Chapter 2
Dissipation and Kullback–Leibler Divergence

In this chapter, we introduce the theoretical framework of the first part of our work, in which we study of the relationship between dissipation and irreversibility quantitatively in microscopic systems in the stationary state.

The relationship between entropy production (dissipation) and irreversibility forms the core of thermodynamics and statistical mechanics. The first studies in dissipation and irreversibility in nonequilibrium processes were done in the context of linear irreversible thermodynamics [16]. In linear regime, the entropy production is linear with the force that drives the system out of equilibrium. However, this relation holds only in linear regime and does not provide a quantitative description of entropy production in terms of the microscopic properties of the system.

With the introduction of fluctuation theorems (see Sect. 1.5), it is possible to derive exact relationships that connect the entropy production of a microscopic system in the NESS with its microscopic properties and, moreover, provide a quantitative tool to measure the time irreversibility of the process. Our work is devoted to clarify this relation and to provide tools to estimate time irreversibility in the NESS from a single stationary trajectory [18, 24, 25]. In this chapter, we describe the theoretical framework of our approach, whereas the estimation techniques and applications to simulations and experimental data are described in further chapters.

This chapter is organized as follows: In Sect. 2.1 we define the notion of time irreversibility in stochastic processes. In Sect. 2.2 we review the notion of dissipation in irreversible processes from the approach of linear irreversible thermodynamics to the new insights provided by fluctuation theorems. In Sect. 2.3 we introduce the concept of relative entropy or Kullback–Leibler divergence and show how it is related to the arrow of time. In Sect. 2.4 we show how dissipation and time irreversibility are quantitatively connected in the NESS. In Sect. 2.5 we study discrete systems and obtain exact expressions for the Kullback–Leibler divergence and the dissipation for two specific stochastic processes.

É. Roldán, *Irreversibility and Dissipation in Microscopic Systems*,
Springer Theses, DOI: 10.1007/978-3-319-07079-7_2,

2.1 What Is Time Irreversibility?

Time irreversibility is the ability to distinguish between a process and its time reversal. In irreversible processes, one can ascertain from the observation of the process if the time is running forward or backwards. The fingerprint of an irreversible processes is therefore the ability to guess the direction of the *arrow of time*.

In the macroscopic world, time irreversibility occurs in different scenarios. One example is a magnetic hysteresis cycle. In such a cycle, the system does not recover its original demagnetized state after a periodic change of the external field. Another example is a glass falling to the ground and smashing into pieces, where the time-reversal process is never observed.

In the microscopic world, the arrow of time is blurred because of thermal fluctuations. Guessing if a process is time reversible or not from a single realization of the process is challenging: an irreversible process can look time reversible when sampled at different frequencies or using insufficient statistics. However, because of the fast relaxation times in the microscale, one can measure the probability to observe a path or its time reversal from the statistics of different trials. This probability can be used to *quantify* the time irreversibility of a process.

To quantify time irreversibility in the microscopic world one therefore needs a metric to compare the probability distributions of forward and backward trajectories. Such metric is the Kullback–Leibler divergence or relative entropy. In this chapter we show how one can quantify the time irreversibility using the Kullback–Leibler divergence and how this quantification is related to the dissipation of the process.

2.2 Average Dissipation in Irreversible Processes

In nonequilibrium processes, an irreversible process is accompanied by a positive entropy production. In linear regime, entropy production in macroscopic systems is a bilinear quadratic form on the macroscopic flows or currents of the systems.

At small scales, an analogous relation between entropy production and currents is found for systems obeying an overdamped equation as shown in Sect. 1.4.3. Using novel results derived in the context of fluctuation theorems, it is possible to measure the entropy production of microscopic systems that are driven arbitrarily far from equilibrium under an arbitrary external protocol.

2.2.1 Linear Irreversible Thermodynamics

Linear irreversible thermodynamics studies systems that are driven not far from equilibrium in the context of linear response theory [16]. Close to equilibrium, entropy production can be expressed as a linear combination of all the different

thermodynamic forces or gradients F_i that are exerted on the system. In this limit, the entropy production per unit volume, σ, equals to

$$\sigma = \sum_i F_i J_i, \tag{2.1}$$

where i runs over all the different forces on the system and J_i is the flux associated to the force F_i. One example is the heat flux J_Q, which is produced by a force that is proportional to the gradient of temperature, $F_Q \propto \nabla \frac{1}{T}$. Another example is the electric current $J_e = I$ that is driven by an electric field E, $F_e \propto E$. In linear regime, the forces are proportional to the fluxes

$$F_j = \sum_k L_{jk} J_k, \tag{2.2}$$

where L_{jk} are the *phenomenological coefficients*. The above relation expresses that, for example, it is possible to induce a heat flow from an electric current, or vice versa. Taking into account (2.2), entropy production per unit volume equals to

$$\sigma = \sum_{i,k} L_{ik} J_i J_k. \tag{2.3}$$

We notice that in linear response, entropy production is a positively defined quadratic form of the currents and it is therefore related to the presence of macroscopically observable flows, and since if $J > 0$, then $\sigma > 0$. This formulation however does not connect the entropy production with the microscopic properties of the system. Using fluctuation theorems, it is possible to obtain a formula that expresses the entropy production for microscopic systems driven arbitrarily far from equilibrium and do a connection between the work dissipated and the microscopic properties of the system.

2.2.2 Entropy Production in Microscopic Systems

As we showed in Sect. 1.4.3, the definition of the entropy associated to a trajectory of a microscopic system allows one to introduce the notion of entropy production in the case of a Brownian particle obeying an overdamped Langevin equation. The following expression relating the ensemble average of the entropy production and the probability flux was derived by Seifert [26],

$$\langle \dot{S}_{\text{prod}}(t) \rangle = k \int dx \frac{j(x,t)^2}{D\rho(x,t)} \geq 0. \tag{2.4}$$

Notice that this expresion was also obtained for Brownian chemical motors described by Langevin equation [21]. We notice that (2.4) expresses a relationship between

entropy production and the probability current, $j(x,t)$. According to (2.4), the entropy production a Brownian particle vanishes if $j(x,t)$ does, which means that irreversibility for these kind of systems is revealed in the flows of the system. Moreover, the entropy production depends on the current as j^2, which is in accordance with linear response theory, as shown in (2.3).

2.2.3 KPB Theorem

Recently, a quantitative relationship between dissipation and irreversibility in non-equilibrium processes for microscopic systems has been derived. The introduction of fluctuation theorems has allowed to express the entropy production of an isolated microscopic system in terms of the microscopic properties of the system [11, 15, 22]. The main result was derived by Kawai et al. [15] and it is known in literature as the KPB (Kawai, Parrondo and van den Broeck) theorem, which we now discuss.

In Refs. [15, 22], the dissipation of microscopic isolated systems that are are brought from an initial equilibrium state at temperature T to a final equilibrium state at the same temperature is investigated. An expression relating the average dissipation (or entropy production) in such processes with the distinguishability between the process and its time reversal is found.

Consider an isolated physical system described by a Hamiltonian $H(q,p;\lambda)$, where (q,p) denotes a point in phase space, and λ is a parameter of the system controlled by an external agent. The system is initially in a canonical equilibrium state at temperature T. Then the system is disconnected from the thermal bath and driven out of equilibrium according to a protocol in which the external agent modifies λ from $\lambda(0)=\lambda_A$ to $\lambda(\tau)=\lambda_B$, following a protocol $\{\lambda(t)\}_{t=0}^{\tau}$. At $t=\tau$ the system is weakly coupled to a thermal bath and relaxes to a canonical state at temperature T.

The dissipation of the process described above (which we call *forward* process) is related to the distinguishability between phase space densities of the forward process and its time reversal (or *backward* process). In the backward process, the system is initially in a canonical equilibrium state at temperature T and driven by the time-reversed protocol $\{\tilde{\lambda}(t)\}_{t=0}^{\tau}=\{\lambda(\tau-t)\}_{t=0}^{\tau}$. For a given trajectory in the forward process that starts in $(q_0,p_0;0)$ and ends in $(q_1,p_1;\tau)$, the corresponding time reversed trajectory is obtained by reversing the position and changing the sign of the momenta, i.e., the time reversed trajectory starts in $(q_1,-p_1;\tau)$ and ends in $(q_0,-p_0;0)$ (see Fig. 2.1). Notice that the time t is taken in the forward process.

