Affine Structures and Stochastic Thermodynamics on the Space of Measures

Lowell F. Thompson

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:
Hong Qian, Chair
Bernard Deconinck
Eric Shea-Brown

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Affine Structures and Stochastic Thermodynamics on the Space of Measures

Lowell F. Thompson

Chair of the Supervisory Committee:
Professor Hong Qian
Applied Mathematics

Kolmogorov's theory of probability emphasizes a given state space Ω and a given probability measure $\mathbb{P}$, then constructs the entire calculus of measurable functions $X : \Omega \to \mathbb{R}$. From this perspective, the properties and dynamics of given families of probability measures are viewed as technical subjects within the general theory. In this work, we show that two fundamental concepts from statistical physics - *entropy* and *energy* - are themselves stochastic objects when one considers change of measure to be the natural representation of real-world dynamics. A relationship between thermodynamics and probability theory is formulated in terms of large deviation principles and affine structures on the space of measures.

TABLE OF CONTENTS

Chapter 1
INTRODUCTION

1.1 Probability, statistics, and statistical physics

For the past three centuries, applied mathematics has rested on the pillar of differential equations. The resulting techniques have been extremely successful, but have also met formidable challenges, particularly when faced with complicated behaviors and complex systems such as nonlinear chaotic dynamics and extremely high-dimensional systems like the atomic motions in a protein molecule. In recent years, the mathematical theory of probability, along with that of time-dependent stochastic processes, has emerged as a new workhorse of applied mathematics [31, 33].

In physics and chemistry, the theory of statistical dynamics has provided the foundation for our current understanding of matter as a high-dimensional system in terms of Newtonian mechanical particles. Since its early development by L. Boltzmann and J.W. Gibbs in the late nineteenth century, this field known as *statistical thermodynamics* has relied - at its core - on classical mechanical ideas. In the past few decades, however, surprising discoveries have led to the suggestion that "the essence of a thermodynamic description is not found in its connection to conservation laws, microscopic reversibility, or the equilibrium state relation they entail, despite the central role those play." [39] Furthermore, it has been hypothesized that these fundamental Newtonian concepts and behaviors may well be the consequences of a probabilistic, entropic description of space, time and numbers. More precisely, while the theories of probability and statistics provide a *kinematic* description of stochastic behavior, Gibbs' theory of statistical ensembles provides an additional layer of description that encompasses the notions of energies, conservation laws, equations of

state and thermodynamics itself [34]. One can think of the physicists' approach as a third alternative to the frequentist and Bayesian schools. It combines the ideas of both schools and with the celebrated limit theorems of probability: The law of large numbers and the large deviation principles [11].

The concept of entropy figures prominently in this new, "de-mechanized" theory of statistical thermodynamics, albeit extended far beyond that of Gibbs' in equilibrium statistical mechanics and Shannon's in information theory. Entropy provides a mathematical description for the change from one probability space to another - from $(\Omega, \mathcal{F}, \mathbb{P}_1)$ to $(\Omega, \mathcal{F}, \mathbb{P}_2)$. This change can be quantified in terms of the Radon-Nikodym derivative as $\ln\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega)\right)$, which is itself a random variable. Recent work in nonequilibrium stochastic thermodynamics has shown that fluctuation theorems and the Jarzynski-Crooks equality naturally arise from the mathematics for change in probability measures [47]. In concert with recent developments in optimal transport and nonlinear partial differential equations, these ideas also point towards a formulation of thermodynamics in terms of "the space of probability measures" and the theory of information geometry [2]. It is interesting to note that Fisher's information, along with the concepts of sufficient statistics, Fisher-Neyman factorization theorem, and Pitman-Koopman-Darmois exponential families, play a central role in the theory of information geometry and have also long been argued by B. B. Mandelbrot [29] as key elements of yet another statistical approach to classical thermodynamics pioneered by L. Szilard [41].

This dissertation explores the mathematical foundations of thermodynamics through spaces of measures, both from a general geometric point of view and through the simpler idea of affine structures and their associated tangent spaces [14].

1.2 Modeling complex systems

At the core of mathematical modeling of real world phenomena is a choice of mathematical representation. Since the birth of calculus and Newtonian mechanics, differential equations representing the *spatial motions* of mechanical particles have been a dominant choice. However, in a wide variety of complex systems, the notion of "counting heads" actually provides

a more natural representation. We shall offer some observations here on the mathematical modeling of complex systems over a wide variety of scales through some emblematic examples:

- An ideal gas can be modeled on the finest level as a classical mechanical system. (This is, by assumption, an exact description of the system.) If the system is large enough, we can use the laws of classical (equilibrium) thermodynamics (e.g., $PV = NRT$) to present the system in terms of key "macroscopic variables", P, V, N, and T, which stand for pressure, volume, number of particles and temperature, respectively. We can also employ mesoscale descriptions like small system thermodynamics [28].

- A chemical reaction network could, in principle, be modeled as an enormous quantum mechanical system. In practice, the smallest scale one actually uses is a large mechanical system of molecules combined with some stochastic reactions. We can also use mesoscale descriptions like a chemical master equation. If the system is large enough, we can use deterministic macroscale models like the law of mass action.

- Gene regulatory networks are, in principle, just enormous chemical reaction networks, and so one could use models from the previous example to study them. However, no one actually knows all of the chemicals and/or reactions involved, and so we are forced to use abstractions such as stochastic gene networks. As in the previous examples, if the system is large enough then we can use macroscale deterministic gene network models instead.

All of these examples share a common theme. They can (at least in principle) be modeled at a more refined level in a fundamentally correct manner, but the fine level model is so large that it is unwieldy. In order to simplify these models, one can perform some sort of coarse-graining procedure to create a mesoscale model. These mesoscale models are lower dimensional (by design), but inevitably become stochastic even if the finer description is

deterministic. Finally, if some measure of system size (typically the number of particles or the volume) becomes sufficiently large, then these stochastic mesoscale models can be replaced with deterministic macroscale models. One good example of this is Kurtz's theorem [24].

If we have a fundamental, microscopic description of our system in hand, then we can always construct the coarser models. However, in practice the microscopic level description is generally unavailable. At best, we might know the fundamental description, but the models involved are so large that using them is computationally infeasible. More often, the best available models are many layers of abstraction away from the fundamental microscopic description.

1.2.1 From mechanics to thermodynamics

Classical physicists (and quantum physicists, to a large extent, but we will ignore such complications here) have staked out the path from microscopic to mesoscopic to macroscopic descriptions of many large systems. This is the field of statistical mechanics and thermodynamics proper.

The microscopic description of a system in classical physics is always essentially the same: A collection of N point masses that can be described by specifying the positions and momenta of all the particles in the phase space $\mathbb{R}^{6N}$. The evolution of this system is governed by a Hamiltonian $H : Q \to \mathbb{R}$, where $Q \subset \mathbb{R}^{6N}$.

Classical (equilibrium) thermodynamics furnishes meso- and macroscopic descriptions of a system. The system can be described by specifying a handful of measurable physical quantities with deterministic values (e.g., energy E, entropy S, volume V and temperature T). In small systems, some of these values can fluctuate according to a given distribution, while in large systems the values are fixed since the deviations can be safely neglected.

Statistical mechanics bridges the gap between these sets of descriptions. Thermodynamic variables are now thought of as functionals on the phase space Q, and thermodynamic states are the intersection of the level sets of these functionals. If a functional is conserved by the

Hamiltonian dynamics, then its value is fixed. Otherwise, the stationary distribution on Q (generally furnished by some ergodicity assumptions about H) leads to fluctuations. When the number of particles becomes very large, this distribution approaches a Dirac-δ function and so the thermodynamic variables become deterministic.

Chapter 2

COARSE-GRAINING AND A LARGE DEVIATIONS APPROACH TO THERMODYNAMICS

2.1 Underlying dynamics and coarse-graining

In this work, we will consider some underlying process $\mathfrak{X}(t)$ on a phase space Q. For the moment, we will be quite vague about the nature of this process and even about the space itself. It will suffice to think of Q as $\mathbb{R}^n$ with some appropriate sigma algebra and assume that $\mathfrak{X}$ is stochastic and has a stable stationary distribution[1] and that the system is ergodic on some appropriate set. Note that the assumption of ergodicity is quite important. If $\mathfrak{X}$ is a diffusion process, then this ergodicity arises almost automatically (with some minor assumptions about the diffusion coefficient). On the other hand, if $\mathfrak{X}$ is a deterministic process (such as a Hamiltonian system) then it will at best be ergodic on level sets of the Hamiltonian. In practice, this is often not true. The theory that follows could likely be extended to such non-ergodic systems, but we make no attempt to do so.

It is safe to remain deliberately vague about the nature of $\mathfrak{X}$ because we will never actually measure it directly. Instead, we will only be able to work with some collection of *observables* $X_i : Q \to \mathbb{R}$. In particular, we will limit ourselves to an extremely special set of observables.

Partition Q into n disjoint, measurable sets

$$\omega_1, \ldots, \omega_n \subseteq Q \tag{2.1}$$

and define the maps

$$X_i : Q \to \mathbb{R} \text{ such that } X_i = \mathbb{1}_{\omega_i}. \tag{2.2}$$

[1] We will use the terms *steady state*, *stationary* and *equilibrium* essentially interchangeably. It is particularly important to emphasize that we will use the word *equilibrium* in the dynamical systems sense as something that does not change in time. With the exception of some discussion in chapter 6, the word equilibrium will not have any relation to detailed balance or concepts from physics or chemistry.

Instead of being able to observe $\mathfrak{X}(t)$ directly, we will be limited to observing the values of each $X_i(\mathfrak{X}(t))$. In other words, instead of knowing the exact configuration of our system at any given time (i.e., knowing a point in Q) we only know the *coarse-grained* state ω_i.

In principle, we would like to study measures on Q, but in practice we will only be able to observe frequencies (and their limiting frequencies) on the finite set of coarse-grained states

$$\Omega = \{\omega_1, \ldots, \omega_n\}. \tag{2.3}$$

In particular, if ρ is a measure on Q (with the appropriate sigma algebra), then there is a corresponding measure μ on $\left(\Omega, 2^{\Omega}\right)$ such that

$$\mu(\{\omega\}) \equiv \rho(\omega). \tag{2.4}$$

In addition to measures on Q and Ω, we are also interested in classical thermodynamic quantities. Two quantities of particular importance are *entropy* and *energy*. We will explore the idea of entropy in much more detail later, but the connection between coarse-graining of probability measures and entropy is fairly well established. It is much less obvious how to translate ideas of energy from the phase space Q to the coarse-grained space Ω. To see how we might go about this, we will take inspiration from statistical mechanics and thermodynamics. We will see that energy should be treated as a random variable on Ω. This idea may seem surprising to someone more familiar with mechanics or classical thermodynamics, where it is common to think of the energy as a constant or parameter, or perhaps as a conserved quantity that can be determined from an initial state.

In mechanics, the energy of a system is typically thought of as a Hamiltonian function $H : Q \to \mathbb{R}$ on the phase space. Unlike in our setup, the dynamics of a mechanical system are entirely governed by this Hamiltonian function. It is important to notice that although the Hamiltonian is a function on all of Q, it is conserved by the dynamics of the system, and so any given mechanical system will have a constant energy for all time. In this sense, the energy of a mechanical system is a single number which can be varied by modifying the initial conditions.

In classical (equilibrium) thermodynamics, the state of a system is quite different: A system is described by specifying the values of a handful of functions on phase space. (For instance, one might specify the energy, entropy and volume.) Some of these values are assumed to be fixed, while others are allowed to vary under some specified distribution. Unlike in mechanics, equilibrium thermodynamics does not attempt to describe the trajectory of a system beyond the idea that some values can fluctuate.

Classical statistical mechanics attempts to bridge the gap between these two descriptions. The (micro)state of a system is a point in phase space, just like in classical mechanics, and the microstates evolve according to some Hamiltonian dynamics. In contrast, thermodynamic states are level sets of some specified functions, and so they are entire regions of phase space. Some of these functions (such as the Hamiltonian H) are conserved, and so their values are fixed along one trajectory, but others are not. In order to understand the values of these fluctuating functions, one typically assumes (and in some very special cases one can prove) that the dynamics are ergodic on some appropriate set. This ergodicity assumption means that the fluctuating functions can be described probabilistically through the invariant distribution of the Hamiltonian system. It is important to note that a Hamiltonian system cannot be ergodic on all of Q, but at most on the level sets of H (or some subsets thereof). This means that any description of a Hamiltonian system with fluctuating H requires more than just an ergodicity assumption. There are many historical approaches to this issue. A common tactic is to think of an "ensemble" of systems with different energies rather than one trajectory through phase space. Another approach is to give up on the idea that our underlying system is a Hamiltonian and assume that the dynamics are actually ergodic on all of Q.

Stochastic (equilibrium) thermodynamics takes the latter approach. The system is assumed to evolve through some ergodic stochastic process (equivalent to our $\mathfrak{X}(t)$). From a physical perspective, the noise in this process might arise because the original mechanical system is immersed in a heat bath. In this context, thermodynamic states are random variables on Q. This neatly avoids the question of how energy can fluctuate in a Hamiltonian

system, but leaves open the question of how to translate the random variable H on Q to a "coarse-grained energy" on Ω.

2.2 Sanov's theorem and the potential of entropic force

In order to define an appropriate energy function on Ω, we will need to look in more depth at how an observer would obtain measurements of the process $\mathfrak{X}(t)$. Rather than observing states in Q directly, we are restricted to observing the states of Ω or (equivalently) the values of the observables X_i. If we take many such measurements, we can obtain an empirical frequency of states in Ω. The rate of convergence can be quantified through Sanov's theorem[2] and we will see that Sanov's large deviation rate function provides a basis for the definitions of both entropy and energy.

We will make two important assumptions about this measurement process. First, suppose that the underlying dynamics of $\mathfrak{X}(t)$ operate on a sufficiently fast time scale so that the system reaches equilibrium[3] before each measurement. Second, assume that we have time to take arbitrarily many measurements. In other words, we assume that the underlying dynamics are much faster than a single measurement, which is in turn much faster than the entire experiment. The former assumption can be relaxed, and we will explore the consequences in a later chapter, but the latter assumption is necessary. These conditions are relatively common in practice. For example, chemical reactions are typically observed in solution. By the time a property of the solution (such as the concentration of a particular species) can be measured, it is not unreasonable to assume that the solution has reached equilibrium. In turn, the process of measuring concentration is fast enough that one can take as many measurements as needed throughout the day.

To make the idea of measuring empirical frequencies more precise, consider the random

---

[2]We will actually phrase the mathematics in terms of Cramér's theorem, but the two are equivalent in this setting.

[3]Remember that we assumed in the previous section that $\mathfrak{X}$ had an ergodic steady state.

vector

$$\mathbf{X} = (X_1, \ldots, X_n) = (\mathbb{1}_{\omega_1}(\mathfrak{X}), \ldots, \mathbb{1}_{\omega_n}(\mathfrak{X}))\,. \tag{2.5}$$

where the X_i are the observables defined in (2.2). Suppose that we obtain M i.i.d. samples of $\mathbf{X}$. This is where the separation of time scales assumptions enter in. Since the process $\mathfrak{X}(t)$ is ergodic, we can obtain i.i.d. samples by taking repeated measurements, but these measurements generally need to be taken sufficiently far apart in order to be independent.

Now define the empirical frequencies

$$\pi_{(M)} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{X}_{(i)}. \tag{2.6}$$

and let $\pi^i_{(M)}$ denote the ith component of $\pi_{(M)}$. From the law of large numbers, we know that the measures $\pi_{(M)}$ converge to the law of $\mathbf{X}$, which we will denote by π. In particular, if $\mathfrak{X}(t)$ has a steady-state distribution ρ^{ss} on Q and each $\mathbf{X}_{(i)}$ is drawn from this steady-state distribution, then we can define the frequency π on Ω by

$$\pi^i \equiv \rho^{ss}(\omega_i). \tag{2.7}$$

This frequency necessarily sums to one (because the vector $\mathbf{X}$ always sums to one), and under some reasonable assumptions about ρ^{ss} and our partition Ω, each π^i will be strictly positive.

Let $P(\Omega)$ denote the set of all such frequencies on Ω (i.e., frequencies that sum to one with strictly positive components) and let $A \subseteq P(\Omega)$. If $\pi \notin \bar{A}$, then we must have $\Pr_\pi(\pi_{(M)}) \to 0$ as $M \to \infty$. (The notation $\Pr_\pi$ is used to emphasize that the probability of drawing a certain sequence of $\mathbf{X}_{(i)}$'s, and therefore a certain empirical measure $\pi_{(M)}$, depends on the law of $\mathbf{X}$, which we have denoted π.) We can characterize the rate of this convergence with a large deviation principle (LDP) and a corresponding rate function.

In particular, Cramèr's theorem [11] tells us that

$$-\inf_{\mu\in \mathrm{Int}A} \varphi(\mu) \le \liminf_{M\to\infty} \frac{1}{M} \ln \Pr_\pi(\pi_{(M)} \in A) \le \limsup_{M\to\infty} \ln \Pr_\pi(\pi_{(M)} \in A) \le -\inf_{\mu\in\bar{A}} \varphi(\mu), \tag{2.8}$$

where $\varphi(\mu)$ is the Legendre-Fenchel transform of $\psi(h)$, which is the logarithmic moment generating function defined by

$$\psi(h) = \ln \mathbb{E}_\pi \left[\exp\left(\sum_{i=1}^{n} h_i X_i \right) \right] \tag{2.9}$$

$$= \ln \sum_{i=1}^{n} \pi^i e^{h^i} \tag{2.10}$$

More explicitly, this means that

$$\varphi(\mu) = \sup_h \{ \langle \mu, h \rangle - \psi(h) \} \tag{2.11}$$

$$= \sum_{i=1}^{n} \mu^i \ln \left(\frac{\mu^i}{\pi^i} \right). \tag{2.12}$$

The supremum in (2.11) is attained when

$$h^i = \ln \frac{\mu^i}{\pi^i}, \tag{2.13}$$

but this formula for h^i is not unique. In general, we can add any constant to each h^i (using the same constant for each i) and still attain the same supremum. In other words, h has a gauge freedom with respect to this additive constant. This gauge freedom arises because the observables X_i are not arbitrary - we know that $\sum_{i=1}^{n} X_i = 1$.

There is a dual relationship between ψ and φ. Not only is φ the Legendre-Fenchel transform of ψ, but ψ is also the Legendre-Fenchel transform of φ. That is,

$$\psi(h) = \sup_{\mu \in P(\Omega)} \{ \langle h, \mu \rangle - \varphi(\mu) \}. \tag{2.14}$$

The supremum in equation (2.14) is attained when

$$\mu^i = \frac{\pi^i e^{h^i}}{\sum_{j=1}^{n} \pi^j e^{h^j}}. \tag{2.15}$$

The variables h^i and μ^i are conjugate variables. The latter already has an obvious interpretation in terms of our coarse-grained system - they are the observed probability mass of the coarse-grained states ω_i. The dual variable h^i, however, does not have a standard

meaning in this context. We will once again take inspiration from statistical mechanics to interpret these quantities.

In this context, $\psi(h)$ might appear to be little more than a useful mathematical tool to calculate the truly important function $\varphi(\mu)$. However, in statistical physics, the function $-k_BT\psi\big(h/k_BT\big)$ in (2.10) is of extreme importance - it is known as the Helmholtz free energy. The above result thus, provides a cogent mathematical basis for a thermodynamically meaningful *partition function*. In particular, if one identifies h^i as the negative mechanical energy (in k_BT units), then ψ is the logarithm of the partition function. In classical thermodynamics, $-k_BT\psi\big(h/k_BT\big)$ is also known as the free entropy of the mechanical system.

This gives us a natural way to define[4] energy as a random variable $H : \Omega \to \mathbb{R}$. We require that $H(\omega_i) \propto -h^i$ and interpret (2.14) as "free energy is equal to mean energy minus entropy." Here, $\langle h, \mu\rangle$ should be thought of as the mean energy, and thus h corresponds to an energy *function* on Ω. The constant of proportionality can be used to fix units for energy (and therefore temperature). It will usually be convenient to set this constant to 1 and to set the additive constant to zero, and so

$$H(\omega; \mu, \pi) = -\ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega). \tag{2.16}$$

It is useful to note that

$$\varphi(\mu) = -\sum_{i=1}^{n} \mu^i H(\omega_i; \mu, \pi). \tag{2.17}$$

It is worth taking a moment to summarize the core ideas of the last two sections and to compare and contrast our approach with that of statistical physics. We take many observations of our system through the observables $X_i = \mathbb{1}_{\{\omega_i\}}$, and then average these observations to obtain a probability frequency μ^i. If we take a sufficiently large number of observations, then we expect the observed frequencies to converge to the steady-state probability mass π^i,

[4]This notation is confusing for (at least) two reasons: First, we have introduced extra minus signs to agree with the notation for energy in physics. For historical reasons, physicists and mathematicians use different signs when computing the Legendre-Fenchel transform. Second, many sources refer to the function φ as H, because this is a standard name for the KL divergence. We plan to emphasize the analogy between our H and a Hamiltonian function, and so we have reversed the typical notation.

and the large deviation rate function $\varphi(\mu)$ characterizes the rate of this convergence. One should think of φ as describing just how surprising it would be to observe the frequency μ instead of π. We then use the dual function ψ and the conjugate variables h to define the energy of the system.

The analogous approach in statistical physics is quite similar, but in some sense in reverse. In this context, we know the energy function H in advance. Suppose, for the moment, that we really do know the correct energy function, which matches the definition in (2.16). Instead of using the the observables X_i from (2.2), we use the functions $Y_i = H(\omega_i)\mathbb{1}_{\{\omega_i\}}$. If we take a sufficiently large number of observations, then we expect the average of these Y_i to converge to $\pi^i H(\omega_i)$, and the large deviation rate function for this process $\hat{\varphi}$ (along with its Legendre transform $\hat{\psi}$) characterizes the rate of this convergence. One can think of $\hat{varphi}(x)$ as describing just how surprising it would be to observe the "energies" x instead of the value we expected. The functions $\hat{\varphi}$ and $\hat{\psi}$ are slightly different than our φ and ψ, but the conjugate relationship between energy and frequency remains the same. This approach is well understood in principle (e.g., [21, 49]), but rarely followed in practice.

One reason[5] why this exact method is not often followed is that the given function H, and therefore $\hat{\psi}$, might not actually match equation (2.16). In particular, both (2.16) and the closest appropriate formula for ψ - given in (2.10) - rely on the equilibrium measure π. In general, this measure might not be known in advance. Even if it is known, physicists tend to ignore it and use a more convenient measure like the counting measure #. Fortunately, the corresponding conjugate variables will still provide the correct probability frequencies, even if H is measured "incorrectly".

With this in mind, we will allow a more general definition of energy, where we permit an arbitrary reference measure λ rather than the stationary measure π. That is, in the absence of a mechanical definition, we will *define* the energy of a stochastic system with an observed

[5]There is another important issue that is somewhat trickier to resolve. In practice, it is often not possible to measure the energy of each state individually. Instead, one can only measure the sum of these energies. This changes the problem in a rather fundamental way because we now only have one observable, and therefore only one conjugate variable. We will forgo a discussion of this issue until chapter 6.

steady-state frequency μ and an arbitrary reference measure λ as $H : \Omega \to \mathbb{R}$ such that

$$H(\omega_i) = -\beta_0^{-1}\left(\ln \frac{\mu^i}{\lambda^i} + C\right), \tag{2.18}$$

where β_0 is a constant of proportionality to set units and C is an arbitrary constant. We will typically just set $\beta_0 = 1$ and $C = 0$.

If we insist that λ be a probability measure, then this is a somewhat restrictive definition of H, but in practice λ is usually non-normalized. (In particular, the counting measure is a standard choice.) In that case, for a fixed stationary distribution μ we can arrive at *any* energy function H by the appropriate choice of λ. This might seem to make our definition of energy useless, since we are now allowing arbitrary random variables, but we will often be interested in situations where we have observed multiple different frequencies, but the reference measure is fixed. In that case, this will give a useful way of relating two energy functions.

In the rest of this document, substantial focus will be placed on the concept of the energy defined in (2.18). Beginning in the next chapter, we will take a much deeper look at this definition of energy, but first we will take a brief diversion to discuss frequencies and measures on Ω.

2.3 Spaces of Measures

The frequencies μ obtained through the process described in the previous section look very much like probabilities, but strictly speaking they are just vectors of numbers that sum to one. It is really only in the limiting case where we take infinitely many measurements that we can identify a frequency with a measure. However, we are not ultimately interested in a theory about frequencies - we are interested in a theory about the underlying process that produced these frequencies. We have made the fairly strict assumption that all of our measurments are obtained from an i.i.d. process with some fixed distribution π, and so we can naturally identify the underlying process with the measure π on $\left(\Omega, 2^{\Omega}\right)$.

Suppose that we go through the process described above and obtain an empirical fre-

quency $\mu \neq \pi$. Conditional on seeing this mean frequency, our sequence of observations will *not* look like it came from an i.i.d. π-distributed process. Instead, it will almost surely look like it came from an i.i.d. μ-distributed process. This result is fairly trivial in our context, but is a special case of a formal technique in large deviations theory known as tilting [11]. In this context, μ is called a *tilted measure*. From now on, we will identify the observed frequency μ with the i.i.d. μ-distributed process, and therefore with the measure μ. This means that we will identify the number μ^i with the probability mass $\mu(\{\omega_i\})$. Such an identification has the useful side effect of simplifying a lot of our notation. In particular, ratios of the form μ^i/π^i appear throughout this work. They can now be thought of as Radon-Nikodym derivatives: $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega_i)$. In particular, this means that the energy function $H(\cdot;\mu,\lambda)$ can be written as

$$H(\omega;\mu,\lambda) = -\beta_0^{-1}\left(\ln\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} + C\right) \tag{2.19}$$

The measures μ obtained through this process are necessarily probability measures, but it will often be useful to refer to arbitrary unsigned measures μ on Ω as well. (In particular, we will allow the arbitrary reference measure λ to be non-normalized.)

As a reminder of some terminology: A measure μ on $\left(\Omega, 2^\Omega\right)$ is called *strictly positive* (s.p.) if $\mu(A) = 0$ implies that $A = \emptyset$; it is called *locally finite* (l.f.) if $\mu(\{\omega\}) < \infty$ for every $\omega \in \Omega$; it is called *normalizable* if $\mu(\Omega) < \infty$; and it is a probability measure if $\mu(\Omega) = 1$.

In this document, we will focus entirely s.p., l.f. measures on $\left(\Omega, 2^\Omega\right)$. Of course, many measures on Q correspond to measures on Ω that are not s.p. and/or not l.f. We have chosen to focus on this particular class of measures for mathematical convenience. It is rarely an issue to insist on s.p. measures; if $\mu(\{\omega_i\})$ is zero, then we should simply have removed ω_i from our partition of Q. Insisting on l.f. measures is slightly more problematic, because many interesting measures on Q necessarily correspond to a non-l.f. measure on Ω. (For example, if $Q = \mathbb{R}^n$, then the Lebesgue measure on Q will lead to at least one coarse-grained state with infinite measure.) We will eventually solve this issue by allowing Ω to be infinite.

It is worth taking a moment to address the choice to focus on *measures* on Ω and not simply discuss frequencies, because it is not standard language in the field. For the moment,

assume that $\Omega = \{\omega_1, \ldots, \omega_n\}$ has cardinality $n < \infty$. We can write any s.p., l.f. measure on Ω in the form

$$\mu = \sum_{i=1}^{n} \mu^i \delta_{\omega_i}, \tag{2.20}$$

where each $\mu^i \in (0, \infty)$. These μ^i are exactly the empirical frequencies obtained through our limiting process. Similarly, any n-tuple of positive real numbers μ^i corresponds to a s.p., l.f. measure μ. Moreover, the usual addition of countably additive measures corresponds exactly to vector addition of these n-tuples. It is therefore common to dispense with the terminology of measures and work solely with vectors in $\mathbb{R}^n$. The one caveat to keep in mind is that we are only allowed to use vectors with strictly positive components, so it is important to check that any vector space operations (in particular, subtraction or linear transformations) do not result in a vector with a negative component. Exactly the same concept applies when Ω is countably infinite, except we must replace $\mathbb{R}^n$ with $\mathbb{R}^{\mathbb{N}}$. In the case of infinite Ω, some care is required in choosing a topology on $\mathbb{R}^{\mathbb{N}}$, and the idea that these vectors represent measures is helpful in finding the most useful topology, but otherwise we can essentially forget that the underlying objects are measures. This approach is particularly common in information geometry, where one tends to think of a probability measure as a point $(p^1, \ldots, p^n)$ on the simplex Σ^{n-1}.

Equivalently, one can also frame everything in terms of functions on Ω. In this case, one identifies the measure μ with the function mapping $\omega \mapsto \mu(\{\omega\})$. This is identical to the previous approach, but it helps to illustrate a point that is often left implicit. We are identifying the measure μ with the function $\mathrm{d}\mu/\mathrm{d}\# : \Omega \to \mathbb{R}$, where $\#$ is the counting measure on Ω. That is, instead of working with measures directly, this approach works with densities *with respect to the counting measure*. This is relevant because the counting measure is not always a good choice of reference measure. Similarly, when studying detailed-balanced Markov processes it is frequently more useful to work with densities with respect to the equilibrium measure. Indeed, we already saw this in the previous section. The natural definition of energy that arises from Sanov's theorem involves a density with respect to π,

not the counting measure. Of course, one can correct this while still identifying measures with vectors or functions, but such an identification tends to slightly obscure the issue.

The two above approaches are perfectly valid as long as Ω is countable. However, if Ω is uncountable, then equation (2.20) is no longer valid, and one can certainly not use $\mathrm{d}\mu/\mathrm{d}\#$ for any reasonable measure on Ω. If Ω is a subset of $\mathbb{R}^n$, then it is common to work with the density $\mathrm{d}\mu/\mathrm{d}\lambda$, where λ is the Lebesgue measure, but for more abstract Ω there is no obvious choice of reference measure. More importantly, even if we confine ourselves to subsets of $\mathbb{R}^n$, it is not unusual to encounter measures that are not absolutely continuous with respect to the Lebesgue measure. It is therefore necessary to take much more care in representing measures on uncountable Ω, and it is generally safer to work with measures directly rather than densities. In this document, however, we will work almost exclusively with countable state spaces, and so it is reasonable to ask why we have bothered with this more abstract formalism. The main reason is that we eventually hope to extend these results to applications with uncountable state spaces. In particular, in population biology it is extremely common to represent genotypes or phenotypes as elements of some subset (typically all) of $\mathbb{R}^n$ and to study how the distribution of these types changes over time. Several technical difficulties arise when trying to write the evolution equations for these distributions in terms of densities, but those difficulties can be avoided by working with measure-valued processes instead ([8, 9]). In addition, the solutions to these problems can be constructed by first discretizing the state space and studying the evolution of atomic measures on this discretization ([9, 10, 13]). One can therefore hope that a thorough understanding of measures on a (necessarily countable) set of atoms will be immediately applicable to more complex biological models.

As discussed earlier, we are interested not just in the properties of one measure on the space $\left(\Omega, 2^{\Omega}\right)$, but in the properties of a wide class of measures. Exactly which measures we allow will have far-reaching implications. There are three reasonable candidates:

1. The set $\mathcal{P}(\Omega)$ of all strictly positive probability measures on Ω.

2. The set $\mathcal{M}_N(\Omega)$ of all strictly positive normalizable measures on Ω.

3. The set $\mathcal{M}(\Omega)$ of all strictly positive, locally finite measures on Ω.

In information geometry and probability theory in general, one almost always uses $\mathcal{P}(\Omega)$. This space is also often encountered in studies of population genetics, where the probability measures represent allele frequencies.

In chemistry and biology, one also frequently encounters measures from $\mathcal{M}_N(\Omega)$. For instance, there is no reason to expect that the concentration of various chemicals or the population density of different species will be normalized, but there will generally be some finite total concentration or population density. One can, of course, divide by the total density $\mu(\Omega)$ to obtain a probability measure, but in general the dynamics of these quantities will depend on the non-normalized totals and not just their relative proportions[6], and so we cannot just normalize everything. With that said, it is often the case that the normalization factor $\mu(\Omega)$ is of limited importance at steady-state. That is, transient dynamics might depend heavily on total density, but long-term properties depend only on the normalized proportion of states. In general, outside of the realm of probability one cannot assume that measures are normalized, but one can hope that steady-state properties depend only on the corresponding normalized distribution.

Non-normalizable measures may seem unusual in physical applications, since one expects quantities like population density or species concentration to remain finite, but there are important examples of non-normalizable measures that we don't wish to rule out. One such class of examples arises in statistical mechanics when the Boltzmann distribution is non-normalizable, such as the asymptotic distribution of particle positions under a Lennard-Jones potential [12]. More importantly, it is frequently useful to treat measures as densities with respect to some non-normalizable reference measure. We have already seen that it is common to use the counting measure # in this context, but it has also recently been shown that the second law of thermodynamics can be naturally formulated for non-detailed balanced Markov processes, and a non-normalizable reference measure naturally arises in

[6]For example, density-dependent population models depend heavily on the total population $\mu(\Omega)$, and this quantity is not usually conserved.

this case [44]. As a general rule, one can expect non-normalizable measures to arise in applications in one of two ways: Either as the limiting case of normalizable measures (as with the Lennard-Jones potential) or as natural reference measures.

We will take the following approach: In general, we will study measures and flows μ_t in the space $\mathcal{M}(\Omega)$, but we will devote considerable effort to tracking the normalization constant $\mu_t(\Omega)$ and determining when it becomes infinite. Furthermore, we will insist that the normalization map $\mathfrak{N} : \mathcal{M}_N(\Omega) \to \mathcal{P}(\Omega)$, defined by

$$\mathfrak{N}(\mu)(A) = \frac{\mu(A)}{\mu(\Omega)} \tag{2.21}$$

preserves important steady-state properties of these flows.

For the purposes of this dissertation, we will generally assume that Ω is not only countable but finite. This neatly avoids several topological complications. In particular, there is only one reasonable choice of topology and vector space structure for $\mathcal{P}(\Omega)$ and for $\mathcal{M}(\Omega)$ (shown in appendix A) and we do not have to worry about difficulties with non-normalizable measures. As we have discussed above, though, we believe that it is important to extend these ideas to cover countably infinite state spaces. In such cases, we need to make non-trivial choices about the topology and geometric structure on $\mathcal{P}(\Omega)$, $\mathcal{M}_N(\Omega)$ and $\mathcal{M}(\Omega)$. In particular, it is very important that if the log partition function defined through (2.10) is finite at some measure, then it should also be finite throughout a small neighborhood of that measure as well. Unfortunately, none of the usual topologies on these spaces meet such a requirement. Much of the work to solve these issues has already been done (see [32]), but it is outside the scope of this document. In this work, we will only deal with infinite Ω in chapter 5, and in that section we will avoid any such topological issues. Everywhere else, we will simply assume that Ω is finite.

Chapter 3

POTENTIAL ENERGY OF GENERALIZED GIBBS MEASURES

In section 2.2, we defined the energy of a steady-state measure μ as

$$H(\omega) = -\ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega), \tag{3.1}$$

where λ is some appropriate reference measure. The equilibrium distribution π of our underlying stochastic process $\mathfrak{X}$ naturally arises as a reference measure, but it is also common to choose a more arbitrary reference like $\lambda = \#$ and write this as $H(\omega) = -\ln \Pr(\omega)$ These related definitions of energy have been used before (e.g., [30, 43]), but they are unusual enough that they merit substantial justification. In the rest of this chapter, we will present several arguments for why it is natural to use such a formula for the energy of a coarse-grained stochastic process.

3.1 Equilibrium potential of mean force

It is widely believed that complex physical systems can ultimately be described as mechanical systems of molecules or atoms or other particles interacting through complicated force fields. However, exact formulas for these force fields are not known for even moderately complicated systems. For example, the study of protein folding through molecular dynamics is ultimately just a classical (or perhaps quantum, but such complications are not germane to this discussion) mechanics problem, but the relevant intermolecular forces are not known explicitly. Instead, the field of molecular dynamics relies on approximations to these fields that have been refined over the past fifty years [27]. Such a setting seems ripe for the application of ideas from statistical mechanics, but it is not *a priori* obvious how to apply such ideas with only an approximate and coarse-grained energy function.

To understand this issue, we will first note an important mathematical equality. Consider a product space $\Omega = \Omega_1 \times \Omega_2$ with product measure (on some appropriate sigma algebra) $\mu = \mu_1 \times \mu_2$, and consider a function $H(x) = H(x_1, x_2)$, where $x \in \Omega$ and each $x_i \in \Omega_i$. We have

$$\begin{aligned} Z(\beta) &= \int_\Omega e^{-\beta H(x)} \, \mathrm{d}\mu \\ &= \int_{\Omega_1} \int_{\Omega_2} e^{-\beta H(x_1, x_2)} \, \mathrm{d}\mu_2 \, \mathrm{d}\mu_1 \\ &= \int_{\Omega_1} e^{-\beta \varphi(x_1; \beta)} \, \mathrm{d}\mu_1, \end{aligned} \tag{3.2}$$

where

$$\varphi(x_1; \beta) = -\beta^{-1} \ln \int_{\Omega_2} e^{-\beta H(x_1, x_2)} \, \mathrm{d}\mu_2. \tag{3.3}$$

This means that if we think of φ as a "potential function" for the system in the coarse-grained state space Ω_1 then we obtain the same partition function $Z(\beta)$. That is, $\varphi(x_1; \beta)$ is the energy of the system with a fluctuating x_2 and a fixed x_1. The important caveat in this argument is that the coarse-grained potential depends on β, whereas the original potential H does not.

J.G. Kirkwood provided a very illuminating interpretation of the potential $\varphi(x_1)$ for a continuous space Ω_1 [23]: It is the potential of a "mean force" acting on an equilibrium system which is fixed at x_1. In mathematical terms,

$$\begin{aligned} -\left.\frac{\mathrm{d}\varphi}{\mathrm{d}x_1}\right|_{x_1} &= -\frac{\int_{\Omega_2} \frac{\partial H(x_1, x_2)}{\partial x_1} e^{-\beta H(x_1, x_2)} \, \mathrm{d}\mu_2}{\int_{\Omega_2} e^{-\beta H(x_1, x_2)} \, \mathrm{d}\mu_2} \\ &= -\int_{\Omega_2} \frac{\partial H(x_1, x_2)}{\partial x_1} p^{eq}(x_2 \mid x_1) \, \mathrm{d}\mu_2, \end{aligned} \tag{3.4}$$

where $p^{eq}(x_2 \mid x_1)$ is the conditional equilibrium probability of x_2 given x_1 and the term $-\frac{\partial H(x_1, x_2)}{\partial x_1}$ is the mechanical force in the x_1 direction with a given x_2. That is, $-\mathrm{d}\varphi/\mathrm{d}x_1$ is the mean force on x_1.

In other words, the negative logarithm of of the marginal equilibrium probability distribution for x_1 is the potential of mean force if one chooses $F(\beta) = -\beta^{-1} \ln Z(\beta)$ as the zero

energy reference point:

$$\varphi(x_1) = -\beta^{-1} \ln \int_{\Omega_2} p^{eq}(x_1, x_2)\, \mathrm{d}\mu_2 + F(\beta). \tag{3.5}$$

It is important to keep in mind that this coarse-grained potential φ is a function of β. This is well known in physical chemistry, where one typically creates a statistical mechanical model using a temperature-dependent potential of mean force from the start rather than a mechanical energy function. A simple calculation lets us rewrite

$$\varphi(\cdot;\beta) = \frac{\partial(\beta\varphi)}{\partial\beta} - \frac{1}{\beta}\left(-\frac{\partial\varphi}{\partial(1/\beta)}\right). \tag{3.6}$$

In the context of physical chemistry, the first term represents an energetic component, while the second term represents an entropic component. These notions naturally arise as soon as the potential function varies with temperature.

This line of reasoning suggests that even if one does not have access to a true mechanical energy function, one can still use the tools of statistical mechanics on a coarse-grained system by using a temperature-dependent potential of mean force. This potential of mean force is simply $-\beta$ times the logarithm of the equilibrium probability of a coarse-grained state (with an appropriate choice of gauge). Such reasoning places the logarithm of probability center-stage as an analogue of energy.

3.2 Deterministic correspondence and large deviation rate

Any mathematical representation of reality necessarily includes both stochastic and determenistic elements. As has been pointed out in [17, 18, 19], it is the interaction between these elements that leads to complex behavior and self-organization. Because of this, it is important to be able to "envision" deterministic dynamics that corresponds to some given stochastic dynamics. The typical approach is to look at the given stochastic system in the limit as some system size parameter grows large or small. For instance, the natural parameter for a stochastic differential equation $\mathrm{d}\mathbf{x}(t) = b(\mathbf{x})\,\mathrm{d}t + \epsilon\,\mathrm{d}B(t)$ is the noise strength ϵ; the

natural parameter in classical statistical mechanics is the system's size (or possibly temperature); and the natural parameter in a Dellbrück-Gillespie process is the system's volume [37].

This is a useful approach when such a natural parameter is available, but it is not immediately obvious how one should proceed when there is no obvious analogue of "system size". An increasingly common approach is to use the modal value of a distribution as a "deterministic counterpart" to a given stochastic system. According to this view, a bimodal distribution corresponds to a bistable system [35]. Note that the idea of using the mean dynamics $\langle \mathbf{x}(t) \rangle$ as a deterministic counterpart of a stochastic $\mathbf{x}(t)$ is widely held but mistaken. For a stochastic differential equation, $\langle \, \mathrm{d}\mathbf{x}(t) \rangle \neq b(\langle \mathbf{x}(t) \rangle$ in general. More importantly, while $\langle x(t) \rangle$ is a determenistic function of t, it is *not* a trajectory of any meaningful, self-contained dynamical system. This point is well illustrated by observing the fact that the differential equation describing $\langle \mathbf{x}(t) \rangle$ typically depends on higher moments such as $\langle \mathbf{x}^2(t) \rangle$. Furthermore, for a discrete system, even when the mean is well-defined it does not usually lie in the same space as $\mathbf{x}(t)$.

In [42], we proposed the following "deterministic counterpart" to a random variable $\mathbf{x}$ with probability mass $p^{ss}_{\mathbf{x}}$:

$$\mathbf{x}_\infty = \lim_{\beta \to \infty} \mathbf{x}_\beta, \tag{3.7}$$

where

$$p^{ss}_{\mathbf{x}_\beta}(x) = \frac{p^{ss}_{\mathbf{x}}(x)^\beta}{Z(\beta)}, \tag{3.8}$$

with normalization constant

$$Z(\beta) = \sum_x p^{ss}_{\mathbf{x}}(x)^\beta. \tag{3.9}$$

In that paper, we focused on random variables and their probability mass functions, but the notion easily generalizes to an arbitrary normalizable measure $\mu \in \mathcal{M}_N(\Omega)$ and its density with respect to some other measure $\lambda \in \mathcal{M}(\Omega)$. We can define

$$\mu_{(\infty)} = \lim_{\beta \to \infty} \mu_{(\beta)}, \tag{3.10}$$

where

$$\mu_{(\beta)}(A) = \frac{\int_A \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^\beta \mathrm{d}\lambda}{Z(\beta)} \tag{3.11}$$

for any measurable set $A \subseteq \Omega$, where

$$Z(\beta) = \int_\Omega \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^\beta \mathrm{d}\lambda. \tag{3.12}$$

In [42], we assumed that $\mu \in \mathcal{P}(\Omega)$ and that $\lambda = \#$. We will follow that convention here, since it makes the interpretation of $\mu_{(\infty)}$ slightly simpler, but the mathematics does not substantively change in the general case.

The measure $\mu_{(\infty)}$ is concentrated on a finite number of states (the states with highest density $\mathrm{d}\mu/\mathrm{d}\lambda$). In particular, if $\mathrm{d}\mu/\mathrm{d}\lambda$ is unimodal, then $\mu_{(\infty)}$ is a delta measure on the modal value, which justifies the name "deterministic counterpart". On the other hand, if $\mathrm{d}\mu/\mathrm{d}\lambda$ is multimodal, then there is no unique deterministic counterpart and $\mu_{(\infty)}$ will be non-zero on the (finitely many) modes of μ.

It is worth noting that similar definitions are often introduced formally as analogues to inverse-temperature without any discussion of deterministic correspondence (e.g., [30, 43]). We have devoted so much space to this idea in order to emphasize that it can arise naturally in a study of stochastic systems without any reference to thermodynamic concepts. The scaling factor β can be thought of as a formal method for introducing temperature to a system, and the measure $\mu_{(\infty)}$ can be thought of as a corresponding zero temperature limit, but it is also useful to think of this as a natural feature of any probabilistic system.

To see how this idea of a deterministic counterpart relates to the function $H(\omega) = -\ln \mathrm{d}\mu/\mathrm{d}\lambda(\omega)$, define $E^* = \min_\Omega H(\omega)$ and for any $E \geq E^*$ define

$$D_E = \{\omega \in \Omega \mid H(\omega) = E\} \text{ and} \tag{3.13}$$

$$X_E = \{\omega \in \Omega \mid H(\omega) \leq E\}. \tag{3.14}$$

Finally, let ω_E be an arbitrary element of D_E whenever D_E is nonempty.

For any $E > E^*$, we know that $\mu_{(\beta)}(\Omega \setminus X_E)$ approaches zero as β goes to infinity. We

would like to know how quickly this quantity decays. We can write

$$\mu_{(\beta)}(\Omega \setminus X_E) = e^{-\beta I(E)+o(\beta)}, \tag{3.15}$$

where

$$\begin{aligned} I(E) &= -\lim_{\beta\to\infty} \frac{1}{\beta} \ln \mu_{(\beta)}(\Omega \setminus X_E) \\ &= -\lim_{\beta\to\infty} \ln \frac{\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega_E)\right)^{\beta} \lambda(D_E)}{\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega_{E^*})\right)^{\beta} \lambda(D_{E^*})} \\ &= H(\omega_E) - H(\omega_{E^*}). \end{aligned}$$

In other words, the energy defined by H describes the rate of convergence of μ to its corresponding deterministic system.

3.3 Informational Representation of Stochastic Change

A stochastic phenomenon can be modeled by a probability space $(\Omega, \mathcal{F}, \mu)$. In real applications, however, one can rarely observe events $\omega \in \Omega$ directly. Instead, one makes observations through a (typically quite small) set of random variables $u : \Omega \to \mathbb{R}$. If one is able to make many measurements of u in a short time, or if we assume that the underlying probability measure μ is stationary, then one can also observe the distribution of u, characterized by its cumulative distribution function $F_u : \mathbb{R} \to [0, 1]$.

Suppose that we attempt to calculate the cdf F_u a second time and finds that it has changed to $F_{\tilde{u}} \neq F_u$. There are two possible explanations for this phenomenon: On the one hand, it is possible that we actually measured a different observable $\tilde{u} : \Omega \to \mathbb{R}$ rather than the original u. On the other hand, it is possible that the underlying measure changed from μ to a different measure $\tilde{\mu}$.

In [36], we discuss some of the implications of these different representations of a change of distribution. The mathematics in that paper relies fairly heavily on using $\Omega = \mathbb{R}^n$, and so we will not present it in full here, but one of the key points is that the random variable $H(\omega) = -\ln \mathrm{d}\mu / \mathrm{d}\tilde{\mu}(\omega)$ arises naturally when one switches between these two representations.

Chapter 4

AFFINE STRUCTURES ON $\mathcal{M}(\Omega)$ AND $\mathcal{P}(\Omega)$

We saw in the previous section that the function $H(\omega) = -\ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega)$ should play a key role in any discussion of the relation between thermodynamics and probability. In particular, in section 3.2 we showed that the set of measures $\mu_{(\beta)}$ defined by

$$\mu_{(\beta)}(A) = \frac{\int_A \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^\beta \mathrm{d}\lambda}{\int_\Omega \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^\beta \mathrm{d}\lambda} = \frac{\int_A e^{-\beta H(\omega)}\,\mathrm{d}\lambda}{\int_\Omega e^{-\beta H(\omega)}\,\mathrm{d}\lambda}. \tag{4.1}$$

provide a key link between stochastic and deterministic dynamics. Our theory should reflect the importance of these measures in some mathematical structure. In particular, we can think of $\mu_{(\beta)}$ as being a curve in $\mathcal{P}(\Omega)$ (parametrized by β). We will insist that curves of this form be "straight lines." So far, we have not endowed the space $\mathcal{P}(\Omega)$ (or $\mathcal{M}(\Omega)$) with any mathematical structure. In order to make the notion of "straight" precise, we will need to give some sort of structure. One setting in which this notion makes sense is that of differential geometry. If we endow $\mathcal{P}(\Omega)$ with a smooth atlas and an appropriate affine connection (and, optionally, a Riemannian metric) then we could ensure that the curves $\mu_{(\beta)}$ were geodesics. Ultimately, this is probably the correct choice, particularly when Ω is infinite, but we will soon see that for finite Ω the full machinery of differential geometry is not required. Instead, we can equip $\mathcal{P}(\Omega)$ with an appropriate *affine structure* and ensure that the curves $\mu_{(\beta)}$ are affine curves. This is really just a special case of the aforementioned geometric approach, but in the author's opinion the concept of an affine space provides more useful intuition. Appendix A discusses how to make this idea more rigorous.

In this section, we will provide a basic overview of affine spaces and show how this can be applied to both $\mathcal{P}(\Omega)$ and $\mathcal{M}(\Omega)$. Throughout this section, we will assume that Ω is finite.

4.1 Review of affine spaces

The following is a brief review of the relevant properties of affine spaces. Proofs of the general properties of these spaces, along with a discussion of the relevant topological issues, are included in appendix A. Any proofs that are not included can be found in various textbooks, such as [14] or [38].

Definition 4.1.1 *An affine space is a triple $(M, V, \oplus)$, where M is a set, V is a (real) vector space (with addition and scalar multiplication denoted by $+$ and $\cdot$) and $\oplus : M \times V \to M$ is a function such that the following hold:*

(a) $p \oplus 0 = p$ for any $p \in M$, where 0 denotes the zero element of V.

(b) $(p \oplus x) \oplus y = p \oplus (x + y)$ for any $p \in M$ and $x, y \in V$.

(c) For every $p \in M$, the map $x \mapsto p \oplus x$ is a bijection.

In other words, if we think of V as a group under vector addition, $\oplus$ is a free and transitive group action of V on M.

We will need the map from part c so often that it will be useful to name it:

Definition 4.1.2 *For any $p \in M$, let $\Phi_p : V \to M$ be defined such that*

$$\Phi_p(x) = p \oplus x. \tag{4.2}$$

The definition of an affine space asserts that each Φ_p is bijective, and so $\Phi_p^{-1} : M \to V$ exists.

If we have an affine space $(M, V, \oplus)$, we will refer to M as the *point space* and V as the *vector space*. If we need to distinguish between $+ : V \times V \to V$ and $\oplus : M \times V \to M$, we will call $+$ *vector addition* and $\oplus$ *affine addition*. Our definition gives a sensible way to add two vectors together (vector addition) and to add a vector to a point (affine addition), but not to add two points together.

