An explanation of the Shannon entropy, with relevance to sediment transport

David Jon Furbish

Emeritus, Vanderbilt University

June 2024

1 Preamble

Entropy seems to be a popular but enigmatic concept from science that is frequently used to qualitatively explain, at a high level, the behavior and configurations of systems, yet which is frequently misunderstood in practice. To complicate things, there are two principal definitions of entropy: the Gibbs entropy from statistical mechanics and thermodynamics, and the Shannon entropy from information theory. Consider a delightful comment attributed to Claude Shannon regarding his work on information entropy:

> [John] Von Neumann told me, "You should call it entropy, for two reasons. In the first place your *uncertainty* function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." (McIrvine and Tribus, 1971; italics mine)

Indeed, qualitative explanations of entropy often center on the idea that it is a measure of the degree of disorganization of things, for example, the scattered arrangement of papers and books in my office (a high-entropy configuration) relative to the well-organized placement of things in my colleague's office (a low-entropy configuration). With that notion, we are comfortable with the idea that a uniform distribution of ink molecules thoroughly dissolved in a beaker of water represents a high-entropy configuration — a maximum disorganization of the molecules relative to the initial, highly organized ink drop that we placed in the beaker. (On the other hand, we might give pause to the idea that the uniformity of the molecules itself represents an organized state, if we use uniformity as a measure of organization.) But then things might seem confusing when we learn that according to the Shannon entropy the uniform distribution of molecules has greater information content than did the initial molecular configuration represented by the drop — particularly when we recall from our studies of particle diffusion that the final, uniform configuration represents a condition where all information regarding the initial configuration, the drop, is irreversibly lost. In approaching this topic, let us therefore start with simple wisdom offered by the philosopher and physicist David Wallace in reference to the meaning of probability:

> When unsure what something is, it often pays to ask what it does. (Wallace, 2012)

This is our approach here: to show that entropy foremost is a measure of uncertainty, and that it provides a useful way to describe how things are organized. We focus on the Shannon entropy from information theory rather than the Gibbs entropy from statistical mechanics and thermodynamics. Indeed, Jaynes (1957a) points out that the Gibbs entropy is actually a special case of the Shannon entropy, and the Shannon entropy certainly is more useful in describing attributes of sediment systems given its focus on information and uncertainty rather than thermodynamic quantities.

Our objective is to highlight elements of this topic that are relevant to sediment systems. We start with basic concepts: the meaning of the *information content* contained in the Shannon entropy and its relation to *uncertainty* represented by a probability distribution. We then examine the foundational idea of Maxwell–Boltzmann counting of particle states, leading to the canonical example of a maximum entropy distribution, the Boltzmann distribution. This is aimed at showing where the idea of a maximum-entropy distribution originated, thus providing the context for the later work of Shannon. Indeed, starting with the Shannon entropy absent a description of the basis of the Boltzmann distribution would risk missing how the Shannon entropy is applicable to mechanical systems. As Jaynes (1957a, p. 622) notes:

> The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the "amount of uncertainty" represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions which make it reasonable.

With this context in place we then turn to the maximum entropy method championed by Edwin Jaynes (1957a, 1957b, 2003), highlighting its emphasis on appealing to mechanical considerations in describing distributions of particle states while being as faithful to what we do not know as we are to what we do know about a system. We also show why the differential entropy can only be viewed as an analogy with the Shannon entropy.

2 Basic Concepts

The Gibbs entropy is defined as

$$S(x) = -k_\mathrm{B} \sum_x p_x(x) \ln p_x(x) \, , \tag{1}$$

where k_B is the Boltzmann constant. The Shannon entropy is defined as

$$H(x) = - \sum_x p_x(x) \ln p_x(x) \, . \tag{2}$$

In the work of Boltzmann and Gibbs, x represents an energy state so that $p_x(x)$ represents the probability mass function of these states, the set of all possible ways to arrange a great number of particles into accessible energy states subject to macroscopic (thermodynamic) constraints. In the work of Shannon (1948a, 1948b), x represents an element of a system of communication, for example an alphabet, so that $p_x(x)$ represents the probability mass function of the occurrence of such elements in relation to transmitted information. In addition, for a continuous random variable x with probability density function $f_x(x)$, the differential entropy is conventionally defined as

$$H(x) = - \int_x f_x(x) \ln f_x(x) \, \mathrm{d}x \, , \tag{3}$$

which is in analogy with (2). We postpone consideration of this continuous case for a later section.

Note straightaway that we are using the natural logarithm in these definitions, consistent with the standard definition of the Gibbs entropy. However, the base of the logarithm may vary depending on the conventions of the field of application.[1] Then notice that according to the law of the unconscious statistician the Gibbs entropy is the expectation — the average — of the quantity $-\ln p_x(x) = \ln[1/p_x(x)]$ multiplied by the Boltzmann constant, and the Shannon entropy is just the average of the quantity $-\ln p_x(x) = \ln[1/p_x(x)]$. Our task now involves clarifying what $\ln[1/p_x(x)]$ represents, and the implications of the similar forms of the definitions (1) and (2). We examine the context of the definitions (1) and (2) in the next section.

To simplify notation we denote $p_i = p_x(x_i)$ for discrete values x_i. We refer to x_i as an outcome, so we may say p_i is the probability that the ith outcome will occur. We now denote an *information function* or *surprisal* as $I(p_i)$. This function has three properties. First, $I(p_i)$ monotonically decreases with p_i. That is, the information represented by the occurrence of a particular outcome decreases as the likelihood of the outcome increases. Or, the occurrence of an unsurprising (likely) outcome represents less information than one whose occurrence is surprising (unlikely). Second, when $p_i = 1$ then $I(p_i = 1) = 0$, which coincides with an outcome whose occurrence is completely certain with no surprise. Third, consider the joint occurrence of two independent outcomes x_1 and x_2 with probabilities p_1 and p_2. Given that the probability of the joint occurrence of the outcomes x_1 and x_2 is $p_1 p_2$, Shannon insisted that the information $I(p_1 p_2)$ represented by the joint occurrence should be given by $I(p_1 p_2) = I(p_1) + I(p_2)$. That is, the total information of the joint occurrence is the sum of the information represented by each outcome. Shannon thus defined $I(p_i) = \log(1/p_i) = -\log p_i$, which satisfies the three desired properties. We may thus write the Shannon entropy as

$$H = \sum_i I(p_i) p_i \,, \tag{4}$$

showing that the entropy is the average information content of an event or series of events taking into account all possible outcomes represented by the distribution p_i.

Because the Shannon entropy usually appears in relation to communication and signal processing, we should be clear about the meaning of information content as used here. In writing text we normally intend the text to *inform* (i.e. provide information to) its audience about a topic, in which case the intended *meaning* of the text must be clear from our construction of the elements — the characters of the alphabet or the words — selected with varying probabilities from the set of possible elements. However, the definition of entropy in (4) does *not* depend on the meaning of the transmitted information. Rather, the information that is characterized by the entropy is actually information about the form of the probability distribution p_i that describes the likelihood of the occurrence of the outcomes (the characters or words), not the meaning of the characters and words per se. These comments apply more broadly to operational applications of information entropy in encoding and transmitting messages, data compression, image processing and so on.

