

We explore an emergent interpretation of the action integral and the Lagrangian in physics, and discuss its connection with the concept of “amount of computation”. We give an abstract definition of action, and argue (1) that it provides a general, model-independent characterization of “amount of computation”, and that (2) the action of physics is a special case of this general action. Much as entropy quantifies the lack of information one has about the *state* of a system, action quantifies the lack of information about the system’s law—or, equivalently, its *behavior*. In this approach, action is to dynamics what entropy is to statics.

---

*From given causes, effects are generated by nature in the most efficient way. ... No natural action can be abbreviated.* [Leonardo]

*An institution will grow until it uses up all available resources.* [Parkinson’s law]

*Send a copy of this message to 16 friends, ... send a dollar to the first four addresses in the list, and without fail in two months you’ll be rich.* [Ponzi et al.]

*You want to get rich in one day? Mind that you don’t end up hanged in one year.* [Leonardo]

## 1 Introduction

What I would like to do here is to add a few strands to a line of thought which threads through Lucian,<sup>1</sup> Huygens, Laplace, Darwin, Boltzmann, Gibbs, Feynman, Chaitin,<sup>2</sup> Jaynes, and Landauer. The theme is how much one can get by conditioning, and the usual conclusion is that “there is no such thing as a free lunch.”

We are taught to regard with awe the variational principles of mechanics. There is something miraculous about them that appeals to our innate teleological longings. And something timeless too: the storms of relativity and quantum mechanics have come and gone, but Hamilton’s principle of least action still shines among our most precious jewels.

But perhaps the reason that these principles have survived such physical upheavals is that they are not physical principles after all! I will argue that they are the expression, in a physical context, of general principles about *computation*, much as entropy is the expression, in the same context, of general principles about *information* (cf. [6]). More specifically, just as entropy measures, on a log scale, the number of possible microscopic *states* consistent with

a given macroscopic description, so I maintain that action measures, again on a log scale, the number of possible microscopic *laws* consistent with a given macroscopic behavior. If entropy measures *in how many different ways you could be in detail* and still be substantially the same, then action measures *how many different recipes you could follow in detail* and still do substantially the same thing.

I also introduce an abstract, model-independent definition of **computational capacity**, which measures the power of a computer in terms of the number of different functions it can compute, and is the counterpart, in “information dynamics”, of the well-known *channel capacity* of information theory (the latter seen as “information statics”). (In spite of the occurrence of similar terms, my approach is quite unrelated to the model-dependent one of [11].)

In terms of computational capacity, the Lagrangian density  $\mathcal{L}$  measures how *degenerate* the underlying physics (seen as a fine-grained computing substrate) is; in other words, for a given certain macroscopic law and a proposed local macroscopic behavior, it counts how many different microscopic trajectory segments are compatible with that law, giving to each trajectory a weight inversely proportional to the the number of alternative microscopic laws that would have produced the same segment. As a consequence, the *action integral*  $S$  of a proposed trajectory measures how demanding this trajectory is with respect to the computational resources of the actual substrate; no wonder that the least demanding proposals are the ones that have the best chances of being accepted. In this light, the principle of least action is not one of parsimony, but of *prodigality*: a system’s natural trajectory is the one which will hog the most computational resources, leaving only a scrap to the potential “user” of the emergent dynamics.

If these arguments (which I give in only a sketchy way, with evidence from a few special cases) are correct and general, then the “principle of least action” is explained as a combinatorial tautology, much as the “survival of the fittest” and the “law of large numbers”.

Do not read me wrong. I am duly impressed by the enormous power of these laws; but when I use them I want to know where it is that they get their power from (*caveat emptor!*)—I just don’t believe in *magic*.

---

<sup>1</sup>The second-century sophist Lucian of Samosata used to poke fun at the Olympian gods. A “fundamentalist” of his time asked him, “How can you doubt the power of the gods? Go to the shrine of such-and-such goddess and see all the votive offering from sailors who, caught in a storm, prayed the goddess and were saved!” “Go to the bottom of the sea,” replied Lucian, “and see all the sailors who prayed the goddess and were *not* saved!”

<sup>2</sup>I am aware that some of Chaitin’s concepts on algorithmic information theory had earlier been proposed by Kolmogorov, for one. However, what I have in mind here is ideas like “You need a pound of axioms to get a pound of theorems,” which go well beyond Kolmogorov’s motif.

## 2 Connect the dots

Let a **state set**  $S$  be given once and for all; most generally, a **dynamics**  $f$  on  $S$  is a rule that, given the current state  $s \in S$  produces its next value  $s' \in S$ . A dynamics may be thought of as a *lookup table* whose format (nature of the entries, number of entries) depends on the nature of the state set  $S$  itself. Thus, if  $S$  is a finite collection of  $N$  symbols—which we may identify with the integers  $1, 2, \dots, N$ , a dynamics on  $S$  is simply a one-column table with  $N$  entries labeled  $1, 2, \dots, N$ , as illustrated in Fig. 1. Since we can independently choose one out of  $N$  possibilities for each of the  $N$  entries, the number of possible dynamics having this format is clearly  $N^N$ . Correspondingly, the size of the table is  $\log N^N = N \log N$ .<sup>3</sup>

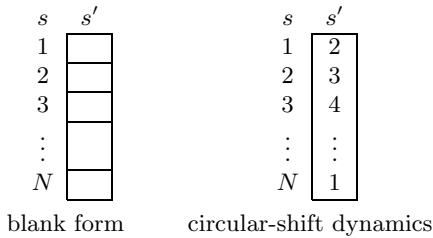


Figure 1: An arbitrary dynamics on the set  $\{1, 2, \dots, N\}$  is specified by filling out the blank form on the left, writing an arbitrary integer between 1 and  $N$  in each of the  $N$  blank entries. On the right, we have filled out this form with the *circular shift* dynamics, which maps  $s$  into  $s + 1 \pmod{N}$ .

Analytical mechanics begins when one recognizes that the dynamical systems of physics, in spite of their great variety, have much in common. Out of all *conceivable* dynamics, only those of certain classes actually occur. The task of analytical mechanics is to identify the constraints that characterize these classes and to learn to take advantage of them. Specifically, if the blank form of Fig. 1 can only be filled subject to certain constraints, then any resulting table is *redundant* and can be *compressed* by a factor that reflects the impact of the constraints themselves.

For instance, if we are only interested in *reversible* dynamics, then the number of possible tables goes down from  $N^N$  to  $N!$ , and, by suitable compression, one could use a smaller form, of size  $\log N! \approx N \log N - N$ .<sup>4</sup> If we are given a compressed table, in order to actually compute the system’s evolution from given initial conditions we have to uncompress it first. Thus, there is a tradeoff between economy of representation and amount of processing.

Consider now a system whose states are naturally labeled by two indices,  $i$  and  $j$ .<sup>5</sup> The table for such a system is more

<sup>3</sup>Intuitively, the  $N$  factor in  $N \log N$  represents the *number* of entries while the  $\log N$  factor represents the *size* of the entry “field”. In fact, this field must have enough room to specify any of  $N$  different choices; if we use binary coding for  $N$ , then the size of the field is  $\log_2 N$  bits.

<sup>4</sup>This is only slightly less than the size  $N \log N$  of the unconstrained case; thus, in terms of table size, the reversibility constraint is rather mild.

<sup>5</sup>For instance, small changes in the indices  $i$  and  $j$  may lead to small changes in behavior.

conveniently written in two-dimensional form,

$$\begin{array}{c} i \backslash j \\ \dots \\ \vdots \\ \dots \boxed{\quad} \dots \\ \vdots \\ \dots \end{array} \quad (1)$$

In ordinary mechanics, the state of a system with one degree of freedom is represented by a pair of real numbers—a position  $q$  and a momentum  $p$ . Moreover, time is continuous and  $q$  and  $p$  are continuous functions of time; thus, instead of specifying the *next value*  $q'$  of  $q$  and the next value  $p'$  of  $p$ , we specify their *rates of change*

$$\dot{q} = dq/dt, \quad \dot{p} = dp/dt. \quad (2)$$

This can be visualized as a table indexed by two variables,  $q$  and  $p$ , and whose entries consist of two fields,  $\dot{q}$  and  $\dot{p}$ ,

$$\begin{array}{c} q \backslash p \\ \dots \\ \vdots \\ \dots \boxed{\dot{q} \quad \dot{p}} \dots \\ \vdots \\ \dots \end{array} \quad (3)$$

Of course, as a mapping from  $\mathbb{R}^2$  to  $\mathbb{R}^2$  ( $\mathbb{R}$  denotes the real numbers), table (3) has an *uncountably infinite* number of entries, and similarly each entry has infinite size. For our purposes, though, we can “renormalize” and think of  $\mathbb{R}$  as a finite set of adequately large size  $N$  (e.g., a computer floating-point variable); then the table has finite size  $N^2 \cdot 2 \log N$ .

For a given dynamical system, to construct the orbit passing through point  $P = (q, p)$  we look up this point in the table, find the corresponding  $(\dot{q}, \dot{p})$ , and use these two numbers as the components of a displacement vector that leads us to a new point  $P' = (q', p') = P + dP$  by the recipe

$$q' = q + \dot{q} dt, \quad p' = p + \dot{p} dt.$$

Intuitively, each point carries with it *two numbers*, which are the incremental coordinates of the next point and so *directly* specify it. We locate and connect one point after the other in a blind-reckoning fashion, completely oblivious to the surrounding landscape. To find the successor to a given point, a single lookup operation is needed; this is called a **vectorial** specification of the dynamics.

Physical systems obey stronger constraints than just continuity. It turns out that for a large class of physical systems each point  $P$  only need carry with it a *single number*, called its **energy**, in order to specify the next point. The recipe for finding the next point  $P'$  is, admittedly, more involved. Given a point  $P$  of energy  $H$ , we must explore its vicinity looking for points having the same energy; these points will be found to lie on a line. Then we look for points having a slightly higher energy  $H + dH$ ; these points will also lie on a line, roughly parallel to the first. The recipe for the next orbit point  $P'$  is

1. From  $P$ , follow the line of constant  $H$ ,

2. in the sense that leaves the line  $H + dH$  on your left,
3. for a distance inversely proportional to the spacing between the two lines.

Thus, in order to determine  $P'$  from  $P$ , we have to look up the contents not only of the  $P$  entry itself but also of a number of its neighbors,<sup>6</sup> so that we can determine the gradient of  $H$  in this neighborhood; this is called a **variational** specification of the dynamics.

