Abstract

Title of dissertation: Equilibrium free energies from nonequilibrium simulations:
Improving convergence by reducing dissipation

Suriyanarayanan Vaikuntanathan, Doctor of Philosophy, 2011

Dissertation directed by: Professor Christopher Jarzynski
Department of Chemistry and Biochemistry
Institute for Physical Science and Technology

The estimation of equilibrium free energy differences is an important problem in computational thermodynamics, with applications to studies of ligand binding, phase coexistence and phase equilibrium, solvation of small molecules, and computational drug design, among others. Recent advances in nonequilibrium statistical mechanics, in particular the discovery of exact nonequilibrium work fluctuation relations, have made it possible to estimate equilibrium free energy differences from simulations of nonequilibrium processes in which a system of interest is driven irreversibly between two equilibrium states.

Estimates of ΔF obtained from processes in which the system is driven far from equilibrium often suffer from poor convergence as a consequence of the dissipation that typically accompanies such processes. This thesis is concerned with this problem of poor convergence, and studies methods to improve the efficiency of such estimators. A central theoretical result that guides the development of these methods is a quantitative connection between dissipation and the extent to which the system “lags” behind the actual equilibrium state, at any point in time of the nonequilibrium process.

The first strategy involves generating “escorted” trajectories in the nonequilibrium simulation by introducing artificial terms that directly couple the evolution of the system to changes in the external parameter. Estimators for ΔF in terms of these artificial trajectories are developed and it is shown that whenever the artificial dynamics manage to reduce the lag, the convergence of the free energy estimate is improved. We demonstrate the effectiveness of this method on a few model systems. In particular, we demonstrate how this method can be used to obtain efficient estimates of solvation free energies of model hard sphere solutes in water and other solvents. In the second strategy,“protocol postprocessing”, the trajectories normally generated in the course of a nonequilibrium simulation are used to construct estimators of ΔF that converge faster than the usual estimators. Again, the connection between dissipation and lag guides the development of this method. The effectiveness of this strategy is also demonstrated on a few model systems.

Equilibrium free energy differences from nonequilibrium computer simulations: Improving convergence by reducing dissipation

by

Suriyanarayanan Vaikuntanathan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:

Professor Christopher Jarzynski, Chair/Advisor
Professor Michael E. Fisher
Professor Devarajan Thirumalai
Professor John D. Weeks
Professor Victor M. Yakovenko

To my parents

Acknowledgments

I begin by acknowledging my advisor Prof. Christopher Jarzynski for his guidance and advice during my PhD studies. Working with him has taught me the importance of critical thinking and clarity in scientific research, and I have greatly benefitted from this. Chris gave me plenty of freedom to explore various topics in nonequilibrium and equilibrium statistical mechanics while at the same time making sure that I stayed on course. I have throughly enjoyed working with him, and I thank him for the illuminating discussions on both the topics covered in this thesis, and on other areas of statistical mechanics.

I would also like to thank Prof. John D. Weeks for an extremely interesting course on simulations and theory of liquids, and for his help with the LMF calculations. I would also like to thank the other faculty in the Physics, Chemistry, Mathematics departments with whom I have taken courses. These courses have provided me a strong foundation, and have helped me immensely in my research. I also thank the Chemical Physics program for supporting me in my pursuits.

During my doctoral research, I had the pleasure of collaborating with David. D. L. Minh and Jordan. M. Horowitz (now at Universidad Complutense de Madrid, Madrid) on various research problems. I have also enjoyed the stimulating atmosphere in the Jarzynski research group and for this I would like to thank its members, Andrew Ballard, Dibyendu Mandal, Shaon Chakrobarti, Haitao Quan, and Rian You. I am also grateful to them for having helped me proofread the thesis. I also thank Saar Rahav, Jodi Basner, Chris Bertrand, Sandeep Somani, Jeetain Mittal, Rick Remsing, Yigit Subasi, Prateek Agarwal, Rajibul Islam and Vijay Kumar Krishnamurthy for very useful scientific discussions.

A special thanks to Varada, Udaya Kiran, Bala, Sidd, Adi, Dikpal, Avinash for their friendship. I cannot imagine how my life in College park would have been without their company. I also thank Divya, Rashmish, Mick, Bhargava, Anusha, SK, Balaji and Anand for livening up my time here at Maryland.

Finally, all this would not have been possible if not for the constant encouragement and support from my family. Uma mami and Ramanathan mama's guidance was crucial in my undergraduate years at IIT Madras. My brother Sankar has supported me in every endeavor I have undertaken and I cannot thank him enough for this. My parents have been a constant source of encouragement and inspiration for me and have backed every decision of mine. I will forever be indebted to them for this.

Funding from the National Science Foundation under CHE-0841557, University of Maryland, College Park, and the Ann G. Wylie Dissertation Fellowship is gratefully acknowledged.

Table of Contents

List of Tables **vii**

List of Figures **ix**

List of Symbols and Abbreviations xiii

1 Introduction **1**

2 Background **8**

2.1 Nonequilibrium work free energy theorem 10

2.2 Fluctuation Relation 14

2.3 Other far from equilibrium estimators of ΔF 19

2.4 Computational Efficiency 21

2.5 Summary 24

3 Dissipation and lag **26**

3.1 Theory 28

3.2 Examples 32

3.2.1 Hamiltonian Dynamics 34

3.2.2 Stochastic dynamics 37

3.3 Summary 40

4 Escorted free energy simulations **42**

4.1 Introduction 42

4.2 Molecular dynamics 44

4.3 Monte Carlo dynamics 50

4.4 Fluctuation Theorem 55

4.5 Computational efficiency and figures of merit 62

4.6 Examples 68

4.6.1 One dimensional model system 68

4.6.2 Cavity Expansion 72

4.6.3 Dipole Fluid 81

4.7 Summary 86

5 Estimating solvation free energies using escorted free energy simulations 90
5.1 Introduction 90
5.2 Model and simulations 95
5.2.1 Simulations with a Lennard-Jones Fluid 96
5.2.2 Water Simulations 103
5.3 Discussion 108
5.3.1 Comparing the effectiveness of the FEP and TFEP approaches 110
5.3.2 Quasi Chemical Theory 118
5.4 Summary 120