Although we can't add two points together, we can make a sensible definition for the difference between two points.

Definition 4.1.3 *If $q, p \in M$, then we define $\ominus : M \times M \to V$ such that*

$$q \ominus p = \Phi_p^{-1}(q). \tag{4.3}$$

That is, if $q \ominus p = x$, then $q = p \oplus x$.

This definition of subtraction has some useful properties.

Lemma 4.1.4 *For any $p, q, r \in M$ and any $x, y \in V$, the following are true:*

(a) $p \oplus (q \ominus p) = q$.

(b) $(p \ominus q) + (q \ominus r) = p \ominus r$.

(c) $(p \ominus q) = -(q \ominus p)$.

(d) $(p \oplus x) \ominus (p \oplus y) = x - y$.

Proof (a) This is true by definition. For any $q, p \in M$, we have $q - p = \Phi_p^{-1}(q)$, which is the unique vector x such that $p \oplus x = q$. Therefore, $p \oplus (q - p) = q$.

(b) Choose $p, q, r \in M$. We have

$$\begin{aligned} r \oplus ((p - q) + (q - r)) &= r \oplus ((q - r) + (p - q)) \\ &= (r \oplus (q - r)) \oplus (p - q) \\ &= q \oplus (p - q) \\ &= p. \end{aligned}$$

Therefore, $(p - q) + (q - r) = p - r$, as desired.

(c) From part (b), we know that $(p-q)+(q-p) = (p-p)$. Since Φ_p is a bijection, $p-p = 0$, so we have $(p - q) + (q - p) = 0$. Therefore, $(p - q) = -(q - p)$ as desired.

(d) Define $z = (p \oplus x) - (p \oplus y)$. We have

$$p \oplus x = (p \oplus y) \oplus z = p \oplus (y + z),$$

so $x = y + z$, and therefore $z = x - y$, as desired. $\square$

We have emphasized that, in an affine space, points and vectors are different objects. However, it is also important to note that they are not *too* different. In particular, if we choose one privileged point (or origin) $p \in M$, then the map Φ_p endows M with a canonical vector space structure. That is, we identify the point q with the vector $q \ominus p$.

Lemma 4.1.5 *Choose $p_0 \in M$ and define $+ : M \times M \to M$ and $\cdot : \mathbb{R} \times M \to M$ such that*

$$p + q = \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right), \quad \textit{and} \tag{4.4}$$

$$\alpha \cdot p = \Phi_{p_0}\left(\alpha \Phi_{p_0}^{-1}(p)\right). \tag{4.5}$$

When equipped with the preceding multiplication and addition operations, $\mathcal{M}$ is a vector space and p_0 is the additive identity element. Moreover, this structure preserves affine addition in the sense that

$$\Phi_{p_0}^{-1}(p \oplus x) = \Phi_{p_0}^{-1}(p) + x. \tag{4.6}$$

Proof We need to show the following:

(a) $p + (q + r) = (p + q) + r$ for any $p, q, r \in M$.

(b) $p + q = q + p$ for any $p, q \in M$.

(c) $p + p_0 = p$ for any $p \in M$.

(d) For every $p \in M$, there exists a $q \in M$ such that $p + q = p_0$.

(e) $\alpha \cdot (\beta \cdot p) = (\alpha\beta) \cdot p$ for every $\alpha, \beta \in \mathbb{R}$ and every $p \in M$.

(f) $1 \cdot p = p$ for every $p \in M$.

(g) $\alpha \cdot (p+q) = \alpha \cdot p + \alpha \cdot q$ for every $\alpha \in \mathbb{R}$ and $p, q \in M$.

(h) $(\alpha + \beta) \cdot p = \alpha \cdot p + \beta \cdot p$ for every $\alpha, \beta \in \mathbb{R}$ and $p \in M$.

These are mostly a matter of unravelling definitions and using the vector space structure of V.

(a) Let $p, q, r \in M$. We know that

$$\begin{aligned} p + (q + r) &= p + \Phi_{p_0}\left(\Phi_{p_0}^{-1}(q) + \Phi_{p_0}^{-1}(r)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\Phi_{p_0}^{-1}(q) + \Phi_{p_0}^{-1}(r)\right)\right)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q) + \Phi_{p_0}^{-1}(r)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right)\right) + \Phi_{p_0}^{-1}(r)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right) + r \\ &= (p + q) + r. \end{aligned}$$

(b) Choose $p, q \in M$. We have

$$\begin{aligned} p + q &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(q) + \Phi_{p_0}^{-1}(p)\right) \\ &= q + p. \end{aligned}$$

(c) Let $p \in M$. We know that $\Phi_{p_0}^{-1}(p_0) = 0$, so

$$\begin{aligned} p + p_0 &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(p_0)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p)\right) \\ &= p. \end{aligned}$$

(d) Choose p and $q = \Phi_{p_0}\left(-\Phi_{p_0}^{-1}(p)\right)$. We have

$$\begin{aligned} p + q &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(-\Phi_{p_0}^{-1}(p)\right)\right)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) - \Phi_{p_0}^{-1}(p)\right) \\ &= \Phi_{p_0}(0) \\ &= p_0. \end{aligned}$$

(e) Let $p \in M$ and $\alpha, \beta \in \mathbb{R}$.

$$\begin{aligned} \alpha \cdot (\beta \cdot p) &= \alpha \cdot \Phi_{p_0}\left(\beta \Phi_{p_0}^{-1}(p)\right) \\ &= \Phi_{p_0}\left(\alpha \Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\beta \Phi_{p_0}^{-1}(p)\right)\right)\right) \\ &= \Phi_{p_0}\left(\alpha\beta \Phi_{p_0}^{-1}(p)\right) \\ &= (\alpha\beta) \cdot p. \end{aligned}$$

(f) Let $p \in M$. We have

$$\begin{aligned} 1 \cdot p &= \Phi_{p_0}\left(1 \cdot \Phi_{p_0}^{-1}(p)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p)\right) \\ &= p. \end{aligned}$$

(g) Let $p, q \in M$ and $\alpha \in \mathbb{R}$. We have

$$\begin{aligned} \alpha \cdot (p + q) &= \alpha \cdot \Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right) \\ &= \Phi_{p_0}\left(\alpha \Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right)\right)\right) \\ &= \Phi_{p_0}\left(\alpha\left(\Phi_{p_0}^{-1}(p) + \Phi_{p_0}^{-1}(q)\right)\right) \\ &= \Phi_{p_0}\left(\alpha \Phi_{p_0}^{-1}(p) + \alpha \Phi_{p_0}^{-1}(q)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\alpha \Phi_{p_0}^{-1}(p)\right)\right) + \Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\alpha \Phi_{p_0}^{-1}(q)\right)\right)\right) \\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}(\alpha \cdot p) + \Phi_{p_0}^{-1}(\alpha \cdot q)\right) \\ &= \alpha \cdot p + \alpha \cdot q. \end{aligned}$$

(h) Let $p \in M$ and $\alpha, \beta \in \mathbb{R}$. We have

$$\begin{aligned}(\alpha+\beta)\cdot p &= \Phi_{p_0}\left((\alpha+\beta)\Phi_{p_0}^{-1}(p)\right)\\ &= \Phi_{p_0}\left(\alpha\Phi_{p_0}^{-1}(p)+\beta\Phi_{p_0}^{-1}(p)\right)\\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\alpha\Phi_{p_0}^{-1}(p)\right)\right)+\Phi_{p_0}^{-1}\left(\Phi_{p_0}\left(\beta\Phi_{p_0}^{-1}(p)\right)\right)\right)\\ &= \Phi_{p_0}\left(\Phi_{p_0}^{-1}\left(\alpha\cdot p\right)+\Phi_{p_0}^{-1}\left(\beta\cdot p\right)\right)\\ &= \alpha\cdot p+\beta\cdot p.\end{aligned}$$

Finally, to see that (4.6) holds, choose $p \in M$ and $x \in V$. By definition, $p_0 \oplus \Phi_{p_0}^{-1}(p) = p$, so

$$\begin{aligned}\Phi_{p_0}^{-1}\left(p\oplus x\right) &= \Phi_{p_0}\left(\left(p_0\oplus\Phi_{p_0}^{-1}(p)\right)\oplus x\right)\\ &= \Phi_{p_0}\left(p_0\oplus\left(\Phi_{p_0}^{-1}(p)+x\right)\right)\\ &= \Phi_{p_0}^{-1}(p)+x.\end{aligned}$$

□

Although each choice of origin gives rise to a different vector space structure on M, these spaces have some important similarities. In particular, constant speed straight curves are the same in each vector space (although their speeds need not be). These curves can therefore be defined without recourse to a vector space structure. With this in mind, we will define the following as an analogue of constant speed straight curves in an affine space:

Definition 4.1.6 *Let $J \subseteq \mathbb{R}$ be an interval and choose $p \in M$ and $x \in V$. The function $c : J \to M$ such that $c(t) = p \oplus (t \cdot x)$ is called an* affine curve.

The image $c(J)$ will be called a *line segment*, or simply a *line* if $J = \mathbb{R}$. There are also analogues of planes and the like, defined in a similar manner, but we will not need them here.

By far the simplest example of an affine space arises when we let $M = V = \mathbb{R}^n$, where we simply forget the operations $+$ and $\cdot$ on M and let $p \oplus x = \iota(p) + x$. (Here, $\iota : M \to V$ is

the natural inclusion map.) One traditionally thinks of M as the set of points in $\mathbb{R}^n$ and of V as the set of arrows (or directions) in $\mathbb{R}^n$. That is, $p, q \in M$ are points, then $q \ominus p$ is an arrow with its base at p and its head at q. If we choose $0 \in M$ as our origin of M, then Φ_0 provides M with the standard vector space structure and makes M and V identical. We will make no distinction between this space with zero chosen as the origin and $\mathbb{R}^n$ itself. Affine curves in this space are quite literally straight lines.

(There is nothing particularly special about $\mathbb{R}^n$ in this example; one can perform the same construction with any vector space. The affine space $(V, V, +)$ should be thought of as the vector space V without a privileged zero element.)

4.2 Application to $\mathcal{M}(\Omega)$

A more interesting example of an affine space arises when we let $M = \mathcal{M}(\Omega)$ and let $V = \mathcal{V}(\Omega)$ be the set of functions on Ω with the usual vector space structure. (There are some important topological decisions to be made when Ω is infinite, but we will defer them until later.) We will define $\oplus : \mathcal{M}(\Omega) \times \mathcal{V}(\Omega) \to \mathcal{M}(\Omega)$ such that

$$(\mu \oplus f)(A) = \int_A e^f \, \mathrm{d}\mu \tag{4.7}$$

for any $A \subseteq \Omega$. For any $\omega \in \Omega$, we have $e^{f(\omega)} \in (0, \infty)$ and $\mu(\{\omega\}) \in (0, \infty)$, and so $(\mu \oplus f)(\{\omega\}) = e^{f(\omega)}\mu(\{\omega\}) \in (0, \infty)$, which means that $\mu \oplus f$ is strictly positive and locally finite. This means that $\oplus$ is well-defined.

Lemma 4.2.1 $(\mathcal{M}(\Omega), \mathcal{V}(\Omega), \oplus)$ *is an affine structure on* $\mathcal{M}(\Omega)$.

Proof To prove this, we need to show three things:

(a) $\mu \oplus 0 = \mu$ for any $\mu \in \mathcal{M}(\Omega)$, where 0 denotes the zero function on Ω.

(b) $\big(\mu \oplus f\big) \oplus g = \mu \oplus \big(f + g\big)$ for any $\mu \in \mathcal{M}(\Omega)$ and $f, g \in \mathcal{V}(\Omega)$.

(c) For every $\mu \in \mathcal{M}(\Omega)$, the map $\Phi_\mu : \mathcal{V}(\Omega) \to \mathcal{M}(\Omega)$ such that $\Phi_\mu(f) = \mu \oplus f$ is a bijection.

To prove (a), choose $\mu \in \mathcal{M}(\Omega)$. For any $A \subseteq \Omega$, we have

$$(\mu \oplus 0)(A) = \int_A e^0 \, \mathrm{d}\mu = \int_A \mathrm{d}\mu = \mu(A),$$

so $(\mu \oplus 0) = \mu$, as desired.

To prove (b), choose $\mu \in \mathcal{M}(\Omega)$ and $f, g \in \mathcal{V}(\Omega)$. For every $A \subseteq \Omega$, we have

$$\begin{aligned}
(\mu \oplus f)(A) &= \int_A e^f \, \mathrm{d}\mu, \text{ so} \\
((\mu \oplus f) \oplus g)(A) &= \int_A e^g \, \mathrm{d}(\mu \oplus f) \\
&= \int_A e^f e^g \, \mathrm{d}\mu \\
&= \int_A e^{f+g} \, \mathrm{d}\mu \\
&= (\mu \oplus (f+g))(A).
\end{aligned}$$

Therefore, $((\mu \oplus f) \oplus g) = \mu \oplus (f+g)$, as desired.

To prove (c), choose $\mu \in \mathcal{M}(\Omega)$ and define Φ_μ such that $\Phi_\mu(f) = \mu \oplus f$. That is, for any $A \subseteq \Omega$, we have

$$\left(\Phi_\mu(f)\right)(A) = \int_A e^f \, \mathrm{d}\mu. \tag{4.8}$$

The map $\exp : \mathcal{V}(\Omega) \to \mathcal{V}(\Omega)$ such that $\exp(f) = e^f$ is one-to-one, and its range is the set $\mathcal{V}^+ = \{f \in \mathcal{V}(\Omega) \mid f(\omega) > 0 \text{ for all } \omega \in \Omega\}$, so exp is a bijection onto $\mathcal{V}^+$. From the Radon-Nikodym theorem, every measure $\nu \in \mathcal{M}(\Omega)$ is of the form $\int_A F \, \mathrm{d}\mu$ with $F \in \mathcal{V}^+$, and therefore of the form $\int_A e^f \, \mathrm{d}\mu$, so Φ_μ is a surjection. The Radon-Nikodym theorem also tells us that F, and therefore f is defined uniquely μ-almost everywhere. Since μ is strictly positive, it is only zero on the empty set and so f is unique. Therefore, Φ_μ is a bijection, as desired.

Moreover, a simple computation shows that

$$\Phi_\mu^{-1}(\nu) \equiv \nu \ominus \mu = \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}. \tag{4.9}$$

□

To characterize the affine curves in $\mathcal{M}(\Omega)$, it will be convenient to identify the measure μ with the vector

$$\mu \cong (\mu(\{\omega_1\}), \ldots, \mu(\{\omega_N\})) \equiv \left(\mu^1, \ldots, \mu^N\right),$$

and to identify the function f with the vector

$$f \cong (f(\omega_1), \ldots, f(\omega_N)) \equiv (f_1, \ldots, f_N),$$

where $\omega_1, \ldots, \omega_N$ is an arbitrary (but fixed) ordering of Ω. If we choose a point $\mu \in \mathcal{M}(\Omega)$ and a function $f \in \mathcal{V}(\Omega)$, then we have

$$[\mu \oplus (t \cdot f)]^i = \mu^i e^{t f_i}. \tag{4.10}$$

In particular, if we choose $\mu = \#$ and $f = \nu \ominus \#$ for some measure $\nu \in \mathcal{M}(\Omega)$ and rename $t \equiv \beta$, then we have

$$[\mu \oplus (\beta \cdot f)]^i = e^{\beta f_i}, \tag{4.11}$$

and so $\# \oplus (\beta\nu \ominus \#) = \nu_{(\beta)}$, where $\nu_{(\beta)}$ is the re-scaled measure defined in chapter 3. This is *almost* saying that $\nu_{(\beta)} = \beta \cdot \nu$. The reason we have had to go through the preceding mathematical contortions is that scalar multiplication is not defined on an affine space - this definition depends on the choice of reference measure.

4.3 Application to $\mathcal{P}(\Omega)$

In a similar spirit, we can define an affine structure on the space $\mathcal{P}(\Omega)$. This time, the case of infinite Ω is much more delicate. We will therefore assume that Ω is finite. The relevant mathematics for infinite Ω is covered in [32], and we hope to explore it in more detail later. For finite Ω, essentially all we need to do is normalize the measures in our definitions from the previous section. To make this precise, however, we need to establish a few technical details.

Let $\mathcal{V}(\Omega)$ be the set of all functions from Ω to $\mathbb{R}$, as before. For any $f, g \in \mathcal{V}(\Omega)$, we will say that $f \sim g$ if and only if there is some $a \in \mathbb{R}$ such that $f(\omega) - g(\omega) = a$ for every $\omega \in \Omega$.

Lemma 4.3.1 *$\sim$ is an equivalence relation.*

Proof To see this, we need to prove the following:

(a) $f \sim f$ for every $f \in \mathcal{V}(\Omega)$.

(b) For any $f, g \in \mathcal{V}(\Omega)$, if $f \sim g$ then $g \sim f$.

(c) For any $f, g, h \in \mathcal{V}(\Omega)$, if $f \sim g$ and $g \sim h$, then $f \sim h$.

To prove (a), choose $f \in \mathcal{V}(\Omega)$. For each $\omega \in \Omega$, we have $f(\omega) - f(\omega) = 0$. Therefore, $f \sim f$.

To prove (b), choose $f, g \in \mathcal{V}(\Omega)$ such that $f \sim g$. That is, there is some $a \in \mathbb{R}$ such that $f(\omega) - g(\omega) = a$ for each $\omega \in \Omega$. We therefore have $g(\omega) - f(\omega) = -a$, so $g \sim f$.

To prove (c), choose $f, g, h \in \mathcal{V}(\Omega)$ such that $f \sim g$ and $g \sim h$. That is, there exist $a, b \in \mathbb{R}$ such that $f(\omega) - g(\omega) = a$ and $g(\omega) - h(\omega) = b$ for all $\omega \in \Omega$. We therefore have $f(\omega) - (h(\omega) + b) = a$, so $f(\omega) - h(\omega) = a + b$. Therefore, $f \sim h$. $\square$

Define $\mathcal{V}_0(\Omega) = \mathcal{V}(\Omega)/\sim$. That is, $\mathcal{V}_0(\Omega)$ is the set of all equivalence classes of $\mathcal{V}(\Omega)$ that differ by a constant. Furthermore, define $+ : \mathcal{V}_0(\Omega) \times \mathcal{V}_0(\Omega) \to \mathcal{V}_0(\Omega)$ and $\cdot : \mathbb{R} \times \mathcal{V}_0(\Omega) \to \mathcal{V}_0(\Omega)$ such that

$$[f] + [g] = [f + g] \text{ and } \alpha \cdot [f] = [\alpha \cdot f], \tag{4.12}$$

for each $[f], [g] \in \mathcal{V}_0(\Omega)$ and each $\alpha \in \mathbb{R}$.

Lemma 4.3.2 *The preceding definitions of $+$ and $\cdot$ are well-defined, and $\mathcal{V}_0(\Omega)$ is a vector space when equipped with these operations, with $[0]$ as the zero element.*

Proof To see that $+$ and $\cdot$ are well-defined, choose $f_1, f_2, g_1, g_2 \in \mathcal{V}(\Omega)$ such that $f_1 \sim f_2$ and $g_1 \sim g_2$ and choose $\alpha \in \mathbb{R}$. We have some $a, b \in \mathbb{R}$ such that $f_1(\omega) - f_2(\omega) = a$ and $g_1(\omega) - g_2(\omega) = b$ for each $\omega \in \Omega$. We therefore have

$$f_1(\omega) + g_1(\omega) = f_2(\omega) + a + g_2(\omega) + b = f_2(\omega) + g_2(\omega) + (a + b),$$

so

$$[f_1] + [g_1] = [f_1 + g_1] = [f_2 + g_2] = [f_2] + [g_2].$$

Likewise, we have

$$\alpha f_1(\omega) = \alpha \left(f_2(\omega) + a \right) = \alpha f_2(\omega) + \alpha a,$$

so

$$\alpha \cdot [f_1] = [\alpha \cdot f_1] = [\alpha \cdot f_2] = \alpha \cdot [f_2].$$

Therefore, the operations $+$ and $\cdot$ are well-defined on $\mathcal{V}_0(\Omega)$.

To prove that $\mathcal{V}_0(\Omega)$ is a vector space, we need only show that $[0]$ is a subspace of $\mathcal{V}(\Omega)$. This means that we must show that for any $\alpha \in \mathbb{R}$ and any $f, g \in \mathcal{V}(\Omega)$ such that $f \sim 0$ and $g \sim 0$, we have $(f + g) \sim 0$ and $(\alpha \cdot f) \sim 0$.

With this in mind, choose $\alpha \in \mathbb{R}$ and $f, g \in \mathcal{V}(\Omega)$ such that $f \sim 0$ and $g \sim 0$. There exist $a, b \in \mathbb{R}$ such that $f(\omega) = a$ and $g(\omega) = b$ for all $\omega \in \Omega$. We therefore have $f(\omega)+g(\omega) = a+b$, so $(f + g) \sim 0$. Likewise, we have $\alpha f(\omega) = \alpha a$, so $(\alpha \cdot f) \sim 0$. Therefore, $\mathcal{V}_0(\Omega)$ is a vector space. $\square$

Now define $\oplus : \mathcal{P}(\Omega) \times \mathcal{V}_0(\Omega) \to \mathcal{P}(\Omega)$ such that

$$\left(\mu \oplus [f] \right)(A) = \left(Z(f, \mu) \right)^{-1} \int_A e^f \, \mathrm{d}\mu, \tag{4.13}$$

for every $A \subseteq \Omega$, where

$$Z(f, \mu) = \int_\Omega e^f \, \mathrm{d}\mu. \tag{4.14}$$

Lemma 4.3.3 *$\oplus$ is well-defined*[1]*.*

Proof We need to show that $\mu \oplus [f] \in \mathcal{P}(\Omega)$ and that the value of $\mu \oplus [f]$ does not depend on the choice of representative function $f \in \mathcal{V}(\Omega)$.

---

[1]This is why we have temporarilly assumed that Ω is finite. If Ω is infinite, then there is no guarantee that either of the integrals in this definition are defined, and so we will need to take much more care with our definition.

First, choose $\mu \in \mathcal{P}(\Omega)$ and $f \in \mathcal{V}(\Omega)$. Since μ is strictly positive and $e^f > 0$ on Ω, it is clear that $\mu \oplus [f]$ is also strictly positive. Since $e^{f(\omega)}\mu(\{\omega\})$ is finite for all $\omega \in \Omega$ and $Z(f, \mu)$ is also finite, we know that $(\mu \oplus [f])(\{\omega\})$ is also finite, and so $\mu \oplus [f]$ is locally finite. The only thing to check is that it is normalized. We have

$$\begin{aligned}\int_\Omega \mathrm{d}\left(\mu \oplus [f]\right) &= \left(Z(f,\mu)\right)^{-1} \int_\Omega e^f \,\mathrm{d}\mu \\ &= \left(\int_\Omega e^f \,\mathrm{d}\mu\right)^{-1} \cdot \int_\Omega e^f \,\mathrm{d}\mu \\ &= 1.\end{aligned}$$

Therefore, $\mu \oplus [f] \in \mathcal{P}(\Omega)$ as desired.

Next, choose $f, g \in \mathcal{V}(\Omega)$ such that $f \sim g$. By definition, there is some $a \in \mathbb{R}$ such that $f - g = a$. We have

$$\begin{aligned}Z(f,\mu) &= \int_\Omega e^f \,\mathrm{d}\mu \\ &= \int_\Omega e^{g+a} \,\mathrm{d}\mu \\ &= e^a \int_\Omega e^g \,\mathrm{d}\mu \\ &= e^a Z(g,\mu).\end{aligned}$$

Therefore, for every $A \subseteq \Omega$ we have

$$\begin{aligned}\left(\mu \oplus [f]\right)(A) &= \left(Z(f,\mu)\right)^{-1} \int_A e^f \,\mathrm{d}\mu \\ &= \left(e^a Z(g,\mu)\right)^{-1} \int_A e^{g+a} \,\mathrm{d}\mu \\ &= e^{-a}\left(Z(g,\mu)\right)^{-1} \int_A e^g e^a \,\mathrm{d}\mu \\ &= \left(Z(g,\mu)\right)^{-1} \int_A e^g \,\mathrm{d}\mu \\ &= \left(\mu \oplus [g]\right)(A).\end{aligned}$$

Therefore, $\mu \oplus [f] = \mu \oplus [g]$ as desired. $\square$

Finally, we can show the following:

Lemma 4.3.4 *The triple* $(\mathcal{P}(\Omega), \mathcal{V}_0(\Omega), \oplus)$ *is an affine space.*

Proof To prove this, we need to show three things:

(a) $\mu \oplus [0] = \mu$ for any $\mu \in \mathcal{P}(\Omega)$, where 0 denotes the zero function on Ω.

(b) $\big(\mu \oplus [f]\big) \oplus [g] = \mu \oplus \big([f] + [g]\big)$ for any $\mu \in \mathcal{P}(\Omega)$ and any $[f], [g] \in \mathcal{V}_0(\Omega)$.

(c) For every $\mu \in \mathcal{P}(\Omega)$, the map $\Phi_\mu : \mathcal{V}_0(\Omega) \to \mathcal{P}(\Omega)$ such that $\Phi_\mu([f]) = \mu \oplus [f]$ is a bijection.

To prove (a), choose $\mu \in \mathcal{P}(\Omega)$. We have

$$Z(0, \mu) = \int_\Omega e^0 \, \mathrm{d}\mu = \int_\Omega \mathrm{d}\mu = \mu(\Omega) = 1.$$

Therefore, for any $A \subseteq \Omega$ we have

$$(\mu \oplus [0])\,(A) = (\mu(\Omega))^{-1} \int_A e^0 \, \mathrm{d}\mu = \int_A \mathrm{d}\mu,$$

so $\mu \oplus [0] = \mu$.

To prove (b), choose $\mu \in \mathcal{P}(\Omega)$ and $[f], [g] \in \mathcal{V}_0(\Omega)$. We have

$$\begin{aligned} Z(f + g, \mu) &= \int_\Omega e^{f+g} \, \mathrm{d}\mu \ \text{ and} \\ Z(g, \mu \oplus [f]) &= \int_\Omega e^g \, \mathrm{d}\,(\mu \oplus [f]) \\ &= (Z(f, \mu))^{-1} \int_\Omega e^g e^f \, \mathrm{d}\mu \\ &= \frac{Z(f + g, \mu)}{Z(f, \mu)}. \end{aligned}$$

For any $A \subseteq \Omega$, we have

$$\begin{aligned}((\mu \oplus [f]) \oplus [g])(A) &= (Z(g, \mu \oplus [f]))^{-1} \int_A e^g \,\mathrm{d}(\mu \oplus [f]) \\ &= \frac{Z(f,\mu)}{Z(f+g,\mu)} (Z(f,\mu))^{-1} \int_A e^g e^f \,\mathrm{d}\mu \\ &= (Z(f+g,\mu))^{-1} \int_A e^{f+g} \,\mathrm{d}\mu \\ &= (\mu \oplus ([f]+[g]))(A),\end{aligned}$$

so $(\mu \oplus [f]) \oplus [g] = \mu \oplus ([f] + [g])$, as desired.

To prove (c), choose $\mu, \nu \in \mathcal{P}(\Omega)$ and define

$$f = \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

We have

$$Z(f,\mu) = \int_\Omega \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \,\mathrm{d}\mu = \nu(\Omega) = 1.$$

For every $A \subseteq \Omega$, we therefore have

$$(\mu \oplus [f])(A) = (Z(f,\mu))^{-1} \int_A e^f \,\mathrm{d}\mu = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \,\mathrm{d}\mu = \int_A \mathrm{d}\nu = \nu(A).$$

Therefore, $\mu \oplus [f] = \nu$. Since ν was arbitrary, Φ_μ is surjective.

Now choose $[f], [g] \in \mathcal{V}_0(\Omega)$ such that $\mu \oplus [f] = \mu \oplus [g]$. For any $A \subseteq \Omega$, we have

$$(\mu \oplus [f])(A) = (Z(f,\mu))^{-1} \int_A e^f \,\mathrm{d}\mu = (Z(g,\mu))^{-1} \int_A e^g \,\mathrm{d}\mu = (\mu \oplus [g])(A).$$

In particular, if we choose $A = \{\omega\}$ for some $\omega \in \Omega$, then we have

$$(Z(f,\mu))^{-1} e^{f(S)} \mu(\{\omega\}) = (Z(g,\mu))^{-1} e^{g(\omega)} \mu(\{\omega\}),$$

so

$$e^{f(\omega)-g(\omega)} = \frac{Z(f,\mu)}{Z(g,\mu)},$$

and therefore

$$f(\omega) - g(\omega) = \ln \frac{Z(f,\mu)}{Z(g,\mu)}.$$

Since ω was arbitrary, this means that $f \sim g$, so $[f] = [g]$. Therefore, Φ_μ is injective, and so it is a bijection as desired.

Moreover, this means that

$$\Phi_\mu^{-1}(\nu) \equiv \nu \ominus \mu = \left[\ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right].$$

□

To characterize the affine curves in $\mathcal{P}(\Omega)$, we will use the same identification as in the previous section. That is, for any $\mu \in \mathcal{P}(\Omega)$ we will identify

$$\mu \cong (\mu(\{\omega_1\}), \ldots, \mu(\{\omega_N\})) \equiv \left(\mu^1, \ldots, \mu^N\right),$$

and for any $[f] \in \mathcal{V}_0(\Omega)$ we will identify

$$[f] \cong [f(\omega_1), \ldots, f(\omega_N)] \equiv [f_1, \ldots, f_N],$$

with the obvious equivalence relation. If we choose a point $\mu \in \mathcal{P}(\Omega)$ and a vector $[f] \in \mathcal{V}_0(\Omega)$, then we have

$$[\mu \oplus (t \cdot [f])]^i = \frac{\mu^i}{Z(tf, \mu)} e^{t f_i}. \tag{4.15}$$

The key point is that this is the same as a normalized affine curve in $\mathcal{M}(\Omega)$. That is:

Lemma 4.3.5 *Define $\mathfrak{N} : \mathcal{M}(\Omega) \to \mathcal{P}(\Omega)$ such that*

$$\mathfrak{N}(\mu)(A) = \frac{\mu(A)}{\mu(\Omega)}.$$

For any $\mu \in \mathcal{M}(\Omega)$ and any $f \in \mathcal{V}(\Omega)$ and any $t \in \mathbb{R}$, we have

$$\mathfrak{N}(\mu \oplus (t \cdot f)) = \mathfrak{N}(\mu) \oplus (t \cdot [f]).$$

Proof Choose $\mu \in \mathcal{M}(\Omega)$ and $f \in \mathcal{V}(\Omega)$ and $t \in \mathbb{R}$ and let $A \subseteq \Omega$. We have

$$\begin{aligned} Z(tf, \mu) &= \int_\Omega e^{tf} \, \mathrm{d}\mu \\ &= \mu(\Omega) \int_\Omega \frac{e^{tf}}{\mu(\Omega)} \, \mathrm{d}\mu \\ &= \mu(\Omega) \int_\Omega e^{tf} \, \mathrm{d}\mathfrak{N}(\mu) \\ &= \mu(\Omega) Z(tf, \mathfrak{N}(\mu)). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathfrak{N}\left(\mu \oplus (t \cdot f)\right)(A) &= \frac{\left(\mu \oplus (t \cdot f)\right)(A)}{\left(\mu \oplus (t \cdot f)\right)(\Omega)} \\ &= \frac{\int_A e^{tf} \, \mathrm{d}\mu}{\int_\Omega e^{tf} \, \mathrm{d}\mu} \\ &= (Z(tf, \mu))^{-1} \int_A e^{tf} \, \mathrm{d}\mu \\ &= (\mu(\Omega) Z(tf, \mathfrak{N}(\mu)))^{-1} \int_A e^{tf} \, \mathrm{d}\mu \\ &= (Z(tf, \mathfrak{N}(\mu)))^{-1} \int_A \frac{e^{tf}}{\mu(\Omega)} \, \mathrm{d}\mu \\ &= (Z(tf, \mathfrak{N}(\mu)))^{-1} \int_A e^{tf} \, \mathrm{d}\mathfrak{N}(\mu) \\ &= \left(\mathfrak{N}(\mu) \oplus (t \cdot [f])\right)(A). \end{aligned}$$

Since A was arbitrary, we have $\mathfrak{N}(\mu \oplus (t \cdot f)) = \mathfrak{N}(\mu) \oplus (t \cdot [f])$ as desired. $\square$

This means that affine lines in $\mathcal{P}(\Omega)$ are just the projections of affine lines in $\mathcal{M}(\Omega)$, where the projection is defined by $\mathfrak{N}$.

Chapter 5

ENTROPY AND CRITICALITY

In sections 3 and 4, we established the curves $\mu_{(\beta)}$ in $\mathcal{M}(\Omega)$. One of the most important features of these curves is that they may not remain normalizable for all β. In particular, if we fix $\mu, \lambda \in \mathcal{M}(\Omega)$ and define $H(\omega;\lambda) = -\ln \mathrm{d}\mu/\mathrm{d}\lambda(\omega)$, then the curve $\mu_{(\beta)}$ is defined by

$$\mu_{(\beta)}(A) = \int_A e^{-\beta H(\omega;\lambda)}\,\mathrm{d}\lambda. \tag{5.1}$$

The normalization factor (if it exists) is

$$Z(\beta;\lambda) \equiv \int_\Omega \mathrm{d}\mu_{(\beta)}. \tag{5.2}$$

It is a simple matter to find examples where $Z(\beta) \notin (0,\infty)$ for some finite β. For instance, if $\Omega = \mathbb{Z}^+$, $\lambda = \#$ and

$$\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} = \frac{x^{-\alpha}}{\zeta(\alpha)}, \tag{5.3}$$

where $\alpha > 1$ and ζ is the Riemann zeta function, then $Z(\beta;\lambda) = \infty$ for all $\beta < 1/\alpha$ [42]. This means that an affine line through some measure $\mu \in \mathcal{M}(\Omega)$ may not lie entirely in $\mathcal{M}_N(\Omega)$. In addition, the curve $\mu_{(\beta)}$ and the function $Z(\beta;\lambda)$ may have different regularity properties at different values of β (and this may depend on the reference measure λ). In the previous example, one can show that $Z(\beta;\lambda)$ is analytic (as a function of β) wherever it is positive and finite [42], but it is not immediately obvious that this should hold for arbitrary μ and λ.

The question of where Z loses regularity is extremely important in statistical physics. As we have already discussed in chapter 3, the function $Z(\beta;\#)$ is analogous to the canonical partition function in statistical mechanics. In that context, real values of β where Z becomes zero or infinite or otherwise ceases to be analytic are known as critical temperatures, and

systems that exhibit such behavior for some $\beta \in (0, \infty)$ are said to undergo a phase transition (see, for example, [48, 46, 26, 7]).

In this section, we will outline one of the major results of our previously published paper [42]. In particular, we will define quantities analogous to Gibbs and Boltzmann entropy in statistical mechanics for a measure μ and show that the existence and value of a critical temperature β_c where Z becomes non-analytic depends on the relationship between entropy and energy. That paper was written early in the development of this theory and it has a slightly different focus and notation. Most relevant to us, that paper worked solely with a probability measure $\mu \in \mathcal{P}(\Omega)$ and with the reference measure $\lambda = \#$. In this document, we will allow $\mu \in \mathcal{M}_N(\Omega)$ - i.e., we will allow any measure with $\mu(\Omega) < \infty$, not just $\mu(\Omega) = 1$ - but we will follow the original paper in choice of reference measure and not allow arbitrary λ. We believe that extending this theory to arbitrary λ is a worthwhile endeavor, but it is outside the scope of this dissertation.

With the preceding discussion in mind, fix a normalizable measure $\mu \in \mathcal{M}_N(\Omega)$ and let $\lambda = \#$ be the counting measure. Define $\mu_{(\beta)}$ as in equation (5.1) for all β such that $Z(\beta) \in (0, \infty)$. We will characterize the domain of Z in detail in short order, but for the moment it is safe to assume that $\mu_{(\beta)}$ is well-defined on some open interval $J \subseteq \mathbb{R}$ containing $[1, \infty)$.

In this section, we will rewrite Z in terms of values of H rather than directly in terms of states in Ω. The mathematics is quite simple, but it involves the definition of some ancillary quantities that have direct analogues in statistical mechanics. We will therefore take a brief diversion to discuss several definitions of entropy in the context of mechanics.

5.1 Gibbs, Boltzmann and Shannon entropies in mechanical systems

The terms discussed in this section can be found in any reasonable text on statistical mechanics. A particularly cogent summary can be found in [20].

Suppose that one has a classical system described by a Hamiltonian $H(\mathbf{x})$, where $\mathbf{x}$ denotes the "microscopic state" of the system. (We will not concern ourselves with what

qualifies as a "microscopic state". One can obtain a perfectly good intuition for the subject by assuming that $\mathbf{x} \in \mathbb{R}^{6N}$ is a vector with the positions and momenta of N particles.) Furthermore, assume that the system is ergodic on level sets of H and that the states are distributed according to the density

$$\rho(\mathbf{x} \mid E) = \frac{\delta(E - H(\mathbf{x}))}{D(E)}, \tag{5.4}$$

where $D(E)$ is a normalization constant describing the density of states given by

$$D(E) = \Lambda \int \delta(E - H(\mathbf{x}))\, \mathrm{d}\mathbf{x}, \tag{5.5}$$

where the integral is over all microscopic states and Λ is a constant expressing some information about the symmetry of the system (e.g., whether or not particles are distinguishable) and giving appropriate units. For mathematical convenience, we will assume that D is smooth. The integrated density of states is given by

$$X(E) = \int \Theta(E - H(\mathbf{x}))\, \mathrm{d}\mathbf{x}, \tag{5.6}$$

where Θ is the Heaviside function. We therefore have $D = \mathrm{d}X/\mathrm{d}E$. It is important to note that X is unitless, but that D has units of inverse energy.

One can use either D or X to define the entropy of the system. The preceding definitions are uncontroversial, but the choice of which to use as a basis for entropy has elicited substantial debate. One option is to define the entropy as the logarithm of X (with associated units). This method was originally proposed by J.W. Gibbs [15] and is therefore known[1] as the Gibbs entropy:

$$S_G = k_B \ln X(E), \tag{5.7}$$

where k_B is Boltzmann's constant.

Alternatively, one can define the entropy of this system as the logarithm of D (with appropriate units). This method is generally attributed to Boltzmann, and so the resulting

[1]This terminology is not entirely standard, but it is relatively common.

entropy is known as the Boltzmann entropy:

$$S_B = k_B \ln \epsilon D(E), \tag{5.8}$$

where k_B is Boltzmann's constant and ϵ is a positive constant with units of energy[2].

In typical statistical mechanical applications (in particular, when the number of particles N goes to infinity) these two definitions of entropy usually coincide. In situations where the two are different, though, there is serious debate over which is a better fundamental definition. One of the key points of contention is that $X(E)$ is always monotonic, while $D(E)$ need not be. This means that one can use S_G to define a temperature that is never negative, but the corresponding temperature using S_B can have any sign. We certainly don't pretend to have settled the debate in this document, but it will turn out that the analogue of Gibbs entropy plays a more important role in our discussion.

Both S_G and S_B are microcanonical entropies - that is, they depend on E, which is assumed to be fixed. It is also possible to define an entropy S in the context of fluctuating energy. In that context, it no longer makes sense to have entropy be a function of E, since our system no longer has a unique energy. Instead, one usually thinks[3] of S as a function of "average energy". This quantity is called the canonical entropy[4]

$$S = \int \rho(\mathbf{x}) \ln \rho(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{5.9}$$

---

[2]If ϵ is sufficiently small, one can think of $\epsilon D(E)$ as being the volume in phase space of an ϵ-width shell around the level set $H(\mathbf{x}) = E$.

[3]There are many mathematically equivalent ways to arive at these definitions. One that is particularly apropos is to define S as a function of the density ρ and then show that for any densities ρ_1, ρ_2 chosen from some suitable family, we have $S(\rho_1) = S(\rho_2)$ whenever $U(\rho_1) \equiv \int H(\mathbf{x})\rho_1(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int H(\mathbf{x})\rho_2(\mathbf{x}) \, \mathrm{d}\mathbf{x} \equiv U(\rho_2)$. One can then think of S as a function of U.

[4]In this case, the definition is standard, but the name is not. The function S is also often called Gibbs canonical entropy, Gibbs entropy, Shannon entropy, or simply entropy. In this document we will generally refer to S as Shannon entropy if there is any potential for confusion, since that is standard in information theory and statistics.

5.2 Gibbs, Boltzmann and Shannon entropies in $\mathcal{M}(\Omega)$

The ideas from the previous section are straightforward to translate to the context of a measure on Ω. We will treat Ω as the phase space of our system and treat the function $H = -\ln \mathrm{d}\mu/\mathrm{d}\lambda$ as an (effective) Hamiltonian (as discussed in chapter 3). The operator $\delta(E-H(\omega))$ makes perfect sense in this context, but the definition of D requires an additional choice. In equation (5.5), we integrated with respect to the Lebesgue measure $\mathrm{d}\mathbf{x}$ on phase space. This is a natural choice in the context of a Hamiltonian system, since H preserves Lebesgue volume, but there is no such natural choice on an arbitrary set Ω. We will integrate with respect to the (already arbitrarily chosen) λ. (In [42], we used $\lambda = \#$ without comment. Some of the proofs in the next section will rely on that choice, but it is worth emphasizing that this is completely arbitrary and that different choices will lead to qualitatively different results.) We therefore have

$$D(E) = \int_\Omega \delta(E - H(\omega))\, \mathrm{d}\lambda. \tag{5.10}$$

(Note that we have chosen Λ to have a value of 1 in the appropriate units. $D(E)$ still has units of inverse energy.) This can be rewritten as

$$D(E) = \lambda(\{\omega \in \Omega \mid H(\omega) = E\}). \tag{5.11}$$

If $\lambda = \#$ then this is the degeneracy of E - i.e., the number of states with energy E. Our analogue of the Boltzmann entropy is therefore

$$S_B(E) = k_B \ln \epsilon D(E), \tag{5.12}$$

where D is defined through (5.11) and k_B and ϵ are constants with appropriate units. Note that, since Ω is discrete, the set $\{\omega \in \Omega \mid H(\omega) = E\}$ is empty for almost all values of E, and so $D(E) = 0$ and $S_B = -\infty$ for those values. Fortunately, this is the only issue that can arise and S_B is finite for all other values of E.

Lemma 5.2.1 *If $\mu \in \mathcal{M}_N(\Omega)$ and $\lambda \in \mathcal{M}(\Omega)$ and there exists $\omega \in \Omega$ such that $H(\omega) = E$, then $D(E)$ is finite.*

Proof Choose $E \in \mathbb{R}$ such that there exists some $\omega \in \Omega$ with $H(\omega) = E$ and define $D_E = \{\omega \in \Omega \mid H(\omega) = E\} \neq \emptyset$. We have $D(E) = \lambda(D_E)$. Since λ is strictly positive and $D_E \neq \emptyset$, we know that $D(E) > 0$. For every $\omega \in D_E$, we have

$$E = H(\omega) = -\ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega) = -\ln \frac{\mu(\{\omega\})}{\lambda(\{\omega\})}, \tag{5.13}$$

and so

$$\lambda(\{\omega\}) = \mu(\{\omega\})e^E. \tag{5.14}$$

If we sum over all $\omega \in D_E$, then we obtain

$$D(E) = \lambda(D_E) = \mu(D_E)e^E \leq \mu(\Omega)e^E < \infty. \tag{5.15}$$

□

Note that the assumption that μ is normalizable is much stronger than necessary, but some restriction is still required on μ. Two trivial examples with $\Omega = \mathbb{Z}^+$ and $\lambda = \#$ will serve to illustrate the point. On the one hand, if $\mu(\{\omega_1\}) \neq \mu(\{\omega_2\})$ whenever $\omega_1 \neq \omega_2$, then $D(E)$ is zero or 1 for all $E \in \mathbb{R}$. For instance, if $\mu(\{\omega\}) = 1/\omega$ for each $\omega \in \mathbb{Z}^+$, then μ is not normalizable but the conclusion of the lemma would still be true. On the other hand, if $\mu = \#$ then $D(0) = \infty$, and the conclusion would certainly be false.

Similarly, we have

$$X(E) = \int_\Omega \Theta(E - H(\omega))\,\mathrm{d}\lambda = \lambda(\{\omega \in \Omega \mid H(\omega) \leq E\}), \tag{5.16}$$

and the corresponding analogue of Gibbs entropy is

$$S_G(E) = k_B \ln X(E), \tag{5.17}$$

where X is defined through (5.16). Once again, if the set $\{\omega \in \Omega \mid H(\omega) \leq E\}$ is empty then $X(E) = 0$ and $S_G(E) = -\infty$. As before, this is the only problem that can arise. For all other values of E, the Gibbs entropy is finite.

Lemma 5.2.2 *If $\mu \in \mathcal{M}_N(\Omega)$ and $\lambda \in \mathcal{M}(\Omega)$ and there exists $\omega \in \Omega$ such that $H(\omega) \leq E$, then $X(E)$ is finite.*

Proof Choose $E \in \mathbb{R}$ such that there exists some $\omega \in \Omega$ with $H(\omega) \leq E$ and define $X_E = \{\omega \in \Omega \mid H(\omega) \leq E\} \neq \emptyset$. We have $X(E) = \lambda(X_E)$. Since λ is strictly positive and $X_E \neq \emptyset$, we know that $X(E) > 0$. Furthermore, we know that

$$X_E = \bigcup_{h \leq E} D_h = \bigcup_{h \in I} D_h, \tag{5.18}$$

where $I = \{h \leq E \mid D_h \neq \emptyset\}$. Since X_E is non-empty, so is I. Since Ω is countable, I is also countable. We therefore have

$$X(E) = \lambda(X_E) = \sum_{h \in I} \lambda(D_h) = \sum_{h \in I} D(h). \tag{5.19}$$

Since μ is normalizable, each term in this sum is finite. For each $h \in I$, we have

$$D(h) = \lambda(D_h) = e^h \mu(D_h), \tag{5.20}$$

and so

$$X(E) = \sum_{h \in I} \lambda(D_h) = \sum_{h \in I} e^h \mu(D_h) \leq e^E \sum_{h \in I} \mu(D_h) = e^E \mu(X_E) < \infty. \tag{5.21}$$

□

Note that the assumption that μ be normalizable is still stronger than necessary (although it is acceptable for our purposes).

It is also important to note that if $\lambda(\Omega)$ is finite (i.e., if $\lambda \in \mathcal{M}_N(\Omega)$), then $X(E) < \infty$ for all $E \in \mathbb{R}$. In particular, this means that if λ is a probability measure, then the Gibbs entropy of any μ is finite. This serves as a useful illustration of the point that these thermodynamic properties are not just properties of the measure μ - they depend on the choice of representation.

Finally, we come to Shannon entropy. This is probably the most familiar of the three in the context of measures, and yet the analogy is not quite as immediate. There are two

important issues: First, one usually defines the Shannon entropy as a function of a *density*, not a measure. This means that we still need to choose a reference measure and define the entropy relative to that reference. In other words, we should be working with the KL divergence between two measures. Second, the Shannon entropy (and KL divergence) are generally only defined for probability measures, not for arbitrary positive measures. Recall from chapter 2, however, that we can safely define equilibrium quantities for normalizable measures only. We will therefore define the entropy of a measure μ (with reference measure λ) as

$$S(\mu;\lambda) = \begin{cases} -\int_\Omega \frac{\mathrm{d}\mathfrak{N}(\mu)}{\mathrm{d}\lambda} \ln \frac{\mathrm{d}\mathfrak{N}(\mu)}{\mathrm{d}\lambda}\,\mathrm{d}\lambda & \text{if } \mu(\Omega) < \infty, \\ \infty & \text{otherwise,} \end{cases} \tag{5.22}$$

where $\mathfrak{N}$ is the normalization function from (2.21). It will often be more convenient to write this as

$$\ln \mu(\Omega) - \int_\Omega \ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\,\mathrm{d}(\mathfrak{N}(\mu)). \tag{5.23}$$

When μ and λ are probability measures, this is just $D_{KL}(\mu||\lambda)$.

$S(\mu;\lambda)$ is not necessarily finite for all $\mu \in \mathcal{M}_N(\Omega)$, even when both μ and λ are probability distributions. For our purposes, it is sufficient to assume that $\mathrm{d}\lambda/\mathrm{d}\#$ is bounded away from zero. In particular, if $\mu \in \mathcal{M}_N(\Omega)$ and λ is the counting measure, then S is finite.

In statistical mechanics, one generally thinks of S as a function of the average energy U of the canonical ensemble of systems. We can derive an analogous relationship in the space of measures by looking at the family of measures $\mu_{(\beta)}$ defined in (5.1) and (5.2). In particular, for any normalizable measure $\mu \in \mathcal{M}_N(\Omega)$ and any reference measure $\lambda \in \mathcal{M}(\Omega)$, we will define the *average energy* of μ at β as

$$U(\beta;\mu,\lambda) = \int_\Omega H(\omega)\,\mathrm{d}\mathfrak{N}(\mu_{(\beta)}) = -\frac{1}{\mu_{(\beta)}(\Omega)} \int_\Omega \ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega)\,\mathrm{d}\mu_{(\beta)}. \tag{5.24}$$

When μ and λ are clear from context, we will simply refer to this as $U(\beta)$. Similarly, we will define the entropy of μ at β as

$$S(\beta;\mu,\lambda) \equiv S(\mu_{(\beta)};\lambda). \tag{5.25}$$

Again, when μ and λ are clear from context, we will refer to this quantity as $S(\beta)$.

We will now derive several classical relations concerning U and S. First, we have

$$\begin{aligned} U(\beta) &= \frac{1}{\mu_{(\beta)}(\Omega)} \int_\Omega H(\omega)\, \mathrm{d}\mu_{(\beta)} \\ &= \frac{1}{Z(\beta)} \int_\Omega H(\omega) e^{-\beta H(\omega)}\, \mathrm{d}\lambda \\ &= \frac{1}{Z(\beta)} \int_\Omega -\frac{\partial}{\partial \beta} e^{-\beta H(\omega)}\, \mathrm{d}\lambda \\ &= -\frac{1}{Z(\beta)} \frac{\partial}{\partial \beta} \int_\Omega e^{-\beta H(\omega)}\, \mathrm{d}\lambda \\ &= -\frac{1}{Z(\beta)} \frac{\partial Z}{\partial \beta} \\ &= \frac{\mathrm{d}}{\mathrm{d}\beta} \left[-\ln Z(\beta) \right]. \end{aligned} \tag{5.26}$$

(We will justify the exchange of derivative and integral in the next section.) The function $-\ln Z(\beta)$ will arise often enough that it is worth naming. We will define the *free entropy*

$$\Phi(\beta) = -\ln Z(\beta). \tag{5.27}$$

It is well known that Φ is strictly concave on the interior of its domain. This means that we can use the inverse function theorem to write β as a function of U. Moreover, Φ has a Legendre transform given by

$$\begin{aligned} \mathcal{L}\left[\Phi\right](U) &= \beta(U) \cdot U - \Phi(\beta(U)) \\ &= \beta(U) \cdot U + \ln Z(\beta) \\ &= \beta \cdot \int_\Omega H(\omega)\, \mathrm{d}\mathfrak{N}(\mu_{(\beta)}) + \ln \mu_{(\beta)}(\Omega) \\ &= \ln \mu_{(\beta)}(\Omega) - \beta \int_\Omega \ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega)\, \mathrm{d}\mathfrak{N}(\mu_{(\beta)}) \\ &= \ln \mu_{(\beta)}(\Omega) - \int_\Omega \ln \frac{\mathrm{d}\mu_{(\beta)}}{\mathrm{d}\lambda}(\omega)\, \mathrm{d}\mathfrak{N}(\mu_{(\beta)}) \\ &= S(\beta). \end{aligned}$$

That is, the entropy S is the Legendre transform of the free entropy.

