International Journal of Theoretical Physics, Vol. 27, No. 4, 1988

Bayesian Approach to Thermostatistics

B. H. Lavenda[1]

Received October 8, 1987

Bayes' theorem is used to derive the dual of the Gibbs formulation of statistical thermodynamics. An asymptotic analysis is performed, akin to Khinchin's use of the central limit theorem, to determine approximate expressions for the moment-generating functions. The prior densities, which are determined by equating the maximum-likelihood estimates with the moment expressions in the asymptotic limit, satisfy Jeffreys' invariant properties of improper prior densities.

1. INTRODUCTION

Most kinetic based theories of statistical thermodynamics start with a reversible molecular description to which a *randomness* hypothesis must be added in order to obtain macroscopic irreversible behavior. More recent theories avoid making any randomness assumption by taking an appropriate asymptotic limit (Bunimovich and Sinai, 1981). Asymptotic theories, at least implicitly, make appeal to certain theorems in probability, such as the law of large numbers and the central limit theorem, which deal with a large number of independent and identically distributed random variables. In light of the conceptual difficulties in going from a causal to a probabilistic description and their inherent lack of uniqueness, it appears worthwhile to concentrate on the element of randomness itself caused by the thermal interaction of material bodies (Szilard, 1925; Mandelbrot, 1956, 1962, 1964; Lavenda, 1987a; Lavenda and Scherer, 1987a).

For instance, when a system is placed in thermal contact with a thermostat, the energy of the system ceases to be a thermodynamic function, since it is no longer uniquely determined in terms of the temperature and number of external parameters needed to specify the state of the system. The energy of the system is said to undergo "fluctuations." Random samples

[1]Università di Camerino, Camerino 62032 (MC), Italy.

0020-7748/88/0040-0451$06.00/0

of such an extensive or observable variable, taken from a population of systems all having the same but unknown value of the conjugate intensive or nonobservable variable, can be used to estimate the latter. Energy and (inverse) temperature are a pair of Laplace conjugate variables. In the limit where the thermostat vanishes, the energy is fixed and a definite temperature cannot be assigned to the system (Landau and Lifshitz, 1969). Alternatively, in the limit of an infinitely large thermostat, the temperature can be determined with unlimited precision, while the energy becomes completely indeterminate. In other words, an infinite thermostat ensures that the heat capacity of the entire system is infinite and the variance of the temperature fluctuations tends to zero. The conjugate variables thus satisfy a fundamental uncertainty relation in which the precision of measurement of one variable, as measured by the variance, increases or decreases at the expense of the precision of measurement of the conjugate variable (Mandelbrot, 1956). At thermodynamic equilibrium, the thermodynamic uncertainty inequality reduces to an equality when the relationship between the conjugate random variables is perfect linear negative (Lavenda, 1987a). Treating one as a fixed but unknown quantity and the other as a fluctuating variable, the variance of the latter reaches its minimum value predicted by the Cramér-Rao lower bound, which is expressed in terms of the inverse of the Fisher information. the idea of "information" in this sense has an intuitive appeal, since the more information we have, the better the estimator or the more *efficient* it becomes. The one that has a variance equal to the inverse of the Fisher information is the most efficient estimator and this occurs at thermodynamic equilibrium (Lavenda, 1987a). The attainment of the Cramér-Rao lower bound means that the conjugate random variables are linear functions of one another.

The structure of statistical thermodynamics or, using Mandelbrot's (1956) terminology, which he attributes to Kramers, "thermostatistics" is in harmony with the theory of maximum likelihood, which was first detailed by Fisher (1922), although it can be traced back to Lambert around 1765 and Daniel Bernoulli, who used it 12 years later without attempting any justification (Barnett, 1973). The distribution of the extensive or observable variables depends on the intensive or estimable variables, which are related to the *state of nature* of the physical system. Consider the simplest case where the fluctuating observable is the energy ε and the conjugate intensive quantity is β, whose true value will be identified with the inverse absolute temperature. In other words, unlike the energy ε, which always has a physical significance, the estimates of the inverse temperature, which are single-valued functions of ε, are not thermodynamic intensive variables except for the true β. Let $\Omega^{\#}(\varepsilon\,|\,\beta)$ stand for the probability density function (pdf) for the observation ε for a given but unknown value of the parameter β. In

thermostatistics, $\Omega^{\#}(\varepsilon\,|\,\beta)$ is known as the canonical Gibbs density

$$\Omega^{\#}(\varepsilon\,|\,\beta) = \exp[-\beta\varepsilon - \log \mathcal{Q}(\beta)]\,\Omega(\varepsilon) \tag{1}$$

which belongs to an *exponential family.* The density $\Omega(\varepsilon)$, known as the "structure function" (Khinchin, 1949), summarizes all the thermodynamic information known about the system prior to observation, and the norming constant

$$\mathcal{Q}(\beta) = \int \exp(-\beta\varepsilon)\,\Omega(\varepsilon)\,d\varepsilon \tag{2}$$

is the partition function.

It is easily deduced from the Gibbs density (1) that it satisfies Neyman's factorization theorem ensuring that ε is a "sufficient" statistic for estimating the unknown parameter β. Any statistic that summarizes all the experimental data relevant to the estimation of the state of nature is said to be a *sufficient* statistic. Alternatively, if a sufficient statistic exists, it can always be found by the method of maximum likelihood. The maximum-likelihood method interchanges the roles of the variable and parameter such that $\Omega^{\#}(\varepsilon\,|\,\beta)$ is considered as a function of β. When viewed in this manner, we will write $\Omega^{\#}(\varepsilon\,|\,\beta) \propto \exp[\mathcal{L}(\beta\,|\,\varepsilon)]$, where $\mathcal{L}(\varepsilon\,|\,\beta) = -\beta\varepsilon - \log \mathcal{Q}(\beta)$ is the log-likelihood function and the proportionality sign stands for the fact that the structure function is extraneous in the method of maximum likelihood, since it drops out in the maximization of the likelihood function or, what is equivalent, the log-likelihood function $\mathcal{L}(\beta\,|\,\varepsilon)$ relative to β. It must be borne in mind that $\exp[\mathcal{L}(\beta\,|\,\varepsilon)]$ is *not* a pdf, since β is not a random variable and $\int \exp[\mathcal{L}(\beta\,|\,\varepsilon)\,d\beta \neq 1$. There is no sense in speaking about the "probability" of any given value of β, although the likelihood of different values of β can be compared. The maximum likelihood value is determined from the likelihood equation $\partial\mathcal{L}(\hat{\beta}\,|\,\varepsilon)/\partial\beta = 0$, where $\partial\mathcal{L}(\hat{\beta}\,|\,\varepsilon)/\partial\beta$ denotes the derivative of $\mathcal{L}$ with respect to β evaluated at the maximum-likelihood value $\hat{\beta}$. Not only does the likelihood equation substantiate Gauss' (1963) assumption of the equivalence of the sample mean and maximum-likelihood value, but, in addition, it equates it with the expected value due to the fact that the distribution belongs to an exponential family. Usually this is true as the number of observations becomes indefinitely large, which can easily be demonstrated by using Chebychev's inequality. We should also note that the statistical property of sufficiency can be used to derive the canonical Gibbs distribution (Mandelbrot, 1956).

It is therefore not surprising that the method of maximum likelihood gives all the well-known results of Gibbsian thermostatistics. As we have mentioned, the fact that the energy or the temperature or the temperature both fluctuate depends on the relative size of the thermostat in comparison

with the size of the system. It therefore cannot be universally valid that β is a fixed, unknown parameter to be estimated in terms of observations made on the energy ε. If for no other reason than the desire for symmetry in nature, energy and temperature should be mutually symmetrical. We are therefore led to consider β as a random variable equipped with a, perhaps unknown, prior density $\omega(\beta)$. This is usually referred to as the "Bayes case" and the dual description can be formulated with the aid of Bayes' theorem or the *principle of inverse probability.* In contrast to "direct probability," for which the random process together with the parameters are known and where probabilistic statements are to be made about the possible outcomes of an experiment, "inverse probability" attempts to infer the random process that has generated the data.

