input  $B$ , that is  $C > 0$ . Then the probability of an error can be made  $2^{-N(C-R)}$ , that is asymptotically small in the limit of  $N \rightarrow \infty$ , 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).

For example, if in a binary channel the probability of every single bit going wrong is  $q$ , then  $A$  is binary random variable with equal probabilities of 0 and 1, so that  $S(A) = \log_2 2 = 1$ . Conditional probabilities are  $P(1|0) = P(0|1) = q$  and  $P(1|1) = P(0|0) = 1 - q$ , so that  $S(A|B) = S(B|A) = S(q) = -q \log_2 q - (1 - q) \log_2 (1 - q)$ . The mutual information  $I(A, B) = S(A) - S(A|B) = 1 - S(q)$ . This is actually the maximum, that is the channel capacity:  $\mathcal{C} = \max_{P(B)} [S(B) - S(B|A)] = 1 - S(q)$ , because the maximal entropy is unity for a binary variable  $B$ . Let us now see how the rate of transmission is bounded from above by the capacity. In a message of length  $N$ , there are on average  $qN$  errors and there are  $N!/(qN)!(N - qN)! \approx 2^{NS(q)}$  ways to distribute them. We then need to devote some  $m$  bits in the message not to data transmission but to error correction. Apparently, the number of possibilities provided by these extra bits,  $2^m$ , must exceed  $2^{NS(q)}$ , which means that  $m > NS(q)$ , and the transmission rate  $R = (N - m)/N < 1 - S(q)$ . The channel capacity is zero for  $q = 1/2$  and is equal to 0.988 bits per symbol for  $q = 10^{-3}$ . The probability of errors is binomial with the mean number of errors  $qN$  and the standard deviation  $\sigma = \sqrt{Nq(1 - q)}$ . If we wish to bound the error probability from above, we must commit to correcting more than the mean number of errors, making the transmission rate smaller than the capacity.

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

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

$$I(A, B) = S(A) - S(\xi) . \quad (44)$$

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

We have seen in the previous section that the mutual information between letters lowered the entropy of the language from the one-letter entropy,

$-\sum_i p(i) \log p(i)$ . That lowering is brought by the knowledge of the conditional probabilities  $p(j|i)$ ,  $p(j|i, k \dots)$ , which is more than knowledge of  $p(i)$ .

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  $A$ :

$$P(B|A) = P(B) \frac{P(A|B)}{P(A)} . \quad (45)$$

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 the example at the beginning of Section 4.4). After that one obtains the conditional probability (45) that allows animal/robot to perceive the environment and function effectively in that habitat.

**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 (45) 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  $A$ :

$$\bar{B} = \int B P(B|A) dB = \frac{A}{\mathcal{S} + \mathcal{N}} = A \frac{SNR}{1 + SNR} , \quad (46)$$

where signal to noise ratio is  $SNR = \mathcal{S}/\mathcal{N}$ . The rule (46) 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 (44):

$$\begin{aligned} I(A, B) &= S(A) - S(A|B) = S(A) - S(B + \xi|B) = S(A) - S(\xi|B) = S(A) - S(\xi) \\ &= \frac{1}{2} [\log_2 2\pi e(\mathcal{S} + \mathcal{N}) - \log_2 2\pi e\mathcal{N}] = \frac{1}{2} \log_2(1 + SNR) . \end{aligned} \quad (47)$$

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  $\rho$  we obtain  $\rho(x) \propto \exp(-\lambda x^2)$ .

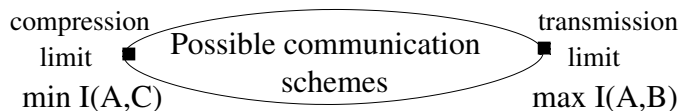
**Examples of redundancy.** How redundant is the genetic code? There are four bases, which must encode twenty amino acids. There are  $4^2$  two-letter words, which is not enough. The designer then must use a triplet code with  $4^3 = 64$  words, so that the redundancy factor is about 3. Number of ways to encode a given amino acid 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 4.2.

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.

Mutual information also sets the limit on the data compression  $A \rightarrow C$ , 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.

### 3.6 Hypothesis testing and relative entropy

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:

$$P(h|e) = P(h) \frac{P(e|h)}{P(e)} . \quad (48)$$

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 international conspiracy to increase the power of governments and the evidence is COVID pandemic. In this case  $P(e|h)$  is high: a pandemic provoking increase of the state power is highly likely *given* such a conspiracy exists. This is presumably why some people stop thinking here and accept the hypothesis. But of course, absent such an event, the prior probability  $P(h)$  could be 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

$$P(h_1|e) = P(h_1) \frac{P(e|h_1)}{P(e)} = P(h_1) \frac{P(e|h_1)}{P(h_1)P(e|h_1) + P(h_2)P(e|h_2)} . \quad (49)$$

Indeed, because our hypothesis are mutually exclusive, the total probability of the evidence consists of two terms:  $P(e) = P(e, h_1) + P(e, h_2) = P(h_1)P(e|h_1) + P(h_2)P(e|h_2)$ . 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:

$$\frac{1}{P(h_1|e)} = 1 + \frac{P(h_2)P(e|h_2)}{P(h_1)P(e|h_1)}. \quad (50)$$

For example, 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 out alternative rather than supports the hypothesis. This is why even good tests, with  $P(e|h_1)$  close to unity and  $P(e|h_2)$  small, are not very reliable at the beginning of a pandemic, when  $P(h_1)$  is small. The same is true for drug test in a mostly clean population. Suppose that a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users (hypothesis  $h_1$ ) and 99% true negative results for clean people (hypothesis  $h_2$ ). If we denote  $e$  the positive test result, then  $P(e|h_1) = 0.99$  and  $P(e|h_2) = 1 - 0.99 = 0.01$ . Suppose that 0.5% of people are 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 than half. The result is more sensitive to specificity approaching unity, when  $P(e|h_2) \rightarrow 0$ , than to sensitivity.

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. Observing our own mental processes gives us both the idea of logic and of statistical inference.

Note a shift in the 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 re-

sulting 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 number  $N$  of trials is sufficient to invalidate our hypothesis? 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:

$$\mathcal{P} = \prod_i q_i^{p_i N} \frac{N!}{\prod_j (p_j N)!}. \quad (51)$$

This is the probability of our hypothetical distribution being true. Considering limit of large  $N$  we obtain a large-deviation-type relation like (24):

$$\mathcal{P} \propto \exp \left[ -N \sum_i p_i \ln(p_i/q_i) \right]. \quad (52)$$

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

$$D(p|q) = \sum_i p_i \ln(p_i/q_i) = \langle \ln(p/q) \rangle. \quad (53)$$

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 (28) 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  $k = 1, \dots, K$ . Then the asymptotic equipartition (the law of large numbers) advises 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 the frequencies we observe in a finite sequence are generally somewhat different from the true probabilities of the  $\{p_k\}$ . Then the positivity of the relative entropy guarantees that the asymptotic equipartition underestimates the probability of the sequence, the true probability is actually high-

her:  $\prod_k p_k^{Nq_k} = \exp(N \sum_k q_k \ln p_k) = \exp[N \sum_k (q_k \ln q_k + q_k \ln(p_k/q_k))] = \exp\{-N[S(q) - D(q|p)]\}$ .

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  $0, 1, \dots, N$  then the number of possible  $K$ -vectors  $\{q_k\}$  is at most  $(N + 1)^K$ , which grows with  $N$  only polynomially, where the alphabet size  $K$  sets the power. 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  $i$ . This can be readily demonstrated using the simple inequality  $\ln x \leq x - 1$ , which turns into equality only for  $x = 1$ .

