KTH Engineering Sciences

Master Thesis

Construction of equilibrium distributions from simulation of nonequilibrium processes with feedback control

Emil Mallmin

Statistical Physics Group, Department of Physics,
School of Engineering Sciences
Royal Institute of Technology, SE-106 91 Stockholm, Sweden

Stockholm, Sweden 2017

Typeset in LaTeX

TRITA-FYS 2017:44
ISSN 0280-316X
ISRN KTH/FYS/–17:44–SE

Printed in Sweden by Universitetsservice US AB, Stockholm June 2017

Abstract

The fluctuation theorems of stochastic thermodynamics have previously been used to derive a reweighting scheme that constructs an equilibrium distribution from realizations of a driven nonequilibrium process. This has been applied as a simulation strategy to calculate equilibrium averages in models where the equilibrium distribution is difficult to simulate directly. Fluctuation theorems have also been generalized to encompass a feedback controller acting on the system. In this thesis, these results are combined to extend the reweighting scheme to the presence of feedback, which widens the range of potential use cases for the method. In the statistical weights, this introduces an information theoretic term involving a virtual "reverse" feedback mechanism not part of the simulation. This becomes a degree of freedom that may be chosen to optimize the convergence of the reweighting scheme. This is done analytically to yield an expression for the information term, which may however become difficult to estimate accurately from simulation data. The reweighting scheme is then applied to a simple Brownian particle in a 1D potential, both in the absence of feedback, and for two special cases of feedback for which the information term is simple to calculate. The reweighting is reliable for the model without feedback, but not for the two cases of feedback considered. The data indicate that these feedback choices are suboptimal in different respects. This suggests that for the extended method to work reliably, both the simulated and reverse feedback mechanisms must be chosen close to optimally.

Sammanfattning

Fluktuationsteoremen inom stokastisk termodynamik har tidigare utnyttjats för att härleda en omviktningsmetod som konstruerar en jämviktsfördelning från realiseringar av en driven ickejämviktsprocess. Detta har använts som en strategi för att beräkna jämviktsmedelvärden i modeller där jämviktsförelningen i sig är svår att simulera. Fluktuationsteoremen har också generaliserats till att inbegripa återkopplad styrning av systemet. I denna uppsats kombineras dessa resultat för att utvidga omviktningsmetoden till återkopplad styrning, vilket potentiellt utökar användbarheten av metoden till fler simuleringssituationer. För de statistiska vikterna tillkommer då en informationsteoretisk term som innefattar en virtuell "reverserad" återkopplingsmekanism som inte direkt ingår i simuleringen av systemet. Denna blir en frihetsgrad som kan väljas för att optimera konvergensegenskaperna för omviktningsmetoden. Detta görs här analytiskt, vilket leder till ett uttryck för informationstermen som dock kan vara svårt att med noggrannhet upskatta från simuleringsdata. Omviktningsmetoden appliceras på en modell av en Brownsk partikel i en endimensionell potential, dels utan återkoppling, och för två specialfall av återkoppling där informationstermen är lätt att beräkna. Omviktningen

för modellen utan återkoppling fungerar väl, men så är inte fallet för de studerade valen av återkoppling. Simuleringsdatat indikerar att dessa fall på olika sätt är suboptimala. Detta föranleder slutsatsen att den utvidgade omviktningsmetoden kan förväntas vara tillförlitlig endast när både den simulerade och den reverserade återkopplingsmekanismen väljs nära optimum.

Preface

This work was done full-time between January and June 2017, and signifies the completion of my Master degree in Physics as well as my five years at KTH.

Foremost, I'd like to thank my supervisor Jack Lidmar. Especially for his help and interest in this Master Thesis project, but also for his encouragement and patience in supervising my Philosophy of Thermal Physics individual project, and for providing high quality teaching during several other Master modules. It has also been a pleasure to share the "thesis experience" (and office space!) with my fellow Master students at the Theoretical Physics Department. The staff and research students at the department have also been most welcoming, and it's been great fun to have lunch and "fredagsfika" together.

Contents

Chapter 1

Introduction

This thesis concerns how the fluctuation theorems of nonequilibrium theory can be used to calculate equilibrium averages from simulations of nonequilibrium processes. Specifically, it explores the possibility to calculate such averages when the simulated physical system is acted upon by a feedback controller.

Fluctuation theorems form a central part of *stochastic thermodynamics*, a theory of "small" physical systems that has emerged over the last two decades in tandem with the development of nano-scale experimental methods. It unifies several conceptually distinct approaches to nonequilibrium physics in the mathematical framework of stochastic (Markov) dynamics [1]. Thermodynamic concepts such as work, heat, and entropy, can be generalized to individual trajectories of such stochastically evolving systems. The fluctuation theorems follow from quite general assumptions. One considers a *forward process*, in which a system that would spontaneously relax towards equilibrium, is driven away from equilibrium by a time-varying parameter. The sequence of parameter values is called a *protocol*, and its effect is to change the energy landscape (the potential) of the system in time. One contrasts this process with an imagined *reverse process* in which the system is driven by the same protocol run in reverse sequence. The relative likelihood of a sequence of system states in the forward process (a *realization*), and the time-reversal of that sequence in the reverse process, is taken as a measure of the irreversibility of the dynamics. Under quite general conditions it equals the exponential of the work dissipated into the thermal environment during the forward process. This leads for example to the *Jarzynski equality* (JE) $\langle e^{-\beta \mathcal{W}} \rangle = e^{-\Delta F}$, relating the work done over many realizations of the forward process to the free energy difference between the initial and final energy landscape [2, 3]. Fluctuation theorems have been derived for quantum systems, but in the following we will only have classical dynamics in mind.

Simulations have played a role in both the discovery and validation of fluctuations theorems [4]. But in the last 10 years or so, methods have been devised that *use* these theoretical results to improve aspects of statistical mechanical (SM) or molecular dynamical (MD) simulations. For example, the JE has been used for the correction of discretization errors in simulation of stochastic differential equations [5], and has inspired variations of Markov Chain Monte Carlo using non-equilibrium work values [6, 7]. The application of interest to this thesis is how a certain weighted average of nonequilibrium realizations is able to recreate an equilibrium distribution that is not directly simulated—the one corresponding to the final value λ_f of the protocol used in the driven process (which itself need not be in equilibrium for λ_f). The weighting factor for each realization i is the exponential of work done by the driving protocol, $w^{(i)} = e^{-\beta\mathcal{W}^{(i)}}$. Often, the energy landscape of an SM or MD simulation is rough and contains energy barriers that prevent the whole of phase space being probed. This poses a challenge to the efficient and accurate calculation of equilibrium quantities. The reweighting scheme offers the opportunity to drive the system into exploring phase space and artificially obtaining the desired equilibrium average from the statistical properties of the work values $\mathcal{W}^{(i)}$ [8, 9]. Another use case is to pool non-equilibrium and equilibrium simulation data, to avoid discarding the non-equilibrium transient in an equilibrium simulation.

However, rather than the practical application of this reweighting scheme to specific model systems, the problem of this thesis concerns the modifications that must be made to the method under the addition of a feedback control mechanism. Instead of letting the driving protocol be *a priori* chosen, it can itself be taken as a stochastic process, one that interacts with the system through feedback loops. The reason this is a potentially useful line of research, is the flexibility that is gained in designing protocols to achieve e.g. the effects suggested in the previous paragraph. Thus, one would be able to utilize the power to play “Maxwell’s demon” in a computer simulation to steer the behaviour of the system in a favourable direction. The fluctuation theorems have been extended to encompass feedback control [10], which has invigorated *information thermodynamics* [11]—the quantitative study of the interconversion of thermodynamic work and information. With feedback, there arises an additional information theoretic term $\mathcal{I}$ in the JE, $\langle e^{-\beta\mathcal{W}-\mathcal{I}}\rangle = e^{-\Delta F}$, and hence also in the reweighting factor, $w^{(i)} = e^{-\beta\mathcal{W}^{(i)}-\mathcal{I}^{(i)}}$. The extension of the reweighting scheme to feedback is a topic that so far has not appeared in publications.

There are two challenges to this extension. The feedback must be such that (i) $\mathcal{I}$ is practically calculable from simulation data, and (ii) the statistical properties of the weights w allow the equilibrium estimation to converge sufficiently quickly for finite data. To address these problems, it is essential to understand the origin of $\mathcal{I}$. It involves the transition probabilities of the stochastic protocol in the forward and reverse process. The forward process is straightforwardly extended to incorporate a feedback mechanism. But the reverse process cannot contain the time-reversal of this feedback since that would be acausal—it would control the system with respect to measurements not yet made. In most of the research literature (with some exceptions), one therefore allows feedback only in the forward process and not in the reverse. This has the advantage of making $\mathcal{I}$ readily interpretable as the information gained through measurements in the forward process. Unfortunately, $\mathcal{I}$

is then not practically calculable from simulation data. There is however the option to retain a causal feedback mechanism in the reverse process, one that is logically independent from the forward feedback. This opens the possibility of choosing the reverse feedback mechanism in such a way that both conditions (i) and (ii) could be satisfied [12]. The situation is peculiar, in that only the forward processes is simulated, so the reverse feedback mechanism is only used after the fact, in the calculation of $\mathcal{I}$. The question is how to best exercise this freedom of choice.

I have aimed to write this thesis in a clear and self-contained way, so that it may hopefully be readily understood by anyone with a basic knowledge of probability theory, thermodynamics and statistical mechanics. Therefore, chapter 2 is included as an elementary introduction to some core concepts like *information* (2.1) and *Markov processes* (2.2).

Chapter 3 contains an exposition of fluctuation theorems, with and without feedback control (3.2 and 3.3). Particularly, section 3.3.2 considers the case where the reverse process also has feedback.

In chapter 4, the reweighting scheme is explained in detail, and some of its statistical properties are discussed (4.1). Then some new analytical results are derived regarding the optimal choice of reverse feedback (4.3). The reweighting scheme is applied in simulation of a 1D Brownian particle, first without feedback (4.4) and then for two special cases of feedback (4.5).

The conclusion drawn from the simulations and analytical work is summarized in chapter 5.

Chapter 2

Mathematics and modelling background

2.1 Information theory

Colloquially, we associate the reduction of uncertainty with a gain of information. This notion was formalized in information theory by Claude Shannon in 1948 [13]. The information gained by witnessing the outcome x, which carried probability $p(x)$, is defined

$$I(x) = -\log p(x). \tag{2.1}$$

If there is only one outcome (no uncertainty) no information is gained; witnessing an unlikely outcome gives more information than a likely outcome. Furthermore, the information in witnessing two *independent* events x, y drawn from $p(x, y) = p(x)p(y)$ is equal to the sum of the information in each separate observation; $I(x, y) = I(x) + I(y)$. In fact, (2.1) is the unique measure with these properties, up to a multiplicative constant which can be absorbed into the choice of base of the logarithm. For $\log_2$ the unit of information is called a *bit*, and for ln a *nat*.

The Shannon entropy H is defined as the expected value of information.

$$H(X) := \langle I \rangle = -\sum_x p(x) \log p(x). \tag{2.2}$$

(The sum may be substituted for an integral) It is maximized for uniform distributions.

There is a host of other information quantities related to the (Shannon) entropy [14]. The *joint entropy* of two random variables X and Y is

$$H(X, Y) := -\sum_x \sum_y p(x, y) \log p(x, y) \tag{2.3}$$

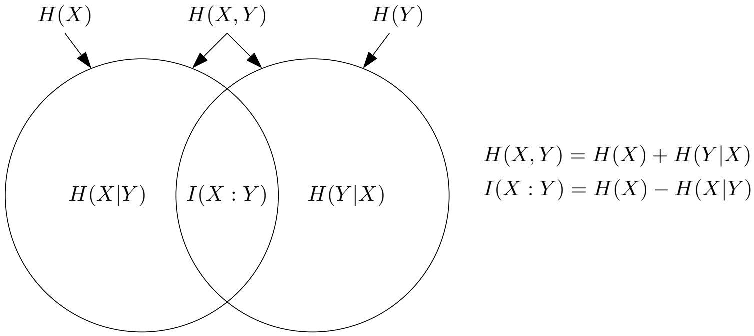


Figure 2.1: Venn diagram of entropies.

(which is of course the entropy of the variable $Z = (X, Y)$). The *conditional entropy* $H(X|Y)$ is defined by

$$H(X|Y) := \sum_y p_Y(y) H(X|Y = y) = -\sum_{x,y} p(x, y) \log p(x|y). \tag{2.4}$$

Then there is the *mutual information* $I(X : Y)$ defined by

$$I(X : Y) := \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{2.5}$$

The relation between these different information quantities can be summarized in a Venn diagram [14] (fig. 2.1).

Another useful information quantity is the *relative entropy*, or *Kullback–Leibler* (KL) *divergence*, defined for two probability distributions $p(x)$ and $q(x)$:

$$D(p \,||\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{2.6}$$

It is implied that it converges only if there is no x for which $p(x) > 0$ but $q(x) = 0$. It can be thought of as measuring the “distance” or “divergence” between probability distributions. It is however *not* a metric since it is neither symmetric nor satisfies a triangle inequality. It does however satisfy positivity,

$$D(p \,||\, q) \geq 0, \tag{2.7}$$

with equality if and only if $p(x) = q(x)$ [14]. It is related to mutual information by

$$I(X : Y) = D(p(x, y) \,||\, p(x)p(y)). \tag{2.8}$$

This grants the interpretation of mutual information as measuring the degree of independence between the random variables, or the amount of information each marginal distribution contains about the other.

