Reversibility, Heat Dissipation, and the Importance of the Thermal Environment in Stochastic Models of Nonequilibrium Steady States

R. A. Blythe

SUPA, School of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom

(Received 30 July 2007; published 8 January 2008)

We examine stochastic processes that are used to model nonequilibrium processes (e.g., pulling RNA or dragging colloids) and so deliberately violate detailed balance. We argue that by combining an information-theoretic measure of irreversibility with nonequilibrium work theorems, the thermal physics implied by abstract dynamics can be determined. This measure is bounded above by thermodynamic entropy production and so may quantify how well a stochastic dynamics models reality. We also use our findings to critique various modeling approaches and notions arising in steady-state thermodynamics.

DOI: 10.1103/PhysRevLett.100.010601 PACS numbers: 05.40.−a, 05.70.Ln, 65.40.G−, 89.70.−a

A theory of nonequilibrium physics is vital if we are to understand such diverse phenomena as geological or biological processes which are inherently dissipative in nature. Although a general theory remains both challenging and elusive, it is now possible to obtain precise experimental data for mesoscopic objects such as RNA strands [1] and optically trapped colloids [2] undergoing irreversible manipulation. In turn this has allowed theoretical developments, such as strikingly general nonequilibrium work relations, to be verified [3].

In this work, we address the fundamental question of how to faithfully model irreversible, dissipative physics with stochastic dynamics. We introduce an irreversibility measure for stochastic processes which, in contrast to standard expressions, respects such basic physics as frame invariance. We find that an explicit prescription for a system's thermal environment—often absent in models—is essential if predictions for entropy production are even to be possible. Using work relations for stochastic systems [4] we find our main result, inequality (4) below, which shows that such predictions always underestimate the true dissipation, unless all relevant processes are modeled. This suggests that a model predicting less dissipation than is observed is incomplete, and one predicting more should be rejected. Since our results hold for arbitrary nonequilibrium states, we gain many insights into theories and models of nonequilibrium steady states (NESS) [5] which cannot be drawn from, for example, a similar expression recently derived for isolated systems constrained initially to be at equilibrium [6].

We begin by reviewing the modeling paradigm introduced by Katz, Lebowitz, and Spohn in their seminal work on fast ionic conductors [7]. One takes the master equation for a (discrete-time) process, $P_{t+1}(\mathcal{C}) = \sum_{\mathcal{C}'} P_t(\mathcal{C}')M(\mathcal{C}' \to \mathcal{C})$, in which $P_t(\mathcal{C})$ is the time-dependent distribution of microstates $\mathcal{C}$. For a reversible, equilibrium system, the transition probabilities $M(\mathcal{C} \to \mathcal{C}')$ are taken to satisfy the detailed balance condition

$$P^*(\mathcal{C})M(\mathcal{C} \to \mathcal{C}') = P^*(\mathcal{C}')M(\mathcal{C}' \to \mathcal{C}) \qquad (1)$$

with respect to the Boltzmann distribution $P^*(\mathcal{C}) = e^{-\beta E(\mathcal{C})}$ where β is inverse temperature and E the internal energy. This relation guarantees microscopic reversibility [8,9] in the steady state, i.e., that any sequence of configurations is witnessed with the same probability as its time reversal. To model irreversible physics, and, in particular, a NESS, one must deliberately violate detailed balance.

There is no obviously correct way to go about this, so following [7] it is commonplace to invoke a local (or generalized [10]) detailed balance principle. In this approach, (1) is taken to apply over some closed subset of configurations, and a nonequilibrium system formed by joining together subsystems that are in contact with heat baths at different temperatures. For illustrative purposes, we take the specific example of hard-core particles in a one-dimensional linear potential, which, if connected to particle reservoirs at different densities, would exhibit the biased diffusion shown in Fig. 1(a). Alternatively, periodic boundary conditions might be imposed [Fig. 1(b)], at the expense of being able to couch the dynamics in terms of a single-valued potential. Note, however, this modeling procedure can be used for all types of particle interaction and in any dimension.

A problem with this approach is that one loses sight of how the system interacts thermally and mechanically with its environment, and could thus be argued to lack a firm physical basis. Furthermore, it is not obvious that alternative approaches, e.g., those based on maximal-entropy

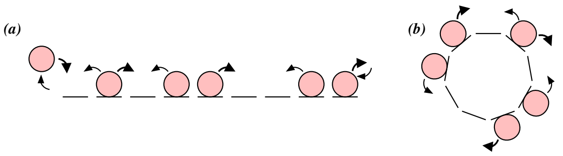

FIG. 1 (color online). Biased diffusion arising from the local or generalized detailed balance principles applied to hard-core particles in a one-dimensional linear potential gradient under (a) open and (b) periodic boundary conditions.

analyses subject to macroscopic flux constraints [11,12], offer more realistic descriptions of nonequilibrium physics than the model-building tradition described above. We address these shortcomings by introducing a framework in which a model system's thermal environment is made completely explicit, which, as we now show, is necessary to establish the degree to which a stochastic dynamics is irreversible.

