Hindawi
Complexity
Volume 2017, Article ID 8715605, 10 pages
https://doi.org/10.1155/2017/8715605

Research Article

Advancing Shannon Entropy for Measuring Diversity in Systems

R. Rajaram,[1] B. Castellani,[2] and A. N. Wilson[3]

[1]*Department of Mathematical Sciences, Kent State University, Kent, OH, USA*
[2]*Department of Sociology, Kent State University, 3300 Lake Rd. West, Ashtabula, OH, USA*
[3]*School of Social and Health Sciences, Abertay University, Dundee DD1 1HG, UK*

Correspondence should be addressed to R. Rajaram; rrajaram@kent.edu

Received 31 January 2017; Revised 5 April 2017; Accepted 23 April 2017; Published 24 May 2017

Academic Editor: Enzo Pasquale Scilingo

From economic inequality and species diversity to power laws and the analysis of multiple trends and trajectories, diversity within systems is a major issue for science. Part of the challenge is measuring it. Shannon entropy H has been used to rethink diversity within probability distributions, based on the notion of information. However, there are two major limitations to Shannon's approach. First, it cannot be used to compare diversity distributions that have different levels of scale. Second, it cannot be used to compare parts of diversity distributions to the whole. To address these limitations, we introduce a renormalization of probability distributions based on the notion of *case-based entropy* C_c as a function of the cumulative probability c. Given a probability density $p(x)$, C_c measures the diversity of the distribution up to a cumulative probability of c, by computing the length or support of an equivalent uniform distribution that has the same Shannon information as the conditional distribution of $\widehat{p}_c(x)$ up to cumulative probability c. We illustrate the utility of our approach by renormalizing and comparing three well-known energy distributions in physics, namely, the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac distributions for energy of subatomic particles. The comparison shows that C_c is a vast improvement over H as it provides a scale-free comparison of these diversity distributions and also allows for a comparison between parts of these diversity distributions.

1. Diversity in Systems

Statistical distributions play an important role in any branch of science that studies systems comprised of many similar or identical particles, objects, or actors, whether material or immaterial, human or nonhuman. One of the key features that determines the characteristics and range of potential behaviors of such systems is the degree and distribution of diversity, that is, the extent to which the components of the system occupy states with similar or different features.

As Page outlined in a series of inquiries [1, 2], including *The Difference* and *Diversity and Complexity*, diversity within systems is an important concern for science, be it making sense of economic inequality, expanding the trade portfolio of countries, measuring the collapse of species diversity in various ecosystems, or determining the optimal utility/robustness of a network. However, an important major challenge in the literature on diversity and complexity, which Page also points out [1, 2], remains: the issue of measurement. Although statistical distributions that directly reflect the spread of key parameters (such as mass, age, wealth, or energy) provide descriptions of this diversity, it can be difficult to compare the diversity of different distributions or even the same distribution under different conditions, mostly because of differences in scales and parameters. Also, many of the measures currently available compress diversity into a single score or are not intuitive [1–4].

At the outset, motivated by examples of measuring diversity in ecology and evolutionary biology from [3, 4], we sought to address these challenges. We begin with some definitions and a review of our previous research.

First, in terms of definitions, we follow the ecological literature, defining *diversity* as the interplay of "richness" and "evenness" in a probability distribution. *Richness* refers to the number of different diversity types in a system. Examples include (a) the different levels of household income in a city, (b) the number of different species in an ecosystem, (c) the diversity of a country's exports, (d) the distribution of

different nodes in a complex network, (e) the various health trends for a particular disease across time/space, or (f) the cultural or ethnic diversity of an organization or company. In all such instances, the greater the number of diversity types (be these types discrete or continuous), the greater the degree of richness in a system. In the case of the current study, for example, *richness* was defined as the number of different energy states.

In turn, *evenness* refers to the uniformity or "equiprobability" of occurrence of such states. In terms of the above examples, *evenness* would be defined as (a) a city where household income was evenly distributed, (b) an ecosystem where the diversity of its species was equal in number, (c) a country with an even distribution of exports, (d) a complex network where all nodes had the same probability of occurrence, (e) a disease where all possible health trends were equiprobable, or (f) a company or organization where people of different cultural or ethnic backgrounds were evenly distributed. In the case of the current study, for example, *evenness* was defined as the uniformity or "equiprobability" of the occurrence of all possible energy states.

More specifically, as we will see later in the paper, we define the diversity of a probability distribution as the number of equivalent equiprobable types required to maintain the same amount of Shannon entropy H (i.e., the number of Shannon-equivalent equiprobable states). Given such a definition, a system with a high degree of richness and evenness would have a higher degree of H, whereas a system with a low degree of richness and evenness would have a low degree of H. In turn, a system with high richness but low evenness (as in the case of a skewed-right system with long tail) would have a lower degree of H than a system with high richness and high evenness.

1.1. Purpose of the Current Study. Recently, we have introduced a novel approach to representing diversity within statistical distributions [5, 6], which overcomes such difficulties and allows the distribution of diversity in any given system (or cumulative portions thereof) to be directly compared to the distribution of diversity within any other system. In effect, it is a *renormalization* that can be applied to any probability distribution to produce a direct representation of the distribution of diversity within that distribution. Arising from our work in the area of complex systems, the approach is based on the notion of *case-based entropy*, C_c [5]. This approach has two major advantages over the Shannon Entropy H, which, as we alluded to above, is one of the most commonly used measures of diversity within probability distributions and which calculates the average amount of uncertainty (or information, depending on one's perspective) present in a given probability distribution. First, C_c can be used to compare distributions that have different levels of scale; and, second, C_c can be used to compare parts of distributions to their whole.