5.3 *Convergence and analyticity of* $Z(\beta)$

In this section, we will reproduce one of the main results from [42]. We will assume throughout that $\lambda = \#$ and $\mu \in \mathcal{P}(\Omega)$.

With the definitions from the previous section in hand, we can rewrite equation (5.2) without direct reference to Ω. In particular, we have

$$Z(\beta) = \int_0^\infty e^{-\beta E} \, \mathrm{d}X(E). \tag{5.28}$$

This is exactly the Laplace-Stieltjes transform of X. It is worth taking a moment to point out the attractive, but incorrect, interpretation of this equation. One can identify $D(E) = \mathrm{d}X/\mathrm{d}E$ and therefore write

$$Z(\beta) = \int_0^\beta e^{-\beta E} \left(\frac{\mathrm{d}X}{\mathrm{d}E} \right) \mathrm{d}E = \int_0^\infty e^{-\beta \left(E - (k_B \beta)^{-1} S_B(E) \right)} \, \mathrm{d}E. \tag{5.29}$$

This is mistaken because $S_B = k_B \ln(\epsilon D(E))$ requires an extra constant ϵ to correct units. We therefore need to write

$$Z(\beta) = \frac{1}{\epsilon} \int_0^\infty e^{-\beta \left(E - (k_B \beta)^{-1} S_B(E) \right)} \, \mathrm{d}E. \tag{5.30}$$

The two equations are identical if we ignore units, but only (5.30) is correct.

A more satisfying interpretation of Z arises if we integrate by parts, obtaining

$$Z(\beta) = \beta \int_0^\infty e^{-\beta E} X(E) \, \mathrm{d}E = \beta \int_0^\infty e^{-\beta \left(E - (k_B \beta)^{-1} S_G(E) \right)} \, \mathrm{d}E. \tag{5.31}$$

In statistical mechanics, both (5.30) and (5.31) are referred to as the "canonical partition function". Which equation is chosen depends entirely on the author's preference between Gibbs and Boltzmann entropies. As we have just seen, our definition of $Z(\beta)$ is compatible with both versions.

One of many advantages of writing $Z(\beta)$ as a Laplace transform is that we can apply several useful theorems from classical analysis. All of the relevant theorems can be found in [45]. First, there exists some value $\beta_c \in [-\infty, \infty]$ such that $Z(\beta)$ converges for all $\beta \in \mathbb{C}$ with

real part greater than β_c and diverges for all $\beta \in \mathbb{C}$ with real part less than β_c. Moreover, $Z(\beta)$ is analytic for all β with real part greater than β_c. The value β_c is called the *abscissa of convergence*.

Second, if the state space Ω is finite, then $Z(\beta)$ is a sum of finitely many terms and therefore converges for any β. In other words, $\beta_c = -\infty$, and so $Z(\beta)$ is analytic on all of $\mathbb{C}$. however, if Ω is infinite then the partition function will not converge for all β. In particular, it cannot converge when $\beta = 0$, because $Z(0) = \lambda(\Omega) = \infty$. However, we do know that $Z(\beta)$ converges when $\beta = 1$, since $Z(1) = \mu(\Omega) < \infty$. For infinite systems, the abscissa of convergence must therefore lie somewhere in $[0, 1]$. Since this abscissa is non-negative, we have

$$\beta_c = \limsup_{E \to \infty} \frac{\ln X(E)}{E}, \tag{5.32}$$

or

$$k_B \beta_c = \limsup_{E \to \infty} \frac{S_G(E)}{E}. \tag{5.33}$$

We now know that the canonical partition function $Z(\beta)$ is analytic for all complex β with real part greater than β_c, where β_c is found as in (5.33). However, we have not yet shown that $Z(\beta)$ cannot be extended analytically beyond $\beta = \beta_c$. For a general Laplace-Stieltjes transform, this might be possible. (In the worst case, a Laplace transform may have a finite abscissa of convergence but still have an analytic continuation to the entire complex plane.) Fortunately, since $X(E)$ is monotonic, $Z(\beta)$ has a singularity at β_c. (This also means that $\beta_c \neq 1$.)

This means that the partition function $Z(\beta)$ has a singularity at some positive β_c if and only if the Gibbs entropy S_G is asymptotic to the energy E in the sense of (5.33). That is, if the Gibbs entropy grows sufficiently quickly as a function of energy, it can become dominant in the computation of $Z(\beta)$ (and therefore $\mu_{(\beta)}$) at a finite temperature.

As a special case, consider the example where $\Omega = \mathbb{Z}^+$, $\lambda = \#$ and μ is the power law distribution defined by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega) = \frac{\omega^{-\alpha}}{\zeta(\alpha)}, \tag{5.34}$$

where $\alpha > 1$ and ζ is the Riemann zeta function. This gives us

$$H(\omega) = \alpha \ln \omega + \ln \zeta(\alpha), \tag{5.35}$$

$$S_G(E) = \frac{k_B}{\alpha} \left(E - \ln \zeta(\alpha)\right) \text{ and} \tag{5.36}$$

$$\beta_c = \frac{1}{\alpha}. \tag{5.37}$$

This means that power law distributions have a finite critical temperature. This result was previously demonstrated in [30]. The frameworks for the two results are similar, but neither is entirely a generalization of the other. Our result gives a sufficient *and necessary* condition for any probability distribution $\mu \in \mathcal{P}(\Omega)$ to have a critical temperature and is not limited to power laws, but the proof in [30] applies to uncountable state spaces Ω.

Chapter 6

THERMODYNAMICS ON $\mathcal{M}(\Omega)$ AND $\mathcal{P}(\Omega)$

In the following chapter, we will look at a special case of great interest in nonequilibrium thermodynamics. Suppose that the underlying process $\mathfrak{X}(t)$ has a steady-state distribution ρ^{eq} with corresponding coarse-grained distribution π on Ω, but after going through the process described in section 2.2, we observe a stationary distribution μ on Ω.

We will think of ρ^{eq} and π as the equilibrium distributions of our system (on Q and Ω respectively) and the measure μ as a non-equilibrium steady state. Note that the distinction between "equilibrium" and "non-equilibrium" is not inherent to the distributions themselves; these terms only have meaning in the context of the underlying dynamics of $\mathfrak{X}(t)$ on Q.

Assuming $\mu \neq \pi$, this is an unusual state of affairs. Exactly how unusual it is can be quantified by Sanov's large deviation rate function (from section 2.2). If we continue to make observations from our system and update our coarse-grained distribution μ, then we expect (because of the law of large numbers) that it will eventually return to π. Such a change in distribution should be thought of as a *spontaneous relaxation* process on Ω.

We are interested in how various "thermodynamic properties" of our system change under this relaxation process. The dynamics of many such thermodynamic properties are discussed extensively in [22]. In this chapter, we will reproduce some of the key results from that paper using notation more in line with the rest of this dissertation.

6.1 Non-equilibrium entropy decomposition

Any meaningful discussion of "thermodynamic properties" must include a definition of a state function for entropy. In our case, the state of the system should be thought of as the (typically non-equilibrium) steady state distribution μ, and so we need a definition of the

entropy of μ. There is a thoroughly standard choice for this definition - the Shannon entropy. That is, we can define

$$S[\mu] = -\int_{\Omega} p(\omega) \ln p(\omega) \, \mathrm{d}\omega, \tag{6.1}$$

where $p \equiv \frac{\mathrm{d}\mu}{\mathrm{d}\#}$ is the probability mass of the steady state measure μ. However, the result from 2.2 supplies us with a more natural choice of entropy function:

$$S[\mu] = -\int_{\Omega} \ln \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right) \, \mathrm{d}\mu \tag{6.2}$$

These two functions are quite similar. In particular, they are both KL divergences between the steady state measure μ and another measure on Ω (either the counting measure $\#$ or the equilibrium distribution π). In the work that follows, we will define the entropy more generally as the KL divergence between μ and some arbitrary (but fixed) reference measure λ. In practice, if we have obtained μ through a very large sequence of i.i.d. observations of the underlying dynamics $\mathfrak{X}(t)$, then we might not actually know π and will therefore be forced to use an arbitrary reference measure. However, it is important to keep in mind that the dynamics of our system ultimately provide us with a *correct* reference measure π.

With this in mind, we will define the entropy of μ as

$$S[\mu;\ \lambda] \equiv -\int_{\Omega} \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \ln \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \right) \, \mathrm{d}\lambda. \tag{6.3}$$

This is identical to the definition in (5.22), but since μ is a probability measure the formula simplifies considerably. When $\lambda = \#$, this is just the Shannon entropy, and when $\lambda = \pi$ this is the Sanov large deviation rate function from section 2.2. In the context of our affine structure on $\mathcal{M}(\Omega)$, we can rewrite (6.3) as

$$S[\mu;\ \lambda] = \mathbb{E}^{\mu} \left[\lambda \ominus \mu \right]. \tag{6.4}$$

The change in entropy as the steady state relaxes from μ to π is

$$\Delta S[\mu, \pi;\ \lambda] \equiv S[\pi;\ \lambda] - S[\mu;\ \lambda] = \Delta S^{(i)} + \Delta S^{(e)}, \tag{6.5}$$

where

$$\Delta S^{(i)} \equiv \int_\Omega \ln\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\,\mathrm{d}\mu = \mathbb{E}^\mu\left[\mu \ominus \pi\right] \text{ and} \tag{6.6}$$

$$\Delta S^{(e)} \equiv -\int_\Omega \ln\left(\frac{\mathrm{d}\pi}{\mathrm{d}\lambda}\right)\,\mathrm{d}\pi + \int_\Omega \ln\left(\frac{\mathrm{d}\pi}{\mathrm{d}\lambda}\right)\,\mathrm{d}\mu = \mathbb{E}^\pi\left[H(\cdot;\ \pi,\lambda)\right] - \mathbb{E}^\mu\left[H(\cdot;\ \pi,\lambda)\right] \tag{6.7}$$

We will refer to $\Delta S^{(i)}$ as the *internal entropy production* and $\Delta S^{(e)}$ as the *entropy exchange*. It is important to note that the internal entropy production is never negative (and is strictly positive as long as $\mu \neq \pi$). This is a straightforward consequence of Jensen's inequality:

$$\begin{aligned}
\Delta S^{(i)} &= \mathbb{E}^\mu\left[\mu \ominus \pi\right] \\
&\geq -\ln \mathbb{E}^\mu\left[e^{-(\mu\ominus\pi)}\right] \\
&= -\ln(\pi(\Omega)) \\
&= 0.
\end{aligned} \tag{6.8}$$

The idea that $\mathbb{E}^\mu\left[e^{-(\mu\ominus\pi)}\right] = \pi(\Omega)$ is surprisingly powerful. From this equation, one can derive wide variety of fluctuation theorems, including the Jarzynski equality and the Crooks' fluctuation theorem. These ideas have been more thoroughly explored in [47]. In contrast to the internal entropy production, the entropy exchange can have any sign.

The internal entropy production depends only on the equilibrium distribution π and the non-equilibrium steady state μ, *not* on the reference measure λ. In contrast, the entropy exchange is reference dependent. Both the quantity and sign of $\Delta S^{(e)}$ can vary based on our choice of λ. However, if we define our entropy through (6.2) by choosing $\lambda = \pi$, then the entropy exchange vanishes and we are left with only (non-negative) internal entropy production.

6.2 Temperature dependence and free energy

If we follow the logic of chapters 2 and 3, then we should think of $H(\omega;\ \pi,\lambda) \propto -\ln\frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(\omega)$ as the equilibrium internal energy function for our system. The constant of proportionality amounts to a choice of units for temperature. Until now, we have set this constant to 1

for convenience, but now it will be useful to explicitly include it. In particular, we will let $H(\omega;\ \pi, \lambda) = -\beta_0^{-1} \ln \frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(\omega)$, where $\beta_0 \in \mathbb{R}^+$. Since this choice was arbitrary, we should also consider the 1-parameter family of energy functions with different choices of β_0. To do so, we will consider the affine line of distributions running from λ to π. That is, we will look at the 1-parameter family of measures

$$\pi_{(\beta)} = \lambda \oplus [-\beta H], \tag{6.9}$$

where $\oplus$ is affine addition[1] in $\mathcal{P}(\Omega)$. Note that $\pi_{(\beta_0)} = \pi$ and $\pi_{(0)} = \lambda$.

We have defined $\pi_{(\beta)}$ in terms of affine addition on $\mathcal{P}(\Omega)$, but it would be much simpler if we could define this family in terms of addition on $\mathcal{M}(\Omega)$ instead. This certainly works when $\beta = \beta_0$, because $\lambda \oplus [-\beta_0 H]$ is identical to $\lambda \oplus (-\beta_0 H)$. However, if $\beta \neq \beta_0$ then in general $\pi_{(\beta)} \neq \lambda \oplus (-\beta H)$. The issue is that addition in $\mathcal{P}(\Omega)$ has a gauge freedom: The term $[-\beta H]$ is only defined up to an arbitrary choice of constant. However, addition in $\mathcal{M}(\Omega)$ does not have this freedom. If we want $\pi_{(\beta)} = \lambda \oplus (-\beta H)$, then we have choose an appropriate additive constant for H. We will denote this constant[2] by $F^{eq}(\beta;\ \pi, \lambda)$ and define it so that

$$\lambda \oplus [-\beta H] = \lambda \oplus -\beta(H - F^{eq}(\beta;\ \pi, \lambda)). \tag{6.10}$$

It is straightforward to check that this relation uniquely defines

$$F^{eq}(\beta;\ \pi, \lambda) = -\beta^{-1} \ln \int_\Omega e^{-\beta H(\omega;\ \pi, \lambda)}\, \mathrm{d}\lambda. \tag{6.11}$$

This quantity is analogous to the equilibrium Helmholtz free energy in statistical mechanics.

Once we have the equilibrium free energy $F^{eq}(\beta;\ \pi, \lambda)$ in hand, we can use affine addition on $\mathcal{M}(\Omega)$ for all of our calculations, instead of the more cumbersome addition on $\mathcal{P}(\Omega)$.

The equilibrium free energy has a particularly useful decomposition (which follows directly from the results of section 5.2):

$$F^{eq}(\beta;\ \pi, \lambda) = U(\beta;\ \pi, \lambda) - \beta^{-1} S(\beta;\ \pi, \lambda), \tag{6.12}$$

[1] For the remainder of this section, the notation $\lambda \oplus [f]$ (with the function in brackets) will denote addition in $\mathcal{P}(\Omega)$, while the notation $\lambda \oplus f$ or $\lambda \oplus (f)$ (without brackets) will denote addition in $\mathcal{M}(\Omega)$.

[2] F^{eq} is a constant function on ω. It generally depends on β as well as π and λ.

where

$$U(\beta;\ \pi, \lambda) = \int_{\Omega} H(\omega;\ \pi, \lambda)\, \mathrm{d}\pi_{(\beta)} \tag{6.13}$$

is the average energy of $\lambda \oplus -\beta(H + F^{eq}(\beta;\ \pi, \lambda))$ (as defined in 5.24) and

$$S(\beta;\ \pi, \lambda) = -\left(\frac{\mathrm{d}F^{eq}(\beta;\ \pi, \lambda)}{\mathrm{d}\beta^{-1}}\right) = S[\pi_{(\beta)};\ \lambda] \tag{6.14}$$

is the entropy of $\lambda \oplus -\beta(H + F^{eq}(\beta;\ \pi, \lambda))$ (as defined in 5.25).

So far in this section, we have not made any reference to the steady-state distribution μ. Notice that the only real distinction between μ and π is at the level of the dynamics of $\mathfrak{X}(t)$. On the level of $\mathcal{M}(\Omega)$ or $\mathcal{P}(\Omega)$, the two are really on an equal footing. (To put this another way, if we obtained μ and π by making coarse-grained observations of our system through equation (2.5), then we would not have enough information to distinguish the equilibrium and non-equilibrium distribution, and so we would be forced to treat the two symmetrically.)

With this in mind, we can go through the same construction using μ in place of π. There is an additional layer of complexity when choosing units of temperature, though. We would like to define our energy as $-\beta_1^{-1}H(\omega;\ \mu, \lambda)$, where $\beta_1 \in \mathbb{R}^+$ is another arbitrary constant. However, if μ is already on the affine line $\pi_{(\beta)}$, then we should not arbitrarilly choose temperature units. Instead, we should choose β_1 so that $\pi_{(\beta_1)} = \mu$. For the purposes of this chapter, assume that this is not the case and so we are free to choose β_1 arbitrarilly[3]. We define the "energy" as $H(\omega;\ \mu, \lambda) = -\beta_1^{-1} \ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(\omega)$ and define the 1-parameter family of measures $\mu_{(\beta)} = \lambda \oplus [-\beta H]$. As before, we have $\mu_{(\beta_1)} = \mu$ and $\mu_{(0)} = \lambda$. Our free energy becomes

$$F^{eq}(\beta;\ \mu, \lambda) = -\beta^{-1} \int_{\Omega} e^{-\beta H(\omega;\ \mu, \lambda)}\, \mathrm{d}\lambda \tag{6.15}$$

and we obtain the relation

$$F^{eq}(\beta;\ \mu, \lambda) = U(\beta;\ \mu, \lambda) - \beta^{-1} S(\beta;\ \mu, \lambda), \tag{6.16}$$

[3]What we should really do is choose a basis for $\mathcal{V}$ and use that basis to define each β. In particular, we could take n affinely independent measures $\{\mu_i\}$ (in addition to our reference measure) and choose a different arbitrary β_i for each of them. The energy for any other measure could then be written as a linear combination of the $\{\beta_i^{-1} H(\cdot;\ \mu_i, \lambda)\}$ without any further arbitrary choice.

where

$$U(\beta;\ \mu,\lambda) = \int_\Omega H(\omega;\ \mu,\lambda)\,\mathrm{d}\mu_{(\beta)} \tag{6.17}$$

and

$$S(\beta;\ \mu,\lambda) = -\left(\frac{\mathrm{d}F^{eq}(\beta;\ \mu,\lambda)}{\mathrm{d}\beta^{-1}}\right) = S[\mu_{(\beta)};\ \lambda] \tag{6.18}$$

It is no longer entirely fair to identify these functions with equilibrium quantities from classical thermodynamics, because we know that μ is not an equilibrium distribution. Instead, we will think of these as "fictitious equilibria". In other words, $F^{eq}(\beta;\ \mu,\lambda)$ is what the free energy of $\mu_{(\beta)}$ would be if μ were really an equilibrium.

If we choose λ carefully, we can drastically simplify many of these quantities. In particular,

$$F^{eq}(\beta;\ \pi,\pi) \equiv 0, \tag{6.19}$$

$$U(\beta;\ \pi,\pi) \equiv 0, \tag{6.20}$$

$$S(\beta;\ \pi,\pi) \equiv 0. \tag{6.21}$$

However, unless $\pi = \mu$ (in which case there is no spontaneous relaxation process) we cannot simultaneously eliminate $F^{eq}(\beta;\ \pi,\lambda)$ and $F^{eq}(\beta;\ \mu,\lambda)$.

6.3 Non-equilibrium free energy

In addition to the equilibrium (or at least fictitious equilibrium) quantities defined above, we need to define non-equilibrium counterparts. It may seem odd to define $F^{eq}(\beta;\ \mu,\lambda)$ for a non-equilibrium steady state μ as we did above and then not use it as the non-equilibrium free energy, but this definition would violate one of the core assumptions of non-equilibrium thermodynamics. In particular, the change in free energy between two steady states should be equal to the temperature β^{-1} and the internal entropy production $\Delta S^{(i)}$ [16]. With the additional requirement that the non-equilibrium free energy of an equilibrium steady state

should be equal to the equilibrium free energy, there is only one viable definition:

$$F^{neq}[\mu, \beta;\ \pi, \lambda] = \beta^{-1}\Delta S^{(i)}[\mu,\ \pi] + F^{eq}(\beta;\ \pi, \lambda) \tag{6.22}$$

$$= \beta^{-1}\mathbb{E}^{\mu}\left[\mu \ominus \pi\right] + F^{eq}(\beta;\ \pi, \lambda). \tag{6.23}$$

The reference measure only appears in the second term, and so the *change in free energy* over a spontaneous relaxation process

$$\Delta F^{neq}[\mu, \beta;\ \pi] = \beta^{-1}\Delta S^{(i)}[\mu,\ \pi] \tag{6.24}$$

is reference independent.

This quantity is *not* the same as the difference in fictitious equilibrium free energies. In particular, we find that

$$\Delta F^{neq}[\mu, \beta;\ \pi] = \left[U(\beta_1;\ \mu, \lambda) - U(\beta_0;\ \pi, \lambda)\right] - \beta_0^{-1}\left[S(\beta_1;\ \mu, \lambda) - S(\beta_1;\ \pi, \lambda)\right] + \overline{W}[\mu, \pi; \lambda], \tag{6.25}$$

where

$$\begin{aligned}\overline{W}[\mu, \pi; \lambda] &\equiv \mathbb{E}^{\mu}\left[W[\mu, \pi]\right] \\ &= \mathbb{E}^{\mu}\left[\beta^{-1}\ln\left(\frac{\mathrm{d}(\lambda \oplus (-\beta H(\cdot; \mu, \lambda)))}{\mathrm{d}(\lambda \oplus (-\beta H(\cdot; \pi, \lambda)))}\right)\right]\end{aligned} \tag{6.26}$$

The quantity W can be thought of as the irreversible work associated with the relaxation process. Despite appearances, it is not actually a function of β (although it does depend on β_0 and β_1 through the definitions of H). We have written it in this manner to emphasize the fact that the work depends on the Radon-Nikodym derivative between two *non-normalized* measures. That is, this formula makes sense because we are working in $\mathcal{M}(\Omega)$ and not just $\mathcal{P}(\Omega)$.

On the basis of these definitions, one can derive a wide variety of thermodynamic relations. We will leave the rest of these results to [22].

Chapter 7

FUTURE WORK

We have begun to explore several different avenues to extend the ideas discussed in this dissertation. Most of these extensions revolve around the set of observables $(X_1, \ldots, X_n)$ from (2.5). One possibility is to allow more general observables X_i instead of requiring them to be the indicator functions $X_i = \mathbb{1}_{\omega_i}$. Cramér's theorem still applies to a fairly broad range of observables, but our interpretation would need some modification. For one thing, the limiting "frequencies" from (2.6) would no longer really be frequencies. In particular, they would not be normalized, and so the limiting objects would not be analogous to probability measures in $\mathcal{P}(\Omega)$. However, if the observables were all non-negative, then it might still be possible to think of these limits as elements of $\mathcal{M}(\Omega)$. For another, the large deviation rate function $\varphi(\mu)$ and the corresponding conjugate variables h_i would be drastically different. Some consequences of these differences are obvious, but others (particularly those involving the affine structures defined in chapter (4)) are more difficult to generalize.

Another approach is to keep the indicator functions X_i but to relax the assumption that the samples be i.i.d.. In this case, the fundamental objects of study are still probability frequencies and measures, but the large deviation rate function φ is quite different. We have begun some analysis of this problem.

7.1 Beyond i.i.d. samples

Suppose that the random variables $\mathbf{X}_{(i)}$ from equation (2.6) are defined as before, but with one exception. Rather than being independently distributed, they are generated from a Markov chain.

There are several equivalent ways to approach this problem. One method is to consider

a continuous time Markov process on Ω with generator and then measure the proportion of time spent in each state. That is, we measure

$$L_i^t = \frac{1}{t}\int_0^t X_i(\mathfrak{X}(s))\,\mathrm{d}s. \tag{7.1}$$

It is a simple matter to show that

$$\lim_{t\to\infty} L_i^t = \pi(\{\omega_i\}). \tag{7.2}$$

One can show [3] that the large deviation rate function for this limit is of the form

$$\varphi(\mu) = \mathbb{E}^\mu[V], \tag{7.3}$$

where $V : \Omega \to \mathbb{R}$ is of the form

$$V(\omega_i) = \sum_{j=1}^n \left(1 - e^{h_j - h_i}\right) q_{ij}. \tag{7.4}$$

The parameters h_i satisfy a system of nonlinear equations (derived in [3]), and if the underlying Markov process is detailed balanced then they have a simple closed form expression:

$$h_i = \frac{1}{2}\ln\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega_i) - \frac{1}{2}\sum_{j=1}^n \ln\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega_i). \tag{7.5}$$

Using this result, we can obtain an explicit formula for the large deviation rate function:

$$\begin{aligned}
\varphi(\mu) &= \sum_{i=1}^n \sum_{j=1,j\neq i}^n \mu^i q_{ij}\left(1 - \exp\left(\frac{1}{2}\ln\frac{\mu^j\pi^i}{\mu^i\pi^j}\right)\right) \\
&= \sum_{i=1}^n \sum_{j=1,j\neq i}^n \mu^i q_{ij}\left(1 - \sqrt{\frac{\mu^j\pi^i}{\mu^i\pi^j}}\right) \\
&= \sum_{i=1}^n \sum_{j=1,j\neq i}^n \frac{q_{ij}}{\pi^j}\left(\mu^i\pi^j - \sqrt{\mu^i\mu^j\pi^i\pi^j}\right) \\
&= \sum_{i=1}^n \sum_{j=i+1}^n \frac{q_{ij}}{\pi^j}\left(\mu^i\pi^j + \mu^j\pi^i - 2\sqrt{\mu^i\mu^j\pi^i\pi^j}\right) \\
&= \sum_{i=1}^n \sum_{j=i+1}^n \frac{q_{ij}}{\pi^j}\left(\sqrt{\mu^i\pi^j} - \sqrt{\mu^j\pi^i}\right)^2.
\end{aligned} \tag{7.6}$$

An equivalent, and perhaps more useful, formula is

$$\varphi(\mu) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{\mu^i}\left(\frac{q_{ij}\sqrt{\pi^i}}{\sqrt{\pi^j}}\right)\sqrt{\mu^j} \tag{7.7}$$

As an alternative approach, we can think of the $\mathbf{X}_{(i)}$ as arising from a discrete time Markov chain and measure the *normalized pair frequencies* along an infinitely long path [11]:

$$\varphi^{(2)}\big(\{\nu_{ij}\};\{\pi_i p_{ij}\}\big) = \sum_{i,j}\nu_{ij}\log\frac{\nu_{ij}}{f_i p_{ij}},\ \ f_i = \sum_k \nu_{ik} = \sum_k \nu_{ki}, \tag{7.8}$$

in which $p_{ij} = \Pr\{X_{n+1} = j | X_n = i\}$ is the transition probability of the Markov chain, which we assume to be aperiodic and irreducible, with ergodic invariant probability $\{\pi_i\}$ given by

$$\sum_i \pi_i p_{ij} = \pi_j.$$

The constraint on the two marginals of ν_{ij} being equal is known as *shift invariance* and is an essential feature of pair frequencies obtained from an infinitely long path. We have $\varphi^{(2)}(\{\nu_{ij}\}) = \infty$ for all $\{\nu_{ij}\}$ without shift invariance.

From $\varphi^{(2)}(\{\nu_{ij};\{\pi_i p_{ij}\})$ it is straightforward, using the method of Lagrange multipliers, to obtain $\varphi^{(1)}(f;\pi)$ for the *normalized singleton frequency*:

$$\begin{aligned}\varphi^{(1)}(\mu;\pi) &= \inf_{\{\nu_{ij}\}}\left\{\varphi^{(2)}\big(\{\nu_{ij}\}\big)\ \middle|\ \sum_k \nu_{ik} = \sum_k \nu_{ki} = \mu_i\right\} \\ &= \sup_{\{\xi_i>0\}}\left\{\sum_i \mu_i \log\frac{\xi_i}{\sum_j p_{ij}\xi_j}\right\}\end{aligned} \tag{7.9}$$

One can show that this is equivalent to the rate function in (7.6).

The Legendre-Fenchel transform of $\varphi^{(1)}(\mu;\pi)$:

$$\psi^{(1)}(h) = \sup_{\mu\in\mathcal{P}(\Omega)}\big\{\langle h,\mu\rangle - \varphi^{(1)}(\mu;\pi)\big\} \tag{7.10}$$

$$= \lim_{n\to\infty}\frac{1}{n}\ln\mathbb{E}\left[\exp\left(n\sum_{i=1}^{n}h_i X_i\right)\right] \tag{7.11}$$

$$= \ln\Lambda\big(\{p_{ij}e^{h_i}\}\big), \tag{7.12}$$

where $\Lambda\big(\mathbf{A}\big)$ is the principal eigenvalue of the positive matrix $\mathbf{A}$.

For continuous time diffusion process in $\mathbb{R}$, (7.9) becomes [11]

$$\varphi^{(1)}(\mu;\pi) = \sup_{\xi(x)>0} -\int_{\mathbb{R}} \left(\frac{\mathcal{L}[\xi](x)}{\xi(x)}\right) d\mu(x), \tag{7.13}$$

in which $\mathcal{L}[\xi]$ is the generator of the diffusion process X_t: $\mathcal{L} = D(x)\frac{d^2}{dx^2} + V(x)\frac{d}{dx}$, $\mathcal{L}^*[\pi] = 0$, and the empirical measure for finite time T

$$\mu^{(T)}(x) = \frac{1}{T}\int_0^T \delta(X_t - x)dt. \tag{7.14}$$

7.2 Geometry of the space of measures

The previous section shows two ways to generalize the setup from this dissertation. Both methods result in a new large deviation rate function. Some of our earlier analysis immediately carries over to the new settings. For instance, it is clear that we should re-define the entropy of a measure μ from (6.3) as $\varphi(\mu)$, where φ is the new rate function. Similarly, we should define the energy function H so that $H(\omega_i) = h_i$, where h_i is the conjugate variable to μ^i in our new large deviation principle.

However, it is not immediately obvious how to generalize other parts of our analysis. In particular, throughout this work we took advantage of the fact that our definition of energy induced an affine structure on $\mathcal{M}(\Omega)$ and $\mathcal{P}(\Omega)$. This is not generally true for other large deviation principles. It is possible, though, to use a more general information geometric approach. As described in appendix A, an affine structure is essentially a special example of a connection from differential geometry. In particular, an affine space is necessarily a flat manifold with a global chart. The affine structures we have discussed are equivalent to the -1 α-divergences on $\mathcal{M}(\Omega)$ and $\mathcal{P}(\Omega)$ (see, for example, [2]), and their geometry has been well explored (at least in the context of finite Ω). At present, our contributions to geometry are limited to a novel pedagogical approach which is relegated to appendix A.

The key insight that should allow us to connect these ideas to more interesting geometric structures comes from [1]. In that work, it was shown that any divergence function on a

manifold M induces a Riemannian metric and a pair of affine connections. A large deviation rate function supplies just such a divergence. (One can think of $\varphi(\mu)$ as the divergence between μ and π. If one replaces π with an arbitrary measure, then one obtains a new divergence.) The geometry we have focused on so far on $\mathcal{M}(\Omega)$ is induced in this manner by the KL divergence, but different divergences will induce different geometries. It is therefore of great interest to find divergence functions that are relevant to thermodynamics or to a particular application and investigate the resulting geometry. We are in the process of exploring the relevant geometry induced by (7.6). One interesting preliminary result is that, in the special case of a 2-state Markov process, we have explicitly calculated the induced Riemannian metric and affine connection and shown that the master equation of the Markov process is the gradient of the cross-entropy $\sum_i \pi^i \ln \mu^i$ under this induced geometry.

We also introduced a new family of divergences on $\mathcal{M}(\Omega)$ in [22] defined by

$$D[\mu, \nu;\ \beta_a, \beta_b, \lambda] = \frac{1}{\beta_a} \int_\Omega \ln \frac{\mathrm{d}\mu}{\mathrm{d}\lambda \oplus \beta_a(\nu \ominus \lambda))}\, \mathrm{d}\mu + \frac{1}{\beta_b} \int_\Omega \ln \frac{\mathrm{d}\nu}{\mathrm{d}\lambda \oplus \beta_b(\mu \ominus \lambda))}\, \mathrm{d}\nu. \tag{7.15}$$

Except for some special cases (in particular, if $\beta_a = \beta_b = 1$ then this reduces to the symmetrized KL divergence), these divergences do not appear to arise from any obvious large deviation principle. In that work, we showed that this quantity was intimately related to the concepts of work and free energy defined in chapter 6 and used it to derive a generalization of Carnot's inequality. It would be interesting to explore the geometry induced by these divergences, and also to investigate whether or not this can be thought of as the consequence of a different large deviation principle.

The second approach from the previous section also provides a useful insight. Important properties of a Markov process can often be much more easily defined in terms of pair frequencies rather than singleton frequencies. This suggests that the spaces $\mathcal{P}(\Omega)$ and $\mathcal{M}(\Omega)$ may be too limited. Instead, it might be more interesting to study the product space of $\mathcal{P}(\Omega)$ and the space of possible transition rates on the edges between states. There have been numerous recent results characterizing the large deviation rate functions of frequencies and "empirical flows" of Markov processes on these spaces (e.g., [4, 6, 5]). These rate functions

could be used to construct an associated divergence, and therefore to induce a geometry.

BIBLIOGRAPHY

[1] S. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences*, 58:183–195, 2010.

[2] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. The American Mathematical Society, 2000.

[3] Paolo Baldi and Mauro Piccioni. A representation formula for the large deviation rate function for the empirical law of a continuous time markov chain. *Statistics and Probability Letters*, 41:107–115, 1999.

[4] Lorenzo Bertini, Alessandra Faggionato, and Davide Gabrielli. From level 2.5 to level 2 large deviations for continuous time Markov chains. *Markov Processes and Related Fields*, 20:545–562, 2014.

[5] Lorenzo Bertini, Alessandra Faggionato, and Davide Gabrielli. Flows, currents, and cycles for markov chains: Large deviation asymptotics. *Stochastic Processes and Their Applications*, 125(7):2786 – 2819, 2015.

[6] Lorenzo Bertini, Alessandra Faggionato, and Davide Gabrielli. Large deviations of the empirical flow for continuous time Markov chains. *Annales de l'I.H.P. Probabilités et Statistiques*, 51(3):867–900, 2015.

[7] R.A. Blythe and M.R. Evans. Lee-Yang zeros and phase transitions in nonequilibrium steady states. *Physical Review Letters*, 89:080601, 2002.

[8] D. A. Dawson. Stochastic evolution equations. *Mathematical Biosciences*, 15:287–316, 1972.

[9] D. A. Dawson. Stochastic evolution equations and related measure processes. *Journal of Multivariate Analysis*, 5:1–52, 1975.

[10] D. A. Dawson. Geostochastic calculus. *The Canadian Journal of Statistics*, 6:143–168, 1978.

[11] Amir Dembo and Ofer Zeitouni. *Large Deviation Techniques and Applications*. Springer-Verlag, Berlin/Heidelberg, Germany, 2nd edition, 2010.

[12] David A. Kessler Erez Aghion and Eli Barkai. From non-normalizable boltzmann-gibbs statistics to infinite-ergodic theory. *Physical Review Letters*, 122:010601, 2019.

[13] Wendell H. Fleming and Michel Viot. Some measure-valued markov processes in population genetics theory. *Indiana University Mathematics Journal*, 28:817–843, 1979.

[14] Jean Gallier. *Geometric Methods and Applications*. Springer, New York, 2nd edition, 2011.

[15] Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics*. Charles Scribner's Sons, London, 1902.

[16] S.R. De Groot and P. Mazur. *Non-equilibrium Thermodynamics*. Dover Publications, New York, 2011.

[17] Hermann Haken. *Synergetics - An introduction: Nonequilibrium phase transitions and self-organization in physics, chemistry and biology*. Springer, Berlin/Heidelberg, Germany, 1983.

[18] Hermann Haken. *Advanced Synergetics: Instability hierarchies of self-organizing systems and devices*. Springer, Berlin/Heidelberg, Germany, 1993.

[19] Hermann Haken. *Information and Self-organization: A macroscopic approach to complex systems*. Springer, Berlin/Heidelberg, Germany, 2010.

[20] Stefan Hilbert, Peter Häanggi, and Jörn Dunkel. Thermodynamic laws in isolated systems. *Physical Review E*, 90:062116, 2014.

[21] Terrel L. Hill. *An Introduction to Statistical Thermodynamics*. Dover Publications, New York, 1986.

[22] Liu Hong, Hong Qian, and Lowell F. Thompson. Representations and divergences in the space of probability measures and stochastic thermodynamics. *Journal of Computational and Applied Mathematics*, 376:112842, 2020.

[23] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3:300–313, 1935.

[24] Thomas G. Kurtz. The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics*, 57(7):2976–2978, 1972.

[25] John M. Lee. *Introduction to Smooth Manifolds*. Springer, New York, 2nd edition, 2013.

[26] T.-D. Lee and C.-N. Yang. Statistical theory of equations of state and phase transitions. ii. lattice gas and ising model. *Physical Review*, 87:410–419, 1952.

[27] Michael Levitt. The birth of computational structural biology. *Nature Structural Biology*, 8:392–393, 2001.

[28] Zhiyue Lu and Hong Qian. Emergence and breaking of duality symmetry in thermodynamic behavior: Repeated measurements and macroscopic limit. *arXiv:2009.12644*, 2020.

[29] Benoit Mandelbrot. On the derivation of statistical thermodynamics from purely phenomenological principles. *Journal of Mathematical Physics*, 5(2):164–171, 1964.

[30] Thierry Mora and William Bialek. Are biological systems poised at criticality? *Journal of Statistical Physics*, 144:268–302, 2011.

[31] David Mumford. The dawning of the age of stochasticity. In *Mathematics Towards The Third Millenium: Convegno internazionale promosso del Centro Linceo Interdisciplinare*, volume 11, pages 107–125. Accademia Nationale dei Lincei, 2000.

[32] Giovanni Pistone and Carlo Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23:1543–1561, 1995.

[33] Hong Qian. Stochastic physics, complex systems and biology. *Quantitative Biology*, 1(1):50–53, Mar 2013.

[34] Hong Qian. Thermodynamic behavior of statistical event counting in time: Independent and correlated measurements. *arXiv:2109.12806*, 2021.

[35] Hong Qian, Ping Ao, Yuhai Tu, and Jin Wang. A framework towards understanding mesoscopic phenomena: Emergent unpredictability, symmetry breaking and dynamics across scales. *Chemical Physics Letters*, 665:153–161, 2016.

[36] Hong Qian, Yu-Chen Cheng, and Lowell F. Thompson. Ternary representation of stochastic change and the origin of entropy and its fluctuations. *arXiv:1902.09536*, 2019.

[37] Hong Qian and Hao Ge. *Stochastic Chemical Reaction Systems in Biology*. Springer, Switzerland, 2021.

[38] H.H. Schaefer. *Topological Vector Spaces*. Springer-Verlag, New York, New York, 2nd edition, 1999.

[39] Eric Smith. Intrinsic and extrinsic thermodynamics for stochastic population processes with multi-level large-deviation structure. *Entropy*, 22(10):1137, 2020.

[40] Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 1. Publish or Perish, Berkeley, 3rd edition, 1999.

[41] Leo Szilard. Über die ausdehnung der phänomenologschen thermodynamik auf die schwankungserscheinungen. *Z. Physik*, 32:753–788, 1925.

[42] Lowell F. Thompson and Hong Qian. Nonlinear stochastic dynamics of complex systems, ii: Potential of entropic force in markov systems with nonequilibrium steady state, generalized gibbs function and criticality. *Entropy*, 18:309, 2016.

[43] Gasper Tkacik, Olivier Marre, Thierry Mora, Dario Amodei, Michael J Berry II, and William Bialek. The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2013:P03011, 2013.

[44] Yue Wang and Hong Qian. Mathematical representation of Clausius' and Kelvin's statements of the second law and irreversibility. *Journal of Statistical Physics*, 179:808–837, 2020.

[45] David V. Widder. *The Laplace Transform*. Princeton University Press, New Jersey, 1946.

[46] C.-N. Yang and T.-D. Lee. Statistical theory of equations of state and phase transitions. i. theory of condensation. *Physical Review*, 87:404–409, 1952.

[47] Ying-Jen Yang and Hong Qian. Unified formalism for entropy production and fluctuation relations. *Phys. Rev. E*, 101:022129, 2020.

[48] Bruno H. Zimm. Contribution to the theory of critical phenomena. *The Journal of Chemical Physics*, 19:1019–1023, 1951.

[49] Robert W. Zwanzig. High-temperature equation of state by a perturbation method i. nonpolar gases. *The Journal of Chemical Physics*, 22:1420–1426, 1954.

Appendix A

GEOMETRY AND AFFINE STRUCTURES

In this appendix, we establish appropriate topologies on an affine structure $(M, V, \oplus)$ and discuss how the connection between affine structures and differential geometry naturally endows M with a smooth structure and a connection, as well as an associated family of Riemannian metrics. The content of this appendix was originally intended as a pedagogical tool, and so it reproduces many standard results from differential geometry that can be found in many textbooks (e.g., [25] or [40]). Throughout, we will assume that M and V are finite dimensional. Because we have restricted ourselves to the finite dimensional case, many of the following results are fairly elementary. A discussion of some of the issues arising in the infinite dimensional case (along with the solutions to many such issues) can be found in [32].

A.1 Topology of affine spaces

To have any hope of doing any analysis on affine spaces, we will need the notions of convergence and continuity. To this end, when we discuss an affine space $(M, V, \oplus)$, we will restrict ourselves to *topological spaces* M and V. In particular, we will insist that the operations of vector addition $+ : V \times V \to V$ and vector scalar multiplication $\cdot : \mathbb{R} \times V \to V$ be continuous (with the usual product topology on all product spaces). Furthermore, we will require each Φ_p and each Φ_p^{-1} to be continuous as well. These turn out to be very serious restrictions.

Lemma A.1.1 *If V is finite dimensional (and we will work exclusively in finite dimensions), then there are only two topologies on V such that $+$ and $\cdot$ are continuous.*

1. *The trivial topology $\{\emptyset, V\}$,*

2. *The topology induced by any vector space isomorphism $T : V \to \mathbb{R}^N$, where $\mathbb{R}^N$ is equipped with the usual topology.*

The proof can be found in most introductions to functional analysis (e.g., [38]). We will assume without further comment that V is equipped with the second of these topologies.

Similarly, the assumption that each Φ_p and Φ_p^{-1} be continuous severely restricts the possible topologies on M. Since we know the topology of V, requiring any bijection between M and V to be a homeomorphism endows M with a unique topology as well. In particular, this means that, for any choice of $p \in M$, there is exactly one topology on M such that $\Phi_p : V \to M$ is a homeomorphism. All that remains is to check that different choices of p result in the same topology.

Lemma A.1.2 *There is a unique topology on M such that each $\Phi_p : V \to M$ is a homeomorphism.*

Proof Choose $p, q, r \in M$. By definition, we have $p \oplus \Phi_p^{-1}(r) = r = q \oplus \Phi_q^{-1}(r)$. It therefore follows that

$$\begin{aligned} r &= p \oplus \left(\Phi_p^{-1}(q) - \Phi_p^{-1}(q) + \Phi_p^{-1}(r)\right) \\ &= \left(p \oplus \Phi_p^{-1}(q)\right) \oplus \left(\Phi_p^{-1}(r) - \Phi_p^{-1}(q)\right) \\ &= q \oplus \left(\Phi_p^{-1}(r) - \Phi_p^{-1}(q)\right). \end{aligned}$$

Since Φ_q is a bijection, we therefore have

$$\Phi_q^{-1}(r) = \Phi_p^{-1}(r) - \Phi_p^{-1}(q).$$

In particular, if we let $r = \Phi_p(x)$, then

$$\left(\Phi_q^{-1} \circ \Phi_p\right)(x) = x - \Phi_p^{-1}(q).$$

This means that $\Phi_q^{-1} \circ \Phi_p$ is continuous. By an identical argument, $\Phi_p^{-1} \circ \Phi_q$ is also continuous.

If we endow M with the topology such that Φ_p is a homeomorphism, then we have

$$\Phi_q = \Phi_p \circ \left(\Phi_p^{-1} \circ \Phi_q\right) \quad \text{and} \quad \Phi_q^{-1} = \left(\Phi_q^{-1} \circ \Phi_p\right) \circ \Phi_p^{-1},$$

which are both compositions of continuous maps, so both are continuous. Therefore, Φ_q is also a homeomorphism.

This means that, at least for finite dimensional V, there is exactly one non-trivial topology on M and V that is compatible with the affine structure, in the sense that vector addition, affine addition and scalar multiplication of vectors are all continuous. From now on, we will assume without comment that the space of points M and space of vectors V in any affine space are always equipped with these topologies.

As an example, take the affine space $(M, V, +)$ with $M = V = \mathbb{R}^n$. The identity map $\mathrm{Id} : V \to \mathbb{R}^n$ is a vector space isomorphism, so we are forced to use the standard topology on $\mathbb{R}^n$ for V. Likewise, if we choose $0 \in M$ as the origin, then the map $\Phi_0 : V \to M$ must be a homeomorphism, but Φ_0 is just the identity map, so M must also be equipped with the standard topology. This gives the comforting result that the affine space $(\mathbb{R}^n, \mathbb{R}^n, +)$ uses the usual topology for both copies of $\mathbb{R}^n$. Indeed, the same is true of $(V, V, +)$ for any vector space V.

Now let's look at the spaces $(\mathcal{M}(\Omega), \mathcal{V}(\Omega), \oplus)$ and $(\mathcal{P}(\Omega), \mathcal{V}_0(\Omega), \oplus)$. To determine the appropriate topologies on $\mathcal{M}(\Omega)$ and $\mathcal{V}(\Omega)$, we first need to find a vector space isomorphism $T : \mathcal{V}(\Omega)V \to \mathbb{R}^N$. To do so, choose an arbitrary bijection $R : \{1, \ldots, N\} \to \Omega$ and define T such that

$$T(f) = ((f \circ R)(1), \ldots, (f \circ R)(N))^T.$$

This is clearly a vector space isomorphism, and so it determines a topology on V. There are no surprises here – this is just the usual topology of pointwise convergence.

Let $\mu^n = \mu(\{(f \circ R)(n)\})$ for each $n = 1, \ldots, N$ (that is, μ^n is the probability mass of $\omega_n \equiv R(n)$) and let λ be the counting measure. The map $\Phi_\lambda : \mathcal{V} \to \mathcal{M}$ is a homeomorphism, as is $\exp \circ T \circ \Phi_\lambda^{-1} : \mathcal{M} \to \mathbb{R}^n$, where $\exp : \mathbb{R}^N \to (\mathbb{R}^+)^N$ maps $(x^1, \ldots, x^N) \mapsto \left(e^{x^1}, \ldots, e^{x^N}\right)$.

This map is given by

$$\left(\exp \circ T \circ \Phi_{\lambda}^{-1}\right)(\mu) = \left(\mu^1, \ldots, \mu^N\right)^T.$$

That is, the map identifying a measure with its vector of probability masses is a homeomorphism, so the requisite topology is the usual one based on open neighborhoods of probability mass. As hoped for, this is the standard topology that one would want to use regardless.

A.2 Differentiation in Affine Spaces

In the previous appendix, we showed that there was a natural choice of topology for an affine space. With that in hand, we can coherently discuss continuous maps between these spaces. That is, if we have two affine spaces $(M, V, \oplus)$ and $(N, W, \oplus)$, then we can decide whether a map $f : M \to N$ is continuous. (Subscripts on $\oplus$ are awkward enough that we will use the same symbol for both addition operations, although one is defined on $M \times V$ and the other is defined on $N \times W$. It should generally be obvious from context which we are using.) We cannot yet, however, say whether such a map is differentiable. If M and V were vector spaces, then we would say that f is differentiable at $x \in M$ if and only if there were a linear operator $L : M \to N$ such that

$$f(x + h) = f(x) + L(h) + o(h), \tag{A.1}$$

where $o(h)$ is a function $g : M \to N$ such that

$$\lim_{h \to 0} \frac{\|g(h)\|}{\|h\|} = 0. \tag{A.2}$$

This definition should strike the reader as suspect, since we have never specified a norm for our vector spaces. However, as long as we restrict ourselves to finite dimensional vector spaces (which we always will) then all norms are equivalent and this definition is the same regardless of our choice of norm.

Unfortunately, neither (A.1) nor (A.2) make any sense in the context of affine spaces. Equation (A.2) is ill-defined since we have no concept of a norm on either M or N. Even if $o(h)$ could be sensibly defined here, equation (A.1) would still be incomprehensible since

neither the notions of addition nor linear operators can be defined. We can solve both of these problems at once by rearranging (A.1). We will say that $f : M \to N$ is differentiable at $p \in M$ if and only if there is a linear operator $L : V \to W$ such that

$$f(q) - f(p) = L(q - p) + o(q - p), \tag{A.3}$$

where $o(q-p)$ represents a function $g : V \to W$ which otherwise satisfies (A.2). Notice that $q - p$ is a vector in V and $f(q) - f(p)$ is a vector in W, so the definitions of linearity and little-oh are sensible and the only addition performed is between vectors in W.

If $M = V$ and $N = W$, with each $\oplus$ as just vector addition, then this definition coincides with the classical version. That is, if $(V, V, +)$ and $(W, W, +)$ are affine spaces, then $f : V \to W$ is differentiable as a map between affine spaces if and only if it is differentiable as a map between vector spaces, where V is given the vector space structure from some isomorphism Φ_p and W is given the vector space structure from some Φ_q, for arbitrary $p \in V$ and $q \in W$. In similar fashion, we can define differentiability of maps between an affine space and a vector space and vice versa. From now on, we will casually conflate the two definitions.

Most of the properties of differential maps between affine spaces (existence/uniqueness, the chain rule, etc.) carry over from classical calculus in the obvious fashion and we will not reproduce their proofs here. Since the linear operator L is unique (if it exists), we will denote $L \equiv Df(p)$ and refer to this operator as "the derivative of f at p". If $Df(p)$ exists for all $p \in M$, we will say that f is differentiable (dropping the "at p").

Lemma A.2.1 *Affine curves are differentiable. In particular, if $c : \mathbb{R} \to M$ is defined by*

$$c(t) = p \oplus (t \cdot x)$$

for some $p \in M$ and some $x \in V$, then we have

$$Dc(t)(s) \equiv c'(t) \cdot s = x \cdot s. \tag{A.4}$$

Proof Choose $t, s \in \mathbb{R}$. We have

$$c(t + s) - c(t) = (p \oplus (t + s)x) - (p \oplus tx) = sx = x \cdot s,$$

as desired.