The KPB theorem relates the average dissipation in the forward process, $\langle W_{\rm diss}\rangle$, with the distinguishability between the forward and the backward phase space densities measured at the same but arbitrary time during the process, $t\in[0,\tau]$,

$$\langle W_{\rm diss}\rangle=\langle W\rangle-\Delta F=kT\int dqdp\;\rho(q,p;t)\ln\frac{\rho(q,p;t)}{\widetilde{\rho}(q,-p;t)}. \tag{2.5}$$

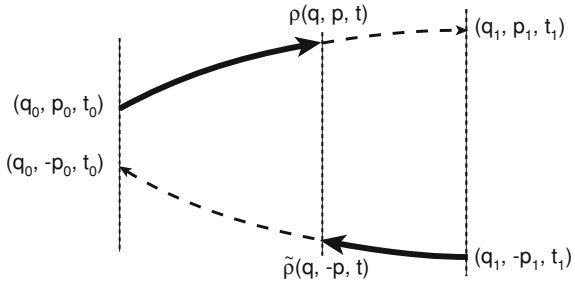

Fig. 2.1 Forward and backward trajectories in phase space. Picture taken from [15]

This result shows that the dissipation of a nonequilibrium process is revealed in the phase space. The right hand side is often called relative entropy or *Kullback–Leibler divergence* (KLD) [4, 17] between the probability distributions $\rho(q, p; t)$ and $\widetilde{\rho}(q, -p; t)$, which is denoted by the letter D,

$$\langle W_{\text{diss}} \rangle = kT \; D[\rho(q, p; t)||\widetilde{\rho}(q, -p; t)]. \tag{2.6}$$

The right hand side of (2.6) measures the difficulty to distinguish whether the microstate of the system at any time is generated from the forward or the backward experiment. Equation (2.6) links directly the average dissipation with the time irreversibility of the process, which can be *quantified* using the KLD. As we will show in Sect. 2.3, the value of the KLD increases when the two probability distributions are more different each other, indicating that the more different are the forward and reverse process the more work is dissipated. The KLD is positive, which ensures that the second law of thermodynamic holds in average, $\langle W_{\text{diss}} \rangle = kT \; D[\rho||\widetilde{\rho}] \geq 0$. The dissipation of microscopic systems is revealed in the phase space and one can get tighter bounds to the dissipation than in the macroscopic case, where $\langle W_{\text{diss}} \rangle \geq 0$.

The KPB theorem can be extended to more general initial (equilibrium) conditions. In Ref. [22] it is proved that the change of the entropy in the system plus the bath averaged over many realizations of the process is equal to

$$\langle S_{\text{prod}} \rangle = k \; D[\rho(q, p; t)||\tilde{\rho}(q, -p; t)]. \tag{2.7}$$

Equation (2.7) is valid for a variety of initial equilibrium conditions, as shown in Ref. [22]: canonical, multi-canonical (several uncoupled systems at different temperatures), and grand-canonical distributions, as well as for different types of baths equilibrating the system at the end of the process. In all these cases, the formula holds when the evolution is isolated and the control parameter follows any arbitrary nonequilibrium protocol. In particular, for canonical initial conditions in the forward and in the backward processes, both at the same temperature T, entropy production equals the average dissipated work $\langle W_{\text{diss}} \rangle = \langle W \rangle - \Delta F$ divided by the temperature T and (2.7) becomes (2.6), $\langle W_{\text{diss}} \rangle = kT \; D[\rho(q, p; t)||\widetilde{\rho}(q, -p; t)]$.

We now reproduce the proof of Eq. (2.7) for the case where both initial conditions of the forward and backward processes are canonical at temperature T. We use the

notation $z = (q, p)$ to denote all the variables in phase space, $\widetilde{z} = (q, -p)$ being the corresponding microstate obtained by changing the sign of all the momenta. We first consider that the KLD $D[\rho(z;t)||\widetilde{\rho}(\widetilde{z};t)]$ can be rewritten as

$$D[\rho(z;t)||\widetilde{\rho}(\widetilde{z};t)] = \int dz\ \rho(z;t)\ln\rho(z;t) - \int dz\ \rho(z;t)\ln\widetilde{\rho}(\widetilde{z};t). \tag{2.8}$$

Secondly, since the system is isolated during its evolution, the phase space density evolves according to Liouville's equation [cf. Eq. (1.28)]

$$\frac{\partial\rho(z;t)}{\partial t} = \mathcal{L}\rho(z;t). \tag{2.9}$$

In addition, the backward phase space density $\widetilde{\rho}(\widetilde{z};t)$ obeys the same Liouville equation considering the derivative with respect to the forward time t [22],

$$\frac{\partial\widetilde{\rho}(\widetilde{z};t)}{\partial t} = \mathcal{L}\widetilde{\rho}(\widetilde{z};t). \tag{2.10}$$

The fact that both ρ and $\widetilde{\rho}$ obey the same Liouville equation implies that the two terms in Eq. (2.8) are invariant in time, which can be proved by using both Liouville's equations and partial integration. A proof of the time invariance of the first term in Eq. (2.8) can be found in Appendix B.1, and the proof for the invariance of the second term is analogous. Consequently, we express (2.8) by evaluating the two terms at any time t, in particular,

$$D[\rho(z;t)||\widetilde{\rho}(\widetilde{z};t)] = \int dz\ \rho(z;0)\ln\rho(z;0) - \int dz\ \rho(z;\tau)\ln\widetilde{\rho}(\widetilde{z};\tau). \tag{2.11}$$

The first term in (2.11) corresponds to (minus) the system entropy in the beginning of the process, in k units. The second term in (2.11) can be interpreted as the system entropy at the end of the process plus the change in the entropy of the environment, as we shall see. We notice the difference between $\rho(z;\tau)$, which is the phase space density at the end of the forward process, which is not at equilibrium, and $\widetilde{\rho}(\widetilde{z};\tau)$, which is the initial (equilibrium) distribution of the backward process.

Let us consider a process where a system that is initially in contact with a thermal bath at temperature T is disconnected from the bath and driven out of equilibrium by an external agent following a protocol $\{\lambda(t)\}_{t=0}^{\tau}$ from $\lambda(0) = \lambda_A$ to $\lambda(\tau) = \lambda_B$. After the process, the system is put in contact with a thermal bath at temperature T' and let relax to an equilibrium state. We assume that during the nonequilibrium driving, the Hamiltonian of the system is $H(z,\lambda)$. In this case, the initial distributions of the forward and backward processes are

$$\rho(z;0) = \frac{e^{-\beta H(z;\lambda_A)}}{Z(T,\lambda_A)}, \quad \widetilde{\rho}(\widetilde{z};\tau) = \frac{e^{-\beta H(\widetilde{z};\lambda_B)}}{Z(T',\lambda_B)}, \tag{2.12}$$

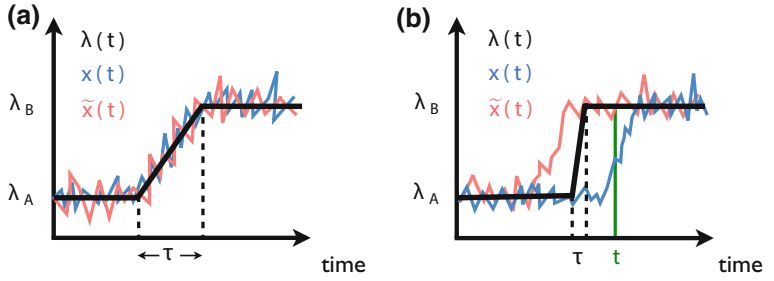

Fig. 2.2 KPB theorem in an example. A system described by a single degree of freedom x is driven by an external agent following a protocol in which a control parameter is changed linearly from $\lambda(0) = \lambda_A$ to $\lambda(\tau) = \lambda_B$. **a** Protocol (*black line*), and the value of x in the forward (*blue*) and backward process (*red*) when the total time of the process is much larger to the relaxation time of the system $\tau \gg \tau_r$. The dynamics is reversible and the phase space densities of forward and backward processes coincide at any time. **b** Same graphs when the time of the process is much smaller than the relaxation time of the system, $\tau \ll \tau_r$. The dynamics is irreversible and the phase space densities of forward and backward processes do not coincide at every time. In this case, at time t indicated in *green* in the figure, $\rho(x,t) \neq \widetilde{\rho}(x,t)$ and we can measure the average dissipation of the forward process with $\langle W_{\rm diss}\rangle = kT\, D[\rho(x,t)||\widetilde{\rho}(x,t)]$

where $Z(t,\lambda) = \int dz e^{-H(z;\lambda)/kT}$. By replacing the above distributions in (2.11) and taking into account that $F(T,\lambda) = -kT \ln Z(T,\lambda)$,

$$k\, D[\rho(z;t)||\widetilde{\rho}(\widetilde{z};t)] = -S(0) + \frac{\langle H\rangle_\tau - F(T',\lambda_B)}{T'}. \tag{2.13}$$