The relevant point here is that the table which gives  $H$  as a function of  $q$  and  $p$ , called the **Hamiltonian** of the system, has only *one* field in each entry rather than two, and thus is half the size of the vectorial table. For a system with  $n$  degrees of freedom, the vectorial table has  $2n$  fields per entry (a  $q$  and a  $p$  for each degree of freedom), but the corresponding Hamiltonian table still has only *one* entry!

In sum, the Hamiltonian formalism

- Recognizes that a certain class of dynamics is characterized by certain constraints; once we are told that a dynamics belongs to this class, then its vectorial lookup table has some redundancy in it and can be compressed by a factor  $2n$ , where  $n$  is the number of degrees of freedom.
- The compression procedure is nonlocal, since in general it entails integrating the vectorial equations of motion over the entire  $q$ - $p$  plane, and thus may be involved and costly.
- On the other hand, the decompression procedure is local, simple, and economical. To reconstruct the contents  $(\dot{q}, \dot{p})$  of the  $(q, p)$  entry of the vectorial table, do the order of  $n$  lookups in the vicinity of the  $(q, p)$  entry of the Hamiltonian table, and do a little linear arithmetic on the data you find there, as specified by Hamilton's equations<sup>7</sup>

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad (i = 1, 2, \dots, n) \quad (4)$$

where  $i$  runs over the degrees of freedom.

In this light, Hamiltonian mechanics is basically a *data compression scheme*. The details are specific to physics, but the general approach has nothing particularly physical to it, and in fact is useful in a lot of other situations in and outside of physics.

“All right!—you’ll say—Let’s grant that Hamiltonian mechanics is a data compression scheme. But the constraints that it takes advantage of and that make it possible for it to work at all are not generic constraints; they are specific of physics. Take that context away, and the compression scheme becomes useless. Thus, after all, the property that

<sup>6</sup>In the most economical numerical-computation scheme, at least two lookup operations (besides that for  $P$ ) are needed.

<sup>7</sup>The *canonical* equations, also called *Hamilton’s* equation, were written by Lagrange in 1908. Likewise, the principle of least action, formulated in its completeness for physics by Lagrange, is often called Hamilton’s principle. The Euler–Lagrange equations for the solution of a class of variational problems were written by Euler in 1744 (cf. [10]).

a system is Hamiltonian *does* express properties that reflect its being a *physical* system.”

How much so? Take a sheet of paper with coordinates  $q$ - $p$  and draw an arbitrary dynamics—a collection of continuous lines on the paper with tick-marks on them indicating the progress of time. I only ask that the trajectory parameters—position and direction as indicated by a line on the paper, and speed as indicated by the tick-marks on the line—change smoothly as one moves about the sheet of paper (“continuity”), and that the lines do not split or merge (“reversibility”<sup>8</sup>). When you are done, I will try to build in 3-D above the paper a surface  $H(q, p)$  that is a Hamiltonian for your dynamics. No matter what dynamics you have drawn, I will have no difficulty in constructing a surface whose contour lines coincide with the lines you have drawn. From this surface, anybody will be able to reconstruct your *lines* by following the Hamiltonian recipe given above; however, the *timing* along these lines, as given by the reconstructed tick-marks, will likely depart from that of your trajectories.

My construction is not unique: I still have room to adjust the surface so that, while the contour lines remain the same, the speeds along the trajectories are altered; for example, by squashing a mountain peak by a factor of two I can slow down by the same factor all the trajectories that are flowing around it. However, even after having done my best to take advantage of this slack, I may only get an approximation of the correct trajectories: if your flow is not Hamiltonian to begin with, the compression scheme will be lossy.

Even though *hardly any*<sup>9</sup> system is exactly Hamiltonian, is there some sense in which one can say that *almost all* of them are approximately Hamiltonian? In other words, does “Hamiltonianity” look like an emergent property? We shall briefly argue for this in the next section.

Note that the two things we took for granted before asking for Hamiltonianity, namely, continuity and reversibility, are very “cheap,” in the sense that the first naturally emerges in a wide range of settings if we just blur our vision[12], and the second also emerges, again in a wide range of settings, if we are willing to let a dynamical system age long enough before we try to “use” it (we will discuss this in a separate paper).

### 3 $T = dS/dE$ holds for almost any system

Consider a continuous system with one degree of freedom (cf. (3)). Let  $T$  be the period of a given orbit of energy  $E$ , and  $dS$  the volume of phase space swept when the energy of the orbit is varied by an infinitesimal amount  $dE$  (Fig. 2). As is well known[1], if the system is Hamiltonian these quantities obey the relation<sup>10</sup>

<sup>8</sup>This has nothing to do with ‘invariance under time reversal’, which is a much more specialized property

<sup>9</sup>I use “hardly any” and “almost all” in the measure-theoretical sense, meaning, respectively, a set of measure zero and a set of measure one.

<sup>10</sup>This is the very relation that makes the compression scheme work: when the trajectories spread, they invariably slow down, so that, given the lines, the timing can no longer be independently assigned.

$$T = dS/dE, \quad (5)$$

Just *how surprising* is this fact?

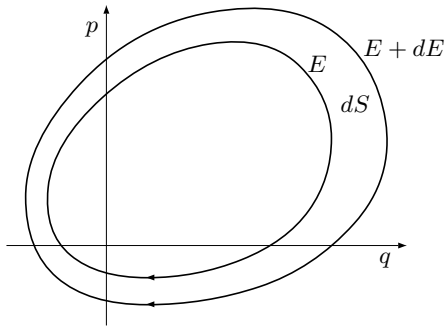


Figure 2: In moving from an orbit of energy  $E$  to an infinitesimally near orbit of energy  $E + dE$ , one sweeps an area  $dS$  of the phase space. If the dynamics is Hamiltonian, the orbit period  $T$  equals the ratio  $dS/dE$  of action-to-energy variation.

One way to answer this question is to go to a more accurate physical level (quantum mechanics) and see how energy and rate of evolution are related there (see [7]). Here instead our approach to the question is to consider systems of a more general nature, but for which quantities analogous to  $T$ ,  $dE$ , and  $dS$  are still meaningful. Under what conditions will the above relation hold for these systems?

Consider, for instance, the class  $\mathcal{X}_N$  of all discrete systems having a finite number  $N$  of states and an invertible but otherwise arbitrary dynamics (cf. Fig. 1 and attendant discussion). Though continuous quantities such as those appearing in (5) may arise from discrete ones in the limit  $N \rightarrow \infty$ , in general relation (5) will not hold (or even be meaningful) for every individual system; however, if one considers the entire class, one may ask whether this relation holds approximately for most systems of the class. Alternatively, one may ask whether this relation holds for a suitably-defined “average” system—treated as a representative of the whole class. This kind of approach is routinely used in statistical mechanics;<sup>11</sup> in our context, however, statistical methods are applied to “ensembles” in which the missing information that characterizes the ensemble concerns a system’s *law* rather than its initial *state*.

We shall show that relation (5) holds for the average element of the “ensemble”  $\mathcal{X}_N$ . Now, this is surprising, as we thought we hadn’t told  $\mathcal{X}_N$  anything about physics!

The systems of class  $\mathcal{X}_N$  have very little structure—basically, just invertibility. Nonetheless, one can still recognize within them the precursors of a few fundamental physical quantities. For instance, the period  $T$  of an orbit is naturally identified with the number of states that are strung along the orbit. Likewise, a volume  $S$  of state space, which in our case is an unstructured “bag” of states, will be measured in terms of how many states it contains. It

<sup>11</sup>For example, given a canonical ensemble for a system consisting of an assembly of many identical subsystems, almost all elements of the ensemble display a subsystem energy distribution that is very close to the Boltzmann distribution; the latter can thus be taken as the “representative” subsystem-energy distribution, even though hardly any element of the ensemble displays that distribution *exactly*.

is a little harder to identify a meaningful generalization of energy; the arguments presented in §3.1 suggest that in this case the correct identification is  $E = \log T$ , and this is the definition that we shall use below. Armed with the above “correspondence rules,” we shall investigate the validity of relation (5).

Each system of  $\mathcal{X}_N$  will display a certain distribution of orbit lengths; that is, one can draw a histogram showing, for  $T = 1, \dots, N$ , the number  $n(T)$  of orbits of length  $T$  (see Fig. 3).

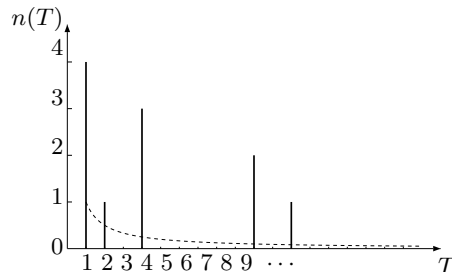


Figure 3: Orbit-length histogram of one particular system. The dashed curve gives the average histogram over the entire set of  $N!$  systems.

If in this histogram we move from abscissa  $T$  to  $T + dT$  we will accumulate a count of  $n(T) dT$  orbits. Since each orbit contains  $T$  points, we will sweep an amount of state space equal to  $dS = T n(T) dT$ ; thus

$$\frac{dS}{dT} = T n(T).$$

On the other hand, since  $E = \log T$ ,

$$\frac{dT}{dE} = T;$$

hence

$$\frac{dS}{dE} = \frac{dS}{dT} \frac{dT}{dE} = T^2 n(T).$$

Therefore, the original relation (5) will hold if and only if the orbit-length distribution is of the form

$$n(T) = 1/T.$$

Do the systems of  $\mathcal{X}_N$  display this distribution?

Observe that, as  $N$  grows, the number of systems in  $\mathcal{X}_N$  grows much faster than the number of possible orbit-length distributions: most distributions will occur many times, and certain distributions may appear with a much greater frequency than others. Indeed, as  $N \rightarrow \infty$ , almost all of the ensemble’s elements will display a similar distribution. In such circumstances, well-known theoretical and practical considerations recommend defining the “typical” distribution as the *mean distribution* over the ensemble, denoted by  $\overline{n(T)}$ .

It turns out that for  $\mathcal{X}_N$  the mean distribution is *exactly*

$$\overline{n_N(T)} = 1/T \quad (6)$$

for any  $N$ , as indicated in Fig. 3. In fact, we construct a specific orbit of length  $T$  by choosing  $T$  states out of  $N$  and

arranging them in a definite circular sequence. This can be done in  $\binom{N}{T} \frac{T!}{T}$  different ways. To know in how many elements of the ensemble the orbit thus constructed occurs, we observe that the remaining  $N - T$  elements can be connected in  $(N - T)!$  ways. Thus, the total number of orbits of length  $T$  found anywhere in the ensemble is

$$\binom{N}{T} \frac{T!}{T} (N - T)! = N! \frac{1}{T}.$$

Divide by the size  $N!$  of the ensemble to obtain  $1/T$ .