The standard way to do this is to compare the left- and right-hand sides of the detailed balance condition (1). For example, logarithms of their ratio appear in a widely used expression for the entropy production attributed to Schnackenberg [13], an action functional that exhibits a Gallavotti-Cohen symmetry [14] and various time-dependent generalizations [4,15], as well as the housekeeping heat [16] that is instantaneously dissipated by a NESS [17]. Their differences, meanwhile, have been proposed to characterize a NESS [18], since instantaneous physical currents (which vanish at equilibrium) can be derived from them. The violation of (1) is thus almost universally used to recognize a dynamics with a dissipative steady state, despite the obvious shortcoming that an observer can witness the former without the latter simply by changing frame.

This difficulty is resolved by realizing that when comparing the probability of a trajectory with that of its time reversal, the latter should not be drawn from the same ensemble as the former, but from an ensemble in which all degrees of freedom in the environment are also time reversed. The dynamics that generate this second ensemble we shall call the reverse process. We may now define the following general measure of reversibility for any stochastic dynamics, i.e., not restricting ourselves to ergodic time-homogeneous Markov chains in discrete time (see also Fig. 2). Let $\mathcal{X}$ denote a trajectory $(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n)$ that visits configuration $\mathcal{C}_i$ at time t_i, possibly other (unspecified) configurations at other times and eventually reaches, with probability $P_T(\mathcal{C}_T)$, configuration $\mathcal{C}_T$ at a time $T > t_n$. Given an initial configuration $\mathcal{C}_0$ that is drawn from a distribution $P_0(\mathcal{C}_0)$, this trajectory is taken to appear under the forward dynamics with probability $P(\mathcal{X}|\mathcal{C}_0)$. This is to be compared with the probability of seeing the time-reversed trajectory $\hat{\mathcal{X}}$, in which the image $\hat{\mathcal{C}}_i$ of $\mathcal{C}_i$ under time reversal (i.e., with all velocities reversed) is seen at time t_i running backwards from time T to 0. Given a starting configuration $\hat{\mathcal{C}}_T$ drawn from a distribution $\hat{P}_T(\hat{\mathcal{C}}_T)$, this reverse trajectory appears with probability $\hat{P}(\hat{\mathcal{X}}|\hat{\mathcal{C}}_T)$. If there is to be any possibility for the ensembles of forward and reverse trajectories to coincide, we must take $\hat{P}_T(\hat{\mathcal{C}}_T) = P_T(\mathcal{C}_T)$, i.e., start the reverse process by immediate time reversal of configurations reached after time T under the forward dynamics. Any other choice requires us to make additional assumptions on the dynamics.

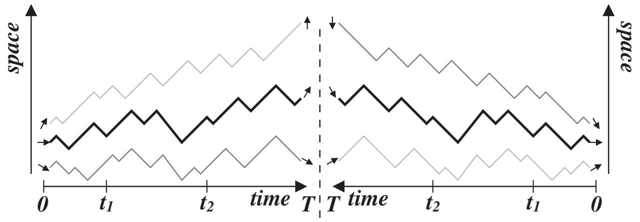

FIG. 2. Comparison of trajectories generated by the forward and reverse processes. Forward trajectories of length T are generated, at which point all velocities (shown as short arrows) are flipped, and the reverse dynamics started. The heaviness of the lines indicates the probability of the trajectories in each ensemble. The central trajectory appears with the same probability in both ensembles, whereas the outer trajectories appear with different probabilities, thus indicating an irreversible dynamics.