6 Protocol Postprocessing 121
6.1 Protocol Postprocessing strategy 123
6.2 Importance Sampling Formalism 126
6.3 Dissipation and Lag 128
6.3.1 Exactly solved models 128
6.3.2 Dissipation Bounds Lag 130
6.4 General Case 132
6.5 Model Systems 134
6.6 Discussion and Conclusion 138

7 Summary and future outlook 140

A Fluctuation theorem for stochastic escorted simulations 144

B Figures of Merit 150

Bibliography 153

List of Tables

4.1 Estimates and figures of merit. Here ΔF_F^{est} denotes the estimate of $\Delta F \equiv F_B - F_A$ from the forward process ($R_A \to R_B$) and ΔF_R^{est} denotes the estimate of $-\Delta F$ from the reverse process ($R_A \leftarrow R_B$). ΔF_{BAR}^{est} denotes the estimate of ΔF obtained from Bennett's Acceptance Ratio method. 77

4.2 Estimates and Figures of Merit for $\gamma = 0.1$. Note that the simulations with the mapping are much more efficient than those without. The forward and reverse work histograms obtained from the simulations without any mappings were so far apart that a reliable estimate of C could not be obtained. 85

5.1 Estimates of ΔF obtained using FEP, and the forward (F) and reverse (R) TFEP calculations along with error bars at different values of P^* and at $T^* = 0.85$ for a hard sphere solute. The BAR column has estimates of ΔF obtained using Bennett's acceptance ratio method. Observe that the TFEP becomes more efficient than FEP as P^* and the contact density increase. At the highest value of P^*, we did not observe a single realization where the region between R_A and R_B is vacant. All estimates of ΔF and $\langle W \rangle$ are in units of ϵ. 99

5.2 Estimates of ΔF obtained using FEP, and the forward (F) and reverse (R) TFEP calculations along with error bars at different values of ϵ_1 and at $P^* = 0.022, T^* = 0.85$. The BAR column has estimates of ΔF obtained using Bennett's acceptance ratio method. Observe that the TFEP becomes more efficient than FEP as ϵ_1 increases. At the highest values of ϵ_1, we did not observe a single realization where the region between R_A and R_B is vacant. All estimates of ΔF and $\langle W \rangle$ are in units ϵ. 102

5.3 Estimates of ΔF along with error bars at different values of ϵ_1 at $R_A = 6.0\mathring{A}$, $R_B = 6.05\mathring{A}$, $P = 1$atm and $T = 300K$ in water . Observe that TFEP becomes more efficient than FEP as ϵ_1 increases. At the highest values of ϵ_1, we did not observe a single realization where the region between R_A and R_B is vacant. All estimates of ΔF are in units kJ/mol. At $T = 300K$, $\beta^{-1} = 2.5\,$kJ/mol. The symbol C has been defined in Eq. 4.80 in Sec 4.5. 107

5.4 Estimates of ΔF along with error bars at different values of ϵ_1 at $R_A = 4.0\mathring{A}$, $R_B = 4.05\mathring{A}$, $P = 1\,\mathrm{atm}$ and $T = 300K$ in water . Observe that the TFEP becomes more efficient that FEP as ϵ_1 increases. All estimates of ΔF are in units kJ/mol. At $T = 300K$, $\beta^{-1} = 2.5$kJ/mol. 109

List of Figures

1.1 The axes schematically represent phase space ($\mathbf{z}$-space). The unshaded ovals denote the statistical state of the system, ρ_t, and the shaded ovals denote the equilibrium state, ρ_t^{eq}, corresponding to the value of external parameter, λ, at various instants of time. As the work parameter λ is switched from A to B, a lag builds up as the state of the system, ρ_t, pursues the equilibrium distribution corresponding to the changing work parameter, ρ_t^{eq}. 5

2.1 A gas of particles inside a container, in contact with a thermal reservoir (not shown) is driven out of equilibrium by switching the position of the piston (compression in this schematic, the dashed lines denote the final position of the piston) at a finite rate. The work performed, averaged over many repetitions of the nonequilibrium process, $\langle W \rangle$, exceeds the free energy difference ΔF between the equilibrium states corresponding to the final and initial positions of the piston. 9

2.2 Snapshots of typical configurations observed in the course of forward (rapid compression) and reverse (rapid expansion) of a gas with $N_p \gg 1$. In forward process, the gas particles stack up against the piston as the gas is compressed rapidly. On the other hand, in the time reversed process, the gas is expanded rapidly and the region around the piston quickly becomes devoid of gas particles. The conjugate twin of a typical trajectory in the forward process is practically never observed in the reverse process. This schematic depicts the time reversal asymmetry inherent to processes with high dissipation. . . . 15

2.3 An illustration of a pair of conjugate trajectories in phase space. The trajectory γ_R was obtained by reflecting γ_F along the q axis. The arrows indicate the direction of time. 18

2.4 A schematic of work distributions observed in a fast switching simulation. The vertical line marks the point where the two distributions intersect ($W = \Delta F$). In the forward (F) simulation, work values, W, are typically sampled from the dominant region of $P_F(W)$. However, in order for the estimate of ΔF to converge, the dominant region of the distribution $P_R(-W)$ needs to adequately sampled. As the two typical regions are far apart, the estimate of ΔF suffers from poor convergence. 22

3.1 A two-component dilute gas, where component 1 (open circles) is confined by the piston, while component 2 (filled circles) is not. . . . 33

3.2 Dissipation ($\beta W_{\text{diss}}(t)$) and lag ($D[\hat{\rho}_t||\hat{\rho}_t^{\text{eq}}]$) are plotted as functions of N_2, with $N_1 = 10$, $d = 3$, $V_0/V_{\lambda(t)} = 5$, and $\beta = 1$. The isothermal limit is achieved as $N_2 \to \infty$. 40

4.1 The potential energy landscape for $\lambda = 0$ (solid line) and $\lambda = 1$ (dashed line). Also depicted are the equilibrium distribution and the flow field, at $\lambda = 0$. 69

4.2 Comparison of estimates of ΔF using Eqs. 2.5 and 4.5. 71

4.3 A schematic of the cavity expansion problem 73

4.4 Running estimate of the probability that the region $R_A \leq r \leq R_B$ is devoid of fluid particles, $P(n = 0) = \exp(-\beta\Delta F)$, from escorted free energy simulations in which R is switched from R_A to R_B, plotted as a function of the number of trajectories used to obtain the estimate. The (green) horizontal line is the estimate of $\exp(-\beta\Delta F)$ obtained using Bennett's Acceptance ratio (BAR) method with $n_F = n_R =$ 50000 trajectories. Observe that the running estimate converges to the BAR estimate in 50000 trajectories. 76