After developing the concept and formalism for case-based entropy for discrete distributions [5], we first applied it to compare complexity across a range of complex systems [6]. In that work, we investigated a series of systems described by a variety of skewed-right probability distributions, choosing examples that are often suggested to exhibit behaviors indicative of complexity such as emergent collectivity, phase changes, or tipping points. What we found was that such systems obeyed an apparent "limiting law of restricted diversity" [6], which constrains the majority of cases in these complex systems to simpler types. In fact, for these types of distribution, the distributions of diversity were found to follow a scale-free 60/40 rule, with 60% or more of cases belonging to the simplest 40% or less of equiprobable diversity types. This was found to be the case regardless of whether the original distribution fit a power law or was long-tailed, making it fundamentally distinct from the well-known (but often misunderstood) Pareto Principle [7].

In the following, we continue to explore the use of case-based entropy in comparing systems described by statistical distributions. However, we now go beyond our prior work in the following ways. First, we extend the formalism in order to compute case-based entropy for continuous as well as discrete distributions. Second, we broaden our focus from complexity/complex systems to diversity in *any* type of statistically distributed system. That is, we start to explore distributions of diversity for systems where richness is not a function of the degree of complexity types.

Third, the discrete indices we used had a degree of subjectivity to them, for example, how should household income be binned and what influence does that have on the distribution of diversity? As such, we wanted to see how well C_c worked for distributions where the unit of measurement was universally agreed upon.

Fourth, we had not emphasized how C_c was a major advance on Shannon entropy H. As known, while H has proven useful, it compresses its measurement of diversity into a single number; it is also nonintuitive; and, as we stated above, it is not scale-free and therefore cannot be used to compare the diversity of different systems; neither can it be used to compare parts of the diversity within a system to the entire system.

Hence, the purpose of the current study, as a demonstration of the utility of C_c, is to renormalize and compare three physically significant energy distributions in statistical physics: the energy probability density functions for systems governed by Boltzmann, Bose-Einstein, and Fermi-Dirac statistics.

2. Renormalizing Probability: Case-Based Entropy and the Distribution of Diversity

The quantity *case-based entropy* [5], C_c, renormalizes the diversity contribution of any probability distribution $P(x)$, by computing the true diversity D of an equiprobable distribution (called the *Shannon-equivalent uniform distribution*) that has the same Shannon entropy H as $P(x)$. C_c is precisely the number of equiprobable types in the case of a discrete distribution, or the length, support, or extent of the variable in the case of continuous distributions, which is required to keep the value of the Shannon entropy the same across the whole or any part of the distribution up to a cumulative probability

c. We choose the Shannon-equivalent uniform distribution for two reasons:

(i) First, it is well known that, on a finite measure space, the uniform distribution maximizes entropy: that is, the uniform distribution has the maximal entropy among all probability distributions on a set of finite Lebesgue measures [8].

(ii) Second, a Shannon-equivalent uniform distribution will, by definition, count the number of values (or range of values) of x that are required to give the same information as the original distribution $P(x)$ if we assume that all the values (or range of values) are equally probable.

Hence, the uniform distribution renormalizes the effect of varying relative frequencies (or probabilities) of occurrence of the values of x without losing information (or entropy). In other words, if all choices of the random variable are equally likely, the number of values (or the length, if it is a continuous random variable) needed for the random variable to keep the same amount of information as the given distribution is a measure of diversity. In a sense, each new value (or type) is counted as adding to the diversity, only if the new value has the same probability of occurrence as the existing values. Diversity necessarily requires the values of the random variable to be equiprobable since lower probability, for example, means that such values occur rarely in the random variable and hence cannot be treated as equally diverse as other values with higher probabilities. Hence, by choosing an equiprobable (or uniform) distribution for normalization, we are counting the true diversity, that is, the number of equiprobable types that are required to match the same amount of Shannon information H as the given distribution.

This calculation (as we have shown elsewhere [5]) can be done for parts of the distribution up to a cumulative probability of c. This means that a comparison of C_c for a variety of distributions is actually a comparison of the variation of the fraction of diversity C_c contributed by values of the random variable up to c.

Since, regardless of the scale and units of the original distribution, c and C_c both vary from 0 to 1, one can plot a curve for C_c versus c for multiple distributions on the same axes. C_c thus provides us with a scale-free measure to compare distributions without omitting any of the entropy information, but by renormalizing the variable to one that has equiprobable values. What is more, it also allows us to compare different parts of the same distribution, or parts to wholes. That is, we can generate a C_c versus c curve for any part of a distribution (normalizing the probabilities to add up to 1 in that part) and compare the C_c curve of the part to the C_c curve of the whole or another part to see if the functional dependence of C_c on c is the same or different. In essence, C_c has the ability to compare distributions in a "fractal" or self-similar way.

In [5], we showed how to carry out the renormalization for discrete probability distributions, both mathematical and empirical. In this paper, as we stated in the Introduction, we make the case for how C_c constitutes an advance over H, in terms of providing a scale-free comparison of probability distributions and also comparisons between parts of distributions. More importantly, we demonstrate how C_c works for continuous distributions, by examining the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac distributions for energy of subatomic particles. We begin with a more detailed review of C_c.