Bayes' theorem, which results from the symmetry of the joint distribution of β and ε, can be phrased as (Jeffreys, 1973) "posterior pdf$\propto$ likelihood$\times$prior pdf," or in symbols

$$\omega^{\#}(\beta\,|\,\varepsilon)\propto\exp[\mathscr{L}(\beta\,|\,\varepsilon)]\,\omega(\beta) \tag{3}$$

The constant of proportionality omitted from (3) is

$$\mathscr{Z}(\varepsilon)=\int\exp[\mathscr{L}(\beta\,|\,\varepsilon)]\,\omega(\beta)\,d\beta \tag{4}$$

The posterior density $\omega^{\#}(\beta\,|\,\varepsilon)$ for the parameter β given the observation ε, as well as the prior density $\omega(\beta)$ for the parameter β, cannot be interpreted as a density in the frequency sense. Rather, it is to be interpreted in the sense of "degree of belief" that some values of β are more "probable" than others. Expressed in words, Bayes' theorem states that the probability that the unknown parameter has the value β given the datum ε is proportional to the product of the likelihood of observing ε given β multiplied by the initial probability of β (Savage, 1962).

Because of the belief that certain values of the parameter β are more "frequent" than others, we are led to consider it as a random variable that is equpped with a prior pdf. The stumbling block in putting Bayes' theorem into practice has always been how to choose the prior pdf $\omega(\beta)$. In the absence of any information about the parameter, it is common practice to use the uniform prior which is formalized by the Bayes–Laplace principle of "insufficient reason" (Jeffreys, 1961). In fact, Boltzmann's derivation of the "most probable" distribution avoids the specification of prior probabilities on the basis of the postulate of equal *a priori* probability, stating that all distributions in energy among the systems are equally probable. Since the number of distinct ways in which one can assign energy values to a collective is not a probability but rather a "thermodynamic probability"

(Planck, 1954), which is maximized with respect to energy and number constraints, in what sense is the derived distribution "most probable"? A more rigorous derivation of the Boltzmann distribution, which avoids the use of Stirling's approximation, is provided by the Darwin-Fowler method of mean values, which is an asymptotic method using saddle point integration (see, for instance, Huang, 1963). As the number of systems comprising the esemble tends to infinity, the most probable distribution of energy among the systems tends to the expected value and the saddle point is formed by an infinitely sharp peak and an infinitely steep valley. The thermodynamic probability is modified so that it is converted into a multinomial distribution, but the prior probabilities are merely introduced for mathematical convenience and have nothing to do with the *a priori* probability of finding a system in a given state. At the end of the calculation, the priors are all set equal to one. The parameter β is estimated from the equivalence of the expected and observed values of the energy that is obtained from the equation bears a striking similarity to the likelihood equation.

The Bayes-Laplace rule of setting $\omega(\beta)\, d\beta \propto d\beta$ or the uniform assessment of a variable was criticized by Jeffreys (1961) in the case where it has a finite range of possible values. Jeffreys offered, as an example, the law connecting the mass and volume of a substance. If the uniform rule is adopted for one of the variables, he concluded that it will be incorrect for the other variable, since the mass and specific volume are reciprocals of one another. In order to make up for such inconsistencies in the Bayes-Laplace rule, Jeffreys (1961) and others made extensive use of *improper* priors to represent "knowing little" based upon certain invariance properties that the priors should have. He suggested that if the parameter may have any value from 0 to ∞, the prior probability of its logarithm should be taken as uniformly distributed. That is, if we set $\vartheta = \log \beta$, then the prior density for ϑ is $\omega(\vartheta)\, d\vartheta \propto d\vartheta$ $(-\infty < \vartheta < \infty)$. Since $d\vartheta = d\beta/\beta$, this implies that $\omega(\beta)\, d\beta \propto d\beta/\beta$ $(0 < \beta < \infty)$ is an *improper* pdf. Observing that the representation of certainty by unity is only one of convention, Jeffreys was led to consider $\int d\beta/\beta = \infty$ as a statement of certainty when improper prior pdf's are used. The fact that both $\int_0^a d\beta/\beta = \infty$ and $\int_a^\infty d\beta/\beta = \infty$ imply that $\Pr\{0 < \beta < a\}/\Pr\{a < \beta < \infty\}$ is indeterminate simply means that nothing can be said about the ratio of the two probabilities. Indeterminacy is taken as a formal representation of ignorance. Moreover, the choice of the improper prior pdf as $\omega(\beta) \propto 1/\beta$ is also seen to be invariant to transformations of the form $\vartheta = \beta^n$, since $d\beta/\beta$ and $d\beta^n/\beta^n$ are always proportional. This would not be true if the uniform distribution were used. Jeffreys cited the measurement of the charge of an electron, where some methods give e while others e^2, and certainly de and de^2 are not proportional. As a generalization of the invariance of the prior under transformations of the

form $\vartheta = \mathcal{T}(\beta)$, with $\mathcal{T}$ a differentiable transformation, Jeffreys showed that the prior should be taken to be proportional to $[\mathcal{I}(\vartheta)]^{1/2}$, where $\mathcal{I}(\vartheta) = -\overline{\partial^2 \mathcal{L}/\partial \vartheta^2}$ is the Fisher information and the overbar denotes the expectation value. Denoting $|\mathcal{J}|$ as the Jacobian of the transformation, namely $|\mathcal{J}|\, d\beta = d\vartheta$, then the fact that $|\mathcal{J}| = [\mathcal{I}(\beta)/\mathcal{I}(\vartheta)]^{1/2}$ implies $\omega(\beta)\, d\beta = \omega(\vartheta)\, d\vartheta$ or that the probability should be independent of the parametrization.

In this paper we develop the dual, or the Bayes representation, of the Gibbs formulation by interchanging the role of the thermodynamic conjugate variables and we show that Jeffrey's invariance properties for choosing prior pdf's are satisfied in statistical mechanics. This will be accomplished "inverting" Bayes' theorem (3) in the asymptotic, large-sample limit, where Laplace's method is applicable. Laplace's method (Sirovich, 1971) can be considered as the real variable analog to the saddle point method that was employed by Darwin and Fowler. We will arrive at the conclusion that since the Gibbs canonical formalism coincides with the method of maximum likelihood, it is the more obvious of the two representations. Furthermore, we shall show that the Gibbs formulation is more amenable to the study of the thermal iteraction of material bodies through the exchange of additive invariants, since ε rather than β is considered as the random variable. However, the maximum-likelihood method (Fisher, 1922) was initially set up against the older Bayes method. And it is Bayes' method that leads to the dual representation, which has a closer correspondence to the fundamental equations of macroscopic thermodynamics than the canonical Gibbs formalism.

The asymptotic analysis is justifiable on the basis that the solution should not depend heavily on the specific type of distribution, sample size, etc., and that what we are really interested in is the thermodynamic limit where most probable and mean values coincide. It has often been stated that "the epistemological value of probability theory is revealed only by limit theorems" (Gnedenko and Kolmogorov, 1954) and we will show that thermostatistics is no exception.

2. ASYMPTOTIC EVALUATION OF BAYES' THEOREM

The asymptotic results of thermostatistics, like large-sample theory in statistics (Chernoff, 1956), can be shown to be based on certain theorems in the theory of probability which make it relatively easy to obtain good approximate results in the limit as the number of systems comprising the ensemble increases without limit or if the sample size is large (Lavenda and Scherer, 1987b). These theorems, like the law of large numbers and the central limit theorem, are extremely elegant and their elgance has surely

been captured by the asymptotic results of thermostatistics. In fact, the "sample size" may be interpreted as the number of subdivisions of the original system or the number of systems in thermal contact with it. The thermodynamic *additivity* property of the energy is translated into the statement that the extensive observables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and identically distributed random variables which for sufficiently large n come under the jurisdiction of the law of large numbers and the central limit theorem. Furthermore, if the process is *reversible*, then the entropy is additive. The additivity of the energy places a lower limit on the number of possible subdivisions. The fact that the true value of β is a "macroscopic" intensive variable means that no better estimate for it can be obtained by making finer and finer subdivisions (Mandelbrot 1956; Lavenda, 1987b). This is none other than Fisher's (1922) criterion that ε be a sufficient statistic for the estimation of β. In other words, the grouping of observations, into, say, the sample mean $\hat{\varepsilon} = (1/n)\sum_i^n \varepsilon_i$, does not cause a loss of "information" (Kullback, 1959).