In the spirit of Landauer's principle [19], we now loosely associate information lost under the dynamics—quantified here by the additional information required to reconstruct the forward trajectory ensemble from the reverse—with irreversibility and dissipation. This amount of information (in natural units) is given by relative entropy of the two ensembles [20],

$$\Delta I = \sum_{\mathcal{C}_0, \mathcal{X}, \mathcal{C}_T} P_0(\mathcal{C}_0) P(\mathcal{X}|\mathcal{C}_0) \ln \frac{P_0(\mathcal{C}_0) P(\mathcal{X}|\mathcal{C}_0)}{P_T(\mathcal{C}_T) \hat{P}(\hat{\mathcal{X}}|\hat{\mathcal{C}}_T)}. \qquad (2)$$

To make contact with thermal physics, we assume that just before the start of the forward and reverse processes any heat baths present are manipulated by a thermostat in such a way that the probability that any particular bath configuration is realized is given by the Boltzmann distribution with a well-defined temperature. Note that this necessarily requires correlations between the system and bath to vanish rapidly—this is the origin of dissipation, as will be seen concretely below. Under such conditions, Jarzynski's detailed fluctuation relation [21]

$$\ln \frac{P(\mathcal{X}, \Delta S_{\text{env}}|\mathcal{C}_0)}{\hat{P}(\hat{\mathcal{X}}, -\Delta S_{\text{env}}|\hat{\mathcal{C}}_T)} = \Delta S_{\text{env}} \qquad (3)$$

applies (in a system of units where Boltzmann's constant is unity). Here, ΔS_{env} is the total entropy increase in the heat baths under the forward dynamics. This is a random variable if the trajectory $\mathcal{X}$ contains insufficient detail to determine how much energy has been exchanged with each heat bath separately. In deriving this formula, it was assumed that the microscopic evolution is Hamiltonian with respect to potentials that are time independent in the baths but may exhibit time dependence in the system of interest. Stochasticity enters from the Boltzmann sampling of bath configurations and any coarse graining in the specification of the trajectory $\mathcal{X}$.

We finally arrive at an important inequality—the main result of this work—by averaging over ΔS_{env}. An application of the log-sum inequality $\sum_i a_i \ln(a_i/b_i) \geq \sum_i a_i \ln(\sum_i a_i / \sum_i b_i)$ (which itself is a consequence of

Jensen's inequality [20]) leads to

$$0 \leq \Delta I \leq \langle \Delta S_{\text{env}} \rangle + S_G(T) - S_G(0), \tag{4}$$

in which $S_G(t) = -\sum_{\mathcal{C}} P_t(\mathcal{C}) \ln P_t(\mathcal{C})$, the Gibbs entropy of the distribution at time t under the forward dynamics. While a similar result was recently given for isolated systems starting at equilibrium [6], our result (4) holds for any initial condition and explicitly requires the system to be open to the environment. Moreover, the thermostatting of the baths means that ΔS_{env} is the true entropy production, which is not always true of isolated systems [22]. As we now discuss, (4) thus provides hitherto unavailable information—spatial and temporal—about heat production in a general nonequilibrium system, e.g., a NESS.

For example, the lower bound is attained only if every forward trajectory appears with the same probability as its time reversal in the reverse ensemble (i.e., no information loss occurs and the process is reversible). This leads to an extended detailed balance relation for a NESS, viz.,

$$P^*(\mathcal{C})M(\mathcal{C} \to \mathcal{C}') = P^*(\mathcal{C}')\hat{M}(\hat{\mathcal{C}}' \to \hat{\mathcal{C}}). \tag{5}$$

Note that one cannot decide on the reversibility of a dynamics until its reversal $\hat{M}(\hat{\mathcal{C}}' \to \hat{\mathcal{C}})$ has been identified (see below for concrete examples). Since equality of forward and reverse trajectory sets implies $P^*(\mathcal{C}) = \hat{P}^*(\hat{\mathcal{C}})$, one finds (5) can be written in a more symmetric form with $\hat{P}^*(\hat{\mathcal{C}}')$ on the right-hand side. This condition is equivalent to (5), as can be shown from conservation of probability $\sum_{\mathcal{C}'} M(\mathcal{C} \to \mathcal{C}') = 1$. The condition (5) can also be stated as a Kolmogorov criterion [8]

$$\begin{aligned} &M(\mathcal{C}_1 \to \mathcal{C}_2)M(\mathcal{C}_2 \to \mathcal{C}_3)\cdots M(\mathcal{C}_T \to \mathcal{C}_1) \\ &\quad = \hat{M}(\hat{\mathcal{C}}_1 \to \hat{\mathcal{C}}_T)\hat{M}(\hat{\mathcal{C}}_T \to \hat{\mathcal{C}}_{T-1})\cdots \hat{M}(\hat{\mathcal{C}}_2 \to \hat{\mathcal{C}}_1) \end{aligned} \tag{6}$$

on every loop in configuration space of length $T \geq 1$. This allows reversibility to be decided without prior knowledge of the stationary distributions $P^*(\mathcal{C})$ or $\hat{P}^*(\hat{\mathcal{C}})$. Equivalence of (5) and (6) is shown in a similar way to the standard case [8].