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

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  $S(p) \leq \bar{l} \leq S(p) + 1$ , that is non-optimality is at most one bit.

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

Relative entropy allows also to generalize the second law for non-equilibrium

processes. Entropy itself can either increase upon evolution towards thermal equilibrium or decrease upon evolution towards a non-equilibrium state, as seen in Section 2.3. 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  $\tau$  when the latter exceeds the correlation time.

## 4 Applications of Information Theory

My brothers are protons, my sisters are neurons  
Gogol Bordello "Superttheory 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.

### 4.1 Distribution from information

So far, we defined entropy and information via the distribution. In practical applications, however, the distribution is usually unknown and we need to guess it from some data. The use of information does that. 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? Here we also include normalization:  $R_0 = r_0 = 1$ . 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, which is the entropy  $S = -\langle \ln \rho \rangle$ . This is to provide for the widest set of possibilities for future use, compatible with the existing information. Looking for the extremum of

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

we differentiate it with respect to  $\rho(x, t)$  and obtain the equation  $\ln[\rho(x, t)] = -1 + \sum_j \lambda_j R_j(x, t)$  which gives the distribution

$$\rho(x, t) = \frac{1}{Z} \exp\left[-1 + \sum_j \lambda_j R_j(x, t)\right]. \quad (54)$$

The normalization factor

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

can be expressed via the measured quantities by using

$$\frac{\partial \ln Z}{\partial \lambda_i} = r_i . \quad (55)$$

The distribution (54) corresponds to the entropy extremum, but how we know that it is the maximum? Positivity of relative entropy proves that. Indeed, consider any other normalized distribution  $g$  which satisfies the constraints:  $\int dx g R_j(x) = r_j$ . Then

$$\int dx g \ln \rho = -1 + \sum_j r_j = \int dx \rho \ln \rho = -S(\rho)$$

so that

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

Gibbs distribution is (54) with  $R_1$  being energy. When it is the kinetic energy of molecules, we have Maxwell distribution; when it is potential energy in an external field, we have Boltzmann distribution. For our initial "candy-in-the-box" problem (think of an impurity atom in a lattice if you prefer physics), let us denote the number of the box with the candy  $j$ . Different attempts give different  $j$  but on average after many attempts we find, say, the mean value  $\langle j \rangle = r_1$ . The distribution giving maximal entropy for a fixed mean is exponential, which in this case is the geometric distribution:  $\rho(j) = (1 - p)p^j$ , where  $p = r_1/(1 + r_1)$  (home exercise). Similarly, if we scatter on the lattice X-ray with wavenumber  $k$  and find  $\langle \cos(kj) \rangle = 0.3$ , then

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

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

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

Note in passing that the set of equations (55) may be self-contradictory or insufficient so that the data do not allow to define the distribution or allow it non-uniquely. For example, consider  $R_i = \int x^i \rho(x) dx$  for  $i = 0, 1, 2, 3$ . Then (54) cannot be normalized if  $\lambda_3 \neq 0$ , but having only three constants  $\lambda_0, \lambda_1, \lambda_2$  one generally cannot satisfy the four conditions. That means that we cannot reach the entropy maximum, yet one can prove that we can come arbitrarily close to the entropy of the Gaussian distribution  $\ln[2\pi e(r_2 - r_1^2)]^{1/2}$ .

If, however, the extremum is attainable, then (54) defines the information still missing after the measurements:  $S\{r_i\} = -\sum_j \rho(j) \ln \rho(j)$ . It is analogous to thermodynamic entropy as a function of (measurable) macroscopic parameters. It is clear that  $S$  have a tendency to decrease whenever we add a constraint by measuring more quantities  $R_i$ . Making an extra measurement  $R$  one changes the distribution from  $\rho(x)$  to (generally non-equilibrium)  $\rho(x|R)$ , which has its own *conditional* entropy

$$S(x|R) = -\int dx dR \rho(R) \rho(x|R) \ln \rho(x|R) = -\int dx dR \rho(x, R) \ln \rho(x|R) . \quad (56)$$

The conditional entropy quantifies my remaining ignorance about  $x$  once I know  $R$ . Measurement decreases the entropy of the system by the mutual information (40,42) — that how much information about  $x$  one gains:

$$\begin{aligned} S(x|R) - S(x) &= -\int \rho(x|R) \ln \rho(x|R) dx dR + \int \rho(x) \ln \rho(x) dx \\ &= \int \rho(x, R) \ln[\rho(x, R)/\rho(x)\rho(R)] dx dR = S(x, R) - S(R) - S(x) . \end{aligned} \quad (57)$$

But all our measurements happen in a real world at a finite temperature. Does it matter? Yes, it determines the energy cost of measurements. Assume that our system is in contact with a thermostat having temperature  $T$ , which by itself does not mean that 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 (57) 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 4.2.

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 losing information is due to separation of trajectories described in Sect. ???. 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 (54) takes the canonical form (16,17) 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 (54) 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.

Mention in passing the suggestions to use relative information and mutual entropy for a more ambitious task of quantifying consciousness, understood as processing information from different channels in an integrated way, irre-

ducible to processing information in the channels separately. Such approach is known as integrated information theory (Tononi 2008).

## 4.2 Exorcizing Maxwell demon

Demon died when a paper by Szilárd appeared, but it continues to haunt the castles of physics as a restless and lovable poltergeist.

P Landsberg, quoted from Gleick "The Information"

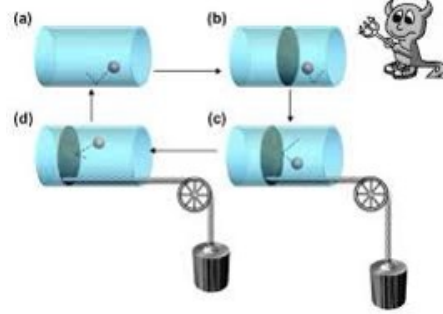
Here we want to address the relation between information and energy, particularly, find out if information has any energy price. Since energy and entropy (information) have different dimensionalities, we need something to relate them. For example, this can be temperature, which is the derivative of the 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 (31) and the mutual information (40,57) can be defined for arbitrary distributions. As we mentioned after (57), 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<sup>13</sup> (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

---

<sup>13</sup>that entropy is not a property of the system but of our knowledge about the system

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). Such work needs to be done for any entropy change by a measurement (57). This is true for an ideal (or demonic) observer, which does not change its state upon measurements. In a general case, the entropy change is the difference between the entropy of the system  $S(A)$  and the entropy of the system interacting with the measuring device  $S(A, M)$ . When there is a change in the free energy  $\Delta F_M$  of the measuring device, the measurement work is

$$W_{meas} \geq T\Delta S + \Delta F_M = T[S(A) - S(A, M)] + \Delta F_M . \quad (58)$$

That guarantees that we cannot break the first law of thermodynamics. But we just turned thermal energy into work. Can we then break the second law by constructing a perpetuum mobile of the second kind, regularly measuring particle position and using its thermal energy to do work? Our demonic engine now includes both the working system A and the measuring device M. To make a full thermodynamic cycle, we need to return the demon's memory to the initial state. What is the energy price of *erasing* information? Such erasure involves compression of the phase space and is irreversible. For example, to erase information in which half of the box the particle is, we may compress the box to move the particle to one half irrespective of where it was. That compression decreases entropy and is accompanied by the heat  $T \ln 2$  released from the system to the environment. If we want to keep the temperature of the system, we need to do exactly that amount of work compressing the box (Landauer 1961). In other words, demon cannot get more

work from using the information  $S(M)$  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:

$$W_{eras} \geq TS(M) - \Delta F_M . \quad (59)$$

Together, the energy price of the cycle,

$$W_{eras} + W_{meas} \geq T[S(A) + S(M) - S(A, M)] = TI , \quad (60)$$

can be recognized as the temperature times what was defined in the Section 3.5 as *the mutual information*. 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 (60) 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 (58,59,60) are versions of the second law of thermodynamics, in which information content and thermodynamic variables are treated on an equal footing.