3. Case-Based Entropy of a Continuous Random Variable

Our impetus for making an advance over the Shannon entropy H comes from the study of diversity in evolutionary biology and ecology, where it is employed to measure the true diversity of species (types) in a given ecological system of study [3, 4, 9, 10]. As we show here, it can also be used to measure the diversity of an arbitrary probability distribution of a continuous random variable.

Given the probability density function $p(x)$ of a random variable x in a measure space X, the Shannon-Weiner entropy index H is given by

$$H = -\int_X p(x) \ln\left(p(x)\right) dx. \tag{1}$$

The problem, however, with the Shannon entropy index H, as we identified in our abstract and Introduction, is that while being useful for studying the diversity of a single system, it cannot be used to compare the diversity across probability distributions. In other words, H is not multiplicative: a doubling of value for H does not mean that the actual diversity has doubled. To address this problem, we turned to the *true diversity* measure D [3, 11, 12], which gives the range of equiprobable values of x that gives the same value of H:

$$D = e^H. \tag{2}$$

The utility of D for comparing the diversity across probability distributions is that, in D, a doubling of the value means that the number of equiprobable ranges of values of x has doubled as well. D calculates the range of such equiprobable values of x that will give the same value of Shannon entropy H as observed in the distribution of x. We say that two probability densities $p_1(x)$ and $p_2(\mathrm{x})$ are Shannon-equivalent if they have the same value of Shannon entropy. Case-based entropy is then the range of values of x for the Shannon-equivalent uniform distribution for $p(x)$. We also note that Shannon entropy can be recomputed from D by using $H = \ln(D)$.

In order to measure the distribution of diversity, we next need to determine the fractional contribution to overall diversity up to a cumulative probability c. In other words, we need to be able to compute the diversity contribution D_c up to a certain cumulative probability c. To do so, we replace H with H_c, the conditional entropy, given that only the portion of the distribution up to a cumulative probability c (denoted by X_c) is observed with conditional probability of occurrence

with density $\widehat{p}_c(x)$ up to a given cumulative probability c. That is,

$$\begin{aligned}\widehat{p}_c(x) &= \frac{p(x)}{\int_{X_c} p(x)\,dx},\\ H_c &= -\int_{X_c} \widehat{p}_c(x)\ln\left(\widehat{p}_c(x)\right),\\ c &= \int_{X_c} p(x)\,dx,\\ D_c &= e^{H_c}.\end{aligned} \tag{3}$$

The value of D_c for a given value of cumulative probability c is the number of Shannon-equivalent equiprobable energy states (or of values of the variable in the x-axis in general) that are required to explain the information up to a cumulative probability of c within the distribution. If $c = 1$, then $D_c = D$ is the number of such Shannon-equivalent equiprobable energy states for the entire distribution itself.

We can then simply calculate the fractional diversity contribution or case-based entropy as

$$C_c = \frac{D_c}{D}. \tag{4}$$

It is at this point that the renormalization (C_c as a function of c) becomes scale independent as both axes range between values of 0 and 1 with the graph of C_c versus c passing through $(0,0)$ and $(1,1)$. Hence, irrespective of the range and scale of the original distributions, all distributions can be plotted on the same graph and their diversity contributions can be compared in a scale-free manner.

To check the validity of our formalism, we calculate D_c for the simple case of a uniform distribution given by $p(x) = \chi_{[0,L]}(x)$ on the interval $X = [0, L]$. Intuitively, if we choose $X_c = [0, c]$, then, owing to the uniformity of the distribution, we expect $D_c = c$ itself. In other words, the diversity of the part $[0, c]$ is simply equal to c, that is, the length of the interval $[0, c]$, and hence the C_c versus c curve will simply be the straight line with slope equal to 1. This can be shown as follows:

$$\begin{aligned}\widehat{p}_c(x) &= \frac{1}{c}\chi_{[0,L]}(x),\\ H_c &= -\int_{[0,c]} \frac{1}{c}\ln\left(\frac{1}{c}\right)dx = \ln(c),\\ D_c &= e^{H_c} = e^{\ln(c)} = c.\end{aligned} \tag{5}$$

With our formulation of C_c complete, we turn to the energy distributions for particles governed by Boltzmann, Bose-Einstein, and Fermi-Dirac statistics.

4. Results

4.1. C_c for the Boltzmann Distribution in One Dimension. We first illustrate our renormalization by applying it to a relatively simple case: that of an ideal gas at temperature T. The kinetic energies E of particles in such a gas are described by the Boltzmann distribution [8]. In one dimension, this is

$$p_{B,1D}(E) = \left(\frac{1}{k_B T}\right)e^{-E/k_B T} = \frac{\beta}{e^{\beta E}}, \tag{6}$$

where k_B is the Boltzmann constant and $\beta = (1/k_B T)$.

The entropy of $p_{B,1D}(E)$ can be shown to be $H_B = 1 - \ln(\beta)$, and hence the true diversity of energy in the range $[0, \infty)$ is given by

$$D_{B,1D} = e^H = e^{1-\ln(\beta)} = \frac{e}{\beta}. \tag{7}$$

The cumulative probability c from $E = 0$ to $E = k$ is then given by

$$c = \int_{[0,k]} p_{B,1D}(E)\,dE = 1 - e^{-\beta k}. \tag{8}$$

Hence, k can be computed in terms of c as

$$k = -\frac{\ln(1-c)}{\beta}. \tag{9}$$

Equation (9) is useful for the one-dimensional Boltzmann case to eliminate the parameter k altogether in (11) to obtain an explicit relationship between C_c and c. It is to be noted that, in most cases, both C_c and c can only be parametrically related through k. The other quantities introduced in Section 3 can then be calculated as follows:

$$\widehat{p}_c(E) = \frac{p_{B,1D}(E)}{c} = \frac{\beta e^{-\beta E}}{1 - e^{-\beta k}}, \tag{10}$$

$$\begin{aligned}H_c &= -\int_{[0,k]} \frac{\beta e^{-\beta E}}{1 - e^{-\beta k}}\ln\left(\frac{\beta e^{-\beta E}}{1 - e^{-\beta k}}\right)dE\\ &= 1 + \ln\left(\frac{c}{\beta}(1-c)^{(1-c)/c}\right),\end{aligned} \tag{11}$$

$$\begin{aligned}D_c &= e^{H_c} = e^{1+\ln((c/\beta)(1-c)^{(1-c)/c})}\\ &= \frac{e}{\beta}\cdot\left(c(1-c)^{(1-c)/c}\right),\end{aligned} \tag{12}$$

$$\begin{aligned}C_c &= \frac{D_c}{D_{B,1D}} = \frac{(e/\beta)\cdot\left(c(1-c)^{(1-c)/c}\right)}{e/\beta}\\ &= c(1-c)^{(1-c)/c}.\end{aligned} \tag{13}$$

We note that, in (13), the temperature factor β cancels out, indicating that the distribution of diversity for an ideal gas in one dimension is independent of temperature. The resulting graph of C_c as a function of c is shown in Figure 1. It is worth noting in passing that C_c reaches 40% when $c \approx 69\%$, indicating that approximately 69% of the molecules in the gas are contained within the lower 40% of *diversity* of energy probability states at all temperatures (here, *diversity* is defined as the number of equivalent equiprobable energy states required to maintain the same amount of

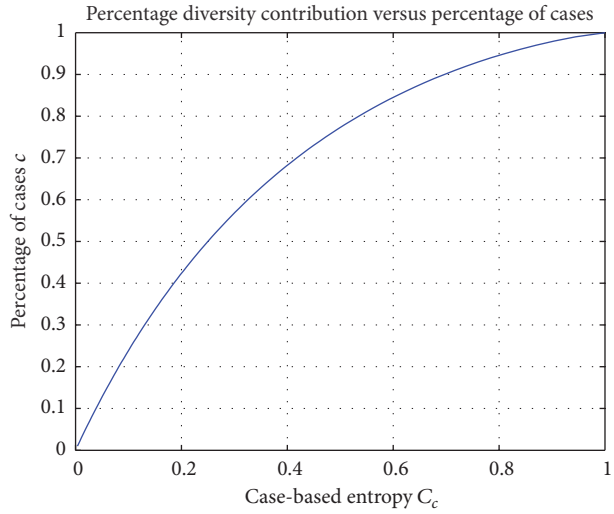


Figure 1: C_c as a function of c for the Boltzmann distribution in one dimension.

Shannon entropy H). Thus, the one-dimensional Boltzmann distribution obeys an interesting phenomenon that we have identified in a wide range of skewed-right complex systems, which (as we briefly discussed in the Introduction) we call *restricted diversity* and, more technically, the 60/40 rule [6]. The independence of temperature in the C_c versus c curve, for the Boltzmann distribution, shows that the effect of increasing T is to shift the mean of the distribution to higher energies and to increase its standard deviation, but not to change its characteristic shape. Still, what is key to our results is that the temperature independence of the C_c curve for the Boltzmann distribution in one dimension validates that our renormalization preserves the fundamental features of the original distribution.

4.2. C_c for the Boltzmann Distribution in Three Dimensions. We now turn to the calculation of C_c for the physically more important case of the Boltzmann distribution in three dimensions [8]:

$$p_{B,3D}(E) = \frac{2\beta^{3/2}E^{1/2}}{\sqrt{\pi}e^{\beta E}}, \tag{14}$$

where the additional factor of $\sqrt{4\beta E/\pi}$ accounts for the density of states.

The cumulative probability c from $E = 0$ to $E = k$ can be computed as follows:

$$c = \int_{[0,k]} p_{B,3D}(E)\, dE = \frac{\sqrt{\pi}\,\mathrm{erf}\left(\sqrt{k\beta}\right) - 2e^{-k\beta}\sqrt{k\beta}}{\sqrt{\pi}}. \tag{15}$$

As we would hope, (15) has the property that as $k \to \infty$, the cumulative probability $c \to 1$.

However, it is difficult to solve (15) for k directly in terms of c. We therefore compute C_c in parametric form with k being the parameter. Also, analytical forms are not possible, so Matlab was used to compute H_c, D_c, and C_c, respectively:

$$\begin{aligned} D_c(k) &= e^{H_c(k)}, \\ D_{B,3D} &= \lim_{k\to\infty} D_c(k), \\ C_c &= \frac{D_c}{D_{B,3D}}. \end{aligned} \tag{16}$$

Thus, C_c can also only be computed in parametric form with parameter k that varies from 0 to ∞. Figure 2 shows the C_c curve thus calculated for the Boltzmann distribution in three dimensions.

Although the temperature independence of this distribution is not immediately evident from Figure 2, one would, following the same logic as for the one-dimensional case, expect the *distribution* of diversity to be the same for all T. That is, as in the one-dimensional case, because changes in T do not affect the original distributions characteristic shape, we expect the renormalized distribution to be independent of temperature. This does, indeed, turn out to be the case. This is illustrated in Figure 2, which overlays the results of the calculations for $T = 50$ K, 500 K, and 5000 K. It is also worth noting that, just like our one-dimensional case, the curve obeys the 60/40 rule of *restricted diversity* [6]: regardless of temperature, over 60 percent of molecules are in the lower 40 percent of *diversity* of energy probability states (here again, *diversity* is defined as the number of equivalent equiprobable energy states required to maintain the same amount of Shannon entropy H).