Affine curves are a special case of affine maps. An affine map $f : M \to N$ is a map such that $(p - q) \mapsto f(p) - f(q)$ is a linear operator (from V to W). It should be clear from the definition that all affine maps are differentiable.

A.3 Smooth Structures

As we have seen, affine structure gives us a way to define differentiable maps on sets without an intrinsic vector space structure. Differential geometry provides a very different (although ultimately quite related) approach. Since our ultimate goal is to study $\mathcal{M}(\Omega)$ and $\mathcal{P}(\Omega)$, it is worth recalling a classical approach to the study of stochastic processes on these spaces. One typically begins by identifying $\mathcal{M}(\Omega)$ or $\mathcal{P}(\Omega)$ with an open subset of $\mathbb{R}^N$. (In particular, $\mathcal{M}(\Omega)$ is typically identified with the positive quadrant of $\mathbb{R}^N$ and $\mathcal{P}(\Omega)$ is identified with an open simplex in this quadrant.) We have no objection to this idea in general. Indeed, the maps Φ_p from an affine space $(M, V, \oplus)$ can be thought of as identifications between a set M and $\mathbb{R}^N$. (Even here, some subtlety is warranted: Each Φ_p identifies M with V, not $\mathbb{R}^N$. As a finite dimensional vector space, we can certainly continue by identifying V with $\mathbb{R}^N$, but there is no canonical way to do so.) What we do object to is the rather arbitrary selection of one particular identification. Instead, we wish to allow a large family of such identifications and to treat them all on the same footing. This is precisely the problem that differential geometry was built to solve.

In this section, we will give a brief overview of basic differential geometry and show how the *tangent bundle* of an affine space, the assembly of all the tangent spaces at different points on a manifold "smooth glued together", can be intimately tied to its associated vector space.

A.3.1 Charts and Atlases

A manifold M can be thought of as a topological space combined with a set of *coordinate charts* (x, U), where $U \subseteq M$ is open and $x : U \to \mathbb{R}^n$ is a homeomorphism onto its image.

(Technically, we need a somewhat regular topology. In particular, we will always assume that M is Hausdorff and second countable. If M is the point set of an affine space, then these requirements are automatically satisfied.) The set U is called the *chart domain*, and we will often refer to x alone as a chart if its domain is obvious from context. If M is also part of an affine space, then it is already endowed with a particular topology. Furthermore, we have already encountered several charts on M. For instance, if $T : V \to \mathbb{R}^n$ is a vector space isomorphism and $p \in M$ and we define $x = T \circ \Phi_p^{-1}$, then (x, M) is a chart. These charts are actually extremely special, and a characteristic feature of affine spaces – their domains are all of M. Most manifolds do not have global charts. For instance, the circle is a one dimensional manifold, but there is no homeomorphism between the circle and any $\mathbb{R}^n$. The existence of such charts makes many structures on M trivial, but also allows for more specialized structures. A chart need not have a global domain, and it need not preserve any sort of vector space structure; for the moment, all we require is that it be a homeomorphism onto its image.

A manifold M is said to have dimension n at p if there is a chart whose domain contains p and whose codomain is $\mathbb{R}^n$. There is a classic, although surprisingly difficult, proof that the dimension of M at p is unique. Moreover, the dimension of M is constant throughout a connected component. As a simple corollary, if M is connected then it has the same dimension at every point, so we say the dimension of M is n. When discussing arbitrary manifolds, we will follow the common (albeit somewhat annoying) convention of using manifold M with dimension n and manifold N with dimension m.

By far the most important manifold we will encounter is $\mathbb{R}^n$ itself. In particular, the identity function $\mathrm{Id} : \mathbb{R}^n \to \mathbb{R}^n$ is a chart, as are its restrictions to any open subset of $\mathbb{R}^n$.

Notice that if (x, U) and (y, V) are charts of $\mathcal{M}$, then the *transition functions*

$$x \circ y^{-1} : y(U \cap V) \to x(U \cap V) \text{ and} \tag{A.5}$$

$$y \circ x^{-1} : x(U \cap V) \to y(U \cap V) \tag{A.6}$$

are also homeomorphisms (between open subsets of $\mathbb{R}^n$). This is vacuously true if $U \cap V$ is

empty.

One should think of a chart as a (local) identification of M with $\mathbb{R}^n$. That is, once we have specified a chart, we will intentionally avoid distinguishing between $p \in M$ and $x(p) \in \mathbb{R}^n$. As long as one is careful to only use results that are independent of the choice of x, this makes life far simpler. This will occasionally prove confusing, particularly in the case of $M = \mathbb{R}^n$ where geometers often use the chart (Id, M) without comment, but it is worth becoming accustomed to.

Similarly, if M and N are manifolds and $f : M \to N$, then we can (at least locally) think of f as a function between Euclidean spaces. In particular, for any $p \in M$, if we choose charts (x, U) on M and (y, V) on M such that $p \in U$ and $f(p) \in V$, then we will identify f with its *coordinate representation* $y \circ f \circ x^{-1} : x(U \cap f^{-1}(V)) \to y(V)$. Since $x(U \cap f^{-1}(V)) \subseteq \mathbb{R}^n$ and $y(V) \subseteq \mathbb{R}^m$, this is just a usual function between Euclidean spaces. If $M = \mathbb{R}^n$ (or a subset of $\mathbb{R}^n$), then we will generally assume $x = \mathrm{Id}$ without comment and write the coordinate representation as $y \circ f : f^{-1}(V) \to y(V)$. Likewise, if $N = \mathbb{R}^m$ (or a subset of $\mathbb{R}^m$), then we will assume that $y = \mathrm{Id}$ without comment and write the coordinate representation as $f \circ x^{-1} : x(U) \to \mathbb{R}^m$. If both M and N are subsets of Euclidean spaces, then we use identity charts on both manifolds and treat f as function between classical spaces. We can now bring the usual tools of real analysis to bear on these coordinate representations. It is particularly useful to note that a function f on M is continuous at p if and only if one (and therefore all) of its coordinate representations is continuous at $x(p)$. This means that continuity is indeed independent of our choice of coordinates.

As a naive first attempt to define smooth maps on manifolds, we could declare that a map f on M is differentiable at p if and only if its coordinate representations are differentiable at $x(p)$. Unfortunately, this definition immediately runs into trouble. For instance, take $M = \mathbb{R}$ and the charts (x, M) and (y, M) where $x(p) = p$ and $y(p) = \sqrt[3]{p}$. Both functions are homeomorphisms onto $\mathbb{R}$, so they are certainly charts. However, if f is a function on M such that $x \circ f \circ x^{-1}$ is smooth, then $y \circ f \circ x^{-1}$ cannot be smooth. This means that differentiability depends, at least to some extent, on the choice of coordinates.

It turns out that this is a fundamental issue – if we allow all homeomorphisms (x, U) to be charts, then we cannot make a suitable definition of differentiability. Instead, we are forced to restrict our definition of "coordinate chart" in some fashion. To this end, we will say that two charts (x, U) and (y, V) are C^∞-related if and only if their transition functions are C^∞. (We will only be interested in infinitely smooth maps from here on. It is relatively straightforward to translate our results to C^r maps, but all proofs would then require tedious counting of derivatives.) An *atlas* $\mathcal{A}'$ of charts on M is a set of charts such that every pair of charts in $\mathcal{A}'$ is C^∞-related. Since we don't want to privilege any particular charts (or worry about keeping track of them all the time), we will always insist on a *maximal atlas* $\mathcal{A}$. That is, $\mathcal{A}$ is the largest possible set of C^∞-related charts that contain $\mathcal{A}'$.

With these definitions in hand, we can now define smooth maps as we wanted to in the first place: A map on M is smooth if and only if one (and therefore all) of its coordinate representations is.

In particular, this means that if (x, U) is a chart on M, then $x : U \to x(U)$ and $x^{-1} : x(U) \to U$ are both smooth.

As an example, consider an n-dimensional vector space V. Any vector space isomorphism $T : V \to \mathbb{R}^n$ is a global chart on V. Moreover, if T and S are two such charts, then $T \circ S^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ and $S \circ T^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ are both linear maps, and therefore C^∞, so T and S are C^∞-related. We can therefore define $\mathcal{A}$ as the maximal atlas containing all vector space isomorphisms from V to $\mathbb{R}^n$. From now on, we will always assume that a vector space V is equipped with this atlas.

Similarly, if $(M, V, \oplus)$ is an affine space, then we would like to endow M with a smooth structure such that the maps Φ_p and Φ_p^{-1} are smooth. To this end, choose $p \in M$ and a smooth chart (x, U) on V. The map $x \circ \Phi_p^{-1} : \Phi_p(U) \to x(U)$ is a homeomorphism, so it is a coordinate chart on M.

Lemma A.3.1 *Any two charts of the form $x \circ \Phi_p^{-1} : \Phi_p(U) \to x(U)$ are C^∞ related. That is, if $p, q \in M$ and (x, U) and (y, W) are charts on V, then the charts $x \circ \Phi_p^{-1}$ and $y \circ \Phi_q^{-1}$*

are C^∞-related.

Proof This is vacuously true if $\Phi_p(U) \cap \Phi_q(W)$ are empty, so suppose that this is not the case. We need to show that the transition maps

$$x \circ \Phi_p^{-1} \circ \Phi_q \circ y^{-1} : y\left(\left(\Phi_q^{-1} \circ \Phi_p\right)(U) \cap W\right) \to x\left(U \cap \left(\Phi_p^{-1} \circ \Phi_q^{-1}\right)(W)\right) \text{ and}$$
$$y \circ \Phi_q^{-1} \circ \Phi_p \circ x^{-1} : x\left(U \cap \left(\Phi_p^{-1} \circ \Phi_q^{-1}\right)(W)\right) \to y\left(\left(\Phi_q^{-1} \circ \Phi_p\right)(U) \cap W\right)$$

are smooth. To see this, note that both $\Phi_p^{-1} \circ \Phi_q : V \to V$ and $\Phi_q^{-1} \circ \Phi_p : V \to V$ are linear maps, and therefore smooth. In particular, the restriction of these maps to any open subset of V are also smooth. This means that each of the above functions is a composition of smooth maps, and therefore smooth. □

From now on, we will assume without comment that every affine space M is endowed with the maximal atlas containing all such charts. With this in mind, we can drop all references to atlases in the future. Whenever we say (x, U) is a chart on M, we mean that it is a smooth chart in the natural atlas defined above.

A.3.2 Standard and Affine Charts

As we have seen, there are a wide variety of possible coordinate charts on any given manifold. (Indeed, that is much of the point of this geometric approach.) However, we are mainly interested in the manifolds $\mathbb{R}^n$, $\mathcal{M}(\Omega)$ and $\mathcal{P}(\Omega)$, and these spaces already come equipped with some standard coordinate systems that will prove useful. The standard coordinate chart on $\mathbb{R}^n$ is just the identity map $y : \mathbb{R}^n \to \mathbb{R}^n$ such that $y(x) = x$. This is (by definition) a chart in our usual maximal atlas on $\mathbb{R}^n$.

The space $\mathcal{M}(\Omega)$ also come with some rather standard charts, which we will take the time to specify here. Recall that an element $\mu \in \mathcal{M}(\Omega)$ is a measure over the finite set Ω. The standard charts on these spaces are only defined up to a choice of order on Ω. With that in mind, choose a bijection $R : \{1, \ldots, n\} \to \Omega$ and, for each $1 \leq i \leq n$, define $\omega^i \equiv R(i)$. We will define the standard chart on $\mathcal{M}(\Omega)$ as $y_R : \mathcal{M}(\Omega) \to \mathbb{R}^n$ such that $y_R^i(\mu) = \mu(\{\omega^i\})$.

In general, the bijection R will be unimportant, so we will simply refer to the standard chart as y.

The space $\mathcal{P}(\Omega)$ is somewhat trickier because it is only $(n-1)$-dimensional. That is, a chart must be of the form $y : U \to \mathbb{R}^{n-1}$. This means that if we want to use probability masses as coordinates then we have to omit one of the states. We will define the standard coordinates as $y_R : \mathcal{P}(\Omega) \to \mathbb{R}^{n-1}$ such that $y_R^i(\mu) = \mu(\{\omega^i\})$. If the bijection R (and therefore the ω_i that is omitted) is obvious from context then we will simply refer to this chart as y.

The charts y are by far the most common coordinate systems for these two spaces, but we will find that they are not always the most convenient. In particular, it will often prove useful to use charts that reflect the affine structure of our spaces. We will therefore use the following construction:

Definition A.3.2 *Let $(M, V, \oplus)$ be an affine space. Choose a point $p_0 \in M$ and a basis v_1, v_2, ..., v_n of V. Every point $p \in M$ can be written uniquely as*

$$p = p_0 \oplus \left(\sum_{i=1}^{n} x^i v_i \right). \tag{A.7}$$

We will define the affine chart $x : M \to \mathbb{R}^n$ *such that $x(p) = (x^1, \ldots, x^n)$, where the x^i are the unique coefficients from* (A.7)*. In general, the base point p_0 and the basis $\{v_i\}$ will be obvious in context, but if we need to be explicit we will write $x(p; p_0, v_1, \ldots, v_n)$ or $x(p; p_0, \{v_i\})$.*

In particular, consider the affine space $(\mathcal{M}(\Omega), \mathcal{V}(\Omega), \oplus)$. If we choose a base point $\pi \in \mathcal{M}(\Omega)$ and the basis $\{\mathbb{1}_{\{\omega^i\}}\}$ of $\mathcal{V}(\Omega)$ (where each $\mathbb{1}_{\{\omega^i\}}$ is the indicator function for the set $\{\omega^i\}$), then the corresponding affine chart is given by $x(\mu) = (x^1, \ldots, x^n)$ with

$$x^i = \ln\left(\frac{\mu^i}{\pi^i}\right), \tag{A.8}$$

where $\mu^i = \mu(\{\omega^i\})$ and $\pi^i = \pi(\{\omega^i\})$. If we choose a general basis $\{v_i\}$, then

$$\sum_{j=1}^{n} v_{ij} x^j = \ln\left(\frac{\mu^i}{\pi^i}\right), \tag{A.9}$$

where $v_{ij} = v_j(\omega^i)$, and therefore

$$x^i = \sum_{j=1}^{n} v^{ij} \ln\left(\frac{\mu^j}{\pi^j}\right), \tag{A.10}$$

where $[v^{ij}]$ is the inverse of $[v_{ij}]$.

Similarly, consider the affine space $(\mathcal{P}(\Omega), \mathcal{V}_0(\Omega), \oplus)$. If we choose a base point $\pi \in \mathcal{P}$ and a basis $\{[v_i]\}$ of $\mathcal{V}_0(\Omega)$ (where each $v_i \in \mathcal{V}(\Omega)$), then the corresponding affine chart is given by $x(\mu) = (x^1, \ldots, x^{n-1})$ with

$$-\ln Z(x;\pi) + \sum_{j=1}^{n-1} v_{ij} x^j = \ln\left(\frac{\mu^i}{\pi^i}\right), \tag{A.11}$$

where $\mu^i = \mu(\{\omega^i\})$, $\pi^i = \pi(\{\omega^i\})$, $v_{ij} = v_j(\omega^i)$ and

$$Z(x;\pi) = \sum_{i=1}^{n} \pi^i \exp\left(\sum_{j=1}^{n-1} v_{ij} x^j\right), \tag{A.12}$$

Unfortunately, unless we choose a rather special basis $\{v_i\}$, we cannot explicitly solve this system for x^i.

A.4 Differentiation on Manifolds

A.4.1 Partial Derivatives

Now that we have a working definition of differentiable maps, we would like to actually calculate their derivatives. As a naive first approach, suppose that $f : M \to N$ is smooth and (x, U) and (y, V) are charts on M and N containing p and $f(p)$ respectively. It would be nice to define the derivative of f as the derivative of its coordinate representation $y \circ f \circ x^{-1}$. Unfortunately, this definition is hopelessly dependent on our choice of charts. To see why this is such a problem, consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and think of x and y as choices of basis. The total derivative of f is supposed to be a linear map from $\mathbb{R}^n \to \mathbb{R}^m$, and so it should not depend on the choice of basis. In contrast, the partial derivatives of f do depend on the choice of basis – they are derivatives in the direction of a basis vector. We can also

specify a matrix representation of the derivative of f – the Jacobian matrix – by choosing particular bases for the domain and range.

In the same spirit, we will define (coordinate dependent) partial derivatives of smooth maps on manifolds and (coordinate dependent) Jacobian matrices. We will start with smooth maps $f : M \to \mathbb{R}$. If (x, U) is a chart on M and $p \in U$, then we will define the ith partial derivative of f at p with respect to the chart x as

$$\left.\frac{\partial f}{\partial x^i}\right|_p = D_i\left(f \circ x^{-1}\right)(x(p)), \tag{A.13}$$

where D_i denotes the ith partial derivative in the classical sense (since $f \circ x^{-1}$ is a function between $\mathbb{R}^n$ and $\mathbb{R}$). Similarly, we will define the Jacobian of f at p with respect to the chart x as

$$Df|_p = D\left(f \circ x^{-1}\right)(x(p)) = \left[\left.\frac{\partial f}{\partial x^1}\right|_p \cdots \left.\frac{\partial f}{\partial x^n}\right|_p\right], \tag{A.14}$$

where $D(f \circ x^{-1})$ is the classical total derivative. Really, we should specify the chart x in our notation, but in practice this is generally obvious from context.

Similarly, if $f : M \to \mathbb{R}^m$ is a smooth map and (x, U) is a chart with $p \in U$, then we define

$$\left.\frac{\partial f^j}{\partial x^i}\right|_p = D_i([f \circ x^{-1}]^j)(x(p)), \tag{A.15}$$

where $[f \circ x^{-1}]^j$ is the jth component of $f \circ x^{-1}$. The Jacobian of f at p with respect to the chart x is then

$$Df|_p = D(f \circ x^{-1})(x(\mu)) = \begin{bmatrix} \left.\frac{\partial f^1}{\partial x^1}\right|_p & \cdots & \left.\frac{\partial f^1}{\partial x^n}\right|_p \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial f^m}{\partial x^1}\right|_p & \cdots & \left.\frac{\partial f^m}{\partial x^n}\right|_p \end{bmatrix}. \tag{A.16}$$

Likewise, if $f : M \to N$ is smooth and (x, U) and (y, V) are charts such that $p \in U$ and $f(p) \in V$, then we define

$$\left.\frac{\partial f^j}{\partial x^i}\right|_p = \left.\frac{\partial [y \circ f]^j}{\partial x^i}\right|_p = D_i\left([y \circ f]^j \circ x^{-1}\right)(x(p)), \tag{A.17}$$

and the Jacobian of f at p is defined just as in (A.15).

We can also extend all of these definitions to all of U in the obvious manner. For instance, $\partial f/\partial x^i|_p$ may be regarded as a function of p with domain U.

As we have emphasized, these definitions depend heavily on the choice of coordinate charts. An important question is how these derivatives change when you choose different charts. In particular, if we have two charts (x, U) and (y, V) and a point $p \in U \cap V$ and we know $\partial f/\partial x^i$ for some function $f : M \to \mathbb{R}$, how can we write $\partial f/\partial y^j$? We have

$$\begin{aligned}
\left.\frac{\partial f}{\partial y^j}\right|_p &= D_j(f \circ y^{-1})(y(p)) \\
&= D_j\left(f \circ x^{-1} \circ x \circ y^{-1}\right)(y(p)) \\
&= \sum_{i=1}^{n} D_i(f \circ x^{-1})((x \circ y^{-1})(y(p))) \cdot D_j([x \circ y^{-1}]^i)(y(p)) \\
&= \sum_{i=1}^{n} D_i(f \circ x^{-1})(x(p)) \cdot D_j([x \circ y^{-1}]^i)(y(p)) \\
&= \sum_{i=1}^{n} \left.\frac{\partial x^i}{\partial y^j}\right|_p \cdot \left.\frac{\partial f}{\partial x^i}\right|_p . \qquad \text{(A.18)}
\end{aligned}$$

This definition extends in the obvious manner to functions $f : M \to \mathbb{R}^m$ and $f : M \to N$.

It is often very useful to think of these definitions in a slightly different light. We have defined the nth partial derivative of f as the nth partial derivative of its coordinate representation $f \circ x^{-1}$. While this leads to extremely satisfying notation, one can also approach derivatives as follows. If $c : J \to M$ is a smooth curve defined on some interval $J \subseteq \mathbb{R}$ with $c(0) = p$ and $f : M \to \mathbb{R}$ is a smooth function, then $f \circ c : J \to \mathbb{R}$ is smooth, and the derivative $(f \circ c)'(0)$ is well defined in the classical sense. In particular, we can choose the curve c such that $c(t) = x^{-1}(x(p) + te_i)$, where e_i is the ith basis vector of $\mathbb{R}^n$. With this definition,

$$\frac{\partial f}{\partial x^i} = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (f \circ c)\,, \qquad \text{(A.19)}$$

where $\mathrm{d}/\,\mathrm{d}t$ is the usual derivative of a function from $\mathbb{R}$ to $\mathbb{R}$. That is, partial derivatives on M are just the derivatives of appropriately chosen curves. In the author's experience, (A.13) is more convenient for computation, but (A.19) provides more useful intuition.

A.4.2 The Tangent Bundle

Much of the study of smooth manifolds amounts to adorning each point with various vector spaces to create new vector bundles. The most important such bundle for our purposes is the tangent bundle TM. There are several ways to construct TM, but we will begin with a method based on equation (A.19). For every $p \in M$, define $C_p^\infty(M)$ as the set of all smooth curves through p. That is,

$$C_p^\infty(M) = \big\{ c : J \to M \mid J \subseteq \mathbb{R} \text{ is an interval,} \\ c \text{ is smooth and } c(0) = p \big\}. \tag{A.20}$$

We don't know how to define the derivatives of these curves in a coordinate independent manner, but we do have the following important result:

Lemma A.4.1 *For any two curves $c_1, c_2 \in C_p^\infty(M)$ and any two charts (x, U) and (y, V) such that $p \in U \cap V$, if*

$$\left.\frac{d}{dt}\right|_0 (x \circ c_1) = \left.\frac{d}{dt}\right|_0 (x \circ c_2)\,,$$

then we also have

$$\left.\frac{d}{dt}\right|_0 (y \circ c_1) = \left.\frac{d}{dt}\right|_0 (y \circ c_2)\,.$$

This means that, although we can't yet meaningfully assign a value to the derivative of a curve, we can say that two curves have the *same* derivative. With this in mind, we will say that two curves in $C_p^\infty(M)$ are equivalent if and only if their derivatives are equal in any (and therefore every) coordinate representation.

We will let T_pM denote the space of all equivalence classes of curves under this relation. This is called the *tangent space* of M at p. For any given chart (x, U) with $p \in M$, we can identify T_pM with $\mathbb{R}^n$ through the map $F_x : T_pM \to \mathbb{R}^n$ defined such that

$$F_x([c]) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_0 (x \circ c)\,. \tag{A.21}$$

This is a well-defined injection by the definition of our equivalence relation. Furthermore, it is a surjection because $F_x([x^{-1}(x(p) + tv)]) = v$ for any $v \in \mathbb{R}^n$. Therefore, F_x is a bijection.

We can use this map to give T_pM a vector space structure by defining

$$[c_1] + [c_2] = F_x^{-1}\left(F_x([c_1]) + F_x([c_2])\right) \text{ and} \tag{A.22}$$

$$\alpha[c_1] = F_x^{-1}\left(\alpha F_x([c_1])\right). \tag{A.23}$$

Furthermore, if (y, V) is another chart with $p \in V$ then we have

$$F_y([c]) = D(y \circ x^{-1})(x(p)) \cdot F_x([c]). \tag{A.24}$$

Since $D(y \circ x^{-1})(x(p))$ is a linear map, this means that the vector space structure induced by F_y is the same as that induced by F_x. Therefore, this vector space structure is coordinate independent.

We will follow traditional notation and denote elements of T_pM by capital letters with the subscript p, such as X_p and Y_p. These elements are called *tangent vectors* at p. If $[c] = X_p$, then we will say that c is a representative curve of X_p.

This definition of T_pM provides excellent geometric intuition for tangent vectors, but is not always the most practical. We will therefore give a very different approach to the same space.

Let $C^\infty(M)$ denote the space of all smooth functions $f : M \to \mathbb{R}$, equipped with the usual pointwise addition and multiplication operations. We will define a *derivation* at p as a linear map $\ell : C^\infty(M) \to \mathbb{R}$ that satisfies

$$\ell(f \cdot g) = \ell(f)g(p) + \ell(g)f(p). \tag{A.25}$$

That is, a derivation is a linear map that satisfies the product rule. We will (temporarily) denote the space of all derivations at p by $\hat{T}_pM$.

We have already encountered many derivations. In particular, if (x, U) is a chart with $p \in U$, then the maps $\partial/\partial x^i|_p : C^\infty \to \mathbb{R}$ such that

$$\left.\frac{\partial}{\partial x^i}\right|_p (f) = \left.\frac{\partial f}{\partial x^i}\right|_p \tag{A.26}$$

are all derivations. That is, partial derivatives are derivations.

If we give $\hat{T}_pM$ the usual pointwise addition and scalar multiplication operations (i.e., $\ell(f+g) = \ell(f) + \ell(g)$ and $\ell(\alpha f) = \alpha\ell(f)$), then it is a vector space. However, it is not immediately clear what the dimension of this space is. Fortunately, a standard theorem of differential geometry states that

Lemma A.4.2 *If M is an n-dimensional manifold, then $\hat{T}_pM$ is an n-dimensional vector space. Moreover, for any coordinate chart (x, U) with $p \in U$, every derivation $\ell \in \hat{T}_pM$ can be written as*

$$\ell = \sum_{i=1}^{n} \ell(x^i) \left.\frac{\partial}{\partial x^i}\right|_p . \tag{A.27}$$

That is, the partial derivatives in any coordinate system form a basis for $\hat{T}_pM$.

(This is one of the few cases where the distinction between C^∞ and C^r is fairly important. If we only required that derivations be C^r for some finite r, then $\hat{T}_pM$ would actually be infinite-dimensional. This is not an insurmountable problem, but it does make life more difficult.)

Notice that if (x, U) and (y, V) are charts with $p \in U \cap V$, then we have (using equation (A.18))

$$\begin{aligned}
\ell(f) &= \sum_{i=1}^{n} \ell^i \left.\frac{\partial f}{\partial x^i}\right|_p \\
&= \sum_{i=1}^{n} \ell^i \left[\sum_{j=1}^{n} \left.\frac{\partial y^j}{\partial x^i}\right|_p \cdot \left.\frac{\partial f}{\partial y^j}\right|_p\right] \\
&= \sum_{j=1}^{n} \left[\sum_{i=1}^{n} \ell^i \left.\frac{\partial y^j}{\partial x^i}\right|_p\right] \cdot \left.\frac{\partial f}{\partial y^j}\right|_p .
\end{aligned}$$

This means that

$$\sum_{i=1}^{n} a^i \left.\frac{\partial f}{\partial x^i}\right|_p = \sum_{i=1}^{n} b^i \left.\frac{\partial f}{\partial y^i}\right|_p \text{ if and only if}$$

$$b^i = \sum_{j=1}^{n} a^j \left.\frac{\partial y^i}{\partial x^j}\right|_p .$$

As the name might suggest, $\hat{T}_pM$ is not so different from T_pM. In particular, if (x, U) is a chart with $p \in U$, then we can define a vector space isomorphism between $\hat{T}_pM$ and T_pM by assigning each derivation ℓ to the equivalence class of the curve

$$t \mapsto x^{-1}\left(x(p) + t\sum_{i=1}^{n} \ell^i e_i\right), \tag{A.28}$$

where $\ell^i = \ell(x^i)$ and e_i is the ith standard basis vector of $\mathbb{R}^n$.

It is relatively straightforward to check that this isomorphism does not depend on the choice of chart. Now that we have this isomorphism, we will routinely identify an equivalence class of curves with its corresponding derivation, and whenever we are working in coordinates (x, U), we will refer to $\ell(f)$ by the coordinate vector $(\ell^1, \ldots, \ell^n)$ such that

$$\ell(f) = \sum_{i=1}^{n} \ell^i \left.\frac{\partial f}{\partial x^i}\right|_p. \tag{A.29}$$

(That is, we will use the derivations $\partial/\partial x^i$ as a basis for T_pM.)

If we want to talk about derivatives on all of M rather than at a single point p, then we need to combine the spaces T_pM in some way. With this goal in mind, we will define the *tangent bundle of* M as

$$TM = \bigsqcup_{p \in M} T_pM. \tag{A.30}$$

That is, the tangent bundle is the disjoint union of tangent spaces. Using our earlier notation, we will typically name elements of the tangent bundle X_p or Y_p. That is, $X_p \in TM$ means that it is a derivation at p. We will also define the standard projection map as $\pi : TM \to M$ such that $\pi(X_p) = p$.

The notion of a tangent bundle allows us to define vector fields. Intuitively, a vector field should just be a tangent vector at every point on a manifold (exactly like the standard picture of a differential equation on $\mathbb{R}^n$). To make this more precise, we define a (smooth) vector field X on a manifold M as a smooth function $X : M \to TM$ such that $\pi(X(p)) = p$. That is, X maps each point to a vector in the tangent space T_pM. Since M and TM are both smooth manifolds, it makes sense to talk about smooth functions between these manifolds.

Fortunately, there is an obvious characterization of vector fields that makes it easy to check smoothness. If (x, U) is a chart on M, then for all $p \in U$ we can write a vector field X as

$$X(p) = \sum_{i=1}^{n} X^i(p) \left.\frac{\partial}{\partial x^i}\right|_p, \tag{A.31}$$

where each X^i is a real-valued function of U. The vector field X is smooth if and only if each X^i is.

The vector field X is naturally identified with a derivation, so we can apply X to any smooth function $f : M \to \mathbb{R}$ in an obvious way. In particular, we say that $X : C^\infty(M) \to C^\infty(M)$ such that $X(f)(p) = X(p)(f)$. In coordinates, this just means that

$$X(f)(p) = \sum_{i=1}^{n} X^i(p) \left.\frac{\partial f}{\partial x^i}\right|_p, \tag{A.32}$$

or equivalently

$$X(f) = \sum_{i=1}^{n} X^i \frac{\partial f}{\partial x^i}. \tag{A.33}$$

It is worth pointing out that a vector field is an example of something called a *section*. In particular, if $\pi : E \to M$ is a vector bundle[1] then a (smooth) function $s : M \to E$ such that $\pi(s(p)) = p$ is called a (smooth) section of the vector bundle. Therefore, vector fields are sections of the tangent bundle.

The connection between vector fields and ordinary differential equations is very important: In coordinate charts every vector field is equivalent to an ordinary differential equation on $\mathbb{R}^n$. To understand this idea, suppose that (x, U) is a chart on M with some point $p \in U$ and $X = \sum_{i=1}^{n} X^i \frac{\partial}{\partial x^i}$ on the domain U. We call a curve $c : J \subset \mathbb{R} \to U$ such that $\left.\frac{\mathrm{d}c}{\mathrm{d}t}\right|_t = X(c(t))$ and $c(0) = p$ the *flow of X through p*. In the coordinates x, this flow is governed by the initial value problem

$$\frac{\mathrm{d}(x \circ c)^i}{\mathrm{d}t}(t) = X^i(c(t)). \tag{A.34}$$

[1]We will not worry too much about the general definition of a vector bundle. Suffice it to say that E is the disjoint union of vector spaces with some nice compatibility properties. In our case, $E = TM$ and π is the usual projection.

The existence and uniqueness of flows on manifolds follows directly from the corresponding theorems for ordinary differential equations on $\mathbb{R}^n$.

A.4.3 Affine Structure and the Tangent Space

In order to identify T_pM with $\mathbb{R}^n$, we used the curves $x^{-1}\left(x(p) + tv\right)$ as representative curves for our equivalence classes. As should be obvious from the definition, these curves are highly dependent on our particular choice of coordinate charts. However, if $(M, V, \oplus)$ is an affine space then we actually have an intrinsic choice of curve: the affine curve $p \oplus tv$.

To see how this works, let's switch to some particularly convenient coordinate charts. In particular, we will choose a base point $p_0 \in M$ and a basis $\{v_i\}$ of V and let $x : M \to \mathbb{R}^n$ be the affine chart defined in A.3.2. That is, we define $x(p) = (x^1, \ldots, x^n)$ such that

$$p = p_0 \oplus \sum_{i=1}^{n} x^i v_i. \tag{A.35}$$

Since $\{v_i\}$ is a basis for V, we can write any $v \in V$ as

$$v = \sum_{i=1}^{n} \ell^i v_i, \tag{A.36}$$

where the ℓ^i are some uniquely defined constants. We therefore have

$$p_0 \oplus tv = p_0 \oplus \left(t \sum_{i=1}^{n} \ell^i v_i \right) = p_0 \oplus \left(\sum_{i=1}^{n} t\ell^i v_i \right). \tag{A.37}$$

This means that

$$x^i(p_0 \oplus tv) = t\ell^i, \tag{A.38}$$

and so

$$\frac{\mathrm{d}}{\mathrm{d}t}(x^i(p_0 \oplus tv)) = \ell^i \tag{A.39}$$

In particular, note that

$$\frac{\mathrm{d}}{\mathrm{d}t}(x(p_0 \oplus tv)) = \frac{\mathrm{d}}{\mathrm{d}t}(x(p_0 \oplus tw))$$

if and only if $v = w$, so each distinct affine curve through p_0 is the representative curve of a distinct tangent vector. Moreover, we can clearly choose v to make $p_0 \oplus tv$ be the

representative curve of any tangent vector X_{p_0} that we want. There is therefore a natural one-to-one correspondence between affine curves through p_0 and tangent vectors in $T_{p_0}M$.

From now on, if $(M, V, \oplus)$ is an affine space, $\{v_i\}$ is a basis of V and $p \in M$, then we will identify the following:

1. The vector $v \in V$ such that
$$v = \sum_{i=1}^{n} \ell^i v_i \tag{A.40}$$

2. The equivalence class of curves $X_p = [p \oplus tv]$.

3. The derivation
$$X_p = \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i}\right|_p,$$
where $x(\cdot\,; p, \{v_i\})$ is an affine chart on M.

This identification lets us think of each tangent space T_pM as a copy of the vector space V.

We will refer to all three of these as *tangent vectors* at p. It is very important to emphasize that the values of ℓ^i depend on the choice of basis $\{v_i\}$, but the vector v does not. That is, the vector v and the derivation X_p will always be identified with the same vector v, regardless of basis, but we will need to choose different ℓ^i for each choice of basis. This is because we have identified tangent vectors with elements of the vector space V, but we have not chosen a canonical representation of V.

It also seems like this identification depends on the base point p, but this dependence is largely illusory. In particular, if we choose $p, q \in M$ and a basis $\{v_i\}$ of V and let $x_p \equiv x(\cdot\,; p, \{v_i\})$ and $x_q \equiv x(\cdot\,; q, \{v_i\})$ be the affine charts at these two base points, then
$$\frac{\partial}{\partial x_p} = \frac{\partial}{\partial x_q}. \tag{A.41}$$
That is, derivatives (and therefore tangent vectors in general) depend on the chosen basis but not on the chosen base point.

Note that this identification furnishes us with a few simple choices of vector fields on an affine space. For example, suppose we fix some point $q \in M$. We can then define a smooth vector field X such that $X(p) = p \ominus q$. It is obvious from this definition that X is independent of the coordinates on M, so we might as well choose a convenient chart. In particular, if we choose a basis $\{v_i\}$ of V then we can work in the affine coordinates $x(\cdot; q, \{v_i\})$. In these coordinates, we identify $p \ominus q$ with $\sum_{i=1}^n x^i(p) \left.\frac{\partial}{\partial x^i}\right|_p$, and so we have

$$X = \sum_{i=1}^{n} x^i \frac{\partial}{\partial x^i}. \tag{A.42}$$

We can find the flow of X through some $p_0 \in M$ by solving the initial value problem

$$\frac{\mathrm{d}x^i}{\mathrm{d}t} = x^i, \text{ with } x^i(0) = x^i(p). \tag{A.43}$$

(The notation for the initial condition is admittedly confusing. Really, this is a differential equation for $(x \circ c)(t)$ and the initial condition should be $x^i(c(0)) = x^i(p)$, but it is standard to drop most references to the curve c.)

More generally, if $A : V \to V$ is a linear map, then we can define a vector field $X(p) = A(p \ominus q)$. The corresponding differential equation in affine coordinates $x(\cdot; q, \{v_i\})$ is then

$$\frac{\mathrm{d}x}{\mathrm{d}t} = Ax. \tag{A.44}$$

This is the affine analogue of linear differential equations. (In fact, linear differential equations on $\mathbb{R}^n$ are exactly this, with $q = 0$.) There isn't really anything special about linear maps A. We can do the same thing with any map $A : V \to V$.

Consider the special case of the affine space $(\mathcal{M}(\Omega), \mathcal{V}(\Omega), \oplus)$. Fix some reference measure $\pi \in \mathcal{M}(\Omega)$ and consider the vector field $X(\mu) = \pi - \mu$. It is most convenient to use the basis $\{v_i\} \equiv \{\mathbb{1}_{\{\omega_i\}}\}$ and corresponding affine chart $x(\cdot; \pi, \{v_i\})$. We then have $\pi \ominus \mu \equiv \sum_{i=1}^n \ln\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega_i)\right) \frac{\partial}{\partial x^i}$. This corresponds to the differential equation

$$\frac{\mathrm{d}x^i}{\mathrm{d}t} = -x^i, \tag{A.45}$$

and so we have $x^i(t) = x_0^i e^t$. In the standard chart $y(\mu) = \mu(\{\omega_i\})$, this translates to

$$y^i(t) = \pi(\{\omega_i\}) \left(\frac{y_0^i}{\pi^i(\{\omega_i\})} \right)^{e^{-t}}. \tag{A.46}$$

These are just affine lines with reparametrized time, but that form is something of a coincidence. These curves are affine lines because the solutions of equation (A.45) are straight lines in $\mathbb{R}^n$. If we use more general maps $A : V \to V$ then the solutions will no longer be lines.

A.5 Connections

A.5.1 Connections

We have already seen that if we have a smooth function on a manifold M we can use the tangent space TM to sensibly talk about first derivatives of this function. In particular, if we choose a coordinate chart (x, U) then we can write a tangent vector $X_p \in TM$ as

$$X_p = \sum_{i=1}^{n} X^i \left. \frac{\partial}{\partial x^i} \right|_p .$$

If $f : M \to \mathbb{R}$ is a smooth function, then taking the derivative of f at p in the direction X_p can be written as

$$X_p f = \sum_{i=1}^{n} X^i \left. \frac{\partial f}{\partial x^i} \right|_p .$$

More generally, we can define a vector field $X : M \to TM$ as

$$X(p) \equiv X_p = \sum_{i=1}^{n} X^i(p) \frac{\partial}{\partial x^i}.$$

The only difference is that the X^i are smooth functions on M instead of constants. We can now write the derivative of f in the direction of X as

$$Xf(p) \equiv X_p f = \sum_{i=1}^{n} X^i(p) \frac{\partial f}{\partial x^i}(p).$$

We would like to be able to take second derivatives in a similar manner. The naive approach fails almost immediately, but it is instructive to see why. If we want to find the second derivative of f along a curve $c : J \to M$, we could try to define it as

$$c''(t) = \lim_{h \to 0} \frac{c'(t+h) - c'(t)}{h}.$$

We can define the first derivatives $c'(t+h)$ and $c'(t)$ unambiguously as tangent vectors. In particular, $c'(t+h) \in T_{c(t+h)}M$ and $c'(t) \in T_{c(t)}M$. Unfortunately, although every term on the right hand side of this definition is well-defined, the entire expression is not. The issue is that $c'(t+h)$ and $c'(t)$ lie in entirely different vector spaces, so there is no sensible way to subtract them. (It does not help to note that they both lie in the vector bundle TM, because this bundle is a *disjoint* union of vector spaces. There is still no way to subtract tangent vectors at different points.)

To solve this issue, we need a way of identifying tangent vectors at different points. Actually, in the general case we need something substantially weaker than that - we need a way to identify tangent vectors at "infinitessimally close" points. Because of the obvious difficulties in defining anything involving infinitessimals, the standard approach in differential geometry skips this identification and only defines the second derivative as a whole, but it is useful to think of this as solving the problem of "identifying" tangent vectors. In particular, notice that identifying two tangent spaces amounts to choosing an isomorphism between them, and we are free to choose a different isomorphism if c moves in a different direction. The space of directions that we could travel in from the point p is just T_pM, which is n-dimensional and the space of isomorphisms between tangent spaces is n^2 dimensional. We should therefore need to make n^3 choices at every point in order to identify a tangent space with all of its "infinitessimally close" neighbors. One should therefore not be surprised to learn that defining second derivatives requires the choice of n^3 smooth functions on M.

To this end, let $\mathscr{T}(M)$ denote the space of all smooth vector fields $X : M \to TM$. Define a *connection* ∇ on M as a map $\nabla : \mathscr{T}(M) \times \mathscr{T}(M) \to \mathscr{T}(M)$ that satisfies the following three properties:

(1) $\nabla(X,Y) \equiv \nabla_X Y$ is linear over $C^\infty(M)$ in the first argument. That is, for any smooth functions $f, g : M \to \mathbb{R}$,

$$\nabla_{fX+gZ} Y = f\nabla_X Y + g\nabla_Z Y.$$

(2) $\nabla_X Y$ is linear over $\mathbb{R}$ in the second argument. That is, for any constants $\alpha, \beta \in \mathbb{R}$,

$$\nabla_X(\alpha Y + \beta Z) = \alpha\nabla_X Y + \beta\nabla_X Z.$$

(3) For any smooth function $f : M \to \mathbb{R}$,

$$\nabla_X(fY) = f\nabla_X Y + (Xf)Y.$$

(One should think of this as the product rule.)

The tangent vector $(\nabla_X Y)(p)$ is the derivative of Y in the direction of X_p at the point p.

Once we have a connection on M, we can define the second derivative of a smooth curve $c : J \to M$ as follows: Let X be a smooth vector field on M such that $X_{c(t)} = c'(t)$ for all $t \in J$. We can then define

$$c''(t) = (\nabla_X X)(c(t)).$$

(This is not an ideal definition. The issue is that we can't always find such a vector field X. For example, if $c(t_1) = c(t_2)$ but $c'(t_1) \neq c'(t_2)$ then we will be stuck. We could repair this definition by somehow restricting our connection to the image of c in some neighborhood of t, but we won't worry about such details here.)

In practice, one always works in local coordinates, so we need to be able to define ∇ in terms of a local coordinate chart. To this end, let (x, U) be a chart on M and let $X, Y \in \mathscr{T}(M)$ be smooth vector fields. We can write

$$X = \sum_{i=1}^{n} X^i \frac{\partial}{\partial x^i} \text{ and } Y = \sum_{i=1}^{n} Y^i \frac{\partial}{\partial x^i},$$

where X^i and Y^i are smooth, real-valued functions on M. One can show that every connection on M can be written as

$$\nabla_X Y = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \left(X^i \frac{\partial Y^k}{\partial x^i} + \Gamma_{ij}^k X^i Y^j \right) \frac{\partial}{\partial x^k},$$

where each $\Gamma^k_{ij} : M \to \mathbb{R}$ is a smooth function on M. The functions Γ^k_{ij} are called the Christoffel symbols of ∇ with respect to x, and ∇ is uniquely determined by these symbols (at least on U).

A.5.2 Identification of Tangent Vectors in an Affine Space

If $(M, V, \oplus)$ is an affine space, then we have a very natural way to identify tangent spaces: We can identify each tangent vector with a vector $v \in V$. We already laid the groundwork for this identification in section A.4.3. In that section, we showed that for any given base point $p \in M$ and basis $\{v_i\}$ of V, there is a natural identification between the vector space V and T_pM. In particular, if we let $x_p : M \to \mathbb{R}^n$ be the corresponding affine chart, then we can identify the tangent vector

$$X_p = \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i_p}\right|_p \tag{A.47}$$

with the vector

$$v = \sum_{i=1}^{n} \ell^i v_i. \tag{A.48}$$

Likewise, if we choose a different base point $q \in M$ and let $x_q(\cdot\,; q, \{v_i\})$ be the affine chart with the same basis and the new base point q, then we can identify

$$X_q = \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i_q}\right|_q \tag{A.49}$$

with the same vector v. This means that we can identify

$$X_p = \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i_p}\right|_p \cong \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i_q}\right|_q = X_q. \tag{A.50}$$

It is somewhat unwieldy to have to use a different chart for every point. Since x_p and x_q are both (global) affine charts with the same basis $\{v_i\}$, we can easily rewrite this in terms of just one chart. In particular, if $x(\cdot\,; r, \{v_i\})$ is an affine chart with *any* base point, then we can identify

$$X_p = \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i}\right|_p \cong \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i}\right|_q = X_q. \tag{A.51}$$

This means that once we choose an affine chart we can think of all the tangent spaces as identical.

The reader is probably quite comfortable with this idea in the context of $\mathbb{R}^n$. Indeed, we are so used to conflating vectors at different base points in $\mathbb{R}^n$ that it is often difficult to notice that they are different mathematical objects. However, in other affine spaces we often work with different coordinate systems, and so these identifications are not as obvious. For instance, consider the space $(\mathcal{M}(\Omega), \mathcal{V}(\Omega), \oplus)$ and the basis $\{\mathbf{1}_{\{\omega^i\}}\}$ of $\mathcal{V}$. Let $\# \in \mathcal{M}(\Omega)$ be the counting measure on Ω and $\mu \in \mathcal{M}(\Omega)$ be some other measure. If we let $x(\cdot\,; \#, \{v_i\})$ be the corresponding affine chart on $\mathcal{M}(\Omega)$ (remember that the base point is irrelevant for this discussion - we chose $\#$ for convenience), then we can use (A.51) to identify derivations at $\#$ with derivations at μ. However, most work with measures is not done in an affine chart; it is done using the standard coordinates $y : \mathcal{M} \to \mathbb{R}^n$, such that $y^i(\mu) = \frac{\mathrm{d}\mu}{\mathrm{d}\#}(\omega^i)$. (In particular, $y^i(\#) = 1$.) In the standard coordinates, we have

$$\frac{\partial}{\partial x^i} = \sum_{j=1}^{n} \frac{\partial y^j}{\partial x^i} \frac{\partial}{\partial y^j} = y^i \frac{\partial}{\partial y^i}, \tag{A.52}$$

so we can identify

$$X_\lambda = \sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial y^i}\right|_\lambda \cong \sum_{i=1}^{n} \ell^i y^i(\mu) \left.\frac{\partial}{\partial y^i}\right|_\mu = X_\mu. \tag{A.53}$$

In other words, we have to weight the coefficients of the tangent vector by the mass at each point. A similar situation arises in every non-trivial affine space.

A.5.3 Trivializations and Induced Connections on Affine Space

We can think of this identification as a vector bundle isomorphism $\phi : TM \to M \times V$. This isomorphism is enough to define a connection, but it is more convenient to instead have a vector bundle isomorphism $\varphi : TM \to M \times \mathbb{R}^n$. We can obtain such a map by further specifying an isomorphism between V and $\mathbb{R}^n$. The obvious method is to choose a basis $\{v_i\}$ of V and then map $v_i \mapsto \mathbf{e}_i$. In order to demonstrate that the choice of basis is not structurally relevant, we will go one step more general and map $v_i \mapsto A\mathbf{e}_i$, where A is an arbitrary element of $\mathrm{GL}(n, \mathbb{R})$.

To make this more explicit, choose a basis $\{v_i\}$ of V and a base point $p_0 \in M$ and let $x(\cdot\,; p_0, \{v_i\})$ be the corresponding affine chart on M. Define $\varphi : TM \to M \times \mathbb{R}^n$ such that

$$\varphi\left(\sum_{i=1}^{n} \ell^i \left.\frac{\partial}{\partial x^i}\right|_p\right) = (p, A\mathbf{v})\,, \tag{A.54}$$

where

$$\mathbf{v} = \sum_{i=1}^{n} \ell^i v_i. \tag{A.55}$$

This is called a trivialization of TM (because $M \times \mathbb{R}^n$ is the trivial bundle).

One of the very nice features of a trivialization is that it induces a connection on M. To construct this connection, we must first choose a basis on $\mathbb{R}^n$. This choice does not affect the final connection (it is already bundled into our choice of A), so we might as well just choose it to be the standard basis. For each i, define $s_i : M \to M \times \mathbb{R}^n$ and $e_i : M \to TM$ such that

$$s_i(p) = (p, \mathbf{e}_i) \text{ and } e_i(p) = \varphi^{-1}(s_i(p)). \tag{A.56}$$

Every vector field $X : M \to TM$ can then be written as

$$X(p) = \sum_{i=1}^{n} X^i(p) \left.\frac{\partial}{\partial x^i}\right|_p = \sum_{i=1}^{n} h^i(p) e_i(p), \tag{A.57}$$

for some uniquely determined functions h^i. (That is, we have declared the functions e_i to be the standard basis vectors for TM.)

We can now define a connection $\nabla : \mathscr{T}(M) \times \mathscr{T}(M) \to \mathscr{T}(M)$ (where $\mathscr{T}(M)$ denotes the space of all smooth vector fields on M) such that

$$\nabla(X, Y) \equiv \nabla_X Y = \nabla_X \left(\sum_{i=1}^{n} h^i e_i\right) = \sum_{i=1}^{n} \left(X h^i\right) e_i. \tag{A.58}$$

In particular, if $X = \partial/\partial x^i$ and $Y = \partial/\partial x^j$ for some fixed i and j, then we have

$$\nabla_X Y = \nabla_X \left(\sum_{i=1}^{n} a_{ij} e_i\right) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial x^i} a_{ij}\right) e_i = \sum_{i=1}^{n} 0 e_i = 0. \tag{A.59}$$

(This follows from the fact that the functions $a_{ij}(p)$ are the *constant* entries of A with respect to the appropriate basis.) Therefore, the Christoffel symbols for ∇ with respect to the affine chart are identically zero. We will write these as

$$\Gamma_{ij}^{k} \equiv 0. \tag{A.60}$$

Note that the Christoffel symbols are, in general, highly dependent on the choice of chart. What we have shown here is that the symbols for ∇ are identically zero with respect to any affine chart, but not necessarily with respect to any other chart.

In general, it is not at all trivial to find a metric (let alone all metrics) that corresponds to a given connection. However, our connection has such a simple form that finding all of the associated metrics is trivial. In general, ∇ is the Levi-Civita connection for a Riemannian metric g if and only if $\Gamma_{ij}^{k} = \Gamma_{ji}^{k}$ and

$$\frac{\partial}{\partial x^{k}} g_{ij} = \sum_{\ell=1}^{n} \left[\Gamma_{ki}^{\ell} g_{\ell j} + \Gamma_{kj}^{\ell} g_{i\ell} \right]. \tag{A.61}$$

If the symbols Γ_{ij}^{k} are at all complicated, then it can be impossible to integrate these differential equations, but in our case it is clear that the only possible metrics have constant coefficients g_{ij}. It is once again worth emphasizing that these coefficients depend heavily on the choice of coordinates. They are constant if we choose $\partial/\partial x^{i}|_{p}$ as the basis vectors for each $T_{p}M$.