Similarly, in the original Maxwell scheme, the demon observes the molecules as they approach the shutter, allowing fast ones to pass from A to B and slow ones from B to A. Creation of the temperature difference with a negligible expenditure of work lowers the entropy precisely by the amount of information that the demon collected. Erasing this information will require work.

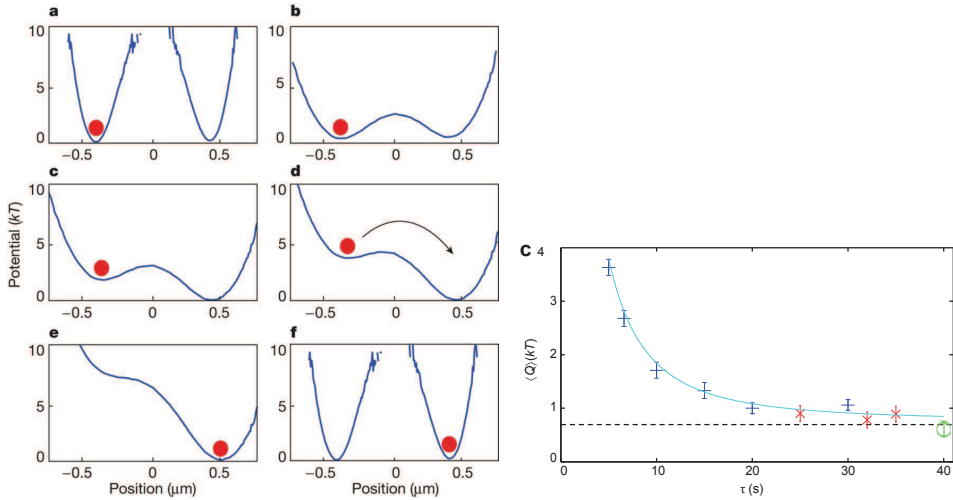
Landauers principle not only exorcizes Maxwells demon, but also imposes the fundamental physical limit on computations. Performing standard operations independent of their history requires irreversible acts (which do not have single-valued inverse). Any Boolean function that maps several input states onto the same output state, such as AND, NAND, OR and XOR, is logically irreversible. When a computer does logically irreversible operation the information is erased and heat must be generated. It is worth stressing that one cannot make this heat arbitrarily small making the process adiabatically slow:  $T \ln 2$  per bit is the minimal amount of dissipation to erase a bit at a fixed temperature<sup>14</sup>.

---

<sup>14</sup>In principle, any computation can be done using only reversible steps, thus eliminating the need to do work (Bennett 1973). That will require the computer to reverse all the steps after printing the answer.

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

**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  $U(x, t)$ :

$$W = Q(\tau) = - \int_0^\tau \dot{x}(t) \frac{\partial U(x, t)}{\partial x} dt . \quad (61)$$

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

$$\langle Q \rangle \geq -T \Delta S = T \ln 2 . \quad (62)$$

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 5.2 where we present a generalization of (58,59).

### 4.3 Renormalization group and information loss

Erase the features Chance installed,  
and you will see the world's great beauty<sup>15</sup>.

A Blok

---

<sup>15</sup>Erase the features Chance installed. Watch by chance do not rub a hole. V Nekrasov

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 loses information step by step and what universal features appear in the process. Most often we lose 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 asymptotic behavior.

As almost everything in this course, the simplest realization of RG refers to summing random numbers, the procedure introduced in the Section 3.1. The small twist is that now we do summation step by step, summing two numbers at every step. Consider a set of random iid variables  $\{x_1 \dots x_N\}$ , each having the probability density  $\rho(x)$  with zero mean and unit variance. The two-step RG reduces the number of random variables by replacing any two of them by their sum and re-scales the sum to keep the variance:  $z_i = (x_{2i-1} + x_{2i})/\sqrt{2}$ . Since summing doubles the variance we divided by  $\sqrt{2}$ . The new random variables each has the following distribution:

$$\rho'(z) = \sqrt{2} \int dx dy \rho(x) \rho(y) \delta(x + y - z\sqrt{2}) . \quad (63)$$

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

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

Since this is a convolution equation, the simplest is to solve it by the Fourier transform,  $\rho(k) = \int \rho(x) e^{ikx} dx$ , which gives

$$\rho(k\sqrt{2}) = \rho^2(k) . \quad (64)$$

We may also say that  $\rho(k)$  is the generating function, which is multiplied upon summation of independent variables. The solution of (64) is  $\rho_0(k) \sim e^{-k^2}$  and

$\rho_0(x) = (2\pi)^{-1/2}e^{-x^2/2}$ . We thus have shown that the Gaussian distribution is a fixed point of repetitive summation and re-scaling of random variables.

To turn that into the central limit theorem, we need also to show that this distribution is linearly stable, that is RG indeed flows towards it. Near the fixed point,  $\rho = \rho_0(1 + h)$ , the transform can be linearized in  $h$ , giving  $h'(k) = 2h(k/\sqrt{2})$ . The eigenfunctions of the linearized transform are  $h_m = k^m$  with eigenvalues  $2^{1-m/2}$ . There are three conservation laws of the transformation (63): the moments  $\int x^n \rho(x) dx$  must be preserved for  $n = 0$  (normalization),  $n = 1$  (zero mean) and  $n = 2$  (unit variance). The moments of  $\rho(x)$  are derivatives of the generation function  $\rho(k)$  at  $k = 0$ . Therefore, the three conservation laws mean that  $h(0) = h'(0) = h''(0) = 0$ , so that only  $m > 2$  are admissible, which means stability, that is deviations from the fixed point decrease. To conclude, in the space of distributions with the same variance, the RG-flow eventually brings us to the distribution with the maximal entropy, forgetting all the information except the invariants - normalization, the mean and the variance.

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

$$\rho(z) = \int \rho(x)\rho(y)\delta(z - x/2 - y/2) dx dy \Rightarrow \rho(k) = \rho^2(k/2).$$

It has the solution  $\rho(k) = \exp(-|k|)$  and  $\rho(x) = (1 + x^2)^{-1}$ , which is the Cauchy distribution mentioned in Section 3.1. In this case, the distribution has an infinite variance, and RG preserves only the mean (which is zero) and normalization. More generally, one can consider  $z_i = (x_{2i-1} + x_{2i})/2^\mu$  and obtain the family of universal distributions,  $\rho(k) = \exp(-|k|^\mu)$ .

## 4.4 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. Life-sustaining substance is entropy: 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 of chemical energy into 100 grams of a chocolate, and you enjoy it even more<sup>16</sup>.

The evolution as a natural selection is an increasingly efficient encoding of information about the environment in the gene pool of its inhabitants. This process is greatly accelerated by sex, which still provides one of the highest transfer rates of information (even though most of it is discarded).

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

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

$T_1$
$s_1 = Q/T_1$
$\downarrow$
$Q$
$\downarrow$
$s_2 = (Q - W)/T_2$
$\downarrow$
$Q - W$
$\downarrow$
$T_2$

When  $T_1 \rightarrow T_2$ , 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.