In addition, it is worth noting that as we might expect, adding more degrees of freedom increases the average energy by a factor of $(1/2)k_BT$ per degree while maintaining the same shape for the distribution of energy. Hence, the current result will still hold true for gas molecules with higher degrees of freedom; that is, the distribution of diversity is always exactly the same for an ideal gas, whether monoatomic or polyatomic.

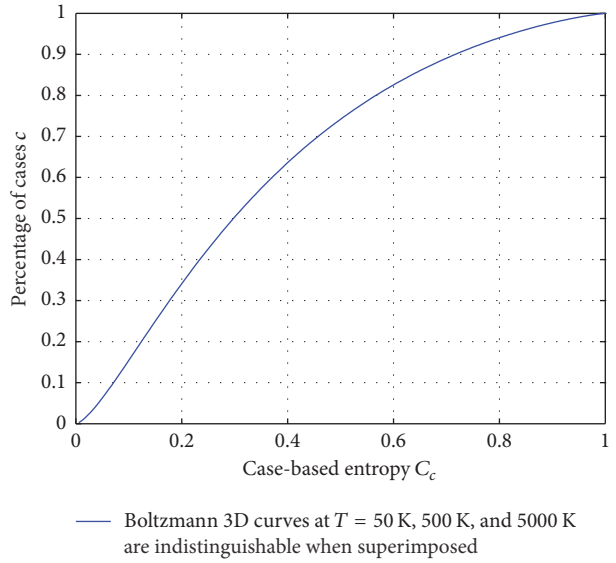


FIGURE 2: C_c versus c for Boltzmann 3D superimposed at three different temperatures: T = 50 K, 500 K, and 5000 K.

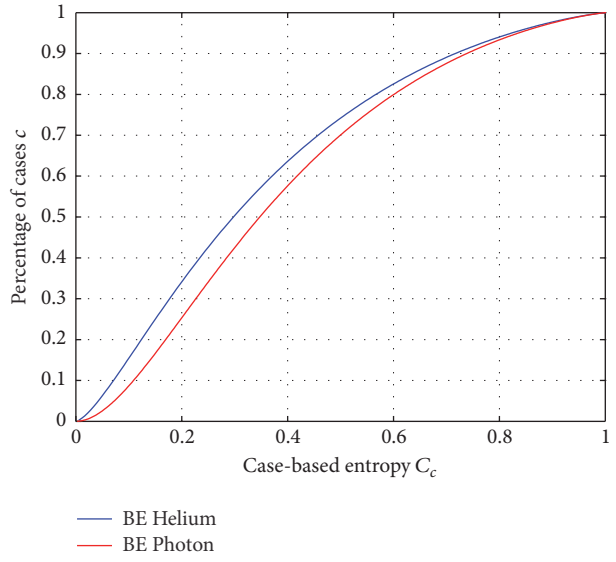


FIGURE 3: C_c versus c for Helium-4 and for photons. Note: the results of calculations carried out at T = 50 K, 500 K, and 5000 K are overlaid.

4.3. The Bose-Einstein Distributions for Massive and Massless Bosons. We now move on to consider the second of our example distributions. The Bose-Einstein distribution gives the energy probability density function for massive bosons above the Bose temperature T_B as

$$p_{\mathrm{HB}}(E) = C \cdot \frac{E^{1/2}}{Be^{\beta E} - 1}, \tag{17}$$

where C is a normalization constant and

$$B = \frac{1}{\zeta(3/2)} \left(\frac{T}{T_B}\right)^{3/2}, \tag{18}$$

where ζ is the Riemann zeta function. In the following calculations, we use the Bose temperature for helium, T_B = 3.14 K.

For massless bosons such as photons, the energy probability density function is [13]

$$p_{\mathrm{BE}} = C \cdot \frac{E^2}{e^{\beta E} - 1}. \tag{19}$$

It is important to note that the "density of states" factors shown in (17) and (19) result in different energy distributions, despite the two types of boson obeying the same statistics.

The conditional probabilities, conditional entropies, true diversities, and case-based entropies for these distributions cannot be calculated analytically but can be calculated numerically. The results of such calculations, using the software Matlab, are shown in Figure 3.

As with the Boltzmann distributions, we find that the distributions of diversity for the two boson systems are

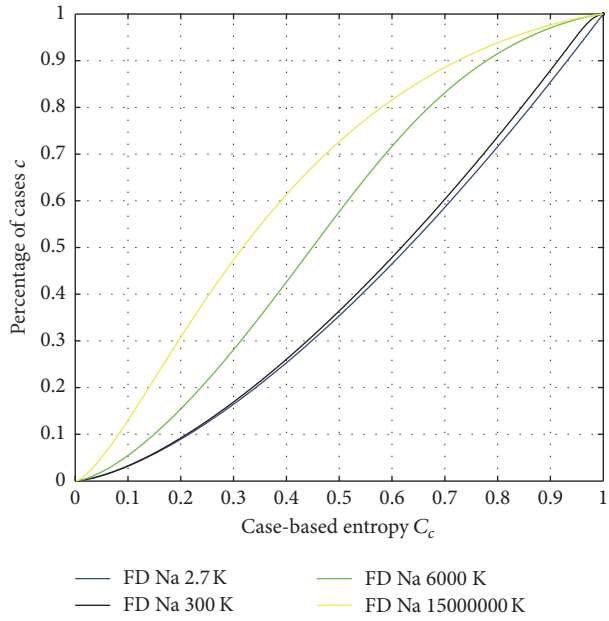


FIGURE 4: Diversity curves for sodium electrons at a range of temperatures with C_c on the x-axis and c on the y-axis.