This means that there is only one connection that is consistent with a given affine structure, and its symbols with respect to any affine chart must be identically zero. There are many Riemannian metrics that are consistent with this structure, but they all must have constant coefficients with respect to any affine chart. This should not really come as a surprise. We declared affine curves to be the analogue of straight lines, which gives us a natural way to define parallel lines, which means that there is a natural choice of connection. However, we do not have a natural method to determine the angles between lines or the length of line segments, so we are still left with many choices for our metric. The metric g should really

be thought of as an inner product on V. Once we choose this inner product, the connection ∇ gives us a way to "move the inner product around M".

It is worth pointing out that we constructed ∇ using a particular choice of coordinates, but a connection is a well-defined object independent of the choice of chart. In particular, we can write the Christoffel symbols with respect to another coordinate chart (say y) as

$$\hat{\Gamma}^k_{ij} = \sum_{\ell=1}^{n} \frac{\partial y^k}{\partial x^\ell} \left[\sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \Gamma^\ell_{\alpha\beta} \frac{\partial x^\alpha}{\partial y^i} \frac{\partial x^\beta}{\partial y^j} + \frac{\partial^2 x^\ell}{\partial y^i \partial y^j} \right] = \sum_{\ell=1}^{n} \frac{\partial y^k}{\partial x^\ell} \cdot \frac{\partial^2 x^\ell}{\partial y^i \partial y^j}. \tag{A.62}$$

Likewise, if we have a Riemannian metric g whose coefficients with respect to an affine chart x are g_{ij}, then the coefficients with respect to some other coordinate chart y are given by

$$\hat{g}_{ij} = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} g_{\alpha\beta} \frac{\partial x^\alpha}{\partial y^i} \frac{\partial x^\beta}{\partial y^j}. \tag{A.63}$$

As a simple example, consider the two-state sample space $\Omega = \{\omega^1, \omega^2\}$ and a continuous time Markov process on this space. That is, we will look at an Ω-valued Markov process $X(t)$. If we let $\mathbf{p}(t) \equiv \begin{bmatrix} p^1(t) & p^2(t) \end{bmatrix}$, where $p^i(t) = \text{Prob}(X(t) = \omega^i)$, then all such processes are governed by the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = \mathbf{p}(t) \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}, \tag{A.64}$$

where a and b are non-negative constants. We will make the further assumption that a and b are strictly positive. This system has an equilibrium

$$\pi = \left(\pi^1, \pi^2\right) = \left(\frac{b}{a+b}, \frac{a}{a+b} \right), \tag{A.65}$$

and a general solution of the form

$$p^i(t) = \pi^i + \left(p^i(0) - \pi^i\right) e^{-(a+b)t}. \tag{A.66}$$

In many ways, these processes are entirely trivial. In addition to being very easy to solve explicitly, they are always detailed balanced, which lends them many useful properties. Nevertheless, it will prove instructive to examine them through the lens of the geometry and affine spaces we have established above.

We can think of a continuous time Markov process as a flow on $\mathcal{P}(\Omega)$. That is, instead of $\mathbf{p}(t)$ we can think of a curve μ_t of probability measures on Ω. Instead of being a solution to (A.64), μ_t is an integral curve of some vector field on $\mathcal{P}(\Omega)$.

If we want to work on $\mathcal{P}(\Omega)$, then it behooves us to choose some coordinate charts. The easy choice is the standard coordinates $y : \mathcal{P}(\Omega) \to \mathbb{R}$, such that $y(\mu) = \mu(\{\omega^1\})$. This would make

$$y(\mu_t) = \pi^1 + (p^1(0) - \pi^1)e^{-(a+b)t}. \tag{A.67}$$

(There is, of course, nothing special about choosing ω^1. We could have used $y(\mu) = \mu(\{\omega^2\})$ instead and achieved similar results.) This has the advantage of mapping directly to our original, classical notation. However, we have already seen that the standard coordinates may obscure some useful features of our system. To remedy this, it may be more convenient to work in an affine chart.

To construct an affine chart, we need a basis for $\mathcal{V}_0(\Omega)$. This is a 1-dimensional vector space, so we just need to pick one (nonzero) element of $\mathcal{V}_0(\Omega)$. All such vectors are of the form $[v]$, where $v(\omega^1) = -v(\omega^2)$, so for convenience we will choose the basis vector $[v]$ with

$$v(\omega^1) = \frac{1}{2} \text{ and } v(\omega^2) = -\frac{1}{2}. \tag{A.68}$$

Likewise, we need to choose a base point in $\mathcal{P}(\Omega)$. For convenience, we will choose the equilibrium measure π. We can now define the affine chart $x : \mathcal{P}(\Omega) \to \mathbb{R}$, where $x(\mu)$ is the unique value such that

$$\mu = \pi \oplus x(\mu)v. \tag{A.69}$$

In particular, if we evaluate both sides of this equation at $\{\omega^1\}$, we obtain

$$y(\mu) = \frac{\int_{\{\omega^1\}} e^{x(\mu)v}\,\mathrm{d}\pi}{\int_\Omega e^{x(\mu)v}\,\mathrm{d}\pi} = \frac{\pi^1 e^{x(\mu)/2}}{\pi^1 e^{x(\mu)/2} + \pi^2 e^{-x(\mu)/2}}. \tag{A.70}$$

Dropping the (μ) notation, we obtain

$$y = \frac{\pi^1 e^{x/2}}{\pi^1 e^{x/2} + \pi^2 e^{-x/2}}, \tag{A.71}$$

and

$$x = \ln\left(\frac{\pi^2}{\pi^1}\right) + \ln\left(\frac{y}{1-y}\right). \tag{A.72}$$

The following formulas will also prove quite useful:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\pi^1\pi^2}{(\pi^1 e^{x/2} + \pi^2 e^{-x/2})^2} = y \cdot (1-y), \tag{A.73}$$

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{y} + \frac{1}{1-y}. \tag{A.74}$$

Since x is an affine chart, the Christoffel symbol for ∇ in the x-coordinates is $\Gamma^1_{11} = 0$ and the associated Riemannian metric g has the constant coefficient g_{11}. This means that in the standard coordinates y, the Christoffel symbol for ∇ is

$$\hat{\Gamma}^1_{11} = \frac{\partial y}{\partial x}\frac{\partial^2 x}{\partial y^2} = \frac{2y-1}{y(1-y)}, \tag{A.75}$$

and the coefficient of the associated Riemannian metric is

$$\hat{g}_{11} = g_{11}\left(\frac{\partial x}{\partial y}\right)^2 = \frac{g_{11}}{y^2(1-y)^2}. \tag{A.76}$$

Appendix B

POTENTIAL OF ENTROPIC FORCE IN MARKOV SYSTEMS WITH NONEQUILIBRIUM STEADY STATE, GENERALIZED GIBBS FUNCTION AND CRITICALITY. (WITH QIAN, H.) *ENTROPY* 18, 309 (2016)

entropy

MDPI

Article

Potential of Entropic Force in Markov Systems with Nonequilibrium Steady State, Generalized Gibbs Function and Criticality

Lowell F. Thompson [1,2,*,†] and Hong Qian [1,*,†]

[1] Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA
[2] Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99352, USA
* Correspondence: lowell.thompson@pnnl.gov (L.F.T.); hqian@u.washington.edu (H.Q.); Tel.: +1-206-543-2584 (H.Q.)
† These authors contributed equally to this work.

Academic Editors: Hermann Haken and Juval Portugali
Received: 3 May 2016; Accepted: 15 August 2016; Published: 18 August 2016

Abstract: In this paper, we revisit the notion of the "minus logarithm of stationary probability" as a generalized potential in nonequilibrium systems and attempt to illustrate its central role in an axiomatic approach to stochastic nonequilibrium thermodynamics of complex systems. It is demonstrated that this quantity arises naturally through both monotonicity results of Markov processes and as the rate function when a stochastic process approaches a deterministic limit. We then undertake a more detailed mathematical analysis of the consequences of this quantity, culminating in a necessary and sufficient condition for the criticality of stochastic systems. This condition is then discussed in the context of recent results about criticality in biological systems.

Keywords: nonequilibrium steady states; stochastic nonequilibrium thermodynamics; generalized potentials; entropy

1. Introduction

This is part II of a series on stochastic nonlinear dynamics of complex systems. Part I [1] presents a chemical reaction kinetic perspective on complex systems in terms of a mesoscopic stochastic nonlinear kinetic approach (e.g., Delbrück–Gillespie processes) as well as a stochastic nonequilibrium thermodynamics (stoc-NET) in phase space. One particularly important feature of the theory in [1] is that it takes the abstract mathematical concepts seriously—that is, it follows what the mathematics tells us [2]. For example, it was shown that the widely employed *local equilibrium assumption* in the traditional macroscopic theory of NET can be eliminated when one recognizes the fine distinction between the set of random events, the $\mathscr{S}$ in a probability space $(\mathscr{S}, \mathcal{F}, P)$ and a random variable that is defined as an observable on the top of the measurable space, $\mathbf{x} : \mathscr{S} \to \mathbb{R}$. The local equilibrium assumption is needed only when one applies the phase space stoc-NET to physically measurable transport processes [3].

The same chemical kinetic approach can be applied to other biological systems. Biological organisms are complex systems with a large number of heterogeneous constituents, which can be thought of as "individuals". To be able to develop a scientific theory for such a complex system with any predictive power, one must use a probabilistic treatment that classifies the individuals into "statistically identical groups" [4–6]. Thermodynamics and statistical mechanics provide a powerful conceptual framework, as well as a set of tools with which one can comprehend and analyze these systems. The fully developed statistical thermodynamic theory taught in college physics classes is mainly a theory of equilibrium systems. The application of its fundamental ideas, however,

is not limited to just equilibrium systems or molecular processes. Stoc-NET [3,7–10], along with the information theoretical approach [6,11–14], is a further development in this area.

One of the key elements of the theory presented in [1] was the nonequilibrium steady state (NESS) potential, or "energy", defined as the minus logarithm of the stationary probability distribution of a kinetic model. In the past, this quantity has appeared repeatedly in the literature [15–19], but most of the studies focus on its computation. In this paper, we attempt to illustrate its central role as a novel "law of force", a necessary theoretical element in the stoc-NET of complex systems.

Once this connection between energy and probability is established, it is possible to formally define probabalistic quantities analagous to other physical variables such as temperature and entropy. (Notice that the term "entropy" is somewhat overloaded. In the context of probability distributions, it typically refers to the Shannon entropy $S = \int p(x) \ln p(x) \, \mathrm{d}x$. In statistical physics, it is more often used to refer to the Gibbs or Boltzmann entropy, both of which are defined in terms of the volume of some region in the phase space of a Hamiltonian system. These various definitions are related but not equivalent. In this work , particularly in Section 4.1, we define analogues of Gibbs and Boltzmann entropies for probability distributions.) In particular, we extend the notion of critical temperatures to the realm of stationary stochastic processes and find a necessary and sufficient condition for the existence of such criticalities. Loosely speaking, at low temperatures, the dynamics of a stochastic process are dominated by energy considerations and become nearly deterministic (i.e., the system is almost always in a ground state). At high temperatures, the dynamics are dominated by entropic considerations and become nearly uniform (i.e., the system traverses all states, regardless of energy). The former occurs for any stationary process, and the rate of approach is given by the energy. In contrast, the entropic effects typically dominate only at infinite temperatures, but some systems can reach uniformity at a finite temperature. We define such a temperature as critical.

Note that we are not presenting an alternative to existing statistical mechanical literature on criticality and phase transition. Instead, we are attempting to generalize these notions from statistical mechanics to a much broader context where the concept of criticality does not yet exist. One can certainly craft stationary distributions from an equilibrium statistical mechanics problem and apply our theory, but this will not produce results that differ from classical approaches.

The paper is organized as follows: In Section 2, we provide a brief historical review of the use of the negative logarithm of a stationary probability distribution as an energy potential. In Section 2.1, we first look at the history of using minus-log-probability to equilibrium chemical thermodynamics and briefly review Kirkwood's fundamental idea of the potential of mean force and the notion of entropic force. In Section 2.2, we describe two recent results identifying the minus-log-probability as "energy": a self-contained and consistent mesoscopic stoc-NET [20], and a precise agreement between its macroscopic limit and Gibbs' theory [21,22]. These two results provide strong evidence for the validity of such an identification. In Section 2.3, we discuss the legitimacy and centrality of stationary distribution in the "entropy inequality" for a Markov process from a mathematical standpoint. In Section 3, we propose a definition of the "corresponding deterministic dynamics" of a stochastic process using power-scaling of probability densities. In Section 3, we show that the rate of convergence to this corresponding deterministic process coincides with the minus-log-probability definition of energy. With the justifications given in Sections 2 and 3, we carry out a more detailed analysis of such a probability distribution in Section 4. In Section 4.1, terms analogous to Boltzmann's and Gibbs' entropy are defined, along with their corresponding microcanonical partition functions. We also discuss the relative merits of these definitions. In Section 4.2, we prove that the system has a critical temperature if and only if the Gibbs' entropy of the system is asymptotic to the energy. In Section 4.3, we discuss several example distributions in order to emphasize some subtleties in the choice of distribution and to illustrate the connection between this theory and equilibrium statistical mechanics. Finally, in Section 5, the ideas from previous sections are related to some recent results on biological systems.

2. A Novel Law of Force: Potential of Entropic Force

In Boltzmann's statistical mechanics, phenomenological thermodynamics is given a Newtonian mechanical basis. Based on the already well developed concepts of mechanical energy and its conservation, Boltzmann [23] derived the relation

$$p^{eq}(x) \propto e^{-U(x)/k_B T}, \tag{1}$$

where $U(x)$ is the mechanical energy of a microstate x and $p^{eq}(x)$ is the probability of state x when the system is in *thermal equilibrium*—a concept which had also already been well established in thermodynamics via the notion of *quasi-stationary processes*. (It is important to distinguish between a mechanical microstate and a thermodynamic state. A thermodynamic state is a state of recurrent motion, defined by an entire level set $\mathcal{A} = \{x \mid U(x) = E\}$. Thus, Boltzmann [23] also introduced his celebrated entropy $S_B(E) = k_B \ln \Omega(E)$, where S_B is the entropy and $\Omega(E)$ is the number of microstates consistent with a given energy E. That is, $\Omega(E)$ is the cardinality of $\mathcal{A}$. In terms of E, then $p^{eq}(E) \propto \Omega(E) e^{-E/k_B T} = e^{-[E-TS(E)]/k_B T}$.) In a thermodynamic equilibrium, there is no net transport of any kind. (In the thermodynamics before Gibbs, macroscopic transport processes were driven by either a temperature or a pressure gradient in the three-dimensional physical space. In Gibbs' macroscopic chemical thermodynamics, a chemical equilibrium has no net flux in the abstract stoichiometric network. In the current mesoscopic, stochastic thermodynamics, an equilibrium has no net probability transport in an appropriate state space. The notion of detailed balance independently arose in physics [24,25], chemistry [26,27] and in probability theory [28].)

It is also worth noting that Boltzmann's mathematical derivation matched the modern *maximum entropy principle* with the constraint of given mean value for energy, which yields an exponential law for the energy distribution. (The mathematical statements of energy conservation $\sum_{k=1}^{N} E_k = C$ and fixed mean energy $\frac{1}{N}\sum_{k=1}^{N} E_k = \bar{c}$ are equivalent when N is given.)

Inspired by Boltzmann's law (1), generalizations of the concept of equilibrium thermodynamic potentials have been proposed in many studies. These generalizations go by a variety of names: generalized thermodynamic potential, kinetic potential, nonequilibrium potential, pseudo-potential, emergent landscape, etc. [15–19,29]. One of the common features of all these names is that the "potential function" is defined by applying Equation (1) in reverse. One *defines* a potential

$$H(x) = -\ln p^{eq}(x) \tag{2}$$

based on the stationary probability, which can be obtained in many statistical models and whose existence can be mathematically proven for a large class of systems. Most importantly, many systems with stationary probability have non-zero transport flux(es).

In fact, this tradition of taking (2) as a legitimate potential function started in equilibrium statistical chemical thermodynamics. Note that according to Equation (1), the term $-k_B T \ln p^{eq}(x)$ is simply the total mechanical energy of state x, which is known a priori. Therefore, there is no reason to define (2) in studies of a pure mechanical system. However, in statistical *chemical* thermodynamics, one usually does not have a full Hamiltonian function for a complex molecule at hand. It is at this juncture that the notion of a *potential of mean force* [30] enters the theory.

2.1. Equilibrium Potential of Mean Force

Physical chemists deal with complex molecules and force fields. Even though in molecular dynamics (MD) a molecule has a classical mechanical representation in terms of atoms as point masses, the precise potential energy is not known. The force fields in MD have therefore been under intense development over the past 50 years [31]. With such complexities, is it even possible to do statistical mechanics?

Let us first note a very important mathematical equality in connection to Equation (1). We consider a function $U(x)$ with $x = (x_1, x_2)$ where $x \in \mathscr{S} = \mathscr{S}_1 \oplus \mathscr{S}_2$, $x_1 \in \mathscr{S}_1$ and $x_2 \in \mathscr{S}_2$. Then,

$$\begin{aligned} Z(T) &= \int_{\mathscr{S}} e^{-U(x)/k_BT} \mathrm{d}x \\ &= \int_{\mathscr{S}_1} \int_{\mathscr{S}_2} e^{-U(x_1,x_2)/k_BT} \mathrm{d}x_1 \mathrm{d}x_2 \\ &= \int_{\mathscr{S}_1} e^{-\varphi(x_1)/k_BT} \mathrm{d}x_1, \end{aligned} \tag{3}$$

$$\varphi(x_1) \;=\; -k_BT \ln \int_{\mathscr{S}_2} e^{-U(x_1,x_2)/k_BT} \mathrm{d}x_2. \tag{4}$$

Notice that if we consider $\varphi(x_1)$ as a "potential function" for the system in (coarse-grained) state x_1, then we can obtain the same $Z(T)$ using Equation (3), which is in the exact same form as in (1). More importantly, we see that $\varphi(x_1)$ is the free energy with fluctuating x_2 and fixed x_1.

After reading the calculations above, one is naturally led to the question, "what does this potential energy function $\varphi(x_1)$ defined in (4) represent?" J. G. Kirkwood answered this question in a very satisfying manner [30]: it is the potential function of a "mean force", in equilibrium, acting on the system which is fixed at x_1:

$$\begin{aligned} -\tfrac{\mathrm{d}\varphi(x_1)}{\mathrm{d}x_1} &= -\frac{\displaystyle\int_{\mathscr{S}_2} \left(\frac{\partial U(x_1,x_2)}{\partial x_1}\right)_{x_2} e^{-U(x_1,x_2)/k_BT} \mathrm{d}x_2}{\displaystyle\int_{\mathscr{S}_2} e^{-U(x_1,x_2)/k_BT} \mathrm{d}x_2} \\ &= -\int_{\mathscr{S}_2} \left(\tfrac{\partial U(x_1,x_2)}{\partial x_1}\right)_{x_2} p^{eq}(x_2|x_1) \mathrm{d}x_2, \end{aligned} \tag{5}$$

in which $p^{eq}(x_2|x_1)$ is the conditional equilibrium probability distribution for x_2 given x_1, and the partial derivative $-(\partial U(x_1,x_2)/\partial x_1)_{x_2}$ is precisely the mechanical force in the x_1 direction, with the given x_2. Averaging over the fluctuating x_2 with distribution $p^{eq}(x_2|x_1)$, Equation (5) is the mean force on x_1.

In other words, Equation (4) states that the negative logarithm of the marginal probability distribution for x_1 is simply the potential of mean force if one chooses the free energy of the entire system, $F(T) = -k_BT \ln Z(T)$, as the zero energy reference point.

$$-k_BT \ln \int_{\mathscr{S}_2} p^{eq}(x_1, x_2) \mathrm{d}x_2 = \varphi(x_1) - F(T). \tag{6}$$

One of the most important facts, as is clear from (4), is that the potential of mean force $\varphi(x_1)$ is itself a function of temperature. In physical chemistry, one usually builds a statistical mechanical model using such a potential of mean force rather than using a mechanical energy function. That is, one uses a free energy function with certain degrees of freedom fixed and averaged over all the others.

Since $\varphi(x_1)$ is temperature dependent, it has its own energy part and entropy part:

$$\varphi(x_1;T) = \underbrace{\frac{\partial(\varphi/T)}{\partial(1/T)}}_{\text{energy}} \underbrace{-T\left(-\frac{\partial\varphi}{\partial T}\right)}_{\text{entropy}}. \tag{7}$$

A potential of mean force can be purely entropic. One of the best known examples is rubber elasticity, which arises from a Gaussian polymer chain [32]. If the temperature is suddenly dropped to zero, the force (and its associated energy) disappears instantly.

Observing this significant conceptual distance between chemical thermodynamics and its mechanical origin, and the essential statistical nature of Gibbs' energy based on minus-log probability in all modeling practices, it is not surprising that some researchers who mainly work with biochemical thermodynamics strongly feel that one could reformulate statistical thermodynamics (at least in

connection to energy) in terms of a "measure of information" and abandon the very term "entropy", along with its root in mechanics [33].

2.2. Nonequilibrium Steady State Potential

For stochastic models of equilibrium systems, therefore, (2) yields a meaningful free energy function in k_BT units. It embodies an exact coarse-graining procedure. For stochastic models of nonequilibrium steady state with non-zero transport flux, we now have sufficient evidence to suggest that

$$H(x) = -\ln p^{ss}(x), \tag{8}$$

where p^{ss} is a stationary distribution, but may or may not be an equilibrium distribution, is also a meaningful energy function. We start with some conceptual discussions.

First, outside classical mechanics, the question "what is a force and how do we quantify it" is highly non-trival and vague. Onsager, however, introduced the notion of a *thermodynamic force* in his theory of irreversible processes [34]. Intuitively, a force is the cause of an action. In Newtonian mechanics, a force is the cause of a change in the vector $\frac{\mathrm{d}}{\mathrm{d}t}\vec{x}$. However, in an "overdamped world", which emcompasses most of chemistry, biology, and society, a force is actually needed to cause a meaningful movement (i.e., a transport).

In terms of the mathematical theory of stochastic dynamics, there is a universal conception for movement or "dynamics": *Given the option to move to one of many states, a system is most likely to move to the state with the highest stationary probability*. One should immediately note that this statement is highly problematic from a rigrous mathematical standpoint. Nevertheless, at least in one class of systems, the above notion is attainable: the class of systems whose dynamics have an invariant measure that is ergodic.

When discussing statistical mechanics, Montroll and Green have stated that [35] "The aim of statistical mechanics is to develop a formalism from which one can deduce the macroscopic behavior of physical systems composed of a large number of molecules from a specification of the component molecular species, the laws of force which govern intermolecular interactions, and the nature of their surroundings". With the rise of equilibrium chemical thermodynamics, it is clear that the "laws of force" themselves can be discovered from the equilibrium distribution. In fact, most such laws of force in biophysical modeling are statistical in nature and can be seen as entropic forces.

Indeed, "[t]o date no one has succeeded in deriving the laws of nonequilibrium phenomena from the [Newtonian] equations of motion merely by allowing the number of particles involved to become infinite. However, considerable success has been achieved by introducing various statistical hypotheses" [35]. Recent studies have shown that if one identifies $H(x)$ as a "generalized Helmhotz or Gibbs energy function", a complete and consistent mesoscopic thermodynamics can be formulated that includes nonequilibrium steady states [3,20]. Furthermore, if one passes the system from mesoscopic to macroscopic by allowing the number of particles involved and the system's volume to become infinite, two macroscopic thermodynamic laws can be derived [21]. If the mesoscopic system is a general chemical reaction network with detailed balance, the macroscopic emergent potential was shown mathematically to be Gibbs' function $G(x)$, where x_i are the concentrations of chemical species and $\partial G/\partial x_i$ are the chemical potentials for the i-th species. The same theory also proves the existence of, and provides an equation for computing, a generalized Gibbs function for an open chemical reaction network under a chemostat, which approaches a nonequilibrium steady state.

2.3. Stationary Distribution and Entropy Inequalities of Markov Processes

Unless stated otherwise, we will exclusively deal with a denumerable state space $\mathscr{S}$ (either finite or infinite) for the remainder of the paper.

A stronger monotonicity result. The strongest version of a monotoic entropy result that we are aware of is [36,37]

$$\frac{\mathrm{d}}{\mathrm{d}t} D\big[\{p_x(t)\}\|\{q_x(t)\}\big] \equiv \frac{\mathrm{d}}{\mathrm{d}t} \sum_{x\in\mathscr{S}} p_x(t) \ln\left(\frac{p_x(t)}{q_x(t)}\right) \leq 0, \tag{9}$$

in which $p_x(t)$ and $q_x(t)$ are two solutions to the Kolmogorov forward equation with different initial distributions. Equation (9) immediately yields a variety of related inequalities:

(i) When $q_x(t) \equiv \pi_x \ \forall t$, where $\{\pi_x\}$ is a stationary distribution of the Markov process, then (9) is the widely known "free energy theorem" [38,39].

(ii) When $q_x(t) \equiv \pi_x \ \forall t$, and $p_i(0) = \delta_{i\ell}$, one has

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{j\in\mathscr{S}} p_{\ell j}(t) \ln\left(\frac{p_{\ell j}(t)}{\pi_j}\right) \leq 0 \ \forall \ell; \tag{10a}$$

therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t} I\big[\mathbf{x}_t\big\|\mathbf{x}_0\big] = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{\ell,j\in\mathscr{S}} \pi_\ell p_{\ell j}(t) \ln\left(\frac{p_{\ell j}(t)}{\pi_\ell \pi_j}\right) \leq 0, \tag{10b}$$

where $I[\mathbf{x}_t\|\mathbf{x}_0]$ is the mutual information between $\mathbf{x}_0$ and $\mathbf{x}_t$ of a stationary Markov process. Similarly,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(-\sum_{\ell,j\in\mathscr{S}} \pi_\ell p_{\ell j}(t) \ln p_{\ell j}(t)\right) \geq 0. \tag{10c}$$

This result was in [40]. The term inside $(\cdots)$ is the conditional Shannon entropy $H[\mathbf{x}_t|\mathbf{x}_0]$ for the stationary $\mathbf{x}_t$. It is also the Kolmogorov–Sinai (KS) entropy of every t steps of the stationary $\mathbf{x}_t$:

$$\lim_{n\to\infty} \frac{1}{n} H\big[\mathbf{x}_0, \mathbf{x}_t, \mathbf{x}_{2t}, \cdots \mathbf{x}_{nt}\big].$$

The result is more easily understood when interpreted this way: KS entropy quantifies the randomness in a "map". The randomness does not decrease with map composition.

(iii) When $p_x(t) \equiv \pi_x$ (and when we then rename $q_x(t)$ as $p_x(t)$), we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{x\in\mathscr{S}} \pi_x \ln\frac{\pi_x}{p_x(t)} \leq 0. \tag{11}$$

To explain this result more intuitively, we note that the sum in (11) can be interpreted as the information lost when predicting π_x from $p_x(t)$. Roughly speaking, if $t_1 < t_2$, then it takes more information to predict the distant future (π_x) from time t_1 than it does from time t_2 because the prediction from $p_x(t_1)$ has to account for the random events that can happen within the time interval $[t_1, t_2]$.

Filtration and entropy monotonicity. Even though the original Shannon entropy used an implicit uniform prior, the necessity for an explicit prior has been widely discussed in information theory [41,42]. (The entropy with respect to an explicit prior is more accurately called the relative entropy or cross-entropy, and its expression is analagous to the free energy in statistical mechanics.) More importantly, for a continuous random variable, the logarithm of a probability density is simply ill-defined mathematically. All the various monotoic "entropy" results in the previous section provide the legitimacy of using $\{\pi_x\}$ as the reference measure for a Markov process. We would like to argue that this is, in fact, necessary.

We consider a Markov process in a more general setting in this section. Let the triple $(\mathscr{S}, \mathcal{F}, P)$ be a probability space; let $(\mathcal{I}, \leq)$ be a totally ordered index set; and let (S, Σ) be a measurable space.

If $X : \mathcal{I} \times \mathscr{S} \to S$ is a stochastic process, then its natural filtration of $\mathcal{F}$ with respect to X is a sequence $\left\{\mathcal{F}_i^{(X)} \mid i \in \mathcal{I}\right\}$ such that

$$\mathcal{F}_i^{(X)} = \sigma\Big\{ X_j^{-1}(A) \;\Big|\; j \in \mathcal{I}, j \leq i, A \in \Sigma \Big\}. \tag{12}$$

That is, $\mathcal{F}_i^{(X)}$ is the smallest σ-algebra on $\mathscr{S}$ that contains all pre-images of Σ-measurable subsets of S for times j up to i. The definition given in (12) yields a monotonic relation

$$\mathcal{F}_j^{(X)} \subseteq \mathcal{F}_i^{(X)} \text{ if } i, j \in \mathcal{I},\ j \leq i. \tag{13}$$

Such a property is called *non-anticipating*; in other words, "when including the future, the dynamics are at least as random as up to now".

The monotonicity in Equation (13) can be expressed in terms of Shannon's information entropy as

$$H[X_0, X_1, \cdots, X_i] \leq H[X_0, X_1, \cdots, X_i, X_{i+1}]. \tag{14}$$

This inequality is true because $H[X_0, \cdots, X_{i+1}] - H[X_0, \cdots, X_i]$ is the conditional Shannon entropy $H[X_{i+1}|X_1, \cdots, X_i]$, which is never negative.

Notice that Equations (13) and (14) are concerned with the *sequences* of $\left\{X_j \mid j \leq i\right\}$, but the "entropy monotonicity" results in statistical physics deal with *individual* X_i and X_{i+1}; and entropy has deterministic values that are different for different times. The relationship among X_i, X_{i+1}, and the filtration is shown as

$$\begin{array}{ccccc} (\mathscr{S}, \mathcal{F}) & \xrightarrow{X_i} & (S, \Sigma) & \xrightarrow{X_i^{-1}} & \left(\mathscr{S}, F_i^{(X)}\right) \\ \| & & \| & \text{time stepping}\downarrow & \\ (\mathscr{S}, \mathcal{F}) & \xrightarrow{X_{i+1}} & (S, \Sigma) & \xrightarrow{X_{i+1}^{-1}} & \left(\mathscr{S}, F_{i+1}^{(X)}\right) \end{array}. \tag{15}$$

We now consider the information lost from X_i to X_{i+1} when the event ω occurs, i.e., $\ln P_{X_{i+1}}(\omega) - \ln P_{X_i}(\omega)$. Its expected value with respect to the stationary, invariant measure $\mu_\pi(\omega)$ is given by

$$\begin{aligned} \mathbb{E}\left[\ln P_{X_{i+1}} - \ln P_{X_i}\right] &= \int_\Omega \ln\left(\frac{\mathrm{d}P_{X_{i+1}}}{\mathrm{d}P_{X_i}}(\omega)\right) \mathrm{d}\mu_\pi(\omega) \\ &= \int_\Omega \ln\left(\frac{\mathrm{d}P_{X_{i+1}}}{\mathrm{d}\mu_\pi}(\omega)\right) \mathrm{d}\mu_\pi(\omega) - \int_\Omega \ln\left(\frac{\mathrm{d}P_{X_i}}{\mathrm{d}\mu_\pi}(\omega)\right) \mathrm{d}\mu_\pi(\omega). \end{aligned} \tag{16}$$

If both X_i and X_{i+1} are real valued (i.e., $S = \mathbb{R}$) with density functions $f_{X_i}(x)$ and $f_{X_{i+1}}(x)$ respectively, then (16) becomes

$$\mathbb{E}\left[\ln P_{X_{i+1}} - \ln P_{X_i}\right] = \int_{\mathbb{R}} \ln\left(\frac{f_{X_{i+1}}(x)}{\pi(x)}\right) \pi(x)\mathrm{d}x - \int_{\mathbb{R}} \ln\left(\frac{f_{X_i}(x)}{\pi(x)}\right) \pi(x)\mathrm{d}x, \tag{17}$$

where $\pi(x) = \mathrm{d}\mu_\pi/\mathrm{d}x$ is the density of the stationary measure. We know that Equation (17) is never negative; therefore, the mean information lost

$$\int_\Omega \ln\left(\frac{\mathrm{d}P_{X_{i+1}}}{\mathrm{d}P_{X_i}}(\omega)\right) \mathrm{d}\mu_\pi(\omega) \geq 0, \tag{18}$$

or equivalently,

$$H\left[X^{ss} \| X_i\right] \geq H\left[X^{ss} \| X_{i+1}\right] \geq 0, \tag{19}$$

where $X^{ss} : \mathscr{S} \to S$ is a random variable distributed according to the stationary distribution π. This is essentially equivalent to the result in Equation (11).

Equation (18) states that information lost from X_i to X_{i+1}, averaged with respect to the invariant density, is always greater than zero, while Equation (19) suggests that "the infinitely distant future has more information to gain from X_i than from X_{i+1}". There is a subtle difference between these statements and the following: "when including the future, the world is at least as random as up to now". The reason for this, we suggest, is that (18) and (19) require the existence of the stationary measure. Knowing the existence of a stationary behavior, "the future is at least as random as now".

3. Deterministic Correspondence and Infinite β

Any representation of reality requires elements of both chance and determinism. These correspond to the stochastic and deterministic components of complex dynamics. As repeatedly pointed out in [43–45], it is the interaction between these two that yields self-orgranization and complex behavior. Therefore, the ability to "envision" a corresponding deterministic dynamics to some given stochastic dynamics, even when there is no obvious "system size parameter", provides a deeper understanding of complex dynamics. The natural parameter for a stochastic differential equation (SDE) $\mathrm{d}\mathbf{x}(t) = b(\mathbf{x})\mathrm{d}t + a\mathrm{d}B(t)$ is the noise strength a; the natural parameter in classical statistical mechanics is the system's size (or one could use the temperature); and the natural parameter in a Delbrück–Gillespie process is the system's volume.

How can one envision such a deterministic correspondence when no obvious natural parameters exist? It is becoming increasingly common to use the modal value of a distribution as a "deterministic" counter part to the stochastic system. According to this view, a bimodal distribution corresponds to a bistable system. Note it is a widely held misconception that the mean dynamics $\langle \mathbf{x}(t) \rangle$ are the deterministic counterpart of a stochastic $\mathbf{x}(t)$. For an SDE, $\langle \mathrm{d}\mathbf{x}(t) \rangle \neq b(\langle \mathbf{x} \rangle)$ in general. More importantly, while $\langle \mathbf{x}(t) \rangle$ is a non-random function of t, it is *not* a trajectory of any meaningful, self-contained dynamical system. This point is best illustrated by the fact that the differential equation describing $\langle \mathbf{x}(t) \rangle$ usually depends on higher moments like $\langle \mathbf{x}^2(t) \rangle$. Moreover, for a discrete system, even if the mean is defined, it does not usually lie in the same space as $\mathbf{x}(t)$.

We propose the following "deterministic" counterpart for a random variable $\mathbf{x}$ with probability mass function $p^{ss}_{\mathbf{x}}$, and we will show that it is intimately related to the energy defined in (8). We will define the "deterministic" variable $\mathbf{x}_\infty$ as

$$\mathbf{x}_\infty = \lim_{\beta \to \infty} \mathbf{x}_\beta, \tag{20}$$

where

$$p^{ss}_{\mathbf{x}_\beta}(x) = \frac{p^{ss}_{\mathbf{x}}(x)^\beta}{Z(\beta)}, \tag{21}$$

with normalization constant

$$Z(\beta) = \sum_x p^{ss}_{\mathbf{x}}(x)^\beta.$$

The random variable $\mathbf{x}_\infty$ will be concentrated on a finite number of states (the most probable ones of $p_{\mathbf{x}}(x)$) with a probability of 1. In particular, if $p^{ss}_{\mathbf{x}}(x)$ is unimodal, then $\mathbf{x}_\infty$ really will be a deterministic system. On the other hand, if $p^{ss}_{\mathbf{x}}(x)$ is multimodal, then there is no unique deterministic counterpart. Applying this idea to a discrete-state Markov process, the corresponding dynamics become a deterministic transformation as discussed in [46].

It is worth noting that similar definitions are often introduced formally as analogues to inverse-temperature without any discussion of deterministic correspondence, e.g., [12,13]. We spend so much time on the concept in order to emphasize that it arises naturally in a study of stochastic systems without any reference to thermodynamic concepts. The scaling factor β should not just be

thought of as a formal method for introducing temperature to a system, but as a natural feature of any probabilistic system.

With this definition in hand, the obvious question becomes "how fast does the limit in (20) converge?" In the next section, we will try to make this question more rigorous. In the process, we will provide more evidence that $H(x)$ is an important quantity.

Large Deviation Principle for Infinite β

We will now investigate the rate of convergence of the limit in (20). This is a question well suited to the methods of large deviation theory. However, before we can use such methods, we need to frame the question somewhat more rigorously. Strictly speaking, we should be dealing with limits of measures rather than limits of random variables.

Let $(\mathscr{S}, \mathcal{F}, P)$ be a discrete probability space with probability mass function p^{ss} and define the family of measures P^β on $(\mathscr{S}, \mathcal{F})$ whose probability mass functions are given by

$$\begin{aligned} p(x,\beta) &= \tfrac{p^{ss}(x)^\beta}{Z(\beta)}, \text{ where} \\ Z(\beta) &= \textstyle\sum_{x\in\mathscr{S}} p^{ss}(x)^\beta. \end{aligned} \tag{22}$$

As we will show later, this is always possible for $\beta \geq 1$. In addition, let (S, Σ) be a measurable space and choose a function $\sigma : \mathscr{S} \to S$. This defines a family of S-valued random variables $\mathcal{O}_\beta$ where

$$\Pr\left\{\mathcal{O}_\beta = z\right\} = P^\beta\left(\left\{x \in \mathscr{S} \,|\, \sigma(x) = z\right\}\right), \tag{23}$$

where $\sigma : \mathscr{S} \to S$ is a measurable map. In particular, if σ is one-to-one, then $\Pr\left\{\mathcal{O}_\beta = z\right\} = p(\sigma^{-1}(z), \beta)$. The random variables $\mathcal{O}_\beta$ are observables on the measurable space $(\mathscr{S}, \mathcal{F})$. In some cases, the distinction between such measurable processes and the underlying measure space becomes vitally important [3]. As we will see here, though, the rate of convergence of these measures is the same whether we phrase the question in terms of observables or the original measure space.

For unimodal distributions, we know that, as β goes to infinity, the distribution of $\mathcal{O}_\beta$ becomes concentrated on a single value $z^* \in S$. However, it is not clear a priori how the rate of this convergence depends on our choice of $\mathcal{O}$. It is conceivable that different observables could lead to different convergence rates. Moreover, we could eschew observables altogether and work solely with the measures P^β. In this section, we will show that the rate of convergence is identical for a wide range of observables and that it is intimately related to $H(x)$.

Case (i): Let $S = \mathbb{R}$. We will not restrict σ to be one-to-one, but we will assume that if $\sigma(x_1) = \sigma(x_2)$ for some $x_1, x_2 \in \mathscr{S}$, then $p^{ss}(x_1) = p^{ss}(x_2)$. We will let $N(x)$ denote the (necessarily finite) number of elements $y \in \mathscr{S}$ such $\sigma(y) = \sigma(x)$. Finally, let $x^* \in \mathscr{S}$ be a state with maximal probability. We know that

$$\lim_{\beta\to\infty} \Pr\left\{\left|\mathcal{O}_\beta - \sigma(x^*)\right| \geq \eta\right\} = 0 \tag{24}$$

for any $\eta \in \mathbb{R}^+$. In fact, $\Pr\left\{\left|\mathcal{O}_\beta - \sigma(x^*)\right| \geq \eta\right\}$ is a non-increasing step function of η. Under reasonable conditions, we can write

$$\Pr\left\{\left|\mathcal{O}_\beta - \sigma(x^*)\right| \geq \eta\right\} = e^{-\beta I_1(\eta) + o(\beta)}, \tag{25}$$

where

$$I_1(\eta) = -\lim_{\beta\to\infty} \ln \Pr\left\{\left|\mathcal{O}_\beta - \sigma(x^*)\right|\right\}. \tag{26}$$

If we define $\hat{x}_\eta = \text{argmax}_{x\in\mathscr{S}}\{|\sigma(x)-\sigma(x^*)|\}$, then we have

$$\begin{aligned} I_1(\eta) &= -\lim_{\beta\to\infty}\frac{1}{\beta}\ln\left(\frac{1}{Z(\beta)}\sum_{x:|\sigma(x)-\sigma(x^*)|\geq\eta} p^{ss}(x)^\beta\right) \\ &= -\lim_{\beta\to\infty}\frac{1}{\beta}\ln\left(\frac{N(\hat{x}_\eta)p(\hat{x}_\eta,\beta)}{N(x^*)p(x^*,\beta)}\right) \\ &= -\ln\left(\frac{p(\hat{x}_\eta,1)}{p(x^*,1)}\right) \\ &= H(\hat{x}_\eta)-H(x^*). \end{aligned}$$

Case (ii): Instead of creating a somewhat arbitrary family of observables $\mathcal{O}_\beta$, we can also work solely with the measures P^β. To make this more convenient, we will introduce some additional notation.

Let $\mathscr{Y} = H(\mathscr{S}) \subset \mathbb{R}$ and let y^* be the minimum value in $\mathscr{Y}$. For any $h > y^*$, let $\mathscr{S}_h = \{x\in\mathscr{S} \mid H(x)<h\}$ and $\mathscr{Y}_h = \{y\in\mathscr{Y}\mid y<h\}$. Let $\lfloor h\rfloor$ denote the minimum value of $\mathscr{Y}\setminus\mathscr{Y}_h$. Finally, define

$$Z_h(\beta) = \sum_{x\in\mathscr{S}_h} p^{ss}(x)^\beta. \tag{27}$$

We know that $P^\beta(\mathscr{S}\setminus\mathscr{S}_h)$ approaches zero as β goes to infinity. Much like the previous case, we would like to know how quickly this quantity decays. We have

$$P^\beta(\mathscr{S}\setminus\mathscr{S}_h) = e^{-\beta I_2(h)+o(\beta)}, \tag{28}$$

where

$$\begin{aligned} I_2(h) &= -\lim_{\beta\to\infty}\frac{1}{\beta}\ln P^\beta(\mathscr{S}\setminus\mathscr{S}_h) \\ &= -\lim_{\beta\to\infty}\frac{1}{\beta}\ln\left(\frac{1-Z_h(\beta)}{Z(\beta)}\right) \\ &= -\lim_{\beta\to\infty}\frac{1}{\beta}\ln\left(\frac{\sum_{x:-\ln p^{ss}(x)=\lfloor h\rfloor} p^{ss}(x)^\beta}{\sum_{x:-\ln p^{ss}(x)=y^*} p^{ss}(x)^\beta}\right) \\ &= \lfloor h\rfloor - y^*. \end{aligned}$$

In fact, this is in some sense just a special case of case (i). If we choose $\sigma = H$ and let $h = \eta + y^*$, then I_1 and I_2 are identical.

Case (iii): One of the key insights from the theory of large deviations is that in the limit as $\beta\to\infty$ the probability $\Pr\{\mathbf{x}_\beta\notin\mathscr{S}_h\}$ is determined by one particular $x^*\notin\mathscr{S}_h$ – the one with $p(x^*,1)\geq p(x,1)$ for all $x\notin\mathscr{S}_h$. Therefore, one has $\lim_{\beta\to\infty}p(x,\beta)\approx e^{-\beta I_3(x)}$ for any $z\in\mathscr{S}$. This is essentially the same as the WKB (Wentzel–Kramers–Brillouin) ansatz. We then have

$$\begin{aligned} I_3(x) &= -\lim_{\beta\to\infty}\frac{1}{\beta}\ln p(x,\beta) \\ &= -\ln p(x,1)+\lim_{\beta\to\infty}\frac{1}{\beta}\ln\sum_{x\in\mathscr{S}} p^\beta(x,1) \\ &= -\ln\left(\frac{p(x,1)}{p(x^*,1)}\right)+\lim_{\beta\to\infty}\frac{1}{\beta}\ln\left[1+\sum_{x\in\mathscr{S},x\neq x^*}\left(\frac{p(x,1)}{p(x^*,1)}\right)^\beta\right] \\ &= -\ln\left(\frac{p(x,1)}{p(x^*,1)}\right) \\ &= H(x)-H(x^*). \end{aligned}$$

Whether we work with observables on top of $(\mathscr{S}, \mathcal{F})$ or if we work with the underlying measure space itself or even if we make substantial approximations as with the WKB ansatz, we always obtain essentially the same rate of convergence. The energy given in (8) describes the rate of convergence of p^{ss} to its corresponding deterministic system.

4. Entropy, Energy and Criticality in Systems with Generalized Potential

The results of the previous section suggest that $H(x) = -\ln p^{ss}(x)$ is a mathematically relevant quantity and that it can reasonably be interpreted as an energy. We will now investigate some of the consequences of this definition in more detail. In particular, we will shed some light on the distinction between Gibbs and Boltzmann entropies and derive a necessary and sufficient condition for the existence of a critical temperature in stationary stochastic systems.

Let us again suppose that our system takes on possible states from a discrete (finite or countably infinite) set $\mathscr{S}$, and let $p^{ss} : \mathscr{S} \to [0,1]$ be the probability mass function describing the chance that event $x \in \mathscr{S}$ ocurrs. As above, we will define the energy of a state $x \in \mathscr{S}$ as

$$H(x) = -\ln p^{ss}(x). \tag{29}$$

In addition, we will avoid substantial difficulties later if we endow H with units of energy. If we do so, then we can no longer simply write $p^{ss}(x) = e^{-H(x)}$. Instead, we need to introduce another parameter β with units of inverse energy. This gives us

$$p^{ss}(x;\beta) = \frac{1}{Z(\beta)} e^{-\beta H(x)}, \tag{30}$$

where the partition function $Z(\beta)$ is defined as

$$Z(\beta) = \sum_{x \in \mathscr{S}} e^{-\beta H(x)}. \tag{31}$$

Note that the partition function is necessarily a dimensionless quantity, as discussed in [47–49]. These distributions are precisely the probability mass functions of the measures P^{β} defined in Section 3.

With this definition, there is a serious concern that the sum in (31) might not converge. Since p^{ss} is a probability distribution, however, we do know that the sum converges for $\beta = 1$ (in fact, we know that $Z(1) = 1$.) We will spend much of the following sections discussing the cases where the sum in (31) diverges, but for the moment we will simply assume that $Z(\beta)$ is well-defined on some subset of $\mathbb{R}$ containing $[1, \infty)$.

In classical statistical mechanics, one typically has the mechanical energy function in hand before p^{ss} and then shows that the system at finite "temperature" β^{-1} has an equilibrium distribution among the states described by (30). Note that when $\beta \to \infty$, the distribution $p^{ss}(x;\beta)$ converges to a uniform probability distribution on the set of states with minimal H. For certain non-convex $H(x)$, the phenomenon of phase transition occurs [50]. This limit gives precisely the deterministic correspondence described in Section 3.

In a classical statistical mechanical problem, $\mathscr{S}$ is a continuous space describing the positions and momenta of all particles in the system, H is a Hamiltonian for this system and $\beta = (k_B T)^{-1}$ is the inverse temperature. One would then be interested in level sets with constant energy h. In particular, Gibbs' and Boltzmann's entropies are concerned with the phase volume and phase surface area of such level sets.

Unlike in a classical problem, though, our state space $\mathscr{S}$ is arbitrary, and, in general, may not be useful as a phase space. In particular, $\mathscr{S}$ often does not come equipped with a metric or even any sort of order. To remedy this, we will define the rank of a state x as

$$\mathfrak{R}(x) = \#\big|\{y \in \mathscr{S} \mid H(x) \geq H(y)\}\big|, \tag{32}$$

where $\#|\cdot|$ denotes cardinality. That is, the rank of x is the number of states which have lower energy than x (or are at least as probable as x). Since $\mathfrak{R}$ depends on x only through $p^{ss}(x)$, we can unambiguously define the rank in terms of energy as $\mathcal{V} : [0, \infty) \to \mathbb{Z}^+$ as

$$\mathcal{V}(h) = \#\big|\{x \in \mathscr{S} \mid H(x) \leq h\}\big|, \tag{33}$$

so that $\mathfrak{R}(x) = \mathcal{V}(H(x))$ for every $x \in \mathscr{S}$.

Notice that $\mathcal{V}$, as opposed to $\mathfrak{R}$, is no longer defined on a discrete space – it is a function of the continuous variable h. However, because $\mathscr{S}$ is discrete, $\mathcal{V}$ can be written as a non-decreasing piecewise constant function.

It is also worth noting that our assumption of a countable state space cannot be easily relaxed in this approach. If $\mathscr{S}$ were uncountable, then one could not hope to order the states by their rank. Indeed, $\mathfrak{R}$ and $\mathcal{V}$ would generally be infinite for almost all input. Such issues arise because p^{ss} is, by assumption, a probability density with respect to the counting measure. We could have instead assumed that p^{ss} was a density with respect to some other measure (e.g., the Lebesgue measure on $\mathscr{S} = \mathbb{R}$), but this would introduce many other subtleties later on.

4.1. Microcanonical Partition Functions and Entropy

If we take the liberty of treating the derivative of a Heaviside function as a Dirac$-\delta$ function, then we can write $\mathcal{V}$ as

$$\mathcal{V}(h) = \int_0^h \mathrm{d}\mathcal{V}(y) = \int_0^h \frac{\partial \mathcal{V}}{\partial y}\,\mathrm{d}y. \tag{34}$$

It is very important to note that $\partial\mathcal{V}(h)/\partial h$ has units of inverse energy. It is tempting (and often quite useful) to define

$$\Omega(h) = \#\big|\{x \in \mathscr{S} \mid H(x) = h\}\big|, \tag{35}$$

and then write

$$\mathcal{V}(h) = \sum_{n=0}^{\infty} \Omega(h_n), \tag{36}$$

where the sum is taken over the values $h_n \leq h$ such that $\Omega(h_n) > 0$. (For any finite h, note that $\mathcal{V}(h) \leq e^h$ because the distribution p^{ss} sums to 1. The number of distinct values $h_n \leq h$ is no greater than $\mathcal{V}(h)$, so it too is finite.) However, one should keep in mind that $\mathrm{d}\mathcal{V}/\,\mathrm{d}h \neq \Omega(h)$. That is, $\mathrm{d}\mathcal{V}/\,\mathrm{d}h$ is not really just a number of states; it is a density. (This is a common point of confusion in probability theory as well. The probability of an event A should always be written as $\int_A \mathrm{d}F = \int_A (\mathrm{d}F/\mathrm{d}x)\,\mathrm{d}x = \int_A f(x)\,\mathrm{d}x$, where F is the cumulative probability measure and $f = \mathrm{d}F/\mathrm{d}x$ is a density with respect to some other measure. When the other measure is a counting measure, however, it is commonplace to replace the integral with a sum and use the probability mass $p(x) = (\mathrm{d}F/\mathrm{d}x)\,\mathrm{d}x$ instead. This is *numerically* correct but often leads to confusion over units.)