Because the ε_i are independent and identically distributed random variables, the likelihood function will factor into a product of n factors for a random sample of size n. This means that the log-likelihood will increase as n and it will ultimately dwarf the comparative term $\log \omega(\beta)$ in equation (4), which is independent of the sample size or the number of subdivisions. As $n \to \infty$, the method of Laplace can be used to obtain an asymptotic expression for the integral in (4). Laplace (Sirovich, 1971) argued that the main contribution to the integral comes from the neighborhood of the global maximum of the log-likelihood function at $\hat{\beta}$. Expanding $\mathscr{L}(\beta \mid \varepsilon)$ in a Taylor series about $\hat{\beta}$, we find

$$\begin{aligned}\mathscr{L}(\bar{\varepsilon}) &\sim \exp[\mathscr{L}(\hat{\beta} \mid \varepsilon)]\, \omega(\hat{\beta}) \int_{-\infty}^{\infty} \exp\left[\frac{1}{2}\frac{\partial^2 \mathscr{L}(\hat{\beta} \mid \bar{\varepsilon})}{\partial \beta^2}(\beta - \hat{\beta})^2\right] d(\beta - \hat{\beta}) \\ &= \exp[\mathscr{L}(\hat{\beta} \mid \varepsilon)]\, \omega(\hat{\beta}) \left\{2\pi\left[-\frac{\partial^2 \mathscr{L}(\hat{\beta} \mid \varepsilon)}{\partial \beta^2}\right]^{-1}\right\}^{1/2} \end{aligned} \tag{5}$$

where the maximum-likelihood estimate $\hat{\beta}$ is a solution of the likelihood equation

$$-\frac{\partial \mathscr{L}(\hat{\beta} \mid \varepsilon)}{\partial \beta} = \varepsilon + \frac{\partial \log \mathscr{Q}(\hat{\beta})}{\partial \beta} = 0 \tag{6}$$

and $\partial^2 \mathscr{L}(\hat{\beta} \mid \varepsilon)/\partial \beta^2$ denotes the second derivative with respect to β evaluated at $\hat{\beta}$. The reason for neglecting higher powers in the Taylor series expansion can be seen in the following way. Since $\sigma_n^{-2}(\hat{\beta}) \equiv -\partial^2 \mathscr{L}(\hat{\beta} \mid \varepsilon)/\partial \beta^2$ is of order n, the term $\exp[-(\beta - \hat{\beta})^2/2\sigma_n^2]$ will be appreciably different from zero for $(\beta - \hat{\beta}) \sim O(1/\sqrt{n})$. In that case, the remainder term is of $O(1/\sqrt{n})$ and is

negligible compared with the second-order term, which is independent of n in the limit as $n \to \infty$. And the fact that the sampling variance $\sigma_n^2(\hat{\beta}) \sim O(n)$, which approaches 0 as $n \to \infty$, has allowed us to replace the limits of integration of 0 to ∞ in equation (5) by $-\infty$ to $+\infty$.

Developing the log-likelihood function in equation (3) in a Taylor series under the same conditions and dividing by the asymptotic form of $\mathcal{Z}$ given by (5), we obtain

$$\omega^{\#}(\beta \mid \varepsilon) \sim \left[\frac{1}{2\pi} \frac{\partial^2 \log \mathcal{Q}(\hat{\beta})}{\partial \beta^2}\right]^{1/2} \exp\left[-\frac{1}{2} \frac{\partial^2 \log \mathcal{Q}(\hat{\beta})}{\partial \beta^2} (\beta - \hat{\beta})^2\right] \tag{7}$$

which is the normal approximation for the posterior density that is valid for sufficiently large n. In the asymptotic limit, the mean value

$$\bar{\beta} = \int \beta \omega^{\#}(\beta \mid \varepsilon)\, d\beta = -\frac{\partial \log \mathcal{Z}(\varepsilon)}{\partial \varepsilon} \tag{8}$$

will coincide with the maximum-likelihood estimate $\hat{\beta} = \beta(\bar{\varepsilon})$, where $\bar{\varepsilon}$ is the expected value of the energy that is obtained by inverting the likelihood equation (6). In other words, the mean value (8) approaches the most likely value, obtained from the implicit relation (6), or equivalently

$$\hat{\beta} = -\partial \log \mathcal{Z}(\bar{\varepsilon})/\partial \varepsilon \tag{9}$$

in the limit as $n \to \infty$. In the next section, we will show that $\hat{\beta}$ is the inverse of the absolute temperature T for temperatures measured in energy units, and therefore $-\log \mathcal{Z}(\bar{\varepsilon})$ corresponds to the entropy [cf. equation (23)]. The negative of the entropy coincides with the logarithm of the maximum-likelihood function (Lavenda and Scherer, 1987a), so that

$$\mathcal{L}(\hat{\beta} \mid \bar{\varepsilon}) = -\hat{\beta}\bar{\varepsilon} - \log \mathcal{Q}(\hat{\beta}) = \log \mathcal{Z}(\bar{\varepsilon}) \tag{10}$$

It is to be appreciated that this relation is valid *only* at the maximum-likelihood estimate and, in general, there will not be any relation connecting the moment-generating functions $\mathcal{Q}(\beta)$ and $\mathcal{Z}(\varepsilon)$ in the Gibbs and Bayes formulations, respectively. Therefore, in view of the asymptotic relation (5), we conclude that

$$\omega(\hat{\beta}) \propto [\mathcal{I}(\hat{\beta})]^{1/2} \tag{11}$$

where $\mathcal{I}(\hat{\beta})$ is the Fisher (1922) information, since the expectation make no difference in random samples of fixed size that are taken from an exponential family (Lindley, 1970). Expression (11) is Jeffreys' choice of the prior pdf.

We can invert Bayes' theorem to obtain an asymptotic expression for the Gibbs density, which, when equated to an analogous asymptotic expansion of (1), will give Khinchin's (1949) important formula for the

structure function [cf. equation (18)]. If β is to be a sufficient statistic to estimate ε, then the posterior pdf $\omega^{\#}(\beta|\varepsilon)$ belongs to an exponential family as ε ranges over its values. From the Neyman factorization theorem for a sufficient statistic, we know that it must have the form

$$\omega^{\#}(\beta|\varepsilon) = \exp[\mathscr{A}(\varepsilon|\beta)]\,\Omega(\beta) \tag{12}$$

where the log-likelihood function $\mathscr{A}(\varepsilon|\beta) = -\beta\varepsilon - \log \mathscr{M}(\varepsilon)$ and the function $\Omega(\beta)$ will be subsequently identified. Thus, Bayes' theorem can be expressed as

$$\Omega^{\#}(\varepsilon|\beta) \propto \exp[\mathscr{A}(\varepsilon|\beta)]\,\omega(\varepsilon) \tag{13a}$$

with the norming constant

$$\mathscr{Z}(\beta) = \int \exp[\mathscr{A}(\varepsilon|\beta)]\,\omega(\varepsilon)\,d\varepsilon \tag{13b}$$

Developing the log-likelihood function $\mathscr{A}(\varepsilon|\beta)$ in a Taylor series expansion about $\bar{\varepsilon}$, which is the solution to the likelihood equation

$$\hat{\beta} = -\frac{\partial \log \mathscr{M}(\bar{\varepsilon})}{\partial \varepsilon} \tag{14}$$

and using

$$\mathscr{Z}(\hat{\beta}) \sim 2\pi\left[\frac{\partial^2 \log \mathscr{M}(\bar{\varepsilon})}{\partial \varepsilon^2}\right]^{1/2} \exp[-\hat{\beta}\bar{\varepsilon} - \log \mathscr{M}(\bar{\varepsilon})]\,\omega(\bar{\varepsilon})$$

which is obtained by evaluating the integral (13b) by Laplace's method, we obtain

$$\Omega^{\#}(\varepsilon|\beta) \sim \left[\frac{1}{2\pi}\frac{\partial^2 \log \mathscr{M}(\bar{\varepsilon})}{\partial \varepsilon^2}\right]^{1/2} \exp\left[-\frac{1}{2}\frac{\partial^2 \log \mathscr{M}(\bar{\varepsilon})}{\partial \varepsilon^2}(\varepsilon - \bar{\varepsilon})^2\right] \tag{15a}$$

for the asymptotic expression of the Gibbs density.