4.5 Graphical verification of the fluctuation theorem and estimation of ΔF. The horizontal line indicates the estimate of ΔF obtained from the acceptance ratio method (Table 4.1). 80

4.6 Work histograms obtained from forward and reverse simulations performed at $\gamma = 0.1$. The degree of overlap between $P_F(W)$ (right) and $P_R(-W)$ (left) provides an indication of the efficiency of the free energy estimate. For unescorted simulations (no mapping) we see no overlap, reflecting considerable dissipation and poor efficiency (Table 4.2). With the mapping given by Eq. 4.91 the overlap is much improved, and with the mean field mapping, Eq. 4.92 the forward and reverse distributions are nearly identical. 89

5.1 A sketch of the potential $V_O^R(r)$ at $R = R_A$. As mentioned in the text, we have set $\sigma_2 = R_A - 2^{(1/6)}\sigma_1$. The potential has an excluded volume interaction for $r < R$. A positive value of ϵ_1 sets the strength of the short-ranged solute-solvent attractions. We are interested in estimating the free energy difference associated with changing R, the radius of the excluded volume interaction, from R_A to R_B while keeping the short-ranged attractions constant. 93

5.2 The solute-solvent radial distribution function $g(r)$ as a function of the distance r from the center of the solute at different values of P^*. Notice the occurrence of drying at $P^* = 0.022$. The density of solvent at the point of contact with solute increases with the reduced pressure P^*. The distances are in units of σ. 94

5.3 The solute-solvent radial distribution function $g(r)$ as a function of the distance r from the center of the solute at different values of ϵ_1. Notice the onset of drying at $\epsilon_1 = 0$. The density of solvent at the point of contact with solute increases with ϵ_1. The distances are in units of σ. 100

5.4 The solute-oxygen radial distribution function $g(r)$ as a function of the distance r of center of the water molecule (oxygen atom) from the center of the solute at different values of ϵ_1. The density of solvent at the point of contact with solute increases with ϵ_1. The length of the simulation box was around $L \sim 30\AA$. 106

5.5 Plots of $\langle x_s^2\rangle/\langle x_s\rangle^2$ at different values of $\beta\epsilon_1$ for the solute described by Eq. 5.1 in a) Lennard-Jones (LJ) solvent with $R_A = 2\sigma$, b) spherically truncated SPC/E water with $R_A = 6\AA$, and c) spherically truncated SPC/E water with $R_A = 4\AA$. We did not plot the value corresponding to $\epsilon_1 = 0$ in case (a) because x_s was negative in this instance. When x_s becomes negative, the trends in $\langle x_s^2\rangle/\langle x_s\rangle^2$ do not accurately represent the trends in $\langle x_s^2\rangle/(N_s(\partial F/\partial R)^2)$. 116

5.6 Schematic of the QCT calculaiton. 118

6.1 Lag in driven nonequilibrium processes. Consider a system driven from state A to state B in a finite-time process. In the above schematic, the ovals represent regions of phase space. The darkly shaded ovals are regions of phase space containing most of the density $\rho^{eq}_{\lambda(t)}$ of the equilibrium state corresponding to the value of the external parameter at time t. The unshaded ovals denote the phase space regions containing most of the density ρ_t actually accessed by the system during the process. In a reversible process, the two would be indistinguishable. Since the system is driven out of equilibrium, however, a lag builds up between ρ_t and $\rho^{eq}_{\lambda(t)}$. This lag is correlated to dissipation and is ultimately responsible for the poor convergence of free energy estimates based on nonequilibrium processes. If one is able to obtain a function $\lambda^*(t)$ with $\lambda^*(0) = A$ such that the equilibrium states $\rho^{eq}_{\lambda^*(t)}$ are closer to the ρ_t (e.g. the lightly shaded ovals), then the convergence of free energy estimates may be improved using Eq. 6.7. 122

6.2 Comparison of free energy estimates for Sun's system: NEDDS simulations were analyzed with importance sampling, Eq. 6.13 (circles), or the Feynman-Kac formalism, Eq. 6.7 (triangles). Standard nonequilibrium work relation estimates, Eq. 2.5 (squares), were performed on slower simulations with the same total time as the NEDDS simulations or by using the analysis protocol as a new sampling protocol (diamonds). The symbols indicate the mean and error bars indicate the standard deviation of 10000 estimates, each based on 50 trajectories. The simulation time step was $\Delta t = 0.001$ and the rate v indicates that λ was incremented by $v\Delta t$ at each time step of the NEDDS simulations. While the switching rates are equivalent, some points are given a small horizontal offset to prevent error bar overlap. The exact free energy is shown as a shaded line. 136

6.3 Comparison of free energy estimates for Hummer's system. The caption for Fig. 6.2 applies here, except that the potential is Hummer's rather than Sun's and each estimate is based on 250 trajectories. . . 137

6.4 Potential energy surface for a 2D system. The contour plot is of $5(x^2-1)^2+5(x-y)^2$ and the red line traces the equilibrium position of the harmonic bias $\frac{15}{2}(x+\cos(\pi\lambda))^2+\frac{15}{2}(y+1-\sin(2\pi\lambda)-2\lambda)^2$ as λ goes from 0 (left) to 1 (right). 138

6.5 Comparison of free energy estimates for a 2D system with $\Delta F = 0$. The caption for Fig. 6.2 applies here, except that the potential is Eq. 6.26 rather than Sun's system, each estimate is based on 250 trajectories, and multidimensional versions of the importance sampling and Feynman-Kac formalisms were used. 139

List of Symbols and Abbreviations

This list describes the symbols and abbreviations that are used commonly in this thesis and references the page number where they are first mentioned.