In considering mechanical systems it is more useful to focus on the idea of entropy as a measure of uncertainty, or our absence of knowledge, rather than information per se. (This becomes essential when we consider the differential entropy given by (3) in Section 4.) This emphasizes the

[1] Normally the base 2 logarithm is used in information theory. Moreover, because $p_x(x)$ in (1) and (2) is dimensionless, taking the logarithm $\ln p_x(x)$ presents no issues. However, because the density $f_x(x)$ in (3) may have dimensions determined by x, care must be given to casting x as a dimensionless quantity, or ensuring that dimensions are correctly preserved in manipulations of (3).

distribution of possible outcomes rather than the outcomes themselves. Thus, consider the canonical example of a Bernoulli trial representing a coin toss. Letting p_i denote the probability of the two possible outcomes $k = [0, 1]$ given by the Bernoulli distribution, then $p_1 = p$ is the probability that $k = 1$ will occur and $p_2 = q = 1 - p$ is the probability that $k = 0$ will occur. The entropy is then given by

$$\begin{aligned} H &= -\sum_{i=1}^{2} p_i \ln p_i \\ &= -p \ln p - q \ln q \\ &= -p \ln p - (1-p) \ln(1-p) \, . \end{aligned} \tag{5}$$

The entropy H systematically varies with p (Figure 1). It is zero at $p = 0$ and $p = 1$ with a maximum at $p = 1/2$. Taking the derivative $\mathrm{d}H/\mathrm{d}p$ and setting the result to zero,

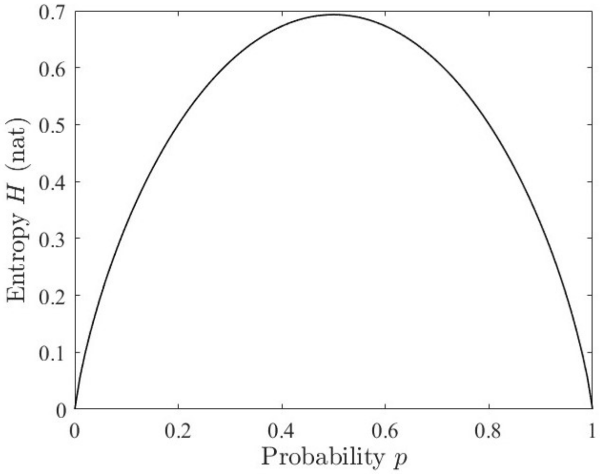

Figure 1: Plot of Shannon entropy H versus the probability p associated with a Bernoulli distribution.

$$\ln p = \ln(1-p) \, . \tag{6}$$

Exponentiation gives $p = 1 - p$, or $p = 1/2$. Thus, the entropy H is a minimum with a perfectly unfair coin ($p = 0, 1$) and it is maximized with a fair coin ($p = 1/2$). Moreover, the entropy does not favor the outcome $k = 0$ or $k = 1$. For example, it is the same for $p = 0.8$ with $q = 1 - 0.8 = 0.2$ and $q = 0.8$ with $p = 1 - 0.8 = 0.2$. Using the language of information, each successive toss of a perfectly unfair coin provides no new information, as the outcome is certain. Each toss of a fair coin provides the maximum information.[2] Note, too, that the Bernoulli distribution with maximum entropy coincides with a uniform distribution.

It should be clear from this example that the uncertainty of the outcome of a coin toss is a maximum when $p = 1/2$, coinciding with the maximum entropy condition. This extends to multiple trials as well as systems of coins. For example, with four trials and $p = 1$ the Bernoulli distribution

[2] Using base 2 logarithms, each toss of a fair coin ($p = 1/2$) provides one bit (one shannon) of information. Two trials give $I(p^2) = I(p) + I(p) = 2$ bits. Because $p = q$, n trials give n bits, regardless of the individual outcomes.

tells us that we must have one sequence of four 1s. We thus have complete knowledge of the system with no uncertainty. Likewise, for $p = 0$ we must have one sequence of four 0s with no uncertainty. But with $p = q = 1/2$ there are $2^4 = 16$ possible unique sequences of 0s and 1s with maximum uncertainty in their order of appearance.

Instead of trials, consider a system of four coins. With $p = 1$ all four must be 1s, or if $p = 0$ all four must be 0s, with no uncertainty. But with $p = 1/2$ we have two coins with 0s facing up and two with 1s facing up. Now with maximum uncertainty we have six possible arrangements: [1 1 0 0], [1 0 1 0], [1 0 0 1], [0 1 1 0], [0 1 0 1] and [0 0 1 1]. As a preview of material in the next section, this configuration involving four coins with two 0s and two 1s may be viewed as a *macrostate*, and the six arrangements may be viewed as *microstates*.

More generally, consider a distribution $p_x(x_i)$ describing an arbitrarily large number of possible states x_i. Then let us recall the comments of Jaynes appearing in the preamble above, "...that a broad distribution represents more uncertainty than does a peaked one..." If $p_x(x_i)$ has a sharp peak centered on a particular value of x_i, then the occurrence of this and immediately surrounding values is unsurprising, like the outcome of an unfair coin, relative to values of x_i with small probabilities $p_x(x_i)$. In contrast, if $p_x(x_i)$ is relatively uniform over all x_i, then our uncertainty about the likely occurrence of any x_i increases, and this uncertainty is what the entropy is measuring.

Experience suggests that confusion can arise from the idea that a high entropy condition has greater information content than a low entropy condition. This has to do with the meaning of information as used in different contexts. With certain physics and geometry problems we sometimes associate "information" with such things as system configuration. Using the example from the preamble, we might imagine the fully dissolved ink molecules as having lost all "information" concerning the configuration of the low entropy, highly organized initial conditions of the molecules (the drop), thus implying that low entropy is associated with high information content. Likewise we might associate spatial randomness of moving particles on a streambed with high entropy, where the statistical uniformity in the arrangement of the particles implies that information associated with a more organized arrangement (e.g. clustering; Roseberry et al., 2012) is missing. These views of entropy and information are reasonable, but they differ from the meaning of information and its relation to entropy in information theory.

For completeness we note that non-traditional definitions of entropy exist. For example the Tsallis (1988) entropy is intended to address non-additive entropies of independent events, and has been applied to various specialized systems. Here, however, we adopt the view of Peterson et al. (2013), who highlight the conclusions of Shore and Johnson (1980). Namely, because the definitions of entropy provided by Gibbs and Shannon uniquely ensure addition and multiplication rules of probability, any other definition of entropy yields a bias in the fitting of data. Peterson et al. (2013) suggest that this offers a "compelling first-principles basis for defining a proper variational principle for modeling distribution functions" (Section 4).

We started this section by presenting the definitions of the Gibbs entropy and the Shannon entropy, then focused on the meaning of the information function $I(p_i) = \ln(1/p_i) = -\ln(p_i)$ appearing in these definitions and its relation to uncertainty represented by the probability distribution p_i. Our next task is to examine the context of the definitions (1) and (2) to gain a sense of where they come from.

3 Macrostates, Microstates and the Boltzmann Distribution

Consider a system containing a fixed number N of particles. For simplicity this system consists of an isolated box containing the N particles. We are interested in two attributes of the particles: their spatial configuration and their energy states. Each particle has three degrees of freedom in movement defined by its xyz coordinate position, and three degrees of freedom defined by its velocity components parallel to the three coordinate directions. The N particles are free to move anywhere within the box, and at any successive instants they are in different configurations. Due to particle–particle collisions the energy of an individual particle changes, so at any instant the N particles have a distribution of energy states. We now want to know the most likely configuration and the most likely distribution of energy states. We start with particle positions.