An analogous asymptotic expression can be obtained from expression (1) by expanding the log-likelihood function $\mathscr{L}(\beta|\varepsilon) = -\beta\varepsilon - \log \mathscr{Q}(\beta)$ about $\hat{\beta}$, which is the solution of the likelihood equation (6). We then obtain

$$\Omega^{\#}(\varepsilon|\beta) \sim \exp[-\hat{\beta}\bar{\varepsilon} - \log \mathscr{Q}(\hat{\beta})]\,\Omega(\bar{\varepsilon}) \exp\left[-\frac{1}{2}\frac{\partial^2 \log \mathscr{Q}(\hat{\beta})}{\partial \beta^2}(\beta - \hat{\beta})^2\right] \tag{15b}$$

Developing ε in a Taylor series about $\hat{\beta}$, we obtain

$$\varepsilon - \bar{\varepsilon} = -\frac{\partial^2 \log \mathscr{Q}(\hat{\beta})}{\partial \beta^2}(\beta - \hat{\beta}) \tag{16}$$

to lowest order (Khinchin, 1949) in the difference $(\beta - \hat{\beta}) \sim O(1/\sqrt{n})$. The asymptotic relationship (16) between ε and β is perfect linear *negative*, implying that the correlation coefficient assumes its extreme value of -1. Now, using the maximum-likelihood estimate $\bar{\varepsilon} = \varepsilon(\hat{\beta})$, we obtain the following relations between the partial derivatives of the log-likelihood function:

$$\frac{\partial \mathscr{L}}{\partial \varepsilon} = \frac{\partial \mathscr{L}}{\partial \beta}\frac{d\beta}{d\varepsilon}, \qquad \frac{\partial^2 \mathscr{L}}{\partial \varepsilon^2} = \frac{\partial^2 \mathscr{L}}{\partial \beta^2}\left(\frac{d\beta}{d\varepsilon}\right)^2 + \frac{\partial \mathscr{L}}{\partial \beta}\left(\frac{d^2\beta}{d\varepsilon^2}\right)$$

Because $\partial \mathscr{L}/\partial \beta = 0$ and $\hat{\beta}$, the second equation gives $\mathscr{I}(\bar{\varepsilon}) = \mathscr{I}(\hat{\beta})(d\hat{\beta}/d\bar{\varepsilon})^2$. Noting that the Jacobian of the one-to-one differentiable transformation $\bar{\varepsilon} = \varepsilon(\hat{\beta})$ is

$$|\mathscr{J}| = |d\bar{\varepsilon}/d\hat{\beta}| = \mathscr{I}(\hat{\beta})$$

we obtain

$$\mathscr{I}(\bar{\varepsilon}) = 1/\mathscr{I}(\hat{\beta}) \tag{17}$$

This is the Cramér–Rao lower bound for the variance, which is valid only at statistical equilibrium; for nonequilium states, the equality must be replaced by an inequality (Lavenda, 1987a). Alternatively, we could have derived (17) from the relation

$$[\mathscr{I}(\hat{\beta})]^{1/2} = |\mathscr{J}|[\mathscr{I}(\bar{\varepsilon})]^{1/2}$$

by noting that $|\mathscr{J}| = \mathscr{I}(\hat{\beta})$.

On the strength of equations (16) and (17), which can be more explicitly stateed as

$$\mathscr{I}(\bar{\varepsilon}) = \sigma^{-2}(\bar{\varepsilon}) = \frac{\partial^2 \log \mathscr{M}(\bar{\varepsilon})}{\partial \varepsilon^2} = \left[\frac{\partial^2 \log \mathscr{Q}(\hat{\beta})}{\partial \beta^2}\right]^{-1} = \mathscr{I}^{-1}(\hat{\beta}) = \sigma^2(\hat{\beta})$$

we obtain

$$\Omega(\bar{\varepsilon}) = \frac{\mathscr{Q}(\hat{\beta})\exp(\hat{\beta}\bar{\varepsilon})}{[2\pi\mathscr{I}(\hat{\beta})]^{1/2}} \tag{18}$$

by equating the two asymptotic expressions for the Gibos density, (15a) and (15b). Expression (18) was first derived by Khinchin [1949, equation (42)] by invoking the central limit theorem for the Gibbs density, inverting (1), and evaluating the resulting expression at the maximum-likelihood values $\beta = \hat{\beta}$ and $\varepsilon = \bar{\varepsilon}$.

Due to the invariance of joint pdf under the exchange of β and ε, we have

$$\omega^{\#}(\beta \,|\, \varepsilon)\omega(\varepsilon) = \Omega^{\#}(\varepsilon \,|\, \beta)\omega(\beta) \tag{19}$$

which is another way of writing Bayes' theorem. If the posterior densities in (19) are the normal densities (7) and (15a), which are related by

$$\omega^{\#}(\beta\,|\,\varepsilon) = |\mathcal{J}|\Omega^{\#}(\varepsilon\,|\,\beta) \tag{20}$$

then $|\mathcal{J}| = \omega(\hat{\beta})/\omega(\bar{\varepsilon})$. Furthermore, since the prior pdf $\omega(\hat{\beta})$ is given by (11) and Cramér–Rao lower bound (17) applies, we get

$$\omega(\bar{\varepsilon}) \propto [\mathcal{I}(\bar{\varepsilon})]^{1/2} \tag{21}$$

which is again Jeffreys' improper prior pdf for a random quantity that can only take on positive values.

3. THERMODYNAMIC RELATIONS IN THE BAYES REPRESENTATION

We have already mentioned that maximum-likelihood values of the moment-generating functions coincide with thermodynamic functions. The relationship between the thermodynamic functions gives rise to equation (10), which connects the moment-generating functions at their maximum-likelihood values. In statistical mechanics, the identification is made by comparing the canonical expression for the entropy, in terms of mean values, to the Gibbs equation. On the strength of the central limit theorem for a large number of independent and identically distributed random quantities, the asymptotic form of the distribution tends to a normal one for which the means and modes coincide. We now consider the relation between the maximum-likelihood values of the moment-generating functions and their relation to thermodynamic quantities in greater detail.

We multiply equation (9) through by $d\bar{\varepsilon}$ to get

$$\hat{\beta}\, d\bar{\varepsilon} = -d \log \mathcal{Z}(\bar{\varepsilon}) \tag{22}$$

provided there is no work done by the external forces on the system, for otherwise the moment-generating function $\mathcal{Z}(\varepsilon)$ would also be a function of the generalized coordinates. Since there is no difficulty in introducing such dependencies, we will treat the case in which no work is done. According to the first law of thermodynamics, the change in the energy is given by $d\bar{\varepsilon} = \delta Q$, where δQ is the "amount of heat" received by the system during the elementary transition. It is clear that we have had to use (9) instead of (8), since ε is a random quantity, to which the first law can obviously not be applied (Khinchin, 1949; Lavenda and Scherer, 1987b). It is, however, applicable to its mean value $\bar{\varepsilon}$. Hence, the quantity $\hat{\beta}\,\delta Q$ is the total differential of a certain thermodynamic function

$$\hat{\beta}\,\delta Q = -d \log \mathcal{Z}(\bar{\varepsilon})$$

which shows that the function $\hat{\beta}$ is the integrating factor for the quantity δQ and whose existence is postulated by the second law. This thermodynamic function is the entropy S and in energy units we have

$$-d \log \mathscr{Z}(\bar{\varepsilon}) = dS(\bar{\varepsilon}) \tag{23}$$

The derivation of the thermodynamic relation (23) in the Bayes representation is more natural than the canonical Gibbs formalism, since the entropy as a function of the internal energy corresponds to a fundamental equation of thermodynamics, in contrast to the more synthetic functional dependence upon $\hat{\beta}$ which results from the canonical formalism (Tisza and Quay, 1963). Furthermore, subtracting the total differential $d(\hat{\beta}\bar{\varepsilon})$ for both sides of (23) leaves it a total differential, viz.,

$$d[\hat{\beta}\bar{\varepsilon} + \log \mathscr{Z}(\bar{\varepsilon})] = \bar{\varepsilon}\, d\hat{\beta} = d[\hat{\beta}\mathscr{F}(\hat{\beta})] \tag{24}$$

where the Helmholtz free energy $\mathscr{F}(\hat{\beta})$ is related to the logarithm of the partition function by

$$\hat{\beta}\mathscr{F}(\hat{\beta}) = \hat{\beta}\bar{\varepsilon} - S(\bar{\varepsilon}) = -\log \mathscr{Q}(\hat{\beta}) \tag{25}$$

This establishes the validity of equation (10).