BAR	Bennett's Acceptance Ratio	23
C	Normalization of the harmonic mean distribution	24
d	Dimensionality of phase (or configuration) space	51
D[f||g]	Relative entropy between the distributions f and g	23
$\mathbf{F}_k(\mathbf{r}_k)$	Force on the k^{th} solvent particle	111
F_λ	Free energy of the equilibrium state λ	10
FEP	Free energy perturbation	2
$g(r)$	Solvent-solute radial distribution function	97
H_λ	Parameter dependent Hamiltonian	2
J_i	Jacobian corresponding to M_i	53
$\mathrm{k_B}$	Boltzmann's constant	2
L	Length of the simulation box	72
LMF	Local molecular field theory	104
$m_i(\mathbf{r}_n)$	Single particle bijective mapping in the i^{th} Monte-Carlo step	74
M_i	Bijective mapping function	52
N_s	Number of nonequilibrium trajectories generated in the simulation	3
N_p	Number of particles	8
P	Pressure	96
P^*	Reduced pressure	96
$P(n=0)$	Probability that the region $R_A \leq r \leq R_B$ is devoid of fluid	74
$P_\lambda(\mathbf{z}\|\mathbf{z}_0)$	Transition probability	50
$P_F(\gamma)$	Distribution of trajectories in the forward process	17
$P_R(\gamma)$	Distribution of trajectories in the reverse process	17
$P_F(W)$	Distribution of W values in the forward process	19
$P_H(W)$	Harmonic mean distribution	23
$P_R(W)$	Distribution of W values in the reverse process	19
QCT	Quasi Chemical Theory	118
$\mathbf{r}_n$	Vector describing the position of the n^{th} fluid molecule	64
$R_A\,, R_B$	Initial and final radius of the hard sphere	72
T	Temperature	2
T^*	Reduced temperature	77
TFEP	Targeted free energy perturbation	55
TI	Thermodynamic Integration	2
$\mathbf{u}(\mathbf{z},\lambda)$	Vector flow field (d-dimensions)	45

$\bar{\mathbf{u}}(\mathbf{r}_k)$	3-dimensional vector field	110
$\mathbf{u}^*\,, M_i^*$	Perfect flow field and mapping functions respectively	63
$\tilde{v}$	Physical dynamics	45
$v_0(r)$	Short-ranged component of the Coulomb potential	104
$v_1(r)$	Long-ranged component of the Coulomb potential	104
W	Work performed in a particular realization of the nonequilibrium process	4
W_{diss}	Total dissipated work	4
$W_{\text{diss}}(t)$	Work dissipated up to time t	28
$\mathbf{z}$	Point in many-dimensional phase or coordinate space	5
$\mathbf{z}^*$	Phase space point obtained from $\mathbf{z}$ by flipping its momentum degrees of freedom	17
$\mathbf{z}'$	Phase space coordinate obtained by subjecting $\mathbf{z}$ to M_i	52
$\{\mathbf{z}_t\}$	Trajectory	12
Z_λ	Partition function of the equilibrium state λ	10
β	Inverse temperature	2
γ	Trajectory	12
γ^*	Conjugate twin of the trajectory γ	17
γ_F, γ_R	A pair of conjugate trajectories in the forward and reverse processes	16
ΔF	Free energy difference	1
$\Delta F^*(t)$	Free energy difference between the states $\lambda^*(0)$ and $\lambda^*(t)$	121
ϵ_1	Strength of the short-ranged solute solvent attractions	95
λ	External parameter	2
$\lambda = A$	Initial value of λ	2
$\lambda = B$	Final value of λ	2
λ_i	Value of λ in the i^{th} step of a Monte-Carlo simulation	51
$\lambda(t)$	Protocol used to switch λ in the nonequilibrium process	10
$\tilde{\lambda}(t)$	Time reversed protocol to switch λ	17
$\lambda^*(t)$	Analysis protocol	121
ρ	Bulk density	97
ρ_t^{eq}	Equilibrium distribution at a time t	5
ρ_t	Distribution of the system at a time t	5
$\rho(\mathbf{z}, t)$	Phase space density of system at time t	29
$\rho^{\text{eq}}(\mathbf{z}, \lambda)$	Equilibrium distribution at λ	10
ρ_c	Contact density	114
σ_c	Cutoff that determines the range of $v_0(r)$	104
$\sigma\,, \epsilon$	Length and energy scales in the Lennard-Jones potential	77
τ	Duration of the nonequilibrium process	10
$\psi(r)$	Self consistent LMF potential	105
$\mathcal{A}(t)$	Auxiliary operator	124
$\mathcal{L}_\lambda$	Liouville operator for physical dynamics	45
$\mathcal{W}_t^*$	“Work” performed in the protocol postprocessing formalism	125
$\langle \ldots \rangle_F$	Average over the forward process	22
$\langle \ldots \rangle_R$	Average over the reverse process	22

Chapter 1

Introduction

Computer simulations are routinely used in condensed matter physics, statistical mechanics, and computational chemistry to investigate the properties of many-body systems, especially model systems not amenable to theoretical treatment or direct experimentation [24, 59]. Simulations of such systems have provided important insights into topics such as phase coexistence and phase equilibria [78], critical phenomena [59], and the protein folding problem [77, 79]. Computing the thermodynamic properties of the model system is the central goal of many such computer studies, and in this context the estimation of equilibrium free energy differences, ΔF, becomes very important [13]. Such estimates of ΔF are crucial for example in identifying stable configurations of proteins [13], computing the excess chemical potential of a molecule in a solvent fluid [103], protein-ligand binding studies [13], and studying fluid-solid and solid-solid phase [23] equilibria. Moreover, by estimating ΔF between a thermodynamic state of interest and an analytically tractable reference state, the absolute free energy of the state of interest can be determined. Given the importance of free energy calculations in computational studies, there is a need to develop techniques that can provide efficient estimates of ΔF from simulations.

Free Energy Perturbation (FEP) [108] and Thermodynamic Integration (TI) [57] were among the first methods developed to estimate ΔF from computer simulations and remain popular to this day [13]. Imagine a system whose equilibrium states at some temperature T are parameterized by the value of an external parameter vector λ. We will generically be interested in computing the free energy difference between the equilibrium states corresponding to $\lambda = A$ and $\lambda = B$, $\Delta F = F_B - F_A$. For instance, if the system of interest is a lattice of Ising spins in a magnetic field h, with nearest neighbor couplings J [10], the external parameter vector can be defined as $\lambda \equiv \{h, J\}$. The equilibrium state of the system at a particular temperature T is then parametrized by λ and we may be interested in computing the free energy difference ΔF between the states $A = \{h_A, J_A\}$ and $B = \{h_B, J_A\}$. Henceforth for simplicity, we will refer to λ as the "external parameter" instead of the "external parameter vector".