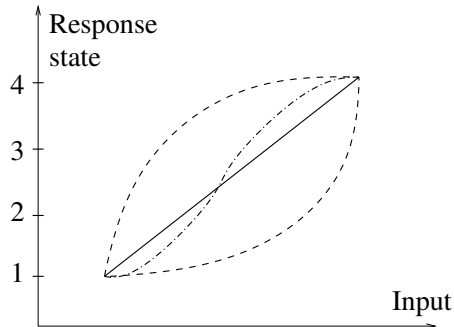
Let us now estimate our rate of information processing and entropy production. A human being dissipates about  $W = 200$  watts of power at  $T = 300 K$ . Since the Boltzmann constant is  $k = 1.38 \times 10^{-23}$ , that gives about  $W/kT \simeq 10^{23}$  bits per second. The amount of information 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. Technically, the problem is to choose thresholds for switching to the next level of response, or equivalently, to choose the function of the input for which we take equidistant thresholds. Suppose that we wish to divide the whole perceivable interval of signals into three regions, encoding them as weak (1,2), medium (2,3) and strong (3,4):

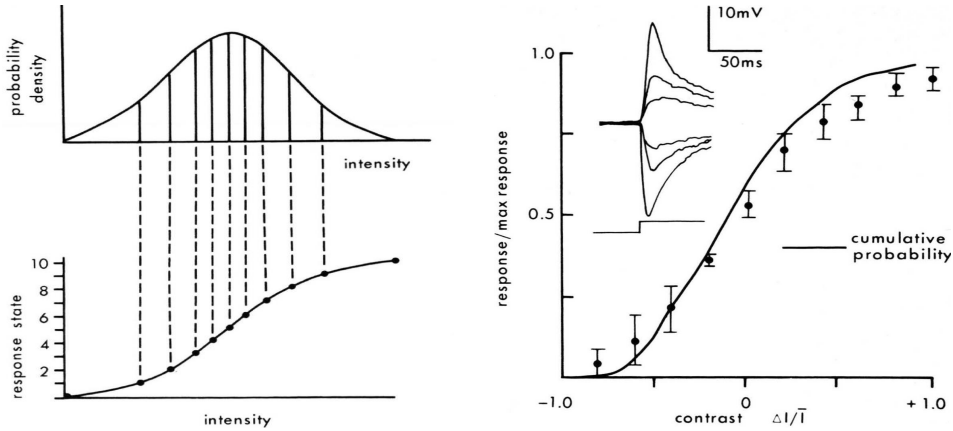
---

<sup>16</sup>Nor we consume matter, only make it more disordered: what we consume has much lower entropy than what comes out of our excretory system



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

If, however, 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:  $I(x, y) = S(y) - S(y|x)$ . Originally, it was assumed that the transmission is error-free, so that the conditional entropy  $S(y|x)$  is zero, but more realistically, we can assume that it is fixed anatomically and does not depend on the input statistics. Therefore, according to Section 3.5, we need to maximize the entropy of the output assuming that the input statistics  $\rho(x)$  is given. Absent any extra constraints except normalization, the entropy is maximal when  $\rho(y)$  is constant. Indeed, since  $\rho(y)dy = \rho(x)dx = \rho(x)dydx/dy = \rho(x)dy/g'(x)$ , then

$$S(y) = - \int \rho(x) \ln[\rho(x)/g'(x)] dx = S(x) + \langle \ln[g'(x)] \rangle, \quad (65)$$

$$\frac{\delta S}{\delta g} = \frac{\partial}{\partial x} \frac{\rho}{g'(x)} = 0 \quad \Rightarrow \quad g'(x) = C\rho(x),$$

as in the Figure. In other words, we choose equal bins for the variable whose probability is flat. Since the probability  $\rho(x)$  is positive, the response function  $y = g(x)$  is always monotonic i.e. invertible. Note that our choice of response function is an exact analog of using longer codewords for less frequent letters. In that way, we utilized only the probability distribution of different signal levels, similar to language encoding which utilizes different frequencies of letters (and not, say, their mutual correlations). We have also applied quasi-static approximation, neglecting dynamics. 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 works when the conditional entropy  $S(y|x)$  is independent of the form of the response function  $y = g(x)$ .

Of course, the eye of any living being provides not a single input signal, but the whole picture. Let us now pass from a single channel to  $N$  inputs and outputs (neurons/channels). Consider a network with an input vector  $\mathbf{x} = (x_1, \dots, x_N)$  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:

$$\rho(\mathbf{y}) = \frac{\rho(\mathbf{x})}{\det[\partial y_i / \partial x_k]}, \quad (66)$$

Making it flat (distribute outputs uniformly) for maximal capacity is not straightforward now. In one dimension, it is enough to follow the gradient to arrive at an extremum, but there are many possible paths to the mountain summit. Maximizing the total mutual information between input and output, which requires maximizing the output entropy, is often (but not always) achieved by minimizing first the mutual information between the output components. For two outputs we may start by maximizing  $S(y_1, y_2) = S(y_1) + S(y_2) - I(y_1, y_2)$ , that is minimize  $I(y_1, y_2)$ . If we are lucky and find encoding in terms of independent components, then we choose for each component the transformation (65), which maximizes its entropy making the respective probability flat. For a good review and specific applications to visual sensory processing see Atick 1992.

**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 impor-

tant 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).

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

$$E = \sum_i S(y_i) - \beta[S(\mathbf{y}) - S(\mathbf{x})] = \sum_i S(y_i) - \beta \ln \det[\partial y_i / \partial x_k]. \quad (67)$$

A coding is considered to yield an improved representation if it possesses a smaller value of  $E$ . The choice of the parameter  $\beta$  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). More sophisticated schemes employ not only mutual information, but also interaction information (??). 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.

## 4.5 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 and thus the amount of information needed to characterize it.

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

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

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  $\mathcal{D}$ . Since  $I(A, B) = S(B) - S(B|A)$ , then minima of  $I(A, B)$  are maxima of  $S(B|A)$ . It is helpful to think of distortion as produced by the added noise  $\xi$  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 (47):

$$R(\mathcal{D}) = I(A, B) = S(B) - S(B|A) = S(B) - S(B - A|A)$$

$$\geq S(B) - S(B - A) = \frac{1}{2} \log_2(2\pi e\sigma^2) - \frac{1}{2} \log_2(2\pi e\mathcal{D}) = \frac{1}{2} \log_2 \frac{\sigma^2}{\mathcal{D}} . \quad (68)$$

It turns into zero for  $\mathcal{D} = \sigma^2$  and goes to infinity for  $\mathcal{D} \rightarrow 0$ . Presenting it as  $\mathcal{D}(R) = \sigma^2 2^{-2R}$ , we see that every extra 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. When the distortion is not a quadratic function, the conditional probability is not Gaussian. In solving practical problems, it must be found solving the variational problem, where one finds a normalized  $P(B|A)$ , which minimizes the mutual information under the condition of a given mean distortion. For that one minimizes the functional

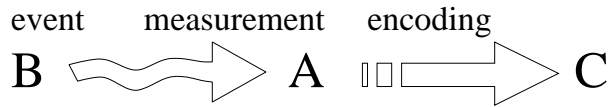
$$F = I + \beta\mathcal{D} = \sum_{ij} P(B_j|A_i)P(A_i) \left\{ \ln \frac{P(B_j|A_i)}{P(B_j)} + \beta d(A_i, B_j) \right\} . \quad (69)$$