Thus, the typical discrete system obeys relation (5). Intuitively, when  $N$  is large enough to make a continuous treatment meaningful, the odds that a system picked at random will closely obey (5) are overwhelming.

### 3.1 Why $E = \log T$

Here we motivate the choice  $E = \log T$  made above.

Finite systems lack the rich topological structure of the state space found in analytical mechanics. Beside invertibility, in general the only *intrinsic*<sup>12</sup> structure that they are left with is the following: *Given two points  $a$  and  $b$ , one can tell whether  $b$  can be reached from  $a$  in  $t$  steps; in particular (for  $t = 0$ ), one can tell whether or not  $a = b$ .* Thus, for instance, one can tell how many orbits of period  $T$  are present, but of these one cannot single out an individual one without actually pointing at it, because they all “look the same”.

To see whether there is a quantity that can be meaningfully called “energy” in this context, let us observe that physical energy is a function  $E$ , defined on the state space, having the following fundamental properties:

1. *Conservation:*  $E$  is constant on each orbit (though it may have the same value on different orbits).
2. *Additivity:* The energy of a collection of weakly-coupled system components equals the sum of the energies of the individual components.
3. *Generator of the dynamics:* Given the constraints that characterize a particular class of dynamical systems, knowledge of the function  $E$  allows one to uniquely reconstruct the dynamics of an individual system of that class.

The proposed identification  $E = \log T$  obviously satisfies property 1.

As for property 2, consider a finite system consisting of two independent components, and let  $a_0$  and  $a_1$  be the respective states of these two components. Suppose, for definiteness, that  $a_0$  is on an orbit of period 3, and  $a_1$  on one of period 7; then the combined system state  $(a_0, a_1)$  is on an orbit of length 21, i.e.,  $\log T = \log T_0 + \log T_1$ . This argument would fail if  $T_0$  and  $T_1$  were not coprime. However, for two randomly chosen integers the expected number of common factors grows extremely slowly with the size of the

<sup>12</sup>That is, independent of the labeling of the points, and thus preserved by any isomorphism.

integers themselves[8] (and, of course, the most likely common factors will be small integers); thus the departure from additivity vanishes in the limit  $T \rightarrow \infty$ .

As for property 3, an individual system of  $\mathcal{X}_N$  is completely identified—up to an isomorphism—by its orbit distribution  $n(T)$ , and thus any “into” function of  $T$  (in particular,  $E = \log T$ ) satisfies this property.

## 4 The power of conditioning

Much as in thermodynamics a “state”—i.e., a *macroscopic* state—is in fact a bag of *microscopic* states, we shall entertain the notion that, in analytical mechanics, a “trajectory” of our dynamics is really a bundle of microscopic trajectories of an underlying, fine-grained dynamics. How does the size of this bundle depend of what we know about the (macroscopic) trajectory?

Let us consider a particle that at every tick of the clock can move one notch right or left on a one-dimensional track. The particle is driven by a computer program. To be specific, there is a given computer with a large amount of memory in it; somebody initializes the memory and then lets the computer run. They fix their attention on a specific data bit. Every billionth cycle of the computer they look at the state of this bit:<sup>13</sup> if it is a 1 they will move the particle to the right; if a 0, to the left. At time 0 we put the the particle at notch 0 and leave the room. At time  $T$  (say, some 10,000 ticks) we are asked, “Where is the particle now?”

How could we possibly know? We know the computer but we do not know what program it is running or what the initial data were. Actually, we know one hard fact: at time  $T$  the particle must be within the interval  $[-T, +T]$ , since it cannot move more than one notch at a time. Beyond that, what can we say? If hard pressed, we may hazard a guess that the particle is probably closer to the middle of the interval than to the very ends.

They let us go. At time  $2T$  they are at it again. This time they tell us where the particle is now: at a certain position  $2X$  (with, say,  $X = 8000$ ); but they ask for the same information as before, namely, “Where was the particle at time  $T$ ?” And, at gunpoint, “A bad answer, and you are dead!”

Let us think rationally. We could simulate the given computer on our laptop PC, trying out all the programs one by one. Perhaps, by the way the computer is designed, there are some positions where the particle can never be at time  $T$  no matter what program is running; that might help us narrow down the choice by a tiny bit. But the computer may have a billion bits, and so may be running any of  $2^{1,000,000,000}$  programs; this is not a promising approach, since we’ll certainly be dead before we’ve tried out even a small fraction.

Then we realize that, no matter how many *programs* there might be, the number of possible *paths* the particle may have followed from the start to time  $T$  is “only”

<sup>13</sup>For boolean values we use the “old style” digits 0 and 1 from the series 0123456789 to stress we are not necessarily thinking of the numerical values 0 and 1).

$2^T$ . Forget the red herring of the  $2^{1,000,000,000}$  programs: our chances of survival are no worse than one in  $2^T$ , and  $T$  is in the thousands rather than in the billions. Even better, the number of *places* where the particle can be at  $T$  (rather than the *paths* from 0 to  $T$ ) can only grow linearly with  $T$ —in fact, it is no more than  $T + 1$ . That is, our chances cannot possibly be worse than one in  $T$  ( $\approx 2^{13}$ )—even if the particle were equally likely to be in any of the possible slots. With a not too large  $T$  in the denominator, perhaps we can put some large constant in the numerator and manage to get an appreciable chance of survival. Let us do some figuring, but without assuming anything more than we know.

The worst scenario—and, by Murphy’s (or Jaynes’s) law—the one that we owe it to ourselves to entertain unless we want to deceive ourselves, is that all particle paths compatible with our information are equally probable. Forget about the computer program, about which we do not know anything besides that it is large, and concentrate on the kinematic constraints. For a specific position  $x$ , how many paths go from event  $P_0$  (particle at the origin at time  $t = 0$ ) through event  $P_1$  (particle at position  $x$  at time  $T$ ) and event  $P_2$  (particle at position  $2X$  at time  $2T$ )?

In general, if the net progress during a time interval  $t$  is  $x$ , and this was done by  $x_+$  steps to the right and  $x_-$  to the left, we must have

$$\begin{aligned} x_+ + x_- &= t, & \text{or} & & x_+ &= (t + x)/2 = \frac{1+\beta}{2}t, \\ x_+ - x_- &= x, & & & x_- &= (t - x)/2 = \frac{1-\beta}{2}t, \end{aligned}$$

where we have called  $\beta$  the average velocity  $x/t$  in that interval. Then, the number of paths is<sup>14</sup>

$$N(\beta, t) = \binom{t}{\frac{1+\beta}{2}t, \frac{1-\beta}{2}t}. \quad (7)$$

Denote by  $H(p)$  the **binary entropy** function

$$H(p) = -p \log p - q \log q,$$

that is, the entropy of the binary distribution  $\{p, q\}$ . This entropy can be rewritten more symmetrically in terms of the the *mean*  $\mu = p - q$  of the distribution, as

$$K(\mu) = H(p(\mu)), \quad \text{where} \quad p = (1 + \mu)/2.$$

Using Stirling’s approximation, we get from (7)

$$\frac{1}{n} \log N(\beta, n) = K(\beta) + O\left(\frac{\log n}{n}\right); \quad (8)$$

the “big-oh” term in (8) vanishes as  $n \rightarrow \infty$ . Some salient features of  $K(\beta)$  are shown in Fig. 4.

If we call  $\beta$  the average velocity of the overall trip (from 0 to  $2T$ ),  $\beta_1$  that of first lap (from 0 to  $T$ ),  $\beta_2$  that the second lap (from  $T$  to  $2T$ ), and  $\epsilon$  the excess of  $\beta_1$  over  $\beta$ , we have

$$\begin{aligned} \beta &= X/T, & \text{and} & & \beta_1 &= \beta + \epsilon, \\ \epsilon &= (x - X)/T, & & & \beta_2 &= \beta - \epsilon. \end{aligned}$$

<sup>14</sup>If  $n = h + k$ , we may write  $\binom{n}{h, k}$  for  $\binom{n}{k}$  to stress the symmetry between  $h$  and  $k$ .

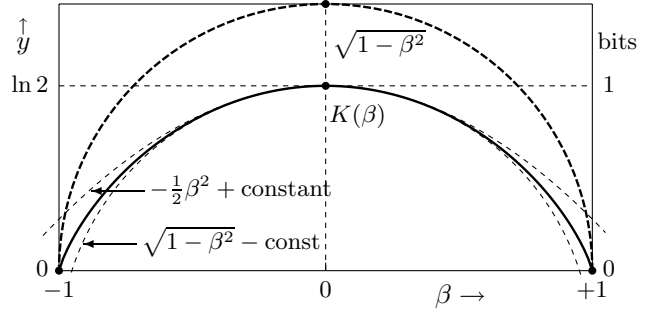


Figure 4: The binary entropy  $K(\beta)$  compared with the parabola osculating it at the apex, and with the relativistic factor  $\sqrt{1 - \beta^2}$  (a half-circle).

Coming back to our question, the number of paths from  $P_0$  to  $P_2$  through  $P_1$  is

$$\begin{aligned} N_{012} &= N_{01}N_{12} = N(\beta_1, T)N(\beta_2, T) \\ &= N(\beta + \epsilon, T)N(\beta - \epsilon, T), \end{aligned}$$

while that from  $P_0$  to  $P_2$  (with no restrictions on intermediate points) is

$$N_{02} = N(\beta, 2T).$$

Thus, the probability  $\mathcal{P}$  that the particle went through  $P_1$ , conditioned by the knowledge that it started at  $P_0$  and ended up at  $P_2$ , is

$$\mathcal{P} = \frac{N_{01}N_{12}}{N_{02}} = \frac{N(\beta + \epsilon, T)N(\beta - \epsilon, T)}{N(\beta, 2T)}.$$

Taking logs and dividing by the number of steps as in (8) we obtain a quantity  $R(\epsilon)$  proportional to the log of the relative frequency of the paths going through  $x$  at  $T$ , namely,

$$R(\epsilon) = K(\beta + \epsilon) + K(\beta - \epsilon) - 2K(\beta).$$

Now, it turns out that, for small (and even not so small) values of  $\beta$ ,

$$K(\beta) \approx -\frac{1}{2}\beta^2 + \text{const} \quad (9)$$

(cf. Fig. 4). Thus, up to an additive constant,

$$R(\epsilon) \approx -[(\beta + \epsilon)^2 + (\beta - \epsilon)^2 - 2\beta^2]/2 = -\epsilon^2/2.$$

The maximum for  $R$  is clearly at  $dR/d\epsilon = 0$ , which occurs for  $\epsilon = 0$ , that is, for  $\beta_1 = \beta_2 \equiv \beta$ . Thus we conclude that the most likely place for the particle to be at time  $T$  is  $X$ , i.e., halfway between its initial position 0 and its final position  $2X$ , as if it had traveled at a uniform velocity  $\beta = X/T$ . In other words, the most likely position for event  $P_2$  to be is on the *straight spacetime line* between  $P_0$  and  $P_1$ , as shown in Fig. 5. This generalizes to any choice of initial and final positions, and of any instant of time for which the position of the particle is guessed (Fig. 6). Incidentally, with the given data, the standard deviation from this straight line at midtime is about 40 notches; with our guessing system, the probability to come out alive is about 0.014, or better than one in a hundred.