Here $\langle H\rangle_\tau$ is an average over $\rho(z,\tau)$, which does not coincide with the average energy on the initial (equilibrium) state of the backward process, $\langle H\rangle_{{\rm eq},\tau}$. On the other hand, when the system relaxes to equilibrium after it is connected to the thermal bath at temperature T' at time $t = \tau$, the system transfers a heat to the environment $Q_{\rm env} = \langle H\rangle_\tau - \langle H\rangle_{{\rm eq},\tau}$. Taking into account this, Eq. (2.13) can be rewritten as

$$\begin{aligned} k\, D[\rho(z;t)||\widetilde{\rho}(\widetilde{z};t)] &= -S(0) + \frac{\langle H\rangle_{{\rm eq},\tau} - F(T',\lambda_B)}{T'} + \frac{Q_{\rm env}}{T'} \\ &= -S(0) + S(\tau) + \Delta S_{\rm bath} = S_{\rm prod}, \end{aligned} \tag{2.14}$$

which proves Eq. (2.7) for the case of initial and final canonical equilibrium states at any temperatures T and T'. In particular, for the case of initial and final conditions at the same temperature T, we recover the expession relating work dissipation and irreversibility (2.6).

In Fig. 2.2 we show an illustrative example of the applicability of KPB theorem. A system is initially in equilibrium at temperature T with an externally-controlled control parameter λ fixed at value λ_A. Then the system is disconnected from the thermal bath and the control parameter is kept constant and equal to λ_A for a time longer than the relaxation time of the system, τ_r. Then the control parameter is changed from

λ_A to λ_B linearly during a time τ in isolated conditions. In the end of the process, the control parameter is held fixed at λ_B for a time longer than τ_r and the system is let to relax by putting the system in contact with a thermal bath of temperature T. If the relaxation time of the system τ_r is very small compared to τ, $\tau_r \ll \tau$, the system relaxes to the equilibrium state at any time t and the process is done reversibly. In this case, the forward and reverse trajectories are indistinguishable and $\rho(x;t) \simeq \widetilde{\rho}(x;t)$ at any time t along the process. If the process is done much faster than the characteristic time scale of the system, $\tau_r \gg \tau$ then the system does not relax tho the equilibrium state during the process as shown in Fig. 2.2b. In this case, forward and backward trajectories are distinguishable, and in general $\rho(x;t) \neq \widetilde{\rho}(x;t)$. The KLD between these two distributions is an estimation of the dissipation of the process.

We remark that the formulas (2.7) and (2.6) connecting dissipation and irreversibility in the microscopic scale were first proved for the case of systems driven out of equilibrium in isolated conditions. The relationship can be extended to any system immersed in a thermal bath at temperature T, described by the overdamped Langevin equation. When such a system is in contact with a thermal bath at temperature T, the system plus the bath can be viewed as an isolated "super system". The KPB theorem implies that the average dissipation of the system plus the bath equals to $\langle W_{\rm diss}\rangle = kT\ D(\rho||\widetilde{\rho})$ where $D(\rho||\widetilde{\rho})$ is calculated in the full phase space (system plus bath) [13]. This relation was tested experimentally using a Brownian particle dragged by an optical tweezer at constant speed and with an electric circuit with an imposed mean current [1, 3]. Notice that both the position of the Brownian particle and the charge inside the resistor of the circuit obey an overdamped Langevin equation with different physical parameters [1].

The KPB theorem can be generalized to other stochastic processes in the microscopic scale using previous results obtained in the framework of fluctuation theorems—which are valid not only in isolated conditions—such as Crooks's theorem [6]. The connection between fluctuation theorems and the KPB theorem, as well as its generalization to other stochastic processes such as nonequilibrium steady states was done in [11]. The relation between irreversibility and entropy production in a generic nonequilibrium steady state is of particular interest in this work, and it is discussed in the next sections.

2.3 Kullback–Leibler Divergence and Irreversibility

The Kullback–Leibler divergence (KLD), or relative entropy, measures the distinguishability of two probability distributions or two random processes [4]. Let us consider a random variable X and let p and q be two different probability distributions of the random variable X. We denote by $p(x)$ and $q(x)$ the probability of the variable X to take the value x when it is distributed according to p and q, respectively. The KLD between the probability distributions p and q is defined as [17]

$$D[p(x)||q(x)] = \int dx\ p(x) \ln \frac{p(x)}{q(x)}. \tag{2.15}$$

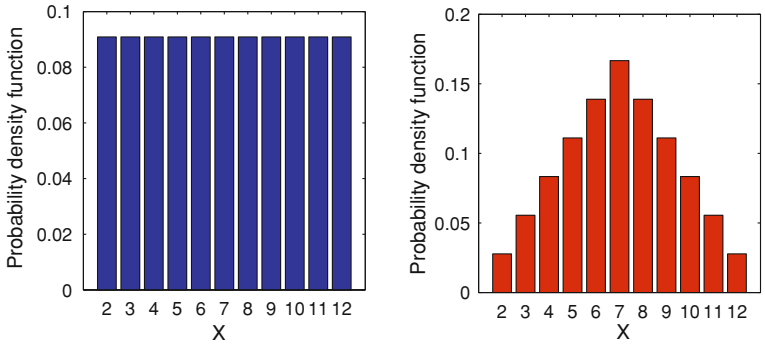

Fig. 2.3 Dice and lottery example of the KLD. *Left* probability density function of a random variable X that can take the values $2, 3, \ldots, 12$ with equal probability as if it were drawn from a lottery. *Right* probability distribution of a random variable whose value is obtained as the sum of the outcome of two dice. One can measure the KLD between these discrete distributions $D(p||q) = \sum_i p_i \ln(p_i/q_i)$, where i runs from 2 to 12. The KLD between the two distributions is not symmetric, since $D(p_{\text{dice}}(x)||p_{\text{lottery}}(x)) = 0.128$ and $D(p_{\text{lottery}}(x)||p_{\text{dice}}(x)) = 0.152$

The KLD is always positive, $D[p(x)||q(x)] \geq 0$, and vanishes if and only if $p(x) = q(x)$ for all x. Therefore, D follows two of the main properties of a mathematical *distance*. However, D is not symmetric with respect to a change of arguments, $D[p(x)||q(x)] \neq D[q(x)||p(x)]$ as it can be seen in the simple example of Fig. 2.3.

The interpretation of the KLD as a measure of distinguishability is a consequence of *Chernoff-Stein lemma* [4]: the probability of incorrectly guessing (via hypothesis testing) that a sequence of n data is distributed according to p when the true distribution is q is asymptotically equal to $e^{-nD[p(x)||q(x)]}$. Therefore, when p and q are similar—in the sense that they overlap significantly—the likelihood of incorrectly guessing the distribution, p or q, is large [4]. In the example of Fig. 2.3, $D(p_{\text{dice}}||p_{\text{lottery}}) < D(p_{\text{lottery}}||p_{\text{dice}})$, meaning that it is easier to incorrectly guess—or equivalently harder to distinguish between distributions—that a sequence is generated from the dice, when true origin is the lottery. In other words, the smaller is the KLD, the more similar are the two distributions and it is harder to distinguish between them using hypothesis testing.

Chernoff-Stein lemma implies that the KLD in (2.6) and (2.7) can be considered as a measure of the arrow of time, since it measures the difficulty to distinguish whether the state (q, p) of the system at time t was generated in the forward on backward experiment [22]. Mathematically speaking, the relative entropy in KPB theorem (2.6),

$$D[\rho(q, p; t)||\widetilde{\rho}(q, -p; t)] = \int dqdp\ \rho(q, p; t) \ln \frac{\rho(q, p; t)}{\widetilde{\rho}(q, -p; t)}, \tag{2.16}$$

has to be understood as a relative entropy between two probability distributions of the *same* random variables (q, p) at time t: $\rho(q, p; t)$, and a second distribution $\sigma(q, p; t)$.

$$D[\rho(q, p; t)||\sigma(q, p; t)] = \int dqdp\ \rho(q, p; t) \ln \frac{\rho(q, p; t)}{\sigma(q, p; t)}, \tag{2.17}$$

such that the value of $\sigma(q, p; t)$ for (q, p) satisfies $\sigma(q, p; t) = \widetilde{\rho}(q, -p; t)$.