One of the main reasons we have introduced this notation with $\mathcal{V}$ is that it gives us a much more convenient way to write $Z(\beta)$. In particular, we can write Z without reference to the individual states x:

$$Z(\beta) = \sum_{x \in \mathscr{S}} e^{-\beta H(x)} = \int_0^\infty e^{-\beta h}\,\mathrm{d}\mathcal{V}(h). \tag{37}$$

This is exactly the Laplace–Stieltjes transform of $\mathcal{V}$.

It is tempting to rewrite Z as

$$Z(\beta) = \int_0^\infty e^{-\beta h}\left(\frac{\partial \mathcal{V}}{\partial h}\right)\mathrm{d}h = \int_0^\infty e^{-\beta\left(h - (k_B\beta)^{-1}k_B \ln \Omega(h)\right)}\,\mathrm{d}h, \tag{38}$$

and to then identify $\partial\mathcal{V}/\partial h$ as the microcanonical partition function and $k_B \ln \Omega(h)$ as the entropy. Unfortunately, this is entirely wrong. Equation (38) relies on the identification of $\frac{\partial \mathcal{V}}{\partial h}$ with $\Omega(h)$, which

is invalid. This method can be salvaged by introducing a factor Δh with units of energy so that the (38) becomes

$$Z(\beta) = \int_0^\infty \frac{1}{\Delta h} e^{-\beta h} \left(\Delta h \frac{\partial \mathcal{V}}{\partial h} \right) \, \mathrm{d}h, \tag{39}$$

and the entropy becomes

$$S_B = k_B \ln \left(\Delta h \frac{\partial \mathcal{V}}{\partial h} \right). \tag{40}$$

In fact, if we choose Δh as a constant, then this is exactly the Boltzmann entropy. Such a solution is somewhat unsatisfying; the introduction of arbitrary constants to correct units generally suggests a deeper misunderstanding. Worse yet, there is no real reason for Δh to be constant so long as it has the correct units.

A much more satisfying interpretation of Z arises if we integrate by parts, obtaining

$$Z(\beta) = \beta \int_0^\infty e^{-\beta h} \mathcal{V}(h) \, \mathrm{d}h = \beta \int_0^\infty e^{-\beta \left(h - (k_B \beta)^{-1} k_B \ln \mathcal{V}(h) \right)} \, \mathrm{d}h. \tag{41}$$

Here, we can interpret $\mathcal{V}(h)$ as the microcanonical partition function and

$$S_G = k_B \ln \mathcal{V}(h) \tag{42}$$

as the entropy. We have chosen the subscripts G and B to emphasize that S_B corresponds to Boltzmann entropy while S_G corresponds to Gibbs entropy.

There has been much debate over the relative merit of these definitions of entropy in statistical mechanics, e.g., [51–54]. While we do not claim to have resolved this question, Equations (38) and (41) suggest that Gibbs entropy is the more natural choice. Furthermore, as we will see in the next section, Gibbs entropy plays a central role in the notion of criticality.

It is worth noting that the terminology surrounding Boltzmann and Gibbs entropy is not entirely consistent. Most notably, some authors, e.g., [55,56], use the phrase "Boltzmann entropy" to refer to the logarithm of the volume of any phase space region corresponding to a suitable macrostate and use "Gibbs entropy" to refer to the quantity $\int p \ln p \, \mathrm{d}x$, where p is some probability density. Using this terminology, Equations (40) and (42) would both be Boltzmann entropies but would use different macrostates.

Instead, we follow the convention used in, e.g., [51–54] and use "Boltzmann entropy" to indicate the logarithm of the volume of a thin shell in phase space and "Gibbs entropy" to indicate the logarithm of the volume of the interior of such a shell. If the quantity $\int p \ln p \, \mathrm{d}x$ is needed, we will refer to it as Shannon entropy.

4.2. Analyticity of Z as a Function of β

The analyticity of $Z(\beta)$, which is analogous to the partition function in statistical mechanics, is intimately related to phase transitions and critical phenomena [57–60]. Our system has a critical temperature (in the statistical mechanical sense of the term) if and only if the partition function is non-analytic for some $\beta \in (0, \infty)$. Since $Z(\beta)$ is a Laplace transform, we have access to some useful theorems from classical analysis (all of which can be found in [61]).

First, there is some value $\beta_c \in [-\infty, \infty]$ such that $Z(\beta)$ converges for all $\beta \in \mathbb{C}$ with real part greater than β_c and diverges for all $\beta \in \mathbb{C}$ with real part less than β_c. The value β_c is called the abscissa of convergence.

Second, if the state space $\mathscr{S}$ is finite then Z is a sum of finitely many terms and therefore converges for any β (i.e., $\beta_c = -\infty$). However, if $\mathscr{S}$ is infinite, then the partition function will not be analytic for all real β. In particular, it cannot converge when $\beta = 0$ because $Z(0) = \#|\mathscr{S}|$. However, by definition, we know that $Z(\beta)$ converges when $\beta = 1$ since $Z(1)$ is the normalization constant of p^{ss}. For infinite systems, the abscissa of convergence must therefore lie somewhere in $[0, 1]$.

Since the abscissa of convergence is non-negative, we have

$$\beta_c = \limsup_{h\to\infty} \frac{\ln \mathcal{V}(h)}{h}, \tag{43}$$

or

$$k_B\beta_c = \limsup_{h\to\infty} \frac{S_G(h)}{h}. \tag{44}$$

We now know that the partition function is analytic for all complex β with real part greater than β_c, where β_c is found as in (44). However, we have not yet shown that $Z(\beta)$ cannot be extended analytically beyond $\beta = \beta_c$. For a general Laplace–Stieltjes transform, this might be possible (in the worst case, a Laplace transform may have a finite abscissa of convergence but still have an analytic continuation to the entire complex plane). Fortunately, since $\mathcal{V}$ is monotonic, $Z(\beta)$ has a singularity at β_c (this also means that $\beta_c \neq 1$).

This means that the partition function $Z(\beta)$ has a singularity at some positive β_c if and only if S_G is asymptotic to h in the sense of (44). That is, if the Gibbs entropy of the system grows sufficiently fast as a function of energy, it can become dominant in the computation of Z (and therefore p^{ss}) at a finite temperature.

4.3. Examples

So far, we have let our system be very general. The arguments above apply equally well to a wide range of systems—from the single electron of a hydrogen atom (where $\mathscr{S}$ is the set of possible orbits) to the configuration of amino acids in a strand of DNA. It is not immediately clear how (44) might be influenced by the structures of $\mathscr{S}$ and p^{ss}. To illustrate the consequences of our result, we will look at a few examples.

First, we will investigate two so-called "non-degenerate" cases where each state has a distinct probability (i.e., $\Omega \equiv 1$). Since we only care about the rank of states, we will suffer no loss of generality by assuming that $\mathscr{S} = \mathbb{Z}^+$ and that the states are ordered so that $p^{ss}(x) > p^{ss}(y)$ whenever $x < y$. As an example, consider the distribution:

$$p^{ss}(x) = 2^{-x}. \tag{45}$$

We have

$$\begin{aligned} H(x) &= x \ln 2, \\ S_G(h) &= k_B \ln h - k_B \ln \ln 2, \text{ and} \\ \beta_c &= 0. \end{aligned} \tag{46}$$

This distribution, therefore, does not have a nonzero critical temperature (which should not be surprising, since it is exponential).

Alternatively, consider a power law distribution.

$$p^{ss}(x) = \frac{x^{-\alpha}}{\zeta(\alpha)}, \tag{47}$$

where $\alpha > 1$ and ζ is the Riemann zeta function. This gives us

$$\begin{aligned} H(x) &= \alpha \ln x + \ln \zeta(\alpha), \\ S_G(h) &= \tfrac{k_B}{\alpha} \left(h - \ln \zeta(\alpha)\right) \text{ and} \\ \beta_c &= \tfrac{1}{\alpha}. \end{aligned} \tag{48}$$

This means that power law distributions do indeed have a finite critical temperature. This result was already demonstrated in [12] but arises as a special case of our work.

These examples highlight the main feature of criticality: a system will be critical if and only if the probability of a state decays too slowly as a function of rank. That is, critical distributions are fat-tailed in "phase space".

We observe a similar result when Ω is not identically 1 ("degenerate" distributions). For example, consider a distribution where, for each $n \in \mathbb{Z}^+$, there are 2^n states with stationary probability 2^{-2n}. That is, for each $h_n = 2n \ln 2$, we have $\Omega(h_n) = 2^n$. In this case,

$$\mathcal{V}(h) = 2\,(2^n - 1) \text{ for } h_n \leq h < h_{n+1}, \tag{49}$$

and we find that $\beta_c = 1/2$. In light of our previous examples, this should not be surprising: when written as a function of rank, p^{ss} decays like x^{-2} so this β_c is exactly what we expect. However, it also illustrates the importance of how we label our state space.

Suppose that we observed the system given above but that we could not identify each individual state. If instead of observing 2^n distinct states, each with probability 2^{-2n}, we only measured 1 state with probability 2^{-n}, we would then calculate the probability distribution $p^{ss}(x) = 2^{-x}$, for which $\beta_c = 0$. Depending on how states are counted, the distribution could either have a finite critical temperature or not! This distinction is exactly why the partition functions in classical and quantum statistical mechanics differ by a factor of $N!$. The classical version overcounts the number of possible microstates because it assumes particles are distinguishable. Without the correction term, this would often lead to substantially different predictions between the two theories. Fortunately, we know that quantum mechanics is the correct theory and so we are able to choose the correct definition of a microstate.

In many applications, however, we do not know what a true microstate looks like. For example, imagine a particle undergoing a random walk on a lattice X and suppose that we can measure only the distance r between a particle and the origin. It would be natural to define a microstate of this system by the distance between the particle and the origin. If $X = \mathbb{Z}^+$, then this is exactly correct. However, if $X = \mathbb{Z}$, then there are really two microstates for each r. Worse yet, if the lattice is two-dimensional (i.e., $X = \mathbb{Z} \times \mathbb{Z}$), then each r corresponds to a different number of microstates and this number grows without bound. As discussed in Section 2.1, we can still find a reasonable interpretation for the energy of such a system. If we treat each r as a microstate, then $H(r)$ is the potential of mean force in the radial direction. However, our notions of entropy and criticality may change drastically depending on how we define our state space.

For a slightly more involved example, consider the so-called "zipper model" (described in, e.g., [62–64]). This is a highly simplified model of, among other things, the conformation of a double-stranded DNA molecule. Suppose there are N base pairs along the DNA molecule (where N can be a positive integer or ∞; if $N = \infty$ then think of the molecule as having a fixed left end but extending infinitely to the right), each of which can either be linked or broken. We will assume that there is only one possible linked configuration for each base pair but that there are G possible broken configurations for each pair, where G is a positive integer. Furthermore, we will suppose that bonds are only broken from left to right. That is, it is possible for a base pair to be in one of the G broken configurations if and only if every base pair to the left is also broken. (This assumption is not entirely necessary, but makes the analysis simpler. Allowing the bonds to break from both ends does not qualitatively alter the behavior of the system, but makes the formulas that follow somewhat more complicated. On the other hand, allowing arbitrary bonds to be broken will make the state space of our system uncountable when the chain becomes infinite. As we will discuss in the next section, this has important consequences.) Suppose that the energy of a linked base pair is 0 and that the energy of any of the G broken configurations for a single base pair is $E > 0$ if all base pairs to the left are broken and infinite otherwise. When $N = \infty$ and $G > 1$, this system has a phase transition at $\beta_c = \ln G/E$. Otherwise, it has no critical temperature [63]. We will show that this critical behaivor is reproduced using (44).

The state space $\mathscr{S}$ of this system is the collection of all possible allowed configurations of linked and broken base pairs. Each configuration consists of m broken base pairs followed by $N-m$ linked base pairs, and there are G^m distinct states for each m. Notice that $\mathscr{S}$ is finite whenever N is and countably infinite when $N=\infty$. The probability of each of these configurations is given by

$$p^{ss}(x;N)=\frac{1}{Q_N}e^{-mE}, \tag{50}$$

where m is the number of broken base pairs in x and Q_N is a constant that depends on N (but not x). Note that it is not immediately obvious from the previous assumptions that $p^{ss}(x;N)$ is well-defined, but one can show that Q_∞ is non-zero and finite for sufficiently large E (in fact, we can solve for Q_∞ exactly, but for our purposes, it is enough to know that it is finite).

Since $\mathscr{S}$ is finite whenever N is, we know that there is no critical temperature for $p^{ss}(\cdot;N)$ when $N<\infty$, so consider the case where $N=\infty$. The possible energy values are $h_m=mE-\ln Q_\infty$ for any $m\in\mathbb{Z}^+$. The Gibbs entropy is therefore

$$S_G(h_m)=k_B\ln\left(\sum_{k=0}^{m}G^k\right)=k_B\ln\left(\frac{G^{m+1}-1}{G-1}\right), \tag{51}$$

if $G>1$ and $S_G(h_m)=k_B\ln(m+1)$ if $G=1$.

Applying (44), we therefore have

$$\beta_c=\limsup_{m\to\infty}\frac{\ln\left(G^{m+1}-1\right)-\ln(G-1)}{mE-\ln Q_\infty}=\lim_{m\to\infty}\frac{(m+1)\ln G}{mE}=\frac{\ln G}{E}, \tag{52}$$

when $G>1$. If $G=1$, we have

$$\beta_c=\limsup_{m\to\infty}\frac{\ln(m+1)}{mE-\ln Q_\infty}=0. \tag{53}$$

These critical temperatures exactly match the known values and the mechanism for this behavior is easy to see. When $G=1$, the phase-volume $\mathcal{V}(h)$ grows linearly with h, but when $G>1$ the phase-volume grows exponentially. This allows the entropy S_G to keep pace with the energy as h grows, leading to a criticality.

The preceding calculations are quite similar to those used in the equilibrium statistical mechanical approach of Kittel [63], but the procedure is very different in spirit. In Kittel's approach, one finds Q_N for arbitrary N, then uses Q_N to calculate a statistic such as the expected number of broken base pairs. Finally, one takes the limit as $N\to\infty$ and demonstrates that this statistic becomes non-analytic at some finite temperature. In particular, Kittel [63] warns that "it is dangerous to write ... the partition function for $N=\infty$; the correct procedure is to evaluate the thermodynamic quantities for finite N and then to examine the limit". In our approach, we start by finding $p^{ss}(x;\infty)$ (up to a constant). Once we have obtained this distribution, we can calculate $S_G(h)$ for the infinite system and directly obtain β_c. The danger that Kittel describes is still present: our method will fail if $p^{ss}(x;\infty)$ is not well-defined.

This example illustrates a general principle. If an equilibrium statistical mechanical problem has a well-defined equilibrium distribution over a countable state space, then both approaches will identify the same critical temperature.

5. Discussion

It is worth taking a moment to discuss not only what we have shown in the previous sections, but also what we have not shown. We have demonstrated that a stationary distribution over a discrete state space has a finite critical temperature if and only if the Gibbs entropy of the distribution (42) satisfies the relation (44). At such a critical temperature, entropic considerations dominate the dynamics

of the stationary process (as opposed to a deterministic system where energetic considerations are dominant). The terminology used here is deliberately suggestive, but one should not take it too far. The novelty of this description of criticality is not as an alternative to (or even as a complete characterization of) the existing statistical mechanics literature on criticality and phase transitions. Instead, we have attempted to generalize some of the notions from statistical mechanics to the broader context of stationary stochastic processes where the notion of criticality does not exist at present.

In particular, notice that there are phase transitions in equilibrium statistical mechanics that do not seem to fit the description given in Section 4. The Lee–Yang theorem, for instance, describes cases where the partition function becomes zero rather than infinite, and two-dimensional Ising models can exhibit various types of phase transitions.

The key point is that we have assumed, from the outset, the existence of a well-defined stationary probability distribution on a countable state space. Such a distribution has a critical temperature β_c if $Z(\beta)$ approaches either zero or infinity as $\beta \to \beta_c$. Because $p^{ss}(x; \beta = 1)$ is a probability mass function, $Z(\beta)$ cannot become zero for any finite β. That is, a Lee–Yang type criticalities can only occur if the stationary distribution p^{ss} is not well-defined for any temperature. Ising models, on the other hand, may have well-defined equilibrium distributions even in the thermodynamic limit. However, these models typically have an uncountable state space when $N \to \infty$. For such a distribution, the proofs of Section 4 do not hold as written and other types of criticalities may be present.

Our theory does agree with existing statistical mechanics in the following sense: if one can find a well-defined equilibrium distribution over a countable state space (either for given parameter values such as N and V or in the thermodynamic limit), then (44) will identify the same critical temperature as standard methods. That is, this theory will not produce any new criticalities in a classical problem.

Mora and Bialek have also discussed this approach in regards to Ising models [12]. In particular, they showed that systems where $p^{ss}(x; N) \propto \mathfrak{R}(x)^{-\alpha}$ follows a power law have a critical temperature given by $\beta_c = 1/\alpha$ when N goes to infinity. Their result utilized the identification of S_G with S_B, which becomes precise in the thermodynamic limit. In the present paper, we have shown that such an identification is unnecessary and that the critical temperature conditions are still exact in "smaller" systems. Moreover, we have found a broader condition for the existence of a critical temperature, of which the power law relationship is a special case.

After Mora and Bialek's paper, there has been much discussion about the idea that biological systems are poised at a critical point. This idea arose because researchers obtained estimates of p^{ss} for a wide range of biological systems and all appeared to follow some sort of power law. Such a distribution would indicate a non-zero abscissa β_c. The result from Section 4.2 does seem like it should indicate a criticality in such cases, but there are some important caveats worth considering.

First, it is notoriously difficult to calculate tail properties (such as β_c) from an estimated distribution. Estimates of p^{ss} are necessarily based on a finite number of samples and therefore cannot give reliable information about arbitrarilly low probability events, which is required to calculate (44).

Second, and much more insidious, many biological processes are not in a true steady state. The formal analogies we have made with statistical mechanics only make sense in the context of stationary systems. If p^{ss} actually varies slowly with respect to some other variable (most importantly time), then our notion of criticality does not necessarily correspond to any interesting feature of the system. For instance, Schwab, Nemenman and Mehta [65] have shown that slowly varying latent variables can give rise to apparent power law distributions, which necessarily have a non-zero β_c, even in conditionally independent systems.

Acknowledgments: Hong Qian is partially supported by National Institutes of Health (NIH) grant R01GM109964.

Author Contributions: Both authors contributed equally to the research and to writing the paper. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qian, H. Stochastic Nonlinear Dynamics of Complex Systems. I: A Chemical Reaction Kinetic Perspective with Mesoscopic Nonequilibrium Thermodynamics. **2016**, arXiv:1605.08070.
2. Guicciardini, N. *Isaac Newton on Mathematical Certainty and Method*; The MIT Press: Cambridge, MA, USA, 2009.
3. Qian, H.; Kjelstrup, S.; Kolomeisky, A.B.; Bedeaux, D. Entropy production in mesoscopic stochastic thermodynamics: Nonequilibrium kinetic cycles driven by chemical potentials, temperatures, and mechanical forces. *J. Phys. Condens. Matter* **2016**, *28*, 153004.
4. Erdi, P.; Lente, G. *Stochastic Chemical Kinetics: Theory and (Mostly) Systems Biological Applications*; Springer: Berlin/Heidelberg, Germany, 2014.
5. Kurtz, T.G. The relationship between stochastic and deterministic models for chemical reactions. *J. Chem. Phys.* **1972**, *57*, 2976–2978.
6. Bialek, W. *Biophysics: Searching for Principles*; Princeton University Press: Princeton, NJ, USA, 2012.
7. Van den Broeck, C.; Esposito, M. Ensemble and trajectory thermodynamics: A brief introduction. *Physica A* **2015**, *418*, 6–16.
8. Zhang, X.-J.; Qian, H.; Qian, M. Stochastic theory of nonequilibrium steady states and its applications (Part I). *Phys. Rep.* **2012**, *510*, doi:10.1016/j.physrep.2011.09.002.
9. Seifert, U. Stochastic thermodynamics, fluctuation theorems, and molecular machines. *Rep. Prog. Phys.* **2012**, *75*, 126001.
10. Jarzynski, C. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Ann. Rev. Condens. Matter Phys.* **2011**, *2*, 329–351.
11. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
12. Mora, T.; Bialek, W. Are Biological Systems Poised at Criticality? *J. Stat. Phys.* **2011**, *144*, 268–302.
13. Tkacik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, M.J.; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**, *2013*, P03011.
14. Bialek, W.; Ranganathan, R. Rediscovering the Power of Pairwise Interactions. **2008**, arXiv:0712.4397.
15. Graham, R.; Haken, H. Generalized thermodynamic potential for Markoff systems in detailed balance and far from thermal equilibrium. *Zeitschrift für Physik* **1971**, *243*, 289–302.
16. Kubo, R.; Matsuo, K.; Kitahara, K. Fluctuation and relaxation of macrovariables. *J. Stat. Phys.* **1973**, *9*, 51–96.
17. Nicolis, G.; Lefever, R. Comment on the kinetic potential and the maxwell construction in non-equilibrium chemical phase transitions. *Phys. Lett. A* **1977**, *62*, 469–471.
18. Yin, L.; Ao, P. Existence and construction of dynamical potential in nonequilibrium processes without detailed balance. *J. Phys. A Math. Gen.* **2006**, *39*, 8593–8601.
19. Feng, H.; Wang, J. Potential and flux decomposition for dynamical systems and non-equilibrium thermodynamics: curvature, gauge field, and generalized fluctuation-dissipation theorem. *J. Chem. Phys.* **2011**, *135*, 234511.
20. Ge, H.; Qian, H. The physical origins of entropy production, free energy dissipation and their mathematical representations. *Phys. Rev. E* **2010**, *81*, 051133.
21. Ge, H.; Qian, H. Mesoscopic Kinetic Basis of Macroscopic Chemical Thermodynamics: A Mathematical Theory. **2016**, arXiv:1601.03159.
22. Ge, H.; Qian, H. Mathematical Formalism of Nonequilibrium Thermodynamics for Nonlinear Chemical Reaction Systems with General Rate Law. **2016**, arXiv:1604.07115.
23. Boltzmann, L. *Lectures on Gas Theory*; University of California Press: Berkeley, CA, USA, 1964.
24. Maxwell, J.C. On the dynamical theory of gases. *Philos. Trans. R. Soc. Lond.* **1867**, *157*, 49–88.
25. Boltzmann, L. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Wissenschaftliche Abhandlungen* **1872**, *1*, 316–402. (In German)
26. Wegscheider, R. Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reactionskinetik homogener Systeme. *Monatshefte für Chemie und Verwandte Teile Anderer Wissenschaften* **1901**, *22*, 849–906. (In German)
27. Lewis, G. N. A new principle of equilibrium. *Proc. Natl. Acad. Sci. USA* **1925**, *11*, 179–183.
28. Kolmogoroff, A. Zur theorie der Markoffschen ketten. *Math. Ann.* **1936**, *112*, 155–160. (In German)

29. Qian, H.; Ge, H. Mesoscopic biochemical basis of isogenetic inheritance and canalization: Stochasticity, nonlinearity, and emergent landscape. *Mol. Cell. Biomech.* **2012**, *9*, doi:10.3970/mcb.2012.009.001.
30. Kirkwood, J.G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.
31. Levitt, M. The birth of computational structural biology. *Nat. Struct. Biol.* **2001**, *8*, 392–393.
32. Hill, T.L. *An Introduction to Statistical Thermodynamics*; Dover: New York, NY, USA, 1960.
33. Ben-Naim, A. *A Farewell to Entropy: Statistical Thermodynamics Based on Information*; World Scientific: Singapore, Singapore, 2008.
34. Onsager, L. Reciprocal relations in irreversible processes. I. *Phys. Rev.* **1931**, *37*, 405–426.
35. Montroll, E.W.; Green, M.S. Statistical mechanics of transport and nonequilibrium processes. *Annu. Rev. Phys. Chem.* **1954**, *5*, 449–476.
36. Lindblad, G. Entropy, information and quantum measurements. *Commun. Math. Phys.* **1973**, *33*, 305–322.
37. Voigt, J. Stochastic operators, information, and entropy. *Commun. Math. Phys.* **1981**, *81*, 31–38.
38. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 1991.
39. Qian, H. Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys. Rev. E* **2001**, *63*, 042103.
40. Yu, B. Tutorial: Information theory and statistics. In Proceedings of the 7th International Conference on Machine Learning and Applications, San Diego, CA, USA, 11–13 December 2008.
41. Hobson, A. A new theorem of information theory. *J. Stat. Phys.* **1969**, *1*, 383–391.
42. Shore, J.E.; Johnson, R.W. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37.
43. Haken, H. *Synergetics—An Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*; Springer: Berlin/Heidelberg, Germany, 1983.
44. Haken, H. *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*; Springer: Berlin/Heidelberg, Germany, 1993.
45. Haken, H. *Information and Self-Organization: A Macroscopic Approach to Complex Systems*; Springer: Berlin/Heidelberg, Germany, 2010.
46. Ye, F.X.-F.; Wang, Y.; Qian, H. Stochastic dynamics: Markov chains and random transformations. *Disc. Contin. Dyn. Syst. B* **2016**, in press.
47. Sack, R.A. Pressure-dependent partition functions. *Mol. Phys.* **1958**, *2*, 8–22.
48. Münster, A. Zur Theorie der generalisierten Gesamtheiten. *Mol. Phys.* **1958**, *2*, doi:10.1080/00268975900100011. (In German)
49. Brown, W. Constant pressure ensembles in statistical mechanics. *Mol. Phys.* **1957**, *1*, 68–82.
50. Ao, P.; Qian, H.; Tu, Y.; Wang, J. A Theory of Mesoscopic Phenomena: Time Scales, Emergent Unpredictability, Symmetry Breaking and Dynamics across Different Levels. **2013**, arXiv:1310.5585.
51. Campisi, M. Construction of microcanonical entropy on thermodynamic pillars. *Phys. Rev. E* **2015**, *91*, 052147.
52. Jaynes, E.T. Gibbs vs. Boltzmann entropies. *Am. J. Phys.* **1965**, *33*, 391–398.
53. Frenkel, D.; Warren, P.B. Gibbs, Boltzmann, and Negative Temperatures. **2014**, arXiv:1403.4299v3.
54. Dunkel, J.; Hilbert, S. Consistent thermostatistics forbids negative absolute temperatures. *Nat. Phys.* **2013**, *10*, 67–72.
55. Goldstein, S.; Lebowitz, J.L. On the (Boltzmann) entropy of non-equilibrium systems. *Phys. D Nonlinear Phenom.* **2004**, *193*, 53–66.
56. Lebowitz, J.L. Boltzmann's entropy and time's arrow. *Phys. Today* **1993**, *46*, 32–38.
57. Zimm, B.H. Contribution to the theory of critical phenomena. *J. Chem. Phys.* **1951**, *19*, 1019–1023.
58. Yang, C.N.; Lee, T.D. Statistical theory of equations of state and phase transitions. I. Theory of condensation. *Phys. Rev.* **1952**, *87*, 404–409.
59. Lee, T.D.; Yang, C.N. Statistical Theory of Equations of State and Phase Transitions. II. Lattice Gas and Ising Model. *Phys. Rev.* **1952**, *87*, 410–419.
60. Blythe, R.A.; Evans, M.R. Lee–Yang zeros and phase transitions in nonequilibrium steady states. *Phys. Rev. Lett.* **2002**, *89*, 080601.
61. Widder, D.V. *The Laplace Transform*; Princeton University Press: Princeton, NJ, USA, 1946.
62. Gibbs, J.H.; DiMarzio, E.A. Statistical mechanics of helix-coil transitions in biological macromolecules. *J. Chem. Phys.* **1959**, *30*, 271–282.
63. Kittel, C. Phase transition of a molecular zipper. *Am. J. Phys.* **1969**, *37*, 917–920.

64. Nagle, J.F. The one-dimensional KDP model in statistical mechanics. *Am. J. Phys.* **1968**, *36*, 1114–1117.
65. Schwab, D.J.; Nemenman, I.; Mehta, P. Zipf's law and criticality in multivariate data without fine-tuning. *Phys. Rev. Lett.* **2014**, *113*, 068102.

Appendix C

REPRESENTATIONS AND DIVERGENCES IN THE SPACE OF PROBABILITY MEASURES AND STOCHASTIC THERMODYNAMICS (WITH HONG, L. & QIAN, H.) *J. COMPUT. APPL. MATH.* 376, 112842 (2020)

Journal of Computational and Applied Mathematics 376 (2020) 112842

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Representations and divergences in the space of probability measures and stochastic thermodynamics

Liu Hong [a,b], Hong Qian [a,*], Lowell F. Thompson [a]

[a] Department of Applied Mathematics, University of Washington, Seattle, WA 98195-3925, USA
[b] Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, 100084, PR China

ARTICLE INFO

Article history:
Received 5 February 2019
Received in revised form 2 March 2020

MSC:
60-xx
80-xx
82-xx

Keywords:
Radon–Nikodym derivative
Affine structure
Space of probability measures
Heat divergence

ABSTRACT

Radon–Nikodym (RN) derivative between two measures arises naturally in the affine structure of the space of probability measures with densities. Entropy, free energy, relative entropy, and entropy production as mathematical concepts associated with RN derivatives are introduced. We identify a simple equation that connects two measures with densities as a possible mathematical basis of the entropy balance equation that is central in nonequilibrium thermodynamics. Application of this formalism to Gibbsian canonical distribution yields many results in classical thermomechanics. An affine structure based on the canonical representation and two divergences are introduced in the space of probability measures. It is shown that thermodynamic work, as a conditional expectation, is indicative of the RN derivative between two energy representations being singular. The entropy divergence and the heat divergence yield respectively a Massieu–Planck potential based and a generalized Carnot inequalities.

1. Introduction

A subtle distinction exists between the prevalent approach to stochastic processes in traditional applied mathematics and the physicist's perspective on stochastic dynamics: In Kolmogorov's theory of stochastic processes, the dynamics are described in terms of a trajectory $\{\mathbf{x}(t) : t \in [0, \infty)\}$. Applied mathematicians treat each of these trajectories as a random event in a large probability space and then study probability distributions over the space of all possible trajectories $\mathbf{x}(t)$. Physists, however, are more accustomed to thinking of a "probability distribution changing with time", $\rho(\mathbf{x}, t)$. In the case of continuous-time Markov processes, $\rho(\mathbf{x}, t)$ is described by the solution to a Fokker–Planck equation or a master equation, while for a Markov chain simply by a stochastic matrix. This latter perspective can perhaps be more rigorously formulated in the *space of probability measures*. The dynamics are then represented as a *change of measure*. The Radon–Nikodym (RN) derivative is a key mathematical concept associated with changes in measures [1]. Interestingly, RN derivative between two measures is also at the heart of the concept of *fluctuating entropy* [2,3].

This "probability distribution changing with time" view is, of course, not foreign to mathematics. Actually in the 1950s, the stochastic diffusion process developed by Feller, Nelson, and others was precisely a such theory [4–7]. That approach, based on solutions to linear parabolic partial differential equations, was formulated in a linear function space. We now know that a more geometrically intrinsic representation for the space of probability measures cannot be linear: There is simply no natural choice of origin. Rather, an *affine space* is more appropriate [8,9].

* Corresponding author.
E-mail addresses: zcamhl@tsinghua.edu.cn (L. Hong), hqian@uw.edu (H. Qian), lfthomps@uw.edu (L.F. Thompson).

https://doi.org/10.1016/j.cam.2020.112842
0377-0427/

Entropy and energy are key concepts in the classical theory of thermodynamics, which is now well understood to have a probabilistic basis. In fact, one could argue that the very notion of "heat" arises only when one treats the motions of deterministic Newtonian point masses as stochastic. In the statistical treatment of thermodynamics, Gibbs' canonical energy distribution is one of the key results that characterize a thermodynamic equilibrium [10]. As we shall see, it figures prominently in the affine space.

The foregoing discussion suggests the possibility of re-thinking thermodynamics and information theory in a novel mathematical framework [11]. Both information theory and thermodynamics are concerned with notions such as entropy, free energy and relative entropy. These concepts are introduced in Section 2 under a single framework based on the Radon–Nikodym derivative, as a random variable relating two different measures. In its broadest context, we are able to capture the essential mathematics used in the theory of equilibrium and nonequilibrium thermodynamics. This approach significantly enriches the scope of "information theory" [12]. The RN derivative should not be treated as an esoteric mathematical concept: It is simply a powerful way to quantify even infinitesimal changes in the probability distributions; it is the calculus for thinking of change in terms of chance [13].

In Section 3, the notion of a temperature, $T = \beta^{-1}$ is introduced through the canonical probability distribution $Z^{-1}(\beta)e^{-\beta U(\omega)}$. It has been shown recently that this Gibbsian distribution has a much broader applications than just thermal physics: It is in fact a limit theorem of a sequence of conditional probability densities under an additive quasi-conservative observable [14]. The focus of this section is to show the centrality of RN derivative in the theory of thermodynamics. The RN derivative is used to describe several results in physics that includes the thermodynamic cycle, equation of states, and the Jarzynski–Crooks equalities.

Next, in Section 4 we equip the space of probability measures with an affine structure and show that the canonical distribution with a random variable $U(\omega)$ and a parameter β becomes precisely an affine line in the space of probability measures when one particular measure $\mathbb{P}$ is chosen as a reference point. With this, the tangent space becomes a linear vector space of random variables and it provides a representation for the space of probability measures. A series of results are obtained. Readers who are more mathematically inclined can skip Section 3, come directly to Section 4, and then go back to Section 3 afterward.

Section 5 contains some discussions.

The presentation of the paper is not mathematically rigorous. The emphasis is on illustrating how the pure mathematical concepts can be fittingly applied in narrating this branch of physics. More thorough treatments of the subject are forthcoming [9].

2. Entropy, relative entropy, and a fundamental equation of information

2.1. Information and entropy

Information theory owes (to a large extent) its existence as a separate subject from the theories of probability and statistics to a singular emphasis on the notion of *entropy* as a quantitative measure of information. It is important to point out at the outset that *information* is a random variable, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, through a Radon–Nikodym derivative $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(\omega)$, $\omega \in \Omega$, between two measures $\mathbb{P}$ and μ that are absolutely continuous w.r.t. each other [2,3,15]. If the $\Omega \subseteq \mathbb{R}^n$ and μ is the Lebesgue measure, then

$$-\ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(\omega)\right) \tag{1}$$

is the *self-information* [16,17], which is a random variable and its expected value is the standard form of Shannon entropy:

$$S[\mathbb{P}] \triangleq -\int_{\Omega} f(x)\ln f(x)\,\mathrm{d}x, \tag{2}$$

in which the Radon–Nikodym derivative is the probability density function, $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu} \equiv f(x)$.

In general, if μ is normalizable, then one has a maximum entropy inequality $S[\mathbb{P}] \leq \ln \mu(\Omega) < +\infty$. Similarly, one has the free energy

$$H[\mathbb{P} \parallel \mu] \triangleq \int_{\Omega} \ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(\omega)\right)\mathrm{d}\mathbb{P}(\omega) \geq -\ln \mu(\Omega). \tag{3}$$

When μ is also a normalized probability measure $\mathbb{P}'$, the $H[\mathbb{P} \parallel \mathbb{P}']$ is called the *relative entropy* or Kullback–Leibler (KL) divergence. The minimum free energy inequality in (3) becomes the better known, but less interesting, $H[\mathbb{P} \parallel \mathbb{P}'] \geq 0$.

From now on, we will drop most references to the underlying space $(\Omega, \mathcal{F})$. Moreover, we will assume that $\Omega \subseteq \mathbb{R}^n$ with the usual σ-algebra and that $\mathbb{P}$ is absolutely continuous w.r.t. the Lebesgue measure. These conditions are not strictly necessary, but they simplify the notation considerably in illustrating our key ideas.

2.2. Fundamental equation of information

With the various forms of entropy introduced above and some straightforward statistical logic, one naturally has the following equation that involves three measures: two probabilistic and the Lebesgue. In particular, let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability measures with density functions $f_1(x)$ and $f_2(x)$ with respect to the Lebesgue measure:

$$\begin{aligned}\Delta S &= S[\mathbb{P}_2] - S[\mathbb{P}_1] = \int_{\mathbb{R}} f_1(x)\ln f_1(x)\mathrm{d}x - \int_{\mathbb{R}} f_2(x)\ln f_2(x)\mathrm{d}x \\ &= \underbrace{\int_{\mathbb{R}} f_1(x)\ln\left(\frac{f_1(x)}{f_2(x)}\right)\mathrm{d}x}_{\Delta S^{(\mathrm{i})}\text{: entropy production}} + \underbrace{\int_{\mathbb{R}}\Big(f_2(x)-f_1(x)\Big)\Big(-\ln f_2(x)\Big)\mathrm{d}x}_{\Delta S^{(\mathrm{e})}\text{: entropy exchange}}. \end{aligned} \tag{4}$$

The entropy production $\Delta S^{(\mathrm{i})}$ is never negative, while the entropy exchange $\Delta S^{(\mathrm{e})}$ has no definitive sign. If $f_2(x)$ is the unique invariant density of some measure-preserving dynamics [18], then $-\ln f_2(x)$ is customarily referred to as the "equilibrium energy function", then $\Delta S^{(\mathrm{e})}$ is the change in the "mean energy", which is related to "heat".

Entropy and free energy in (2) and (3) have their namesakes in the theory of statistical equilibrium thermodynamics [10]. The Second Law, in terms of entropy maximization or free energy minimization, has its statistical basis precisely in the two inequalities associated with S and H. The $\Delta S^{(\mathrm{i})}$ term on the rhs of (4), however, is a nonequilibrium free energy associated with a *nonequilibrium distribution*, either due to a spontaneous fluctuation or a man-made perturbation [19]. In the theory of stochastic dynamics, one uses a probability distribution $\rho(x,t)$ to represent the state of a system; thus any ρ that differs from the equilibrium distribution is a nonequilibrium distribution. In applications to laboratory systems, the ρ can only be obtained from a data-based statistical approach. This approach can rely on either a time scale separation, or a system of many independent and identically distributed subsystems, or a fictitious ensemble. Ideal gas theory and the Rouse model of polymers are two successful examples of the second type [19].

Eq. (4) in fact has the form of the *fundamental equation of nonequilibrium thermodynamics*. It states that if $f_2(x)$ is uniform, then $\Delta S = \Delta S^{(\mathrm{i})} \geq 0$; and if one identifies $U(x) \triangleq -T\ln f_2(x)$, where T is a positive constant, then one can introduce $F[\mathbb{P}] \triangleq \mathbb{E}^{\mathbb{P}}[U] - TS[\mathbb{P}]$, and $\Delta F = T\Delta S^{(\mathrm{i})} \geq 0$. Unifying the various forms of the Second Law to a single concept of entropy production was a key idea of the Brussel school of thermodynamics [20].[1] See [3,11,21,22], and the references cited within, for the theory of entropy production of Markov processes.

2.3. Two results on relative entropy

With regards to relative entropy, there are two results worth discussing.

First, as the expected value of the logarithm of the Radon–Nikodym derivative $\xi \equiv \ln\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}(\omega)\right)$, the relative entropy between two probability measures can be written as

$$H\big[\mathbb{P}_1 \parallel \mathbb{P}_2\big] = \int_{\mathbb{R}} f_1(x)\ln\left(\frac{f_1(x)}{f_2(x)}\right)\mathrm{d}x = \mathbb{E}^{\mathbb{P}_1}\big[\xi(\omega)\big], \tag{5}$$

with respective probability density functions $f_1(x) = \frac{\mathrm{d}\mathbb{P}_1(x)}{\mathrm{d}x}$ and $f_2(x) = \frac{\mathrm{d}\mathbb{P}_2(x)}{\mathrm{d}x}$. The non-negativity of the $H[\mathbb{P}_1 \parallel \mathbb{P}_2]$ can actually be framed as a consequence of a stronger result, an equality

$$\mathbb{E}^{\mathbb{P}_1}\left[e^{-\xi(\omega)}\right] = 1, \tag{6}$$

and an inequality for convex exponential function:

$$\mathbb{E}^{\mathbb{P}_1}\big[\xi(\omega)\big] \geq -\ln\mathbb{E}^{\mathbb{P}_1}\left[e^{-\xi(\omega)}\right] = 0. \tag{7}$$

Eq. (6) implies that the Second Law and entropy production could even be formulated through equalities rather than inequalities. Indeed, variations of (6) have found numerous applications in thermodynamics, such as Zwanzig's free energy perturbation method [23], the Jarzynski–Crooks relation [24,25], and the Hatano–Sasa equality [26].

Second, if the density f_2 contains an unknown parameter θ, then $f_2(x;\theta)$ is the likelihood function for θ. In this case, with respect to the change of measure,

$$\begin{aligned}\mathcal{I}_\ell(\theta) &\triangleq -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^\ell}{\partial\theta^\ell}\ln f_2(\omega;\theta)\,\middle|\,\theta\right] \\ &= -\int_{\mathbb{R}} f_2(x;\theta)\frac{\partial^\ell}{\partial\theta^\ell}\ln f_2(x;\theta)\mathrm{d}x\end{aligned}$$

[1] The second author would like to acknowledge an enlightening discussion with M. Esposito in the spring of 2011 at the Snogeholm Workshop on Thermodynamics, Sweden.

$$= -\int_{\mathbb{R}} \left\{ \left(\frac{f_2(x;\theta)}{f_1(x)} \right) \frac{\partial^\ell}{\partial \theta^\ell} \ln \left(\frac{f_2(x;\theta)}{f_1(x)} \right) \right\} f_1(x) \mathrm{d}x$$
$$= \mathbb{E}^{\mathbb{P}_1} \left[e^{-\xi(\omega)} \frac{\partial^\ell}{\partial \theta^\ell} \xi(\omega;\theta) \right]. \tag{8}$$

$\mathcal{I}_0(\theta)$ is the Shannon entropy of $X_2(\theta)$, $\mathcal{I}_1(\theta) \equiv 0$, and $\mathcal{I}_2(\theta)$ is the *Fisher Information* for $X_2(\theta)$:

$$\mathcal{I}_2(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \ln f_2(X_2;\theta) \right)^2 \middle| \theta \right]. \tag{9}$$

3. Canonical distribution and thermodynamics

In many applications, stochastic dynamics exhibit a separation of slow and fast time scales [27,28]. In mechanical systems with sufficiently small friction, the dynamics are organized as fast Hamiltonian dynamics with slow energy dissipation through heat. The theory of thermodynamics arises in this context when the mechanical motions of point masses are described stochastically. It can be shown then that the probability distribution for the energy E of a small mechanical system in equilibrium with a large heat bath takes a particularly canonical form

$$p_E(y) = \frac{\Omega^{(B)}(y) e^{-\beta y}}{Z(\beta)}, \tag{10}$$

in which $\beta^{-1} = T$ is the temperature of the heat bath [10,14]. In fact, if $\mathbf{x}$ denotes random variable in an appropriate state space and $U(x)$ is the mechanical energy function, then one has distribution $f_{\mathbf{x}}(x) \propto e^{-\beta U(x)}$, and

$$p_E(y)\mathrm{d}y = \int_{y<U(x)\le y+\mathrm{d}y} f_{\mathbf{x}}(x)\mathrm{d}x = \left(\frac{\Omega^{(B)}(y) e^{-\beta y}}{Z(\beta)} \right) \mathrm{d}y, \tag{11}$$

in which

$$\Omega^{(B)}(y) = \frac{1}{\mathrm{d}y} \int_{y<U(x)\le y+\mathrm{d}y} \mathrm{d}x = \frac{\mathrm{d}\Omega^{(G)}(y)}{\mathrm{d}y}, \quad \Omega^{(G)}(y) = \int_{U(x)\le y} \mathrm{d}x. \tag{12}$$

$\ln \Omega^{(B)}$ and $\ln \Omega^{(G)}$ are called Boltzmann's entropy and Gibbs' entropy in statistical mechanics [29]. They are related via $\mathrm{d}\Omega^{(G)}(y) = \Omega^{(B)}(y)\mathrm{d}y$. That is, $\Omega^{(G)}$ is a cumulative distribution function and $\Omega^{(B)}$ is its density function.

Note that the expected value of any function of the energy $U(x)$ (e.g., $g(U)$) is invariant under different representations as a result of the rules of changes of variable for integration. For example, if $\mathbf{x}$ is a state space representation and E is the energy representation, then

$$\int g\big(U(x)\big) f_{\mathbf{x}}(x)\mathrm{d}x = \int_{\mathbb{R}} g(y) \left(\frac{e^{-\beta y}}{Z(\beta)} \right) \Omega^{(B)}(y)\mathrm{d}y$$
$$= \int_{\mathbb{R}} g(y) p_E(y)\mathrm{d}y.$$

In contrast, the thermodynamic entropy in statistical mechanics is not invariant under different representations [30]:

$$-\int p_{\mathbf{x}}(x) \ln p_{\mathbf{x}}(x)\mathrm{d}x = -\int_{\mathbb{R}} p_E(y) \ln \left(\frac{p_E(y)}{\Omega^{(B)}(y)} \right) \mathrm{d}y \tag{13a}$$
$$\neq -\int_{\mathbb{R}} p_E(y) \ln p_E(y)\mathrm{d}y. \tag{13b}$$

The rhs of (13a) is precisely the negative free energy with non-normalized $\Omega^{(G)}(y)$ as the reference measure (which has density $\Omega^{(B)}(y)$). The missing term from (13a) to (13b)

$$\int_{\mathbb{R}} p_E(y)\Big(-\ln \Omega^{(B)}(y) \Big)\mathrm{d}y. \tag{14}$$

is contributed by the reference measure. It is mean-internal-energy like. We see that while $\ln \Omega^{(B)}(y)$ is widely considered as an "entropic" term, it actually plays the role of an energetic term in the energy representation in (13a). In terms of this measure-theoretic framework, the distinction between entropy and energy is always relative. This has long been understood in the work of J. G. Kirkwood on the *potential of mean force*, which is itself temperature dependent [31].

3.1. Thermodynamics under a single temperature

Equilibrium statistical thermodynamics. In terms of the canonical distribution in (10), an equilibrium system under a constant temperature $T = \beta^{-1}$ has its mechanical energy distributed according to the canonical distribution $p^{\mathrm{eq}}(y) =$

$Z^{-1}(\beta)\Omega(y)e^{-\beta y}$. (We have dropped the superscript in $\Omega^{(B)}(y)$ to avoid cluttering.) The *mean internal energy* associated with the $p^{\rm eq}(y)$ is then the expected value

$$\overline{U}(\beta) = \int_{\mathbb{R}} y \left(\frac{\Omega(y)e^{-\beta y}}{Z(\beta)} \right) \mathrm{d}y = -\frac{\mathrm{d}\ln Z(\beta)}{\mathrm{d}\beta}, \tag{15a}$$

which can be decomposed into an equilibrium *free energy* and an *entropy*, $\overline{U}(\beta) = F^{\rm eq}(\beta) + \beta^{-1}S(\beta)$, where:

$$\underbrace{F^{\rm eq}(\beta) = -\beta^{-1}\ln Z(\beta)}_{\text{free energy}} \text{ and } \underbrace{S(\beta) = -\left(\frac{\mathrm{d}F^{\rm eq}(\beta)}{\mathrm{d}\beta^{-1}}\right)}_{\text{entropy}}. \tag{15b}$$

One can verify that the $S(\beta)$ is the same as (13a), but not (13b).

Nonequilibrium statistical thermodynamics. For deep mathematical reasons that will become clear in Section 4, discussions of nonequilibrium systems should begin in the full state space. Intuitively, the canonical energy representation $p^{\rm eq}(E)$ based on a given energy function $U(x)$ is a "projection" in the space of probability measures that is nonholographic.

Consider a system outside statistical equilibrium with a nonequilibrium probability measure $\mu^{\rm neq}$. Suppose that this measure is absolutely continuous w.r.t. some other probability measure $\mathbb{P}$, with density $\rho(x) = \frac{\mathrm{d}\mu^{\rm neq}}{\mathrm{d}\mathbb{P}}(x)$. The measure $\mu^{\rm neq}$ possesses a nonequilibrium free energy functional (a potential that can cause change) given by

$$F^{\rm neq}\big[\rho;\beta\big] \triangleq F^{\rm eq}(\beta) + \beta^{-1}\int_\Omega \rho(x)\ln\left(\frac{\rho(x)}{p^{\rm eq}(x)}\right)\mathrm{d}x \tag{16a}$$

$$= \beta^{-1}\int_\Omega \rho(x)\ln\left(\frac{\rho(x)}{e^{-\beta U(x)}}\right)\mathrm{d}x. \tag{16b}$$

One should recognize the fraction in (16b) as a Radon–Nikodym derivative of ρ w.r.t. the non-normalized *canonical equilibrium measure* $e^{-\beta U(x)}$. The minimum free energy inequality in (3) takes the form $F^{\rm neq}\big[\rho;\beta\big] \geq F^{\rm eq}(\beta)$ for any distribution ρ. In fact, $\beta\{F^{\rm neq}\big[\rho;\beta\big] - F^{\rm eq}(\beta)\}$ is the entropy production associated with the spontaneous *relaxation process* of the distribution ρ tending to $p^{\rm eq}$.

The $F^{\rm neq}\big[\rho;\beta\big]$ also has another expression:

$$F^{\rm neq}\big[\rho;\beta\big] = \underbrace{\beta^{-1}\int_\Omega \rho(x)\ln\rho(x)\mathrm{d}x}_{\text{neg-entropy}} + \int_\Omega \rho(x) \underbrace{\Big(-\beta^{-1}\ln p^{\rm eq}(x) + F^{\rm eq}(\beta)\Big)}_{\text{internal energy of state } x,\ F^{\rm eq} \text{ as reference}} \mathrm{d}x. \tag{17}$$

Eq. (17) is very telling: The internal energy of a system in state x is given in the second term with a fixed energy gauge (*i.e.*, the arbitrary constant in the $U(x)$) according to the equilibrium $F^{\rm eq}$, where $U(x) = F^{\rm eq}(\beta) - \beta^{-1}\ln p^{\rm eq}(x)$. This fact implies that a change in the energy function from $U_1(x)$ to $U_2(x)$ necessarily involves a change of gauge. Mechanical work in classical thermodynamics can be understood as a consequence of *gauge invariance*. One particular β defines an autonomous, time-homogeneous stochastic dynamical system with a unique $p^{\rm eq}$. All the energetic discussions in such a system are with respect to the equilibrium free energy $F^{\rm eq}(\beta)$, which fixes a choice for the energy gauge. In the theory of probability, the gauge invariance is achieved through the notion of conditional probability and the law of total probability.