Lastly, A useful inequality in information theory is Jensen's inequality which states that for any convex function[1] $f(x)$,

$$\langle f(x)\rangle \geq f(\langle x\rangle). \tag{2.9}$$

2.2 Markov processes

2.2.1 The Markov property

A *stochastic process* $\boldsymbol{X} = \{\boldsymbol{x}(t) : t \in T\}$ is a collection of random variables indexed by time, where the indexing set T may be either discrete (e.g. $\{0, 1, 2, 3, \ldots\}$) or continuous (e.g. $[0, \infty)$). A *realization* of a stochastic process is a sequence of outcomes $X = \{x(t) : t \in T\}$. Probabilities are associated with observing a particular outcome x_n at time t_n, i.e. the event (x_n, t_n). Assume by convention $t_N > t_{N-1} > \ldots > t_1$. Then, the probability of observing for those times the sequence of outcomes $x_N, x_{N-1}, \ldots, x_1$ is denoted[2]

$$P_{(N)}(x_N, t_N;\, x_{N-1}, t_{N-1};\, \ldots\, ;\, x_1, t_1). \tag{2.10}$$

A stochastic process is a *Markov process* if it has the *Markov property*: the probability of a future state conditioned on a history of states, depends only on the most recent state:

$$P_{(1|N-1)}(x_N, t_N|x_{N-1}, t_{N-1};\, \ldots\, ;\, x_1, t_1) = P_{(1|1)}(x_N, t_N|x_{N-1}, t_{N-1}). \tag{2.11}$$

Then, all probability functions can be constructed from the knowledge of an initial distribution at time t_1, and the transition probabilities for all other times,

$$\begin{aligned} &P_{(N)}(x_N, t_N; x_{N-1}, t_{N-1};\, \ldots\, ;\, y_1, t_1) = \\ &\qquad P_{(1|1)}(x_N, t_N|x_{N-1}, t_{N-1}) \cdots P_{(1|1)}(x_2, t_2|x_1, t_1) P_{(1)}(x_1, t_1). \end{aligned} \tag{2.12}$$

2.2.2 Jump and diffusion processes

The realizations of a continuous time Markov process may be continuous, or a series of discrete jumps, or most generally a combination of both. A Markov process with only discrete jumps in called a *jump process*. (This is necessarily the case if the state space is discrete). Suppose the state is known with certainty to be x_0 at time $t = t_0$. Define $P(x, t) := P_{(1|1)}(x, t|x_0, t_0)$ for $t > t_0$. Then the stochastic dynamics of a jump process is described by a *Master equation* (ME),

$$\frac{\partial P(x,t)}{\partial t} = \int dx'\, [W(x|x', t)P(x', t) - W(x'|x, t)P(x, t)], \tag{2.13}$$

with initial condition $P(x, t_0) = \delta(x - x_0)$. $W(x'|x, t)$ is the probability per unit time to jump from state x to x' at time t. The Master equation is a balance equation

[1] A convex function is one that satisfies $f(ta + (1-t)b) \geq tf(a) + (1-t)f(b), 0 \leq t \leq 1$.

[2] $P_{(n|m)}(\underbrace{\cdots}_{n}|\underbrace{\cdots}_{m})$ notationally distinguishes all different probability distributions.

equating the change of probability of a state to the net "flow" of probability entering that state.

A Markov process with continuous realizations is called a *diffusion process*. The probability $P(x, t)$ then obeys a *Fokker-Planck equation* (FPE), which for a state-space of dimension one is

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x,t)P(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[B(x,t)P(x,t)]. \tag{2.14}$$

A characterizes drift, the movement of the mean state, and B the variance around that mean. For example, with $A(x, t) = a$ and $B(x, t) = 2D$ with initial state $x_0 = 0, t_0 = 0$, the solution for $t > 0$ is

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left[-(at-x)^2/2Dt\right]. \tag{2.15}$$

For a completely general Markov process $\partial_t P(x, t)$ equals the sum of the right hand side of the ME and FPE. As one might intuit, in a suitable limit of infinitesimally small jumps it is possible to approximate a diffusion process with a jump process (but not the other way around) [15].

2.2.3 Stationarity, detailed balance, and ergodicity

A stochastic process is called *stationary* if

$$P_{(N)}(x_N, t_N+\tau; \ldots; x_1, t_1+\tau) = P_{(N)}(x_N, t_N; x_{N-1}, t_{N-1}; \ldots; x_1, t_1) \tag{2.16}$$

for all τ. Since $P_{(N)}$ can then only depend on time differences, it follows in particular that

$$P_{(1)}(x,t) =: P_s(x). \tag{2.17}$$

An important concept related to stationarity is the condition of *detailed balance*. It holds if for every pair of states x and x', the probability of the transition $x \to x'$ is exactly compensated by the reverse transition $x' \to x$. For a stationary process this means

$$P_{(2)}(x', t+dt; x, t) = P_{(2)}(x, t+dt; x', t). \tag{2.18}$$

If the process is described by the ME (2.13), the instantaneous transition rates become time-independent and satisfy

$$W(x'|x)P_s(x) = W(x|x')P_s(x'). \tag{2.19}$$

Detailed balance is clearly a sufficient but *not* necessary condition for the stationary distribution $P(x, t) = P_s(x)$ to be a solution of the the Master equation.

In physics, the states are often in dynamical coordinates, i.e. a position and momentum pair, so the reverse transition balancing $x \to x'$ should be $x'^* \to x^*$, where the asterisk denotes velocity reversal. Instead of (2.19) one then has

$$W(x'|x)P_s(x) = W(x^*|x'^*)P_s(x'), \tag{2.20}$$

called the principle of *extended* detailed balance [16]. Under conditions of Markovity and stationarity, one can also show that extended detailed balance follows if the process is a probability distribution described by an ensemble of identical systems evolving deterministically, e.g. as in Hamiltonian mechanics.

A stochastic process is called *ergodic* if time averages in the limit of infinite time coincide with ensemble averages over the unique stationary distribution of the process. A sufficient condition for ergodicity is $\lim_{t\to\infty} P(x,t) = P_s(x)$, i.e. there is spontaneous relaxation towards equilibrium [15].

2.2.4 Langevin equations

The theory of stochastic processes began with the study of Brownian motion. The paradigm case is a colloidal particle in suspension, for example a very small plastic bead in water. The effect of the incessant collision of the water molecules with the bead is split into two parts. The first is a frictional force, due to the average (hydrodynamical) effect of the suspension. The second is a fluctuating force, since the bead is small enough to experience the random deviations from the hydrodynamical mean due to the atomic composition of the suspension. Lastly, there may also be an additional conservative force e.g. due to an applied electrical field. With x the position of the bead, Newton's second law is

$$m\ddot{x} = F_{\text{tot}}(x,\dot{x},t) = F_{\text{fric}}(\dot{x}) + F_{\text{cons}}(x,t) + F_{\text{fluct}}(t). \tag{2.21}$$

Such an equation is called a *Langevin equation*, and is explicitly written

$$m\ddot{x} = -\gamma\dot{x} - \frac{\partial V(x,t)}{\partial x} + \sqrt{B}\xi(t). \tag{2.22}$$

$\xi(t)$ is a realization of a stochastic variable $\boldsymbol{\xi}$ representing the random fluctuations. Since it is due to a physical process, it should have a finite correlation time, but often the noise is idealised as uncorrelated, with statistical properties

$$\langle\xi(t)\rangle = 0, \qquad \langle\xi(t)\xi(t')\rangle = \delta(t-t'). \tag{2.23}$$

Consistency with the Boltzmann-distribution in equilibrium requires $B = 2m\gamma k_B T$ [17]. In the strong friction limit $\gamma \gg m$, and with $\alpha = m\gamma$, one obtains the *overdamped* Langevin equation

$$\dot{x} = -\frac{m}{\alpha}\frac{\partial V(x,t)}{\partial x} + \sqrt{2\alpha k_B T}\xi(t). \tag{2.24}$$

This is an example from physics of a *stochastic differential equation* (SDE) on the form

$$\dot{x} = a(x,t) + b(x,t)\xi(t). \tag{2.25}$$

Adding a wildly fluctuating noise to a differential equation is not strictly rigorous, so the SDE is formally interpreted as an integral equation. The integral over the noise must then be defined, which can be done in different ways, leading to the Itô or the Stratonovich stochastic calculus. However, for $b(x,t) = \text{const.}$ these coincide [15]. It can then be shown that (2.25) with $b(x,t) = b^2$ leads to the FPE for $P(x,t)$ with $A(x,t) = a(x,t)$ and $B(x,t) = b^2$. For example, the Langevin equation

$$\dot{x} = -a + \sqrt{2D}\xi(t) \tag{2.26}$$

corresponds to the FPE with solution (2.15).

A discretization of (2.25) is

$$x_{t+1} = x_t + \Delta t\, a(x_t,t) + \Delta t\, b^2 W_t, \tag{2.27}$$

where W_t is sampled from $N(0,1)$. This thesis will use the Master Equation approach to model a stochastic process. Langevin equations will however be used as a notational short-hand to describe the models. The discretization (2.27) breaks detailed balance, which will be important for the simulation of fluctuation theorem. Therefore, instead the Metropolis-Hastings algorithm will be used.

2.3 The Metropolis-Hastings algorithm

Monte-Carlo algorithms refers to a broad set of techniques used to solve numerical or simulation problems by judicious use of random number generation.

One important problem in the simulation of stochastic processes is to draw samples x from a complicated but (up to a normalization factor) known probability density $\pi(x)$, for which the dynamical model that produces it is not explicitly known. The Monte Carlo approach is to construct a stochastic dynamics that will eventually (and hopefully quickly) produce the distribution $\pi(x)$. One way is to choose an ergodic Markov process satisfying detailed balance with the stationary distribution $\pi(x)$. In the Metropolis–Hastings algorithm [18] the transition probability $W(y|x)$ is split into a proposal probability $T(y|x)$ to propose the state change y given x, and an acceptance function $A(y|x)$ to accept the proposed change. Let

$$A(y|x) = \frac{S(x,y)}{T(y|x)\pi(x)}, \tag{2.28}$$

where $S(x,y)$ is any function that is symmetric in its arguments and such that $A(y|x) \leq 1$ for all y,x. Then $A(y|x)$ will yield the desired properties[3] [19]. Specifically, the Metropolis choice is

$$A(y|x) = \min\left\{1, \frac{T(x|y)\pi(y)}{T(y|x)\pi(x)}\right\}. \tag{2.29}$$

With this choice one need not know the normalization constant of the distribution π.

In statistical mechanical simulations, $\pi(x)$ is typically the Maxwell-Boltzmann distribution $P^{\mathrm{eq}}(x) = e^{-\beta V(x)}/Z$, where Z is the partition function that may be difficult to calculate if the potential V is complicated. The simulation algorithm can then be stated:

1. Choose x_0 and set $t = 0$
2. Generate a proposal x' from x_t according to $T(x'|x_t)$
3. Generate a random number $r \sim U(0,1)$
4. If $r < e^{-\beta(V(x')-V(x_t))}$, let $x_{t+1} := x'$. Else, let $x_{t+1} := x_t$.
5. Increment t, return to 2.

[3] One-line proof shows detailed balance: $W(y|x)\pi(x) = A(y|x)T(y|x)\pi(x) = S(x,y) = S(y,x) = W(x|y)\pi(y)$

Chapter 3

Stochastic thermodynamics and fluctuation theorems

3.1 Introduction

The origin of irreversibility is a fundamental question to the philosophy of physics, to which there is no universally accepted answer among philosophers and physicists [1, 20]. As became clear in the 1870s exchange between Loschmidt and Boltzmann, there is an apparent contradiction between the deterministic and time-reversal invariant dynamics at the microscopic scale (originally classical Hamilton dynamics, and later quantum dynamics), and the decidedly time-asymmetric and irreversible processes that we encounter at a human scale, and is described astoundingly well by thermodynamics. Through the better part of 150 years physicists have worked to connect the micro- and macroscopic scales through intermediate theories. One might say that statistical mechanics has expounded that connection, but in the process it has introduced probabilistic concepts that so far have not formally been reduced to mechanical concepts.

Nonetheless, in adopting a probabilistic approach there have been considerable progress toward describing processes far from equilibrium, especially in the last few decades. A number of conceptually differing perspectives on nonequilibrium physics have become unified in the mathematical framework of stochastic Markov processes [1]. From the perspective of *coarse graining*, if we only have access to meso- or macroscopic states of a system, then the observed dynamics on this spatial scale will be appear stochastic, and arguably Markovian on the relevant time scale. From the perspective of *interventionism*, it is (at least practically) impossible to shield a system from interactions with an environment that cannot be modelled in detail, which leads to a seemingly stochastic behaviour even if the physical object of interest could in principle be described microscopically. The combination of the theory of stochastic processes with thermal physics has resulted in stochastic thermodynamics [21, 22], a theory where thermal concepts such as work, heat, and entropy are generalized to individual trajectories of a stochastically evolving system. In this framework, it is possible to have spontaneous evolution from nonequilibrium to equilibrium ensembles, and to define an entropy that is, on average, increasing. As will be explained in this chapter, one can quantify the degree of irreversibility through an entropy production.

3.2 Fluctuation theorems

The fluctuation theorems are equalities holding arbitrarily far from equilibrium, involving the statistical properties of the micro-trajectories. Historically, several different fluctuation theorems were independently found, derived in different contexts. For example, the Jarzynski equality of 1998 [2] was derived in a classical Hamiltonian framework; the Crooks fluctuation theorem of 1999 in a stochastic dynamics set-up [23]. It was later realized that these and the several different other theorems could be unified in the framework of stochastic dynamics and derived by simple means [4]. The unifying idea behind the fluctuation theorems is that the relative likelihood of witnessing a sequence of events, to witnessing the same sequence "backwards" or in "reverse order", constitutes a quantitative measure of the irreversibility of the process.

3.2.1 General identities

Consider any stochastic process $\boldsymbol{X}$ (not necessarily Markovian) from time $t = 0$ to τ, which is called the *forward process*. The probability of a realization X, a particular sequence of states, is given by $\mathcal{P}^F[X]$. Denote the reverse sequence of states by $\overline{X} := \hat{T}X$, where $\hat{T}$ is a time-reversal operator. To the forward process, one constructs a corresponding "reverse process", with distribution $\mathcal{P}^R[Y]$ on the same state space. At this point there is no *a priori* requirements for what this process should be, and there are several choices that can lead to interpretable results, although most certainly won't [24, p. 39]. However, it is to be interpreted in some sense as a time-reversal of the forward stochastic dynamics.

The basic relation is the *definition* of a quantity $\sigma[X]$ called *entropy production* by

$$\frac{\mathcal{P}^F[X]}{\mathcal{P}^R[\overline{X}]} = e^{\sigma[X]}. \tag{3.1}$$

Following additional physical assumptions about the forward and reverse processes, $\sigma[X]$ can be related to a stochastic generalization of thermodynamics quantities (demonstrated in the next section). From the definition follows

$$\langle \sigma \rangle_F = D(\mathcal{P}^F[X] \,||\, \mathcal{P}^R[\overline{X}]) \geq 0, \tag{3.2}$$

where D is the KL divergence (2.7), and therefore $\langle \sigma \rangle_F \geq 0$ with equality if and only if $\mathcal{P}^F[X] = \mathcal{P}^R[\overline{X}]$. This also implies $\sigma[X] = 0$.