Let us recall a property of the KLD that we use throughout our work [4]. If we have two random variables X, Y and two joint probability distributions $p(x, y)$ and $q(x, y)$, then the *chain rule* holds (see Appendix B.2),

$$\begin{aligned} D[p(x, y)||q(x, y)] &= D[p(x)||q(x)] + D[p(y|x)||q(y|x)] \\ &\geq D[p(x)||q(x)]. \end{aligned} \tag{2.18}$$

The inequality in Eq. (2.18) implies that it is harder to distinguish between p and q when we consider only the marginal distributions, $p(x)$ and $q(x)$, instead of the full joint distributions, $p(x, y)$ and $q(x, y)$. If X, Y describe the state of a physical system, Eq. (2.18) indicates that the KLD is smaller when calculated using a partial description of the system, given by the variable X, than when using full information (X and Y). The bound in (2.18) is an equality when the variable Y carries redundant information with respect to the variable X, for example, when Y is obtained as a function of X, $Y = f(X)$ for any function f.

When not all the degrees of freedom of the system can be sampled, we say that *partial information* of the physical system is available. Let x be *any* collection of m position and n momenta of the system $x = (q_1, \ldots, q_m; p_1, \ldots, p_n)$, where m and n can be different and $3n+3m$ is smaller than the total number of degrees of freedom of the system. Since x describes in general only a part of the physical system we say that x contains *partial information* of the system. Because of the chain rule, when the state of the system is described with partial information given by x, the KPB theorem (2.6) turns into an inequality

$$\langle W_{\text{diss}} \rangle \geq kT\ D[\rho(x; t)||\widetilde{\rho}(\tilde{x}; t)], \tag{2.19}$$

where $\tilde{x} = (q_1, \ldots, q_m; -p_1, \ldots, -p_n)$. We notice that even ignoring the full information of phase space (2.19) $D[\rho(x; t)||\widetilde{\rho}(\tilde{x}; t)]$ still gives at least a lower bound to the average dissipation that is in accordance with the second law of thermodynamics, $\langle W_{\text{diss}} \rangle \geq kT\ D[\rho(x; t)||\widetilde{\rho}(\tilde{x}; t)] \geq 0$. Consequently, when the system is described using only a reduced set of variables of the phase space, the KLD in (2.19) provides a lower bound to the dissipation.[1]

[1] However, a recent work shows that, if the neglected information contains an external driving, the entropy production estimated in the coarse grained system can be bigger than the real entropy production [8].

A particular set of *redundant* variables do not provide any statistical information about the direction of the arrow of time of the process, and the value of the KLD remains the same if these variables are not sampled [11, 24, 25]. There are two groups of variables that provide *redundant* information to measure the irreversibility with the KLD. First, variables which are time reversible whose distribution is the same in the forward and backward processes. Second, system variables that are obtained as a function of other variables.

Now suppose that at every time t, we cannot measure the microstate $(q, p; t)$ of the system but we can only detect that the system is in a specific subset of the phase space. In this situation, the state of the system can be described using a *coarse-grained* random variable X that indicates in which subset the system is at every time t. For example, the position of a Brownian particle in one dimension x can be coarse-grained by introducing a new variable α that indicates if the value of the position is positive or negative, for example $\alpha = 0$ if $x \leq 0$ and $\alpha = 1$ if $x > 0$. If we phase space is partitioned in K non overlapping subsets $\{\mathcal{X}_j\}_{j=1}^K$ the coarse-grained forward and backward phase space densities are

$$\rho_j(t) = \int_{\mathcal{X}_j} dqdp\ \rho(q, p; t); \quad \widetilde{\rho}_j(t) = \int_{\widetilde{\mathcal{X}}_j} dqdp\ \widetilde{\rho}(q, -p; t), \tag{2.20}$$

where $\widetilde{\rho}_j$ is identical to ρ_j except a change of sign in all the momenta. These distributions measure the probability of the system to be in the region j of the phase space. At an arbitrary time t during the nonequilibrium process, the KLD between the forward and backward coarse-grained distributions is

$$D[\rho(t)||\widetilde{\rho}(t)] = \sum_{j=1}^{K} \rho_j(t) \ln \frac{\rho_j(t)}{\widetilde{\rho}_j(t)}, \tag{2.21}$$

Since the coarse-grained description of the state of the system is a partial description of its microstate, the KLD in Eq. (2.21) bound from below the KLD using full information of the phase space, by virtue of the chain rule. Therefore using a coarse-grained description of the microstate of the system, we can only bound from below the average dissipation [15],

$$\langle W_{\text{diss}} \rangle \geq kT\ D[\rho(t)||\widetilde{\rho}(t)]. \tag{2.22}$$

The chain rule allows one to rewrite the KPB theorem using the KLD between the forward and reverse distributions of *trajectories* in phase space. For isolated systems, the evolution is deterministic, except for the last stage where the system is connected to the bath, and the point $z = (q, p)$ at time t determines the whole trajectory of the system $\{z(t)\}_{t=0}^{\tau}$. Then $z(t)$ and $\{z(t)\}_{t=0}^{\tau}$ carry the same information and the KLD of their respective probability densities are equal by virtue of the chain rule. Equation (2.6) can be rewritten in terms of *path probabilities*. Let $\mathcal{P}(\{z(t)\}_{t=0}^{\tau})$ be the probability to observe a trajectory $\{z(t)\}_{t=0}^{\tau} = \{q(t), p(t); t\}_{t=0}^{\tau}$

in the forward process. The corresponding time-reversed path in the backward process starts in $(q(\tau), -p(\tau); \tau)$ and ends in $(q(0), -p(0); 0)$ as shown in Fig. 2.1. In a more compact notation, the time-reversed trajectory is defined as $\{\tilde{z}(\tau - t)\}_{t=0}^{\tau} = \{q(\tau - t), -p(\tau - t); \tau - t\}_{t=0}^{\tau}$. The probability to observe such a trajectory in the backward process is denoted by $\widetilde{\mathcal{P}}(\{\tilde{z}(\tau - t)\}_{t=0}^{\tau})$. The KLD between forward and backward trajectory distributions equals to the KLD between phase space densities at every time t during the process,

$$D[\rho(z;t)||\widetilde{\rho}(\tilde{z};t)] = D[\mathcal{P}(\{z(t)\}_{t=0}^{\tau})||\widetilde{\mathcal{P}}(\{\tilde{z}(\tau - t)\}_{t=0}^{\tau})]. \tag{2.23}$$

Because of this, the dissipation in isolated microscopic systems can be expressed in terms of the distinguishability between the trajectory distributions of forward and backward processes. Equation (2.6) can be rewritten as [11, 24, 25]

$$\langle W_{\rm diss}\rangle = kT\ D[\mathcal{P}(\{z(t)\}_{t=0}^{\tau})||\widetilde{\mathcal{P}}(\{\tilde{z}(\tau - t)\}_{t=0}^{\tau})]. \tag{2.24}$$

The above KLD has to be understood as a KLD between two distributions of a stochastic process [cf. (B.18) in Appendix B.2]. As noticed before in (2.17), the right hand side in (2.24) is a KLD between $\mathcal{P}$ and a different trajectory distribution $\mathcal{Q}$,

$$\langle W_{\rm diss}\rangle = kT \int \mathcal{D}(\{z(t)\}_{t=0}^{\tau})\mathcal{P}(\{z(t)\}_{t=0}^{\tau}) \ln \frac{\mathcal{P}(\{z(t)\}_{t=0}^{\tau})}{\mathcal{Q}(\{z(t)\}_{t=0}^{\tau})}, \tag{2.25}$$

where $\mathcal{Q}$ is such that $\mathcal{Q}(\{z(t)\}_{t=0}^{\tau}) = \mathcal{P}(\{\tilde{z}(\tau - t)\}_{t=0}^{\tau})$.

We now recall that using Crooks's fluctuation theorem, we can arrive to an expression of the average dissipative work in terms of the forward and backward work distributions. Integrating Crooks's theorem (1.70), $W - \Delta F = \ln \frac{\rho(W)}{\widetilde{\rho}(-W)}$, where $\rho(W)$ $[\widetilde{\rho}(W)]$ is the probability density of the work done on the system along the forward (backward) process [6, 11], one immediately gets

$$\langle W_{\rm diss}\rangle = kT\ D[\rho(W)||\widetilde{\rho}(-W)]. \tag{2.26}$$

Notice that the work W is a functional of the trajectory $\{z(t)\}_{t=0}^{\tau}$ [see (1.59)] containing less information than the trajectory itself. As indicated by the chain rule (2.18), the KLD of work distributions should in principle be smaller than the KLD of trajectory distributions, and therefore bound the dissipation from below. On the contrary, the KLD is the same, indicating that all the irreversibility of the process is captured by the dissipative work [11]. Equation (2.26) indicates that the irreversibility in the work reveals the entropy production.