The symmetrical structure between the Gibbs canonical representation, in terms of the Gibbs pdf $\Omega^{\#}(\varepsilon\,|\,\beta)$, and the Bayes' representation, in terms of the posterior pdf $\omega^{\#}(\beta\,|\,\varepsilon)$, leads to a number of conjugate thermodynamic relations. In the Bayes representation, the requirement that the likelihood equations (9) and (14) be identical implies $\mathscr{M}(\varepsilon) = \mathscr{Z}(\varepsilon)$. Alternatively, in the Gibbs representation, the equivalence of the likelihood equations (6) and

$$\varepsilon + \frac{\partial \log \mathscr{Z}(\hat{\beta})}{\partial \beta} = 0 \tag{26}$$

leads to the conclusion that $\mathscr{Z}(\beta) = \mathscr{Q}(\beta)$.

Therefore, the expressions for the moment-generating functions (2) and (13b) must be one and the same, implying that

$$\Omega(\varepsilon) = \omega(\varepsilon)/\mathscr{Z}(\varepsilon) \tag{27a}$$

and, in particular,

$$\Omega(\bar{\varepsilon}) = \omega(\bar{\varepsilon})e^{S(\bar{\varepsilon})} \tag{27b}$$

which is Khinchin's expression (18) for the structure function. Likewise, expressing the moment-generating function (4) as the Laplace transform of some "structure function" $\Omega(\beta)$,

$$\mathscr{Z}(\varepsilon) = \int e^{-\beta\varepsilon}\Omega(\beta)\, d\beta \tag{28}$$

and requiring it be identical to (4) yields

$$\Omega(\beta) = \omega(\beta)/\mathcal{Q}(\beta) \tag{29a}$$

and, in particular,

$$\Omega(\hat{\beta}) = \omega(\hat{\beta}) \exp[\hat{\beta}\mathcal{F}(\hat{\beta})] \tag{29b}$$

Formula (29b) can be derived in an analogous manner to Khinchin's expression (18) for the structure function. Introducing the normal approximation (7) for the posterior density $\omega^{\#}(\beta \mid \varepsilon)$ into (12) and rearranging gives

$$\Omega(\beta) = \left[\frac{1}{2\pi} \frac{\partial^2 \log \mathcal{Q}(\hat{\beta})}{\partial \beta^2}\right]^{1/2} \exp\left[-\frac{1}{2} \frac{\partial^2 \log \mathcal{Q}(\hat{\beta})}{\partial \beta^2} (\beta - \hat{\beta})^2\right]$$
$$\times \exp[\beta\varepsilon + \log \mathcal{Z}(\varepsilon)]$$

We choose for the "parameter" ε the simple root of equation (9). In particular, for $\beta = \hat{\beta}$, we obtain the important formula

$$\Omega(\hat{\beta}) = \left[\frac{1}{2\pi} \mathcal{I}(\hat{\beta})\right]^{1/2} \mathcal{Z}(\bar{\varepsilon}) \exp(\hat{\beta}\bar{\varepsilon}) \tag{30}$$

which is the Bayesian analog of Khinchin's expression (18) for the structure function. On the strength of equation (10) relating the two generating functions at their maximum likelihood values, the two structure functions (18) and (30) are related by

$$\Omega(\bar{\varepsilon})\Omega(\hat{\beta}) \propto \exp(\hat{\beta}\bar{\varepsilon}) \tag{31}$$

If expression (19) were to be evaluated at the maximum-likelihood values where (27b) and (29b) hold, then we would obtain $\omega(\bar{\varepsilon}) = \Omega^{\#}(\bar{\varepsilon} \mid \hat{\beta})$, using the pair of relations (12) and (29b), while we would get $\omega(\hat{\beta}) = \omega^{\#}(\hat{\beta} \mid \bar{\varepsilon})$ where (1) and (27b) to be used. It would therefore appear that the choice of improper prior densities would be reflected in improper posterior densities. However, these relations hold for the maximum-likelihood values of the parameters, which are related by a thermal equation of state. These particular values modify the functional form of the distribution and one can no longer speak about Laplace conjugate variables or to distinguish between improper and proper densities, because the pdf's will no longer be normalizable. It suffices to cite an ideal gas, where the product $\hat{\beta}\bar{\varepsilon} = \text{const}$ and the posterior density $\Omega^{\#}(\bar{\varepsilon} \mid \hat{\beta}) \propto 1/\bar{\varepsilon}$ [cf. equations (53) ff.]. Jeffreys (1961, p. 195) contended that when n is sufficiently large, the likelihood function is nearly proportional to (7) and it is immaterial whether the prior density $\omega(\beta)$ is evaluated at the actual value β or at $\beta = \hat{\beta}$. Furthermore, the relation between the generating functions and thermodynamic potentials can only be made at the maximum-likelihood values of

the parameters, so a similar situation is encountered as in the Boltzmann relation of entropy to the nonnormalizable "thermodynamic probability" (Fowler, 1936).

Consider now a composite system formed from two subsystems, which we will label by the subscripts 1 and 2. The variables related to the composite system will not have any subscript. For the composite system we have $\varepsilon_1 + \varepsilon_2 = \bar{\varepsilon}$ and

$$\mathcal{Q}(\beta) = \mathcal{Q}_1(\beta)\mathcal{Q}_2(\beta) \tag{32}$$

This relation is derived from the fact that the generating function $\mathcal{Q}(\beta)$ is the Laplace transform (2) of the structure function $\Omega(\varepsilon)$, which obeys the fundamental law of composition (Khinchin, 1949)

$$\Omega(\bar{\varepsilon}) = \int \Omega_1(\varepsilon_1)\Lambda_2(\bar{\varepsilon} - \varepsilon_1)\, d\varepsilon_1 \tag{33}$$

The integral can be taken between infinite limits, since there is no divergence difficulty; the integrand is different from zero only for $0 < \varepsilon' < \varepsilon$. The entropy of the composite system is

$$\begin{aligned} S(\bar{\varepsilon}) &= \hat{\beta}\bar{\varepsilon} + \log \mathcal{Q}(\hat{\beta}) \\ &= \hat{\beta}\bar{\varepsilon}_1 + \log \mathcal{Q}_1(\hat{\beta}) + \hat{\beta}\bar{\varepsilon}_2 + \log \mathcal{Q}_2(\hat{\beta}) \end{aligned} \tag{34}$$

Additivity only applies if both systems are at the same temperature. However, if the functions $\hat{\beta}\bar{\varepsilon}_1 + \log \mathcal{Q}_1(\hat{\beta})$ and $\hat{\beta}\bar{\varepsilon}_2 + \log \mathcal{Q}_2(\hat{\beta})$ have minima at $\hat{\beta} = \hat{\beta}_1$ and $\hat{\beta} = \hat{\beta}_2$, respectively, with $\hat{\beta}_1 \neq \hat{\beta}_2$, then it follows that

$$S(\bar{\varepsilon}) \geq S_1(\bar{\varepsilon}_1) + S_2(\bar{\varepsilon}_2) \tag{35a}$$

or

$$\mathcal{Z}(\bar{\varepsilon}) \leq \mathcal{Z}_1(\bar{\varepsilon}_1)\mathcal{Z}_2(\bar{\varepsilon}_2) \tag{35b}$$

which follows from the identity (23).

Inequality (35a) implies that the entropy of the composite system cannot be less than the entropies of the subsystems. From the definition of the Helmholtz free energy (25) and inequality (35a) it follows that

$$\begin{aligned} \hat{\beta}\mathcal{F}(\hat{\beta}) &= \hat{\beta}\bar{\varepsilon} - S(\bar{\varepsilon}) \\ &\leq \hat{\beta}\bar{\varepsilon}_1 - S(\bar{\varepsilon}_1) + \hat{\beta}\bar{\varepsilon}_2 - S(\bar{\varepsilon}_2) \\ &= \hat{\beta}\{\mathcal{F}_1(\hat{\beta}) + \mathcal{F}_2(\hat{\beta})\} \end{aligned} \tag{36}$$

The free energy of the composite system cannot be greater than the free energies of the subsystems. The equality sign applies to the case where both subsystems are at the same temperature [cf. inequality (51)].