Eq. (3.1) leads to a general formula for path-ensemble averages of any path-functional $\mathcal{F}[X]$ [23]:

$$\begin{aligned}\left\langle \mathcal{F}e^{-\sigma}\right\rangle_F &= \int dX\,\mathcal{P}^F[X]\mathcal{F}[X]e^{-\sigma[X]} \\ &= \int d\overline{X}\,\mathcal{P}^R[\overline{X}]\overline{\mathcal{F}}[\overline{X}] \\ &= \langle \overline{\mathcal{F}}\rangle_R,\end{aligned} \tag{3.3}$$

where we use $|dX/d\overline{X}| = 1$ and define $\overline{\mathcal{F}}[\overline{X}] = \mathcal{F}[X]$. Now, if we choose $\mathcal{F} = 1$, we recover the *integral fluctuation theorem*,

$$\left\langle e^{-\sigma}\right\rangle = 1. \tag{3.4}$$

Defining $\rho[\overline{X}] = -\sigma[\overline{X}]$, and choosing $\mathcal{F}[X] = \delta(\sigma[X] - s)$, one finds

$$\frac{P^F(\sigma = s)}{P^R(\rho = -s)} = e^{+s}. \tag{3.5}$$

Suppose one had the symmetry[1] $\sigma[X] = -\sigma[\overline{X}]$. Then in a simplified notation,

$$\frac{P^F(\sigma)}{P^R(-\sigma)} = e^{+\sigma}. \tag{3.6}$$

This is known as a *transient* or *detailed fluctuation theorem.* (Note that the only assumptions made in these derivations are the definition (3.1), and that $\hat{T}$ is such that $|dX/d\overline{X}| = 1$.)

3.2.2 Stochastic derivation of the Jarzynski equality

For notational simplicity, the process is taken to be discrete in time, or rather we can think of strobe-lighting a continuous process. (This will not be an essential restriction since diffusion processes can be approximated by jump processes). For notational brevity, we let t take integer values from 0 to τ: $X = (x_0, x_1, \ldots, x_\tau)$ and $\overline{X} := \hat{T}X = (x^*_\tau, x^*_{\tau-1}, \ldots x^*_0)$. The asterisk denotes a velocity reversal.

A crucial restriction is to let both the forward and reverse processes be Markovian (we assume a Master equation stochastic dynamics),

$$\mathcal{P}^F[X] = P^F_{\text{start}}(x_0)\prod_{t=1}^{\tau} P^F(x_t|x_{t-1};t), \tag{3.7}$$

$$\mathcal{P}^R[\overline{X}] = P^R_{\text{start}}(x^*_\tau)\prod_{t=1}^{\tau} P^R(x^*_{t-1}|x^*_t;t). \tag{3.8}$$

[1] A more sophisticated statement of the symmetry underlying this fluctuation theorem is found in Ref. [25]

(For the reverse process, we count time t backwards, $\tau, \tau - 1, \ldots, 0$) Next, one must specify the initial and transition probabilities. One important class of non-equilibrium processes are those which spontaneously approach an equilibrium state (i.e. are ergodic Markov), but are being externally driven away from equilibrium. This is modelled by two features. First, by letting the time-dependence of the forward transition probabilities be due to a time varying control parameter λ_t,

$$P^F(x_t|x_{t-1};t) := P(x_t|x_{t-1},\lambda_t). \tag{3.9}$$

The specification of this parameter for the full duration is called a *protocol* and is denoted $\Lambda = (\lambda_1, \ldots, \lambda_\tau)$. Second, for any *fixed* value of λ one assumes that an extended detailed balance holds,

$$P(x_t|x_{t-1},\lambda_t)P^{\text{eq}}(x_{t-1};\lambda_t) = P(x_{t-1}^*|x_t^*,\lambda_t^*)P^{\text{eq}}(x_t;\lambda_t), \quad \forall : 0 < t \leq \tau. \tag{3.10}$$

It is important to stress that the process *as such* does not satisfy (extended) detailed balance (which would imply equilibrium) since the transition rates here are time-dependent through Λ. The stationary distribution P^{eq} is the Maxwell-Boltzmann distribution[2],

$$P^{\text{eq}}(x;\lambda) = \frac{1}{Z_\lambda}e^{-\beta E(x,\lambda)} = e^{\beta[F(\lambda)-\beta E(x,\lambda)]}. \tag{3.11}$$

This choice of stationary distribution is the assumption making the explicit contact with thermal physics.

Under the stated conditions, it is natural to consider the dynamics of the reverse process as differing from the forward process in being driven by the time-reversed protocol $\overline{\Lambda} = (\lambda_\tau^*, \ldots, \lambda_1^*)$. The conjugation denoted by the asterisk would for example mean the sign change of any magnetic field used as a control parameter.

$$P^R(x_{t-1}^*|x_t^*;\tau - t) := P(x_{t-1}^*|x_t^*,\lambda_t^*). \tag{3.12}$$

Next one considers the case where both the forward and reverse processes begin in equilibrium,

$$P_{\text{start}}^F(x) = P^{\text{eq}}(x;\lambda_0) \tag{3.13}$$

$$P_{\text{start}}^R(x) = P^{\text{eq}}(x;\lambda_f^*). \tag{3.14}$$

Note that λ_0 only appears in the forward process and λ_f^* (which we will also write as $\lambda_{\tau+1}^*$) only in the reverse process. However, the interpretation of the final result becomes clear if one imagines that after the state x_τ at parameter λ_τ is reached in the forward process, there is an additional final step in the protocol, $\lambda_\tau \to \lambda_f$. Importantly, any change of state due to this final step is not included in X. (One may do similarly for the reverse process with regards to λ_0^* for symmetry, but it will not matter.) Therefore, the forward process does not necessarily end in

[2]There is of course the possibility to consider a system which at different times is connected to different heat reservoirs, but we'll leave this complication aside.

equilibrium [3]. Occasionally we'll write the total protocols as $\Lambda^F = (\lambda_0, \Lambda, \lambda_f)$ and $\Lambda^R = (\lambda_f^*, \overline{\Lambda}, \lambda_0^*)$. The forward and reverse processes can be depicted:

$$\begin{array}{c}\text{Forward:}\\ \begin{array}{rccccccccccc} X: & x_0 & \rightarrow & x_1 & \rightarrow & x_2 & \cdots & x_{\tau-1} & \rightarrow & x_\tau & & \\ & \uparrow & & \uparrow & & \uparrow & \cdots & \uparrow & & \uparrow & & \\ \Lambda^F: & \lambda_0 & & \lambda_1 & & \lambda_2 & \cdots & \lambda_{\tau-1} & & \lambda_\tau & & \lambda_f \end{array}\\ \longrightarrow \text{time}\end{array} \tag{3.15}$$

$$\begin{array}{c}\text{Reverse:}\\ \begin{array}{cccccccccccl} x_0^* & \leftarrow & x_1^* & \leftarrow & x_2^* & \cdots & x_{\tau-1}^* & \leftarrow & x_\tau^* & & & :\overline{X} \\ & \nwarrow & & \nwarrow & & & & \nwarrow & & \nwarrow & & \\ \lambda_0^* & & \lambda_1^* & & \lambda_2^* & \cdots & \lambda_{\tau-1}^* & & \lambda_\tau^* & & \lambda_f^* & :\Lambda^R \end{array}\\ \longleftarrow \text{time}\end{array} \tag{3.16}$$

The process can be viewed as evolving in two separate steps [26]. First, a change in parameter $\lambda_{t-1} \rightarrow \lambda_t$ at fixed x_{t-1} leading to an energy change identified as stochastic work,

$$\delta W_t := E(x_{t-1}, \lambda_t) - E(x_{t-1}, \lambda_{t-1}). \tag{3.17}$$

Thereafter, there is the state transition $x_{t-1} \rightarrow x_t$ at fixed λ_t which incurs an energy change that is identified as stochastic heat,

$$\delta Q_t := E(x_t, \lambda_t) - E(x_{t-1}, \lambda_t). \tag{3.18}$$

From these assumptions $\sigma[X]$ can be physically interpreted. In terms of probabilities it is written:

$$\sigma[X] = \ln \frac{P^{\text{eq}}(x_0; \lambda_0)}{P^{\text{eq}}(x_\tau^*; \lambda_f^*)} + \sum_{t=1}^{\tau} \ln \frac{P(x_t | x_{t-1}, \lambda_t)}{P(x_{t-1}^* | x_t^*, \lambda_t^*)}. \tag{3.19}$$

The first term is the internal entropy change

$$\Delta S_i / k_B := [-\ln P^{\text{eq}}(x_\tau^*; \lambda_f^*)] - [-\ln P^{\text{eq}}(x_0; \lambda_0)], \tag{3.20}$$

since $-\ln P^{\text{eq}}(x, \lambda)$ is the Boltzmann entropy. With detailed balance and the definition (3.18), the second term (the sum) becomes

$$\Delta S_e[X] / k_B := -\beta \mathcal{Q}[X], \tag{3.21}$$

to be understood as the entropy change of the environment, where the total heat

$$\mathcal{Q}[X] := \sum_{t=1}^{\tau} \delta Q_t = \sum_{t=1}^{\tau} E(x_t, \lambda_t) - E(x_{t-1}, \lambda_t). \tag{3.22}$$

Hence

$$\sigma[X] = (\Delta S_i + \Delta S_e)/k_B =: \Delta S_{\text{tot}}[X]/k_B, \tag{3.23}$$

the total entropy production for the trajectory X. With the following additional and obvious extensions of thermodynamic quantities to microscopic trajectories,

$$\Delta F := F(\lambda_0) - F(\lambda_f), \tag{3.24}$$

$$\Delta E := E(x_\tau, \lambda_f) - E(x_0, \lambda_0), \tag{3.25}$$

$$\mathcal{W}[X] := \sum_{t=1}^{\tau+1} \delta W_t = \sum_{t=1}^{\tau+1} E(x_{t-1}, \lambda_t) - E(x_{t-1}, \lambda_{t-1}), \tag{3.26}$$

the thermodynamics expressions $\Delta E = \mathcal{Q} + \mathcal{W}$ and $\Delta F = \Delta E - T\Delta S_i$ hold true, and one can write

$$\sigma[X] = \beta(\Delta E - \Delta F - \mathcal{Q}) = \beta(\mathcal{W} - \Delta F) =: \beta\mathcal{W}_d, \tag{3.27}$$

which is the dissipative work, i.e. the excess work done on the system which does not add to the free energy.

The integral fluctuation theorem becomes

$$\left\langle e^{-\beta\mathcal{W}[X]} \right\rangle = e^{-\beta\Delta F}, \tag{3.28}$$

which is the *Jarzynski equality*, originally derived by Christopher Jarzynski [2]. It is interpreted as follows: Choose a forward protocol $\Lambda^F = (\lambda_0, \Lambda, \lambda_f)$, which drives the system away from equilibrium at λ_0, to a nonequilibrium state arbitrarily far from equilibrium, ending with λ_f. An amount of work $W[X^{(i)}]$ will be performed for the particular outcome $X^{(i)}$. Repeat this N times, yielding a distribution of work values. Take the work average according to (3.28); as N becomes large, this will yield (a simple function of) the free-energy difference between the start and final configuration. The surprising fact is that the statistical properties of work done in a nonequilibrium processes contain information about free energy which is a function of equilibrium. Note, that if one so wishes, one may consider the forward process to proceed to equilibrium at λ_f, but this does not affect the work, and thus neither the validity of (3.28).

Applying Jensen's inequality (2.9) on the form $\log(\langle x\rangle) \geq \langle\log(x)\rangle$ to the Jarzynski equality yields two expressions of the Second Law

$$\langle \mathcal{W}_d \rangle \geq 0, \quad \langle \Delta S_{\text{tot}} \rangle \geq 0. \tag{3.29}$$

One can check that the symmetry $\sigma[X|\Lambda] = -\sigma[\overline{X}|\overline{\Lambda}]$ holds, and hence the detailed fluctuation theorem (3.6) becomes

$$\frac{P^F(\Delta S_{\text{tot}})}{P^R(-\Delta S_{\text{tot}})} = e^{\Delta S_{\text{tot}}/k_B} \tag{3.30}$$

An entropy producing process in exponentially more likely than its reverse entropy absorbing process: a microscopic extension of the Second Law.

Specifically, since ΔF is not stochastic,

$$\frac{P^F(W)}{P^R(-W)} = e^{\beta(W-\Delta F)}. \tag{3.31}$$

The latter is known as *Crook's fluctuation theorem* [23], which clearly implies the Jarzynski equality.

3.3 Fluctuation theorems with feedback

In the previous section, the protocol was deterministic and independent of X. In the last ten years, there has been research extending fluctuation relations to processes where the protocol is dynamically determined by feedback with the system. This was first found for one feedback cycle by Sagawa and Ueda [10], and later independently generalized (acc. [27]) to repeated discrete feedback by Horowitz and Vaikuntanathan [28] and Fujitani and Suzuki [29]. Further directions include continuous feedback and extensions to non-Markovian process, as well as the analogous results for quantum systems, all of which go beyond the present thesis, however.

In the discrete repeated feedback formalism of Sagawa and Ueda [30], at specified times, measurements on the system are made and used to update the driving protocol until the next time of measurement. Specifically, the forward process includes a feedback mechanism, while the reverse does not. Section 3.3.1 summarizes this approach and some important related results. However, for the framing of problem of this thesis, it is more convenient to use a slightly different formalism and explicitly consider the possibility of feedback in the reverse process as well. This reformulation is done in section 3.3.2.

3.3.1 Forward feedback only

In Ref. [10], the feedback is executed by making measurements on the system at discrete times, and deciding the value of the parameter up to the next measurement based on all previous measurements. Let $\mathcal{X}(t)$ be a continuous process with $t \in [0, \tau]$. For notational brevity, suppose the measurements are made at $t = 0, 1, 2, \ldots, \tau$. The states at these times is denoted as before by $X = (x_0, x_1, \ldots, x_\tau)$ and the result of measurement on these states (except the last) by $Y = (y_0, y_1, \ldots, y_{\tau-1})$. The measurement y may be an estimation of the state x, or a coarser physical quantity determined by x; it is described by the measurement probability $P(y|x)$. At each measurement time t, the control parameter $\lambda_t = \lambda(Y_{t-1})$ is determined until the next time of measurement[3], where $Y_t = (y_0, \ldots, y_t)$. The

[3]One may imagine that λ_t determines not a fixed parameter, but a particular time-evolution of parameters to be actuated in the time interval leading up to the next measurement, see e.g. Ref. [31].

probability of of a particular realization of states and measurements in the forward process with feedback is

$$
\begin{aligned}
\mathcal{P}^F[X,Y] &= P^{\text{eq}}(x_0;\lambda_0)\prod_{t=1}^{\tau} P(x_t|x_{t-1},\lambda(Y_{t-1}))P(y_{t-1}|x_{t-1}) \\
&= \mathcal{P}_c[Y|X]\mathcal{P}^F[X|\Lambda(Y)],
\end{aligned} \tag{3.32}
$$

with the definition [30]

$$
\mathcal{P}_c[Y|X] := \prod_{t=1}^{\tau} P(y_{t-1}|x_{t-1}). \tag{3.33}
$$

Only in absence of feedback or for one single feedback loop does $\mathcal{P}_c[Y|X]$ equal $\mathcal{P}^F[Y|X](= \mathcal{P}^F[X,Y]/\mathcal{P}^F[Y])$, since X depends on Y via $\Lambda(Y)$. It nevertheless holds that $\int \mathcal{D}Y\, \mathcal{P}_c[Y|X] = 1$.