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

$$P(B_j|A_i) = \frac{P(B_j)}{Z(A_i, \beta)} e^{-\beta d(A_i, B_j)} , \quad (70)$$

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  $\beta$  reflects our priorities: at small  $\beta$  the conditional probability is close to the unconditional one, that is we minimize information without much regard to the distortion. On the contrary, large  $\beta$  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/infering  $B$ , 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

$$I(C, A) - \beta I(C, B) . \tag{71}$$

The coding  $A \rightarrow C$  is also generally stochastic, characterized by  $P(C|A)$ . The quality of the coding is determined by the rate, that is by the average number of bits per message needed to specify an element in the codebook without confusion. This number per element  $A$  of the source space  $\{A\}$  is bounded from below by the mutual information  $I(C, A)$  which we thus want to minimize. Effective coding utilizes the fact that mutual information is usually sub-extensive in distinction from entropy which is extensive. Note the difference from the Section 3.5, 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 (71). 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  $\beta$  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  $\beta$  grows, we are pushed toward detailed quantization. By varying  $\beta$  one can explore the tradeoff between the preserved meaningful information and compression at various resolutions. Comparing with the rate distortion theory functional (70), we recognize that we are looking for the conditional probability of the mapping  $P(C|A)$ , that is we explicitly want to treat some pixels  $A_i$  as more relevant than the others.

However, the constraint on the meaningful information is now nonlinear in  $P(C|A)$ , so this is a much harder variational problem. Indeed, (71) can be written as follows:

$$\begin{aligned}
 I(C, A) - \beta I(C, B) &= \sum_{ij} P(C_j|A_i)P(A_i) \ln \frac{P(C_j|A_i)}{P(C_j)} \\
 &\quad - \beta \sum_{jk} P(B_k|C_j)P(C_j) \left\{ \ln \frac{P(B_k|C_j)}{P(B_k)} \right\} . \quad (72)
 \end{aligned}$$

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

$$P(B_k|C_j) = \frac{1}{P(C_j)} \sum_i P(A_i)P(B_k|A_i)P(C_j|A_i) , \quad (73)$$

where the conditional probability of the measurements,  $P(B_k|A_i)$ , is assumed to be known. The variation of (72) with respect to the encoding conditional probability,  $P(C_j|A_i)$ , now gives the equation (rather than an explicit expression):

$$P(C_j|A_i) = \frac{P(C_j)}{Z(A_i, \beta)} \exp \left[ -\beta \sum_k P(B_k|A_i) \log \frac{P(B_k|A_i)}{P(B_k|C_j)} \right] , \quad (74)$$

Technically, the system of equations (73,74) is usually solved by iterations, for instance, via deep learning (one of the paradigms for unsupervised learning). Doing compression procedure many times,  $A \rightarrow C_1 \rightarrow C_2 \dots$  is used in multi-layered Deep Learning Algorithms. Here knowledge of statistical physics helps in several ways, particularly in identifying phase transitions (with respect to  $\beta$ ) 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.

## 4.6 Information is money

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

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

$$\Lambda(f) = \ln 2 + \frac{n}{N} \ln f + \left(1 - \frac{n}{N}\right) \ln(1 - f) . \quad (75)$$

As  $N \rightarrow \infty$  we approach the mean geometric 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  $\lim_{N \rightarrow \infty} (n/N) = p$  (it is called self-averaging quantity because it is again a sum of random numbers). Differentiating  $\Lambda(f)$  with respect to  $f$  you find that the maximal growth rate corresponds to  $f = p$  (proportional gambling) and equals to

$$\lambda(p) = \ln 2 + p \ln p + (1 - p) \ln(1 - p) = S(u) - S(p) , \quad (76)$$

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. An

important lesson is that we maximize not the averaged return but its logarithm, which is a geometric mean. The geometric mean is less than the arithmetic mean. Therefore, we may have a situation when the arithmetic growth rate is larger than unity while the geometric mean is smaller than unity. That would be a disaster, since the probability to loose it all will tend to unity as  $N \rightarrow \infty$ , even though the mean returns grows unbounded.

It is straightforward to generalize (76) 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

$$\lambda(p, g) = \sum_i p_i \ln(p_i g_i) . \quad (77)$$

Here you have a formidable opponent - the track operator, who actually sets the payoffs. Knowing the probabilities, an ideal operator can set the payoffs,  $g_i = 1/p_i$ , to make the game fair and your rate zero. More likely is that the real operator has business sense to make the racecourse profitable by setting the payoffs a bit lower to make your  $\lambda$  negative. 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  $g_i = 1/Z q_i$  with  $Z > 1$ , then

$$\lambda(p, q) = -\ln Z + \sum_i p_i \ln(p_i/q_i) = -\ln Z + D(p|q) . \quad (78)$$

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, (78) 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 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). The same strategy is used by many plants, where the fraction of the seeds do not germinate in the same year they were dispersed; the fraction increases together with the environment variability.

Bacteria, plants and 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:

$$I(A, B) = \int dA P(A) \int dB P(B|A) \log_2 \frac{P(B|A)}{P(B)}. \quad (79)$$

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

$$\lambda = \int dA P(A) \int dB P(B|A) r(A, B). \quad (80)$$

Therefore, we look for the maximum of the functional  $F = \lambda - aI$ , which gives similarly to (69,70)

$$P(B|A) = \frac{P(B)}{Z(A, \beta)} e^{\beta r(A, B)}, \quad (81)$$

where  $\beta = a^{-1} \ln 2$  and the partition function  $Z(A, \beta) = \int dB P(B) e^{\beta r(A, B)}$  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  $\beta$  reflects relative costs of the information and the metabolism. If information is hard to get, one chooses small  $\beta$ , which makes  $P(B|A)$  weakly dependent of  $r(A, B)$  and close to unconditional probability. If information is cheaper, (81) 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 or 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. When you earn enough money, it may be a good time to start thinking about the nature of money itself. Money appeared first as a measure of value, it acquired probabilistic aspect with the development of credit. These days, when most of it is in bits, it is clear that this is not matter (coins, banknotes) but information. Moreover, the total amount of money grows on average, but could experience sudden drops when the crisis arrives. Yet in

payments money behaves as energy, satisfying the conservation law. I have a feeling that we need a new concept for describing money, which has properties of both entropy and energy. Free energy (which combines energy and entropy additively) probably cannot play this role, since the money as a universal medium of exchange is essentially a social construct. For example, cash payments are guaranteed by governments, but credit card payments are guaranteed usually by private banks, so these two kinds of money are not identical. Add to this non-bank money like cryptocurrencies and we start to understand that the value of money depends essentially on how many people agree to use it. It is a challenge to devise a conceptual framework able to handle both material and ephemeral sides of money, but it seems that the information theory is a right place to start.

## 5 Stochastic processes

In this Section we 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.

### 5.1 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,  $\tau$  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

$$P(\mathbf{x}, t + \tau) = \frac{1}{2d} \sum_{i=1}^d [P(\mathbf{x} + \mathbf{e}_i, t) + P(\mathbf{x} - \mathbf{e}_i, t)] . \quad (82)$$

We can write is as a finite difference approximation

$$P(\mathbf{x}, t + \tau) - P(\mathbf{x}, t) = \frac{1}{2d} \sum_{i=1}^d [P(\mathbf{x} + \mathbf{e}_i, t) + P(\mathbf{x} - \mathbf{e}_i, t) - 2P(\mathbf{x}, t)] . \quad (83)$$

The diffusion equation appears in the continuum limit taken while keeping constant the ratio  $\kappa = a^2/2d\tau$ :This

$$(\partial_t - \kappa \Delta)P(\mathbf{x}, t) = 0 . \quad (84)$$

Its solution is as follows:

$$\rho(\mathbf{x}, t) = P(\mathbf{x}, t)a^{-d} \approx (2\pi)^{-d} \int e^{i\mathbf{k}\mathbf{x} - t\kappa k^2} d^d k = (4\pi\kappa t)^{-d/2} \exp\left(-\frac{x^2}{4\kappa t}\right). \quad (85)$$

Note that (85,84) 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,  $\partial_t \rho(\mathbf{x}, t) = -\text{div } \mathbf{j}$  with the current  $\mathbf{j} = -\kappa \nabla \rho$ .