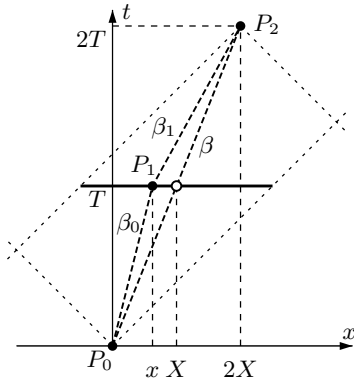


Figure 5: For a particle that performs a symmetric random walk, starts at  $P_0 = (0, 0)$ , and ends up at  $P_2 = (2T, 2X)$ , what is the spatial probability distribution at  $t = T$ ? All paths are comprised within the forward light-cone from  $P_0$  and the backward light-cone from  $P_2$ ; thus, the range of positions at  $T$  is restricted to the solid line in the middle. The particle’s mean velocity from  $P_0$  to  $P_2$  is  $\beta = X/T$ ; the mean for the subpath  $P_0P_1$  is  $\beta_1$ ; that for  $P_1P_2$ ,  $\beta_2$ . The requested distribution has both mean and maximum at  $x = X$  (hollow dot).

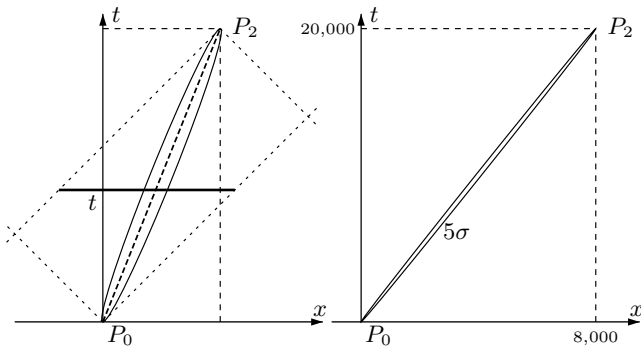


Figure 6: For any given time  $t$ , the probability distribution for the particle to be at  $x$  has a mean indicated by the line of slope  $\beta$  and a standard deviation  $\sigma$  indicated by the width of the surrounding ellipse. The figure on the right uses the data from the example ( $T = 10,000$ ,  $X = 8,000$ ); even when showing a  $5\sigma$  deviation, the width of the ellipse is barely noticeable.

Note that we have never suggested that the particle must have had some “momentum” that made it try to preserve its “speed”, or that, since the particle moved to the right 9 times out of 10, the computer program, seen as a random-number generator, must have been one characterized by a certain biased “statistics”. We did not make hypotheses about where the particle was when we were not seeing it; all we used was the little we were told, namely, where the particle was at  $t = 0$  and  $t = 2T$ , and that it was hopping right or left one notch at the time (the latter is a crucial piece of information). Our prediction is the “flattest” that one could make out of the available data, and thus is just a *tautological rewording*—though an illuminating one—of these data.

## 5 Enter the Lagrangian

It will not have escaped the reader’s notice that the quantity

$$L_{\text{comb}} = -K(\beta)$$

(‘comb’ for “combinatorial”) coincides, for small  $\beta$  (as in (9)), with the *Lagrangian of the free particle*, which in Newtonian mechanics is given by

$$L = \frac{1}{2}mv^2$$

(an additive constant is irrelevant to the Lagrangian, and, for a single particle, the size of the  $m$  factor is also irrelevant). Over the entire range of  $\beta$ ,  $L_{\text{comb}}$  is amazingly close to the *relativistic Lagrangian*,

$$L = -mc^2\sqrt{1 - (v/c)^2},$$

if we set  $\beta = v/c$ , i.e., equate the speed of light  $c$  with the maximum speed of propagation of information in the random walk—one notch per tick of the clock. Does this coincidence have any significance?

To answer this question we shall proceed on three fronts, namely

- Examine the role of the Lagrangian in physics.
- Study the form of  $L_{\text{comb}}$  in less trivial combinatorial models: systems of interacting particles, and programs whose parameters depend on space and time.
- Explore the connections between the ordinary physical Lagrangian, the combinatorial Lagrangian introduced here, and the so far poorly quantified concepts of “amount of computation” and “computation capacity”.

To give an idea of what we are after, let us observe that in Fig. 5, the number of microscopic paths from  $P_0$  to  $P_2$  is given by the product of two binomial coefficients (cf. (7)), one for  $P_0P_1$  and one for  $P_1P_2$ . The first factor grows very rapidly as  $P_1$  gets closer to the  $t$ -axis; a greedy algorithm would try to position  $P_1$  right on the  $t$ -axis to get the greatest benefit from the first stretch. However, as we move  $P_1$  leftwards, the second factor shrinks *even more rapidly*. When the day of reckoning comes, this greedy policy will be punished; the best reward will go to the policy that positions  $P_1$  so as to maximize the number of paths through it. Minimizing  $-K(\beta)$  (or  $\frac{1}{2}\beta^2$ , for small  $\beta$ ) is a means to this goal.

I argue that, deep down, this must be the reason why the Lagrangian approach works. It is only a matter of convention that, by taking minus the log of path count, the issue becomes one of *minimizing a sum* rather than maximizing a product; what is important is that, in this interpretation, we do not really need any intermediaries to give us prize or punishment; *the act is its own reward*. (In Feynman’s approach[2], for comparison, there still is one level of mediation, which happens to work but still remains mysterious, between action and probability.) Minimizing the action integral  $I$  leads to the path with the best chances because the Lagrangian of a path segment is *nothing but* (up to

a cosmetic “minus log” dress-up) the very chances of that segment. As in Darwinian evolution, we may think of a lot of intermediate fitness indicators, but the ultimate indicator of the chances of survival is—*by definition*—how many individuals will survive. A tautology, but a useful one. As in evolution, so in Lagrangian mechanics local optimization via local indicators may get us trapped into dead ends: local optimization through local indicators has its risks. On the other hand, global optimization through local indicators has (in the worst scenario) an exponential cost. Pick one!

As introductions to the subject I recommend (a) *The Parsimonious Universe* by Hildebrandt and Tromba[3]—a wonderful nontechnical book rich in historical motivation and examples from the natural sciences—(b) the classic textbook by Lanczos[5], and, of course, (c) Feynman’s monograph on path integrals in quantum mechanics[2].

The first examples of variational problems were connected with *statics*; for example, the equilibrium position of a chain, first solved by Johann Bernoulli in 1690. There the approach was to assign a “penalty” to each link of the chain equal to its distance from ground (potential energy); the goal was to minimize the overall penalty subject to the geometric constraint that the links remain—by definition of ‘chain’—connected to one another. And, indeed, the minimum-penalty configuration—the *catenary*—happens to be the equilibrium configuration. An important property of this configuration is that small changes from it leave the penalty *unchanged in first order*; conversely, configurations that are “penalty-indifferent” in this sense—or **stationary**—are equilibrium solutions (though some may be unstable or may correspond to a relative *maximum* rather than minimum of the penalty). Once we look at the statistical-mechanical explanation of this solution, we realize that the catenary is the configuration of least energy not because it was trying to be parsimonious, but because it spent *all it could* in the thermal market.

The novelty of the Lagrangian approach is the search for a minimum in *dynamical* problems, where a penalty based only on position is not sufficient to single out the actual trajectories—the Earth does not fall on the Sun even though the potential energy is less there. What other observable properties of a system should be used in assessing the “tax” in order to obtain agreement with observation? Potential energy is assessed on the basis of position; do we also have to take into account functions of velocity, acceleration, etc., and where do we stop? It turns out that in all clean-cut cases<sup>15</sup> velocity is enough; indeed, a **Lagrangian** tax  $L$  defined as the difference between a kinetic energy  $T$  and a potential energy  $U$ ,

$$L = T(\text{velocity}) - U(\text{position})$$

works well for the dynamics of many-particle systems when the interactions between particles are only pairwise; moreover, here kinetic energy is not an arbitrary function of velocity, to be assigned *ad hoc* for each system as in the

<sup>15</sup>Exceptions mostly arise where the “real” dynamics is hidden under layers of macroscopic phenomenology: arcade video-games can come up with quite arbitrary trajectories.

case of potential energy, but is invariably (at low speeds) a *quadratic* function of the velocities.<sup>16</sup>

Thus, velocity appears to be a *bona fide* part of the state of a system. Specifying only the initial position of a cannon ball does not allow us to predict its trajectory; but specifying also the initial velocity allows us to do so *uniquely*. Information about acceleration, on the other hand, is like knowledge and the Koran: if it is right, it agrees with the Lagrangian and we can throw it away; if it does not agree with the Lagrangian, it is wrong and we must throw it away.

How do we use the Lagrangian? How do we compute a trajectory from it? Let us use an analogy first. If we disregard the one-time pick-up charge, the taxicab fare  $I_{\text{taxi}}$  for a trip  $s$  from a given origin  $x_0$  to a given destination  $x_1$  is determined by a taximeter which computes the following “taxicab action”

$$\int_s L_{\text{taxi}} dt, \quad (10)$$

where, up to a currency-conversion factor, the “taxicab Lagrangian” is

$$L_{\text{taxi}} = |v| + 1. \quad (11)$$

Integrated over time, the first term of (11) gives the geometric length of the trip (odometer reading) while the second term gives the elapsed time (clock reading). If we fix for a moment the departure and arrival times  $t_0$  and  $t_1$ , we may ask, Among all possible trips (following the same route but going through the same place at different speeds counts as different trips) which are the cheapest? Since  $t_1 - t_0$  is fixed and thus the elapsed time term gives a fixed contribution, the cheapest trip is that which follows the shortest route—speed doesn’t count.<sup>17</sup> If we now ask the same question for different arrival times  $t_1$ , the term 1 in (11) comes into play, and we get that the cheapest trip is that which follows the shortest route *and* proceeds at the highest possible speed.<sup>18</sup>

While the rider’s interest is to minimize the cost of the trip, the driver’s interest is to *maximize* the take-home pay

$$\int_{\text{workday}} L_{\text{taxi}} dt. \quad (12)$$

Here the second term in  $L_{\text{taxi}}$  gives again a constant contribution, while the first term gives a contribution that is proportional to the speed and *independent* of the route. In a first approximation, the driver’s optimal policy is thus, “Always go at infinite speed, never mind the route!”<sup>19</sup>

If we make  $L_{\text{taxi}}$  also time- and space-dependent ( $v$  is constrained by bottleneck in tunnel at rush hour, night rate applies after 10:00 pm, etc.), deriving an explicit prescription

<sup>16</sup>Since we are dealing with continuous and differentiable variables, all that this “first-order quadraticness” means is that kinetic energy bottoms out when the speed is zero.