If the microstate of the system is not known are every time t but only a trajectory containing the evolution of a subset of variables of the phase space, $\{x(t)\}_{t=0}^{\tau}$, the KLD between forward and backward trajectories yields a lower bound to the dissipation,

$$\langle W_{\rm diss}\rangle \geq kT\ D[\mathcal{P}(\{x(t)\}_{t=0}^{\tau})||\widetilde{\mathcal{P}}(\{\tilde{x}(\tau - t)\}_{t=0}^{\tau})]. \tag{2.27}$$

The bound saturates when the variables that are sampled capture the same information about the irreversibility as the work. The information of the irreversibility is contained in the variables that interact with the work-performing device which are the *footprints* of irreversibility [11].

When the state of the system cannot be sampled at every time t along the process but only every finite time interval, Δt, the sampled trajectory contains less information than the path of the system along the process (which coincides with the limit $\Delta t \to 0$). In this case, even if the microstate of the system $(q(t), p(t); t)$ is known every Δt, the KLD gives a lower bound to the dissipation. In Ref. [10], Gomez-Marin et al. studied the dynamics of an overdamped Brownian particle that moves in one dimension dragged by a harmonic potential that moves at constant velocity. They measured the KLD between trajectories of the position of the particle $\{x(t + i\Delta t)\}_{i=0}^{\tau/\Delta t}$ in the forward process and the backward process, the latter consisting in moving the trap at the opposite velocity as in the forward process. When reducing the sampling time Δt, the KLD approaches asymptotically to the entropy production in the system.

In summary, when considering the KLD between forward and backward distributions of trajectories, the irreversibility is partially captured by the KLD when one or more of the following phenomena occur:

- **Coarse graining**: Only a coarse-grained description of the system is available, that is, we only know in which of an ensemble of phase space subsets the system is.
- **Partial information**: Not all the variables of phase space can be sampled but only a subset of variables in phase space.
- **Finite time sampling**: The trajectory cannot be sampled at any time but only at a finite sampling frequency.

When the information about the system is not full in the sense that one of the three above mentioned shortcomings occur, the KLD between forward and backward trajectories is a lower bound to the average entropy production in the system. In Fig. 2.4 we show an illustrative example where the three types of lack of information can occur.

In the majority of experimental contexts, only system variables are measurable, neglecting the variables of the bath and therefore using partial information of the system. In this case, the KLD yields a lower bound to the dissipation $\langle W_{\text{diss}} \rangle \geq kT \; D(\rho_F || \rho_B)$. In [13], this result is illustrated with an overdamped Brownian particle in a dragged harmonic trap immersed in a thermal bath at temperature T. It is shown that in isothermal conditions, a lower bound to the dissipation is obtained using the KLD $\langle W_{\text{diss}} \rangle \geq kT \; D(\rho_F || \rho_B)$. The bound is tighter when decreasing the friction coefficient γ, and therefore the when the coupling between the Brownian particle and the reservoir is weaker. When $\gamma \to 0$, the system is uncoupled to the bath and therefore isolated, and the equality is met.

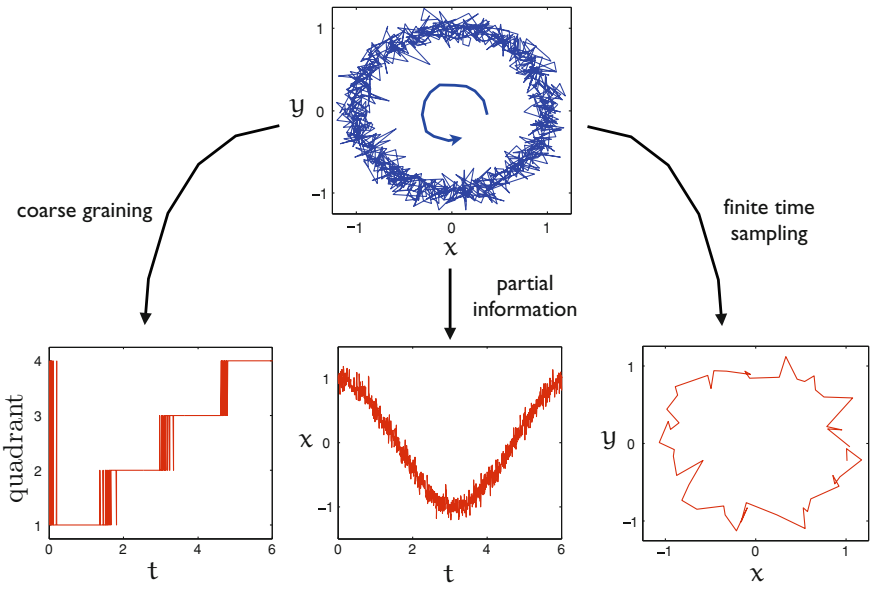

Fig. 2.4 Illustrative example of possible information shortcomings. In the *top panel*, we show the trajectory of a microscopic system that reaches a limit cycle in the xy plane. In the *bottom panels* we show the trajectory of the system when using partial descriptions of the microstate of the system: measuring the quadrant in which the system is at any time (*bottom left panel*), measuring only the variable x at any time (*bottom center panel*) and sampling the trajectory in xy plane every 20 data (*bottom right panel*)

2.4 Dissipation and Irreversibility in the Nonequilibrium Stationary State

The main goal of our work is to explore the existing quantitative relation between dissipation and irreversibility for microscopic systems for the case of nonequilibrium processes that reach a nonequilibrium stationary state (NESS). As we revised in Sect. 1.5.3, a fluid under a constant shear, a gas in a piston that is moved sinusoidally or a molecular motor driven by the ATP hydrolysis are only a few examples of nonequilibrium processes that reach a NESS.

We now proceed to apply the above results to stationary trajectories. Consider a long process where a microscopic system reaches a nonequilibrium stationary state (NESS) after a possible initial transient. In the NESS, the external parameter is held fixed, $\lambda(t) = \lambda$ or it is time-symmetric; the system is kept out of equilibrium due to the existence of baths at different temperatures (a possibility that is included in the hypothesis used in [22] to prove (2.7)) or different chemical potentials, external constant forces, etc. In the steady state, the protocol and its time reversal are identical $\lambda(t) = \tilde{\lambda}(t) = \lambda$. For simplicity, we will denote by $\mathcal{P}$ the trajectory probability

density in the stationary state. In the long time limit, $\tau \to \infty$, we can neglect the contribution of the transient to the entropy production and rewrite (2.7) for the entropy production per unit of time $\dot{S}_{\rm prod}$ in the NESS [20] as

$$\langle \dot{S}_{\rm prod} \rangle = k \lim_{\tau\to\infty} \frac{1}{\tau} D\left[\mathcal{P}\left(\{z(t)\}_{t=0}^{\tau}\right) \middle| \middle| \mathcal{P}\left(\{\tilde{z}(\tau - t)\}_{t=0}^{\tau}\right)\right]. \tag{2.28}$$

As we showed in Sect. 1.5.3, a similar expression can be obtained from the steady state fluctuation theorem, which holds in the long time limit (1.75),

$$\langle \dot{S}_{\rm prod} \rangle = k \lim_{\tau\to\infty} \frac{1}{\tau} \left\langle \ln \frac{\rho_\tau(S)}{\rho_\tau(-S)} \right\rangle, \tag{2.29}$$

$$= k \lim_{\tau\to\infty} \frac{1}{\tau} D[\rho_\tau(S)||\rho_\tau(-S)]. \tag{2.30}$$

where $p_\tau(S)$ is the probability to observe an entropy production $S_{\rm prod} = S$ in the interval $[0, \tau]$. Notice that the average in (2.29) is done over all possible values of the entropy production by averaging with $\rho_\tau(S)$, which yields the KLD in (2.30). Comparing (2.28) and (2.30) we arrive at

$$D\left[\mathcal{P}\left(\{z(t)\}_{t=0}^{\tau}\right) \middle| \middle| \mathcal{P}\left(\{\tilde{z}(\tau - t)\}_{t=0}^{\tau}\right)\right] = D[\rho_\tau(S)||\rho_\tau(-S)], \tag{2.31}$$

for $\tau \to \infty$. Consequently, although S is another observable that is obtained as a function of the microstate of the system, the KLD calculated with S yields the same value as the one calculated with full information of the system. Therefore entropy production captures all the information about the time irreversibility of the NESS.