The introduction of feedback complicates the construction of a reverse process. A simple-minded time-reversal of the forward process with feedback will be acausal, because reverse feedback depends on future measurement [28, 32]. The approach following [10, 30] is to have no feedback in the reverse process. Instead, the protocol of the reverse process is the time-reversal of a protocol drawn from the ensemble of protocols generated by the forward process. Since the forward protocol is strictly determined by measurements, $\Lambda = \Lambda(Y)$ is drawn from the distribution $\mathcal{P}[Y]$, and therefore so is $\overline{\Lambda(Y)}$. The probability in the reverse process of a realization of states and the measurements determining the protocol is

$$
\mathcal{P}^R[\overline{X},Y] = \mathcal{P}^R[\overline{X}|\overline{\Lambda(Y)}]\mathcal{P}[Y]. \tag{3.34}
$$

Recalling that the mutual information[4]

$$
\mathcal{I}[X:Y] := \ln\frac{\mathcal{P}[Y|X]}{\mathcal{P}[Y]}, \tag{3.35}
$$

Ref. [30] makes a similar definition

$$
\mathcal{I}_c[X:Y] := \ln\frac{\mathcal{P}_c[Y|X]}{\mathcal{P}[Y]}, \tag{3.36}
$$

which is called *transfer entropy* [33]. $\mathcal{I}_c$ quantifies only the correlation of X and Y that is due to the measurement procedure; $\mathcal{I}$ additionally takes into account that they are correlated by the feedback control. Only in the case of one single

[4]It is more often $I = \langle\mathcal{I}\rangle$ rather than $\mathcal{I}$ that is called mutual information.

measurement and feedback update, or no feedback, does $\mathcal{I}_c = \mathcal{I}$. Actually, $\mathcal{I}_c$ is the more meaningful term [30]. One shows straightforwardly that it satisfies

$$\langle e^{-\mathcal{I}_c} \rangle = 1, \tag{3.37}$$

from which Jensen's inequality establishes $\langle \mathcal{I}_c \rangle \geq 0$. Defining $\sigma[X, Y]$ by

$$\frac{\mathcal{P}^F[X, Y]}{\mathcal{P}^R[\overline{X}, Y]} = e^{\sigma[X,Y]}, \tag{3.38}$$

one readily finds

$$\sigma[X, Y] = \sigma[X|Y] + \mathcal{I}_c[X, Y], \tag{3.39}$$

where $\sigma[X|Y]$ is essentially the same as (3.1),

$$\sigma[X|Y] := \ln \frac{\mathcal{P}^F[X|\Lambda(Y)]}{\mathcal{P}^R[\overline{X}|\overline{\Lambda(Y)}]}. \tag{3.40}$$

Given the assumptions made in Section 3.2.2 one obtains the *generalized Jarzynski equality* [10, 32]

$$\left\langle e^{\beta(\Delta F - W[X|Y] - \mathcal{I}_c[X:Y])} \right\rangle_F = 1. \tag{3.41}$$

The average is over all realizations of X and Y in the forward process. Jensen's inequality then gives

$$\langle \mathcal{W}_d \rangle \geq -\langle \mathcal{I}_c \rangle. \tag{3.42}$$

Since $\langle I_c \rangle \geq 0$, this means that it is possible for the dissipative work to be negative on average. This is the effect of the feedback on the system, which can use information gathered by measurement to control the system in a way that reliably extracts work from it. That is, the feedback works like a Maxwellian demon. The "rescue" of the Second Law comes from the consideration of the physical implementation of the feedback mechanism, and the necessary work associated with performing the measurements, and returning the feedback controller to its original state (memory erasure). See e.g. [11, 34] for reviews.

From (3.38) and (3.39) it follows that

$$D(\mathcal{P}^F \,||\, \mathcal{P}^R) = \langle W_d \rangle + \langle I_c \rangle \geq 0. \tag{3.43}$$

The case when equality holds is optimal in the sense that all information acquired through measurement has been "used" to extract work. This has been considered by Horowitz and Parrondo [27, 35]. The properties of the KL divergence imply

$$\mathcal{P}^F[X, Y] = \mathcal{P}^R[\overline{X}, Y], \quad \forall X, Y. \tag{3.44}$$

This means in particular that an "optimal" feedback is such that there are no state trajectories $\overline{X}$ generated in the reverse process, for which X is never generated by the forward process (except possibly for a set of measure zero). A process saturating the inequality (3.43) is called feedback reversible.

In Ref. [10] is defined a quantity γ called the "efficacy parameter",

$$\gamma := \int d\overline{Y}\mathcal{P}[\overline{Y}|\overline{\lambda(Y_{t-1})}], \tag{3.45}$$

for which it is shown that

$$\left\langle e^{\beta(\Delta F-\mathcal{W})}\right\rangle = \gamma. \tag{3.46}$$

$-\ln\gamma$ is a measure of correlation between $\mathcal{W}_d$ and $\mathcal{I}_c$, since if the their joint distribution is Gaussian,

$$\langle\mathcal{W}_d\mathcal{I}_c\rangle - \langle\mathcal{W}_d\rangle\langle\mathcal{I}_c\rangle = -\ln\gamma. \tag{3.47}$$

The larger γ, the more anticorrelated are $\mathcal{W}_d$ and $\mathcal{I}_c$, so that gaining more information lowers the dissipative work.

3.3.2 Forward and reverse feedback

Another formulation of the feedback set up, on which the work of this thesis is based, is to consider the protocol Λ itself to be a stochastic Markov process (the initial and final values λ_0 and λ_f remain fixed) [12]. The relation of this formalism to that of [30] outlined in the previous section is described in appendix A.1.

Schematically, we now have (compare (3.15) and (3.16) on page 17):

$$\begin{array}{c}\text{Forward:}\\ \begin{array}{cccccccccccc} X: & x_0 & \rightarrow & x_1 & \rightarrow & x_2 & \cdots & x_{\tau-1} & \rightarrow & x_\tau & & \\ & \uparrow & \searrow & \uparrow & \searrow & \uparrow & \cdots & \uparrow & \searrow & \uparrow & & \\ \Lambda^F: & \lambda_0 & \rightarrow & \lambda_1 & \rightarrow & \lambda_2 & \cdots & \lambda_{\tau-1} & \rightarrow & \lambda_\tau & & \lambda_f \end{array}\\ \longrightarrow \text{time}\end{array} \tag{3.48}$$

$$\begin{array}{c}\text{Reverse:}\\ \begin{array}{cccccccccccc} x_0 & \leftarrow & x_1^* & \leftarrow & x_2^* & \cdots & x_{\tau-1}^* & \leftarrow & x_\tau^* & & & :\overline{X} \\ & \nwarrow & \downarrow & \nwarrow & \downarrow & \cdots & \downarrow & \nwarrow & \downarrow & \nwarrow & & \\ \lambda_0^* & & \lambda_1^* & \leftarrow & \lambda_2^* & \cdots & \lambda_{\tau-1}^* & \leftarrow & \lambda_\tau^* & \leftarrow & \lambda_f^* & :\Lambda^R \end{array}\\ \longleftarrow \text{time}\end{array} \tag{3.49}$$

The joint probability of system trajectory and protocol is

$$\begin{aligned}\mathcal{P}^F[X,\Lambda] &:= P^{\text{eq}}(x_0;\lambda_0)\prod_{t=1}^{\tau}P(x_t|x_{t-1},\lambda_t)P(\lambda_t|x_{t-1},\lambda_{t-1})\\ &= \mathcal{P}^F[X|\Lambda]\mathcal{P}_c[\Lambda|X]\end{aligned} \tag{3.50}$$

$$\begin{aligned}\mathcal{P}^R[\overline{X},\overline{\Lambda}] &= P^{\text{eq}}(x_\tau;\lambda_f)\prod_{t=1}^{\tau}P(x_t^*|x_{t-1}^*,\lambda_t^*)\tilde{P}(\lambda_t^*|x_t^*,\lambda_{t+1}^*)\\ &= \mathcal{P}^R[\overline{X}|\overline{\Lambda}]\widetilde{\mathcal{P}}_c[\overline{\Lambda}|\overline{X}],\end{aligned} \tag{3.51}$$

where

$$\mathcal{P}_c[\Lambda|X] = \prod_{t=1}^{\tau} P(\lambda_t|x_{t-1}, \lambda_{t-1}), \text{ etc.} \tag{3.52}$$

Here, $P(\lambda_t|x_{t-1}, \lambda_{t-1})$ defines the feedback mechanism of the forward process, while $\tilde{P}(\lambda_t^*|x_t^*, \lambda_{t+1}^*)$ is a logically independent *causal* feedback in the reverse process[5]. With the definition

$$\frac{\mathcal{P}^F[X,\Lambda]}{\mathcal{P}^R[\overline{X},\overline{\Lambda}]} =: e^{\sigma[X,\Lambda]}, \tag{3.53}$$

the analogous derivation to 3.2.2 gives

$$\sigma[X,\Lambda] = \beta(\Delta F - \mathcal{W}[X|\Lambda]) - \mathcal{I}[X,\Lambda], \tag{3.54}$$

where

$$\begin{aligned}\mathcal{I}[X,\Lambda] &:= \ln \frac{\mathcal{P}_c[\Lambda|X]}{\widetilde{\mathcal{P}}_c[\overline{\Lambda}|\overline{X}]} \\ &= \sum_{t=1}^{\tau} \ln \frac{P(\lambda_t|x_{t-1}, \lambda_{t-1})}{\tilde{P}(\lambda_t^*|x_t^*, \lambda_{t+1}^*)} \\ &= \ln \frac{P(\lambda_1|x_0, \lambda_0)}{\tilde{P}(\lambda_\tau^*|x_\tau^*, \lambda_f^*)} + \sum_{t=1}^{\tau-1} \ln \frac{P(\lambda_{t+1}|x_t, \lambda_t)}{\tilde{P}(\lambda_t^*|x_t^*, \lambda_{t+1}^*)} \\ &=: J + \sum_{t=1}^{\tau-1} \delta I_t\end{aligned} \tag{3.55}$$

Several of the results of section 3.3.1 remain structurally identical. The generalized Jarzynski equality becomes

$$\left\langle e^{\beta(\Delta F - \mathcal{W}[X|\Lambda]) - \mathcal{I}[X,\Lambda]} \right\rangle_F = 1, \tag{3.56}$$

and it is therefore the case that

$$\beta\langle \mathcal{W}_d \rangle + \langle \mathcal{I} \rangle \geq 0. \tag{3.57}$$

The average is over the forward process, which does not make use of the feedback mechanism $\tilde{P}$ of the reverse process. Remarkably, $\tilde{P}$ only enters (3.56) and (3.57) through $\mathcal{I}$, yet they still hold for any choice of causal and Markovian reverse feedback! However, for $\mathcal{I}[X,\Lambda]$, there is no integral fluctuation theorem like (3.37), and it is possible that $\langle \mathcal{I} \rangle \leq 0$. Insofar as it is the logarithm of probabilities,

$$\mathcal{I}[X,\Lambda] = (-\ln \widetilde{\mathcal{P}}_c[\overline{\Lambda}|\overline{X}]) - (-\ln \mathcal{P}_c[\Lambda|X]), \tag{3.58}$$

it is an information theoretic quantity, and in a sense the difference in information gained by the reverse versus forward feedback. For ease of reference I will call

[5] On notation: the modifier $\widetilde{\ }$ is used rather than ^{R} or $\overline{\ }$ to stress that this is an independent feedback mechanism not necessarily interpreted as a time-reversal of another.

$\mathcal{I}[X, \Lambda]$ the "(feedback) information", δI_t is the "(feedback information) bulk term", and J the "(feedback information) jump or boundary term".

Chapter 4

Calculation of equilibrium distributions from nonequilibrium trajectories

4.1 The reweighting scheme

The Jarzynski equality has been extensively used in the determination of free energy differences [24]. But it can in theory be used more generally to obtain any equilibrium averages from a driven non-equilibrium process. Returning to Eq. (3.3) on page 15, with all the assumptions of section 3.2.2, one can recover the equilibrium distribution at λ_f by choosing $\mathcal{F}[X] = \delta(x - x_\tau)$ [26, 36],

$$\begin{aligned}\left\langle \delta(x - x_\tau) e^{\beta(\Delta F - W)} \right\rangle_F &= \langle \delta(x - x_\tau) \rangle_R \\ &= \int dx^*_\tau \, \delta(x - x_\tau) P^{\mathrm{eq}}(x_\tau; \lambda_f) \prod_{t=1}^{\tau} \int dx^*_{t-1} \, P(x^*_{t-1} | x^*_t, \lambda^*_t) \\ &= P^{\mathrm{eq}}(x; \lambda_f).\end{aligned} \tag{4.1}$$

Hence, even without ending in equilibrium, the statistical properties of the forward process contains the full information about equilibrium at the final parameter value. More generally, by choosing $\mathcal{F}[X] = A(x_\tau)$ one finds

$$\langle A(x) \rangle^{\mathrm{eq}}_{\lambda_f} = \left\langle A(x_\tau) e^{\beta(\Delta F - W[X])} \right\rangle_F = \frac{\left\langle A(x_\tau) e^{-\beta W[X]} \right\rangle_F}{\left\langle e^{-\beta W[X]} \right\rangle_F}, \tag{4.2}$$

where the free energy is replaced using the Jarzynski identity (3.28). Building on the feedback formalism introduced in section 3.3.2.

$$\begin{aligned}\langle A(x_\tau)\rangle_R &= \int \mathcal{D}\overline{X}\,\mathcal{D}\overline{\Lambda}\,A(x_\tau)P^{\mathrm{eq}}(x_\tau;\lambda_f)\prod_{t=1}^{\tau}P(x^*_{t-1}|x^*_t,\lambda^*_t)\tilde{P}(\lambda^*_t|x^*_t,\lambda^*_{t+1})\\ &= \int dx_\tau\,A(x_\tau)P^{\mathrm{eq}}(x_\tau;\lambda_f)\prod_{t=1}^{\tau}\int d\lambda^*_t\int dx^*_{t-1}\,P(x^*_{t-1}|x^*_t,\lambda^*_t)\tilde{P}(\lambda^*_t|x^*_t,\lambda^*_{t+1})\\ &= \int dx_\tau\,A(x_\tau)P^{\mathrm{eq}}(x_\tau;\lambda_f)\\ &= \langle A(x)\rangle^{\mathrm{eq}}_{\lambda_f}.\end{aligned} \tag{4.3}$$

The simplification from the second to the third line comes from taking the integrals in the order $x_0, \lambda_1, x_1, \lambda_2, \ldots, x_{\tau-1}, \lambda_\tau$. This also points out the fact that both causality and Markovity of the protocol is necessary to derive the end result (unless one makes very special assumptions about the stochastic dynamics). The extension of (4.2) to feedback is

$$\langle A(x)\rangle^{\mathrm{eq}}_{\lambda_f} = \frac{\left\langle A(x_\tau)e^{-\beta W[X,\Lambda]-I[X,\Lambda]}\right\rangle_F}{\left\langle e^{-\beta W[X,\Lambda]-I[X,\Lambda]}\right\rangle_F}. \tag{4.4}$$