We now consider system 1 as the small system and system 2 as the reservior such that $\varepsilon_1 \ll \varepsilon$. The probability

$$\Pr(a < \varepsilon_1 < b) = \frac{1}{\Omega(\bar{\varepsilon})} \int_{a<\varepsilon_1<b} \Omega_1(\varepsilon_1)\Omega_2(\bar{\varepsilon} - \varepsilon_1)\, d\varepsilon_1 \tag{37}$$

for the random quantity ε_1 to lie between any two values a and b, with $a < b$, implies that [Khinchin, 1949, formula (27)]

$$p_1(\varepsilon_1) = \frac{\Omega_1(\varepsilon_1)\Omega_2(\bar{\varepsilon} - \varepsilon_1)}{\Omega(\varepsilon)} \tag{38}$$

is the pdf for the random variable ε_1. The classical derivation of the canonical form of the pdf for $p_1(\varepsilon_1)$ (Blanc-Lapierre and Tortrat, 1956) involves a Taylor series expansion of $\Omega_2(\bar{\varepsilon} - \varepsilon_1)$ about $\bar{\varepsilon}$, and using the fact that $\varepsilon_1 \ll \bar{\varepsilon}$ results in

$$p_1(\varepsilon_1) \sim \frac{1}{\mathcal{Q}_1(\hat{\beta})} \Omega_1(\varepsilon_1) \exp(-\hat{\beta}\varepsilon_1) \tag{39}$$

The derivation has also employed the approximate relation

$$\frac{\Omega_2(\bar{\varepsilon})}{\Omega(\bar{\varepsilon})} = \frac{\omega_2(\bar{\varepsilon})}{\omega(\bar{\varepsilon})} \exp[S_2(\bar{\varepsilon}) - S(\bar{\varepsilon})] \sim \frac{\mathcal{Q}_2(\hat{\beta})}{\mathcal{Q}(\hat{\beta})} = \frac{1}{\mathcal{Q}_1(\hat{\beta})} \tag{40}$$

and (Gibbs, 1902)

$$\frac{\partial \log \Omega_2(\bar{\varepsilon})}{\partial \varepsilon} = \hat{\beta} \tag{41}$$

The approximate relation (40) implies that $\omega_2(\bar{\varepsilon}) \sim \omega(\bar{\varepsilon})$, while, in view of (9) and (27b), equation (41) implies that the prior density $\omega_2(\bar{\varepsilon})$ is essentially a constant, which is not true in general.

Rather, consider the prior density of subsystem 1, which, according to (27a), can be written as

$$\omega_1(\varepsilon_1) = \exp[\log \mathcal{Z}_1(\varepsilon_1)]\, \Omega_1(\varepsilon_1) \tag{42}$$

Developing $\log \mathcal{Z}_1(\varepsilon_1)$ in a Taylor series expansion about $\bar{\varepsilon}_1$ gives

$$\log \mathcal{Z}_1(\varepsilon_1) = \log \mathcal{Z}_1(\bar{\varepsilon}_1) + \frac{\partial \log \mathcal{Z}_1(\bar{\varepsilon}_1)}{\partial \varepsilon} (\varepsilon_1 - \bar{\varepsilon}_1) + O(1) \tag{43}$$

Notice that the first-order term is a quantity of the order of magnitude of $O\sqrt{n}$, which dominates over the remainder as $n \to \infty$, we get $\log \mathcal{Z}_1(\varepsilon_1) \simeq -\hat{\beta}\varepsilon_1 - \log \mathcal{Q}_1(\hat{\beta})$ when (9) and (10) are introduced in (43). Hence,

expression (42) can be written in the canonical form

$$\omega_1(\varepsilon_1) \simeq \Omega_1(\varepsilon_1) \frac{\exp(-\hat{\beta}\varepsilon_1)}{\mathcal{Q}_1(\hat{\beta})} \tag{44}$$

which allows us to identify the prior $\omega_1(\varepsilon_1)$ with the pdf $p_1(\varepsilon_1)$ given by the canonical form (39).

Symmetry considerations between the Gibbs and Bayes representations lead us to consider (29a) written in the form

$$\omega_1(\beta) = \exp[\log \mathcal{Q}_1(\beta)]\, \Omega_1(\beta) \tag{45}$$

as the pdf for the random quantity β. Expanding $\log \mathcal{Q}_1(\beta)$ in a Taylor series about the most likely value of the intensity $\hat{\beta}$ yields

$$\log \mathcal{Q}_1(\beta) = \log \mathcal{Q}_1(\hat{\beta}) + \frac{\partial \log \mathcal{Q}_1(\hat{\beta})}{\partial \beta}(\beta - \hat{\beta}) + O(1) \tag{46}$$

Since $(\beta - \hat{\beta})$ is a quantity whose order of magnitude is the order of $1/\sqrt{n}$, the first-order term, which is $O\sqrt{n}$, dwarfs the remainder as $n \to \infty$. With the aid of the likelihood equation (6), or equivalently (26), and equation (10), expression (46) reduces to $\log \mathcal{Q}_1(\beta) \simeq -\beta\bar{\varepsilon}_1 - \log \mathcal{Z}_1(\bar{\varepsilon}_1)$, and introducing this into (45) gives

$$\omega_1(\beta) \simeq \Omega_1(\beta) \frac{\exp(-\beta\bar{\varepsilon}_1)}{\mathcal{Z}_1(\bar{\varepsilon}_1)} \tag{47}$$

as the canonical pdf for the random intensity β. Expression (41) is the Bayesian analog of the Gibbsian relation (44) and has an equally important role. In fact, the pdf $p_1(\beta)$ for the random quantity β can be written in the form

$$p_1(\beta) = \frac{\Omega_1(\beta)\Omega_2(\beta)}{\Omega(\beta)} \sim \Omega_1(\beta) \frac{\exp(-\beta\bar{\varepsilon}_1)}{\mathcal{Z}_1(\bar{\varepsilon}_1)} \tag{48}$$

analogous to expression (38) for the pdf $p_1(\varepsilon_1)$, by noticing that $\omega_2(\beta) \sim \omega(\beta)$ (Khinchin, 1949, p. 91) as $n \to \infty$ and approximating the partition function by (44). The condition $\omega_2(\beta) \sim \omega(\beta)$ implies that the heat capacity of subsystem 2 is approximately equal to the total heat capacity of the composite system.

In the Gibbs formulation, the energy of the subsystem is a random variable whose distribution depends upon a fixed but unknown constant representing the state of nature, while in the Bayes representation the energy of the subsystem is fixed and the conjugate intensitive quantity is the random variable that at equilibrium has the same value in each of the subsystems. Therefore, the method of composite systems, based on the exchange of

additive invariants (Carathéodory, 1909; Landsberg, 1956), is more readily adaptable to the Gibbs formulation than to the Bayes one and this may in part explain why historically the former has preceded the latter.

Let us now reverse the process and consider a composite system initially characterized by an intensity β. We do work on the system by using a thermal engine to create a temperature difference such that $T_2 > T_1$. Subsystems 1 and 2 are brought into a state characterized by the intensities $\hat{\beta}_1 = \hat{\beta} - 1/2\Delta\beta$ and $\hat{\beta}_2 = \hat{\beta} + 1/2\Delta\beta$, respectively, where $\Delta\beta = \hat{\beta}_2 - \hat{\beta}_1 < 0$. Thermodynamic additivity no longer holds and, in particular, we must replace (33) by

$$\begin{aligned} \mathcal{Q}_1(\hat{\beta}_1)\mathcal{Q}_2(\hat{\beta}_2) &= \frac{\exp(\hat{\beta}_1\varepsilon_1 + \hat{\beta}_2\varepsilon_2)}{\mathcal{Z}_1(\varepsilon_1)\mathcal{Z}(\varepsilon_2)} \\ &\leq \frac{\exp[\varepsilon_1(\hat{\beta} - 1/2\Delta\beta) + \varepsilon_2(\hat{\beta} + 1/2\Delta\beta)]}{\mathcal{Z}(\varepsilon)} \\ &= \mathcal{Q}(\hat{\beta}) \exp(1/2\Delta\varepsilon\Delta\beta) \leq \mathcal{Q}(\hat{\beta}) \end{aligned} \tag{49}$$

where $\Delta\varepsilon = \varepsilon_2 - \varepsilon_1$. The first inequality in (49) is a consequence of (36b), while the second inequality follows from the fact that Carnot's principle is satisfied.