The FEP method is based on the following identity by Zwanzig [108]

$$\langle e^{-\beta\Delta H}\rangle_A = e^{-\beta\Delta F}, \tag{1.1}$$

where $\Delta H = H_B - H_A$ denotes the change in the energy of system when λ is switched from A to B (we will use H_λ to denote the Hamiltonian that describes the system when the external parameter is at λ), $\beta^{-1} = k_B T$, and $\langle \dots \rangle_\lambda$ denotes an average over the canonical distribution that describes the equilibrium state λ.

The thermodynamic integration method on the other hand is based on the following identity by Kirkwood [57]

$$\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda = \frac{\partial F}{\partial \lambda}, \tag{1.2}$$

which can be integrated to give

$$\int_A^B d\lambda \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda = \Delta F, \tag{1.3}$$

where the integral is performed over a path in parameter space connecting $\lambda = A$ to $\lambda = B$.

While many of the methods in use to estimate ΔF rely either on the TI (Eq. 1.2) or the FEP (Eq. 1.1) identity (or variants thereof) and involve sampling from a thermal ensemble, or a "biased" thermal ensemble in the case of umbrella sampling [98], there has been recent interest in the use of methods that estimate ΔF from simulations in which the system is driven irreversibly between two equilibrium states. These estimators are based on the nonequilibrium work fluctuation relations [17,18,48,49] and are valid in principle for systems driven arbitrarily far from equilibrium. In this approach, one repeatedly simulates a thermodynamic process during which the parameter λ is "switched" at a finite rate from A to B, with initial conditions sampled from equilibrium. ΔF is then estimated using the identity [49]

$$e^{-\beta \Delta F} = \left\langle e^{-\beta W} \right\rangle \approx \frac{1}{N_s} \sum_{n=1}^{N_s} e^{-\beta W_n}. \tag{1.4}$$

Here angular brackets denote an ensemble average over realizations of the process, W_n is the work performed on the system during the n'th of N_s such simulations, and the approximation becomes an equality as $N_s \to \infty$. This relation reduces to Eq. 1.1

and Eq. 1.3 in the opposite limits of sudden switching and quasi-static isothermal switching, respectively. The nonequilibrium approach is especially relevant in the context of single molecule force spectroscopy [9,14,42,43,64]. In these experiments and analogous simulations [46], one end of a molecule or a molecular complex is anchored while a force is applied to the other end (using laser tweezers or atomic force microscopes in experiments and using constraint potentials in simulations). This nonequilibrium process is used to induce and probe rare events such as protein and nucleic acid unfolding and ligand dissociation. Nonequilibrium fluctuation relations [17,18,49] such as Eq. 1.4 can then be used to extract equilibrium thermodynamic information, for example the potential of mean force along a reaction coordinate, from the data obtained in such processes [42,43,80].

While Eq. 1.4 can in principle be used to estimate ΔF from simulations of arbitrarily short duration ("fast switching" [36]), in practice we pay a penalty in the form of poor convergence [33,51,58], as the number of simulations needed to obtain a reliable free energy estimate using Eq. (1.4) increases rapidly with the *dissipated work*,

$$W_{\text{diss}} \equiv \langle W \rangle - \Delta F \geq 0, \tag{1.5}$$

that accompanies fast switching simulations. This dissipation is a consequence of the second law of thermodynamics, and reflects the *lag* that develops as the system pursues – but is unable to keep pace with – the equilibrium state corresponding to the continually changing value of the work parameter, λ (Fig 1.1) [37,82,99,104]. We can diminish the lag by running longer simulations in which λ is varied slowly,

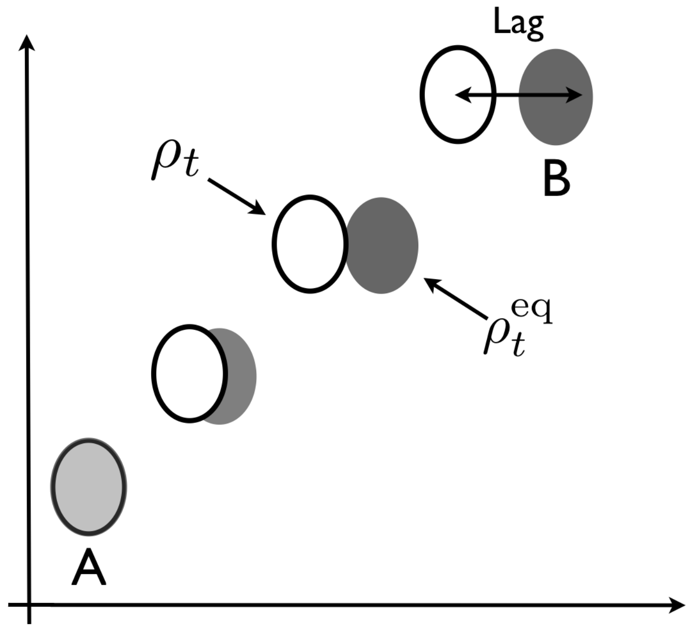

Figure 1.1: The axes schematically represent phase space (**z**-space). The unshaded ovals denote the statistical state of the system, ρ_t, and the shaded ovals denote the equilibrium state, ρ_t^{eq}, corresponding to the value of external parameter, λ, at various instants of time. As the work parameter λ is switched from A to B, a lag builds up as the state of the system, ρ_t, pursues the equilibrium distribution corresponding to the changing work parameter, ρ_t^{eq}.

but this increases the computational cost per simulation.