Following [37], one can extend Eq. (4.2) to multiple *time-slices*. This holds without modification also for our case with feedback, Eq. (4.4). Given the full realization (X,Λ), imagine that the process had been cut off at time $t \le \tau$, whereupon the final jump in the protocol $\lambda_t \to \lambda_f$ is executed. The work and feedback information in this process is

$$W_t := \sum_{s=1}^{t}\delta W_s + E(x_t,\lambda_f) - E(x_t,\lambda_t), \tag{4.5}$$

$$I_t := J_t + \sum_{s=1}^{t-1}\delta I_s, \quad J_t := \ln\frac{P(\lambda_1|x_0,\lambda_0)}{P^R(\lambda^*_t|x^*_t,\lambda^*_f)}. \tag{4.6}$$

Choose any $\alpha_t(x)$ such that

$$\sum_{t=1}^{\tau}\alpha_t(x) = 1, \quad \forall\, x. \tag{4.7}$$

Then

$$\begin{aligned}
\langle A(x)\rangle_{\lambda_f}^{\mathrm{eq}} &= \left\langle \left(\sum_{t=1}^{\tau} \alpha_t(x)\right) A(x) \right\rangle_{\lambda_f}^{\mathrm{eq}} \\
&= \sum_{t=1}^{\tau} \left\langle \alpha_t(x_t) A(x_t) e^{\beta(\Delta F - W_t[X_t,\Lambda_t]) - I_t[X_t,\Lambda_t]} \right\rangle_F \\
&= \frac{\sum_{t=1}^{\tau} \left\langle \alpha_t(x_t) A(x_t) e^{-\beta W_t[X_t,\Lambda_t] - I_t[X_t,\Lambda_t]} \right\rangle_F}{\sum_{t=1}^{\tau} \left\langle \alpha_t(x_t) e^{-\beta W_t[X_t,\Lambda_t] - I_t[X_t,\Lambda_t]} \right\rangle_F}.
\end{aligned} \tag{4.8}$$

The final line gets rid of the free energy by requiring that the expression must give the correct average of 1 for $A(x) = 1$. The weighting factor is

$$w_t := e^{-\beta W_t[X_t,\Lambda_t] - I_t[X_t,\Lambda_t]}. \tag{4.9}$$

The naïve estimator for the equilibrium average using N trials in the forward process is

$$\langle A(x)\rangle_{\lambda_f}^{\mathrm{eq}} \approx \frac{\sum_{t=1}^{\tau} \sum_{i=1}^{N} \alpha_t(x_t^{(i)}) A(x_t^{(i)}) w_t^{(i)}}{\sum_{t=1}^{\tau} \sum_{i=1}^{N} \alpha_t(x_t^{(i)}) w_t^{(i)}}. \tag{4.10}$$

The simplest multiple time-slice weighting is $\alpha_t(x) = 1/\tau$, which gives equal importance to all times,

$$\langle A(x)\rangle_{\lambda_f}^{\mathrm{eq}} \approx \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{N} \sum_{i=1}^{N} A(x_t^{(i)}) u_t^{(i)}, \quad u_t^{(i)} := \frac{w_t^{(i)}}{\frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{N} \sum_{i=1}^{N} w_t^{(i)}}. \tag{4.11}$$

As Ref. [37] points out, the naive choice of weighting all time-slices equally by $\alpha = \text{const.}$, will be suboptimal if one expects only later time-slices to carry much information about the final equilibrium distribution. This is a possible issue for finite times, but any choice satisfying (4.7) will converge asymptotically in the limit of infinite realizations.

With or without feedback, to make the reweighting robust, one wants the weights w_t not to fluctuate too much; if a few realization carries a huge weight compared to all the others, it alone will dominate in the equilibrium estimation, and we do not expect it to be accurate unless it really makes use of a large number of realizations.

The weight is essentially $e^{-\sigma}$, disregarding the normalization constant eliminating ΔF. Its variance is

$$\mathrm{Var}\left(e^{-\sigma}\right) = \left\langle e^{-2\sigma}\right\rangle - \left\langle e^{-\sigma}\right\rangle^2 = \left\langle e^{-2\sigma}\right\rangle - 1, \tag{4.12}$$

where we use the integral fluctuation theorem $\langle e^{-\sigma}\rangle = 1$. If we can suppose the process is close to (feedback) reversibility, and σ is small, then we can Taylor expand,

$$1 = \left\langle e^{-\sigma}\right\rangle \approx 1 - \langle\sigma\rangle + \tfrac{1}{2}\langle\sigma^2\rangle \tag{4.13}$$

so that

$$\langle \sigma^2 \rangle \approx 2\langle \sigma \rangle. \tag{4.14}$$

Taylor expanding the variance,

$$\mathrm{Var}\left(e^{-\sigma}\right) \approx -2\langle \sigma \rangle + 2\left\langle \sigma^2 \right\rangle \approx 2\langle \sigma \rangle. \tag{4.15}$$

At least for small σ, minimizing the mean entropy production is a proxy for minimizing the weight variance.

4.2 Uses of the reweighting scheme in simulations

The reweighting scheme (4.2) has been used by Minh and Chodera [37] in a simple 1D model of a colloidal particle to calculate a number of equilibrium quantities. Recently, Yang et. al. [8] have employed a similar reweighting scheme to a 1D colloidal particle, and the simulation of a Lennard-Jones fluid.

One can think of multiple uses for the calculation of equilibrium averages by sampling from a non-equilibrium process:

- To improve time efficiency of MD or SM simulations of equilibrium, by pooling data attained after equilibrium is reached with data from the non-equilibrium transient.
- To temporarily modify the energy landscape (reduce energy barriers) in an equilibrium simulation to allow the system to sample more of phase space, and to compensate for this intervention [8].
- To calculate equilibrium quantities for a sequence of parameter values using a protocol traversing that sequence. (This could be compared to techniques such as (multi-)histogram reweighting that generate equilibrium averages for a range of parameters from simulation at a single parameter value.)

For all such purposes, one could imagine employing a protocol with feedback designed in some way that optimizes either procedure.

In this case, one will need to compute either $\mathcal{I}_c$ (3.36) or more generally $\mathcal{I}$ (3.55). The former case requires explicit knowledge of a marginal path probability $\mathcal{P}^F[\Lambda]$ which is not easily attained from simulation data. However, one might be able to find a reverse feedback that makes $\mathcal{I}[X, \Lambda]$ calculable, so that the reweighting can be done in practice.

4.3 Optimizing the reverse feedback

In any of the use cases suggested in section 4.2, one would need to design an appropriate forward feedback. If we then consider this feedback to be fixed, one can minimize $\langle \sigma \rangle$ with respect to the reverse feedback mechanism, i.e. optimizing towards feedback reversibility. Since the work only depends on the forward process, this is equivalent to the functional optimization of $\langle \mathcal{I} \rangle$ with respect to $\tilde{P}(\lambda'|x,\lambda)$. (For brevity we take $x = x^*$, $\lambda = \lambda^*$.) See appendix A.2 for the full calculations. The result is that

$$\tilde{P}(\lambda'|x,\lambda) = P(\lambda|x,\lambda') \cdot \frac{1}{\alpha(x,\lambda)} \sum_{t=1}^{\tau-1} P_t^F(x,\lambda'), \qquad (\lambda \neq \lambda_f), \tag{4.16}$$

where $\alpha(x,\lambda)$ normalizes the probability. For a given x, the optimal reverse transition probability $\lambda \to \lambda'$ is proportional to the forward transition $\lambda' \to \lambda$ weighted by the average over time to observe λ' in the forward process.

Although the transition probabilities of the state x are time independent (for fixed λ), there is nothing to prevent the forward or reverse feedback transition probabilities from being time dependent. In this case, the optimal reverse feedback is

$$\tilde{P}(\lambda'|x,\lambda;t) = P(\lambda|x,\lambda';t) \cdot \frac{P_t^F(x,\lambda')}{\alpha(x,\lambda,t)}. \tag{4.17}$$

Even if the forward feedback is not explicitly time-dependent, the reverse feedback should optimally still be, as seen by its dependence on P_t^F. For a given x, the likelihood of the reverse transition $\lambda \to \lambda'$ should be proportional to the forward transition $\lambda' \to \lambda$ weighted by the probability of λ' in the forward process.

Inserting the time-dependent optimum into the feedback information,

$$\mathcal{I} = J + \sum_{t=1}^{\tau-1} \delta I_t = \ln \frac{P(\lambda_1|x_0,\lambda_0;0)}{P(\lambda_\tau|x_\tau,\lambda_f;\tau)} + \sum_{t=1}^{\tau} (-\ln P_t^F(x_t,\lambda_t)) + \sum_{t=1}^{\tau} \ln \alpha(x_t,\lambda_{t+1},t) \tag{4.18}$$

However, given that the final feedback step does not enter $\mathcal{P}^F[X,\Lambda]$ it is another degree of freedom to optimize[1], with the result

$$P(\lambda_\tau|x_\tau,\lambda_f;\tau) = P_\tau^F(\lambda_\tau|x_\tau), \tag{4.19}$$

or, if we take the transition not to depend on x, as may be preferable for use in simulation,

$$P(\lambda_\tau|x_\tau,\lambda_f;\tau) = P_\tau^F(\lambda_\tau). \tag{4.20}$$

However, for numerical simulations, the expression (4.18) is difficult to calculate empirically due to all the joint distributions $P_t^F(x,\lambda)$ which would require a very large number of process realizations.

One special case that can be treated is when the forward feedback is deterministic. However, due to the boundary term, having the first transition $P(\lambda_1|x_0,\lambda_0;0)$ be deterministic (a delta function) is problematic since the discontinuity cannot always be cancelled by a discontinuity in $P(\lambda_\tau|x_\tau,\lambda_f;\tau)$ due to the difference in

[1] For $t = \tau$ one can instead skip the substitution (4.17) and optimize w.r.t. $\tilde{P}(\lambda'|x,\lambda;\tau)$ directly, with the same results.

arguments. So we must suppose $P(\lambda_1|x_0, \lambda_0; 0)$ to have some width (i.e. not be deterministic), and then $P(\lambda_\tau|x_\tau, \lambda_f; \tau)$ can be chosen as the optimum (4.20). For $0 < t < \tau$ we put

$$P(\lambda|x, \lambda'; t) = \delta(\lambda - \xi(\lambda'; x, t)), \tag{4.21}$$

where ξ is some function we assume to be invertible in its λ argument. The optimal reverse is then

$$\tilde{P}(\lambda'|x, \lambda; t) = \delta(\lambda' - \xi^{-1}(\lambda; x, t)), \tag{4.22}$$

which one can convince oneself of is also the only well defined option. Then

$$\delta I_t = -\ln\left|\frac{\partial \xi}{\partial \lambda}(\lambda_t; x_t, t)\right|. \tag{4.23}$$

4.4 Simulation: driven Brownian 1D-particle

4.4.1 Model

The model is an overdamped Langevin equation

$$\dot{x} = -\mu \frac{\partial V(x, \lambda(t))}{\partial x} + b\eta(t) \tag{4.24}$$

with a linear protocol

$$\lambda(t) = \lambda_0 + (\lambda_f - \lambda_0) \cdot t/\tau. \tag{4.25}$$

For this study, it is important to have the initial state x_0 sampled from the equilibrium distribution at λ_0. Therefore, it is particularly convenient to use a potential for which the Boltzmann-distribution can be sampled from exactly. A "trick" is to construct the equilibrium distribution first as a sum of Gaussians, and then construct the potential that gives rise to this distribution. Take a sum of N Gaussian distributions, with mean μ_i and variance σ_i^2, each weighted with a factor $0 < p_i < 1$, so that $\sum_i^N p_i = 1$. To sample from this distribution, one first chooses one of the Gaussians according to their probabilities p_i, and then generates $x \sim N(\mu_i, \sigma_i^2)$ for that Gaussian. The corresponding potential is

$$V(x) = -\frac{1}{\beta} \log\left(\sum_i \frac{p_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-(x-\mu_i)^2/2\sigma_i^2\right)\right). \tag{4.26}$$

The specific parameter values in the simulation is given by table 4.1. The protocol paramters used are $\lambda_0 = -1.5$, $\lambda_f = 1.5$. The potential for $\lambda = -1.5, 0, 1.5$ is plotted in fig. 4.1. The symmetry of this model implies $\Delta F = 0$. The Gaussian associated with $\lambda(t)$ will be referred to as the moving "trap".

The initial state of the particle is sampled exactly. Its evolution is simulated using the Metropolis-Hastings algorithm, and the protocol is updated linearly each time step. In the presented results, unless otherwise stated, the number of time

i	p_i	μ_i	σ_i
1	0.2	-1	0.6
2	0.2	$+1$	0.6
3	0.6	$\lambda(t)$	0.4

Table 4.1: Values of the potential parameters

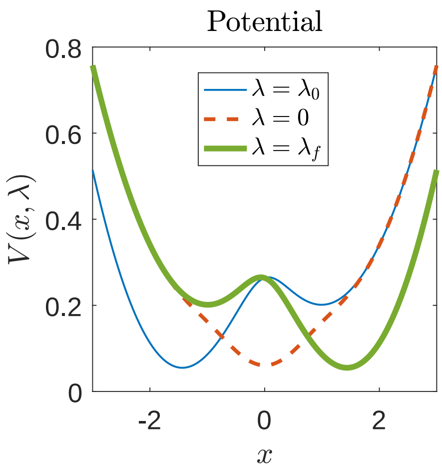


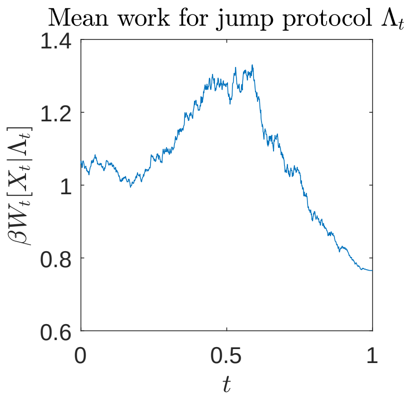
Figure 4.1: The potential plotted for the initial, mid-way, and final value of the protocol.

steps is $T = 1000$ (the total time τ is normalized to 1 in the graphs) and the number of trajectories is $N = 500$.