3.2. A clarification of Eq. (16)

A discussion on the meaning of the expression in (16) is in order. To do that, let us only consider discrete x_k, and the corresponding

$$F^{\rm neq}[\rho;\beta] = \sum_k \rho(x_k)\left[\beta^{-1}\ln\left(\frac{\rho(x_k)}{p^{\rm eq}(x_k)e^{-\beta F^{\rm eq}(\beta)}}\right)\right]. \tag{18}$$

For a particular state z, if $\rho(x) = \delta_{x,z}$, then $F^{\rm neq}[\rho;\beta] = F^{\rm eq}(\beta) - \beta^{-1}\ln p^{\rm eq}(z)$, which represents the traditional potential energy of the system in the state z. A question then naturally arises: Why is $F^{\rm neq}[\rho;\beta]$ the average of

$$\beta^{-1}\ln\left(\frac{\rho(x_k)}{p^{\rm eq}(x_k)e^{-\beta F^{\rm eq}}}\right), \tag{19a}$$

but not

$$\beta^{-1}\ln\left(\frac{1}{p^{\rm eq}(x_k)e^{-\beta F^{\rm eq}}}\right)? \tag{19b}$$

Actually, (19b) is the potential energy for a deterministic initial state x_k. It is natural, therefore, the average would be carried out over (19b) if the initial state of the system were a *mixture of heterogeneous states* (mhs). However, if the initial state is a *stochastic fluctuating state* (sfs), then the entropy of assimilation applies [32] and the $F^{\rm neq}[\rho;\beta]$ in (16a) is the

average carried out over (19a). The change from mhs to sfs is analogous to a change from the Lagrangian to the Eulerian representation in fluid mechanics; in stochastic terms, the potential for an sfs to do work is lower than an mhs [33].

3.3. Work, heat, and Jarzynski–Crooks' relation

We now consider the case where the distribution $\rho(x)$ in (16) arises from the equilibrium distribution $p^{\rm eq}(x)$ as the consequence of a temperature change from T_a to T_b: $\rho(x) = Z^{-1}(\beta_a)e^{-\beta_a U(x)}$, and the $p^{\rm eq}(x) = Z^{-1}(\beta_b)e^{-\beta_b U(x)}$. Note that in the energy representation they can be written as $\rho_E(y) = Z^{-1}(\beta_a)\Omega(y)e^{-\beta_a y}$ and $p_E^{\rm eq}(y) = Z^{-1}(\beta_b)\Omega(y)e^{-\beta_b y}$; they share the same Gibbs entropy $\ln \Omega(y)$ determined by $U(x)$ as in (12). Then

$$F^{\rm neq}\big[\rho;\,\beta_b\big] - F^{\rm eq}(\beta_b) = \beta_b^{-1}\int_\Omega \rho(x)\ln\left(\frac{\rho(x)}{p^{\rm eq}(x)}\right)\mathrm{d}x \tag{20a}$$

$$= \beta_b^{-1}\int_{\mathbb{R}} \rho_E(y)\ln\left(\frac{\rho_E(y)}{p_E^{\rm eq}(y)}\right)\mathrm{d}y \tag{20b}$$

$$= \Big[\overline{U}(\beta_a) - \beta_b^{-1}S(\beta_a)\Big] - \Big[\overline{U}(\beta_b) - \beta_b^{-1}S(\beta_b)\Big]. \tag{20c}$$

The equation from (20a) to (20b) utilizes a key property of a Radon–Nikodym derivative: *When it exists*, it is invariant under a change of measure.

Eq. (20c) is not widely discussed, but it is a highly meaningful result. It contains the essence of Crooks' equality in time-inhomogeneous Markov processes [25]. It implies that at the instant of switching from T_a to T_b, the system has internal energy $\overline{U}(\beta_a)$, entropy $S(\beta_a)$, and nonequilibrium free energy

$$F^{\rm neq}[\rho;\,\beta] = \overline{U}(\beta_a) - T_b S(\beta_a). \tag{21}$$

Assuming that both $\rho(x)$ and $p^{\rm eq}(x)$ have the same $\Omega(y)$, Eq. (20) gives the free energy change that is expected to be the maximum reversible work that can be extracted. We now explicitly consider a change from $\rho(x)$ to $p^{\rm eq}(x)$ that involves changing the mechanical energy function from $U_1(x)$ to $U_2(x)$. Even though the corresponding canonical energy distributions are $\rho_E(y) = Z_1^{-1}(\beta_a)\Omega_1(y)e^{-\beta_a y}$ and $p_E^{\rm eq}(y) = Z_2^{-1}(\beta_b)\Omega_2(y)e^{-\beta_b y}$, these RN derivative $\frac{\mathrm{d}\rho_E}{\mathrm{d}p^{\rm eq}}(\omega)$ can be infinity! Thus in this case one has to start with the full distributions on the state space:

$$\beta_b^{-1}\int_\Omega \rho(x)\ln\left(\frac{\rho(x)}{p^{\rm eq}(x)}\right)\mathrm{d}x = \Big[\overline{U}_1(\beta_a) - \overline{U}_2(\beta_b)\Big] - \beta_b^{-1}\Big[S_1(\beta_a) - S_2(\beta_b)\Big] + \int_\Omega \rho(x)\big[U_2(x) - U_1(x)\big]\mathrm{d}x. \tag{22}$$

The last term in (22) is identified as the irreversible work associated with the isothermal relaxation process with mechanical change from $U_1(x)$ to $U_2(x)$,

$$\overline{\mathcal{W}}_{12}(\beta_a) = \int_\Omega \rho(x)\mathcal{W}_{12}(x)\mathrm{d}x, \tag{23}$$

in which $\mathcal{W}_{12}(x)$ should be considered as the logarithm of the Radon–Nikodym derivative between two non-normalized measures

$$\mathcal{W}_{12}(x) = \beta_a^{-1}\ln\left(\frac{e^{-\beta_a U_1(x)}}{e^{-\beta_a U_2(x)}}\right) = \beta_b^{-1}\ln\left(\frac{e^{-\beta_b U_1(x)}}{e^{-\beta_b U_2(x)}}\right). \tag{24}$$

$\mathcal{W}_{12}(x)$ is actually not a function of β; work done in an isothermal process is independent of the temperature. In the canonical energy representation of $U_1(x)$, then,

$$\overline{\mathcal{W}}_{12}(\beta_a) = \int_\Omega \rho(\mathrm{x})\big[U_2(x) - U_1(x)\big]\mathrm{d}x = \int_{\mathbb{R}}\left(\frac{\Omega_1(y)e^{-\beta_a y}}{Z_1(\beta_a)}\right)\left\{\frac{\displaystyle\int_{y<U_1(x)\le y+\mathrm{d}h} U_2(x)\mathrm{d}x}{\displaystyle\int_{y<U_1(x)\le y+\mathrm{d}h}\mathrm{d}x} - y\right\}\mathrm{d}y. \tag{25}$$

The first term inside $\{\cdots\}$ is a conditional expectation: $\mathbb{E}^{\rm eq}\big[U_2(\mathrm{x})\big|U_1(\mathrm{x}) = y\big]$, where $\mathbb{E}^{\rm eq}$ is the expectation in terms of the equilibrium measure $p^{\rm eq}(x)$.

The transferred irreversible heat is

$$\mathcal{Q}(\beta_b) \triangleq \beta_b^{-1}\left\{S_1(\beta_a) - S_2(\beta_b) + \int_\Omega \rho(x)\ln\left(\frac{\rho(x)}{p^{\rm eq}(x)}\right)\mathrm{d}x\right\}. \tag{26}$$

Then the relation

$$S_2(\beta_b) - S_1(\beta_a) + \frac{\mathcal{Q}(\beta_b)}{T_b} = \Delta S^{(\mathrm{i})} = \int_\Omega \rho(x) \ln\left(\frac{\rho(x)}{p^{\mathrm{eq}}(x)}\right) \mathrm{d}x \geq 0 \tag{27}$$

is known as the Clausius inequality in thermodynamics. The equality is a special case of the fundamental equation of nonequilibrium thermodynamics.

Concerning the work $\mathcal{W}_{12}(x)$ in (24), we have Jarzynski–Crooks' relation [24,25]:

$$\int_\Omega \left(\frac{e^{-\beta_a U_1(x)}}{Z_1(\beta_a)}\right) e^{-\beta_a \mathcal{W}_{12}(x)} \mathrm{d}x = \int_\Omega \frac{e^{-\beta_a U_2(x)}}{Z_1(\beta_a)} \mathrm{d}x = \frac{Z_2(\beta_a)}{Z_1(\beta_a)}. \tag{28}$$

Note that the work is performed under β_b, but the rhs of (28) is evaluated at β_a. The original Jarzynski–Crooks' equality emphasized path-wise average over a stochastic trajectory, but Eq. (28) is an ensemble average over a single step, which can be generalized to many different other forms [34].

The concept of exergy. In Eq. (21), equilibrium internal energy and entropy under temperature T_a, $\overline{U}(T_a)$ and $S(T_a)$ are assembled with temperature $T_b \neq T_a$ to form a nonequilibrium free energy $F^{\mathrm{neq}} = \overline{U}(T_a) - T_b S(T_a)$, which plays a central role in our analysis of canonical systems. This quantity has been extensively discussed in the literature on thermodynamics: *Exergy* of a system is "the maximum fraction of an energy form which can be transformed into work". The remaining part is the waste heat [35]. After a system reaches equilibrium with its surrounding, its exergy is zero. Therefore, the concept of exergy epitomizes a nonequilibrium quantity [36]. Its identification to the entropy production in Eq. (20) implies its importance in information energetics. Even though the term "exergy" was coined as late as in 1956, the idea had been already in the work of Gibbs.

Mechanical work of an ideal gas. For an ideal gas with total mechanical energy $U(x) = U_p(x_1) + U_k(x_2)$, where U_p and U_k are potential and kinetic energy functions, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are position and momentum state variables,

$$U(x) = \sum_{i=1}^{N} \left\{ \frac{x_{2,i}^2}{2m_i} + H_V\left(x_{1,i}\right) \right\}, \tag{29}$$

in which $H_V(z) = 0$ when $0 < z < V$ and $H_V(z) = +\infty$ when $z \leq 0$ or $z \geq V$. The V represents the "volume" of a box containing the ideal gas. Then

$$\Omega(E, V) = \frac{V^N}{\mathrm{d}E} \int_{E < U_k(x_2) \leq E + \mathrm{d}E} \mathrm{d}x_2 = V^N \tilde{\Omega}(E, N), \tag{30}$$

in which the $\tilde{\Omega}$ is independent of V. Therefore, the mechanical work associated with a change in $V_1 = V \to V_2 = V + \Delta V$ is given by

$$\beta^{-1} \ln\left(\frac{\Omega(E, V_2)}{\Omega(E, V_1)}\right) = NT \ln\left(\frac{V + \Delta V}{V}\right) = \frac{NT\Delta V}{V} = \hat{p}\Delta V, \tag{31}$$

where $\hat{p} = Nk_BT/V$ is the pressure of an ideal gas. (We have set Boltzmann's constant $k_B \equiv 1$ throughout the present paper.)

3.4. Application to heat engines and thermodynamic cycles

Carnot cycle. Applying Eqs. (24) and (26) twice for *thermomechanical* (*i.e.*, temperature and mechanical) changes from $\{T_a, U_1\}$ to $\{T_b, U_2\}$ and from $\{T_b, U_2\}$ back to $\{T_a, U_1\}$, we derive the celebrated Carnot efficiency for a heat engine. For each of the processes described in the left column below, the energetic status of the system is shown in the right column:

$$\text{adiabatic switching } \{T_a, U_1\} \to \{T_b, U_1\}: \quad F_1^{\mathrm{neq}}(T_b) = \overline{U}_1(T_a) - T_b S_1(T_a), \tag{32a}$$

$$\text{isothermal relaxation } \{T_b, U_1\} \to \{T_b, U_2\}: \quad \overline{U}_1(T_a) - \overline{U}_2(T_b) = \mathcal{Q}_{12}(T_b) - \overline{\mathcal{W}}_{12}, \tag{32b}$$

$$\text{equilibrium under } T_b: \quad F_2^{\mathrm{eq}}(T_b) = \overline{U}_2(T_b) - T_b S_2(T_b), \tag{32c}$$

$$\text{adiabatic switching } \{T_b, U_2\} \to \{T_a, U_2\}: \quad F_2^{\mathrm{neq}}(T_a) = \overline{U}_2(T_b) - T_a S_2(T_b), \tag{32d}$$

$$\text{isothermal relaxation } \{T_a, U_2\} \to \{T_a, U_1\}: \quad \overline{U}_2(T_b) - \overline{U}_1(T_a) = \mathcal{Q}_{21}(T_a) - \overline{\mathcal{W}}_{21}, \tag{32e}$$

$$\text{equilibrium under } T_a: \quad F_1^{\mathrm{eq}}(T_a) = \overline{U}_1(T_a) - T_a S_1(T_a). \tag{32f}$$

In (32f), the system is returned to the equilibrium state under T_a. Without loss of generality, let $T_a > T_b$. In the ideal Carnot cycle, one assumes that the processes of switching the temperatures are adiabatic without free energy dissipation. That is, the $F_1^{\mathrm{neq}}(T_b)$ in (32a) is strictly equal to $F_1^{\mathrm{eq}}(T_a)$ in (32f), with a reversible change of gauge reference, and similarly

the $F_2^{\text{neq}}(T_a)$ in (32d) is strictly equal to $F_2^{\text{eq}}(T_b)$ in (32c). In the two processes of isothermal relaxation, irreversible heat $\mathcal{Q}_{12}(T_b) = T_b\{S_1(T_a) - S_2(T_b) + \Delta S_{12}^{(\text{i})}\}$ and $\mathcal{Q}_{21}(T_a) = T_a\{S_2(T_b) - S_1(T_a) + \Delta S_{21}^{(\text{i})}\}$ each contain an entropy production term,

$$\Delta S_{jk}^{(\text{i})} = \int_{\mathbb{R}} p_j^{\text{eq}}(y) \ln\left(\frac{p_j^{\text{eq}}(y)}{p_k^{\text{eq}}(y)}\right) \mathrm{d}y \geq 0. \tag{33}$$

In a Carnot cycle with *quasi-static* processes, they are assumed to be zero. Then, the total work done by the system over the cycle is

$$\begin{aligned} W &= -\left(\overline{\mathcal{W}}_{12} + \overline{\mathcal{W}}_{21}\right) = -\mathcal{Q}_{12}(T_b) - \mathcal{Q}_{21}(T_a) \\ &= T_b\left\{S_2 - S_1 - \int_{\mathbb{R}} p_1^{\text{eq}} \ln\left(\frac{p_1^{\text{eq}}}{p_2^{\text{eq}}}\right) \mathrm{d}y\right\} + T_a\left\{S_1 - S_2 - \int_{\mathbb{R}} p_2^{\text{eq}} \ln\left(\frac{p_2^{\text{eq}}}{p_1^{\text{eq}}}\right) \mathrm{d}y\right\} \\ &\leq T_b\Big[S_2(T_b) - S_1(T_a)\Big] + T_a\Big[S_1(T_a) - S_2(T_b)\Big], \end{aligned} \tag{34}$$

in which the reversible heat being absorbed at T_a is $Q_h = T_a[S_1(T_a) - S_2(T_b)] > 0$, and the heat being expelled at T_b is $Q_l = T_b[S_2(T_b) - S_1(T_a)] < 0$. Thus the Carnot (first-law) efficiency

$$\eta_{Carnot} = \frac{W}{Q_h} \leq 1 - \frac{T_b}{T_a}. \tag{35}$$

On the other hand, since the rhs of (34) is the maximum possible work, the second-law, exergy efficiency

$$\eta_{exergy} = \frac{W}{(T_a - T_b)[S_1(T_a) - S_2(T_b)]} = \frac{W}{Q_h\left(1 - \frac{T_b}{T_a}\right)} \leq 1. \tag{36}$$

Stirling cycle. There are many different realizations of heat engines in terms of thermodynamic cycles. We now consider the Stirling cycle below.

$$\begin{aligned} \text{isothermal working } \{T_a, U_1\} \to \{T_a, U_2\}: &\quad \overline{U}_1(T_a) - \overline{U}_2(T_a) = \overline{\mathcal{Q}}_{12}(T_a) - \overline{\mathcal{W}}_{12}, && (37a)\\ \text{isochoric cooling } \{T_a, U_2\} \to \{T_b, U_2\}: &\quad \overline{U}_2(T_a) - \overline{U}_2(T_b) = \mathcal{Q}_2(T_a, T_b), && (37b)\\ \text{equilibrium under } \{T_b, U_2\}: &\quad F_2^{\text{eq}}(T_b) = \overline{U}_2(T_b) - T_b S_2(T_b), && (37c)\\ \text{isothermal working } \{T_b, U_2\} \to \{T_b, U_1\}: &\quad \overline{U}_2(T_b) - \overline{U}_1(T_b) = \overline{\mathcal{Q}}_{21}(T_b) - \overline{\mathcal{W}}_{21}, && (37d)\\ \text{isochoric heating } \{T_b, U_1\} \to \{T_a, U_1\}: &\quad \overline{U}_1(T_b) - \overline{U}_1(T_a) = \mathcal{Q}_1(T_b, T_a), && (37e)\\ \text{equilibrium under } \{T_a, U_1\}: &\quad F_1^{\text{eq}}(T_a) = \overline{U}_1(T_a) - T_a S_1(T_a). && (37f) \end{aligned}$$

After two isothermal processes in (37a), (37d), the system is still in the equilibrium states with free energy $F_2^{\text{eq}}(T_a) = \overline{U}_2(T_a) - T_a S_2(T_a)$ and $F_1^{\text{eq}}(T_b) = \overline{U}_1(T_b) - T_b S_1(T_b)$ respectively. Notice the difference between the equilibrium free energy above and the non-equilibrium free energy functions $F_2^{\text{neq}}(T_a)$ and $F_1^{\text{neq}}(T_b)$ defined in (32a) and (32d). The irreversible heats for the two isothermal processes are

$$\overline{\mathcal{Q}}_{12}(T_a) = T_a\left[S_1(T_a) - S_2(T_a) + \int_\Omega \rho_1(x; T_a) \ln\left(\frac{\rho_1(x; T_a)}{\rho_2(x; T_a)}\right) \mathrm{d}x\right], \tag{38}$$

$$\overline{\mathcal{Q}}_{21}(T_b) = T_b\left[S_2(T_b) - S_1(T_b) + \int_\Omega \rho_2(x; T_b) \ln\left(\frac{\rho_2(x; T_b)}{\rho_1(x; T_b)}\right) \mathrm{d}x\right]. \tag{39}$$

Meanwhile, those for the isochoric cooling and heating processes are

$$\mathcal{Q}_2(T_a, T_b) = T_b\left[S_2(T_a) - S_2(T_b) + \int_\Omega \rho_2(x; T_a) \ln\left(\frac{\rho_2(x; T_a)}{\rho_2(x; T_b)}\right) \mathrm{d}x\right], \tag{40}$$

$$\mathcal{Q}_1(T_b, T_a) = T_a\left[S_1(T_b) - S_1(T_a) + \int_\Omega \rho_1(x; T_b) \ln\left(\frac{\rho_1(x; T_b)}{\rho_1(x; T_a)}\right) \mathrm{d}x\right]. \tag{41}$$

Summarizing the whole heat cycle, we find that

$$\begin{aligned} W &= -\left(\overline{\mathcal{W}}_{12} + \overline{\mathcal{W}}_{21}\right) = -\overline{\mathcal{Q}}_{12}(T_a) - \mathcal{Q}_2(T_a, T_b) - \overline{\mathcal{Q}}_{21}(T_b) - \mathcal{Q}_1(T_b, T_a) \\ &\leq (T_a - T_b)\Big[S_2(T_a) - S_1(T_b)\Big]. \end{aligned} \tag{42}$$

This will lead to the same conclusions on the first-law and second-law efficiency for the Stirling cycle.

Realization of a reversible cycle. The Carnot cycle and Stirling cycle considered above are not truly reversible, once $U_1 \neq U_2$ or $T_a \neq T_b$. To achieve the theoretical maximal efficiency, we need to construct a reversible heat cycle through a series of quasi-static processes, each of which involves only an infinitesimal change in either U or T. Taking the Stirling

cycle as an example. In the first isothermal working step, we insert $N-1$ intermediate states between $\{T_a, U_1\}$ and $\{T_a, U_2\}$, that are $\{T_a, U_1+\triangle U\}, \{T_a, U_1+2\triangle U\}, \ldots, \{T_a, U_1+(N-1)\triangle U\}$ with $\triangle U = (U_2-U_1)/N$. In the limit of $N\to\infty$, $\triangle U\to 0$, which means each transition between two adjacent states can be treated as a quasi-static process. Therefore, the whole step between $\{T_a, U_1\}$ and $\{T_a, U_2\}$ becomes reversible with the help of those intermediate states. Applying similar procedure to other three steps, we will achieve a true thermodynamically reversible Stirling cycle by requiring an infinitesimal change in either U or T for each sub-step.

3.5. Work as a conditional expectation in energy representation

Consider once again two distributions $\rho(x)$ and $p^{\mathrm{eq}}(x)$ with respective energy representations, $\rho_E(y) = Z_1^{-1}(\beta_a)\Omega_1(y)e^{-\beta_a y}$ and $p_E^{\mathrm{eq}}(y) = Z_2^{-1}(\beta_b)\Omega_2(y)e^{-\beta_b y}$. The key thermodynamic quantity that arises in (22), the irreversible work, cannot be expressed in terms of the six quantities: $\Omega_1(y), Z_1(\beta), \Omega_2(y), Z_2(\beta)$, and β_a, β_b. We note that

$$\int_\Omega \rho(x)\big[U_2(x)-U_1(x)\big]\mathrm{d}x = \int_{\mathbb{R}}\left(\frac{\Omega_1(y)e^{-\beta_a y}}{Z_1(\beta_a)}\right)\left\{\overline{U}_{2|U_1=y}-y\right\}\mathrm{d}y, \tag{43}$$

in which

$$\overline{U}_{2|U_1=y} = \frac{\displaystyle\int_{y<U_1(x)\le y+\mathrm{d}h} U_2(x)\mathrm{d}x}{\displaystyle\int_{y<U_1(x)\le y+\mathrm{d}h}\mathrm{d}x}, \tag{44}$$

is a conditional expectation of $U_2(x)$ given $U_1(x)=y$. The energy functions $U_1(x)$ and $U_2(x)$ are only two observables on the probability space and they certainly do not provide a full description of the probability space. Actually, knowing the canonical energy distributions $\rho_E(y)$ and $p_E^{\mathrm{eq}}(y)$ is not equivalent to knowing their joint probability distribution; the missing information on their correlation is captured precisely in (44).

The lhs of (43) can also be expressed as

$$\begin{aligned}&\int_\Omega \rho(x)\Big[U_2(x)-U_1(x)\Big]\mathrm{d}x\\ &= \frac{1}{\beta_a}\left[\ln\frac{Z_1(\beta_a)}{Z_2(\beta_a)} + \int_\Omega \rho(x)\ln\left\{\frac{e^{-\beta_a U_1(x)}Z_2(\beta_a)}{Z_1(\beta_a)e^{-\beta_a U_2(x)}}\right\}\mathrm{d}x\right].\end{aligned} \tag{45}$$

The term inside $\{\cdots\}$ indeed can be understood as a Radon–Nikodym derivative between the two probability measures, which is well-defined on the entire σ-algebra $\mathcal{F}$ as well as the restricted joint σ-algebra $\mathcal{F}_{U_1,U_2}$. However, it is singular on the further restricted σ-algebra $\mathcal{F}_{U_1}$ or $\mathcal{F}_{U_2}$.

3.6. The role and consequence of determinism

Consider a sequence of measures μ_ϵ and two real-valued continuous random variables $\mathbf{x}(\omega)$ and $\mathbf{y}(\omega)$, with corresponding probability density functions $p_\epsilon(x)$ and $q_\epsilon(x)$. Their relative entropy is then

$$H\big[\mathbf{x}\parallel\mathbf{y};\,\mu_\epsilon\big] = \int_\Omega p_\epsilon(x)\ln\left(\frac{p_\epsilon(x)}{q_\epsilon(x)}\right)\mathrm{d}x. \tag{46}$$

If the sequence of measures μ_ϵ tends to a singleton with corresponding $p_\epsilon(x)\to\delta(x-z)$ and $q_\epsilon(x)\to\delta(x-y^*)$ as $\epsilon\to 0$, we call the limit *deterministic*.

It can be shown under rather weak conditions, or more properly through the theory of large deviations, that as $\epsilon\to 0$ the $p_\epsilon(x)$ and $q_\epsilon(x)$ have asymptotic forms

$$\ln p_\epsilon(x) = -\frac{\varphi_p(x)}{\epsilon} + O(\ln\epsilon),\ \ \ln q_\epsilon(x) = -\frac{\varphi_q(x)}{\epsilon} + O(\ln\epsilon), \tag{47}$$

in which $\varphi_p(z) = \varphi_q(y^*) = 0$. This asymptotic relation is known as the large deviations principle in the theory of probability [37]. Therefore,

$$\ln H\big[\mathbf{x}\parallel\mathbf{y};\,\mu_\epsilon\big] \sim \frac{\varphi_q(z)}{\epsilon} + O(\ln\epsilon), \tag{48}$$

as $\mathbf{x}\to z$. Even though $\mathbf{y}\to y^*$, the relative entropy in (46) provides the φ_q as a function of z fully supported on $\mathbb{R}^n$. If the q_ϵ is an invariant measure of a stochastic dynamical system, then the $\varphi_q(z)$ is thought of as a "deterministic energy function", which can be obtained as the asymptotic limit of determinism. The normalization of $e^{-\varphi_q(x)/\epsilon}$, however, is lost in the $\ln\epsilon$-order term. This corresponds to a certain gauge freedom.

A combination of the determinism with the canonical distribution immediately yields a key relationship that is well known in thermodynamics. Specifically, if the probability density function

$$\frac{\Omega^{(B)}(E)e^{-\beta E}}{Z(\beta)} = \frac{e^{-\beta E+\ln\Omega^{(B)}(E)}}{Z(\beta)} \to \delta\big(E-E^*\big), \tag{49}$$

in an asymptotic limit, then one has the *equation of state*

$$\left[\frac{\mathrm{d}}{\mathrm{d}E}\Big(\beta E-\ln\Omega^{(B)}(E)\Big)\right]_{E=E^*} = 0. \tag{50}$$

A system in macroscopic thermodynamic equilibrium possesses one less degree of freedom [10]. Eq. (50) implies

$$\beta = \frac{\mathrm{d}\ln\Omega^{(B)}(E^*)}{\mathrm{d}E} = \frac{\frac{\mathrm{d}}{\mathrm{d}E}\Omega^{(B)}(E^*)}{\Omega^{(B)}(E^*)}, \tag{51}$$

in which

$$\begin{aligned}\Omega^{(B)}(E) &= \frac{1}{\mathrm{d}E}\int_{E<U(x)\le E+\mathrm{d}E}\mathrm{d}x \;=\; \oint_{U(x)=E}\frac{\mathrm{d}\Sigma\cdot\hat{\mathbf{n}}}{\|\nabla U(x)\|}\\ &= \int_{U(x)\le E}\nabla\cdot\left(\frac{\nabla U(x)}{\|\nabla U(x)\|^2}\right)\mathrm{d}x,\end{aligned} \tag{52}$$

$$\begin{aligned}\frac{\mathrm{d}\Omega^{(B)}(E)}{\mathrm{d}E} &= \frac{1}{\mathrm{d}E}\int_{E<U(x)\le E+\mathrm{d}E}\nabla\cdot\left(\frac{\nabla U(x)}{\|\nabla U(x)\|^2}\right)\mathrm{d}x\\ &= \oint_{U(x)\le E}\nabla\cdot\left(\frac{\nabla U(x)}{\|\nabla U(x)\|^2}\right)\frac{\mathrm{d}\Sigma\cdot\hat{\mathbf{n}}}{\|\nabla U(x)\|}.\end{aligned} \tag{53}$$

Therefore,

$$\frac{\mathrm{d}\ln\Omega^{(B)}(E)}{\mathrm{d}E} = \frac{\displaystyle\oint_{U(x)=E}\nabla\cdot\left(\frac{\nabla U(x)}{\|\nabla U(x)\|^2}\right)\frac{\mathrm{d}\Sigma\cdot\hat{\mathbf{n}}}{\|\nabla U(x)\|}}{\displaystyle\oint_{U(x)=E}\frac{\mathrm{d}\Sigma\cdot\hat{\mathbf{n}}}{\|\nabla U(x)\|}}. \tag{54}$$

That is, the equilibrium β is the average of

$$\nabla\cdot\left(\frac{\nabla U}{\|\nabla U\|^2}\right) = \frac{\|\nabla U\|\nabla^2 U-2\nabla U\cdot\nabla\|\nabla U\|}{\|\nabla U\|^3}, \tag{55}$$

on the level-surface $\{x : U(x) = E^*\}$. For a given energy function $U(x)$, or an observable [14], Eq. (54), which generalizes the virial theorem in classical mechanics, provides the function $\beta(E)$.

4. The space of probability measures

4.1. Affine structure, canonical distribution and its energy representation

We will now give a brief, non-rigorous introduction to the theory developed in [9]. Let $\mathcal{M}$ be the set of all probability measures on $(\Omega,\mathcal{F})$ that are absolutely continuous w.r.t. some probability measure $\mathbb{P}$ (and therefore absolutely continuous w.r.t. each other) and let $\mathcal{V}$ be an appropriate set of real-valued functions on Ω. (Note that any choice of $\mathbb{P}$ in $\mathcal{M}$ would do; one only cares that all measures in $\mathcal{M}$ are absolutely continuous w.r.t each other.) One now defines $\oplus\colon \mathcal{M}\times\mathcal{V}\to\mathcal{M}$ such that

$$\big(\mu\oplus g\big)(A) = \frac{\displaystyle\int_A e^g\mathrm{d}\mu}{\displaystyle\int_\Omega e^g\mathrm{d}\mu}, \tag{56}$$

for any $A\in\mathcal{F}$. Assuming the denominator is finite (which requires some assumptions on $\mathcal{V}$), the positivity of e^g implies that $(\mu\oplus g)$ is also absolutely continuous w.r.t. $\mathbb{P}$. Since $(\mu\oplus g)(\Omega)=1$, it is a probability measure. These two facts mean that $(\mu\oplus g)\in\mathcal{M}$, so the operation $\oplus$ is well-defined. Note that $\mu\oplus g=\mu\oplus(g+c)$ for any constant c, so this addition is not actually one-to-one. We can remedy this issue by restricting $\mathcal{V}$ to functions that sum to zero, or we can replace each function with an equivalence class of functions that differ by a constant. One can then show that $(\mathcal{M},\mathcal{V},\oplus)$ is an affine structure on $\mathcal{M}$ [8,9]. If one chooses a particular measure $\mathbb{P}\in\mathcal{M}$ as the origin, then any other measure $\mu\in\mathcal{M}$ will have a Radon–Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{P}}(\omega)$, and $\mu=(\mathbb{P}\oplus g)$ where $g=\ln\big(\frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{P}}\big)$.

Let $J \subseteq \mathbb{R}$ be an interval and $U \in \mathcal{V}$. The function $p: J \to \mathcal{M}$ such that $p(\beta) = \mathbb{P} \oplus (-\beta U)$ is an affine straight line. More explicitly, we have the family of probability densities

$$\frac{e^{-\beta U(\omega)}}{Z(\beta)}\mathbb{P}(\mathrm{d}\omega), \tag{57}$$

where $Z(\beta)$ is the normalization factor.

In Kolmogorov's theory, the real-valued function $U(\omega)$, when thought of as a random variable, has its own probability density function w.r.t. the Lebesgue measure:

$$\mathbb{P}\Big\{y < U(\omega) \le y + \mathrm{d}h\Big\} = \frac{\Omega_U(y)}{Z(\beta)}e^{-\beta y}\mathrm{d}h, \tag{58}$$

in which $\ln \Omega_U(y)$ is the Gibbs entropy associated with function $U(\omega)$, defined in Eq. (12):

$$\Omega_U(y) = \frac{1}{\mathrm{d}h}\int_{y<U(\omega)\le y+\mathrm{d}h} \mathbb{P}(\mathrm{d}\omega). \tag{59}$$

The relation between the distributions in (57) and (58) establishes a map between the observables in the tangent space $\mathcal{V}$ of $\mathcal{M}$ and the standard probability density functions. (This is analogous to the dual relation between the Koopman operator on the space of observables and the Perron–Frobenius operator on the space of densities in dynamical systems theory.) We call (57) the *canonical representation* for the space of probability measures (SoPMs), and (58) its *energy representation*. Note that the energy representation of a given probability measure is not unique. The choice of U depends on both $\mathbb{P}$ and β.

A pair of observables. We now discuss the notions of joint, marginal, and conditional probability in terms of the canonical representation in $\mathcal{V}$, with a fixed "origin" $\mathbb{P}$, which should be thought of as the $\mathbb{P}$ in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, *à la* Kolmogorov. The SoPMs $\mathcal{M}$ then is represented by observables $U(\omega) \in \mathcal{V}$, the tangent space of $\mathcal{M}$.

Consider two observables $\big(U_1(\omega), U_2(\omega)\big)$, where $U_2 \neq aU_1 + b$. The corresponding "flat plane" can be parametrized as

$$\frac{e^{-\beta_a U_1(\omega)-\beta_b U_2(\omega)}}{Z_{1,2}(\beta_a,\beta_b)}\mathbb{P}(\mathrm{d}\omega),\ (\beta_a,\beta_b) \in \mathbb{R}^2. \tag{60}$$

We note that each observable induced a restricted σ-algebra on $\mathbb{R}$: $\mathcal{F}_{U_1}$ and $\mathcal{F}_{U_2}$ respectively, and the joint observable induces $\mathcal{F}_{U_1,U_2} = \sigma(\mathcal{F}_{U_1} \cup \mathcal{F}_{U_2})$. With respect to $\mathcal{F}_{U_1,U_2}$, the distribution in (60) can expressed on as:

$$\frac{\Omega_{1,2}(y_1,y_2)}{Z_{1,2}(\beta_a,\beta_b)}e^{-\beta_a y_1-\beta_b y_2}\mathrm{d}y_1\mathrm{d}y_2, \tag{61a}$$

in which

$$\Omega_{1,2}(y_1,y_2) = \frac{1}{\mathrm{d}y_1\mathrm{d}y_2}\int_{y_1<U_1(\omega)\le y_1+\mathrm{d}y_1,\ y_2<U_2(\omega)\le y_2+\mathrm{d}y_2} \mathrm{d}\mathbb{P}(\omega) \tag{61b}$$

$$= \frac{\partial^2}{\partial y_1 \partial y_2}\int_{U_1(\omega)\le y_1,\ U_2(\omega)\le y_2} \mathrm{d}\mathbb{P}(\omega). \tag{61c}$$

We note that the marginal distribution

$$\int_{\mathbb{R}} \frac{\Omega_{1,2}(y_1,y_2)}{Z_{1,2}(\beta_a,\beta_b)}e^{-\beta_a y_1-\beta_b y_2}\mathrm{d}y_2 = \frac{1}{Z_1(\beta_a)}\left(\int_{\mathbb{R}}\Omega_{1,2}(y_1,y_2)\mathrm{d}y_2\right)e^{-\beta_a y_1}. \tag{62}$$

This implies that

$$\frac{1}{Z_1(\beta_a)}\int_{\mathbb{R}}\Omega_{1,2}(y_1,y_2)\mathrm{d}y_2 = \frac{1}{Z_{1,2}(\beta_a,\beta_b)}\int_{\mathbb{R}}\Omega_{1,2}(y_1,y_2)e^{-\beta_b y_2}\mathrm{d}y_2. \tag{63}$$

Since the rhs of (63) is not a function of β_b, we have the following equality:

$$\frac{\partial \ln Z_{1,2}(\beta_a,\beta_b)}{\partial \beta_b} = -\frac{\displaystyle\int_{\mathbb{R}} y_2\Omega_{1,2}(y_1,y_2)e^{-\beta_b y_2}\mathrm{d}y_2}{\displaystyle\int_{\mathbb{R}}\Omega_{1,2}(y_1,y_2)e^{-\beta_b y_2}\mathrm{d}y_2}. \tag{64}$$

Eq. (63) can also be re-arranged into

$$\frac{\displaystyle\int_{\mathbb{R}}\Omega_{1,2}(y_1,y_2)e^{-\beta_b y_2}\mathrm{d}y_2}{\displaystyle\int_{\mathbb{R}}\Omega_{1,2}(y_1,y_2)\mathrm{d}y_2} = \frac{Z_{1,2}(\beta_a,\beta_b)}{Z_1(\beta_a)}, \tag{65a}$$

$$\frac{\int_{\mathbb{R}} \Omega_{1,2}(y_1, y_2) e^{-\beta_b y_2} \left(e^{\beta_b y_2}\right) \mathrm{d}y_2}{\int_{\mathbb{R}} \Omega_{1,2}(y_1, y_2) e^{-\beta_b y_2} \mathrm{d}y_2} = \frac{Z_1(\beta_a)}{Z_{1,2}(\beta_a, \beta_b)}. \tag{65b}$$

Relative entropy between two random variables. The relative entropy between two measures $\mu_1 = \left(\mathbb{P} \oplus (-\beta_a U_1)\right) \in \mathcal{M}$ and $\mu_2 = \left(\mathbb{P} \oplus \left(-\beta_b U_2(\omega)\right)\right) \in \mathcal{M}$, when transformed into the energy representation, is given by:

$$\begin{aligned} \int_\Omega \ln\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(\omega)\right) \mathrm{d}\mu_1(\omega) &= \int_\Omega \frac{e^{-\beta_a U_1(\omega)}}{Z_1(\beta_a)} \ln\left(\frac{Z_2(\beta_b)}{Z_1(\beta_a)} e^{-\beta_a U_1(\omega) + \beta_b U_2(\omega)}\right) \mathrm{d}\mathbb{P}(\omega) \\ &= \int_{\mathbb{R}} \frac{\Omega_{U_1}(h) e^{-\beta_a h}}{Z_1(\beta_a)} \ln\left(\frac{\Omega_{U_2}(h) e^{-\beta_a h}}{Z_1(\beta_a) \Omega_{U_1}(h)}\right) \mathrm{d}h \qquad (66a) \\ &\quad + \beta_b \int_\Omega U_2(\omega) \left(\frac{e^{-\beta_a U_1(\omega)}}{Z_1(\beta_a)}\right) \mathrm{d}\mathbb{P}(\omega) \; + \; \ln Z_2(\beta_b). \qquad (66b) \end{aligned}$$

Note that the first term in (66b) again contains the $\overline{U}_{2|U_1=h_1}$ that appeared in (25) and (44). It cannot be expressed in terms of the energy representations of μ_1 and μ_2. Unless $g_2 = ag_1 + b$, the two measures μ_1 and μ_2, with densities $\mathrm{d}\mu_1 = e^{g_1}\mathrm{d}\mathbb{P}$ and $\mathrm{d}\mu_2 = e^{g_2}\mathrm{d}\mathbb{P}$, do not share the same restricted σ-algebra.

4.2. Entropy divergence in the SoPMs

Consider two probability measures $\mu_1, \mu_2 \in \mathcal{M}$ in the SoPMs, with Radon–Nikodym derivatives w.r.t. $\mathbb{P}$ given by $f_1(\omega)$ and $f_2(\omega)$ respectively. One can introduce the following divergence on $\mathcal{M}$:

$$d^2\left(\mu_1, \mu_2\right) = \int_\Omega \left(f_1(\omega) - f_2(\omega)\right) \left(\frac{\ln f_1(\omega) - \ln f_2(\omega)}{f_1(\omega) - f_2(\omega)}\right) \left(f_1(\omega) - f_2(\omega)\right) \mathbb{P}(\mathrm{d}\omega). \tag{67}$$

This divergence can also be rewritten as the sum of two non-negative terms in the form of relative entropy, a symmetrized version of the latter:

$$d^2\left(\mu_1, \mu_2\right) = \int_\Omega \ln\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(\omega)\right) \mu_1(\mathrm{d}\omega) + \int_\Omega \ln\left(\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}(\omega)\right) \mu_2(\mathrm{d}\omega). \tag{68}$$

From this second form, it is clear that d is symmetric with respect to μ_1 and μ_2 and is zero if and only if $\mu_1 = \mu_2$ on $\mathcal{F}$. This form also has the advantage of making it clear that d is invariant with respect to the choice of an origin $\mathbb{P}$. Note that, despite our notation, this quantity is not a metric because it does not satisfy the triangle inequality. It is only a local metric. That is, if μ_1, μ_2 and μ_3 are sufficiently close together then $d(\mu_1, \mu_2) + d(\mu_2, \mu_3) \geq d(\mu_1, \mu_3)$.

Divergence in energy representation. If μ_1 and μ_2 are written in their respective energy representations, *i.e.* $f_{E1}(y_1) = Z_1(\beta_a)\Omega_1(y_1)e^{-\beta_a y_1}$ and $f_{E2}(y_2) = Z_2(\beta_b)\Omega_2(y_2)e^{-\beta_b y_2}$. Then from Eq. (68), we have

$$\begin{aligned} d^2\left(\mu_1, \mu_2\right) \; &= \; \beta_b \int_{\mathbb{R}} \left(\frac{\Omega_1(y_1) e^{-\beta_a y_1}}{Z_1(\beta_a)}\right) \overline{U}_{2|U_1=y_1} \, \mathrm{d}y_1 - \beta_a \overline{U}_1(\beta_a) - \beta_b \overline{U}_2(\beta_b) \\ &\quad + \beta_a \int_{\mathbb{R}} \left(\frac{\Omega_2(y_2) e^{-\beta_b y_2}}{Z_2(\beta_b)}\right) \overline{U}_{1|U_2=y_2} \, \mathrm{d}y_2. \end{aligned} \tag{69}$$

There are three interesting special cases:

Different β's and same Ω. If $\Omega_1(y) = \Omega_2(y) = \Omega(y)$,

$$d^2\left(\mu_1, \mu_2\right) = \left(\beta_b - \beta_a\right)\left(\overline{U}(\beta_a) - \overline{U}(\beta_b)\right). \tag{70}$$

Different Ω's and same β. With same $\beta_a = \beta_b = \beta$ but different Ω's,

$$\begin{aligned} d^2\left(\mu_1, \mu_2\right) \; &= \; \beta \int_{\mathbb{R}} \left(\frac{\Omega_1(y_1) e^{-\beta y_1}}{Z_1(\beta)}\right) \left[\overline{U}_{2|U_1=y_1} - y_1\right] \mathrm{d}y_1 \qquad (71a) \\ &\quad + \beta \int_{\mathbb{R}} \left(\frac{\Omega_2(y_2) e^{-\beta y_2}}{Z_2(\beta)}\right) \left[\overline{U}_{1|U_2=y_2} - y_2\right] \mathrm{d}y_2 \qquad (71b) \\ &= \beta\left(\overline{\mathcal{W}}_{12}(\beta) + \overline{\mathcal{W}}_{21}(\beta)\right). \qquad (71c) \end{aligned}$$

Here, following (22) and (23), we have identified the terms in (71a) and (71b) as $\overline{\mathcal{W}}_{12}(\beta)$ and $\overline{\mathcal{W}}_{21}(\beta)$, respectively.

Different Ω's and β's.

$$d^2\left(\mu_1, \mu_2\right) = \beta_b \overline{\mathcal{W}}_{12}(\beta_a) + \beta_a \overline{\mathcal{W}}_{21}(\beta_b) + \left(\beta_b - \beta_a\right)\left(\overline{U}_1(\beta_a) - \overline{U}_2(\beta_b)\right). \tag{72}$$

Eq. (72) implies an inequality that, being different from (35) and (36), is based on Massieu–Planck potential:

$$-\left(\frac{\overline{\mathcal{W}}_{12}}{T_b}+\frac{\overline{\mathcal{W}}_{21}}{T_a}\right)\le\left(\overline{U}_1-\overline{U}_2\right)\left(\frac{1}{T_b}-\frac{1}{T_a}\right). \tag{73}$$

4.3. Heat divergence

One can also introduce another related divergence on $\mathcal{M}$. For fixed β_a, $\beta_b>0$, define:

$$\begin{aligned} d_\beta^2(\mu_1,\mu_2) &= \frac{1}{\beta_a}\int_\Omega f_1(\omega)\ln\left(\frac{f_1(\omega)}{f_2^{(\beta_a)}(\omega)}\right)\mathbb{P}(\mathrm{d}\omega)+\frac{1}{\beta_b}\int_\Omega f_2(\omega)\ln\left(\frac{f_2(\omega)}{f_1^{(\beta_b)}(\omega)}\right)\mathbb{P}(\mathrm{d}\omega)\\ &= \int_\Omega\left(\frac{e^{-\beta_a U_1(\omega)}}{Z_1(\beta_a)}-\frac{e^{-\beta_b U_2(\omega)}}{Z_2(\beta_b)}\right)\Big(U_2(\omega)-U_1(\omega)\Big)\mathbb{P}(\mathrm{d}\omega)\\ &+\frac{1}{\beta_a}\ln\left(\frac{Z_2(\beta_a)}{Z_1(\beta_a)}\right)-\frac{1}{\beta_b}\ln\left(\frac{Z_2(\beta_b)}{Z_1(\beta_b)}\right), \end{aligned} \tag{74}$$

in which

$$f_1(\omega)=\frac{e^{-\beta_a U_1(\omega)}}{Z_1(\beta_a)}\ \text{and}\ f_2(\omega)=\frac{e^{-\beta_b U_2(\omega)}}{Z_2(\beta_b)} \tag{75}$$

are the densities of μ_1 and μ_2 with respect to $\mathbb{P}$ and

$$f_2^{(\beta_a)}(\omega)=\frac{e^{-\beta_a U_2(\omega)}}{Z_2(\beta_a)},\ f_1^{(\beta_b)}(\omega)=\frac{e^{-\beta_b U_1(\omega)}}{Z_1(\beta_b)}. \tag{76}$$

The same caveats as before apply: This is not a metric on $\mathcal{M}$ because it does not satisfy the triangle inequality, but it is a local metric in the sense that the triangle inequality is satisfied when all measures are sufficiently close together. We shall call $d_\beta(\cdot,\cdot)$ in (74) the *heat divergence*. In terms of

$$\mathcal{W}_{12}(\omega)=\frac{1}{\beta_a}\ln\left(\frac{e^{-\beta_a U_1(\omega)}}{e^{-\beta_a U_2(\omega)}}\right),\ \mathcal{W}_{21}(\omega)=\frac{1}{\beta_b}\ln\left(\frac{e^{-\beta_b U_2(\omega)}}{e^{-\beta_b U_1(\omega)}}\right), \tag{77}$$

we have

$$\begin{aligned} d_\beta^2(\mu_1,\mu_2) &= \mathbb{E}^{\mu_1}\big[\mathcal{W}_{12}(\omega)\big]+\beta_a^{-1}\ln\mathbb{E}^{\mu_1}\Big[e^{-\beta_a\mathcal{W}_{12}(\omega)}\Big]+\mathbb{E}^{\mu_2}\big[\mathcal{W}_{21}(\omega)\big]\\ &+\beta_b^{-1}\ln\mathbb{E}^{\mu_2}\Big[e^{-\beta_b\mathcal{W}_{21}(\omega)}\Big]. \end{aligned} \tag{78}$$

Using the Jarzynski–Crooks relation from (28), Eq. (78) implies

$$\overline{\mathcal{W}}_{12}(\beta_a)+\overline{\mathcal{W}}_{21}(\beta_b)+F_1(\beta_a)-F_2(\beta_a)+F_2(\beta_b)-F_1(\beta_b)\ge 0. \tag{79}$$

This result generalizes Carnot's inequality.

4.4. Infinitesimal entropy metric associated with $\Delta\beta$

Consider an infinitesimal change in $\beta\to\beta+\Delta\beta$ and corresponding $\mathrm{d}\mu=e^{-\beta U}\mathrm{d}\mathbb{P}\to\mathrm{d}(\mu+\Delta\mu)=e^{-(\beta+\Delta\beta)U}\mathrm{d}\mathbb{P}$. Then we have

$$\begin{aligned} d^2\big(\mu,\mu+\Delta\mu\big) &= (\Delta\beta)^2\int_0^\infty\frac{\Omega(y)e^{-\beta y}}{Z(\beta)}\left[\left(\frac{\mathrm{d}\ln Z}{\mathrm{d}\beta}\right)+y\right]^2\mathrm{d}y\\ &= (\Delta\beta)^2\int_0^\infty\frac{\Omega(y)e^{-\beta y}}{Z(\beta)}\Big(y-\mathbb{E}[U]\Big)^2\mathrm{d}y\\ &= (\Delta\beta)^2\,\mathrm{Var}\big[U\big]. \end{aligned} \tag{80}$$

This is a very important relation that connects the *entropy divergence* with *temperature* and *energy fluctuations*. Furthermore, we have

$$d^2\big(\mu,\mu+\Delta\mu\big)=(\Delta\beta)^2\left(-\frac{\mathrm{d}^2\ln Z(\beta)}{\mathrm{d}\beta^2}\right)\ =\ (\Delta\beta)^2\left(\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}\big[U\big]\right). \tag{81}$$

The term inside $(\cdots)$ on the rhs is called the *heat capacity* in thermodynamics. Internal energy $\mathbb{E}[U]$ is a "slope" and the $\mathrm{Var}[X_\beta]$ is a curvature of the "potential function" $-\ln Z(\beta)$.

4.5. A mathematical remark

Log-mean-exponential inequality and equality. We see that both entropy divergence in (68) and heat divergence in (78) are based on a very general inequality involving the log-mean-exponential of a random variable $\xi(\omega)$ [38]: Jensen's inequality.

$$\mathbb{E}\Big[\xi(\omega)\Big] + \beta^{-1} \ln \mathbb{E}\left[e^{-\beta\xi(\omega)}\right] \geq 0. \tag{82}$$

In (68), the two ξs are the information $\ln \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(\omega)$ and $\ln \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}(\omega)$; and in (78), the two ξs are the work $\mathcal{W}_{12}(\omega) = \beta_a^{-1} \ln \frac{e^{-\beta_a U_1(\omega)}}{e^{-\beta_a U_2(\omega)}}$ and $\mathcal{W}_{21}(\omega) = \beta_b^{-1} \ln \frac{e^{-\beta_b U_2(\omega)}}{e^{-\beta_b U_1(\omega)}}$. They are all different forms of Radon–Nikodym derivatives. In the entropy divergence, the second, log-mean-exponential term in (82) is zero according to the Hatano–Sasa equality. In the heat divergence case, the same term gives a Jarzynski–Crooks' free energy difference.