Consider now a Brownian particle under the action of a random force  $\mathbf{f}$  and in an external field  $V(\mathbf{q})$ . The momentum and coordinate satisfy the equations

$$\dot{\mathbf{p}} = -\lambda \mathbf{p} + \mathbf{f} - \partial_{\mathbf{q}} V, \quad \dot{\mathbf{q}} = \mathbf{p}/M. \quad (86)$$

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  $2\lambda T M = \int \langle f_i(0) f_i(t) \rangle dt$  means that the thermostat is in equilibrium.

Consider the over-damped limit  $\lambda^2 M \gg \partial_{qq}^2 V$ , where we can neglect the acceleration term on timescales exceeding the force correlation time  $\tau$ :  $\lambda \mathbf{p} \gg \dot{\mathbf{p}}$ . For example, if we apply to a charged particle an electric field  $\mathbf{E} = -\partial_{\mathbf{q}} V$  constant in space, then  $\partial_{qq}^2 V = 0$ ; averaging (coarse-graining) over times exceeding  $\tau$ , 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 (86) on  $\mathbf{q}$  is reduced to the first-order equation:

$$\lambda \mathbf{p} = \lambda M \dot{\mathbf{q}} = \mathbf{f} - \partial_{\mathbf{q}} V. \quad (87)$$

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  $V$ ,

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

and the density  $\rho(q, t)$  satisfies the diffusion equation. The dynamical equation without any randomness,  $\lambda M \dot{\mathbf{q}} = -\partial_{\mathbf{q}} V$ , 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  $\partial_t \rho = -\text{div } \rho \mathbf{w} = -\partial_{q_i} w_i \rho$ . Together, diffusion and advection give the so-called Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \kappa \nabla^2 \rho + \frac{1}{\lambda M} \frac{\partial}{\partial q_i} \rho \frac{\partial V}{\partial q_i} = -\text{div } \mathbf{J}. \quad (88)$$

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:

$$\rho(q) \propto \exp[-V(q)/\lambda M\kappa] = \exp[-V(q)/T] . \quad (89)$$

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.

## 5.2 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  $V(x, t)$ :

$$\dot{x} = -\partial_x V + \eta . \quad (90)$$

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 (88):

$$\partial_t\rho = T\partial_x^2\rho + \partial_x(\rho\partial_x V) = -\hat{H}_{FP}\rho . \quad (91)$$

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  $\psi$ -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 (91) 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:

$$\rho(x', t'; x, t) = \langle x' | \exp[(t - t')H_{FP}] | x \rangle . \quad (92)$$

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  $\partial_x^2$  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:

$$\begin{aligned} \rho(x', t; x, 0)e^{-V(x)/T} &= \rho(x, t; x', 0)e^{-V(x')/T} . \\ \langle x' | e^{-tH_{FP}-V/T} | x \rangle &= \langle x | e^{-tH_{FP}-V/T} | x' \rangle = \langle x' | e^{-V/T-tH_{FP}^\dagger} | x \rangle . \end{aligned} \quad (93)$$

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

$$H_{FP}^\dagger \equiv \left( \frac{\partial V}{\partial x} - T \frac{\partial}{\partial x} \right) \frac{\partial}{\partial x} = e^{V/T} H_{FP} e^{-V/T} , \quad (94)$$

i.e.  $e^{V/2T} H_{FP} e^{-V/2T}$  is hermitian, which can be checked directly. The quantum-mechanical notations thus allowed us to translate the detailed balance from the property of transition probabilities to that of the evolution operator.

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 (91) because of the extra term:  $\partial_t \rho = -\beta \rho \partial_t V$ . The distribution needs some time to adjust to the potential changes and is generally dependent on the history of these. For example, if we suddenly broaden the potential well, it will take diffusion (with diffusivity  $T$ ) to broaden the distribution. Still, can we find some use of the Gibbs factor and also have anything generalizing the detailed balance relation (93) 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 (91) 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  $t$ :

$$W_t = \int_0^t dt' \frac{\partial V(x(t'), t')}{\partial t'} . \quad (95)$$

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  $\rho_0$  and run (90) 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 (88), 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

$$\partial_t \rho = \beta^{-1} \partial_x^2 \rho + \partial_x (\rho \partial_x V) - \partial_W \rho \partial_t V , \quad (96)$$

Since  $W_0 = 0$  then the initial condition for (96) is

$$\rho(x, W, 0) = Z_0^{-1} \exp[-\beta V(x, 0)] \delta(W) . \quad (97)$$

While we cannot find  $\rho(x, W, t)$  for arbitrary  $V(t)$  we can multiply (96) by  $\exp(-\beta W)$  and integrate over  $dW$ . Since  $V(x, t)$  does not depend on  $W$ , we get the closed equation for  $f(x, t) = \int dW \rho(x, W, t) \exp(-\beta W)$ :

$$\partial_t f = \beta^{-1} \partial_x^2 f + \partial_x (f \partial_x V) - \beta f \partial_t V , \quad (98)$$

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 (97). 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 phase volume defines probability only in equilibrium, yet the work divided by temperature is an analog of the entropy change (production), and the exponent of it is an analog of the phase volume change. Let us stress that  $f(x, t)$  is not a probability distribution. In particular, its integral over  $x$  is not unity but the mean phase volume change, which remarkably is expressed via equilibrium partition functions at the ends (Jarzynski 1997):

$$\int f(x, t) dx = \int \rho(x, W, t) e^{-\beta W} dx dW = \langle e^{-\beta W} \rangle = \frac{Z_t}{Z_0} = \frac{\int e^{-\beta V(x, t)} dx}{\int e^{-\beta V(x, 0)} dx}. \quad (99)$$

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  $\langle x^n \exp(-\beta W_t) \rangle$ <sup>17</sup>. 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 at 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).

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

$$\langle e^{-\beta W_d} \rangle = \int dW_d \rho(W_d) \exp(-\beta W_d) = 1, \quad (100)$$

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:

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

---

<sup>17</sup>I thank R. Chetrite for this derivation.

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  $f \rightarrow 0$ , expand (99) 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:

$$\langle e^{-\beta W_d - I} \rangle = \langle e^{-\Delta S} \rangle = 1 . \quad (101)$$

(Sagawa and Uedo, 2012; Sagawa 2012). We have considered such a case in Section 4.2, where we denoted  $W_d = Q$  and used  $\langle W_d \rangle \geq -IT = -T\Delta S$ . The exponential equality (101) is a generalization of this inequality and (60).

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 (101) 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):

$$\rho^\dagger(-\Delta S) = \rho(\Delta S)e^{-\Delta S} . \quad (102)$$

Integrating (102) one obtains (99,100,101). That remarkable relation allows also one to express the mean entropy production via the relative entropy (53) between probabilities of the forward and backward evolution:

$$\langle \Delta S \rangle = \left\langle \ln[\rho(\Delta S)/\rho^\dagger(-\Delta S)] \right\rangle . \quad (103)$$

Entropy fluctuations were unobservable in classical macroscopic thermodynamics, but they are often very important in modern applications to nano and bio objects.