This thesis is concerned with this problem of poor convergence of ΔF estimates from fast switching nonequilibrium simulations due to dissipation and lag. General strategies to improve the efficiency of these estimates are introduced. The next chapter reviews the theoretical underpinnings of Eq. 1.4, and discusses other nonequilibrium estimators of ΔF. We will elaborate on the reasons behind the poor performance of fast switching nonequilibrium estimators and revisit the assertion that fast switching nonequilibrium estimates of ΔF are inefficient due to high dissipation. Chapter 3 presents an exact quantitative relation between dissipation and lag for systems driven away from equilibrium. This relation allows us to correlate the poor performance of fast-switching nonequilibrium simulations with the lag. In the subsequent chapters, methods aimed at improving the efficiency of ΔF estimates by reducing the lag are introduced. In particular, Chapter 4 introduces a method, *escorted free energy simulations*, in which the equations of motion ordinarily used to simulate the evolution of the system are modified with artificial terms that couple the evolution of the system to changes in the external parameter λ. A generalization of Eq. 1.4 that allows us to estimate the free energy difference in terms of these artificial trajectories is derived. Using the connection between dissipation and lag, we show that whenever these artificial terms manage to reduce the lag, the method provides an improved estimator of the free energy difference. We illustrate the effectiveness of our method by (a) estimating the free energy difference in a one dimensional model system, (b) estimating the free energy difference associated with growing a hard sphere solute in a fluid, and (c) estimating the free energy difference associated with introducing an electric field in a model dipole fluid. The free energy estimation problem described in (b) is rather important in computational thermodynamics. Hence in Chapter 5 this problem is considered in detail. In particular we compute the free energy cost of growing hard sphere solutes in water and Lennard-Jones fluids and compare the effectiveness of the free energy estimates obtained using the new method to that obtained from Eq. 1.4. Chapter 6 develops another method, *protocol postprocessing*, in which the trajectories normally generated in the course of a nonequilibrium simulation are used to construct estimators of ΔF that converge faster than the estimator obtained from Eq. 1.4. Again, the connection between dissipation and lag becomes useful in the development of this method. We end the thesis by suggesting directions for future research.

Chapters 3, 4, 6 are based in full or in part on the following publications.

- Chapter 3: S. Vaikuntanathan, C. Jarzynski "Dissipation and Lag in Irreversible Processes", *Euro. Phys. Lett,* **87**, *600005*, 2009.

- Chapter 4: S. Vaikuntanathan, C. Jarzynski "Escorted Free Energy Simulations: Improving Convergence by Reducing Dissipation", *Phys. Rev. Lett* **100**, *190601*, 2008, and S. Vaikuntanathan, C. Jarzynski "Escorted Free Energy Simulations", *J. Chem. Phys* **134**, *054107*, 2011.

- Chapter 6: D. D. L. Minh, S. Vaikuntanathan "Density-Dependent Analysis of Nonequilibrium Paths Improves Free Energy Estimates II. A Feynman-Kac Formalism ", *J. Chem. Phys* **134**, *034117*, 2011.

Research work described in the following publications is not described in this thesis.

- J. M. Horowitz, S. Vaikuntanathan "Nonequilibrium Detailed Fluctuation Theorem for Repeated Discrete Feedback", *Phys. Rev. E* **82**, *061120*, 2010

- S. Vaikuntanathan, C. Jarzynski "Modeling Maxwells demon with a microcanonical Szilard engine", *Phys. Rev. E* **83**, *061120*, 2011

Chapter 2

Background

For the purpose of illustrating the general ideas discussed in this chapter, it is useful to imagine a system of N_p gas particles confined inside a container with a piston (see Fig 2.1), in contact with a thermal reservoir. Consider a process in which the system is prepared in a state of thermal equilibrium, after which the piston is moved from its initial location to a predetermined final location at a speed v, compressing the gas in the process. If the piston is moved quasi-statically and the gas remains in equilibrium with the reservoir throughout the process, the second law of thermodynamics stipulates that the average work performed on the gas, $\langle W \rangle$, is equal to ΔF, the free energy difference between the equilibrium states corresponding to the final and initial positions of the piston [60]. When the piston is moved at a finite rate, driving the system out of equilibrium in the process, thermodynamics does not provide a prescription to obtain an estimate of ΔF. Rather, the second law of thermodynamics just tells us that $\langle W \rangle > \Delta F$. However, recent

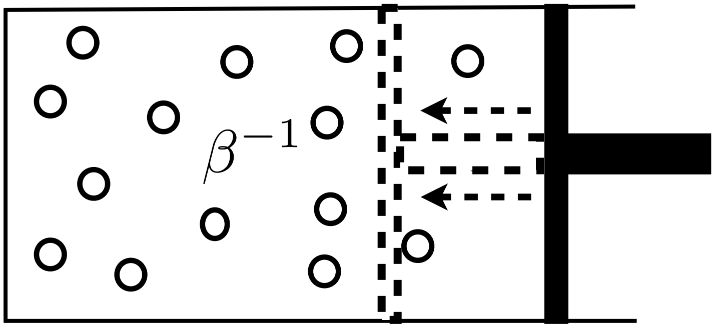

Figure 2.1: A gas of particles inside a container, in contact with a thermal reservoir (not shown) is driven out of equilibrium by switching the position of the piston (compression in this schematic, the dashed lines denote the final position of the piston) at a finite rate. The work performed, averaged over many repetitions of the nonequilibrium process, $\langle W \rangle$, exceeds the free energy difference ΔF between the equilibrium states corresponding to the final and initial positions of the piston.

advances in nonequilibrium statistical mechanics have shown that it possible to estimate equilibrium free energy differences from such nonequilibrium processes (see for example Eq. 2.5) [17,18,48,49,53]. Besides providing a method to compute ΔF from nonequilibrium processes, these results are interesting in their own right as they have clarified important issues regarding irreversibility, and the applicability of the second law of thermodynamics to microscopic systems. In this chapter, we briefly review these results and show how they can be of use in computational thermodynamics in the context of free energy estimation. We will end this chapter with a discussion on the efficiency of nonequilibrium estimators of ΔF.

2.1 Nonequilibrium work free energy theorem

We begin by specifying the framework that we will use to describe processes such as the one illustrated in Fig. 2.1, and we discuss an exact relation, Eq. 2.5, that is valid for these processes. This framework will be used throughout this thesis.