3.1 Configurations

To illustrate the essential points involved, we use an example involving a small (manageable) number of particles N. We also choose a small box, and further imagine it to be "one-dimensional" such that we can focus just on the x coordinate positions of the particles. With $N = 4$, suppose the particles are constrained to be in four possible coordinate positions: x_1, x_2, x_3 and x_4. Now consider all possible arrangements of the four particles in these four coordinate positions (Figure 2). Each possible configuration is a *macrostate*. In general, for N particles distributed among m coordinate positions the number of macrostates is given by

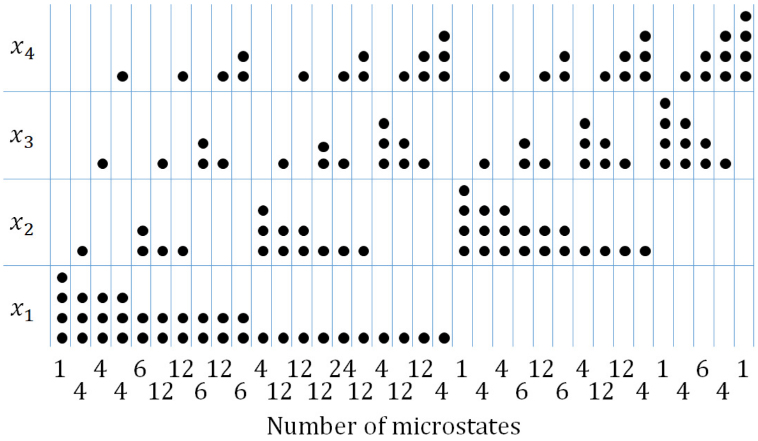

Figure 2: Schematic illustration of 35 macrostates involving four particles and four coordinate positions, together with the number of microtates in each macrostate.

$$N_{\mathrm{e}} = \frac{(N+m-1)!}{N!(m-1)!}\,. \tag{7}$$

In our example with $N = 4$ and $m = 4$ there are $N_{\mathrm{e}} = 35$ macrostates.

Classical particles are *distinguishable* with respect to coordinate position and velocity, and thus energy. Here we focus just on coordinate position. To see the significance of this we label the four particles as A, B, C and D. Now consider the first macrostate in which the number of particles $n_1 = 4$ in position x_1 (Figure 3). Because all particles are in the same coordinate state x_1, they

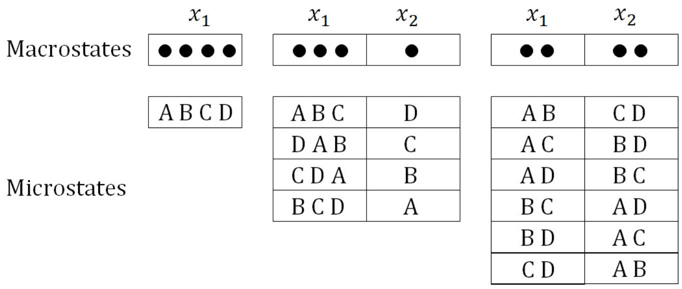

Figure 3: Schematic illustration of the first, second and fifth macrostates in Figure 00, showing associated microstates with four distinguishable particles A, B, C and D. Note that the ordering of particles within an individual coordinate position is immaterial, and that the first macrostate at position x_1 also is a microstate.

are *indistinguishable.* In contrast, consider the second macrostate with $n_1 = 3$ in x_1 and $n_2 = 1$ in x_2. The single particle at position x_2 could be A, B, C or D. But any three particles in the same position x_1 are indistinguishable with respect to coordinate position. Thus, for the second macrostate there are a total of four possible arrangements of the four distinguishable particles in two coordinate states, x_1 and x_2. Each of these four possible arrangements is a *microstate.* In general, for N particles distributed among m coordinate positions the number of microstates in a macrostate is given by

$$n_{\mathrm{e}} = \frac{N!}{n_1!n_2!n_3!...n_m!}\,, \tag{8}$$

where $n_1, n_2, n_3, ..., n_m$ denotes the numbers of particles in the coordinate states $x_1, x_2, x_3, ..., x_m$. Notice in our example that each macrostate with all four particles in one coordinate state has one microstate (Figure 2). That is, the macrostate coincides with the microstate. Also notice that the macrostate with one particle in each of the four coordinate states has 24 microstates.

This enumeration of distinguishable particle states is referred to as Maxwell–Boltzmann counting (or Maxwell–Boltzmann statistics). The numerator in (8) gives the total number of ways to uniquely order N distinguishable particles into distinct coordinate positions. The denominator removes from the counting those instants in which, with $n_i > 1$, the particles are indistinguishable with respect to coordinate position. A second type of counting, referred to as Bose–Einstein statistics, focuses on indistinguishable particles, essentially starting with (7). Bose–Einstein statistics converge to Maxwell-Boltzmann statistics at high temperatures or with sufficiently small particle number densities.

With reference to Figure 2, all 35 macrostates are *admissible* representations of the system for the given constraints, and in turn, each of the 256 microstates is an admissible representation of the system. Four macrostates have one microstate, 12 macrostates have four microstates, six macrostates have six microstates, 12 macrostates have 12 microstates, and one macrostate has 24 microstates. Following Gibbs (1902), we designate this set of 256 microstates as an *ensemble* of microstates.[3] We now make a far-reaching assumption — the foundational assumption of classical statistical mechanics. Namely, we assume that all microstates are independent and equally likely. That is, each microstate has a probability equal to 1/256 of occurring. Using all digits,

[3] In effect we imagine, as did Gibbs (1902), "a great number of independent systems, identical in nature, but differing in phase, that is, in their condition with respect to configuration and velocity."

the probability is $4/256 = 0.015625$ that the system has all particles clumped at one coordinate position. The probability is $48/256 = 0.1875$ that the system has a clump of three particles at one coordinate position and one at another position. The probability is $36/256 = 0.140625$ that the system has two particles in each of two positions. The probability is $144/256 = 0.5625$ that the system has two particles in one position and one particle in each of two positions. And, the probability is $24/256 = 0.09375$ that the system has one particle in each of four positions. These probabilities sum to unity. Thus clumping is less likely than is a more uniform configuration. As we will see below, clumping represents a relatively low entropy configuration and uniformity represents a relatively high entropy configuration.

We can go beyond this. Because the microstates are assumed to be equally probable, we can calculate expected proportions of particles in each coordinate position by averaging over the 256 microstates. This amounts to summing the product $n_i n_\mathrm{e}$ in each coordinate position, then dividing by the total sum of this product over all coordinate positions. More directly, we can observe in Figure 2 that the four coordinate positions contain identical sets of n_i. We therefore discover that the expected distribution $p_x(x_i)$ of coordinate states x_i is uniform and equal to $1/m = 1/4$. That is, upon selecting a particle at random, the probability that it is located within any one of the four coordinate positions is $p_x(x_i) = 1/4$. Let us note that this averaging over all microstates is doable for small numbers of macrostates and microstates. But as we will see next, this is not the strategy that Boltzmann adopted.