<sup>17</sup>Speed *does* count in the taximeter Lagrangian (11), but the dependency is such that in this fixed-time problem it disappears from the solution.

<sup>18</sup>Naively, we would strive for infinite speed. In practice the  $|v|$  term should be complemented by a highly nonlinear term in  $v$  (probability of accident or breakdown, the car just can’t make it—or even legal speed limits!).

<sup>19</sup>Since the driver is in charge, it is not surprising that this is close to reality. Here, the annoying one-time charge comes to our help, as it gives the driver an incentive to *also* follow the shortest route to our destination, so as to get a fresh customer as soon as possible.

from the taximeter Lagrangian becomes in practice very complicated, as we have to evaluate a lot of different discrete possibilities before knowing which is the “best”. However, if variables and parameters change in a continuous way, as they do in the stylization of physics called analytical mechanics, in principle it is possible to *recognize* a locally optimal solution by the fact that small changes to it do not in first order affect the action integral. As in the case of the Hamiltonian, an explicit vectorial law (in which direction and at what rate do I move from the present state?) can be obtained by judicious sampling of the Lagrangian lookup table in the vicinity of the given state (cf. §2). Though the Lagrangian approach is more general and versatile than the Hamiltonian one, we pay for that by a more elaborate decompression scheme, which works in two steps. First, given your current *state*, you sample the Lagrangian in its neighborhood and construct the quantities that go into the *Euler–Lagrange* equation—which, for the case of one degree of freedom, is

$$\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} = 0.$$

This equation yields a lookup table that relates the quantity we want, namely  $\ddot{q}$  (the acceleration), to those we know,  $q$  and  $\dot{q}$ . However, the resulting table may not be indexed by  $(q, \dot{q})$ , and, consequently, we may have to sample this table in turn in order to reconstruct an explicit relationship  $\ddot{q} = f(q, \dot{q})$ .

## 6 On two levels<sup>20</sup>

You may think it impertinent that I took the random walk—no inertia, total dependence on external whims—as a model of the motion of a free particle governed purely by inertia. We certainly need to test these ideas with more realistic models of dynamics. To this purpose, and in order to establish a link between action and amount of computation, we shall introduce a computer-like system to which the Lagrangian and Hamiltonian formalisms can be applied verbatim. By moving back and forth between the computational and the physical interpretation of this system we will be able to establish correspondence rules between physical and computational constructs.

### 6.1 Chains and strings

chains

Consider a linear **chain** of **dots** running along the  $x$ -axis with two dots for every unit of length (Fig. 7). The displacement of a dot from a horizontal reference line will be recorded along the  $q$ -axis. A dot is connected to its two neighbors by **links** of slope  $\pm 1$ . Thus, on a macroscopic scale the chain will appear as a continuous string with slope in the  $x$ - $q$  plane never exceeding  $\pm 1$ .

For brevity, coordinates having integer values (1,2,3,...) will be called **even**; those having half-integer values ( $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots$ ), **odd**. We are interested in the class of chain dynamics obeying the following constraints (Fig. 8):

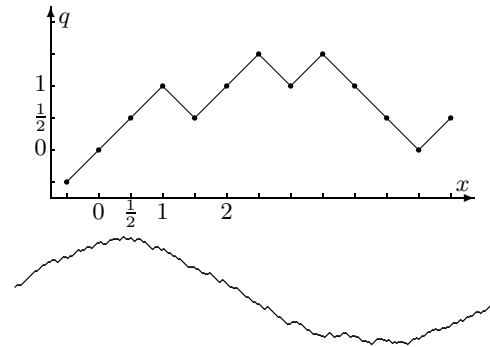


Figure 7: A chain with links of slope  $\pm 1$ , stretching along the  $x$ -axis and with displacements recorded along the  $q$ -axis (top). Macroscopically, the chain will look like a continuous string with slope in the  $x$ - $q$  plane never exceeding  $\pm 1$  (bottom).

- At even times, the candidates for a move are the dots at even places (solid dots); at odd times, those at odd places (hollow dots).
- If a candidate can hop up or down one unit in the  $q$  direction while retaining links of slope  $\pm 1$  with its neighbors, then it is up to the specific dynamics to decide whether it will do so (**flip**). Otherwise, the candidate will remain in place (**rest**).

Thus a dynamics is just a table that, for any  $q, x, t$  that are all three odd or all three even, specifies whether the candidate dot is *permitted* to flip. Whether a dot will actually flip depends not only on whether the adjacent links allow it to flip, but also on whether the permission bit allows it to take advantage of this possibility.

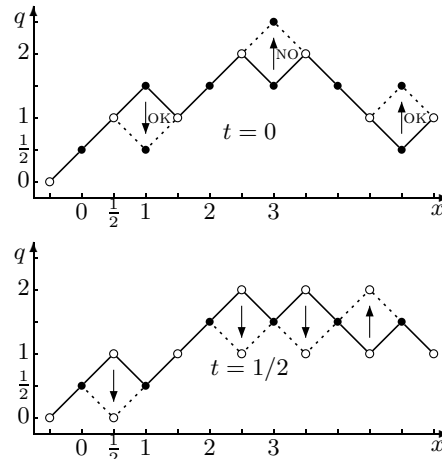


Figure 8: Two successive configurations of a discrete chain with links of slope  $\pm 1$ . At even steps, the even dots (solid) try to move; at odd steps, the odd dots (hollow). The arrows indicate the moves that are possible without breaking the links. A possible move will actually be made (the dot will **flip**) only if the dynamics gives permission; otherwise the dot will **rest**. In this figure, permission was not granted at  $t = 0$  for the dot at  $x = 3$ , so that at time  $t = \frac{1}{2}$  we find it at the same position.

One can think of the present setting as a special-purpose *computer*; the *data* are the states ( $\pm 1$ ) of the chain’s links; the *program* is the permission table itself. That is, we view

<sup>20</sup>This phrase is borrowed from Kadanoff[4].

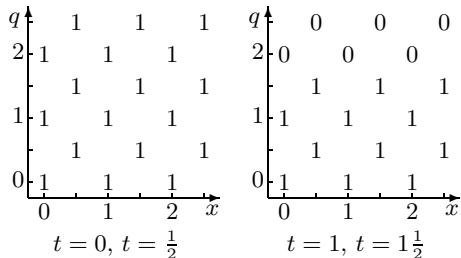


Figure 9: At its finest level, the program of the chain computer is just a binary table containing a permission bit for each point  $(q, x, t)$  such that all three coordinates are even or all odd (“body-centered cubic lattice”). Each frame display, in an interleaved fashion, data for two consecutive values of  $t$  (even  $t$  at even  $q, x$ ; odd  $t$ , at odd). For  $q < 2$ , this dynamics always grants permission; for  $q \geq 2$ , it grants permission on two half-steps and then withholds it for two more half-steps.

spacetime (here including for convenience  $q$  as another spatial coordinate) as a read-only memory containing a bit at each point. The *processor* is just the mechanism that interprets the data as a linked chain and the program as a dynamics, flipping a dot if and only if its links (the general kinematic constraints) and the permission bit (the specific dynamics) permit it. Note that this is a *reversible* computer, which can be operated indifferently in “forward” or “reverse”.<sup>21</sup>

A sequence of chain configurations compatible with a given dynamics is a *trajectory* of that dynamics. A sequence that just obeys the kinematic links (and thus is a trajectory of *some* dynamics) will be called a *history*.

What kinds of behavior can such a chain display as one ranges over the possible dynamics? At one extreme, if the dynamics never grants permission, then no flips can possibly occur: in the *identity* dynamics, any initial configuration remains forever unchanged. At the other extreme, suppose that permission is always granted. Then one can immediately verify that a configuration consisting of all +1 links (Fig. 10a) or all -1 links cannot move at all. On the other hand, a configuration consisting of regularly alternating +1 and -1 links, and thus having an average slope of 0, will steadily march vertically at unit speed, moving upwards or downwards depending on whether the black dots are in the valleys or on the peaks at even times (Fig. 10c,d). It turns out that, for an arbitrary chain configuration, this permissive, “flip-when-ever-you-can” dynamics gives exact *wave equation* behavior; this will be proved in §6.2, where we also discuss other dynamics that can be “programmed” in our chain computer.

In sum, we have defined a whole class of dynamics; the general format is given by the kinematic constraints (try to flip at even or odd places at, respectively, even or odd times, and only if the links allow it), while the specifics is given by the permission table (flip only if the permission bit is set).

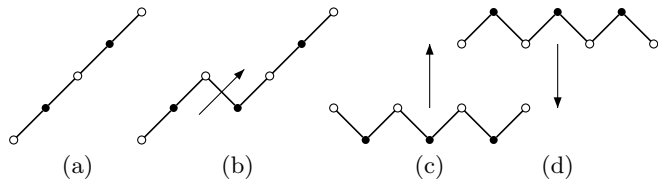


Figure 10: Examples, when permission to flip is always granted: A chain with slope +1 (or -1) cannot move at all (a); a kink on it will propagate at unit speed (b). A chain with alternating +1 and -1 links will march at unit speed upwards or downwards—depending on whether the black dots start in valleys or on ridges (c,d).

## 6.2 Means and flows

To get from a discrete chain to a continuous string, so as to get differentiable quantities which we can use in variational arguments, we must go to a *macroscopic level*; i.e., consider space averages over lengths that are large with respect to the lattice spacing, and time averages over a large number of steps. Ideally, coarse-grained averaging should commute with the dynamics; that is, if we average the microscopic state variables and then let this average evolve for a certain time using an appropriate macroscopic dynamics we should get the same result as if we had let them evolve over the same time using the microscopic dynamics on the microscopic data and then done the averaging.