When one does not observe the entire microscopic trajectory $\{z(t)\}_{t=0}^{\tau}$ in (2.28) but the trajectory followed by one or several observables of the system $x(t)$, the KLD only provides a lower bound to the entropy production, as we discussed in Sect. 2.3. Equation (2.31) indicates that the equality is recovered if the observables determine in a unique way the entropy production or the dissipated work.

We are interested in exploring this formula in simulations and experiments where a microscopic system reaches a NESS and we are given a single stationary trajectory or time series produced by the system. In an experimental context, the observables are usually sampled at a finite frequency. The output is then a time series of data or discrete trajectory, $\mathbf{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, where $\hat{x}_i$ can be the value of a single or several observables of the system. In this case, we are interested in estimating the entropy production *per data* produced by the underlying physical process, which we denote by $\langle \dot{S}_{\rm prod} \rangle$ in the rest of the chapter. Entropy production per data is related to the KLD rate *per data*, which we define below.

For the sake of simplicity, we now consider random discrete processes, but the discussion below holds also for continuous random processes. Given an infinitely long realization or time series sampled from a random discrete process X_i $(i = 1, 2, \ldots)$, which can be multi-dimensional, we define by $p(x_1^m)$ the probability that a given sequence of m consecutive data is equal to $x_1^m = (x_1, x_2, \ldots, x_m)$. We define

the $m-$th order KLD for this random process X_i by the distinguishability (KLD) between $p(x_1^m)$ and the probability $p(x_m^1)$ to observe the reverse sequence of data $x_m^1 = (x_m, x_{m-1}, \ldots, x_1)$.

$$D_m^X = D[p(x_1^m)||p(x_m^1)] = \sum_{x_1,\ldots,x_m} p(x_1^m) \ln \frac{p(x_1^m)}{p(x_m^1)}. \tag{2.32}$$

The *KLD rate* for the process X_i is defined as the growth rate of D_m^X with the number of data,

$$d^X = \lim_{m\to\infty} \frac{D_m^X}{m}. \tag{2.33}$$

Because of the finite time sampling and given that x may not contain the information of the entropy production (2.31), the chain rule (2.18) implies that the KLD rates bounds from below the entropy production per data

$$\frac{\langle \dot{S}_{\text{prod}} \rangle}{k} \geq d^X. \tag{2.34}$$

The above bound is saturated if the random variable is the microstate of the system $X = \{\mathbf{q}, \mathbf{p}\}$ and the sampling rate is infinite or X determines uniquely the entropy production in the process.

Equation (2.34) is our basic result. It reveals a striking connection between physics and the statistics of a time series. The left-hand side, $\langle \dot{S}_{\text{prod}} \rangle / k$, is a purely physical quantity, whereas the right-hand side, d^X, is a statistical magnitude depending solely on the observed data, but not on the physical mechanism generating the data. This means that if we are given a stationary time series of any random variable X produced by a microscopic system we can bound from below the average entropy production rate in the physical mechanism that generated the data. In particular (2.34) can be used to study the minimum amount of entropy produced in a symmetry restore such as the erasure of a bit, which yields Landauer's principle relating entropy production and logical irreversibility in computing machines [2, 15, 19] as we will see in Chap. 6. Equation (2.34) extends this principle and suggests that we can determine the average dissipation of an arbitrary NESS, even ignoring any physical detail of the system.

2.5 Discrete Systems

We first study the bound to the entropy production provided by the KLD in discrete nonequilibrium stationary processes, namely in Markov chains satisfying detailed balance condition and in Hidden Markov models.

2.5.1 Markov Chains Obeying Local Detailed Balance

We first analyze how the bound (2.34) is expressed for Markovian time series that obey the detailed balance condition by deriving analytical expressions for both entropy production and the KLD rate.

Among all the stochastic processes, *Markov* processes are the most important ones in physics, chemistry and biology [27]. Let us consider a stochastic process of a random (discrete or continuous) variable X. Such a process is said to be *Markovian* if the probability to observe any sequence of n data x_1^n at times $t_1^n = (t_1, t_2, \ldots, t_n)$ satisfies the following property,

$$\rho(x_{n+1}, t_{n+1}|x_1, t_1, \ldots, x_n, t_n) = \rho(x_{n+1}, t_{n+1}|x_n, t_n), \tag{2.35}$$

where the bar | denotes conditioned probability. Therefore, in a Markov process, the probability to observe a value of the process at a given time t_{n+1} only depends on the state of the process one step before, at time t_n. In order to know the probability to observe a sequence x_1^n, we only need to know $p(x_1, t_1)$ and the transition probability $p(x_2, t_2|x_1, t_1)$ for successive times t_1, t_2 and any value of x_1, x_2, since the following property derives from (2.35)

$$\rho(x_1, t_1; \ldots; x_n, t_n) = p(x_1, t_1) \cdot p(x_2, t_2|x_1, t_1) \cdots p(x_n, t_n|x_{n-1}, t_{n-1}). \tag{2.36}$$

Nuclei decay, the voltage in an RC circuit, or the motion of a molecular motor in a microtubule can be described by Markov processes. Concerning microscopic physics, the position of a Brownian particle described by the overdamped Langevin equation can be considered as a Markov process [27]. The probability a random *discrete* variable X to take the value x_i at time t, $p_i = p_i(t)$ obeys the *Master equation*,

$$\dot{p}_i = \sum_j k_{j\to i} p_j - k_{i\to j} p_i, \tag{2.37}$$

where $k_{i\to j}$ is the rate from state i to state j, i.e., the number of times that the transition $i \to j$ occurs per unit of time. Therefore, $k_{i\to j} \geq 0$ for all i, j. The Master equation (2.37) can be seen as a gain-loss equation for state i. The term $k_{j\to i} p_j$ accounts for the net incoming probability from any state j to state i and the term $-k_{i\to j} p_i$ accounts for the losses from state i to any other state j. The net change of p_i due to an exchange with state j is defined as the *current*

$$J_{j\to i} = k_{j\to i} p_j - k_{i\to j} p_i, \tag{2.38}$$

which allows one to write the master equation as a balance equation

$$\dot{p}_i = \sum_j J_{j\to i}. \tag{2.39}$$

A Markov process reaches a stationary state when the probability to be at any state i does not change in time, and reaches a stationary value, $p_i^{\rm ss}$. Equivalently, in the stationary state, $\dot{p}_i = 0$, which implies by (2.37) the balance condition $\sum_j J_{j\to i} = 0$ for all i. If the following (stronger) condition called *detailed balance condition* is satisfied for every pair of states i and j,

$$\frac{k_{i\to j}}{k_{j\to i}} = \frac{p_j^{\rm ss}}{p_i^{\rm ss}}, \tag{2.40}$$

the system also reaches a stationary state. Notice however that detailed balance condition on the transition rates $k_{i\to j}$ *does not* ensure that the stationary state is an equilibrium state. For a physical system that is in contact with a thermal bath at temperature T during the process, and where the potential energy of the state i is V_i, detailed balance condition is written as

$$\frac{k_{i\to j}}{k_{j\to i}} = \exp\left(-\frac{V_j - V_i}{kT}\right), \tag{2.41}$$

and it has to be satisfied for every i, j in order to reach a stationary equilibrium state. Equation (2.41) is compatible with stationary probabilities that are weighed by the Boltzmann factor, $p_i^{\rm ss} = p_i^{\rm eq} \propto \exp(-\beta V_i/kT)$. In fact, local detailed balance condition is derived by imposing equilibrium stationary probabilities in detailed balance condition (2.40).

When both the state x_1^n and the times t_1^n are discrete, the Markov process is called a *Markov chain* [27]. A simple example of a three-state Markov chain is shown in Fig. 2.5. A physical system jumps at discrete times $t = 0,\ \Delta t,\ 2\Delta t,\ \ldots$ between three states labeled by an index $s = 1, 2, 3$. The transition rates $k_{i\to j}$ and the balance equation for the probability to stay in one of the three states are illustrated in Fig. 2.5.

Let us now consider a Markov chain X_i where the random variable can only take discrete values and we can only sample the value of the random variable at a finite frequency. For a Markov chain, the probability distribution to observe a sequence x_1^m, $p(x_1^m)$, factorizes as $p(x_1^m) = p(x_1)p(x_2|x_1)\cdots p(x_m|x_{m-1})$, which also holds if we reverse the arguments, i.e., for $p(x_m^1)$. Substituting these expressions into equation (2.33), we get

$$d^X = \sum_{x_1,x_2} p(x_1,x_2)\ln\frac{p(x_2|x_1)}{p(x_1|x_2)} = D_2^X - D_1^X = D_2^X, \tag{2.42}$$

since $D_1^X = 0$ when comparing a trajectory and its reverse. Therefore, d^X only depends on transition probabilities if X is a random Markovian process. This expression was also derived for Markov chains in [9].