independent of temperature. Although the curves for the two types of boson are very similar, it is evident that the distributions of diversity do differ to some extent. For helium-4 bosons, a slightly larger fraction of particles are contained in lower diversity energy states than is the case for photons, with 60% of atoms contained in the approximately 37% of the lowest diversity states, as compared to approximately 42% for photons. In other words, using C_c, we are able to identify, even in such instances where intuition might suggest it to be true, common patterns within and across these different energy systems, as well as their variations. With this point made, we move to our final energy distribution.

4.4. The Fermi-Dirac Distribution. The final distribution we use to illustrate our approach is the Fermi-Dirac distribution:

$$p_{\mathrm{FD}}(E) = C \cdot \frac{E^{1/2}}{e^{\beta(E-\mu)} + 1}, \tag{20}$$

where C is again a normalization constant and μ is the Fermi energy [13]. In the following, we calculate distributions for sodium electrons, for which $\mu = 3.4$ eV. Once again, $\widehat{p}$, H_c, D_c, and C_c cannot be calculated analytically and so we rely on numerical calculations using Matlab.

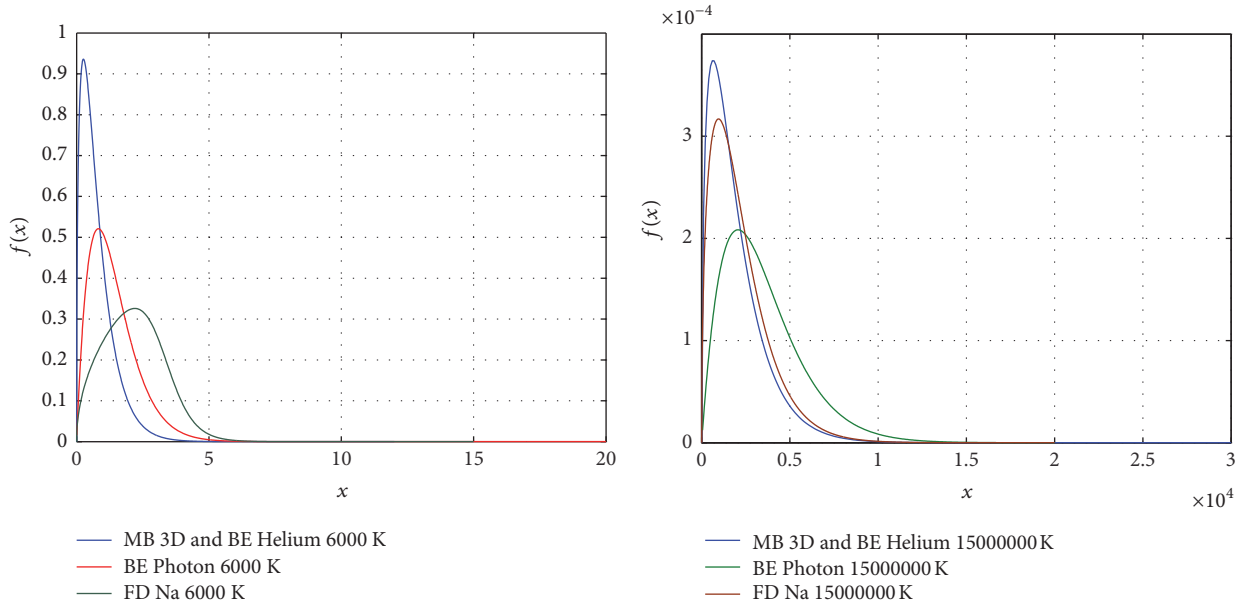
The Fermi-Dirac distribution differs from the previous examples in that it is not simply scaled by changes in energy. Instead, its shape changes, transforming from a skewed-left distribution, with a sharp cut-off at the Fermi energy at low temperatures, to a smooth, skewed-right distribution at high temperatures. Thus, unlike the situation for Boltzmann and Bose-Einstein distributions, one would expect the distributions of diversity for fermions such as electrons to be dependent on temperature. Figure 4 compares the results of calculating C_c as a function of c for electrons in sodium at temperatures of 2.7 K (the temperature of space), 300 K (representing temperatures on earth), 6000 K (the temperature of the surface of the sun), and 15×10^6 K (the temperature of the core of the sun).

This figure shows that the degree of diversity is the highest for fermions at low temperatures; for example, at 2.7 K, fully 70% of the lowest equiprobable diversity states are need to contain 60% of the particles, compared with only approximately 38% at $15{\times}10^6$ K. It also shows that, for sodium electrons, the diversity curve at normal temperatures on earth (300 K) is almost identical to that at very low temperatures. That is, a room temperature Fermi gas of sodium electrons has a distribution of diversity very similar to that of a "Fermi condensate."

5. Using C_c to Compare and Contrast Systems

With our renormalization complete for all three distributions, we sought next to demonstrate, albeit somewhat superficially, the utility of C_c for comparing and contrasting systems, given how widely known the results are for these three classic energy distributions. To begin with, it is usual to assume that, in the limit of high T, both Bose-Einstein and Fermi-Dirac distributions reduce to Boltzmann distributions, and so the physical properties of both bosons and fermions in this limit should be those of an ideal gas.