Consider a classical system described by a Hamiltonian $H(\mathbf{z}; \lambda)$, or $H_\lambda(\mathbf{z})$, where $\mathbf{z}$ specifies a point in many dimensional phase or configuration space and λ denotes an external parameter. For example, in the system described in Fig 2.1, λ specifies the position of the position. At a temperature T, the equilibrium state of this system is parameterized by λ and described by the distribution

$$\rho^{\text{eq}}(\mathbf{z}, \lambda) = \frac{1}{Z_\lambda} \exp[-\beta H(\mathbf{z}, \lambda)], \tag{2.1}$$

with free energy $F_\lambda = -\beta^{-1} \ln Z_\lambda$, where as usual $\beta^{-1} = k_B T$. We are interested in the difference $\Delta F = F_B - F_A$ between two equilibrium states at the same temperature T but at different parameter values, $\lambda = A$ and $\lambda = B$. To estimate ΔF, we will imagine a process in which the system is initially prepared in the equilibrium state A by allowing it to equilibrate with a thermal reservoir at a temperature T, after which λ is switched from $\lambda(0) = A$ to $\lambda(\tau) = B$ according to a specific protocol $\lambda(t)$. The system may be either isolated or in contact with the thermal reservoir while the value of λ is switched. Note that in general the system will be out of equilibrium at time $t = \tau$, that is, its statistical state will not correspond to the distribution $\rho^{\text{eq}}(\mathbf{z}, B)$.

When the system in question is macroscopic, the second law of thermodynamics predicts that the work performed on the system during this process will be no less than the free energy difference $\Delta F = F_B - F_A$, even if the system ends the process out of equilibrium:

$$W \geq \Delta F. \tag{2.2}$$

To establish this, let us imagine that from $t = \tau$ to some later time $t = \tau^*$, the parameter is held fixed at $\lambda = B$, allowing the system to re-equilibrate with the reservoir [1]. Thus we now have a process during which the system begins in the equilibrium state A (at $t = 0$) and ends in equilibrium state B (at $t = \tau^*$). During this process the change in the entropy of the system is $\Delta S = S_B - S_A$, while the change in the entropy of the reservoir is $-Q/T$, where Q is the heat absorbed by the system from $t = 0$ to $t = \tau^*$.

Since the combined change in the entropy must be non-negative, we get

$$\Delta S - \frac{Q}{T} \geq 0. \tag{2.3}$$

We can now use the first law of thermodynamics, $\Delta E = W + Q$, where ΔE denotes the change in the internal energy of the system, and the macroscopic definition of Helmholtz free energy, $F = E - TS$, to rewrite Eq. 2.3 in the form given by Eq. 2.2.

Finally, since no work is performed on the system during the re-equilibration state ($\tau \leq t \leq \tau^*$) we can simply interpret W in Eq. 2.2 as the work performed during the process from $t = 0$ to $t = \tau$.

When the system is microscopic, we expect Eq. 2.2 to hold on average,

$$\langle W \rangle \geq \Delta F, \tag{2.4}$$

[1] If the system was isolated from the reservoir during the switching interval $0 \leq t \leq \tau$, we assume that it is brought back into contact with the reservoir during the interval $\tau \leq t \leq \tau^*$.

where $\langle \dots \rangle$ denote an average over infinitely many repetitions of the process. Thus the average work places an upper bound on the free energy difference ΔF. However, when the full statistical distribution of work values is considered, it is possible to obtain an estimate for ΔF (not just an upper bound) using the identity [48, 49]

$$e^{-\beta \Delta F} = \left\langle e^{-\beta W} \right\rangle \quad . \tag{2.5}$$

In both Eqs. 2.4 and 2.5, we use the following expression for the work performed on the system during a particular realization of the process:

$$W = \int_0^\tau \dot{\lambda} \, \frac{\partial H_\lambda}{\partial \lambda}(\mathbf{z}_t, \lambda(t)) \, \mathrm{d}t, \tag{2.6}$$

where the trajectory $\gamma = \{\mathbf{z}_t\}$ describes the microscopic evolution of the system during this realization. The definition of work in the equation above can be connected to the mechanical definition of work (product of force and displacement) by interpreting λ as a generalized coordinate and $\partial H / \partial \lambda$ as its conjugate generalized force. Eq. 2.5 is commonly referred to as the nonequilibrium work relation and relates the distribution of nonequilibrium work values to the equilibrium free energy difference ΔF. The second law of thermodynamics, Eq. 2.4, follows from Eq. 2.5 using Jensen's inequality [15], which states that for any convex function f, and a random variable x,

$$\langle f(x) \rangle \geq f(\langle x \rangle) \tag{2.7}$$

where $\langle \dots \rangle$ denotes an average over values of the random variable. Applying Eq .2.7 to the nonequilibrium work relation, we obtain the second law.

In the limit of infinitely fast switching, $\dot{\lambda} \to \infty$, the system does not have time to respond to the external perturbation. Thus the work performed is given by $W = \Delta H = H(\mathbf{z}, B) - H(\mathbf{z}, A)$, where the point $\mathbf{z}$ is sampled from the equilibrium state A, and the average $\langle \dots \rangle$ in Eq. 2.5 is simply an average over the initial equilibrium state A. In other words, Eq. 2.5 reduces to the FEP identity Eq. 1.1. In the opposite limit of quasi-static isothermal switching, the work performed along any trajectory is equal to

$$W = \int_0^\tau \dot{\lambda} \, \langle \partial H_\lambda / \partial \lambda \rangle \, \mathrm{d}t, \tag{2.8}$$

due to adiabatic averaging [49,100]. Eq. 2.5 then reduces to the TI identity, Eq. 1.3.

When λ is switched slowly (but not quasi-statically), and the system remains near equilibrium throughout the process, the distribution of work values obtained in the process is Gaussian [49,93]. In this near-equilibrium limit, Eq. 2.5 is equivalent to a fluctuation-dissipation relation [37,49]

$$\langle W \rangle - \Delta F = \frac{\beta}{2} \sigma_W^2, \tag{2.9}$$

where $\sigma_W^2 \equiv \langle (W - \langle W \rangle)^2 \rangle$. Eq. 2.9 relates the work dissipated in a near equilibrium process to the fluctuations in the work values.

When dealing with processes in which systems are driven out of equilibrium, and especially when one is interested in simulating such processes, it becomes necessary to explicitly model the evolution of the system. For example if the system is isolated and not in contact with the thermal reservoir when the external parameter is being switched, the evolution of the system can be modeled by Hamilton's equations. On the other hand, if the system remains in contact with the thermal reservoir, other dynamics such as Langevin dynamics, Monte-Carlo dynamics, and the Andersen and Nosé-Hoover thermostats [24] are more appropriate choices to model the evolution. One may then wonder whether the nonequilibrium work relations and related results discussed below are valid only for a particular choice or a restrictive set of dynamics.