The upper bound in (4) is reached only if the stochastic dynamics faithfully models all dissipative processes in the physical system. This we have already seen from the fact that if one cannot work out from the trajectory $\mathcal{X}$ how much energy has been exchanged with each bath, ΔS_{env} in (3) is a random variable and ΔI underestimates the true entropy change. As for isolated systems starting at equilibrium [6], the log-sum inequality implies that ΔI further decreases under spatial coarse graining. Since in the present, more general context, ΔI contains temporal information, coarse graining in time, or reduction of a non-Markov dynamics to a Markov process, has the same effect. We thus suggest that this reduction in ΔI could reveal the amount of heat dissipated at the finer-grained scale, and that differences between a model's prediction for entropy production and that measured in a real system might allow deficiencies in the model to be identified.

Finally, we use relation (4) to gain new insights into the stochastic modeling approaches described earlier. The physics of the open system [constructed using generalized detailed balance [10] and as illustrated here in Fig. 1(a)] is consistent with that described above. In particular, the interpretation of $\ln M(\mathcal{C} \to \mathcal{C}')/M(\mathcal{C}' \to \mathcal{C})$ as being proportional to the energy exchanged with a reservoir holds, as long as one is confident that all dissipative processes are captured by the Markov dynamics of particles hopping on a lattice, and further that one can unambiguously identify which bath exchanges energy at any given transition in the system of interest. Note that since particle velocities are not included in the model $\hat{\mathcal{C}} \equiv \mathcal{C}$; also $\hat{M} \equiv M$ as the potential is time independent. We also see explicitly that dissipation results from a continuous thermostatting of the reservoirs that enables particles to enter or leave the system with a constant probability in every time step.

Models in which a current is induced by periodic boundary conditions [see Fig. 1(b)] are more subtle. There are at least two ways in which such dynamics may be realized in a manner consistent with (3). First, one can apply a change of frame to unbiased diffusion on a ring, and then discretize: clearly, this yields a reversible dynamics. Alternatively, one can fashion a time-homogeneous Markov process by coarse graining the response to a rotating potential over one period of its motion. For concreteness, and to keep track of all energy fluxes, we consider a dynamics in which the energy function $E_t(\mathcal{C})$ is static during each time step, and changed instantaneously between them. As in [23], the dynamics is assumed to satisfy (1) while the potential is static. One can compute transition probabilities over the course of one period of rotation of a potential (e.g., a square well) in either direction, and show that typically the forward and reverse dynamics, M and $\hat{M}$, are not simply related to each other, nor do they satisfy (6) [24]. Because of the coarse graining, the irreversibility measured by ΔI underestimates the true dissipation in the system.

We remark that in our framework, coarse graining generically leads to (and is in fact the only mechanism for) the appearance of nonconservative forcing in the system of interest. By contrast, such forces are central to models based on Langevin equations (see, e.g., [15,17,25]), are associated with the dissipation of housekeeping heat [16], and have been argued to differ fundamentally from those due to a moving potential. Since coarse graining blurs this distinction, it is not clear in what sense it is meaningful. We also remark that the nonconservative forces considered in [4,15,17] are assumed not to change under time reversal, which even in simple models is not the case for forces due to a moving potential. As well as this, it seems often to have been assumed that trajectories are sufficiently detailed that the upper bound in (4) is in fact an equality, and further that housekeeping heat can be defined on a per-trajectory basis in terms of the instantaneous state of a system and its environment [4,17]. Such a definition conflicts with the macroscopic quantity described in [16] if the latter is interpreted as the heat exported by some sequence of dissipative steady states if one could somehow switch between them without incurring additional entropy costs. For these costs to be removed when averaging over all microscopic realizations of an arbitrary switching process, one finds that details of the history of this process must appear in the single-trajectory expressions, contrary to [4,17]. We thus contend that far greater clarity about the meaning of central quantities in the putative framework of steady-state thermodynamics [16] is necessary.

Finally, we examine modeling approaches in which transition probabilities are obtained from maximal-entropy inference subject to macroscopic flux constraints [12]. If this is to be interpreted as a general recipe for deriving a stochastic dynamics, then we have shown the need to derive both the forward and the reverse dynamics, the latter obtained from time reversal of all driving forces, using this procedure. If all macroscopic fluxes simply change sign under time reversal, the outcome will be a reversible dynamics, and so—at least within the framework put forward here—one needs to argue for time-asymmetric macroscopic constraints to realize a dissipative dynamics. However, the theory developed in [12] is intended to apply to internal portions of a larger sheared system, and as such are in contact with nonequilibrium reservoirs, not the thermostatted heat baths described here. It would be interesting to try and interpret our definition of reversibility in this more general context.