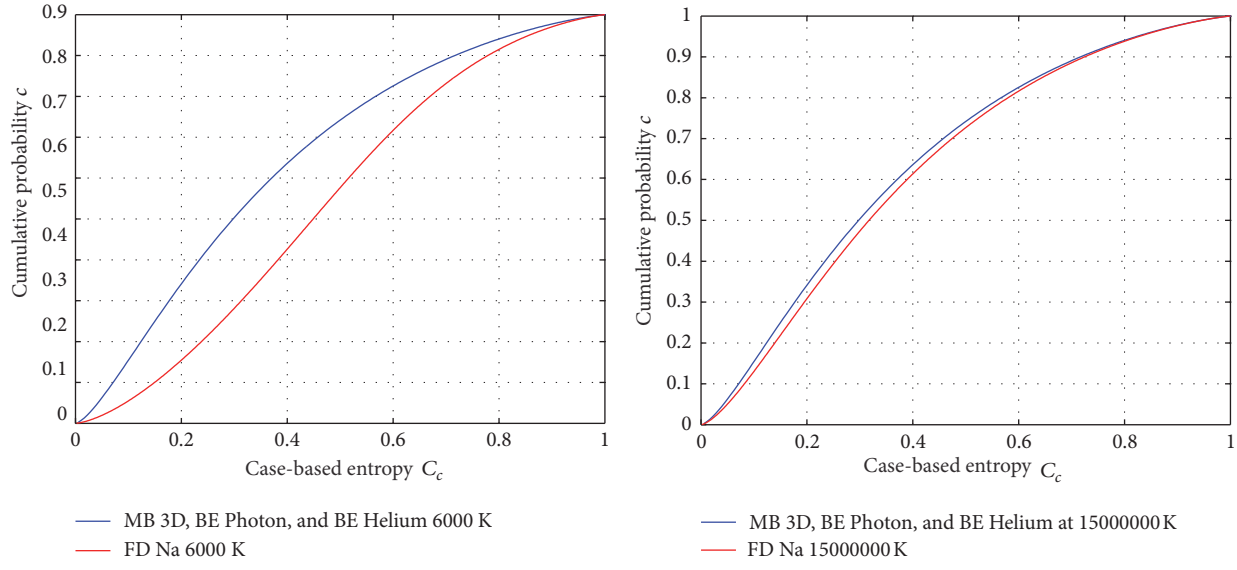
In Figures 5 and 6, we show a comparison of all three energy distributions for temperatures of 6000 K and 15×10^6 K (the Bose-Einstein distribution for massless bosons is included for comparison). In these figures, it appears that, by 6000 K, the Bose-Einstein distribution for helium-4 is indistinguishable from the 3D Boltzmann distribution. Also, while the Fermi-Dirac distribution has clearly not reduced to the Boltzmann distribution even at 15×10^6 K, it appears to be trending towards it.



(a) Energy density curves for Maxwell-Boltzmann 3D, Bose-Einstein Helium, Bose-Einstein Photon, and Fermi-Dirac Na 6000 K

(b) Energy density curves for Maxwell-Boltzmann 3D, Bose-Einstein Helium, Bose-Einstein Photon, and Fermi-Dirac Na 15000000 K

Figure 5: Energy density curves.



(a) C_c versus c curves for Maxwell-Boltzmann 3D, Bose-Einstein Helium, Bose-Einstein Photon, and Fermi-Dirac Na 6000 K

(b) C_c versus c curves for Maxwell-Boltzmann 3D, Bose-Einstein Helium, Bose-Einstein Photon, and Fermi-Dirac Na 15000000 K

Figure 6: C_c versus c curves.

However, comparison of the diversity distributions suggests that even when the energy probability density functions appear to coincide, significant physical differences remain between the systems. Figure 7 compares all the diversity curves calculated in the present work.

It is clear from Figure 7 that the distributions of *diversity* for a classical ideal gas and for both Bose-Einstein and Fermi-Dirac distributions are significantly different. Because these renormalized distributions are independent of temperature, this suggests that there is no limit in which the Bose-Einstein distribution for the photon becomes completely indistinguishable from the Boltzmann distribution. Even more strikingly, the distribution of diversity in a system obeying Fermi-Dirac statistics only approaches that of bosonic systems at extremely high temperatures, similar to those at the core of the sun. At lower temperatures, the Fermi gas has substantially higher degrees of diversity than all the other systems. This is because, at lower temperatures, most of the

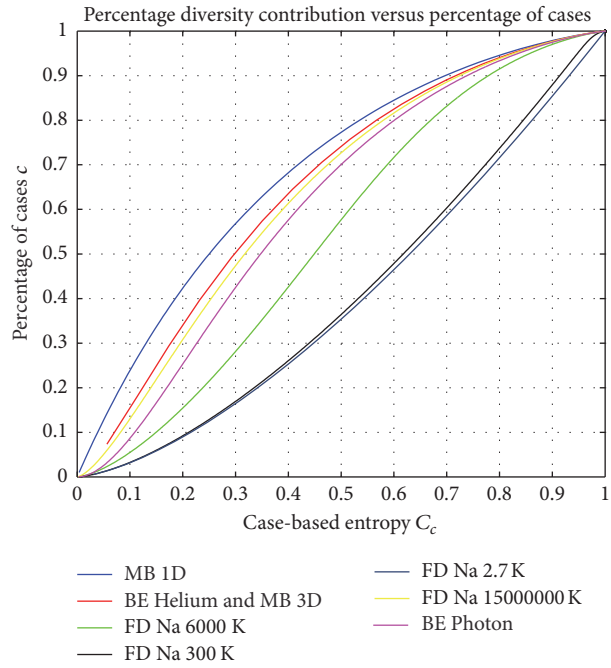


Figure 7: Superposition of all diversity curves for Boltzmann 1D, Boltzmann 3D, Bose-Einstein Helium, Bose-Einstein Photon, and Fermi-Dirac Na at 2.7 K, 300 K, 6000 K, and 15000000 K.

fermions are yet to surpass the barrier created by the Fermi energy and hence are all restricted to the lower end of the energy.