Since $\hat{\beta}_1 > \hat{\beta}_2$, there will be a heat transfer from subsystem $2 \to 1$. In the presence of a heat flow without any work being done, we have $\varepsilon_2 - \varepsilon_1 = \delta Q > 0$. The total entropy change will therefore be given by

$$dS_{\text{tot}} = \left(\frac{1}{T_1} - \frac{1}{T_2}\right)\delta\mathcal{Q} = -\Delta\varepsilon\ \Delta\beta \geq 0 \tag{50}$$

which, when multiplied by the lower temperature T_1 is commonly referred to as the disipation.

If $\hat{\beta}_1$ and $\hat{\beta}_2$ are the values for which the functions S_1 and S_2 reach a minimum, while S has a minimum at $\hat{\beta}$, then inequality (49) can be expressed as

$$\hat{\beta}_1\mathcal{F}_1(\hat{\beta}_1) + \hat{\beta}_2\mathcal{F}_2(\hat{\beta}_2) \geq \hat{\beta}\mathcal{F}(\hat{\beta}) \tag{51}$$

which is actually what is implied by inequality (37). Although equation (34) is a macroscopic thermodynamic relation, $S = S(\bar{\varepsilon})$, the canonical entropy $S = S(\hat{\beta})$ has been implicitly used in going from equality (34) to inequality (35a). But implicit in inequality (35a) is the difference in temperatures of the two subsystems, which is not reflected in the inequality (36). Notwithstanding the fact that the thermodynamic entropy $S(\bar{\varepsilon})$ and the canonical entropy $\mathcal{S}(\hat{\beta})$ are related to one another by the likelihood equation $\hat{\beta} = \beta(\bar{\varepsilon})$, there is not a complete harmony between the two when it comes to extremum principles. This is one important advantage of the

Bayes representation, in which the fundamental thermodynamic dependences are preserved.

4. ON THE ASYMPTOTIC EQUIVALENCE OF THE MAXIMUM-LIKELIHOOD AND MOMENTS METHODS

Asymptotic approaches, such as the Darwin–Fowler method for deriving the canonical Gibbs density, rely on the fact that the most probable value and average values coincide as $n \to \infty$. The asymptotic equivalence between the maximum-likelihood estimate of the parameter β that is obtained from the likelihood equation (6) and its average value given by (8) will allow us to obtain an explicit expression for the prior $\omega(\beta)$. This will show that some of the most common distributions of statistical mechanics sustain Jeffreys' choice of the prior pdf given by (11). Alternatively, by assuming that Jeffreys' choice of the prior is valid, we will obtain an asymptotic equivalence between the maximum-likelihood estimate and the average value.

As a first example, consider an ideal monatomic gas whose partition function per particle is $\log \mathcal{Z}(\beta) \propto -\frac{3}{2}\log\beta$. The maximum-likelihood estimator found from the likelihood equation (6) is $\hat{\beta} = 3/2\bar{\varepsilon}$. The logarithm of the asymptotic expression (2.1) can be written as

$$\log \mathscr{L}(\bar{\varepsilon}) \sim -S(\hat{\beta}) + \log \omega(\hat{\beta}) - \frac{1}{2}\log \frac{\partial^2 S(\hat{\beta})}{\partial \beta^2} \tag{52}$$

where we have made the identification $\mathscr{L}(\hat{\beta}\,|\,\bar{\varepsilon}) = -S(\hat{\beta}) = \frac{3}{2}\log\hat{\beta} + \text{const.}$ Differentiating (52) a single time gives

$$\frac{\partial \log \mathscr{L}(\bar{\varepsilon})}{\partial \bar{\varepsilon}} = -\frac{3}{2\bar{\varepsilon}^2}\left[\frac{d \log \omega(\hat{\beta})}{d\beta} + \frac{5}{2\hat{\beta}}\right] \tag{53}$$

which, according to (9), is equal to $-\hat{\beta}$. We thus obtain the prior pdf as

$$\omega(\hat{\beta}) \propto 1/\hat{\beta} \propto [\mathscr{I}(\hat{\beta})]^{1/2}$$

which is none other than Jeffreys' (1961) improper pdf (11) for $\hat{\beta}$ based on the invariance property that the prior be invariant with respect to powers of $\hat{\beta}$. Alternatively, if we had chosen the prior pdf according to (11), then we would have obtained the asymptotic equivalence of the most probable and average values of the parameter.

Moreover, since (17) applies at thermodynamic equilibrium, $\omega(\hat{\beta}) \propto 1/\omega(\bar{\varepsilon})$, so that $\omega(\bar{\varepsilon}) \propto 1/\bar{\varepsilon}$. Then relation (27b) between the prior density and the structure function gives $\Omega(\bar{\varepsilon}) \propto (\bar{\varepsilon})^{1/2}$, which is the correct form of the structure function for a system with three degrees of freedom (Perrin, 1939). Likewise, from (29b) we find that $\Omega(\hat{\beta}) \propto (\hat{\beta})^{1/2}$.

Furthermore, the variance has the value predicted by the Cramér-Rao lower bound. Differentiating (53) gives

$$\frac{\partial^2 \log \mathscr{Z}(\bar{\varepsilon})}{\partial \bar{\varepsilon}^2} = \frac{3}{2\bar{\varepsilon}^2} = \left[\frac{\partial^2 \log \mathscr{Q}(\hat{\beta})}{\partial \hat{\beta}^2}\right]^{-1}$$

or $\sigma^2(\hat{\beta}) = \sigma^{-2}(\bar{\varepsilon})$.

As a second example, we consider the harmonic oscillator with frequency ν. The partition function is $\mathscr{Q}(\beta) = 1/\sinh(\frac{1}{2}h\nu\beta)$, where h is Planck's constant. The likelihood equation gives

$$\hat{\beta} = \frac{2}{h\nu} \coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right) \tag{54}$$

as the maximum-likelihood estimate of the parameter β. The asymptotic form of the moment-generating function ((5) is

$$\mathscr{Z}(\bar{\varepsilon}) \sim \left[\frac{2\pi}{\mathscr{I}(\hat{\beta})}\right]^{1/2} \exp\left[-\left(\frac{2}{h\nu}\bar{\varepsilon}\right)\coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right)\right]$$
$$\times \sinh\left[\coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right)\right]\omega(\hat{\beta}) \tag{55}$$

where

$$\mathscr{I}(\hat{\beta}) = \left(\frac{1}{2}h\nu\right)^2 \operatorname{csch}^2\left[\coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right)\right]$$

is the Fisher information. Using the criterion pdf $\omega(\hat{\beta})$ in (55) must be chosen such that

$$\frac{\partial \log \mathscr{Z}(\bar{\varepsilon})}{\partial \bar{\varepsilon}} = -\frac{2}{h\nu}\coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right) + \frac{1}{1-[(2/h\nu)\bar{\varepsilon}]^2}\left[\left(\frac{2}{h\nu}\right)^2\bar{\varepsilon} + \frac{\partial \log \omega(\hat{\beta})}{\partial \hat{\beta}}\right] \tag{56}$$

is equal to the negative of the maximum-likelihood estimate (54). This gives again expression (11) for the prior pdf. And differentiating (56) gives

$$\frac{\partial^2 \log \mathscr{Z}(\bar{\varepsilon})}{\partial \bar{\varepsilon}^2} = \left(\frac{2}{h\nu}\right)^2 \sinh^2\left(\frac{2}{h\nu}\hat{\beta}\right) = \left[\frac{\partial^2 \log \mathscr{Q}(\hat{\beta})}{\partial \hat{\beta}^2}\right]^{-1}$$

showing that the distribution has the minimum variance predicated by the Cramér-Rao lower bound. This is true of all equilibrium distributions where the conjugate variables are perfectly negatively correlated.