In summary, we have argued, by examining what it means for a stochastic process to be reversible, that the presence of dissipation in a model steady state can only be decided once a reverse process, which demands knowledge of the environment, is known. Using Jarzynski's detailed fluctuation theorem (3) and results from information theory, we have specified a physical environment that allows information loss to be bounded above by the thermodynamic entropy production, extending to a much larger class of nonequilibrium systems a result of [6]. This allows physical mechanisms by which the system is driven and heat dissipated away to be identified in otherwise abstract models of a NESS, illustrating with the particular examples shown in Fig. 1. We note that although the standard detailed balance condition (1) is satisfied in all these models, only in some is the steady state actually dissipative. Although we have couched our discussion in terms of discrete-time Markov processes, everything we have said also applies in the continuous-time limit.

While we have mostly taken a theoretical perspective, we hope that the main result (4) will be useful experimentally, e.g., to determine whether a stochastic model captures all relevant dissipative processes, as we have proposed. The hypothesis that decreases in ΔI under coarse graining relate to dissipation at a given scale could also be tested explicitly. Finally, we see, from the difficulty in discriminating between nonconservative forces and those due to coarse graining a moving potential, for example, that in the field of nonequilibrium statistical mechanics conceptual problems remain.

I thank Mike Evans, Andy Jackson, and Wilson Poon for useful comments and the Royal Society of Edinburgh for financial support.

[1] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, Science **296**, 1832 (2002).

[2] G. M. Wang, E. M. Sevick, E. Mittag, D. J. Searles, and D. J. Evans, Phys. Rev. Lett. **89**, 050601 (2002).

[3] C. Bustamante, J. Liphardt, and F. Ritort, Phys. Today **58**, No. 7, 43 (2005).

[4] R. J. Harris and G. M. Schütz, J. Stat. Mech. (2007) P07020.

[5] See B. Schmittmann and R. K. P. Zia, *Statistical Mechanics of Driven Diffusive Systems*, Phase Transitions and Critical Phenomena Vol. 17 (Academic, London, 1995); G. M. Schütz, *Exactly Solvable Models for Many-Body Systems Far From Equilibrium*, Phase Transitions and Critical Phenomena Vol. 19 (Academic, London, 2000); R. B. Stinchcombe, Adv. Phys. **50**, 431 (2001); M. R. Evans and T. Hanney, J. Phys. A **38**, R195 (2005); R. A. Blythe and M. R. Evans, J. Phys. A **40**, R333 (2007) for reviews.

[6] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, Phys. Rev. Lett. **98**, 080602 (2007).

[7] S. Katz, J. L. Lebowitz, and H. Spohn, Phys. Rev. B **28**, 1655 (1983).

[8] F. P. Kelly, *Reversibility and Stochastic Networks* (Wiley, Chichester, 1979).

[9] M. F. Chen, *From Markov Chains to Non-Equilibrium Particle Systems* (World Scientific, Singapore, 1992); N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1992).

[10] T. Bodineau and B. Derrida, C.R. Physique **8**, 540 (2007).

[11] R. Dewar, J. Phys. A **36**, 631 (2003).

[12] R. M. L. Evans, Phys. Rev. Lett. **92**, 150601 (2004); R. M. L. Evans, J. Phys. A **38**, 293 (2005).

[13] J. Schnakenberg, Rev. Mod. Phys. **48**, 571 (1976).

[14] J. L. Lebowitz and H. Spohn, J. Stat. Phys. **95**, 333 (1999).

[15] U. Seifert, Phys. Rev. Lett. **95**, 040602 (2005).

[16] Y. Oono and M. Paniconi, Prog. Theor. Phys. Suppl. **130**, 29 (1998).

[17] T. Hatano and S. Sasa, Phys. Rev. Lett. **86**, 3463 (2001); T. Speck and U. Seifert, J. Phys. A **38**, L581 (2005).

[18] B. Schmittmann and R. K. P. Zia, J. Phys. A **39**, L407 (2006).

[19] R. Landauer, IBM J. Res. Dev. **44**, 261 (2000).

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, Hoboken, NJ, 2006).

[21] C. Jarzynski, J. Stat. Phys. **98**, 77 (2000).

[22] B. Palmieri and D. Ronis, Phys. Rev. E **75**, 011133 (2007).

[23] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[24] R. A. Blythe (unpublished).

[25] J. Kurchan, J. Phys. A **31**, 3719 (1998).