Thus, the transformation from the usual probability distribution to a distribution of case-based entropy (C_c versus c) has allowed us to make direct scale-free comparisons, of the ways in which the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac energy distributions are similar or differ both internally (as a function of temperature T) and across distributions. It appears that, except for extremely high temperatures, the Fermi-Dirac distribution has a larger value of C_c than the others. This means that there are a larger number of Shannon-equivalent equiprobable states of energy for the Fermi-Dirac distribution as compared to the others. A speculative explanation could be that Pauli's exclusion principle does not allow for more than one fermion to occupy the same quantum state, thereby restricting the accumulation of fermions in the same state (i.e., more diversity).

6. Conclusion

As we have hopefully shown in this paper, while Shannon entropy H has been used to rethink probability distributions in terms of diversity, it suffers from two major limitations. First, it cannot be used to compare distributions that have different levels of scale. Second, it cannot be used to compare parts of distributions to the whole.

To address these limitations, we introduced a renormalization of probability distributions based on the notion of *case-based entropy* C_c (as a function of the cumulative probability c). We began with an explanation of why we rethink probability distributions in terms of diversity, based on a Shannon-equivalent uniform distribution, which comes from the work of Jost and others on the notion of *true diversity* in ecology and evolutionary biology [4, 9, 10]. With this approach established, we then reviewed our construction of case-based entropy C_c. Given a probability density $p(x)$, C_c measures the diversity of the distribution up to a cumulative probability of c, by computing the length or support of an equivalent uniform distribution that has the same Shannon information as the conditional distribution of $\widehat{p}_c(x)$ up to a cumulative probability c.

With our conceptualization of C_c complete, we used it to renormalize and compare three physically significant energy distributions in physics, namely, the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac distributions for energy of subatomic particles. We chose these three distributions for three key reasons: (1) we wanted to see if C_c works for continuous distribution; (2) where the focus was on diversity of types and not on their rank order in terms of complexity; and (3) where the unit order of measure was both objective and widely accepted. Based on our results, we concluded that C_c is a vast improvement over H as it provides an intuitively useful, scale-free comparison of probability distributions and also allows for a comparison between parts of distributions as well.

The renormalization obtained will have a different shape for different distributions. In fact, a bimodal, right skewed, or other kinds of distributions will lead to a different C_c versus c curve. There are two interesting points of inquiry in future papers, namely, (a) how the shape of the original distribution influences the C_c versus c curve and (b) whether we can reconstruct the original shape of the distribution given the C_c versus c curve. Because of the scale-free nature of C_c,

all distributions can be compared in the same plot without reference to their original scales. In our future work, we will endeavor to connect the shape of the C_c versus c curve to the shape of the original distribution. This will allow us to locate portions of the original distribution (irrespective of their scale), where diversity is concentrated, and portions where it is sparse, even though the original distributions cannot be plotted on the same graph due to huge variation in their scales.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the following colleagues at Kent State University: (1) Dean Susan Stocker, (2) Kevin Acierno and Michael Ball (Computer Services), and (3) the Complexity in Health and Infrastructure Group for their support. They also wish to thank Emma Uprichard and David Byrne and the ESRC Seminar Series on Complexity and Method in the Social Sciences (Centre for Interdisciplinary Methodologies, University of Warwick, UK) for the chance to work through the initial framing of these ideas.

References

[1] S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton University Press, 2008.

[2] S. E. Page, *Diversity and Complexity*, Princeton University Press, 2008.

[3] M. O. Hill, "Diversity and evenness: a unifying notation and its consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.

[4] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.

[5] R. Rajaram and B. Castellani, "An entropy based measure for comparing distributions of complexity," *Physica A. Statistical Mechanics and Its Applications*, vol. 453, pp. 35–43, 2016.

[6] B. Castellani and R. Rajaram, "Past the power law: complex systems and the limiting law of restricted diversity," *Complexity*, vol. 21, no. 2, pp. 99–112, 2016.

[7] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[8] M. C. Mackey, *Time's Arrow: The Origins of Thermodynamic Behavior*, Springer Verlag, Germany, 1992.

[9] T. Leinster and C. A. Cobbold, "Measuring diversity: the importance of species similarity," *Ecology*, vol. 93, no. 3, pp. 477–489, 2012.

[10] J. Beck and W. Schwanghart, "Comparing measures of species diversity from incomplete inventories: an update," *Methods in Ecology and Evolution*, vol. 1, no. 1, pp. 38–44, 2010.

[11] R. H. Macarthur, "Patterns of species diversity," *Biological Reviews*, vol. 40, pp. 510–533, 1965.

[12] R. Peet, "The measurement of species diversity," *Annual Review of Ecological Systems*, vol. 5, pp. 285–307, 1974.

[13] C. H. Tien and J. H. Lienhard, *Statistical Thermodynamics*, Hemisphere, 1979.