At the microscopic level, knowledge of the *positions*  $q(x)$  completely specifies the system’s state; that is, the assignment of  $q(x)$  at a given instant completely determines the subsequent microscopic evolution of the system under a given dynamics. More specifically, initial rate-of-change information—which dots are going to flip at the initial step—is not part of the initial specifications but is dictated by the dynamics itself.

The situation is different at the macroscopic level. As a smeared-out, macroscopic state variable,  $q(x)$  does not capture enough information about the state of the system to uniquely determine its macroscopic evolution. In fact, as one can see in Fig. 10c,d, two chains with the same  $q(x)$  may move one up and the other down! More macroscopic data than just  $q(x)$ , namely, the rates of change  $\partial q(x, t)/\partial t$  (or some equivalent information) are needed to give a self-contained macroscopic dynamics, and, as we shall see, they are indeed sufficient.

Coming back to the microscopic level, let us fix our attention on one candidate dot (Fig. 8) and the unit-space cell surrounding it; this cell contains two links, one on the right and one on the left of the dot itself, whose values we shall call respectively  $\rho_{\rightarrow}$  and  $\rho_{\leftarrow}$ . From these we construct, by symmetrization, the new quantities

$$\begin{aligned} \rho &= (\rho_{\rightarrow} + \rho_{\leftarrow})/2, \\ j &= (\rho_{\rightarrow} - \rho_{\leftarrow})/2. \end{aligned} \tag{13}$$

On a unit-cell scale, the possible values for  $\rho$  and  $j$  are

$$\begin{array}{c|cc} & \rho_{\leftarrow} & \\ \hline \rho & -1 & +1 \\ \hline \rho_{\rightarrow} & -1 & 0 \\ & +1 & +1 \end{array}, \quad \begin{array}{c|cc} & \rho_{\leftarrow} & \\ \hline j & -1 & +1 \\ \hline \rho_{\rightarrow} & -1 & 0 \\ & +1 & +1 \end{array}.$$

<sup>21</sup>More than that, given appropriate interlocks it can go forward, backward, or idle independently at each point of the chain.

Thus, along the chain, we have the two microscopic sequences of links,  $\rho_{\rightarrow}(x)$  and  $\rho_{\leftarrow}(x)$ , and the two derived sequences  $\rho(x)$  and  $j(x)$ ; for example, for Fig. 8 we have

$$\begin{array}{cccccc} \rho_{\rightarrow} & \boxed{+1} & \boxed{+1} & \boxed{+1} & \boxed{-1} & \boxed{-1} & \boxed{-1} \\ \rho_{\leftarrow} & \boxed{+1} & \boxed{-1} & \boxed{+1} & \boxed{+1} & \boxed{-1} & \boxed{+1} \\ \rho & \begin{array}{|c|c|c|c|c|c|} \hline +1 & 0 & +1 & 0 & -1 & 0 \\ \hline 0 & +1 & 0 & -1 & 0 & -1 \\ \hline \end{array} & & & & & \\ j & & & & & & \end{array} \quad (14)$$

Under coarse graining,  $\rho(x, t)$  represents the average slope of the of chain (in the  $x$ - $q$  plane) while  $-j(x, t)$  represents its average velocity (the ‘‘slope’’ in the  $t$ - $q$  plane):

$$\begin{aligned} \frac{\partial q}{\partial x} &= \rho = \frac{\rho_{\rightarrow} + \rho_{\leftarrow}}{2}, \\ \frac{\partial q}{\partial t} &= -j = -\frac{\rho_{\rightarrow} - \rho_{\leftarrow}}{2}. \end{aligned} \quad (15)$$

To integrate the dynamics we have to determine the evolution of  $\rho$  and  $j$ , or, equivalently, or  $\rho_{\rightarrow}$  and  $\rho_{\leftarrow}$ . It turns out that if permission to flip is always granted, then the two sequences  $\rho_{\rightarrow}$  and  $\rho_{\leftarrow}$  of (14) shift respectively rightwards and leftwards one position at every step *without interacting*. In other words, the dependency on  $x$  and  $t$  is such that

$$\begin{aligned} \rho_{\rightarrow}(x, t) &= \rho_{\rightarrow}(x - t), \\ \rho_{\leftarrow}(x, t) &= \rho_{\leftarrow}(x + t). \end{aligned} \quad (16)$$

In this case, then, the sequence  $\rho$  can be thought of as the superposition of two traveling strings that move at unit speed in opposite directions. As we know, this is a solution of the one-dimensional wave equation. Thus, we conclude that under this ‘‘permissive’’ dynamics *the string*  $q(x, t)$  *strictly obeys the one-dimensional wave equation*

$$\frac{\partial^2 q}{\partial x^2} - \frac{\partial^2 q}{\partial t^2} = 0. \quad (17)$$

on any scale, with no damping whatsoever.<sup>22</sup>

If permissions are assigned randomly and independently with a density  $1 - s$ , then, with reference to Fig. 11, where the two trains  $\rho_{\rightarrow}$ ,  $\rho_{\leftarrow}$  are shown running on opposite tracks, when two cars pass by one another they will swap contents with probability  $\eta$ . The result is that, for small  $\eta$ , these two quantities are related by

$$\begin{aligned} \partial_t \rho_{\rightarrow} &= -\partial_x \rho_{\rightarrow} - s(\rho_{\rightarrow} - \rho_{\leftarrow}), \\ \partial_t \rho_{\leftarrow} &= -\partial_x \rho_{\leftarrow} + s(\rho_{\rightarrow} - \rho_{\leftarrow}), \end{aligned} \quad (18)$$

and thus  $\rho$ ,  $j$ , and  $q$  all satisfy the *telegraph* equation

$$\partial_{tt} u = \partial_{xx} u - 2s\partial_t u,$$

which is essentially the wave equation with damping proportional to  $s$ . For fixed  $s$ , if we scale  $t$  by  $a$  and  $x$  by  $\sqrt{a}$ , then in the limit as  $k \rightarrow \infty$  the telegraph equation turns into the *diffusion* equation,

$$\partial_t u = \frac{1}{2s} \partial_{xx} u.$$

<sup>22</sup>Norman Margolus and I came upon this model around 1983. On a computer simulation, it is impressive to see the full interplay of elasticity and inertia emerge from such simple discrete primitives.

Nonrandom tables, and especially tables with a fractal structure, need not lead to an emerging macroscopic dynamics. But in all cases we get a *linear* dynamics, since the  $\rho_{\rightarrow}$  and  $\rho_{\leftarrow}$  tokens of Fig. 11 are switched without even being looked at—they do not interact! Note that, in a measure-theoretical sense, *almost all* permission tables yield a uniform and independent distribution of permissions with density 1/2. In this sense, the typical dynamics of our class is that of a heavily damped string.

$$\begin{array}{cccccc} \rho_{\rightarrow} & \boxed{+1} \rightarrow & \boxed{+1} \rightarrow & \boxed{+1} \rightarrow & \boxed{-1} \rightarrow & \boxed{-1} \rightarrow & \boxed{-1} \rightarrow \\ \rho_{\leftarrow} & \leftarrow \boxed{+1} & \leftarrow \boxed{-1} & \leftarrow \boxed{+1} & \leftarrow \boxed{+1} & \leftarrow \boxed{-1} & \leftarrow \boxed{+1} \end{array}$$

Figure 11: The sequences  $\rho_{\rightarrow}$  and  $\rho_{\leftarrow}$  may be thought of as trains traveling in opposite directions at unit speed. A request to flip (in the original chain) corresponds to two train cars passing by one another; if the permission is denied, then the two cars swap contents, so that a bit that was traveling rightwards now starts moving leftwards and vice versa. This leads to scattering in the  $\rho_{\rightarrow}-\rho_{\leftarrow}$  picture, corresponding to *damping* in the  $q$  picture.

### 6.3 Inertia of the lumped string

We shall now focus our attention on the harmonic string behavior, as supported by the permissive dynamics in our toy computer.

Let us close the string into a loop of length  $m$  (‘‘periodic boundary conditions’’  $q(x + m) = q(x)$ ), and consider the position of its center of mass (on the  $q$ -axis) as a function of time. Whatever the configuration of the string, this point will move at a constant velocity  $\beta$  (since  $\int j dx$  is strictly conserved) in the range  $-1 \leq \beta \leq +1$ , with  $2n + 1$  discrete possible values

$$\beta = i/m, \quad \text{for } i = -m, \dots, -1, 0, +1, \dots, +m.$$

We will have

$$\begin{aligned} \rho_{\rightarrow} + \rho_{\leftarrow} &= 0, & \text{or} & & \rho_{\rightarrow} &= -\beta, \\ \rho_{\rightarrow} - \rho_{\leftarrow} &= -2\beta, & & & \rho_{\leftarrow} &= \beta. \end{aligned} \quad (19)$$

Of the  $m$  elements of the train  $\rho_{\rightarrow}$ , a fraction  $\rho_{\rightarrow}^{\pm}$  will consist of  $+1$ s and the rest,  $\rho_{\rightarrow}^{-}$ , of  $-1$ s, that is,

$$\begin{aligned} \rho_{\rightarrow}^+ + \rho_{\rightarrow}^- &= 1, \\ \rho_{\rightarrow}^+ - \rho_{\rightarrow}^- &= \rho_{\rightarrow}, \end{aligned} \quad (20)$$

and similarly for  $\rho_{\leftarrow}$ . Thus, the number of configurations compatible with velocity  $\beta$  is (using the same notation as in §4)

$$\mathcal{N} = \binom{m}{\rho_{\rightarrow}^+} \binom{m}{\rho_{\rightarrow}^-}, \quad \text{with} \quad \begin{aligned} \rho_{\rightarrow}^+ &= (1 - \beta)/2, \\ \rho_{\rightarrow}^- &= (1 + \beta)/2, \end{aligned} \quad (21)$$

or

$$\ln \mathcal{N} \approx 2mK(\beta) \approx -m\beta^2 + \text{const}, \quad (22)$$

which is essentially the same dependency on  $\beta$  as (9). This will of course give the correct results as the Lagrangian of the ‘‘lumped-string’’ free particle; moreover, the appearance

of a mass factor  $m$  makes it possible to study the interactions between particles of different masses<sup>23</sup> and see how the number of trajectories varies as a function of different masses and velocities.

## 6.4 Refraction

Here we'll briefly consider the case of a nonhomogeneous medium. If the permission table withholds permission for the two consecutive steps that make up a time unit (cf. top part of Fig. 9), then the evolution of the chain is frozen for that time unit. These “suspensions” may be randomly scattered through time, reducing the effective speed by a factor of  $n$  (see [13] and [9] for examples of this approach in lattice-gas dynamics). Suppose (Fig. 12) that the particle starts at the origin, traverses free space (“index of refraction”  $k = 1$ ) until time  $t$ , and then traverses a stretch of time with “index of refraction”  $k = n$ , finally landing at  $Q$ . What is the most likely value for  $q$  when the particle enters the denser medium? Equivalently, what is the “refraction angle”?