Let us now consider a slightly more formal analysis, loosely following the presentation of Furbish and Schmeeckle (2013) and unpublished notes of A. M. Steane, to provide a view of the original arguments of Boltzmann and in turn the basis for appealing to entropy in describing probability distributions. We start by assuming large N and large $n_1, n_2, n_3, ...$ in (8). As a point of reference, for $N = 10$ particles distributed among $m = 2$ coordinate states, there are a total of $N_\mathrm{e} = 11$ macrostates and a total of $n_\mathrm{e} = 1\,024$ microstates. For $N = 10$ and $m = 5$, there are $N_\mathrm{e} = 1\,001$ macrostates and $n_\mathrm{e} = 9\,765\,625$ microstates. And, for $N = 10$ and $m = 10$, there $N_\mathrm{e} = 92\,378$ macrostates and $n_\mathrm{e} = 10^{10}$ microstates. One can imagine that these numbers become unimaginably large for gas systems involving the Avogadro number of $6.022\,140\,76 \times 10^{23}$ particles per mole and a great number of coordinate states.

Using Stirling's approximation of factorials we write (8) as

$$\ln n_\mathrm{e} = N \ln N - N - \sum_i (n_i \ln n_i - n_i)\,. \tag{9}$$

Now note that $n_i = Np_i$, where for simplicity of notation the probability mass function $p_i = p_x(x_i)$. Then (9) becomes

$$\begin{aligned} \ln n_\mathrm{e} &= N \ln N - N - N \sum_i [p_i(\ln p_i + \ln N) - p_i] \\ &= N \ln N - N - N \sum_i p_i \ln p_i - N \ln N \sum_i p_i + N \sum_i p_i\,. \end{aligned} \tag{10}$$

By the definition of a probability distribution the last two sums equal unity, so

$$\frac{\ln n_\mathrm{e}}{N} = -\sum_i p_i \ln p_i\,. \tag{11}$$

Observe that the sum in (11) is the Shannon entropy of the distribution p_i, which is notable in that we have reached this point just based on the counting of particle states. Also notice that the Boltzmann constant k_B is not involved. That is, the formulation that follows launches from the Shannon entropy rather than the Gibbs entropy. As an aside, if we interpret (11) in terms of information content, and using our example with four particles, then the right side of (11) gives the average information content of the configuration of particles in the four coordinate states x_i depending on the set of p_i and thus the macrostate.

We now want to know the set of p_i that gives the maximum number of microstates n_e, which is the same as the set of n_i, and thus the macrostate, that maximizes n_e. In this problem, maximizing n_e is the same as maximizing $\ln n_\mathrm{e}$. And to be clear, maximizing n_e is the same as maximizing the entropy given by the right side of (11). To proceed we use the method of Lagrange multipliers. We denote $f(p_i) = \ln n_\mathrm{e}/N$ so that

$$f(p_i) = -\sum_i p_i \ln p_i \,. \tag{12}$$

To constrain the problem we specify that the probabilities sum to unity,

$$\sum_i p_i = 1 \tag{13}$$

We now write

$$f(p_i) = -\sum_i p_i \ln p_i + \alpha \left(\sum_i p_i - 1 \right) \,, \tag{14}$$

with Lagrange multiplier α. The function f has a stationary value when

$$\frac{\partial f}{\partial p_j} = 0 \,. \tag{15}$$

Here, p_j denotes the set of p_i representing the maximum, and we note that the derivative $\partial f/\partial p_j = 0$ for all $p_{i \neq j}$. Taking derivatives over the system of equations then leads to

$$\ln p_j + 1 - \alpha = 0 \,. \tag{16}$$

Thus, the set of p_j (the macrostate) representing the maximum number of microstates is

$$p_j = e^{\alpha - 1} \,. \tag{17}$$

Because $e^{\alpha-1}$ is a constant, $p_j = p_x(x_j)$ must be a uniform distribution. Specifically, because the set of p_j must sum to unity over m discrete coordinate positions,

$$\sum_{j=1}^{m} p_j = \sum_{j=1}^{m} e^{\alpha-1} = e^{\alpha-1} m = 1 \,. \tag{18}$$

From this it follows that

$$p_j = \frac{1}{m} \,. \tag{19}$$

That is, the most probable configuration of the set of n_i and thus x_i is uniform. Moreover, among all possible configurations, this uniform configuration of particle coordinate states is the one with the maximum entropy.

Recall in our example with small N and m that we inferred the most likely particle configuration as being uniform by averaging the particle positions over all equally probable microstates. This is possible for small N and m, but is not manageable with large N and m. Here we emphasize that, instead, Boltzmann aimed at determining the macrostate representing the most microstates. We examine the justification for this strategy when we consider energy in the next section.

Here is a point of interest with important implications for interpreting sediment systems. The arguments presented above apply to arbitrarily large numbers N and accessible coordinate positions x, y and z. Consider the air particles in an ordinary classroom. These particles are, by chance, continuously transitioning from one configuration microstate to another among a great number possible. At any instant there is a small but finite probability that all particles, by chance, will become clumped within a small volume in the room, representing a low-probability microstate. The reason we do not see this configuration occur is not because it is not possible. Rather, this configuration and others like it are highly improbable. Most possible microstates are virtually identical, and are essentially uniform in the configuration of particle locations. Collectively these microstates are highly probable, and this is what we experience from one instant to the next. For fun, we can use the binomial distribution to estimate the likelihood of "observing" all air molecules within a small volume at some instant. This involves N trials with $k = N$ successes with the probability of a success equal to $p = 1/m$. Partitioning a classroom of volume $\sim 1\,000$ m^3 into cubic meter volumes gives $p \sim 10^{-3}$. With 10^{25} molecules per cubic meter, $N \sim 10^{28}$. So assuming a grand amount of time to wait for it (here we are neglecting the time element of the calculation), the probability of observing all air molecules within a cubic meter is $\sim 0.001^{10^{28}}$. I think we can take our chances that we will not suddenly be within a vacuum while listening to a lecture.

Consider a counterfactual. Suppose our classroom contains a relatively small number of air molecules representing rarefied conditions in the outer atmosphere of Earth. With $N = 100$, the probability of observing all of them at some instant within a cubic meter is now $\sim 0.001^{100}$. This is still a small number, but not as unimaginably small as the example above. On the other hand, and more importantly, the probability is ~ 0.905 that any cubic meter contains no molecules at any instant, and the probability is ~ 0.091 that any cubic meter contains one molecule at any instant. Because the 100 molecules must be moving among the 1 000 cubic meter partitions, this means we would certainly perceive the *fluctuations* in the particle configurations (microstates) from one instant to the next. This is at the heart of the matter in considering the small numbers of sediment particles involved in rarefied (non-continuum) transport conditions. That said, we now turn to particle energies, which was the focus of Boltzmann's efforts in this problem.

3.2 Energy States

Again consider a system containing a fixed number N of particles. We now imagine that these particles can be arranged in a great number of different energy states, $\epsilon_1, \epsilon_2, \epsilon_3, ...$, and we let $n_1, n_2, n_3, ...$ denote the numbers of particles in these states. As with particle configurations, there are a great number of ways to arrange N particles into the different energy states, and we want to know the most likely arrangement. In contrast to particle configurations, here we impose an additional constraint on the possible arrangements. Namely, we insist that the total energy E of the N particles is fixed. By fixing the system volume, the number of particles N and the total energy E of the particles, we are considering what is referred to as a *microcanonical ensemble* of possible microstates. In the language of thermodynamics, this coincides with an isolated system with fixed mass, pressure and temperature.