We now relate d^X in Eq. (2.42) with the entropy production when the system reaches a NESS, because it is in contact with several thermal baths. In this situation, the local detailed balance condition is satisfied. We call $V(x_i)$ is the energy of the

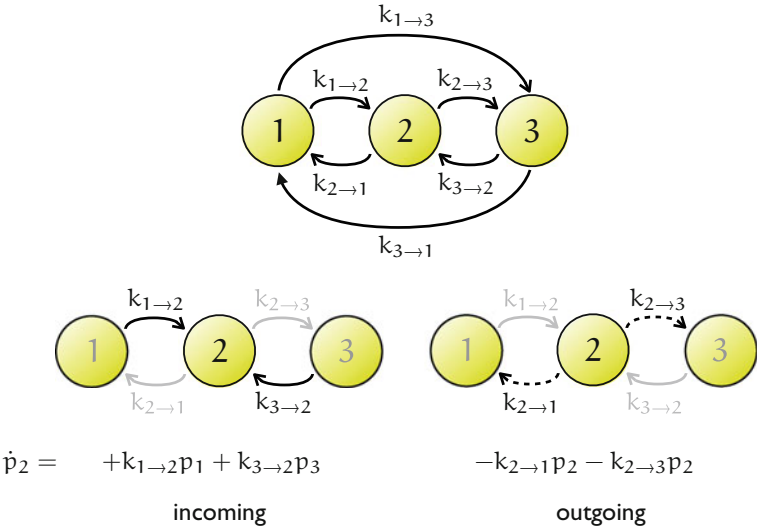

Fig. 2.5 Example of a Markov chain. A system can jump randomly between three states labeled by 1, 2 and 3. When the system is at state i, its state in the next step is $j \neq i$. The transition rates are indicated in the *top* figure. In the *bottom* figure, the incoming and outgoing contributions for the time derivative of p_2 are illustrated (*solid lines* for incoming and *dashed* for outgoing). In the *bottom line* the Master equation for the probability to be in state 2 is written

state x_i, and T_{x_1,x_2} is the temperature of the bath that activates the transitions $x_1 \to x_2$ and $x_2 \to x_1$. The local detailed balance condition reads in this case

$$\frac{p(x_2|x_1)}{p(x_1|x_2)} = \exp\left(-\frac{V(x_2) - V(x_1)}{k\, T_{x_1,x_2}}\right). \tag{2.43}$$

Inserting (2.43) into (2.42),

$$\begin{aligned} d^X &= \sum_{x_1,x_2} p(x_1, x_2) \frac{V(x_1) - V(x_2)}{k\, T_{x_1,x_2}} \\ &= \sum_{x_1,x_2} p(x_1, x_2) \frac{Q_{x_1,x_2}}{k\, T_{x_1,x_2}} = \frac{\langle \dot{S}_{\text{prod}} \rangle}{k}, \end{aligned} \tag{2.44}$$

where $Q_{x_1,x_2} = V(x_1) - V(x_2)$ is the heat dissipated to the corresponding thermal bath in the jump $x_1 \to x_2$, and $\dot{S}_{\text{prod}}$ is the total entropy production per data. Therefore, Eq. (2.34) is reproduced, with equality, in the case of a physical system obeying local detailed balance, if we have access to all the variables describing the system. The same conclusion is reached if we induce the NESS by means of non-conservative constant forces.

Equation (2.42) can be explored further by means of the current from the state x_1 to the state x_2 as the net probability flow from x_1 to x_2, $J_{x_1 \to x_2} = p(x_1, x_2) - p(x_2, x_1)$.

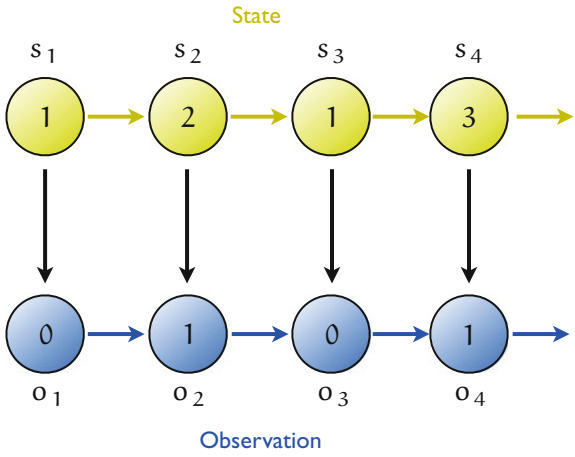

Fig. 2.6 Example of a Hidden Markov chain. The state of a system s changes in discrete time steps according to the Markovian process described in Fig. 2.5. At every time, the observed state o is obtained according to the following rule: If $s_i = 1$, then $o_i = 0$, else $o_i = 1$. The sequence of observations $o_1, o_2, o_3, \ldots$ does not form a Markov chain while the (hidden) state does

If the system is not far from equilibrium the current tends to zero, and the following condition is satisfied $J_{x_1 \to x_2} \ll p(x_1, x_2)$, yielding

$$\frac{\langle \dot{S}_{\text{prod}} \rangle}{k} = d^X = D_2^X \simeq \sum_{x_1, x_2} \frac{(J_{x_1 \to x_2})^2}{2p(x_1, x_2)}. \tag{2.45}$$

This expression is well known from linear irreversible thermodynamics [cf. Eq. (2.4)]. Equation (2.45) implies that the time asymmetry of a Markovian process not far from equilibrium is revealed by the currents or probability flows that can be observed. In other words, a Markovian process without flows is time reversible. This is not the case for non-Markovian time series, where irreversibility can show up even in the absence of currents, as shown in the next section.

2.5.2 *Hidden Markov Processes*

In many experimental situations, a physical process is Markovian at a micro- or mesoscopic level of description, but the observed time series only contain a subset of the relevant observables, being non-Markovian in general. This is the case in biological systems, where one can only register the behavior of some mechanical and maybe a few chemical variables, while most of the relevant chemical variables cannot be monitored. These kind of non-Markovian time series obtained from an underlying Markov process are called *Hidden Markov processes* [23]. If the time and the state of the system are both discrete, the process is called a *Hidden Markov chain*. A simple example of a Hidden Markov chain is shown in Fig. 2.6.

We now show how to calculate the KLD rate between a specific case of hidden Markov chains semi-analytically. We focus on a simple case where an underlying

Markov process is described by two observables X and Y; however we only observe X whose evolution is described by a hidden Markov chain. The KLD rate for the observable X is

$$\begin{aligned} d^X &= \lim_{m\to\infty} \frac{1}{m} \sum_{x_1^m} p(x_1^m) \ln \frac{p(x_1^m)}{p(x_m^1)} \\ &= \lim_{m\to\infty} \frac{1}{m} \sum_{x_1^m} p(x_1^m) \ln \frac{\sum_{y_1^m} p(x_1^m, y_1^m)}{\sum_{y_m^1} p(x_m^1, y_m^1)}. \end{aligned} \tag{2.46}$$

where we have expressed the marginal distribution in X by summing the joint distribution in X, Y to all the possible values that Y can take, $p(x_1^n) = \sum_{y_1^n} p(x_1^n, y_1^n)$. The chain rule ensures that the KLD for the random variable X is smaller than the KLD calculated with full information given by x and Y, $d^X \leq d^{X,Y}$. To compute d^X analytically, it is convenient to write d^X as a difference between two terms,

$$d^X = h_r^X - h^X, \tag{2.47}$$

where

$$\begin{aligned} h^X &= -\lim_{m\to\infty} \frac{1}{m} \sum_{x_1^m} p(x_1^m) \ln p(x_1^m), \\ &= -\lim_{m\to\infty} \frac{1}{m} \sum_{x_1^m} p(x_1^m) \ln \sum_{y_1^m} p(x_1^m, y_1^m), \end{aligned} \tag{2.48}$$

is called *Shannon entropy rate*. On the other hand,

$$\begin{aligned} h_r^X &= -\lim_{m\to\infty} \frac{1}{m} \sum_{x_1^m} p(x_1^m) \ln p(x_m^1), \\ &= -\lim_{m\to\infty} \frac{1}{m} \sum_{x_1^m} p(x_1^m) \ln \sum_{y_m^1} p(x_m^1, y_m^1), \end{aligned} \tag{2.49}$$