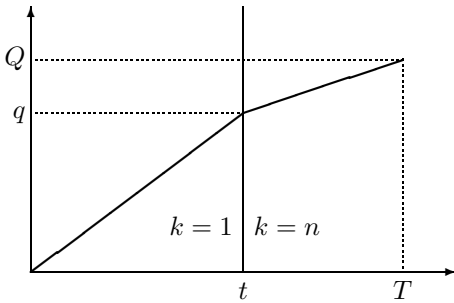


Figure 12: A lumped-string particle travels until  $t$  through a 100% permissive medium (“vacuum”); thereafter, the permission rate is granted only every  $n$ th time slot (“index of refraction”  $n$ ). If the average velocity in the denser medium is  $v$ , the string must be moving at an actual speed  $v' = nv$  when moving at all, and the path statistics that enters in the Lagrangian must be that corresponding to this higher speed  $v'$ .

We will derive the Lagrangian directly from the microscopic dynamics, always assuming that the Lagrangian must be an indicator of the number of trajectories compatible with the given data. In the denser medium, whatever the macroscopic velocity  $v$ , the particle must be following a sequence of configuration appropriate for a velocity  $v' = nv$ , with the only difference that on time slots when permissions are withheld the current configuration will be held frozen. Thus, in the denser medium, the Lagrangian must reflect the path statistics of this higher “internal” velocity rather than that of the apparent velocity  $v$ .

Multiplying the velocity by a factor of  $n$  in (19) shifts the statistics in (21) so that a factor of  $n^2$  appears in (22); rescaling time by a factor  $1/n$ , and then moving this factor from the  $dt$  term to the  $L$  term in  $LdT$  adds to the Lagrangian a factor  $1/n$ , so that the correct Lagrangian for index of refraction  $n$  is

$$L_n(v) = \frac{1}{n} n^2 L_{\text{vacuum}}(v) = nmv^2.$$

<sup>23</sup>Interactions between chains require a slightly more complex CPU for our computer, able to look further than just first neighbors when considering a flip.

This expression for the Lagrangian indeed gives, upon extremizing the action integral, the correct actual trajectory with  $q/t = n(Q - q)/(T - t)$ .

## 6.5 The string as an extended system

We shall now look at a chain of indefinite length undergoing harmonic motion as a spatially extended system.

If we denote by  $\mathcal{K}$ ,  $\mathcal{U}$ ,  $\mathcal{H}$ , and  $\mathcal{L}$  the densities of respectively kinetic energy, potential energy, total mechanical energy (Hamiltonian), and Lagrangian, the correspondence with analytical mechanics for the harmonic chain, given by (15) and (17), is completed by setting

$$\begin{aligned} \mathcal{K} &= j^2, & \mathcal{H} &= \mathcal{K} + \mathcal{U}, \\ \mathcal{U} &= \rho^2, & \mathcal{L} &= \mathcal{K} - \mathcal{U}, \end{aligned} \quad (23)$$

whence

$$\mathcal{H} = j^2 + \rho^2 = \frac{1}{2}(\rho_{\rightarrow}^2 + \rho_{\leftarrow}^2), \quad (24)$$

$$\mathcal{L} = j^2 - \rho^2 = -\rho_{\rightarrow}\rho_{\leftarrow}. \quad (25)$$

We can now go back and forth between the two levels, and find the microscopic interpretation, in this computational model, of the usual macroscopic concepts. For example, the quantity  $\rho$  represents the amount of stretch of the chain from the flat configuration in which links are randomly oriented up and down—and its square represent the energy stored in this “spring”. The quantity  $j$  represents the momentum of the string (the difference between the number of the “valleys”, where the chain moves up, and that of “peaks”, where it moves down), and its square represents the energy stored in the “inertia”. The sum of these squares is of course the total mechanical energy  $\mathcal{H}$ , which from the rightmost equality in (24) can also be seen as the sum of the energies of the two traveling waves  $\rho_{\rightarrow}$  and  $\rho_{\leftarrow}$ .

For a combinatorial interpretation of  $\mathcal{H}$ , consider a piece of chain of length  $m$  and proceed as in (21), but with  $\rho_{\pm}^{\pm}$  and  $\rho_{\pm}^{\mp}$  more generally given by

$$\begin{aligned} \rho_{\rightarrow}^{\pm} &= (1 + \rho_{\rightarrow})/2, \\ \rho_{\leftarrow}^{\pm} &= (1 + \rho_{\leftarrow})/2. \end{aligned} \quad (26)$$

We obtain

$$\mathcal{N} = \binom{m}{\rho_{\rightarrow}^{\pm}, m} \binom{m}{\rho_{\leftarrow}^{\pm}, m}$$

or

$$\begin{aligned} \frac{1}{n} \log \mathcal{N} &= K(\rho_{\rightarrow})K(\rho_{\leftarrow}) \\ &\approx -\frac{1}{2}(\rho_{\rightarrow}^2 + \rho_{\leftarrow}^2) = -\mathcal{H} + \text{constant}. \end{aligned}$$

Thus, in this model of the elastic string, the energy  $\int \mathcal{H}(x)dx$  measures, on a log scale, the number of microscopic chain configurations compatible with a given macroscopic assignment of positions and velocities. A low-energy state is “cheap” because it is “common”—there are so many ways to achieve it. Conversely, high-energy states are “rare”. In fact, the four states of maximal energy ( $j = 0$ ,  $\rho = \pm 1$ , as in Fig. 10a; or  $j = \pm 1$ ,  $\rho = 0$ , as in Fig. 10c,d) are each represented by a single microscopic configuration.

Since the underlying cellular automaton is microscopically reversible, its fine-grained entropy is strictly constant—rare states map into rare states, common ones into common ones; in this model, thence, energy conservation is just a macroscopic expression of microscopic reversibility.

Note that  $\mathcal{H}$  depends on the coarseness of graining; vibrations of a wavelength shorter than this grain do not contribute to  $\mathcal{H}$ , and may be viewed as thermalized degrees of freedom.

From (25), the wave equation (17) follows immediately by the Euler–Lagrange equation

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \frac{\partial q}{\partial t}} + \frac{d}{dx} \frac{\partial \mathcal{L}}{\partial \frac{\partial q}{\partial x}} = 0. \quad (27)$$

By (25),  $\mathcal{L} = -\rho_{\rightarrow}\rho_{\leftarrow}$ . Expanding  $\rho_{\rightarrow}\rho_{\leftarrow}$  by (20) yields

$$\rho_{\rightarrow}\rho_{\leftarrow} = (\rho_{\rightarrow}^+\rho_{\leftarrow}^+ + \rho_{\rightarrow}^-\rho_{\leftarrow}^-) - (\rho_{\rightarrow}^+\rho_{\leftarrow}^- + \rho_{\rightarrow}^-\rho_{\leftarrow}^+). \quad (28)$$

In the above equation, the four terms in parentheses represent the probabilities that two consecutive chain links, the first from  $\rho_{\rightarrow}$  and the other from  $\rho_{\leftarrow}$ , form

$$\begin{array}{ll} \rho_{\rightarrow}^+\rho_{\leftarrow}^+ & \text{a +1 slope} \\ \rho_{\rightarrow}^-\rho_{\leftarrow}^- & \text{a -1 slope} \\ \rho_{\rightarrow}^+\rho_{\leftarrow}^- & \text{a ridge} \\ \rho_{\rightarrow}^-\rho_{\leftarrow}^+ & \text{a valley} \end{array}$$

(cf. Fig. 10).

If a chain’s evolution obeys the permissive rule, a ridge yields a downward flip (cf. Fig. 8); a valley, an upward flip; and  $\pm 1$  slopes yield a rest. Thus, on any small patch of a proposed macroscopic spacetime history,

$$\begin{aligned} \mathcal{L}_{\text{actual}} &= \text{density of flips} - \text{density of rests} \\ &= 2(\text{density of flips}) + \text{constant}. \end{aligned} \quad (29)$$

Consequently, given spacetime boundary conditions for the macroscopic string configuration (e.g., the entire string at  $t_0$  and  $t_1$ ), the histories that are actual solutions of the permissive dynamics are those that fill in the intervening spacetime area with the *least number of flips*.

By the same token, the actual solutions—always for the permissive dynamics—are those that maximize the number of rests. Now, in this dynamics, the rests points are all and only those spacetime points where no motion was possible to begin with, because of kinematic constraints. Thus a rest is a point where withholding permission would not have made any difference! For any particular rest point in any particular history, a table with a 0 at that point instead of a 1 would have given the same history; consequently, that history is also a trajectory of this variant table. In conclusion, the actual data flow is that which *maximizes* the number of possible tables that would have given the same data flow—in other words, it is the flow that is maximally indifferent to the rule.

Actually, what we have called a “permissive” rule because permission to flip is always granted is really a *prescriptive* rule—If you can flip, you *must*! For an analogy, imagine a mythical country where the law essentially says, If you are

in a position to pay taxes then you must! Well, in that country it happens that the citizens, left to their own devices, naturally find all ways to do essentially the same things they would like to do, but without putting themselves in a position to have visible taxable income. Of course, I am talking about a mythical country.

The above intuitive comments are not about Lagrangian behavior in general; they are about the way the Lagrangian policy expresses itself in the highly distinguished case of the “flip-when-you-can” table. For other dynamics, the correspondence rules between microscopic rule, path statistics, and emergent behavior may well be more subtle and harder to interpret. Nonetheless, one hopes that examples of this kind will help understand how concepts are connected and *why*.

## 7 Action as computation capacity

Intuitively, we would like to define the “computation capacity” of a programmable machine as the number (on a log scale, for a number of practical reasons) of distinct *behaviors* the machine can achieve over the range of possible settings of the machine or *programs*. While different behaviors must come from different programs, two or more programs may yield the same behavior. We are interested in measuring the variety of actual behaviors, as constricted to the variety of settings. For example, a “ten-speed” bicycle has two levers, one with two settings (*A* and *B*) and one with five, giving ten settings overall. However, the *A* and *B* gear-ratio ranges overlap, as indicated in Fig. 13, giving a smaller number of effectively distinct “speeds”.

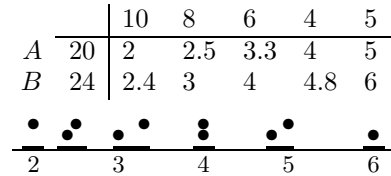


Figure 13: In a “ten-speed” bicycle, the two five-speed ranges substantially overlap, giving fewer than ten effectively distinct gear ratios. Here there are only nine distinct ratios, and, on a coarser grain, only six or seven practically different “speeds”.