Recall that classical particles are distinguishable with respect to coordinate position and velocity, and thus energy. Absent a gravitational field, particle velocities in a microcanonical ensemble are independent of coordinate positions, so here we may focus just on energy. Using our example from above, consider all possible arrangements of $N = 4$ particles in different energy states $\epsilon_1, \epsilon_2, \epsilon_3, ..., \epsilon_m$ such that the total energy of each arrangement is $E = 10$ (Figure 4). For simplicity of illustration we are setting the energy states to integer values: $\epsilon_1 = 1, \epsilon_2 = 2, \epsilon_3 = 3, ...$ Each of these nine arrangements is a macrostate. In turn, for distinguishable particles we can specify the number of microstates associated with each macrostate using (8), where now the energy ϵ_m can be no larger than the total energy E, which must be much smaller than the energy associated with the speed of light. As before, in our example a macrostate with all four particles in one energy state coincides with a microstate.

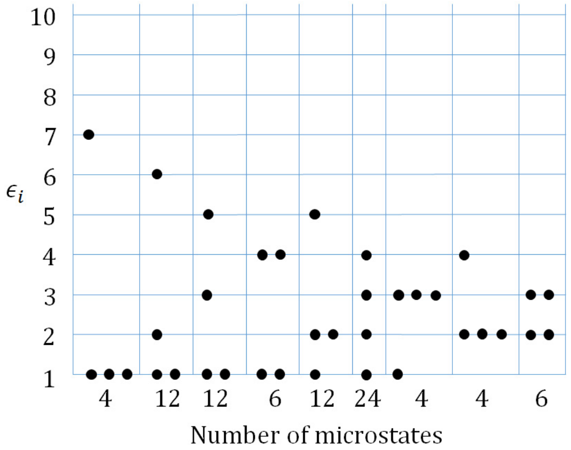

Figure 4: Schematic illustration of nine macrostates involving four particles together with the number of microstates in each macrostate. The total energy $E = 10$.

Here we make a key observation. Suppose that we had not constrained the total energy of the particles, but instead allowed them to randomly occupy all energy states up to and including $E = 10$. According the (7) the number of possible macrostates would be $N_\mathrm{e} = 715$, and as a consequence there would be a great number of possible microstates. Thus, because each macrostate is constrained to possess a total energy $E = 10$, we are now focused on a subset of the imagined possible (unconstrained) macrostates, namely, those in which the sum of the particle energies is fixed. Concomitantly, the total number of microstates is reduced, and we may refer to these as *accessible* microstates.

With reference to Figure 4, all nine macrostates are possible representations of the system for the given constraints, and in turn, each of the 84 microstates is a possible representation of the system. We again appeal to the foundational assumption of classical statistical mechanics, that all microstates are independent and equally likely. Thus each microstate has a probability equal to 1/84 of occurring. As with particle configurations we calculate the expected proportions of particles in each energy state by averaging over the 84 microstates, and we discover that the distribution $p_\epsilon(\epsilon_i)$ of energy states ϵ_i is decidedly not uniform (Figure 5). In fact, this bounded exponential-like distribution provides a glimpse of the Boltzmann distribution, which comes next.

Closely following the development above we again have

$$\frac{\ln n_\mathrm{e}}{N} = -\sum_i p_i \ln p_i \,, \tag{20}$$

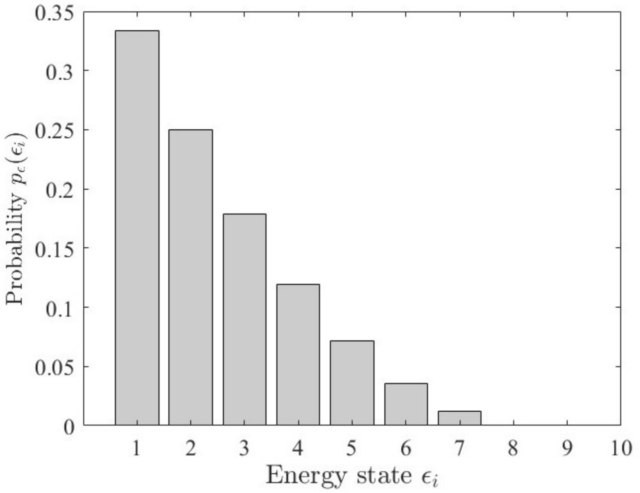

Figure 5: Probability distribution $p_\epsilon(\epsilon_i)$ of energy states ϵ_i based on the 84 microstates involving four particles in Figure 4.

where now n_e is the number of energy microstates in a macrostate and $p_i = p_\epsilon(\epsilon_i)$ is the probability mass function of the particle energy states ϵ_i. As before we want to know the set of p_i that gives the maximum number of microstates n_e. We again denote $f(p_i) = \ln n_\mathrm{e}/N$ so that

$$f(p_i) = -\sum_i p_i \ln p_i \,. \tag{21}$$

To constrain the problem we again specify that the probabilities sum to unity,

$$\sum_i p_i = 1 \,, \tag{22}$$

and then add a second constraint, that with fixed N and E the distribution of energy states has finite mean $\mu_\epsilon = E/N$,

$$\sum_i \epsilon_i p_i = \mu_\epsilon \,. \tag{23}$$

We now write

$$f(p_i) = -\sum_i p_i \ln p_i + \alpha\left(\sum_i p_i - 1\right) + \lambda\left(\sum_i \epsilon_i p_i - \mu_\epsilon\right) \,, \tag{24}$$

with Lagrange multipliers α and λ. We again take derivatives $\partial f/\partial p_j$ over the system of equations to give

$$\ln p_j + 1 - \alpha - \lambda p_j = 0 \,. \tag{25}$$

Thus, the set of p_j (the macrostate) representing the maximum number of microstates is

$$p_j = e^{\alpha - 1} e^{\lambda \epsilon_j} \,, \tag{26}$$

which represents the Boltzmann distribution.

Based on separate arguments the Lagrange multiplier $\lambda = -1/k_{\mathrm{B}}T$ where T is absolute temperature. The quantity $e^{\alpha-1}$ is a constant, and we treat it as a normalization factor. That is, because the distribution p_j must sum to unity we now write (26) as

$$p_j = \frac{e^{-\epsilon_j/k_{\mathrm{B}}T}}{\sum_{j=1}^{m} e^{-\epsilon_j/k_{\mathrm{B}}T}} \, . \tag{27}$$

The denominator in (27) is referred to as a *partition function.* In addition, if $\langle N_j \rangle$ denotes the expected number of particles in the jth energy state, then $\langle N_j \rangle = Np_j$ so that

$$\frac{\langle N_j \rangle}{N} = \frac{e^{-\epsilon_j/k_{\mathrm{B}}T}}{\sum_{j=1}^{m} e^{-\epsilon_j/k_{\mathrm{B}}T}} \, . \tag{28}$$

The Boltzmann distribution is the starting point for deriving the Maxwell–Boltzmann distributions of particle speeds, velocities, momenta and kinetic energies for ideal gases at thermal equilibrium. In addition the Boltzmann distribution can be generalized in essentially the same form to the case of a canonical ensemble in which the number of particles, the system volume and the temperature are fixed, and to the case of a grand canonical ensemble in which the system volume and the temperature are fixed but the system can exchange particles with its surroundings (e.g. Tolman, 1938).