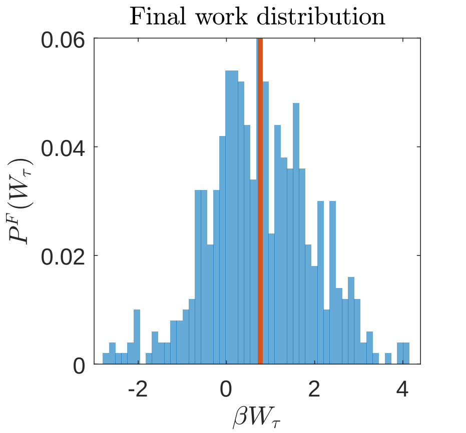
The calculation of the estimated equilibrium distribution is done using the estimator (4.11), with $A(x)$ an approximation of a delta function (a vertical bar with with $6/M$) for each x in the interval $[-3, 3]$ with $M = 100$ equidistant points.

4.4.2 Results

Three aspects of the simulation results are noteworthy: The distribution of work values, shown in fig. 4.2; the appearance of the equilibrium estimate compared to the exact distribution, fig. 4.4; and the distribution of weights, fig. 4.5.



(a) The mean total work (including the jump $\lambda_t \to \lambda_f$) plotted for each partial process $0 : t$. By contrast, the cumulative work from $0 : \tau$ grows approximately linearly (not plotted).



(b) Distribution of total work values (including the jump $\lambda_\tau \to \lambda_f$) for the full length process. Red lines indicates the mean.

Figure 4.2: Work values.

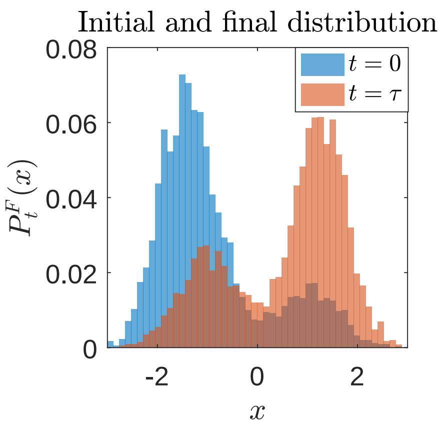


Figure 4.3: Comparison of the initial state (sampled from the exact distribution) and the empirical final distribution after full simulation (for $N = 4000$). There is visible lag at $t = \tau$, and not equilibrium (which would show a perfect mirroring of $t = 0$).

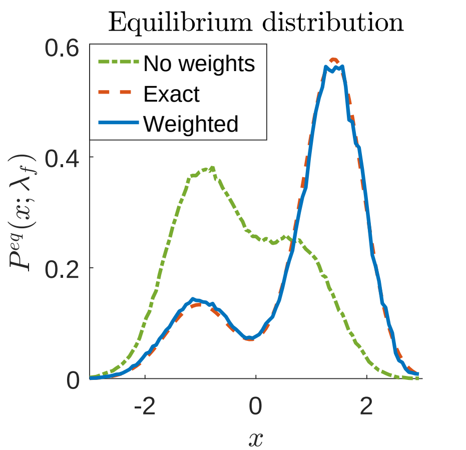


Figure 4.4: Comparison of the exact equilibrium distribution to the estimate using all time-slices ("Weighted"). The effect of the the reweighting is apparent by contrast to "No weights", which has $w_t^{(i)} = 1$.

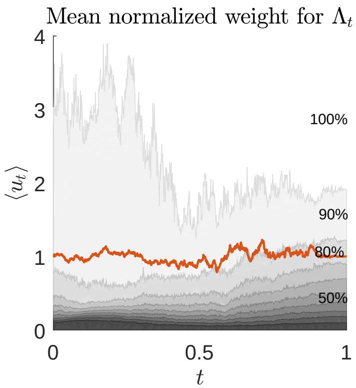


Figure 4.5: The average of normalized weights (4.11). The integral fluctuation theorem implies that the average weight should be exactly 1 at all times. The gray shaded regions plot the cumulative distribution function of normalized weights in 10% increments: about 20% of the weights are larger than the average.

4.4.3 Discussion

The mean work for time-slice t is largest for $t \approx 0.5$ (fig. 4.2a). This is expected since looking at the potential (fig. 4.1) at time $t = 0.5$, it has a minimum at $x = 0$, whereas $x = 0$ is a global maximum at parameter value λ_f. Therefore the jump $\lambda_t \to \lambda_f$ is costly in work. The work distribution fig. 4.2b is approximately Gaussian with a positive expected value. There is a considerable likelihood of extracting positive work from the system (i.e. W is negative), however the mean is positive as the statistical Second Law (3.29) demands (recall $\mathcal{W}_d = \mathcal{W}$ since $\Delta F = 0$).

The final state distribution, fig. 4.3 is close to, but visibly different from, the exact equilibrium distribution at λ_f (which would be a mirroring around $x = 0$ of the initial distribution). This shows the effect of lag (is verified by increasing N). The reweighting plotted in fig. 4.4 matches the exact distribution without systematic error. Here, all time slices are weighed equally (by α). Even using only the earliest 10% of time-slices, one obtains a fairly accurate equilibrium estimate, and the same for using only the last 10% (neither plotted). The reason is that for all parameter values, there is still access to all regions of phase space.

4.5 Simulation: feedback-driven Brownian 1D-particle

4.5.1 Model: deterministic feedback

The model is the same as in section 4.4 with parameters given by table 4.1, but using a different protocol involving feedback. The set-up is to chose a "reasonable" and relatively weak forward feedback, for which compensating reverse feedback can be found.

One such feedback mechanism is to have the position of the trap on average move from λ_0 to λ_f without excessive fluctuations, but to also let it follow the instantaneous position of the particle to some extent. This should lower the amount of work needed in the process. To design such a feedback, we imagine that the trap position is affected by two linear forces, one pulling it to the final value λ_f, and the other pulling it toward the particle's position $x(t)$. To this end we adopt an overdamped equation,

$$\dot{\lambda}(t) = -\kappa(\lambda - \lambda_f) - \nu(\lambda - x). \tag{4.27}$$

Since $\langle x \rangle \approx \langle \lambda \rangle$, the distance of $\langle \lambda(t) \rangle$ to λ_f will decreases exponentially in time.

The discretization of (4.27) used is

$$\lambda_{t+1} = \lambda_t - \Delta t\kappa(\lambda_t - \lambda_f) - \Delta t\nu(\lambda - x_t). \tag{4.28}$$

We define the new parameters $D = \Delta t\, \kappa\lambda_f$ and $p = \Delta t\, \nu$. Equation (4.28) can then be written

$$\lambda_{t+1} = \lambda_t(1 - p) + px_t + D(1 - \lambda_t/\lambda_f) =: \xi(\lambda_t; x_t). \tag{4.29}$$

We impose the restrictions that $0 \leq p \leq 1$ and $D \geq 0$. This feedback mechanism is then interpreted thus: first, the trap moves to a position at the fractional distance p between λ_t and x_t. Then, there is an added drift with strength D that moves the trap in the direction of λ_f. The simulation results are presented for a relatively week feedback with $p = D = 0.005$.

The forward transition probability is

$$P(\lambda_{t+1}|x_t, \lambda_t) = \delta(\lambda_{t+1} - \xi(\lambda_t; x_t)). \tag{4.30}$$

We have

$$\xi^{-1}(\lambda_{t+1}; x_t) = \frac{1}{1 - p - D/\lambda_f}(\lambda_{t+1} - px_t - D), \tag{4.31}$$

so with the reverse feedback described in section 4.3,

$$\delta I_t = \ln \alpha, \quad \alpha := |1 - p - D/\lambda_f|^{-1}. \tag{4.32}$$

For the boundary transitions we choose (for convenience)

$$P(\lambda_1|x_0, \lambda_0; 0) := \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-(\lambda_1 - x_0)^2/2\sigma_0^2\right], \tag{4.33}$$

with $\sigma_0 = 1$, and

$$\tilde{P}(\lambda_\tau | x_\tau, \lambda_f; \tau) = P^F_\tau(\lambda_\tau), \tag{4.34}$$

which is estimated empirically from by a histogram of the final distribution of λ.

4.5.2 Results: deterministic feedback

The average protocol is shown in fig. 4.6. The efficacy parameter diverges from 1 as is shown in fig. 4.7, which indicates the effect of feedback. Figure 4.8 shows how the average work becomes negative with time due to feedback, but the linearly growing information (due to $\ln \alpha$) keeps $\langle \sigma \rangle \geq 0$. The results of reweighting is shown in fig. 4.9. It uses the time slices from $t = 0.3$ to 1 to avoid a bias from the time period in which the effect of feedback has not manifested (as judged from fig. 4.7)—reweighting using only early time-slices would reproduce the results of section 4.4. However, it could *not be concluded* that the inclusion of $\mathcal{I}$ in the reweighting yields a systematically more accurate reproduction. Related to this is the observation that the integral fluctuation is severely violated, so that average weights for different time slices differ by many orders of magnitude. The reweighting here is not robust under different selections of time-slices.

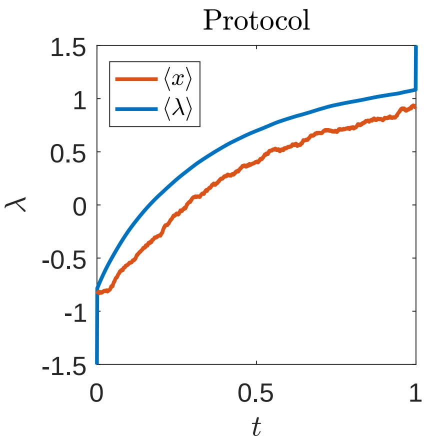


Figure 4.6: The mean protocol and particle position over time. Except the jumps at beginning and end, the protocol is clearly exponential.

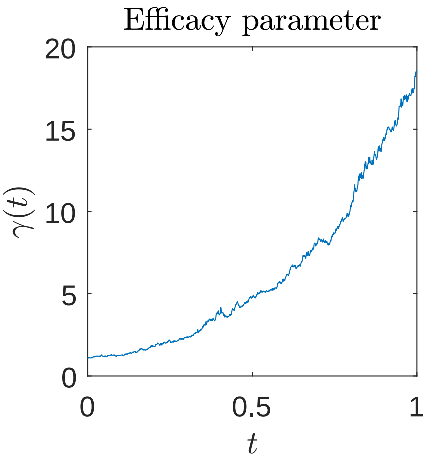


Figure 4.7: The divergence from $\gamma = 1$ quantifies the effect of feedback over time.

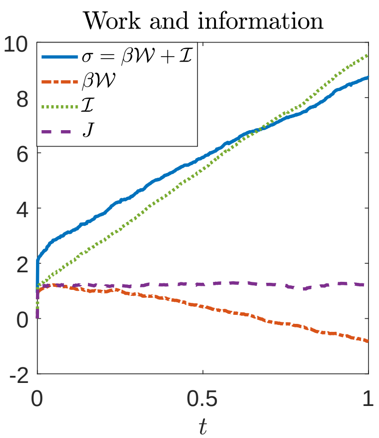


Figure 4.8: The mean work becomes negative due to feedback, but the entropy production grows linearly.

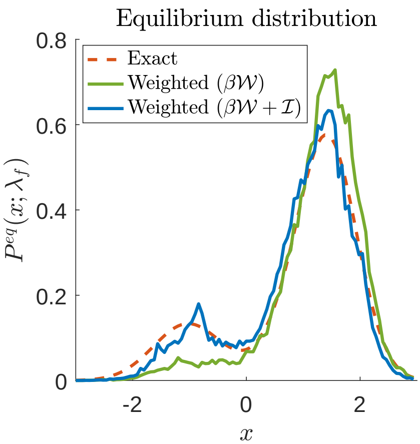


Figure 4.9: The reweighting is done using time-slices from $t = 0.3$ to avoid a bias from the initial time period where the effect of feedback is small.

4.5.3 Model: detailed-balanced feedback

In this model, the forward feedback is the same as (4.27) but with a Gaussian noise η.

$$\dot{\lambda}(t) = -\kappa(\lambda - \lambda_f) - \nu(\lambda - x) + b\eta. \tag{4.35}$$

One can Itô descretize this Langevin equation to find $P(\lambda_{t+1}|x_t, \lambda_t)$. Then, one can write $P(\lambda_{t+1}|x_t, \lambda_t) = P_d(\lambda_{t+1}|y_t, \lambda_t)P_m(y_t|x_t)$, where P_d is the deterministic feedback (4.30) and P_m is a measurement probability that is Gaussian. However, the difficulty is the choice of a reverse process, and the optimum found in section 4.3 does not lead to a readily calculable $\mathcal{I}$. Instead, we let choose reverse feedback such that instantaneous detailed balance holds for fixed x,

$$P(\lambda_{t+1}|x_t, \lambda_t; t)e^{-\beta U(x_t,\lambda_t)} = \tilde{P}(\lambda_t|x_t, \lambda_{t+1}; t)e^{-\beta U(x_t,\lambda_{t+1})}, \tag{4.36}$$

with the potential

$$U(x, \lambda) = \tfrac{\kappa}{2}(\lambda - \lambda_f)^2 + \tfrac{\nu}{2}(\lambda - x)^2. \tag{4.37}$$

It follows that the transitions $P(x_{t+1}, \lambda_{t+1}|x_t, \lambda_t)$ of the joint process (X, Λ) also satisfy detailed balance with energy $V(x, \lambda) + U(x, \lambda)$.

The time-evolution of the protocol is simulated using the Metropolis–Hastings algorithm to preserve detailed balance. However, (4.36) is only invoked for $0 < t < \tau$. For the boundary transitions, the same transitions as in section 4.5.1 are used, i.e. (4.33) and (4.34). The graphs presented are for $\kappa/\beta = 8$ and $\nu/\beta = 1$.

4.5.4 Results: detailed-balanced feedback

In many regards the results are similar to that of the deterministic feedback. However, in the average weights no not drop off toward zero, although they become increasingly dominated by large fluctuations with time (fig. 4.15). The reweighting using information is still not distinctly more accurate than using only work, with a typical result in fig. 4.14.

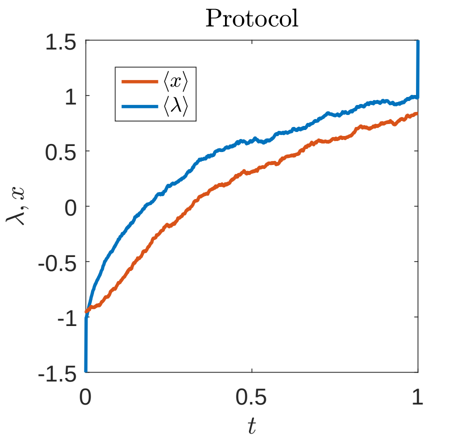


Figure 4.10: Mean protocol and particle position over time.