Just as Shannon proposed a way to quantify “variety of *state*”, and appropriately called it *entropy* as a natural generalization from physics, here we propose a way to quantify “variety of *behavior*”, and we choose to call it *action* because we surmise that the well-known physical quantity called ‘action’ is but a special case of this. In the same sense as entropy measures (or defines) “amount of information”, so action measures (or defines) “amount of computation”.

It was not Shannon that invented entropy—that had been done by Clausius—or connected it with the number of microscopic states or “complexions” of a system—that was done by Boltzmann. Shannon’s originality was to abstract those aspects of the “incompleteness of description” of a system that can be quantified in a universal and objective way. In brief, entropy is a property of a *distribution*: give me a distribution—from any discipline—and it will have an

entropy. The entropy of the state of a physical system is a special case of this, insofar as that state can be identified with a distribution. Change the information about the system and the distribution will change even if the system has not changed, and the entropy will correspondingly change. Thus entropy is not, strictly speaking, a quantity, but an *operator*—or a recipe, if you wish—which yields a number when applied to a certain scenario. Without any contradiction, to one and the same system there may be associated several quantities all having the “dimension” of entropy, just as a desk may be described by several parameters all having the dimension of “length”, namely its length, height, depth, etc. In sum, entropy is a *yardstick* for measuring a certain type of quantity. Similarly, to a given computational scenario there may be associated many quantities all having the “dimension” of action. Action is another yardstick, for measuring another type of quantities.

In physics, as well as in more abstract settings, one can speak of “state” before introducing a dynamics; in fact, one can compare different dynamics operating on the same state set. Thus entropy, thought of as amount of incompleteness in the specification of a system’s state, is meaningful independently of dynamical considerations: the subject matter of entropy is *state*. The subject matter of action is a concept that is dual to that of state, namely, *law*: action is meant to measure the amount of incompleteness in the specification of a system’s law. Ultimately, a dynamics can be defined, extensionally, by just listing the set of trajectories or histories that are “legal” or compatible with that dynamics, contrasted with all other potential trajectories or histories. Therefore, “uncertainty as to trajectory” and “uncertainty as to law” are related concepts.

Consistently with the more abstract setting we would like to establish, it will be convenient to use the terms “computer”, “program”, and “computation” as an alternative to “system”, “dynamics”, and “trajectory”.

We define in a most general way a **computer** as a *probability distribution over a family  $\mathbb{H}$  of “computational histories”* (we will specify in a moment what kind of mathematical object  $\mathbb{H}$  is). Without any qualifiers, the **action** of a computer is defined as the entropy of this distribution. Any conditioning of this distribution leads to a new distribution, and thus, technically speaking, to a new computer. Thus, action can be seen as the operator that associates to any conditioning of a distribution of histories the entropy of the corresponding conditioned distribution.

If we wish, we may choose two mutually independent sets of random variables over  $\mathbb{H}$ , one collectively called the **program** and the other the **input**, and a third set called the **output**, and consider the marginal distributions over these sets. These are just narrower ways to define a computer; that is, an input/output relationship ignores all internal details and characterizes a computer purely in terms of the function it computes—a “black box”; a program is just a set of switches used to parameterize this function.

What is a history? Let a **network**  $G$  (Fig. 14) be an acyclic, directed, colored graph. That is,  $G$  consists of a collection of arcs called **signals** and a collection of nodes called **events**. To each arc there is associated a set called its

**state alphabet** (this is an arc’s “color”). To each node there are associated a set of signals called its **inputs** and a set of signals called its **outputs**. If a signal  $i$  is an input of a given node and a signal  $j$  an output of the same node,  $j$  is said to be **later** than  $i$ . We require the relation ‘later’ to be a *partial order* (this is the meaning of “acyclic”). If the network is finite, since the graph is acyclic some signals will not be inputs of any nodes—these are the network’s **outputs**; similarly, the *inputs* are those signals that are not outputs of any nodes. A network’s **history** is simply an assignment of states to every arc, each taken from the arc’s state alphabet.

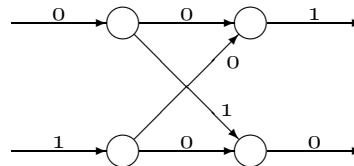


Figure 14: A *network* with a particular history on it. While each arc has a state alphabet associated with it (for instance, the Boolean set  $\{0, 1\}$  for all arcs), no properties are assigned to nodes. A *computer* is defined by a distribution on the set of all possible histories.

To visualize the above definition, a network can be imagined as an ordinary *combinational network* (Fig. 14) in which all information about the nature of the gates has been erased, and only the nature of the signals (e.g., binary) and their interconnection topology has been retained. A network turns into a computer as soon as we assign a distribution over its histories. A computer may be **deterministic, invertible, local**, etc., depending on the specifics of this distribution.

We shall immediately illustrate the above definitions with an example, namely the “canonical” computer, consisting of an ordinary PROM (Programmable Read-Only Memory) or EPROM (Erasable PROM). This is a network consisting of a single node with  $m$  input lines called collectively the *address*,  $n$  output lines called the *output*, and  $k$  more input lines (where  $k = n2^m$ ) called the *program*. The distribution of an ideal PROM is the following. Write the program as a table of  $2^m$  rows and  $m$  columns. For every program  $p$  (i.e., an assignment of binary values to all  $k$  program lines) and address  $x$ , take the  $2^n$  histories corresponding to the possible values of the output  $y$ , and mark that in which  $y$  coincides with the  $x$ th row of  $p$  (the collection of marked histories represents the “correct” behavior, as per specs, of the ROM). Finally, assign equal weights to all marked histories (there are  $2^{m+k}$  of them) and normalize to 1, giving 0 weight to all others.

What is the action  $A$ , or the “computational worth”, of this canonical computer? Using base-2 logs for evaluating entropy, we get

$$A = k + m = n2^m + m.$$

The overwhelming contribution is given by  $n2^m$ , which is simply the number of PROM program bits; this is the log of the *number of distinct functions* that our computer can compute, or the number of “different things it can do”.

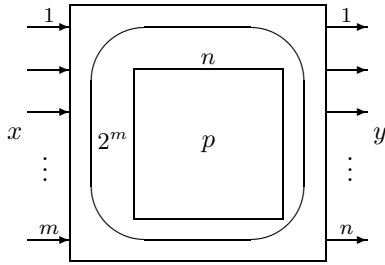


Figure 15: The PROM as a canonical computer, with  $m$  lines of address  $x$ ,  $n$  lines of output  $y$ , and  $n2^m$  bits of program  $p$ .

The small correction  $m$  is proportional to the depth of the addressing decoding tree.

Note that in a real PROM chip the programmable bits take up most of the chip's real estate; the decoding tree occupies a much smaller area and consists of a fixed—nonprogrammable—gates. Thus, our estimator  $A$  of computational worth is fully consistent with the market reality, where PROM chips of different organization (for example, 32-vs 8-bit wide) but with the same number of programmable bits and the same speed have essentially the same size, cost essentially the same, and have a comparable market share. We started with an abstract, model-independent definition, which did not know about transistors, area, speed, etc., and we ended up with a gauge that makes perfect sense.

Suppose now that we want to estimate the degradation of computational worth if the chip becomes less-than-ideal; for instance, a stuck address or output line, a stuck program bit, a probability of error in programming, addressing, or read-out—or even a change in the distribution of possible address patterns (e.g., three-quarters of the accesses are expected to be in the lower half of the address range). We just modify the probability distribution so as to reflect the new information about the “law” embodied by the chip and the usage context, and from the same formula we get the *action* of the new situation.

Finally, we can ask problems of the following kind. Given a certain macroscopic specification of a computational task, for example, an input/output relationship defined in terms of a joint input/output distribution, and given a certain computer as defined here, what is the mutual information between the two? In other words, how much do we have to condition the computer's distribution in order to obtain that particular input/output distribution? Returning to ordinary parlance, just as a computer is *powerful* if there are a lot different things that it can do, a computational task is easy if there are a lot of different ways to write a program for it. In this scheme of things, if my interpretation of analytical mechanics as emergent from an underlying fine-grained dynamics is correct, the action obtained by integrating the Lagrangian of a dynamics over a particular proposed trajectory tells us how “natural” that particular trajectory is for that dynamics—how good the fit between the two is.

## 8 Acknowledgments

This research was funded in part by NSF (DMS-9596217) and by the I.S.I. Foundation (Turin, Italy). Part of this

work was carried out during the 1997 Elsas-Bailey–I.S.I. Foundation research meeting on quantum computation.

I would like to thank Lev Levitin, Sandu Popescu, and Zac Walton for stimulating discussions.

## List of references

- [1] ARNOLD, Vladimir, *Mathematical Methods of Classical Mechanics*, Springer–Verlag 1978.
- [2] FEYNMAN, Richard, and A. HIBBS, *Quantum Mechanics and Path Integrals*, McGraw–Hill 1965.
- [3] HILDEBRANDT, Stefan, and Anthony TROMBA, *The Parsimonious Universe*, Springer–Verlag 1996.
- [4] KADANOFF, Leo, “On Two Levels,” *Physics Today* (September 1986), 7–9
- [5] LANCZOS, Cornelius, “The Variational Principles of Mechanics”, 4th edition, Dover.
- [6] LANDAUER, Rolf, “Information is physical”, *Physics Today* **44** (May 1991), 23–29.
- [7] MARGOLUS, Norman, and Lev LEVITIN, “The maximum speed of dynamical evolution”, to appear in *Physica D*.
- [8] SCHROEDER, Manfred, *Number Theory in Science and Communication*, second enlarged edition, Springer–Verlag 1986.
- [9] SMITH, Mark, “Representation of geometrical and topological quantities in cellular automata,” *Physica D* **45** (1990), 271–277.
- [10] SOURIAU, Jean–Marie, *Structure of Dynamical Systems*, Birkhäuser 1997.
- [11] TYAGI, Akhilesh, “A principle of least computational action”, *Workshop on Physics and Computation*, IEEE Computer Society Press (1993), 262–266.
- [12] TOFFOLI, Tommaso, “How cheap can mechanics' first principles be?” *Complexity, Entropy, and the Physics of Information* (W. H. ZUREK ed.), Addison–Wesley 1990, 301–318.
- [13] TOFFOLI, Tommaso, and Norman MARGOLUS, *Cellular Automata Machines—A New Environment for Modeling*, MIT Press 1987.