is called the *cross entropy rate*. Since the underlying process is Markovian, the probability distribution in X, Y factorizes (2.36) $p(x_1^m, y_1^m) = p(x_1, y_1)p(x_2, y_2|x_1, y_1)\cdots p(x_m, y_m|x_{m-1}, y_{m-1})$ and both Shannon and cross entropy can be expressed in terms of the trace of a product of random transition matrices $\mathbf{T}$ [12, 14]. These are square $M \times M$ random matrices, where M is the number of values that the variable y can take on, and their entries are given by

$$\mathbf{T}(x_1, x_2)_{y_1 y_2} = p(x_2, y_2|x_1, y_1). \tag{2.50}$$

There are a total number of N^2 transition matrices, where N is the number of values that x can take on. Note the different role played by each variable in this formalism:

x_i are parameters defining the matrix (making $\mathbf{T}$ a random matrix), whereas y_i are subindices of the matrix elements. The Shannon and cross entropy can be expressed in terms of these matrices,

$$h^X = -\lim_{m\to\infty} \frac{1}{m} \left\langle \ln \mathrm{Tr} \left[\prod_{i=1}^{m-1} \mathbf{T}(x_i, x_{i+1}) \right] \right\rangle, \tag{2.51}$$

$$h_r^X = -\lim_{m\to\infty} \frac{1}{m} \left\langle \ln \mathrm{Tr} \left[\prod_{i=1}^{m-1} \mathbf{T}(x_{m-i+1}, x_{m-i}) \right] \right\rangle, \tag{2.52}$$

where $\langle\cdot\rangle$ denotes the average over the random process X_i, which is weighted by $p(x_1^m)$. For sufficiently large m, Eqs. (2.51) and (2.52) are self-averaging [12], meaning that we do not need to calculate the average but just compute the trace for a single stationary trajectory. For any sufficiently long time series $\mathbf{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ with n large, the following expressions converge to $-h$ and $-h_r$ almost surely [12] [2]

$$\hat{\lambda}^{\mathbf{x}} = \frac{1}{n} \ln \left\| \prod_{i=1}^{n-1} \mathbf{T}(\hat{x}_i, \hat{x}_{i+1}) \right\| \simeq -h^X, \tag{2.53}$$

$$\hat{\lambda}^{\tilde{\mathbf{x}}} = \frac{1}{n} \ln \left\| \prod_{i=1}^{n-1} \mathbf{T}(\hat{x}_{n-i+1}, \hat{x}_{n-i}) \right\| \simeq -h_r^X, \tag{2.54}$$

where $\|\cdot\|$ is any matrix norm that satisfies $\|\mathbf{A}\cdot\mathbf{B}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$. In particular, the trace satisfies this condition for positive matrices. In the context of random matrix theory, $\hat{\lambda}^{\mathbf{x}}$ and $\hat{\lambda}^{\tilde{\mathbf{x}}}$ are known as *maximum Lyapunov characteristic exponents* [5] and measure the asymptotic rate of growth of a random vector when being multiplied by a random sequence of matrices. In practice, we can estimate d^X semi-analytically as

$$\hat{d}^{\mathbf{x}} = \hat{\lambda}^{\mathbf{x}} - \hat{\lambda}^{\tilde{\mathbf{x}}}. \tag{2.55}$$

Here $\hat{\lambda}^{\mathbf{x}}$ and $\hat{\lambda}^{\tilde{\mathbf{x}}}$ are estimated using (2.53) and (2.54) with a single time series $\mathbf{x}$ of size n, following a numerical technique introduced in Ref. [5] to calculate Lyapunov characteristic exponents:

1. We generate a random stationary time series $\mathbf{x} = \{\hat{x}_1^n\}$ and compute the matrices $\mathbf{T}$ analytically.
2. A random unitary vector is multiplied by those matrices in the order given by (2.53) and normalized every l data, keeping track of the normalization factor.
3. The product of these factors divided by n yields $\hat{\lambda}^{\mathbf{x}}$.
4. For $\hat{\lambda}^{\tilde{\mathbf{x}}}$, the same procedure is repeated but using the reversed time series $\tilde{\mathbf{x}} = \{\hat{x}_n^1\}$.
5. The KLD is estimated using Eq. (2.55).

[2] A sequence of a random variable X, given by $X_1, X_2, \ldots$, is said to converge almost surely to x when the probability that the sequence satisfies $\lim_{n\to\infty} X_n = x$ is equal to 1.

The technique is semi-analytical since the transition probabilities are known analytically but a single random stationary time series $\mathbf{x}$ is necessary to estimate d^X with the multiplication of n transition matrices that are chosen according to $\mathbf{x}$.

Let us recall that the estimator $\hat{d}^{\mathbf{x}}$ cannot be applied to empirical time series unless we know the Markov model generating the data. Consequently, it is not useful in practical situations. However, we will use it to test the performance of the estimators of the KLD introduced in the next chapters. On the other hand, one can also analytically estimate of Eqs. (2.51) and (2.52) by using the replica trick, in an analogous way as done in Ref. [7]. The calculation is cumbersome and is explained in Appendix B.3.

References

1. D. Andrieux, P. Gaspard, S. Ciliberto, N. Garnier, S. Joubaud, A. Petrosyan, Phys. Rev. Lett. **98**, 150601 (2007)
2. D. Andrieux, P. Gaspard, Proc. Nat. Acad. Sci. **105**, 9516–9521 (2008)
3. D. Andrieux, P. Gaspard, S. Ciliberto, N. Garnier, S. Joubaud, A. Petrosyan, J. Stat. Mech: Theory Exp. **2008**, P01002 (2008)
4. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley-interscience, New York, 2006)
5. A. Crisanti, G. Paladin, A. Vulpiani, *Products of Random Matrices in Statistical Physics* (Springer Series in Solid State Sciences) vol. 104 (Springer-Verlag, Berlin, 1993) p. 166
6. G. Crooks, Phys. Rev. E **60**, 2721 (1999)
7. M. De Oliveira, A. Petri, Phys. Rev. E **53**, 2960 (1996)
8. M. Esposito, J.M.R. Parrondo, In preparation (2013)
9. P. Gaspard, J. Stat. Phys. **117**, 599–615 (2004)
10. A. Gomez-Marin, J.M.R. Parrondo, C. Van den Broeck, Phys. Rev. E **78**, 011107 (2008)
11. A. Gomez-Marin, J.M.R. Parrondo, C. Van den Broeck, EPL-Europhys. Lett. **82**, 50002–54000 (2008)
12. T. Holliday, A. Goldsmith, P. Glynn, IEEE Trans. Inf. Theory **52**, 3509–3532 (2006)
13. J. Horowitz, C. Jarzynski, Phys. Rev. E **79**, 21106 (2009)
14. P. Jacquet, G. Seroussi, W. Szpankowski, Theoret. Comput. Sci. **395**, 203–219 (2008)
15. R. Kawai, J.M.R. Parrondo, C.V. den Broeck, Phys. Rev. Lett. **98**, 80602 (2007)
16. D. Kondepudi, I. Prigogine, *From Heat Engines to Dissipative Structures* (Wiley, New York, 1998)
17. S. Kullback, R.A. Leibler, Ann. Math. Stat. **22**, 79–86 (1951)
18. L. Lacasa, A. Nuñez, É. Roldán, J.M.R. Parrondo, B. Luque,Eur. Phys. J. B-Condensed Matter and Complex Systems **85**, 1–11 (2012)
19. R. Landauer, IBM J. Res. Dev. **5**, 183–191 (1961)
20. C. Maes, Séminaire Poincaré **2**, 29–62 (2003)
21. J.M.R. Parrondo, B. Cisneros, Appl. Phys. A Mater. Sci. Process. **75**, 179–191 (2002)
22. J.M.R. Parrondo, C. Van den Broeck, R. Kawai, New J. Phys. **11**, 073008 (2009)
23. L.R. Rabiner, Proc. IEEE **77**, 257–286 (1989)
24. É. Roldán, J.M.R. Parrondo, Phys. Rev. Lett. **105**, 150607 (2010)
25. É. Roldán, J.M.R. Parrondo, Phys. Rev. E **85**, 031129 (2012)
26. U. Seifert, Phys. Rev. Lett. **95**, 40602 (2005)
27. N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, vol. 1 (North holland, 1992)

http://www.springer.com/978-3-319-07078-0

Irreversibility and Dissipation in Microscopic Systems

Roldán, É.

2014, XXI, 211 p. 75 illus., 41 illus. in color., Hardcover
ISBN: 978-3-319-07078-0