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.

### 5.3 Stochastic Web surfing and Google's PageRank

When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness... As was natural, this was followed by an excessive depression. The certitude that some ... precious books were inaccessible seemed almost intolerable.

Approaching the conclusion of the course, we cannot any more avoid the question: can we find an objective and quantitative measure not only of the amount of information, but also of its importance? We need to know which are the most precious books in the Library. By this time, it should come as no surprise for the reader that such measures can also be found using the statistical approach.

For an efficient information retrieval from the Web Library, webpages need to be ranked by their importance to order search results. A reasonable way to measure the importance of a page is to count the number of links that refer to it. Not all links are equal though — those from a more important page must bring more importance. On the other hand, a link from a page with many links must bring less importance (here probability starts creeping in). One then comes to the two rules: i) every page relays its importance score to the pages it links to, dividing it equally between them, ii) the importance score of a page is the sum of all scores obtained by links. For Internet with  $n$  pages, we organize all their scores into a vector  $\mathbf{p} = \{p_1, \dots, p_n\}$  which we normalize:  $\sum_{i=1}^n p_i = 1$ . According to the above rules,  $p_i = \sum_j p_j / n_j$  where  $n_j$  is the number of outgoing links on page  $j$ , which links to the page  $i$ . In other words, we are looking for the eigenvector of the hyperlink matrix,  $\mathbf{p}\hat{A} = \mathbf{p}$ , where the matrix elements  $a_{ij} = 1/n_j$  if  $j$  links to  $i$  and  $a_{ij} = 0$  otherwise. Does a unique eigenvector with all non-negative entries and a unit eigenvalue always exist? If yes, how to find it?

The iterative algorithm to find the score eigenvector is called PageRank<sup>18</sup> (Brin and Page 1998). The algorithm equates the score  $p_i$  of a page with the probability that a person randomly clicking on links will arrive at this page. It starts from ascribing equal probability to all pages,  $p_i(0) = 1/n$ , and generates the new probability distribution by applying the above rules of the score relay:

$$\mathbf{p}(t+1) = \mathbf{p}(t)\hat{A}. \quad (104)$$

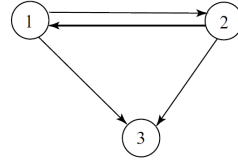
We recognize that this stochastic process is a Markov chain, which means that the future is determined by the present state and the transition probability, but not by the past. We thus interpret  $\hat{A}$  as the matrix of transition probabilities between pages for our stochastic surfer. In later modifications, one fills the elements of  $\hat{A}$  not uniformly as  $1/n_j$  but use information about actual frequencies of linking that can be obtained from access logs. Could our self-referential rules lead to a vicious circle or the iterations converge at  $t \rightarrow \infty$ ? It better be convergent fast to be of any use for the instant-gratification generation. It is clear that if the largest eigenvalue  $\lambda_1$  of  $\hat{A}$  was larger than unity, than the iterations would diverge; if  $\lambda_1 < 1$ , then

---

<sup>18</sup>"Page" relates both to webpage and to Larry Page, who with Sergei Brin invented the algorithm and created Google.

the iterations would converge to zero. We need the largest eigenvalue to be unity and correspond to a single eigenvector, so that the iterations converge. How fast it converges then will be determined by the second largest eigenvalue  $\lambda_2$  (which must be less than unity).

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



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

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

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

with the largest absolute value of a positive square matrix is positive, and belongs to a positive eigenvector, where all of the vector's elements are positive. All other eigenvectors are smaller in absolute value (Markov 1906, Perron 1907).

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

$$\mathbf{p}(t+1) = \hat{G}\mathbf{p}(t). \quad (105)$$

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

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

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

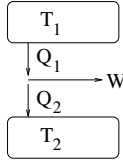
## 6 Conclusion

This Chapter attempts to compress the book to its most essential elements.

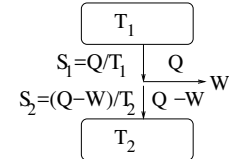
### 6.1 Take-home lessons

1. Thermodynamics studies restrictions imposed by hidden on observable. It deals with two fundamental extensive quantities. The first one (energy)  $E$  is conserved for a closed system, and its changes are divided into work (due to observable degrees of freedom) and heat (due to hidden ones). The second quantity (entropy)  $S$  can only increase for a closed system and reaches its maximum in thermal equilibrium, where the system entropy is a convex function of the energy. All available states lies below this convex curve in  $S - E$  plane.

2. Convexity of the dependence  $E(S)$  allows us to introduce temperature as the derivative of the energy with respect to the entropy. Extremum of the entropy means that the temperatures of the connected subsystems are equal in equilibrium. The same is true for the energy derivatives with respect to volume and other extensive variables. The entropy increase (called the second law of thermodynamics) imposes restrictions on thermal engine efficiency, that is the fraction of heat used for work:

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


Similarly, if information processing generates  $\Delta S$ , its energy price is as follows:

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


3. Need in statistics appear due to incomplete knowledge: We can measure only part of the degrees of freedom; even if we measure them all, we do it with a finite precision. Statistical physics defines the (Boltzmann) entropy of a closed system as the log of the phase volume,  $S = \log \Gamma$  and assumes (for the lack of any knowledge) the uniform distribution  $w = 1/\Gamma$  called microcanonical. For a subsystem, the (Gibbs) entropy is defined as the mean phase volume:  $S = -\sum_i w_i \log w_i$ ; the probability distribution is then obtained requiring maximal entropy for a given mean energy:  $\log w_i \propto -E_i$ . Information theory generalizes this approach, see 13 below.

4. 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. If we then assume that before every collision the 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 the mutual information. The lesson is that if we follow precisely all the degrees of freedom, the entropy is conserved and no information is lost. But if we follow only part of them, the entropy of that part will generally grow — whatever information we had is getting less relevant with time.

5. The total entropy growth can appear even if we follow all the degrees of freedom, but do it with a finite precision. In this case, we essentially consider evolution of finite phase-space regions. Instability leads to separation of trajectories, which spread over the whole phase space under a generic reversible Hamiltonian

dynamics, very much like flows of an incompressible liquid are mixing (we may say metaphorically, that for unstable systems, any extra digit in precision adds a new degree of freedom). Spreading and mixing in phase space correspond to the approach to equilibrium and entropy growth. On the contrary, to deviate a system from equilibrium, one adds external forcing and dissipation, which makes its phase flow compressible and distribution non-uniform.

6. Basic mathematical object we use in our discrete thinking is the sum of independent random numbers  $X = \sum_{i=1}^N y_i$ . Three concentric statements were made. The weakest one is that  $X$  approaches its mean value  $\bar{X} = N\langle y \rangle$  exponentially fast in  $N$ . The next statement is that the distribution  $\mathcal{P}(X)$  is Gaussian in the vicinity  $N^{-1/2}$  of the maximum. For even larger deviations, the distribution is very sharp:  $\mathcal{P}(X) \propto e^{-NH(X/N)}$  where  $H \geq 0$  and  $H(\langle y \rangle) = 0$ . Applying this to the log of the probability of a given sequence,  $\lim_{N \rightarrow \infty} p(y_1 \dots y_N) = -NS(Y)$ , we learn two lessons: i) the probability is independent of a sequence for most of them (almost all events are almost equally probable), ii) the number of typical sequences grows exponentially and **the entropy is the rate**.