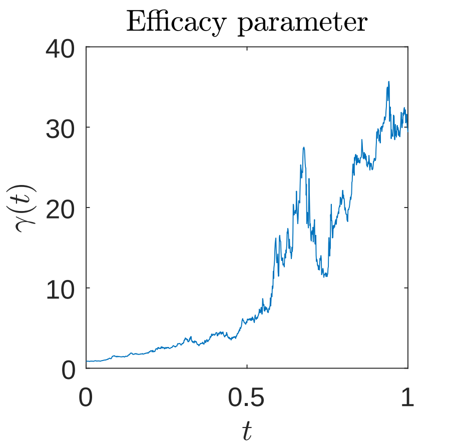


Figure 4.11: Efficacy paramter demonstrates effectiveness of feedback.

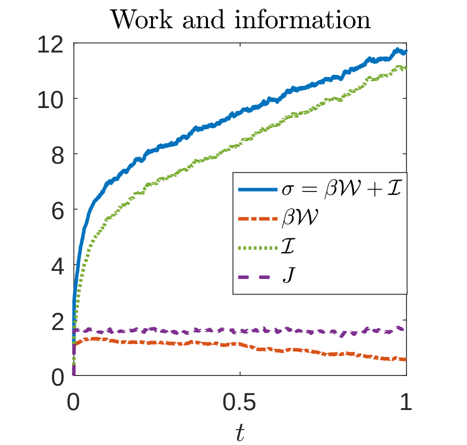


Figure 4.12: Mean work and information values

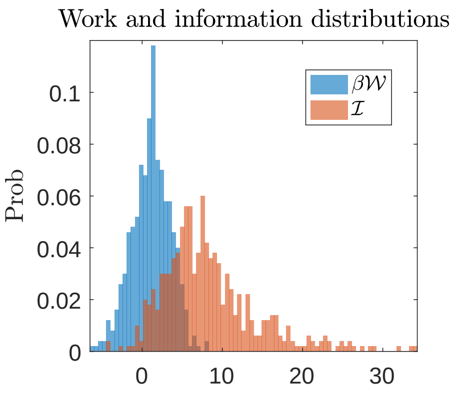


Figure 4.13: Distribution of work and feedback information at $t = \tau$.

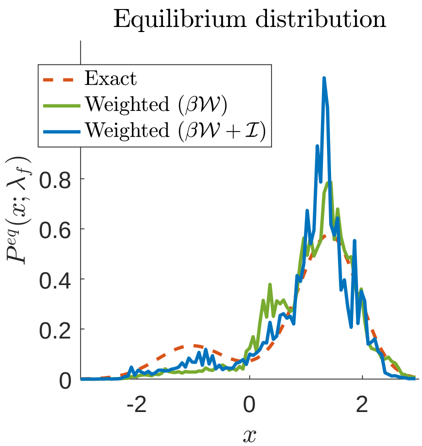


Figure 4.14: Reweighting results with time slices from $t = 0.3$.

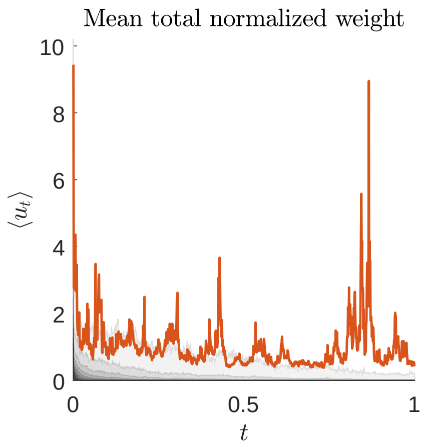


Figure 4.15: Average weight is on the order of 1 for all time slices, and is increasingly due to large fluctuations with time.

4.5.5 Discussion

Neither in the model with deterministic feedback nor that with detailed-balanced feedback was it found that reweighting using both work and feedback information eliminated the systematic error due to feedback in the reweighting using only work.

If $\langle \sigma \rangle$ grows in time, then fluctuations in σ must increase in time to preserve the integral fluctuation theorem. At the same time, the less σ is clustered around the mean at a given time, the fewer realization will contribute to the calculated average at that time, given the exponential weight. Therefore, for finite simulations making use of multiple time-slices, one would like $\langle \sigma \rangle$ not to grow too rapidly, i.e. one is close to feedback reversibility.

In the case of deterministic feedback, the integral fluctuation theorem is empirically not satisfied, with the result that the statistical weights w_t tend to zero with time. In the present case, both the magnitude and rate of change of $\langle I \rangle$ is much larger than $\beta\langle \mathcal{W} \rangle$. Therefore, one would expect that it is the fluctuations of $\mathcal{I}$ more so than $\mathcal{W}$ that would have to grow in time. Since $\delta I_t = \text{const.}$, all fluctuations must pertain to J_t. Theoretically, this is acceptable. But practically, it is appears difficult to choose $\tilde{P}(\lambda_\tau|x_\tau, \lambda_f; \tau)$ such that the integral fluctuation theorem is approximately satisfied for a finite number of realizations. Given that this process is not nearly feedback reversible, the "optimum" $\tilde{P}(\lambda_\tau|x_\tau, \lambda_f; \tau) = P_\tau(\lambda_\tau)$ need not be the best choice. For this reason, the J term was also calculated for a "non-optimal" choice, a Gaussian distribution for λ_τ with mean λ_f and a variance $\tilde{\sigma}_0^2$. However, this did not lead to significant improvement of the reweighting. Theoretically, one can expect the deterministic feedback in this model to be far from optimal: the reverse feedback has $\lambda_t \propto \lambda_{t+1} - px_t - D$, which would drive λ away from x exponentially and without bound. The reverse processes thus produces trajectories whose reversals are practically impossible in the forward process, and hence it cannot be close to feedback reversibility.

The failure of the detailed-balance choice of reverse feedback is a little more surprising, given that the calculation of $\mathcal{I}$ is very analogous to $\mathcal{W}$. In contrast to the deterministic feedback, the average weight is at least on the order of 1 for most times (fig. 4.15), although it becomes dominated by large fluctuations more quickly than in the case without feedback (cf. fig. 4.5). That $\langle \mathcal{I} \rangle$ grows faster than $\beta\langle \mathcal{W} \rangle$ decreases (fig. 4.12), and $\mathcal{I}$ has a wide distribution (fig. 4.13), supports the claim that the poor results are due being far from feedback reversibility.

Some aspects of the model may be criticized. The protocols fig. 4.6 and fig. 4.10 do not end at λ_f on average. This can be remedied by increasing the strength of D or κ, or adjusting the first linear force to $-\kappa(\lambda - \lambda_f - \delta)$, where δ is chosen so that $\langle \lambda_\tau \rangle = \lambda_f$. However, the distribution of λ_τ is still wide, and one may argue that it should preferentially be sharply focused around λ_f. Within the studied model, this can be done by making the D and p, or κ and ν, time-dependent, such that the D,κ increases in time relative to p,ν. If one uses all time-slices, the majority of data will nonetheless have large jumps $\lambda_t \to \lambda_f$. These suggested variations were indeed experimented with for a range of parameter values, without appreciable improvements to the reweighting result.

In conclusion, even for a simplistic model, naive choices of feedback cannot be expected to be compensated for using this reweighting technique. However, it is not at all established in general that the reweighting with scheme for feedback cannot be made practically useful. But the approach must likely be a very deliberate design of both forward and reverse feedback.

4.6 Suggested future directions

Given the difficulty in naive simulation, one direction would be to first study an analytically tractable model, for which one may (hopefully) find constraints on the feedback mechanisms that guarantee that σ is small. A candidate would be a Browninan 1D particle in a harmonic potential (a trap), where λ is the position of the trap to be moved from λ_0 to λ_f in a short enough time that the lag is appreciable. An optimal fixed protocol has been found in this case [38], which is linear but with jumps at the first and last time instances. One could then add a dynamical model to λ (e.g. deterministic or a Langevin equation) and optimize it to minimize the average work. Minimizing work fluctuations by feedback control for a Harmonic potential has been previously considered [39]. Here, one would also want to minimize the feedback information, with the reverse feedback as additional degrees of freedom.

For simulations without feedback, the reweighting worked well in the simple model studied, and its application to more complicated (and interesting) models could be pursued. A study of the pooling of the non-equilibrium transient and equilibrium measurement for a less ideal and more interesting model, seems to me a potentially very useful direction.

Chapter 5

Summary and Conclusions

The question of interest has been the possibility to simulate driven (classical) stochastic thermodynamic processes with and without feedback, and to recover from the simulated non-equilibrium trajectories an estimation of the equilibrium distribution at a particular parameter value of the potential.

Two components of established theory have been combined: (i) the method of reweighting non-equilibrium trajectories for a given fixed protocol, based on Jarzynski's equality, and (ii) the generalized Jarzynski equality, which adds to the entropy production an information term $\mathcal{I}_c$ due to measurements and feedback in the forward process.

For use in simulation the term $\mathcal{I}_c$ is not readily calculable. However, it could be replaced by a more general term $\mathcal{I}$, which involves the transition probabilities of two independent feedback mechanisms, one in the forward process, and one in the reverse. Since the reverse feedback mechanism is not involved in a simulation of the forward process, in the reweighting scheme it becomes a degree of freedom subject to a constraint: the protocol, viewed as a stochastic process due to feedback, must be causal and Markovian for both feedback mechanisms. The opening this provides is the possibility to choose a reverse feedback mechanism that makes $\mathcal{I}$ calculable from simulation data, and has the statistical properties that makes the reweighting scheme converge for a reasonable number of simulation trajectories.

One question, that still remains largely open, is to derive additional constraints to guide the choice of a reverse process. It was argued that for systems where the entropy production is small, minimizing $\langle \mathcal{I} \rangle$ for a given forward feedback mechanism, with respect to the reverse mechanism, will implicitly minimize the flucuations of the statistical weights, thereby optimizing the convergence properties of the reweighting scheme. A formally simple relationship between the reverse and forward transition rates for the stochastic protocol could be derived through this optimization. Nonetheless, $\mathcal{I}$ may still be difficult to estimate accurately from the empirical probability distributions involved. For the special case of deterministic feedback, however, the optimal reverse feedback lead to a simply calculable $\mathcal{I}$.

The reweighting method was applied to a model of a 1D Brownian particle in a simple potential, the position of whose minimum was controlled by the protocol. First, this was simulated without feedback. The equilibrium distribution at the final value of the protocol was calculated using the reweighting scheme, and was found to converge to the exact distribution, in agreement with previous studies. Then, two special cases of feedback were simulated: a deterministic feedback (except for an initial random step), and a stochastic feedback where the forward and reverse feedback transition rates satisfied a detailed balance condition. First, it was seen that the reweighting using only work no longer yields a correct equilibrium estimate when feedback is present. But it could not be demonstrated that the extended reweighting scheme incorporating the feedback information gives a consistently more accurate result. The reason for this, it was concluded, was that neither case comes close to being a reversible process, which is when the extended reweigthing can be expected to converge appropriately.

The reweighting by choice of a non-simulated reverse feedback might still prove workable in practice. However, it is likely the case that the design of both the forward and reverse feedback mechanisms must be made deliberately in a way that not only makes $\mathcal{I}$ readily calculable, but also optimizes the convergence properties of the reweighting method.

Appendix A

Detailed calculations

A.1 Stochastic protocol

Here, we rewrite the joint process of system and measurements to system and protocol. Starting from

$$\mathcal{P}^F[X,Y] = P^{\text{eq}}(x_0;\lambda_0)\prod_{t=1}^{\tau} P(x_t|x_{t-1},\lambda(Y_{t-1}))\prod_{t=1}^{\tau} P(y_{t-1}|x_{t-1}) \tag{A.1}$$

we let $\lambda_t = \lambda(Y_{t-1})$ be a new random variable since Y_{t-1} is random. Then define

$$\mathcal{P}[\Lambda|Y] = \prod_{t=1}^{\tau} \delta(\lambda_t - \lambda(Y_{t-1})) \tag{A.2}$$

We suppose that the function $\lambda(Y_{t-1}) = \lambda'(y_{t-1}, \lambda(Y_{t-2}))$, i.e. $\lambda_t = \lambda'(y_{t-1}, \lambda_{t-1})$, for some function λ'. The joint distribution of system trajectory and protocol is related to the joint distribution of system trajectory and measurements by

$$\begin{aligned}
\mathcal{P}^F[X,\Lambda] &:= \int \mathcal{D}Y\,\mathcal{P}[\Lambda|Y]\mathcal{P}[X,Y] \\
&= P^{\text{eq}}(x_0;\lambda_0)\prod_{t=1}^{\tau} P(x_t|x_{t-1},\lambda_t)\prod_{t=1}^{\tau}\int dy_{t-1}P(y_{t-1}|x_{t-1})\delta(\lambda_t - \lambda(Y_{t-1})) \\
&= P^{\text{eq}}(x_0;\lambda_0)\prod_{t=1}^{\tau} P(x_t|x_{t-1},\lambda_t)P(\lambda_t|x_{t-1},\lambda_{t-1})
\end{aligned} \tag{A.3}$$

where we have defined

$$P(\lambda_t|x_{t-1},\lambda_{t-1}) = \int dy_{t-1}\, P(y_{t-1}|x_{t-1})\delta(\lambda_t - \lambda'(y_{t-1},\lambda_{t-1})). \tag{A.4}$$

It is normalized to unity and is therefore a proper transition probability. For fixed X, Λ is a Markov process, and I will call this feedback Markovian. This is not the same as Ref. [30] means by Markovian feedback, which is that $\lambda(Y_{t-1}) = \lambda(y_{t-1})$.