Consider the justification for choosing the macrostate with the most microstates versus attempting to average over all microstates. Substituting our notation in the comments of Edwin Schrödinger (1952, p. 6) on this matter,

> ...all [micro]states of the assembly are embraced — without overlapping — by the classes [macrostates] described by all different admissible sets of numbers n_i... The number of single [micro]states, belonging to this class [macrostate], is obviously (00)... The present method [of the most probable distribution] admits that, on account of the enormous largeness of the number N, the total number of distributions (i.e., the sum of all n_{e}'s) is very nearly exhausted by the sum of those n_{e}'s whose number sets n_i do not deviate appreciably from the set which gives n_{e} its maximum value (among those, of course, which comply with (00)). In other words, if we regard this set of occupation numbers as obtaining always, we disregard only a very small fraction of all possible distributions — and this has "a vanishing likelihood of ever being realized".

This points to our earlier assertion that most microstates appear similar to those in the one macrostate having the most microstates. That is, one must not imagine that a system is constrained just to the set of microstates in the one macrostate identified by the maximization method. There are a great number of macrostates that are virtually (statistically) identical to the one macrostate identified, and as a consequence there are a great number of virtually identical microstates. Thus, the system continuously fluctuates among a great number of possible microstates, but the distribution of energy states does not deviate appreciably from the *expected* distribution. This likewise means that conditions described by the Maxwell–Boltzmann distributions of particle speeds, velocities and kinetic energies continuously fluctuate, but these fluctuations do not appreciably vary from the expected conditions. Numerical simulations of manageably small numbers of classical particles moving about in an isolated box nicely illustrate this point, and are readily available on numerous websites. Conversely, because the numbers are relatively small, the fluctuations are readily apparent in such simulations. To complete the story of a Gibbs ensemble, we summarize in Appendix A the idea of a phase space as used in classical statistical mechanics.

4 Maximum Entropy Distributions

With respect to sediment transport research, entropy really is about uncertainty, and about being as faithful to our uncertainty as we are to what we think we know — notably in view of the necessary coarsening of physics to the sediment particle scale and larger. The celebrated physicist and mathematical probabilist Edwin Jaynes championed these points — that entropy and uncertainty are synonymous, and that an important element of our intellectual honesty in doing science resides in formalizing this uncertainty. For example, in reference to using a criterion of maximum entropy in selecting a probability distribution Jaynes suggests:

> The maximum entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information [p. 623]. [T]he maximization of entropy is not an application of a law of physics, but merely a method of reasoning which ensures that no unconscious arbitrary assumptions have been introduced [p. 630]. (Jaynes, 1957a)

Jaynes (1957a, 1957b) elaborated the significance of the fact that the Gibbs entropy in statistical mechanics and the Shannon entropy in information theory are essentially one and the same, differing only by a constant. This similarity inspired Jaynes to champion the use of a maximum entropy criterion in choosing a probability distribution, leading to what is now known as the maximum entropy method (aka MaxEnt or MEM). The key idea of the maximum entropy method, whether viewed as a method of statistical mechanics or as one of inferential statistics, is that it provides an unbiased choice of a distribution by honoring only what is known mechanically about a system. That is, this unbiased choice is a maximally noncommittal choice that is faithful to what we do not know; it is therefore the most reasonable choice in the absence of additional information (Jaynes, 1957a; Williamson, 2010, pp. 25 and 51). Importantly, mechanical constraints imposed on the system are part of the choice of the distribution, as opposed to empirical fitting without regard to such constraints. The maximum entropy method has been applied in a remarkable variety of fields (Shore and Johnson, 1980; Ramirez and Carta, 2006; Verkley and Lynch, 2009; Singh, 2011; Peterson et al., 2013; Golan and Harte, 2022), including sediment transport (Furbish and Schmeeckle, 2013; Furbish et al., 2016, 2021b).

In using the maximum entropy method, constraints imposed on the system normally translate to constraints imposed on the moments of the distribution. In this case the method leads to a distribution that is among the exponential family (e.g. exponential, Gaussian). However, applications of the maximum entropy method to non-exponential distributions, including heavy-tailed distributions, are of particular interest in many problems (Peterson et al., 2013). Applying this method to heavy-tailed distributions presents a special challenge (Furbish et al., 2021b) in that the first or second moment, or both of these moments, may be undefined for such distributions. Hereafter we focus on the differential entropy.

If x denotes a continuous random variable distributed as $f_x(x)$ with support $x \in [0, \infty)$, then the differential entropy of x is conventionally defined as

$$H(x) = -\int_0^\infty f_x(x) \ln f_x(x) \, \mathrm{d}x \, , \tag{29}$$

where it is understood that $f_x(x) \ln f_x(x) = 0$ when $f_x(x) = 0$. In turn, let $g_j(x)$ denote a measurable quantity of x with $j = 0, 1, 2, ..., n$. We then assume that

$$\mathrm{E}\left[g_j(x)\right] = \int_0^\infty g_j(x) f_x(x) \, \mathrm{d}x = a_j \, , \tag{30}$$

with finite a_j. For example, if $g_0(x) = g_0 = 1$, then (30) gives $a_0 = 1$. That is, the density $f_x(x)$ integrates to unity. If $g_1(x) = x$, then (30) gives the mean of the distribution, $a_1 = \mu_x$. If $g_2(x) = (x - \mu_x)^2$, then (30) gives the variance, $a_2 = \sigma_x^2$. Note, however, that $g_j(x)$ need not be selected just to obtain the usual moments of a distribution. Indeed, (30) may represent a constraint imposed by a function $g_j(x)$ that does not coincide with a moment of $f_x(x)$. This is essential for heavy-tailed distributions whose first or second moment, or both of these moments, is undefined (Peterson et al., 2013; Furbish et al., 2016c). Whether obtained from the method of Lagrange multipliers (as in Section 3 above) or from a variational method, the maximum entropy distribution is then given by

$$f_x(x) = \exp\left[\sum_{j=0}^{n} \lambda_j g_j(x)\right], \tag{31}$$

where $\lambda_0, \lambda_1, \lambda_2, ...$ are Lagrange multipliers introduced in the problem of maximizing the entropy $H(x)$. Moreover, we set $g_0(x) = g_0 = 1$ with $a_0 = 1$, which guarantees that the probability density $f_x(x)$ integrates to unity.