In the high-temperature limit we have

$$[\mathscr{I}(\hat{\beta})]^{1/2} = \left(\frac{2}{h\nu}\right)^{-1} \operatorname{csch}\left[\coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right)\right] \sim \left[\frac{2}{h\nu}\coth^{-1}\left(\frac{2}{h\nu}\bar{\varepsilon}\right)\right]^{-1} = \frac{1}{\hat{\beta}}$$

giving back the ideal gas result, for which Jeffreys' second rule and its generalization coincide. Therefore, in the high-temperature limit or, in general, for a monatomic ideal gas we have that $\bar{\varepsilon}\hat{\beta}$ is constant and

$$\frac{d\bar{\varepsilon}}{\bar{\varepsilon}}+\frac{d\hat{\beta}}{\hat{\beta}}=0$$

Since $\bar{\varepsilon}$ is capable of taking any value from 0 to ∞, we take its prior density $\omega(\bar{\varepsilon})\, d\bar{\varepsilon} \propto d\bar{\varepsilon}/\bar{\varepsilon}$. And, since the same is true of $\hat{\beta}$, we have two consistent statements of the same form, which would not have been true had we chosen the Bayes–Laplace rule $\omega(\bar{\varepsilon})\, d\bar{\varepsilon} \propto d\bar{\varepsilon}$. And for any other power n, $d\bar{\varepsilon}/\bar{\varepsilon}$ and $d\bar{\varepsilon}^n/\bar{\varepsilon}^n$ are always proportional. This invariance property is characteristic of Maxwell–Boltzmann statistics, for which $\Omega(\bar{\varepsilon}) \propto 1/\Omega(\hat{\beta})$.

As we have seen, the invariance property also holds for Bose particles in the high-temperature limit. However, the same is not true for Fermi particles. Consider a Fermi oscillator with two levels: 0 and ε_0. The partition function is

$$\mathcal{Q}(\beta)=1+e^{-\beta\varepsilon_0}$$

which gives

$$\mathcal{L}((\beta\,|\,\varepsilon))=-\beta\varepsilon-\log(1+e^{-\beta\varepsilon_0})$$

as the expression for the log-likelihood function. The likelihood equation (6) gives the maximum-likelihood estimator as

$$\hat{\beta}=\frac{1}{\varepsilon_0}\log\left(\frac{\varepsilon_0}{\bar{\varepsilon}}-1\right) \tag{57}$$

The Fisher information is given by

$$\mathcal{I}(\hat{\beta})=-\frac{\partial^2\mathcal{L}(\hat{\beta}\,|\,\bar{\varepsilon})}{\partial\beta^2}=\bar{\varepsilon}^2\exp(\hat{\beta}\varepsilon_0)$$

where we recall that the use of the expectation makes no difference in random samples of any fixed size taken from an exponential family. The logarithm of the asymptotic expression for the moment-generating function $\mathcal{Z}(\varepsilon)$ is

$$\log\mathcal{Z}(\bar{\varepsilon})\sim-\frac{\bar{\varepsilon}}{\varepsilon_0}\log\left(\frac{\varepsilon_0}{\bar{\varepsilon}}-1\right)-\log\left(\frac{\varepsilon_0}{\varepsilon_0-\bar{\varepsilon}}\right)+\log\omega(\hat{\beta})-\frac{1}{2}\log\mathcal{I}(\hat{\beta})$$

Setting the first moment equal to the negative of the maximum-likelihood estimate, (57), we obtain the prior pdf as

$$\omega(\hat{\beta})\propto[\mathcal{I}(\hat{\beta})]^{1/2}=\frac{\varepsilon_0}{2\cosh(\frac{1}{2}\hat{\beta}\varepsilon_0)}$$

In the high-temperature limit, we find that $\omega(\hat{\beta}) \propto [\mathcal{J}(\hat{\beta})]^{1/2} = \text{const}$ and there is no invariance property for the prior pdf of the Fermi oscillator. This, however, is precisely the Bayes–Laplace rule, which Jeffreys (1961) considers as an unacceptable representation of the ignorance concerning the value of the parameter.

We have shown that the prior pdf's of the temperature for some of the most common statistical ensembles are *improper* "uniformative" pdf's. From the data acquired through observations on the energy, we are able to make a better "guess" of the conjugate, intensive variable which is described by a *proper* pdf. The initial state of ignorance is to be attributed to the isolated nature of the system and, in order to define a temperature at all, we must suppose it to have been in thermal contact with a thermostat at some very distant time in the past (Mandelbrot, 1962). However, we are in no way restricted to isolated systems and uniformative priors. For open systems, maintained in nonequilibrium states by external constraints, we expect to have informative priors.

In a state of equilibrium, the conjugate thermodynamic variables are perfectly linear and negatively correlated. According to regression theory, if we want to predict the value of the random variable β from values of the random variable ε, we conclude that at equilibrium the error of linear prediction is zero. This means that the correlation coefficient will be greater than its equilibrium value, -1, for nonequilibrium states and consequently the conjugate thermodynamic variables will only fluctuate about their equilibrium equations of state. If the unobserved random "error" or "disturbance" is due to thermal fluctuations that are modeled as Brownian motion, we can conclude that such forms of disturbances have no effect at equilibrium.

ACKNOWLEDGMENTS

This work was supported in part by the Italian Ministry of Public Instruction and the Consiglio Nazionale delle Richerche.

REFERENCES

Barnett, V. (1973). *Comparative Statistical Inference*, p. 131, Wiley, London.

Blanc-Lapierre, A., and Tortrat, A. (1956). In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pp. 145–170.

Bunimovich, L. A., and Sinai, Ya. G. (1981). *Communications in Mathematical Physics*, **78**, 479.

Carathéodory, C. (1909). *Mathematical Annalen*, **67**.

Carathéodory, C. (1925). *Sitzungber. Preuss. Akal. Wiss.* **1925**, 39.

Chernoff, H. (1956). *Annals of Mathematical Statistics*, **27**, 1.

Fisher, R. A. (1922). *Phylosophical Transactions of the Royal Socity of London A*, **222**, 309.

Fisher, R. A. (1925). *Proceedings of the Cambridge Philosophical Society*, **22**, 700.

Fowler, R. H. (1936). *Statistical Mechanics*, 2nd ed., Cambridge University Press, Cambridge.
Gauss, K. F. (1963). *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, pp. 253-259, Dover, New York.
Gibbs, J. W. (1980). *Elementary Principles in Statistical Mechanics*, Scribner, New York (Reprinted, Yale University Press, New Haven, 1948).
Gnedenko, B. V., and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Reading, Massachusetts.
Huang, K. (1963). *Statistical Mechanics*, pp. 206-213, Wiley, New York.
Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Chapter 3, Clarendon, Oxford.
Jeffreys, H. (1973). *Scientific Inference*, 3rd ed., p. 31, Cambridge University Press, Cambridge.
Khinchin, A. I. (1949). *Mathematical Foundations of Statistical Mechanics*, p. 37, Dover, New York.
Kullback, S. (1959). *Information Theory and Statistics*, p. 18, Dover, New York.
Landau, L. D., and Lifshitz, E. M. (1969). *Statistical Physics*, 2nd ed., p. 353, Pergamon Press, Oxford.
Landsberg, P. T. (1956). *Reviews of Modern Physics*, **28**, 363.
Lavenda, B. H. (1987a). *International Journal of Theoretical Physics*, **26**, 1069.
Lavenda, B. H. (1987b). On the phenomenological basis of statistical thermodynamics, *Journal of the Physics and Chemistry of Solids*, in press.
Lavenda, B. H., and Scherer, C. (1987a). The role of statistical inference in equilibrium and nonequilibrium thermodynamics, *Rivista Nuovo Cimento*, in press.
Lavenda, B. H., and Scherer, C. (1987b). The statistical inference approach to generalized thermodynamics: I. Statistics and II. Thermodynamics, *Nuovo Cimento B*, in press.
Lindley, D. V. (1970). *Introduction to Probability and Statistics*, Part 2, p. 139, Cambridge Univeristy Press, Cambridge.
Mandelbrot, B. (1956). *IRE Transactions on Information Theory*, **IT-2**, 190.
Mandelbrot, B. (1962). *Annals of Mathematical Statistics*, **33**, 1021.
Mandelbrot, B. (1964). *Journal of Mathematical Physics*, **5**, 164.
Perring, F. (1938). *Mécanique Statistique Quantique*, Chapter 3, Section 12, Gauthier-Villars, Paris.
Planck, M. (1954). *Treatise on Thermodynamics*, 3rd ed., Dover, New York.
Savage, L. J. (1962). In *The Foundations of Statistical Inference*, p. 15, Methuen, London.
Sirovich (1971).
Tisza and Quay (1963).
Szilard, L. (1925). *Zeitschrift für Physik*, **32**, 753.