A.2 Optimal reverse feedback calculations

A.2.1 Homogeneous feedback process

We minimize the variation of

$$\langle \mathcal{I} \rangle = \int \mathcal{D}X\, \mathcal{D}\Lambda\, \mathcal{P}^F[X,\Lambda] \ln \frac{\mathcal{P}_c[\Lambda|X]}{\widetilde{\mathcal{P}}_c[\overline{\Lambda}|\overline{X}]} \tag{A.5}$$

with respect to $\tilde{P}(\lambda'|x,\lambda)$ under the normalization constraint

$$\int d\lambda'\, \tilde{P}(\lambda'|x,\lambda) = 1, \quad \forall\, x,\lambda. \tag{A.6}$$

Using the method of Lagrange multipliers, we set the variation of

$$\mathcal{L} := \int \mathcal{D}X\, \mathcal{D}\Lambda\, \mathcal{P}^F[X,\Lambda] \ln \frac{\mathcal{P}_c[\Lambda|X]}{\widetilde{\mathcal{P}}_c[\overline{\Lambda}|\overline{X}]} - \int dy\, d\mu\, \alpha(y,\mu) \left(1 - \int d\mu'\, \tilde{P}(\mu'|y,\mu)\right) \tag{A.7}$$

with respect to the three-variable function $\tilde{P}$ to zero,

$$\frac{\delta \mathcal{L}[\tilde{P}]}{\delta \tilde{P}(\lambda'|x,\lambda)} = 0. \tag{A.8}$$

The variation of the constraint term is trivially $\alpha(x,\lambda)$. The variation of $-\mathcal{I}$ is

$$\begin{aligned} \frac{\delta \ln \widetilde{\mathcal{P}}_c[\overline{\Lambda}|\overline{X}]}{\delta \tilde{P}(\lambda'|,x,\lambda)} &= \sum_{t=1}^{\tau} \frac{\delta \ln \tilde{P}(\lambda_t|x_t,\lambda_{t+1})}{\delta \tilde{P}(\lambda'|x,\lambda)} \\ &= \frac{1}{\tilde{P}(\lambda'|x,\lambda)} \sum_{t=1}^{\tau} \delta(\lambda_t - \lambda')\delta(x_t - x)\delta(\lambda_{t+1} - \lambda). \end{aligned} \tag{A.9}$$

Define

$$f(\lambda',x,\lambda) = \sum_{t=1}^{\tau} \langle \delta(\lambda_t - \lambda')\delta(x_t - x)\delta(\lambda_{t+1} - \lambda) \rangle_F\,. \tag{A.10}$$

Then condition (A.8) becomes

$$-\frac{f(\lambda',x,\lambda)}{\tilde{P}(\lambda'|x,\lambda)} + \alpha(x,\lambda) = 0. \tag{A.11}$$

The normalization (A.6) implies

$$\alpha(x,\lambda) = \int d\lambda'\, f(\lambda',x,\lambda), \tag{A.12}$$

so that

$$\tilde{P}(\lambda'|x,\lambda) = \frac{f(\lambda',x,\lambda)}{\int d\lambda'\, f(\lambda',x,\lambda)}. \tag{A.13}$$

From the definition (A.10)

$$\begin{aligned} f(\lambda',x,\lambda) &= \sum_{t=1}^{\tau} P^F(\lambda_t = \lambda', x_t = x, \lambda_{t+1} = \lambda) \\ &= \sum_{t=1}^{\tau} P^F(\lambda_{t+1} = \lambda | x_t = x, \lambda_t = \lambda') P^F(x_t = x, \lambda_t = \lambda') \\ &= \delta(\lambda_f - \lambda) P^F_\tau(x,\lambda') + P(\lambda|x,\lambda') \sum_{t=1}^{\tau-1} P^F_t(x,\lambda'), \end{aligned} \tag{A.14}$$

where

$$P^F_t(x,\lambda) = \int \mathcal{D}X_t\, \mathcal{D}\Lambda_t\, \mathcal{P}^F[X_t,\Lambda_t]\delta(x_t - x)\delta(\lambda_t - \lambda) \tag{A.15}$$

is the probability that the joint process is in state (x,λ) at time t.

Assuming $\lambda \neq \lambda_f$ we find

$$\tilde{P}(\lambda'|x,\lambda) = P(\lambda|x,\lambda') \cdot \frac{1}{\alpha(x,\lambda)} \sum_{t=1}^{\tau-1} P^F_t(x,\lambda'). \tag{A.16}$$

A.2.2 Nonhomogeneous feedback process

Now, suppose that both forward and reverse feedback transition probabilities are time-dependent. Then $\tilde{P}(\lambda'|x,\lambda;s)$ is a four-variable function and we must include the variation of s (time) in the functional optimization. Similarly we define

$$\begin{aligned} f(\lambda',x,\lambda,s) &:= \sum_{t=1}^{\tau} \langle \delta(\lambda_t - \lambda')\delta(x_t - x)\delta(\lambda_{t+1} - \lambda)\delta_{ts} \rangle_F \\ &= P(\lambda|x,\lambda';s) P^F_s(x,\lambda') \end{aligned} \tag{A.17}$$

and

$$\begin{aligned} \alpha(x,\lambda,s) &:= \int d\lambda'\, f(\lambda',x,\lambda,s) \\ &= \int d\lambda'\, P(\lambda|x,\lambda';s) P^F_s(x,\lambda') \end{aligned} \tag{A.18}$$

so again it's found that $\tilde{P} = f/\alpha$:

$$\tilde{P}(\lambda'|x, \lambda; t) = P(\lambda|x, \lambda'; t) \cdot \frac{P_t^F(x, \lambda')}{\alpha(x, \lambda, t)}. \tag{A.19}$$

Using the optimal feedback (A.19) we can find the corresponding expression for $\mathcal{I}$. The bulk term is

$$\delta I_t := \ln \frac{P(\lambda_{t+1}|x_t, \lambda_t; t)}{\tilde{P}(\lambda_t|x_t, \lambda_{t+1}; t)} = -\ln P_t^F(x_t, \lambda_t) + \ln \alpha(x_t, \lambda_{t+1}, t), \tag{A.20}$$

and the boundary term

$$\begin{aligned} J &:= \ln \frac{P(\lambda_1|x_0, \lambda_0; 0)}{\tilde{P}(\lambda_\tau|x_\tau, \lambda_f; \tau)} \\ &= \ln \frac{P(\lambda_1|x_0, \lambda_0; 0)}{P(\lambda_\tau|x_\tau, \lambda_f; \tau)} - \ln P_\tau^F(x_\tau, \lambda_\tau) + \ln \alpha(x_\tau, \lambda_f, \tau). \end{aligned} \tag{A.21}$$

Actually, the second line offers nothing new over the definition; because the forward process is only executed up to time τ, the forward transition probability $P(\lambda|x, \lambda'; \tau)$ is never used (i.e. not defined). So whether one takes the transition P or $\tilde{P}$ at τ to be independent is a matter of choice. Proceeding with the latter,

$$\mathcal{I} = J + \sum_{t=1}^{\tau-1} \delta I_t = \ln \frac{P(\lambda_1|x_0, \lambda_0; 0)}{P(\lambda_\tau|x_\tau, \lambda_f; \tau)} + \sum_{t=1}^{\tau}(-\ln P_t^F(x_t, \lambda_t)) + \sum_{t=1}^{\tau} \ln \alpha(x_t, \lambda_{t+1}, t). \tag{A.22}$$

Logically, given that $P(\lambda_\tau|x_\tau, \lambda_f; \tau)$ does not enter $\mathcal{P}^F[X, \Lambda]$, we can again optimize with respect to this function, or to be clear, with respect to $g(x, \lambda) := P(\lambda|x, \lambda_f; \tau)$. The variational problem follows the exact same step as before with the result

$$g(x, \lambda) = \frac{P_\tau^F(x, \lambda)}{\int d\lambda\, P_\tau^F(x, \lambda)} = \frac{P_\tau^F(x, \lambda)}{P_\tau^F(x)} = P_\tau^F(\lambda|x). \tag{A.23}$$

If one takes there to be no dependence on x, $g(x, \lambda) = g(\lambda)$, it is similarly found

$$g(\lambda) = P_\tau^F(\lambda). \tag{A.24}$$

A.2.3 Deterministic feedback

Consider now the special case when the forward feedback is deterministic,

$$P(\lambda|x, \lambda') = \delta(\lambda - \xi(\lambda'; x)). \tag{A.25}$$

Then

$$\begin{aligned}
\tilde{P}(\lambda'|x,\lambda) &= \frac{\delta(\lambda - \xi(\lambda';x)) \sum_{t=1}^{\tau-1} P_t^F(x,\lambda')}{\int d\lambda' \, \delta(\lambda - \xi(\lambda';x)) \sum_{t=1}^{\tau-1} P_t^F(x,\lambda')} \\
&= \delta(\lambda' - \xi^{-1}(\lambda;x)) \frac{\sum_{t=1}^{\tau-1} P_t^F(x,\lambda')}{\sum_{t=1}^{\tau-1} P_t^F(x,\xi^{-1}(\lambda;x))} \\
&= \delta(\lambda' - \xi^{-1}(\lambda;x))
\end{aligned} \tag{A.26}$$

where $\xi^{-1}(\lambda;x)$ is the inverse of $\xi(\lambda;x)$ with respect to λ if we assume it exists. Deterministic feedback can however not be applied to the transition probabilities in J which would diverge.

For non-homogenous feedback, one can have deterministic feedback for $0 < t < \tau$,

$$\begin{aligned}
\tilde{P}(\lambda'|x,\lambda;t) &= \delta(\lambda - \xi(\lambda';x,t)) \cdot \frac{P_t^F(x,\lambda')}{\int d\lambda' \, \delta(\lambda - \xi(\lambda';x,t)) P_t^F(x,\lambda')} \\
&= \delta(\lambda' - \xi^{-1}(\lambda;x,t)).
\end{aligned} \tag{A.27}$$

Using the rule $\delta(g(x)) = \sum_{x_0 \text{ zeros of } g} |g'(x_0)|^{-1} \delta(x - x_0)$, the bulk term is

$$\delta I_t = \ln \frac{\delta(\lambda_{t+1} - \xi(\lambda_t;x_t,t))}{\delta(\lambda_t - \xi^{-1}(\lambda_{t+1};x_t,t))} = -\ln \left| \frac{\partial \xi}{\partial \lambda}(\lambda_t;x_t,t) \right|. \tag{A.28}$$

Bibliography

[1] J. Uffink, *Compendium of the foundations of classical statistical physics* (North-Holland, 2007) pp. 923–1074.

[2] C. Jarzynski, *A nonequilibrium equality for free energy differences*, **78**, 0 (1997), 9610209.

[3] C. Jarzynski, *Nonequilibrium work theorem for a system strongly coupled to a thermal environment*, Journal of Statistical Mechanics: Theory and Experiment **2004**, P09005 (2004), 0407340v2.

[4] U. Seifert, *Stochastic thermodynamics, fluctuation theorems and molecular machines.*, Reports on progress in physics. Physical Society (Great Britain) **75**, 126001 (2012), arXiv:1205.4176v1.

[5] D. A. Sivak, J. D. Chodera and G. E. Crooks, *Using nonequilibrium fluctuation theorems to understand and correct errors in equilibrium and nonequilibrium simulations of discrete Langevin dynamics*, Physical Review X **3**, 1 (2013), 1107.2967.

[6] A. J. Ballard and C. Jarzynski, *Replica exchange with nonequilibrium switches.*, Proceedings of the National Academy of Sciences of the United States of America **106**, 12224 (2009), 0702048.

[7] J. P. Nilmeier *et al.*, *Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation*, Proceedings of the National Academy of Sciences **108**, E1009 (2011).

[8] C. Yang *et al.*, *Equilibrium sampling by reweighting nonequilibrium simulation trajectories*, Physical Review E **93**, 1 (2016), 1512.08616.

[9] D. D. L. Minh and J. D. Chodera, *Optimal estimators and asymptotic variances for nonequilibrium path-ensemble averages*, Journal of Chemical Physics **131** (2009), 0907.4776.

[10] T. Sagawa and M. Ueda, *Generalized Jarzynski equality under nonequilibrium feedback control*, Physical Review Letters **104**, 1 (2010), 0907.4914.

[11] J. M. R. Parrondo, J. M. Horowitz and T. Sagawa, *Thermodynamics of information*, Nature Physics **11**, 131 (2015).

[12] J. Lidmar, personal communication, 2017.

[13] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal **27**, 379 (1948).

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second edi ed. (Wiley Interscience, 2006).

[15] C. Gardiner, *Stochastic Methods*, 4 ed. (Springer-Verlag Berlin Heidelberg, 2009).

[16] N. van Kampen, *Stochastic Processes in Physics and Chemistry*, Third edit ed. (Elsevier B.V., 2007).

[17] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).

[18] W. K. Hastings, *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika **57**, 97 (1970).

[19] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, 2008).

[20] L. Sklar, *Physics and Chance* (Cambridge University Press, 1993).

[21] T. Tomé and M. J. De Oliveira, *Stochastic approach to equilibrium and nonequilibrium thermodynamics*, Physical Review E **91**, 1 (2015), `arXiv:1503.04342v1`.

[22] C. Bechinger, F. Sciortino and P. Ziherl, *Physics of Complex Colloids,* International School of Physics Enrico Fermi (IOS Press, 2013).

[23] G. E. Crooks, *Entropy production fluctuation theorems and the nonequilibrium work relation for free energy differences*, Physical Review E **60**, 2721 (1999), `9901352`.

[24] R. Klages, W. Just and C. Jarzynski (Eds.), *Nonequilibrium Statistical Physics of Small Systems* (Wiley-VCH Verlag, 2013).

[25] B. H. Shargel, *The Measure-theoretic Identity Underlying Transient Fluctuation Theorems*, Journal of Physics A: Mathematical and Theoretical **43**, 135002 (2010), `1001.4128`.

[26] G. E. Crooks, *Path-ensemble averages in systems driven far from equilibrium*, Physical Review E **61**, 2361 (2000), `9908420`.

[27] J. M. Horowitz and J. M. R. Parrondo, *Thermodynamic reversibility in feedback processes*, EPL (Europhysics Letters) **95**, 10005 (2011), `1104.0332`.

[28] J. M. Horowitz and S. Vaikuntanathan, *Nonequilibrium detailed fluctuation theorem for repeated discrete feedback*, Physical Review E **82**, 1 (2010), 1011.4273.

[29] Y. Fujitani and H. Suzuki, *Jarzynski equality modified in the linear feedback system*, Journal of the Physical Society of Japan **79**, 1 (2010).

[30] T. Sagawa and M. Ueda, *Nonequilibrium Thermodynamics of Feedback Control*, Control **021104**, 1 (2011), arXiv:1105.3262v1.

[31] A. Kundu, *Nonequilibrium fluctuation theorem for systems under discrete and continuous feedback control*, Physical Review E **86**, 1 (2012).

[32] M. Ponmurugan, *Generalized detailed fluctuation theorem under nonequilibrium feedback control*, Physical Review E **82**, 1 (2010), 1005.4311.

[33] T. Schreiber, *Measuring information transfer*, Physical Review Letters **85**, 461 (2000), 0001042.

[34] T. Sagawa, *Thermodynamics of Information Processing in Small Systems*, PhD thesis, University of Tokyo, 2013.

[35] J. M. Horowitz and J. M. R. Parrondo, *Designing optimal discrete-feedback thermodynamic engines*, New Journal of Physics **13** (2011), 1110.6808.

[36] E. Boksenbojm, B. Wynants and C. Jarzynski, *Nonequilibrium thermodynamics at the microscale: Work relations and the second law*, Physica A: Statistical Mechanics and its Applications **389**, 4406 (2010), 1002.1230.

[37] D. D. L. Minh and J. D. Chodera, *Estimating equilibrium ensemble averages using multiple time slices from driven nonequilibrium processes: Theory and application to free energies, moments, and thermodynamic length in single-molecule pulling experiments*, Journal of Chemical Physics **134** (2011), 1010.5467.

[38] T. Schmiedl and U. Seifert, *Optimal finite-time processes in stochastic thermodynamics*, Physical Review Letters **98**, 1 (2007), 0701554.

[39] G. Xiao and J. Gong, *Suppression of work fluctuations by optimal control: An approach based on Jarzynski's equality*, Physical Review E **90**, 1 (2014), 1408.3461.