As a point of reference, a fixed mean with $g_1(x) = x$ and no other constraint leads to the result

$$f_x(x) = e^{\lambda_0} e^{\lambda_1 x}\,. \tag{32}$$

The Lagrange multipliers are then obtained as follows. By the definition of a probability density,

$$e^{\lambda_0} \int_0^\infty e^{\lambda_1 x}\,\mathrm{d}x = 1\,, \tag{33}$$

which leads to $e^{\lambda_0} = -\lambda_1$. Alternatively, (33) may be written as

$$f_x(x) = \frac{e^{\lambda_1 x}}{\int_0^\infty e^{\lambda_1 x}\,\mathrm{d}x}\,, \tag{34}$$

where it becomes clear that e^{λ_0} is a normalization factor that ensures the probability density integrates to unity. In turn, by the definition of the mean,

$$-\lambda_1 \int_0^\infty x e^{\lambda_1 x}\,\mathrm{d}x = \mu_x\,, \tag{35}$$

which leads to $\lambda_1 = -1/\mu_x$ and the exponential distribution,

$$f_x(x) = \frac{1}{\mu_x} e^{-x/\mu_x}\,, \tag{36}$$

where it becomes clear that the Lagrange multiplier λ_1 enforces the constraint of a fixed mean. With support $x \in \mathbb{R}$ the Gaussian distribution is similarly obtained as the maximum entropy distribution with the additional constraint imposed by a fixed second moment (variance).

We must add caveats regarding the interpretation of the differential entropy. First, if the random variable x has dimensions, then the probability density $f_x(x)$ has dimensions and the differential entropy given by (29) is not dimensionally sound. This means that x must be dimensionless or recast in dimensionless form, or care must be given to ensuring that dimensions are preserved in manipulating (29). For example, notice that the expression for the maximum entropy distribution given by (31) is dimensionally sound so long as the product $\lambda_j g_j(x)$ is dimensionless. This occurs because the maximization procedure removes the logarithm of $f_x(x)$.

Second, the Shannon entropy is formally restricted to discrete random variables and is a positive quantity. The differential entropy therefore must be viewed as an analogy with the Shannon entropy.[4] Recall that the definite integral of a continuous function can be obtained from the limit of a Riemann sum as

$$\int_{-\infty}^{\infty} f_x(x)\,\mathrm{d}x = \lim_{\Delta x \to 0} \sum_{i=-\infty}^{\infty} f_x(x_i)\Delta x\,. \tag{37}$$

By interpreting $f_x(x_i)\Delta x = p_i$ as the discrete probability associated with the small interval Δx, we now write a Shannon-like entropy as

$$H^*(x) = -\sum_{i=-\infty}^{\infty} f_x(x_i)\Delta x \ln[f_x(x_i)\Delta x]\,.$$

This is dimensionally sound and may be interpreted as a Shannon entropy so long as the interval Δx is specified. In effect it represents a discretization of the probability density $f_x(x)$. However, expanding the logarithm gives

$$\begin{aligned} H^*(x) &= -\sum_{i=-\infty}^{\infty} f_x(x_i)\Delta x[\ln f_x(x_i) + \ln(\Delta x)] \\ &= -\sum_{i=-\infty}^{\infty} f_x(x_i)\Delta x \ln f_x(x_i) - \sum_{i=-\infty}^{\infty} f_x(x_i)\Delta x \ln(\Delta x)\,. \end{aligned} \tag{38}$$

Taking the limit as $\Delta x \to 0$ then leads to

$$H^*(x) = -\int_{-\infty}^{\infty} f_x(x) \ln f_x(x)\,\mathrm{d}x + \infty\,, \tag{39}$$

in that $\ln(\Delta x) \to -\infty$ as $\Delta x \to 0$. That is, the differential entropy is not a limit of the Shannon entropy as the interval Δx goes to zero in a Riemann sense. The differential entropy is offset from the Shannon entropy by an infinite amount, and it may be defined as $H(x) = \lim_{\Delta x \to 0} H^*[x + \ln(\Delta x)]$. Jaynes (1963) addresses these points and offers a modified definition of the differential entropy that views it as the limit of an increasingly dense discrete distribution, analogous to a Riemann sum. Operationally the maximum entropy method reduces to the result given by (31).

As a point of reference, using (29) the differential entropy of the exponential distribution is

$$H(x) = 1 - \ln\left(\frac{1}{\mu_x}\right) = 1 + \ln \mu_x\,. \tag{40}$$

This illustrates that the differential entropy cannot be interpreted in terms of the Shannon information content, as $H(x)$ becomes negative with small values of the mean μ_x. Nonetheless, this decrease corresponds with decreasing uncertainty; the mode of the distribution at $x = 0$ sharpens with decreasing μ_x such that most values of x are close to the origin. Conversely, as $H(x)$ increases

[4] Shannon presented the differential entropy as an analogue of the discrete entropy but did not elaborate the relation between the two.

with increasing μ_x, uncertainty in the occurrence of any value x increases with flattening of the distribution to a more uniform condition.

The maximum entropy method has been used to describe the energetics of rarefied particle motions on hillslopes (Furbish et al., 2021b) as well as bed load particle velocities (Furbish and Schmeeckle, 2013; Furbish et al, 2016) to suggest how the probability distributions of these motions reflect mechanical constraints. This work is focused on the idea that mechanical constraints imposed on the system are part of the choice of the distribution, as opposed to empirical fitting without regard to such constraints. With respect to rarefied particle motions on hillslopes, the analysis is satisfying in that it reveals how the transition from a light-tailed form of the distribution of travel distances to a heavy-tailed form involves no real change in the physics of particle–surface interactions, thus reinforcing the physical (versus mathematical) interpretation of nonlocal transport (Furbish and Haff, 2010; Doane, 2018; Doane et al., 2018; Furbish et al., 2021a, 2021b). However, with respect to bed load particle motions the analysis is less satisfying in that we have information about the dynamics of these motions that cannot (yet) be readily cast as a mechanical constraint in the maximum entropy method.

A Phase Space

To complete the story of a Gibbs ensemble, consider Figure 2, Figure 3 and Figure 4. These give a static view of particles in different states. As described above, however, individual particles are continuously moving among accessible coordinate and energy states such that the system is continuously transitioning from one microstate to another among a great number possible — an ensemble of microstates — subject to imposed constraints. This leads to the idea of a phase space.

We need to appeal to generalized coordinates and velocities versus our familiar view of things. Starting with the familiar view, the instantaneous state of a classical particle is defined by its position and velocity. We are then accustomed to specifying the state of a particle using three position coordinates and three velocity components, representing six degrees of freedom. A system consisting of N particles thus involves $6N$ degrees of freedom. Letting subscripts $1, 2, 3, ...$ denote individual particles, then their coordinate states are normally specified as a set of tuples: (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3), ..., (x_N, y_N, z_N). Similarly, particle velocities are specified as a set: (u_1, v_1, w_1), (u_2, v_2, w_2), (u_3, v_3, w_3), ..., (u_N, v_N, w_N). The instantaneous state of a system is then specified by a set of N points located in the spaces defined by three coordinate axes and three velocity axes.