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 to show that even when the mean of a random quantity is zero, its mean exponent is positive. That provides for an exponential separation of trajectories in an incompressible flow and exponential growth of the density of an element in a compressible flow.

8. Since uncertainty or the lack of knowledge plays such a prominent role, we wish to quantify it. The measure of uncertainty is the amount of information needed to remove it. 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  $\log_2$  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 6 above states that the number of typical sequences grows with the rate equal to the entropy  $S$ . The number of typical binary sequences of length  $N$  is then  $2^{NS}$ , which is smaller than  $2^N$ . The efficient encoding of the typical sequences thus involves words with lengths from unity to  $NS$ , which is less than  $N$  if the probabilities of 0 and 1 are not equal. That means that the entropy is both the mean and the fastest rate of the reception of information brought by long messages/measurements. To squeeze out all the unnecessary bits, encoding is used both in industry and in nature where sources often bring highly redundant



information, like in visual signals.

10. If the transmission channel  $B \rightarrow A$  makes errors, then the message does not completely eliminate uncertainty; what remains is the conditional entropy  $S(B|A) = S(A, B) - S(A)$ , which is the mean rate of growth of the number of possible errors. Sending extra bits to correct these errors lowers the transmission rate from  $S(B)$  to the mutual information  $I(A, B) = S(B) - S(B|A)$ , which is the mean difference of the uncertainties before and after the message. The great news is that one can still achieve an asymptotically error-free transmission if the transmission rate is lower than  $I$ . The maximum of  $I$  over all source statistics is the channel capacity, which is the maximal rate of asymptotically error-free transmission. In particular, to maximize the capacity of sensory processing, the response function of a living beings or a robot must be a cumulative probability of stimuli.

11. Very often our goal is not to transmit as much information as possible, but to compress it and process as little as possible, looking for an encoding with a minimum of the mutual information. For example, the rate distortion theory looks for the minimal rate  $I$  of information transfer under the restriction that the signal distortion does not exceed the threshold  $\mathcal{D}$ . This is done by minimizing the functional  $I + \beta\mathcal{D}$ . Another minimization task could be to separate the signal into independent components with as little as possible (ideally zero) mutual information between them.

12. The conditional probability allows for hypothesis testing by the Bayes' rule:  $P(h|e) = P(h)P(e|h)/P(e)$ . That is the probability  $P(h|e)$  that the hypothesis is correct after we receive the data  $e$  is the prior probability  $P(h)$  times the support  $P(e|h)/P(e)$  that  $e$  provide for  $h$ . Taking a log and averaging we obtain familiar  $S(h|e) = S(h) - I(e, h)$ . If our hypothesis concerns the probability distribution itself, then the difference between the true distribution  $p$  and the hypothetical distribution  $q$  is measured by the relative entropy  $D(p|q) = \langle \log_2(p/q) \rangle$ . This is yet another rate — with which the error probability grows with the number of trials.  $D$  also measures the decrease of the transmission rate due to non-optimal encoding: the mean length of the codeword is not  $S(p)$  but bounded by  $S(p) + D(p|q)$ . Mutual information is a particular case of relative entropy, they are both invariant with respect to arbitrary transformations of variables in a continuous case, which facilitates their ever-widening area of applications.

13. Since so much hangs on getting the right distribution, how best to guess it from the data? This is achieved by maximizing the entropy under the given data — "the truth and nothing but the truth". That explains and makes universal the 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:  $TI(A, M)$ . Another side of the physical nature of information is that there is the (Bekenstein) limit on how much entropy one can squeeze inside a given a radius; surprisingly, the limit is proportional to the area rather than the volume and is realized by black holes — our gates outside of this world.

15. The Renormalization Group is a best so far known way to forget information. Apart from forgetting in the first step, it involves the second step of renormalization. The focus is on the change of the probability distribution and the appearance of an asymptotic distribution after many steps. We find that the entropy of the partially averaged and renormalized distribution is the proper measure of forgetting in simple cases, like adding random numbers on the way to the central limit theorem. In physical systems with many degrees of freedom, it can be the mutual information defined in two ways: either between remaining and eliminated degrees of freedom or between different parts of the same system. In particular, it shows us examples of the area law, when  $I$  is sub-extensive.

16. The last lesson is two progressively more powerful forms of the second law of thermodynamics, which originally was  $\langle \Delta S \rangle \geq 0$ . The first new form,  $\langle e^{-\Delta S} \rangle = 1$ , is the analog of a Liouville theorem. The second form relates the probabilities of forward and backward process:  $\rho^\dagger(-\Delta S) = \rho(\Delta S)e^{-\Delta S}$ .

## 6.2 Epilogue

The central idea of this course is that learning about the world means building a model, which is essentially finding an efficient representation of the data. Optimizing information transmission or encoding may seem like a technical problem, but it is actually the most important task of science, engineering and survival. Science works on more and more compact encoding of the strings of data, which culminates in formulating a law of nature, potentially describing infinity of phenomena. The mathematical tool we learnt here is an ensemble equivalence in the thermodynamic limit, its analog is the use of typical sequences in communication theory. The result is two roles of entropy: it defines maximum transmission and minimum compression.

Another central idea is that entropy is not a property of the physical world, but is an information we lack about it. And yet the information is physical — it has an energetic value and a monetary price. Indeed, the difference between work and heat is that we have information about the former but not the later. That means that one can turn information into work and one needs to release heat to erase information. We also have learnt that one not only pays for information but

can turn information into money as well. The physical nature of information is manifested in the universal limit on how much of it we can squeeze into a space restricted by a given area.

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 2 can find them in Dorfman "An Introduction to Chaos in Nonequilibrium Statistical Mechanics". Detailed information theory with proofs can be found in Cowen & Thomas "Elements of Information Theory", whose Chapter 1 gives a concise overview. 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. On quantum information the comprehensive books are those by Preskill and Nielsen&Chuang. Numerous references scattered through the text, like (Zipf 1949), give you the most compact encoding of what is to google to find details.

Mention briefly one important subject left out of this course. Our focus was largely (though not entirely) on finding a data description that is good on average. Yet there exists a closely related approach that focuses on finding the shortest description and ultimate data compression for a given string of data. The Kolmogorov complexity is defined as the shortest binary computer program able to compute the string. It allows us to quantify how much order and randomness is in a given sequence — truly random sequence cannot be described by an algorithm shorter than itself, while any order allows for compression. Complexity is (approximately) equal to the entropy if the string is drawn from a random distribution, but is actually a more general concept, treated in courses on Computer Science. Another fundamental issue not treated here is the dramatic difference between the classical and quantum classifications of computational complexity.

Taking a wider view, I invite you to reflect on the history of our attempts to realize limits of possible, from heat engines through communication channels to computations. Will the next step be to study the natural limits of thinking and feeling?

Looking back one may wonder why accepting the natural language of information took so much time and was so difficult for physicists and engineers. Generations of students (myself including) were tortured by "paradoxes" in the statistical physics, which disappear when information language is used. I suspect that the resistance was to a large extent caused by the misplaced desire to keep scientist out of science. A dogma that science must be something "objective" and only related to the things independent of our interest in them obscures the simple fact that science is a form of human language. True, we expect it to be objectively independent of personality of this or that scientist as opposite, say, to literature, where we celebrate the difference between languages (and worlds) of Tolstoy and

Chekhov. However, science is the language designed by and for humans, so that it necessarily reflects both the way body and mind operate and the restrictions on our ability to obtain and process the data. Presumably, omnipresent and omniscient being would have no need in the statistical information approach described here. One may also wonder to what extent essential presence of scientist in science may help us to understand the special status of measurement in quantum mechanics.

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