Eq. (82) should be recognized as "mean internal energy minus free energy". Thus it should be some kind of entropy:

$$\mathbb{E}^{\mathbb{P}}\Big[\xi(\omega)\Big] + \beta^{-1} \ln \mathbb{E}^{\mathbb{P}}\Big[e^{-\beta\xi(\omega)}\Big] = \mathbb{E}^{\mathbb{P}}\left[\ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}'}(\omega)\right)\right], \tag{83}$$

in which $\mathbb{P}' = \mathbb{P} \oplus (-\beta\xi)$ is the affine sum of $\mathbb{P}$ and $(-\beta\xi)$. Eq. (83) could be argued as the *fundamental equation for isothermal processes* under a single temperature $T = \beta^{-1}$. The implication of this interesting "Jensen's equality" to the affine geometry of the SoPMs is currently being explored.

5. Discussion

It has been well established, through the work of Gibbs, Carathéodory, and many others, that geometry has a role in the theory of equilibrium thermodynamics [39–41]. Classical thermodynamics is not based on the theory of chance, but there is no doubt that the notion of entropy has its root in the theory of probability. In the present work, we propose that the space of probability measures as a natural setting in which thermodynamic concepts can be established logically. In particular, an affine structure is naturally related to the canonical probability distribution studied by Boltzmann and Gibbs in their statistical theories, and almost all thermodynamic potentials are different forms of Radon–Nikodym derivatives associated with *changes of measures*. Even the fundamental equation of nonequilibrium thermodynamics, together with the distinctly nonequilibrium notion of entropy production, naturally emerges.

Statistical mechanics, as a scientific theory, differs from Kolmogorov's axiomatic theory of probability in one essential point: The latter demands a complete probability space and a normalized probability measure, while in the former every probability distribution is a *conditioned probability* under many known and unknown conditions. More importantly, the probability of the conditions, themselves as random events, are usually not knowable. In the theory of the space of measures, we see that one mechanical system with a given energy function $U(\omega)$ corresponds to a straight line, and the fixing of the origin in $\mathcal{M}$ in terms of $\mathbb{P}$ or the normalization in terms of $Z(\beta)$ [which translates to the arbitrary constant in $U(\omega)$] amount to the idea of gauge fixing. Thermodynamic work then arises in the rotation from $U_1(\omega)$ to $U_2(\omega)$. In the theory of probability, associated with any "change" is a *change of measure*: Radon–Nikodym derivatives simply provide the calculus to quantity the *fluxion*! In Newtonian mechanics, change in space is absolute; but in probability, it is a complex matter, and it is all relative.

The probability theory of large deviations is now a recognized mathematical foundation for statistical thermodynamics [37,42,43]. Such a theory is concerned with the deterministic *thermodynamic limit*. In Section 3.6, we see that the combination of our theory and a deterministic limit gives rise to the concept of *macroscopic equations of state* in classic thermodynamics [10].

Equilibrium mean internal energy $\overline{U}(\beta)$ depends on both the intrinsic properties of a system and its external environment. This is most clearly shown through the canonical distribution that is determined by $U(\omega)$ and β. The decomposition in Eq. (15), a simple example of the much more general (83), connects the internal energy with "work" and "heat", or the "usable energy" and "useless energy", or entropy production and entropy change. These are all just different interpretations under different perspectives.

Acknowledgments

We thank Yu-Chen Cheng and Ying-Jen Yang for many helpful discussions, and Professors Jin Feng (University of Kansas) and Hao Ge (Peking University) for advices. L.H. acknowledges the financial supports from the National Natural Science Foundation of China (Grants 21877070) and Tsinghua University Initiative Scientific Research Program (Grants 20151080424). H.Q. acknowledges the Olga Jung Wan Endowed Professorship for support.

References

[1] A.N. Kolmogorov, S.V. Fomin, Introductory Real Analysis, Silverman, R.A. Transl., Dover, New York, 1968.
[2] H. Qian, Mesoscopic nonequilibrium thermodynamics of single macromolecules and dynamic entropy-energy compensation, Phys. Rev. E 65 (2001) 016102.
[3] F.X.F. Ye, H. Qian, Stochastic dynamics II: Finite random dynamical systems, linear representation, and entropy production, Discrete Contin. Dyn. Syst. B 24 (2019) 4341–4366.
[4] W. Feller, The general diffusion operator and positive preserving semi-group in one dimension, Ann. of Math. 60 (1954) 417–436.
[5] E. Nelson, An existence theorem for second order parabolic equations, Trans. Amer. Math. Soc. 88 (1958) 414–429.
[6] D.Q. Jiang, M. Qian, M.P. Qian, Mathematical Theory of Nonequilibrium Steady States, Springer, New York, 2004.
[7] G.A. Pavliotis, Stochastic Processes and Applications, Springer, New York, 2014.
[8] J. Gallier, Geometric Methods and Applications, second ed., Springer, New York, 2011.
[9] L.F. Thompson, Affine structures, geometry and thermodynamics, 2020, Manuscript in preparation.
[10] W. Pauli, Pauli Lectures on Physics: Vol. 3, Thermodynamics and the Kinetic Theory of Gases; Vol. 4, Statistical Mechanics, The MIT Press, Cambridge, MA, 1973.
[11] H. Qian, Y.-C. Cheng, L.F. Thompson, Ternary representation of stochastic change and the origin of entropy and its fluctuations, 2019, arXiv:1902.09536.
[12] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, New York, 1991.
[13] Y.-J. Yang, H. Qian, Unified formalism for entropy production and fluctuation relations, Phys. Rev. E 101 (2020) 022129.
[14] Y.-C. Cheng, H. Qian, Y. Zhu, Asymptotic behavior of a sequence of conditional probability distributions and the canonical ensemble, 2020, arXiv:1912.11137.
[15] H. Qian, Information and entropic force: physical description of biological cells, chemical reaction kinetics, and information theory, Sci. Sin. Vitae 47 (2017) 257–261 (in Chinese).
[16] A.N. Kolmogorov, Three approaches to the quantitative definition of information, Int. J. Comput. Math. 2 (1968) 157–168.
[17] M. Tribus, Thermostatics and Thermodynamics: An Introduction to Energy, Information and States of Matter, with Engineering Applications, D. van Nostrand, New York, 1961.
[18] M.C. Mackey, The dynamic origin of increasing entropy, Rev. Modern Phys. 61 (1989) 981–1015.
[19] H. Qian, Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations, Phys. Rev. E 63 (2001) 042103.
[20] I. Prigogine, Introduction to Thermodynamics of Irreversible Processes, Charles C. Thomas Pub., Springfield, IL, 1955.
[21] H. Qian, Thermodynamics of the general diffusion process: Equilibrium supercurrent and nonequilibrium driven circulation with dissipation, Eur. Phys. J. Spec. Top. 224 (2015) 781–799.
[22] H. Qian, S. Kjelstrup, A.B. Kolomeisky, D. Bedeaux, Entropy production in mesoscopic stochastic thermodynamics - Nonequilibrium kinetic cycles driven by chemical potentials, temperatures, and mechanical forces, J. Phys. Condens. Matter. 28 (2016) 153004.
[23] R.W. Zwanzig, High-temperature equation of state by a perturbation method. I. Nonpolar gases, J. Chem. Phys. 22 (1954) 1420–1426.
[24] C. Jarzynski, Nonequilibrium equality for free energy differences, Phys. Rev. Lett. 78 (1997) 2690–2693.
[25] G. Crooks, Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences, Phys. Rev. E 60 (1999) 2721–2726.
[26] T. Hatano, S.I. Sasa, Steady-state thermodynamics of Langevin systems, Phys. Rev. Lett. 86 (2001) 3463–3466.
[27] P. Ao, H. Qian, Y. Tu, J. Wang, A theory of mesoscopic phenomena: Time scales, emergent unpredictability, symmetry breaking and dynamics across different levels, 2013, arXiv:1310.5585.
[28] R. Balian, From Microphysics to Macrophysics: Methods and Applications of Statistical Physics, Vol. I, Springer, New York, 1991.
[29] D. Frenkel, P.B. Warren, Gibbs, Boltzmann, and negative temperatures, Amer. J. Phys. 83 (2015) 163–170.
[30] R.M. Noyes, Entropy of mixing of interconvertible species: Some reflections on the Gibbs paradox, J. Chem. Phys. 34 (1961) 1983–1985.
[31] J.G. Kirkwood, Statistical mechanics of fluid mixtures, J. Chem. Phys. 3 (1935) 300–313.
[32] A. Ben-Naim, Mixing and assimilation in systems of interaction particles, Amer. J. Phys. 55 (1987) 1105–1109.
[33] J.M.R. Parrondo, J.M. Horowitz, T. Sagawa, Thermodynamics of information, Nat. Phys. 11 (2015) 131–139.
[34] T.M. Hoang, R. Pan, J. Ahn, J. Bang, H.T. Quan, T. Li, Experimental test of the differential fluctuation theorem and a generalized Jarzynski equality for arbitrary initial states, Phys. Rev. Lett. 120 (2018) 080602.
[35] J. Honerkamp, Statistical Physics: An Advanced Approach with Applications, Springer, New York, 2002.
[36] L. Chen, C. Wu, F. Sun, Finite time thermodynamic optimization or entropy generation minimization of energy systems, J. Non-Equilib. Thermodyn. 24 (1999) 327–359.
[37] H. Touchette, The large deviation approach to statistical mechanics, Phys. Rep. 478 (2009) 1–69.
[38] H. Qian, Nonequilibrium potential function of chemically driven single macromolecules via Jarzynski-type log-mean-exponential heat, J. Phys. Chem. B 109 (2005) 23624–23628.
[39] G.A. Maugin, Continuum Mechanics Through the Eighteenth and Nineteenth Centuries: Historical Perspectives from Bernoulli to Hellinger, Springer, New York, 2014, pp. 137–147.
[40] L. Pogliani, M.N. Berberan-Santos, Constantin Carathéodory and the axiomatic thermodynamics, J. Math. Chem. 28 (2000) 313–324.
[41] P. Salamon, B. Andresen, J. Nulton, A.K. Konopka, The mathematical structure of thermodynamics, in: A.K. Konopka (Ed.), Handbook of Systems Biology, CRC Press, Boca Raton, 2006.
[42] R.S. Ellis, Entropy, Large Deviations, and Statistical Mechanics, Springer, New York, 2006.
[43] H. Ge, H. Qian, Mesoscopic kinetic basis of macroscopic chemical thermodynamics: A mathematical theory, Phys. Rev. E 94 (2016) 052150.

Appendix D

TERNARY REPRESENTATION OF STOCHASTIC CHANGE AND THE ORIGIN OF ENTROPY AND ITS FLUCTUATIONS. (WITH QIAN, H. & CHENG, Y.-C.) *ARXIV:1902.09536* (2019)

Ternary Representation of Stochastic Change and the Origin of Entropy and Its Fluctuations

Hong Qian,* Yu-Chen Cheng,† and Lowell F. Thompson‡
Department of Applied Mathematics, University of Washington, Seattle, WA 98195, U.S.A.

A change in a stochastic system has three representations: Probabilistic, statistical, and informational: (i) is based on random variable $u(\omega) \to \tilde{u}(\omega)$; this induces (ii) the probability distributions $F_u(x) \to F_{\tilde{u}}(x)$, $x \in \mathbb{R}^n$; and (iii) a change in the probability measure $\mathbb{P} \to \tilde{\mathbb{P}}$ under the same observable $u(\omega)$. In the *informational representation* a change is quantified by the Radon-Nikodym derivative $\ln\left(\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega)\right) = -\ln\left(\frac{\mathrm{d}F_u}{\mathrm{d}F_{\tilde{u}}}(x)\right)$ when $x = u(\omega)$. Substituting a random variable into its own density function creates a fluctuating entropy whose expectation has been given by Shannon. Informational representation of a deterministic transformation on $\mathbb{R}^n$ reveals entropic and energetic terms, and the notions of configurational entropy of Boltzmann and Gibbs, and potential of mean force of Kirkwood. Mutual information arises for correlated $u(\omega)$ and $\tilde{u}(\omega)$; and a nonequilibrium thermodynamic entropy balance equation is identified.

I. INTRODUCTION

A change according to classical physics is simple: if one measures x_1 and x_2 which are traits, in real numbers, of a "same" type, then $\Delta x = x_2 - x_1$ is the mathematical representation of the change; $\Delta x \in \mathbb{R}^n$. How to characterize a change in a complex world? To represent a complex, stochastic world [1], the theory of probability developed by A. N. Kolmogorov envisions an abstract space $(\Omega, \mathcal{F}, \mathbb{P})$ called a *probability space*. Similar to the Hilbert space underlying quantum mechanics [2], one does not see or touch the objects in the probability space, $\omega \in \Omega$, nor the $\mathbb{P}$. Rather, one observes the probability space through functions, say $u(\omega)$, called random variables which map $\Omega \to \mathbb{R}^n$. The same function maps the probability measure $\mathbb{P}$ to a cumulative probability distribution function (cdf) $F_u(x)$, $x \in \mathbb{R}^n$.

Now a change occurs; and based on observation(s) the $F_u(x)$ is changed to $F_{\tilde{u}}(x)$. In the current statistical data science, one simply works with the two functions $F_u(x)$ and $F_{\tilde{u}}(x)$. In fact, the more complete description in the *statistical representation* is a joint probability distribution $F_{u\tilde{u}}(x_1, x_2)$ whose marginal distributions are $F_u(x_1) = F_{u\tilde{u}}(x_1, \infty)$ and $F_{\tilde{u}}(x_2) = F_{u\tilde{u}}(\infty, x_2)$.

If, however, one explores a little more on the "origin" of the change, one realizes that there are two possible sources: A change in the $\mathbb{P}$, or a change in the $u(\omega)$. If the $\mathbb{P} \to \tilde{\mathbb{P}}$ while the $u(\omega)$ is fixed, then according to the measure theory, one can characterize this "change of measure" in terms of a Radon-Nikodym (RN) derivative $\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega)$ [3–5]. In the rest of this paper, we will assume that all the measures under consideration are absolutely continuous with respect to each other and that all measures on $\mathbb{R}^n$ are absolutely continuous with respect to the Lebesgue measure. This ensures that all RN derivatives are well-defined. Note, this is a mathematical object that is defined on the invisible probability space. It actually is itself a random variable, with expectation, variance, and statistics.

What is the relationship between this *informational representation* of the change and the observed F_u and $F_{\tilde{u}}$? For $u(\omega), \tilde{u}(\omega) \in \mathbb{R}$, the answer is [52]:

$$\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega) = \left[\frac{\mathrm{d}F_u}{\mathrm{d}F_{\tilde{u}}}\Big(u(\omega)\Big)\right]^{-1}. \tag{1}$$

On the rhs of (1), $\frac{\mathrm{d}F_u}{\mathrm{d}F_{\tilde{u}}}(x)$ is like a probability density function, which is only defined on $\mathbb{R}$. However, substituting the random variable $u(\omega)$ into the probability density function, one obtains the lhs of (1). Putting a random variable back into the logarithm of its own density function to create a new random variable is the fundamental idea of fluctuating entropy in stochastic thermodynamics [6, 7], and the notion of self-information [8–10]. Its expected value then becomes the Shannon information entropy or intimately related relative entropy [11]. The result in (1) can be generalized to $u, \tilde{u} \in \mathbb{R}^n$. In this case,

$$\frac{\mathrm{d}F_u}{\mathrm{d}F_{\tilde{u}}}\big(x_1, \cdots, x_n\big) = \frac{\frac{\partial^n F_u}{\partial x_1 \cdots \partial x_n}}{\frac{\partial^n F_{\tilde{u}}}{\partial x_1 \cdots \partial x_n}}. \tag{2}$$

In the rest of the paper, we shall consider the $u(\omega) \in \mathbb{R}$. But the results are generally valid for multidimensional $u(\omega)$.

In this paper, we present key results based on this informational representation of stochastic change. We show all types of entropy are unified under the single theory. The discussions are restricted on very simple cases; we only touch upon the stochastic change with a pair of correlated $u(\omega) \to \tilde{u}(\omega)$, which have respective generated σ-algebras that are non-identical in general. The notion of "thermodynamic work" will appear then [5].

The informational and probabilistic representations of stochastic changes echo the Schrödinger and Heisenberg pictures in quantum dynamics [12]: in terms of wave functions in the abstract, invisible Hilbert space and in terms of self-adjoint operators as observables.

II. INFORMATIONAL REPRESENTATION OF STOCHASTIC CHANGE

Statistics and information: Push-forward and pull-back. Consider a sequence of real-valued random variables of a

* hqian@u.washington.edu
† yuchench@u.washington.edu
‡ lfthomps@u.washington.edu

same physical origin and their individual cdfs:

$$F_{u_1}(x), F_{u_2}(x), \cdots, F_{u_T}(x). \tag{3}$$

According to the axiomatic theory of probability built on $(\Omega, \mathcal{F}, \mathbb{P})$, there is a sequence of random variables

$$u_1(\omega), u_2(\omega), \cdots, u_T(\omega), \tag{4}$$

in which each $u_k(\omega)$ maps the $\mathbb{P}(\omega)$ to the *push-forward measure* $F_k(x)$, $x \in \mathbb{R}$. Eq. 5 illustrate this with u and $\tilde{u}$ stand for any u_i and u_j.

$$\begin{array}{ccc} \mathbb{P}(\omega) & \xrightarrow{u(\omega)} & F_u(x) \\ \Big\downarrow \frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega) = \big[\frac{\mathrm{d}F_u}{\mathrm{d}F_{\tilde{u}}}\big(u(\omega)\big)\big]^{-1} & \searrow \tilde{u}(\omega) & \\ \tilde{\mathbb{P}}(\omega) & \xrightarrow{u(\omega)} & F_{\tilde{u}}(x) \end{array} \tag{5}$$

Informational representation, thus, considers the (3) as the push-forward of a sequence of measures $\mathbb{P}_1 = \mathbb{P}, \mathbb{P}_2, \cdots, \mathbb{P}_T$, under a single observable, say $u_1(\omega)$. This sequence of $\mathbb{P}_k$ can be represented through the fluctuating entropy inside the $[\cdots]$ below:

$$\mathrm{d}\mathbb{P}_k(\omega) = \left[\frac{\mathrm{d}F_{u_1}}{\mathrm{d}F_{u_k}}\big(u_1(\omega)\big)\right]^{-1} \mathrm{d}\mathbb{P}(\omega). \tag{6}$$

Narratives based on information representation have rich varieties. We have discussed above the information representation cast with a single, common random variable $u(\omega)$: $\mathbb{P}_k(\omega) \xrightarrow{u} F_k(x)$. Alternatively, one can cast the information representation with a single, given $F^*(x)$ on $\mathbb{R}$: $\mathbb{P}_k(\omega) \xrightarrow{u_k} F^*(x)$ for any sequence of $u_k(\omega)$. Actually there is a corresponding sequence of measures $\mathbb{P}_k$, whose push-forward are all $F^*(x)$, on the real line independent of k. Then parallel to Eq. 5 we have a schematic:

$$\begin{array}{ccc} \mathbb{P}(\omega) & \xrightarrow{u_k(\omega)} & F_{u_k}(x) \\ \Big\downarrow \frac{\mathrm{d}\mathbb{P}_k}{\mathrm{d}\mathbb{P}}(\omega) = \big[\frac{\mathrm{d}F_{u_k}}{\mathrm{d}F^*}\big(u_k(\omega)\big)\big]^{-1} & \searrow u_1(\omega) & \\ \mathbb{P}_k(\omega) & \xrightarrow{u_k(\omega)} & F^*(x) \end{array} \tag{7}$$

If the invariant $F^*(x)$ is the uniform distribution, *e.g.*, when one chooses the Lebesgue measure on $\mathbb{R}$ with $F^*(x) = x$, then one has

$$-\ \mathbb{E}^{\mathbb{P}}\left[\ln\left(\frac{\mathrm{d}F_{u_k}}{\mathrm{d}x}\big(u_k(\omega)\big)\right)\right] = -\int_{\mathbb{R}} f_{u_k}(x) \ln f_{u_k}(x) \mathrm{d}x.$$

This is precisely the *Shannon entropy*! More generally with a fixed $F^*(x)$, one has

$$\mathbb{E}^{\mathbb{P}}\left[\ln\left(\frac{\mathrm{d}F_{u_k}}{\mathrm{d}F^*}\big(u_k(\omega)\big)\right)\right] = \int_{\mathbb{R}} f_{u_k}(x) \ln\left[\frac{f_{u_k}(x)}{f^*(x)}\right] \mathrm{d}x$$
$$= -\int_{\mathbb{R}} f_{u_k}(x) \ln f^*(x) \mathrm{d}x + \int_{\mathbb{R}} f_{u_k}(x) \ln f_{u_k}(x) \mathrm{d}x. \tag{8}$$

This is exactly the *relative entropy* w.r.t. the stationary $f^*(x)$. In statistical thermodynamics, this is called *free energy*. $-\beta^{-1} \ln f^*(x)$ on the rhs of (8) is called *internal energy*, where β stands for the physical unit of energy. The integral is the mean internal energy.

Essential facts on information. Several key mathematical facts concerning the information, as a random variable defined in (1), are worth stating.

First, even though the $u(\omega)$ appears in the rhs of the equation, the resulting lhs is independent of the $u(\omega)$: It is a random variable created from $\mathbb{P}$, $\tilde{\mathbb{P}}$, and random variable $\tilde{u}(\omega)$, as clearly shown in Eq. 5. This should be compared with a well-known result in elementary probability: For a real-valued random variable $\eta(\omega)$ and its cdf $F_\eta(x)$, the constructed random variable $F_\eta\big(\eta(\omega)\big)$ is a uniform distribution on $[0, 1]$ independent of the nature of $\eta(\omega)$.

Second, if we denote the logarithm of (1) as $\xi(\omega)$, $\xi(\omega) = \ln \frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}$, then one has a result based on the change of measure for integration:

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}\big[\eta(\omega)\big] &= \int_\Omega \eta(\omega) \mathrm{d}\mathbb{P}(\omega) \ = \ \int_\Omega \eta(\omega) \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\tilde{\mathbb{P}}}(\omega)\right) \mathrm{d}\tilde{\mathbb{P}}(\omega) \\ &= \int_\Omega \eta(\omega) e^{-\xi(\omega)} \mathrm{d}\tilde{\mathbb{P}}(\omega) \\ &= \mathbb{E}^{\tilde{\mathbb{P}}}\left[\eta(\omega) e^{-\xi(\omega)}\right]. \end{aligned} \tag{9a}$$

And conversely,

$$\mathbb{E}^{\tilde{\mathbb{P}}}\big[\eta(\omega)\big] = \mathbb{E}^{\mathbb{P}}\left[\eta(\omega) e^{\xi(\omega)}\right]. \tag{9b}$$

In particular, when the $\eta = 1$, the log-mean-exponential of fluctuating ξ is zero. The incarnations of this equality have been discovered numerous times in thermodynamics, such as Zwanzig's free energy perturbation method [13], the Jarzynski-Crooks equality [15, 36], and the Hatano-Sasa equality [16].

Third, one has an inequality,

$$\ln E^{\mathbb{P}}\big[\xi(\omega)\big] \le \ln E^{\mathbb{P}}\left[e^{\xi(\omega)}\right] = 0. \tag{10}$$

As we have discussed in [5], this inequality is the mathematical origin of almost all inequalities in connection to entropy in thermodynamics and information theory.

Fourth, let us again consider a real-valued random variable $\eta(\omega)$, with probability density function $f_\eta(x)$, $x \in \mathbb{R}$, and its information, *e.g.*, fluctuating entropy $\xi(\omega) = -\ln f_\eta\big(\eta(\omega)\big)$ [6]. Then one has a new measure $\tilde{\mathbb{P}}$ whose $\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega) = e^{\xi(\omega)}$, and

$$\begin{aligned} \int_{x_1 < \eta(\omega) \le x_2} \mathrm{d}\tilde{\mathbb{P}}(\omega) &= \int_{x_1 < \eta(\omega) \le x_2} e^{\xi(\omega)} \mathrm{d}\mathbb{P}(\omega) \\ &= \int_{x_1 < \eta(\omega) \le x_2} \big[f_\eta\big(\eta(\omega)\big)\big]^{-1} \mathrm{d}\mathbb{P}(\omega) \\ &= \int_{x_1}^{x_2} \big(f_\eta(y)\big)^{-1} f_\eta(y) \mathrm{d}y \\ &= x_2 - x_1. \end{aligned} \tag{11}$$

This means that under the new measure $\tilde{\mathbb{P}}$, the random variable $\eta(\omega)$ has an unbiased uniform distribution on the $\mathbb{R}$. Note that the measure $\tilde{\mathbb{P}}(\omega)$ is non-normalizable if $\eta(\omega)$ is not a bounded function.

Entropy is the greatest "equalizer" of random variables, as physical observables inevitably biased!

The forgoing discussion leaves no doubt that entropy (or negative free energy) $\xi(\omega)$ is a quantity to be used in the form of $e^{\xi(\omega)}$. This clearly points to the origin of partition function computation in statistical mechanics. In fact, it is fitting to call the tangent space in the affine structure of the space of measures its "space of entropies" [5]; which represents the *change* of information.

III. CONFIGURATIONAL ENTROPY IN CLASSICAL DYNAMICS

By classical dynamics, we mean the representation of dynamical change in terms of a deterministic mathematical description, with either discrete space-time or continuous space-time. The notion of "configurational entropy" arose in this context in the theories of statistical mechanics, developed by L. Boltzmann and J. W. Gibbs, either as Boltzmann's *Wahrscheinlichkeit* W, the Jacobian matrix in a deterministic transformations [17] as a non-normalizable density function, or its cumulative distribution. Boltzmann's entropy emerges in connection to macroscopic observables, which are chosen naturally from the conserved quantities in a microscopic dynamics.

In our present approach, the classical dynamics is a deterministic map in the space of observables, $\mathbb{R}^n$. The informational representation demands a description of the change via a change of measures, and we now show the notion of configurational entropy arises.

Information in deterministic change. Let us now consider a one-to-one deterministic transformation $\mathbb{R} \to \mathbb{R}$, which maps the random variable $u(\omega)$ to $v(\omega) = g^{-1}\big(u(\omega)\big)$, $g'(x) > 0$. Then

$$\frac{\mathrm{d}F_v(x)}{\mathrm{d}F_u(x)} = \left(\frac{f_u(g(x))}{f_u(x)}\right) g'(x). \tag{12}$$

Applying the result in (1) and (5), the corresponding RN derivative

$$-\ln\left[\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega)\right] = \Big[\ln f_u(g(x)) - \ln f_u(x) + \ln g'(x)\Big]_{x=v(\omega)} \tag{13a}$$

$$= \underbrace{-\ln f_u\big(v(\omega)\big) + \ln\left(\frac{\mathrm{d}g}{\mathrm{d}x}\big(v(\omega)\big)\right)}_{\text{information of } v \text{ under observable } u} \tag{13b}$$

$$\underbrace{-\Big(-\ln f_u\big(u(\omega)\big)\Big)}_{\text{information of } u} \tag{13c}$$

in which the information of $v = g^{-1}(u)$, under observable $u(\omega)$, has two distinctly different contributions: *energetic part* and *entropic part*.

The energetic part represents the "changing location", which characterizes movement in the classical dynamic sense: A point to a point. It experiences an "energy" change where the internal energy is defined as $-\beta^{-1}\ln f_u(x) = \varphi(x)$.

The entropic part represents the resolution for measuring information. This is distinctly a feature of dynamics in a continuous space, it is related to the Jacobian matrix in a deterministic transformation: According to the concept of Markov partition developed by Kolmogorov and Sinai for chaotic dynamics [18], there is the possibility of *continuous entropy production* in dynamics with increasing "state space fineness". This term is ultimately related to Kolmogorov-Sinai entropy and Ruelle's folding entropy [19].

Entropy and Jacobian matrix. We note that the "information of v under observable u"

$$-\ln f_u\big(v(\omega)\big) + \ln\left(\frac{\mathrm{d}g}{\mathrm{d}x}\big(v(\omega)\big)\right) \neq -\ln f_v\big(v(\omega)\big)! \tag{14}$$

This difference precisely reflects the effect of "pull-back" from $\mathbb{R}$ to the Ω space, there is a breaking symmetry between $u(\omega)$ and $v(\omega)$ in (13a), when setting $x = v(\omega)$. Actually, the full "information entropy change" associated with the deterministic map

$$-\ln f_v\big(v(\omega)\big) + \ln f_u\big(u(\omega)\big) = -\ln\left(\frac{\mathrm{d}g}{\mathrm{d}x}\big(v(\omega)\big)\right), \tag{15}$$

as expected. The rhs is called configurational entropy. Note this is an equation between three random variables, *i.e., fluctuating entropies*, that is valid for all ω. For a one-to-one map in $\mathbb{R}^n$, the above $|\mathrm{d}g(x)/\mathrm{d}x|$ becomes the absolute value of the determinant of $n \times n$ Jacobian matrix, which has a paramount importance in the theory of functions, integrations, and deterministic transformations. The matrix is associated with an invertible local coordinate transform, $y_i = g_i(\mathbf{x})$, $1 \leq i \leq n$, $\mathbf{x} \in \mathbb{R}^n$:

$$\frac{\mathscr{D}[y_1, \cdots, y_n]}{\mathscr{D}[x_1, \cdots, x_n]} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{bmatrix}, \tag{16}$$

whose determinant is the local "density change". Classical Hamiltonian dynamics preserves the Lebesgue volume.

Measure-preserving transformation. A change with information preservation means the lhs of (13a) being zero for all $\omega \in \Omega$. This implies the function $g(x)$ necessarily satisfies

$$-\ln f_u(x) + \ln\left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right) = -\ln f_u\big(g(x)\big). \tag{17}$$

Eq. 17 is actually the condition for g preserving the measure $F_u(x)$, with density $f_u(x)$, on $\mathbb{R}$ [18]:

$$f_u(x) = f_u(g(x))\left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right) = f_v(x), \tag{18a}$$

$$F_u(x) = \int_{-\infty}^{x} f_u(z)\mathrm{d}z = F_u\big(g(x)\big). \tag{18b}$$

In this case, the inequality in (14) becomes an equality.

In terms of the internal energy $\varphi(x)$, then one has

$$\beta = \frac{1}{\varphi\big(g(x)\big) - \varphi(x)} \ln\left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right), \tag{19}$$

which should be heuristically understood as the ratio $\frac{\Delta S}{\Delta E}$ where $\Delta S = \ln \mathrm{d}g(x) - \ln \mathrm{d}x$ and $\Delta E = \varphi\big(g(x)\big) - \varphi(x)$. For an infinitesimal change $g(x) = x + \varepsilon(x)$, we have

$$\beta = \frac{\varepsilon'(x)}{\varphi'(x)\varepsilon(x)}. \tag{20}$$

Entropy balance equation. The expected value of the lhs of (13a) according to measure $\mathbb{P}$ is non-negative. In fact, consider the Shannon entropy change associated with $F_{\tilde{u}}(x) \to F_u(x)$:

$$\int_{\mathbb{R}} f_{\tilde{u}}(x) \ln f_{\tilde{u}}(x) \mathrm{d}x - \int_{\mathbb{R}} f_u(x) \ln f_u(x) \mathrm{d}x \tag{21a}$$

$$= \underbrace{\int_{\mathbb{R}} f_{\tilde{u}}(x) \ln\left(\frac{f_{\tilde{u}}}{f_u}\right) \mathrm{d}x}_{\text{entropy production } \Delta S^{(\mathrm{i})}} + \underbrace{\int_{\mathbb{R}} \Big[f_{\tilde{u}}(x) - f_u(x)\Big] \ln f_u(x) \mathrm{d}x}_{\text{entropy exchange } \Delta S^{(\mathrm{e})}}$$

$$= \mathbb{E}^{\mathbb{P}}\left[\ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\tilde{\mathbb{P}}}(\omega)\right)\right] \tag{21b}$$

$$+\mathbb{E}^{\mathbb{P}}\left[\ln\left(\frac{\mathrm{d}F_u}{\mathrm{d}x}\right)\big(\tilde{u}(\omega)\big) - \ln\left(\frac{\mathrm{d}F_u}{\mathrm{d}x}\right)\big(u(\omega)\big)\right]. \tag{21c}$$

This equation in fact has the form of the *fundamental equation of nonequilibrium thermodynamics* [20, 21]: $\Delta S = \Delta S^{(\mathrm{i})} + \Delta S^{(\mathrm{e})}$. The entropy production $\Delta S^{(\mathrm{i})}$ on the rhs is never negative, the entropy exchange $\Delta S^{(\mathrm{e})}$ has no definitive sign. If $f_u(x) = \frac{\mathrm{d}F_u(x)}{\mathrm{d}x} = C$ is a uniform distribution, then $\Delta S^{(e)} = 0$ and entropy change is the same as entropy production.

Contracted description, endomorphism, and matrix volume. We have discussed above the $\mathbb{R}^n \to \mathbb{R}^n$ deterministic invertible transformation $\mathbf{y} = g(\mathbf{x})$ and shown that the determinant of its Jacobian matrix is indeed the informational entropy change. The informational representation, in fact naturally, allows us to also consider a non-invertible transformation in the space of observable $u(\omega) \in \mathbb{R}^n$ through a much lower dimensional $g(\mathbf{x}) \in \mathbb{R}^m$, $m \ll n$. In this case, the original $\mathbb{R}^n$ is organized by the m-dimensional observables, called "macroscopic" (thermodynamic) variables in classical physics, or "principle components" in current data science and model reduction [22].

The term *endomorphism* means that a deterministic map $u \to v = g^{-1}(u)$ in (12) is many-to-one. In this case, a simple approach is to divide the domain of u into invertible parts. Actually, in terms of the probability distributions of $g(\mathbf{x}) = (g_1, \cdots, g_m)(\mathbf{x})$, the non-normalizable density function

$$W(y_1, \cdots, y_m) = \frac{\partial^m}{\partial y_1 \cdots \partial y_m} \int_{g_1(\mathbf{x}) \le y_1, \cdots, g_m(\mathbf{x}) \le y_m} \mathrm{d}\mathbf{x}, \tag{22}$$

now plays a crucial role in the information representation. In fact, the W in (22) is the reciprocal of the matrix volume, *e.g.,* the absolute value of the "determinant" of the rectangular matrix [23]

$$\det\left(\frac{\mathscr{D}[y_1, \cdots, y_m]}{\mathscr{D}[x_1, \cdots, x_n]}\right), \tag{23}$$

which can be computed from the product of the singular values of the non-square matrix. Boltzmann's *Wahrscheinlichkeit*, his thermodynamic probability, is when $m = 1$. In that case,

$$W(y) = \frac{\mathrm{d}}{\mathrm{d}y} \int_{g(\mathbf{x}) \le y} \mathrm{d}\mathbf{x} = \oint_{g(\mathbf{x}) = y} \frac{\mathrm{d}\Sigma}{\|\nabla g(\mathbf{x})\|}, \tag{24}$$

in which the integral on rhs is the surface integral on the level surface of $g(\mathbf{x}) = y$.

One has a further, clear physical interpretation of $-\ln W(y_1, \cdots, y_m)$ as a *potential of entropic force* [24], with force:

$$\frac{\partial \ln W}{\partial y_\ell} = \frac{1}{W} \frac{\partial W}{\partial y_\ell}, \ \ (1 \le \ell \le m). \tag{25}$$

In the polymer theory of rubber elasticity, a Gaussian density function emerges due to central limit theorem, and Eq. 25 yields a three-dimensional Hookean linear spring [25].

IV. CONDITIONAL PROBABILITY, MUTUAL INFORMATION AND FLUCTUATION THEOREM

In addition to describing change by $\Delta x = x_2 - x_1$, another more in-depth characterization of a pair of observables (x_1, x_2) is by their functional dependency $x_2 = g(x_1)$, if any. In connection to stochastic change, this leads to the powerful notion of conditional probability, which we now discuss in terms of the informational representation.

First, for two random variables $(u_1, u_2)(\omega)$, the conditional probability distribution on $\mathbb{R} \times \mathbb{R}$:

$$F_{u_1|u_2}(x; y) = \frac{\mathbb{P}\big\{y < u_2(\omega) \le y + \mathrm{d}y, u_1(\omega) \le x\big\}}{\displaystyle\int_{y < u_2(\omega) \le y + \mathrm{d}y} \mathrm{d}\mathbb{P}}. \tag{26}$$

Then it generates an "informational" random variable

$$\xi_{12}(\omega) = \ln\left[\frac{\partial F_{u_1|u_2}(x; y)/\partial x}{\mathrm{d}F_{u_1}(x)/\mathrm{d}x}\right]_{x = u_1(\omega), y = u_2(\omega)}, \tag{27}$$

which is a conditional information, whose expected value is widely know in information theory as *mutual information* [26]:

$$\begin{aligned}
\mathbb{E}\left[\xi_{12}(\omega)\right] &= \int_{\mathbb{R}^2} f_{u_1 u_2}(x, y) \ln\left[\frac{\partial F_{u_1|u_2}(x; y)/\partial x}{\mathrm{d}F_{u_1}(x)/\mathrm{d}x}\right] \mathrm{d}x\mathrm{d}y \\
&= \int_{\mathbb{R}^2} f_{u_1 u_2}(x, y) \ln\left[\frac{f_{u_1|u_2}(x; y)}{f_{u_1}(x)}\right] \mathrm{d}x\mathrm{d}y \\
&= \int_{\mathbb{R}^2} f_{u_1 u_2}(x, y) \ln\left[\frac{f_{u_1, u_2}(x, y)}{f_{u_1}(x) f_{u_2}(y)}\right] \mathrm{d}x\mathrm{d}y \\
&= \mathrm{MI}(u_1, u_2).
\end{aligned} \tag{28}$$

We note that $\xi_{12}(\omega)$ is actually symmetric w.r.t. u_1 and u_2. The term inside the $[\cdots]$ in (27

$$\frac{f_{u_1|u_2}(x;y)}{f_{u_1}(x)} = \frac{f_{u_1,u_2}(x,y)}{f_{u_1}(x)f_{u_2}(y)} = \frac{f_{u_2|u_1}(x;y)}{f_{u_2}(x)}. \tag{29}$$

In fact the equality $\xi_{12}(\omega) = \xi_{21}(\omega)$ is the Bayes' rule. Furthermore, $\mathrm{MI}(u_1, u_2) \geq 0$. It has been transformed into a distance function that satisfies triangle inequality [27, 28].

In statistics, "distance" is not only measured by their dissimilarity, but also their statistical dependence. This realization gives rise to the key notion of *independent and identically distributed* (i.i.d.) random variables. In the informational representation, this means u_1 and u_2 have same amount of information on the probability space, *and* they have zero mutual information.

Finally, but not the least, for $F_{u_1u_2}(x, y)$, one can introduce a $\tilde{F}_{u_1u_2}(x, y) = F_{u_1u_2}(y, x)$. Then entropy production,

$$\xi(\omega) = \ln\left[\frac{\mathrm{d}F_{u_1u_2}}{\mathrm{d}\tilde{F}_{u_1u_2}}(x,y)\right]_{x=u_1(\omega),y=u_2(\omega)} \tag{30a}$$

$$= \ln\left[\frac{\frac{\partial^2 F_{u_1u_2}(x,y)}{\partial x\partial y}}{\frac{\partial^2 F_{u_1u_2}(y,x)}{\partial x\partial y}}\right]_{x=u_1(\omega),y=u_2(\omega)}, \tag{30b}$$

satisfies the fluctuation theorem [35, 37]:

$$\frac{\mathbb{P}\{a < \xi(\omega) \leq a + \mathrm{d}a\}}{\mathbb{P}\{-a - \mathrm{d}a < \xi(\omega) \leq -a\}} = e^a, \tag{31}$$

for any $a \in \mathbb{R}$. Eq. 31 characterizes the statistical asymmetry between $u_1(\omega)$ and $u_2(\omega)$, not independent in general. Being identical and independent is symmetric, so is a stationary Markov process with reversibility [35].

V. CONCLUSION AND DISCUSSION

The theory of information [26] as a discipline that exists outside the field of probability and statistics owes to its singular emphasis on the notion of *entropy* as a quantitative measure of information. Our theory shows that it is indeed a unique *representation* of stochastic change that is distinctly different from, but complementary to, the traditional mathematical theory of probability. In statistical physics, there is a growing awareness of a deep relation between the notion of thermodynamic free energy and information [29]. The present work clearly shows that it is possible to narrate the statistical thermodynamics as a subject of theoretical physics in terms of the measure-theoretic information, as suggested by some scholars who studied deeply thermodynamics and the concept of entropy [30]. Just as differential calculus and Hilbert space providing the necessary language for classical and quantum mechanics, respectively, the Kolmogorovian probability, including the informational representation, provides many known results in statistical physics and chemistry with a deeper understanding, such as phase transition and symmetry breaking [31], Gibbsian ensemble theory [32], nonequilibrium thermodynamics [35–38], and the unification of the theories of chemical kinetics and chemical thermodynamics [39]. In fact, symmetry principle [40] and emergent probability distribution via limit laws [41] can both be understood as providing legitimate measures *a priori*, $\mathbb{P}(\omega)$, for the physical world or the biological world. And stochastic kinematics *dictates* entropic force and its potential function, the free energy [42]. [53]

For a long time the field of information theory and the study of large deviations (LD) in probability were not integrated: A. Ya Khinchin's book was published the same year as the Sanov theorem [43, 44]. Researchers now are agreed upon that Boltzmann's original approach to the canonical equilibrium energy distribution [45], which was based on a maximum entropy argument, is a part of the contraction of Sanov theorem in the LD theory [46]. In probability, the contraction principle emphasizes the LD rate function for the sample mean of a random variable.[54] In statistics and data science, the same mathematics has been used to justify the *maximum entropy principle* which emphasizes a bias, as a conditional probability, introduced by observing a sample mean [47]. In all these work, Shannon's entropy and its variant relative entropy, as a single numerical characteristic of a probability distribution, has a natural and logical role. Many approaches have been further advanced in applications, *e.g.*, surprisal analysis and maximum caliber principle [48, 49].

The idea that information can be itself a stochastic quantity originated in the work of Tribus, Kolmogorov [8, 9], and probably many other mathematically minded researchers [50]. In physics, fluctuating entropy and entropy production arose in the theory of nonequilibrium stochastic thermodynamics. This development has significantly deepened the concept of entropy, both to physics and as the theory of information. The present work further illustrates that the notion of information, together with fluctuating entropy, actually originates from a *perspective* that is rather different from that of strict Kolmogorovian; with complementarity and contradistinctions. In pure mathematics, the notion of change of measures goes back at least to 1940s [51], if not earlier.

[1] Qian, H. (2016) Nonlinear stochastic dynamics of complex systems, I: A chemical reaction kinetic perspective with mesoscopic nonequilibrium thermodynamics. arXiv1605.08070.

[2] Dirac, P. A. M. (1930) *The Principles of Quantum Mechanics*, Oxford Univ. Press, U.K.

[3] Kolmogorov, A. N. and Fomin, S. V. (1968) *Introductory Real Analysis*, Silverman, R. A. transl. Dover, New York.

[4] Ye, F. X.-F. and Qian, H. (2019) Stochastic dynamics II: Finite random dynamical systems, linear representation, and entropy production. *Disc. Cont. Dyn. Sys. B* to appear.

[5] Hong, L., Qian, H. and Thompson, L. F. (2019) Representations

and metrics in the space of probability measures and stochastic thermodynamics. *J. Comp. Appl. Math.* submitted.

[6] Qian, H. (2001) Mesoscopic nonequilibrium thermodynamics of single macromolecules and dynamic entropy-energy compensation. *Phys. Rev. E* **65**, 016102.

[7] Seifert, U. (2005) Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.* **95**, 040602.

[8] Tribus, M. (1961) *Thermostatics and Thermodynamics: An Introduction to Energy, Information and States of Matter, with Engineering Applications.* D. van Nostrand, New York.

[9] Kolmogorov, A. N. (1968) Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, **2**, 157–168.

[10] Qian, H. (2017) Information and entropic force: Physical description of biological cells, chemical reaction kinetics, and information theory (In Chinese). *Sci. Sin. Vitae* **47**, 257–261.

[11] Hobson, A. (1969) A new theorem of information theory. *J. Stat. Phys.* **1**, 383–391.

[12] Louisell, W. H. (1973) *Quantum Statistical Properties of Radiation*, Wiley Interscience, New York.

[13] Zwanzig, R. W. (1954) High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **22**, 1420–1426.

[14] Jarzynski, C. (1997) Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **78**, 2690–2693.

[15] Crooks, G. (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* **60**, 2721–2726.

[16] Hatano, T and Sasa S.-I. (2001) Steady-state thermodynamics of Langevin systems. *Phys. Rev. Lett.* **86**, 3463–3466.

[17] Gibbs, J. W. (1902) *Elementary Principles in Statistical Mechanics*. Scribner's Sons, New York.

[18] Mackey, M. C. (1989) The dynamic origin of increasing entropy. *Rev. Mod. Phys.* **61**, 981–1015.

[19] Ruelle, D. (1996) Positivity of entropy production in nonequilibrium statistical mechanics. *J. Stat. Phys.* **85**, 1–23.

[20] de Groot, S. R. and Mazur, P. (2011) *Non-Equilibrium Thermodynamics*, Dover, New York.

[21] Qian, H., Kjelstrup, S., Kolomeisky, A. B. and Bedeaux, D. (2016) Entropy production in mesoscopic stochastic thermodynamics: Nonequilibrium kinetic cycles driven by chemical potentials, temperatures, and mechanical forces. *J. Phys. Cond. Matt.*, **28**, 153004.

[22] Kutz, J. N. (2013) *Data-Driven Modeling & Scientific Computation*. Oxford Univ. Press, U.K.

[23] Ben-Israel, A. (1999) The change of variables formula using matrix volume. *SIAM J. Matrix Anal.* **21**, 300–312.

[24] Kirkwood, J. G. (1935) Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3**, 300–313.

[25] Hill, T. L. (1987) *Statistical Mechanics: Principles and Selected Applications*, Dover, New York.

[26] Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory*. John Wiley & Sons, New York.

[27] Dawy, Z., Hagenauer, J., Hanus, P. and Mueller, J. C. (2005) Mutual information based distance measures for classification and content recognition with applications to genetics. In *IEEE International Conf. Communication*, vol. 2, pp. 820–824.

[28] Steuer, R., Daub, C. O., Selbig, J. and Kurths, J. (2005) Measuring distances between variables by mutual information. In *Innovations in Classification, Data Science, and Information Systems*. Baier, D. and Wernecke, K. D. eds., Springer, Berlin, pp. 81–90.

[29] Parrondo, J. M. R., Horowitz, J. M. and Sagawa, T. (2015) Thermodynamics of information, *Nature Phys.* **11**, 131–139.

[30] Ben-Naim, A. (2008) *A Farewell to Entropy: Statistical Thermodynamics Based on Information*. World Scientific, Singapore.

[31] Ao, P., Qian, H., Tu, Y., and Wang, J. (2013) A theory of mesoscopic phenomena: Time scales, emergent unpredictability, symmetry breaking and dynamics across different levels. arXiv:1310.5585.

[32] Cheng, Y.-C., Wang, W. and Qian, H. (2018) Gibbsian statistical ensemble theory as an emergent limit law in probability. arXiv:1811.11321.

[33] Jaynes, E. T. (2003) *Probability Theory: The Logic of Science*, Cambridge Univ. Press, U.K.

[34] Wang, J. (2015) Landscape and flux theory of nonequilibrium dynamical systems with application to biology. *Adv. Phys.* **64**, 1–137.

[35] Jiang, D. Q., Qian, M. and Qian, M.P. (2004) *Mathematical Theory of Nonequilibrium Steady States, Lecture Notes in Mathematics*, Vol. 1833. Springer-Verlag, Berlin.

[36] Jarzynski, C. (2011) Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Ann. Rev. Cond. Matt. Phys.* **2**, 329 (2011).

[37] Seifert, U. (2012) Stochastic thermodynamics, fluctuation theorems, and molecular machines. *Rep. Prog. Phys.* **75**, 126001 (2012).

[38] van den Broeck, C. and Esposito, M. (2015) Ensemble and trajectory thermodynamics: A brief introduction. *Physica A* **418**, 6.

[39] Ge, H. and Qian, H. (2016) Mesoscopic kinetic basis of macroscopic chemical thermodynamics: A mathematical theory. *Phys. Rev. E* **94**, 052150.

[40] Yang, C. N. (1996) Symmetry and physics. *Proc. Am. Philos. Soc.* **140**, 267–288.

[41] Anderson, P. W. (1971) More is different. *Science* **177**, 393–396.

[42] Qian, H. (2017) Kinematic basis of emergent energetic descriptions of general stochastic dynamics. arXiv:1704.01828.

[43] Khinchin, A. Ya. (1957) *Mathematical Foundations of Information Theory*, Dover, New York.

[44] Sanov, I. N. (1957) On the probability of large deviations of random variables. *Matematicheskii Sbornik* **42**, 11–44.

[45] Sharp, K. and Matschinsky, F. (2015) Translation of Ludwig Boltzmann's paper "On the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium". *Entropy* **17**, 1971–2009.

[46] Touchette, H. (2009) The large deviation approach to statistical mechanics. *Phys. Rep.* **478**, 1–69.

[47] Shore, J. E. and Johnson, R. W. (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Tran. Info. Th.* **IT-26**, 26–37.

[48] Levine, R. D. (1978) Information theory approach to molecular reaction dynamics. *Annu. Rev. Phys. Chem.* **29**, 59–92.

[49] Pressé, S., Ghosh, K., Lee, J. and Dill, K. A. (2013) Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **85**, 1115–1141.

[50] Downarowicz, T. (2011) *Entropy in Dynamical Systems*, Cambridge Univ. Press, U.K.

[51] Cameron, R. H. and Martin, W. T. (1944) Transformations of Wiener integrals under translations. *Ann. Math.* **45**, 386–396.

[52] $\tilde{F}_u(x) =$
$\int_{u(\omega)\le x} \mathrm{d}\tilde{\mathbb{P}}(\omega) \;=\; \int_{u(\omega)\le x} \frac{\mathrm{d}F_{\bar{u}}}{\mathrm{d}F_u}\big(u(\omega)\big)\mathrm{d}\mathbb{P}(\omega)$
$= \int_{-\infty}^{x} \frac{\left[\frac{\mathrm{d}F_{\bar{u}}(u)}{\mathrm{d}u}\right]}{\left[\frac{\mathrm{d}F_u(u)}{\mathrm{d}u}\right]} f_u(u)\mathrm{d}u \;=\; \int_{-\infty}^{x} \left[\frac{\mathrm{d}F_{\bar{u}}(u)}{\mathrm{d}u}\right]\mathrm{d}u \;=\; F_{\bar{u}}(x).$
The equality in (1) is restricted on all the $B \in \sigma\big(u^{-1}(\mathcal{B})\big) \subseteq \mathcal{F}$.

[53] There is a remarkable logic consistency between fundamental physics pursuing symmetries in appropriate spaces under appropriate group actions and the notion of invariant measures in a theory of stochastic dynamics as the legitimate prior probability [33, 34].

[54] Even though the mathematical steps in Boltzmann's work on ideal gas are precisely these in the LD contraction computation, their interpretations are quite different: Total mechanical energy conservation figured prominently in Boltzmann's work, which is replaced by the conditioning on a given sample mean in the LD theory. The subtle relation between *conservation law* and *conditional probability* was discussed in [32].