Let us now instead choose generalized coordinates and velocities denoted by q and $\dot{q} = \mathrm{d}q/\mathrm{d}t$. These coordinates and velocities are identified with individual particles, but not in the sense above. Rather, they refer to degrees of freedom, so now the subscripts $1, 2, 3, ...$ refer to these degrees of freedom rather than the particles per se. We thus map $x_1 = q_1$, $y_1 = q_2$, $z_1 = q_3$ and $x_2 = q_4$, $y_2 = q_5$, $z_2 = q_6$, then so on to $x_N = q_{3N-2}$, $y_N = q_{3N-1}$, $z_N = q_{3N}$. Similarly with velocities, $u_1 = \dot{q}_1$, $v_1 = \dot{q}_2$, $w_1 = \dot{q}_3$ and $u_2 = \dot{q}_4$, $v_2 = \dot{q}_5$, $w_2 = \dot{q}_6$, then so on to $u_N = \dot{q}_{3N-2}$, $v_N = \dot{q}_{3N-1}$, $w_N = \dot{q}_{3N}$. We then imagine a $6N$-dimensional phase space (a hyperspace) whose axes consist of the $6N$ degrees of freedom. In contrast to the previous specification of the state of a system involving N points in three-dimensional coordinate and velocity spaces, the state of a system involving N particles is completely specified by a *single* point in the $6N$-dimensional phase space. Moreover, with distinguishable particles this point in the phase space coincides with a microstate of the system such that its *phase trajectory* reflects movement of the system among accessible microstates. In addition, the states of an ensemble of systems consist of a set of points in this phase space, each moving along a phase trajectory.

As impossible as it is to envision a $6N$-dimensional phase space, we can actually gain the essence of what this is point of view is describing by examining a low-dimensional system. Following Kittel (1958), let us imagine a two-particle system where each particle is free to move with varying velocities in one dimension parallel to x. Thus each particle has two degrees of freedom and the system has $2N = 4$ degrees of freedom. Assuming positions and velocities are independent, we focus on velocities. We now have $u_1 = \dot{q}_1$ and $u_2 = \dot{q}_2$. Assuming microcanical conditions the total kinetic energy E of the system is fixed, so

$$\dot{q}_1^2 + \dot{q}_2^2 = C \, , \tag{41}$$

where $C = 2E/m$. This is just the equation of a circle in the $\dot{q}_1\dot{q}_2$ phase space. Thus, for a given total kinetic energy within the infinitesimal interval $\mathrm{d}E$ at E, the probability distribution $f_{\dot{q}_1,\dot{q}_2}(\dot{q}_1, \dot{q}_2)$ of velocities $\dot{q}_1$ and $\dot{q}_2$ looks like a thin cylinder with radius $(2E/m)^{1/2}$ in the $\dot{q}_1\dot{q}_2$ phase space. Because each microstate is equally probable, and because we have no information regarding initial states, the distribution is uniform over the cylinder. If instead we consider a three-particle system such that $\dot{q}_1^2 + \dot{q}_2^2 + \dot{q}_3^2 = C$, then the probability distribution of the velocities looks like a spherical shell in the $\dot{q}_1\dot{q}_2\dot{q}_3$ phase space.

References

[1] Doane, T. H. (2018) Theory and application of nonlocal hillslope sediment transport. PhD thesis, Vanderbilt University, Nashville, Tennessee.

[2] Doane, T. H., Furbish, D. J., Roering, J. J., Schumer, R., and Morgan, D. J. (2018) Nonlocal sediment transport on steep lateral moraines, eastern Sierra Nevada, California, USA. *Journal of Geophysical Research – Earth Surface*, 123, 187–208, https://doi.org/10.1002/2017JF004325.

[3] Furbish, D. J. and Haff, P. K. (2010) From divots to swales: Hillslope sediment transport across divers length scales. *Journal of Geophysical Research – Earth Surface*, 115, F03001, https://doi.org/10.1029/2009JF001576.

[4] Furbish, D. J. and Schmeeckle, M. W. (2013) A probabilistic derivation of the exponential-like distribution of bed load particle velocities. *Water Resources Research*, 49, 1537–1551, https://doi.org/10.1002/wrcr.20074.

[5] Furbish, D. J., Schmeeckle, M. W., Schumer, R., and Fathel, S. L. (2016) Probability distributions of bed load particle velocities, accelerations, hop distances, and travel times informed by Jaynes's principle of maximum entropy. *Journal of Geophysical Research – Earth Surface*, 121, 1373–1390, https://doi.org/10.1002/2016JF003833.

[6] Furbish, D. J., Roering, J. J., Doane, T. H., Roth, D. L., Williams, S. G. W., and Abbott, A. M. (2021a) Rarefied particle motions on hillslopes – Part 1: Theory. *Earth Surface Dynamics*, 9, 539–576, https://doi.org/10.5194/esurf-9-539-2021.

[7] Furbish, D. J., Williams, S. G.W., and Doane, T. H. (2021b) Rarefied particle motions on hillslopes – Part 3: Entropy. *Earth Surface Dynamics*, 9, 615–628, https://doi.org/10.5194/esurf-9-615-2021.

[8] Gibbs, J. W. (1902) *Elementary Principles in Statistical Mechanics.* Yale University Press, New Haven, Connecticut.

[9] Golan, A. and Harte, J. (2022) Information theory: A foundation for complexity science. *Proceedings of the National Academy of Sciences*, 119, e2119089119, https://doi.org/10.1073/pnas.2119089119.

[10] Jaynes, E. T. (1957a) Information theory and statistical mechanics. *Physical Review*, 106, 620–630.

[11] Jaynes, E. T. (1957b) Information theory and statistical mechanics. II. *Physical Review*, 108, 171–190.

[12] Jaynes, E. T. (1963). Information theory and statistical mechanics. In K. Ford (ed.), *Statistical Physics*, Benjamin, New York, p. 181.

[13] Jaynes, E. T. (2003) *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge.

[14] Kittel, C. (1958) *Elementary Statistical Physics.* John Wiley & Sons, New York.

[15] McIrvine, E. C. and Tribus, M. (1971) Energy and Information. *Scientific American*, 225, 179–188.

[16] Peterson, J., Dixit, P. D., and Dill, K. A. (2013) A maximum entropy framework for nonexponential distributions. *Proceedings of the National Academy of Sciences*, 110, 20380–20385.

[17] Ramirez, P. and Carta, J. A. (2006) The use of wind probability distributions derived from the maximum entropy principle in the analysis of wind energy. A case study. *Energy Conservation and Management*, 47, 2564–2577.

[18] Roseberry, J. C., Schmeeckle, M. W., and Furbish, D. J. (2012) A probabilistic description of the bed load sediment flux. 2. Particle activity and motions. *Journal of Geophysical Research – Earth Surface*, 117, F03032.

[19] Schrödinger, E. (1952) *Statistical Thermodynamics.* Cambridge University Press, Cambridge.

[20] Shannon, C. E. (1948a) A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423.

[21] Shannon, C. E. (1948b) A mathematical theory of communication. *Bell System Technical Journal*, 27, 623–656.

[22] Shore, J. and Johnson, R. (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26, 26–37.

[23] Singh, V. P. (2011) Derivation of power law and logarithmic velocity distributions using the Shannon entropy. *Journal of Hydrologic Engineering*, 16, 478–483, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000335.

[24] Tolman, R. C. (1938) *The Principles of Statistical Mechanics.* Oxford University Press, New York.

[25] Tsallis, C. (1988) Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52, 479–487.

[26] Verkley, W. T. M. and Lynch, P. (2009) Energy and enstrophy spectra of geostrophic turbulent flows derived from a maximum entropy principle. *Journal of Atmospheric Sciences*, 66, 2216–2236, https://doi.org/10.1175/2009JAS2889.1.

[27] Wallace, D. (2012) *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation.* Oxford University Press, Oxford.

[28] Williamson, J. (2010) *In Defense of Objective Bayesianism.* Oxford University Press, Oxford, UK.