The validity of Eq. 2.5 hinges on the condition that the dynamics used to model the evolution of the system must preserve the canonical distribution when λ is held fixed [48]. This is not a restrictive condition [53]. Consider for example the case that the system is isolated and its dynamics are Hamiltonian. An ensemble of trajectories evolving under Hamilton's equations at fixed λ with initial conditions sampled from the equilibrium distribution Eq. 2.1 continue to be described by the same equilibrium distribution at later times [49]. On the other hand, if the system is in contact with a thermal reservoir, the dynamics that are commonly used to model the evolution of a system (Langevin, Monte-Carlo dynamics for example) are designed to generate phase space points, $\mathbf{z}$, eventually distributed according to Eq. 2.1 when λ is held fixed. In other words, the equilibrium distribution is a stationary solution of the dynamics for fixed λ thus ensuring that the aforementioned condition is satisfied [54].

2.2 Fluctuation Relation

Irreversible thermodynamic processes are both dissipative ($\langle W\rangle - \Delta F \geq 0$) and asymmetric under time reversal. It is useful to illustrate this point with a concrete

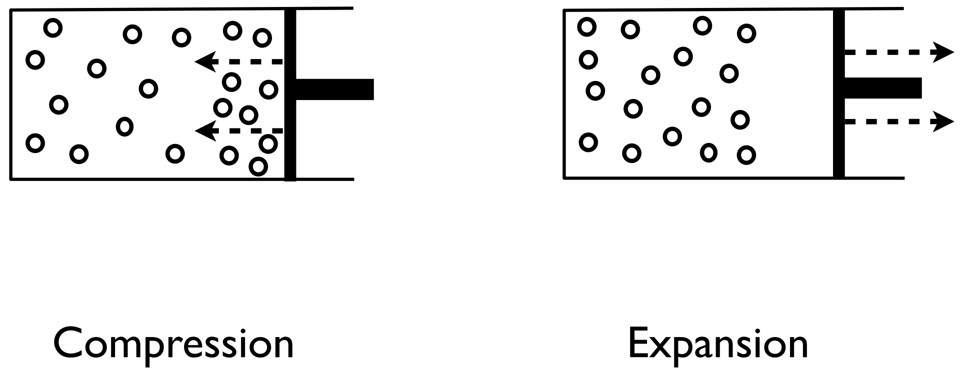

Figure 2.2: Snapshots of typical configurations observed in the course of forward (rapid compression) and reverse (rapid expansion) of a gas with $N_p \gg 1$. In forward process, the gas particles stack up against the piston as the gas is compressed rapidly. On the other hand, in the time reversed process, the gas is expanded rapidly and the region around the piston quickly becomes devoid of gas particles. The conjugate twin of a typical trajectory in the forward process is practically never observed in the reverse process. This schematic depicts the time reversal asymmetry inherent to processes with high dissipation.

example. Consider again the system composed of N_p gas particles enclosed inside a container with a piston and imagine a pair of forward (F) and reverse (R) processes. The system starts from equilibrium in both processes. In the forward process, the piston is moved from A to B at a speed v, while in the reverse process the piston is moved from B to A at the same speed. The trajectories in these processes occur in conjugate pairs. If a trajectory γ_F is a solution of the equations of motion in the forward process, its conjugate twin, γ_R, obtained by running γ_F backwards (see Fig 2.3 and Eq. 2.10), is a solution of the equations of motion in the reverse process.

Let us now consider the case that the number of gas particles is macroscopic, $N_p \gg 1$, and imagine a pair of forward and reverse processes in which the position of the piston is switched rapidly. The dissipation in this process is macroscopic. Let $\gamma_F^{typical}$ denote a typical trajectory of the forward process. While its conjugate twin is a solution of the equations of motion of the reverse process, it is practically never observed in the reverse process (see Fig 2.2). In other words the process is asymmetric under time reversal. On the other hand, if the process is carried out quasi-statically ($v \to 0$) and isothermally, dissipation is eliminated, $W = \Delta F$, and conjugate twin of $\gamma_F^{typical}$ is in turn typical to the reverse process: the process is symmetric under time reversal.

If the system is microscopic, statistical fluctuations become important and it might be possible to observe the conjugate twin of $\gamma_F^{typical}$ in the reverse process even for high switching speeds v. The notion of time reversal asymmetry can be generalized in such cases and can be quantified by computing the likelihood that the conjugate twin of a typical trajectory in the forward process is observed in the reverse process. This likelihood decreases (time reversal asymmetry increases) with increasing dissipation. At the heart of this connection between dissipation and time-reversal asymmetry is the Fluctuation Theorem, Eq 2.11 below, which relates the probability densities associated with observing a pair of conjugate trajectories in the forward and reverse processes.

Before stating this theorem, we will formally define the reverse (R) process as one where the system is initially prepared in the state B, after which the value of λ is varied according to the protocol $\tilde{\lambda}(t) \equiv \lambda(\tau - t)$ from $\tilde{\lambda}(0) \equiv B$ to $\tilde{\lambda}(\tau) \equiv A$. Let $\gamma_F = \{\mathbf{z}_F(t)\}$ denote a trajectory in the forward process starting from $\mathbf{z}_F(0)$ and ending at $\mathbf{z}_F(\tau)$, and let the trajectory $\gamma_R \equiv \gamma^* = \{\mathbf{z}_R(t)\}$ denote its conjugate twin [51,53], with

$$\mathbf{z}_R(t) = \mathbf{z}_F{}^*(\tau - t), \tag{2.10}$$

where $\mathbf{z}^*$ is obtained from $\mathbf{z}$ by reversing the signs of the momentum degrees of freedom (Fig 2.3).

Let $P_F(\gamma)$ ($P_R(\gamma)$) denote the probability density in trajectory space in the forward (reverse) process. The pair of densities, $P_F(\gamma)$ and $P_R(\gamma)$, satisfy the following fluctuation theorem by Crooks [17,18]

$$\frac{P_F(\gamma_F)}{P_R(\gamma_R)} = e^{\beta(W - \Delta F)}, \tag{2.11}$$

where W denotes the work done on the system as it evolves along the trajectory γ_F. If λ is switched quasi-statically and isothermally, the system remains in equilibrium throughout and $W = \Delta F$ for every trajectory. Eq. 2.11 then tells us that the

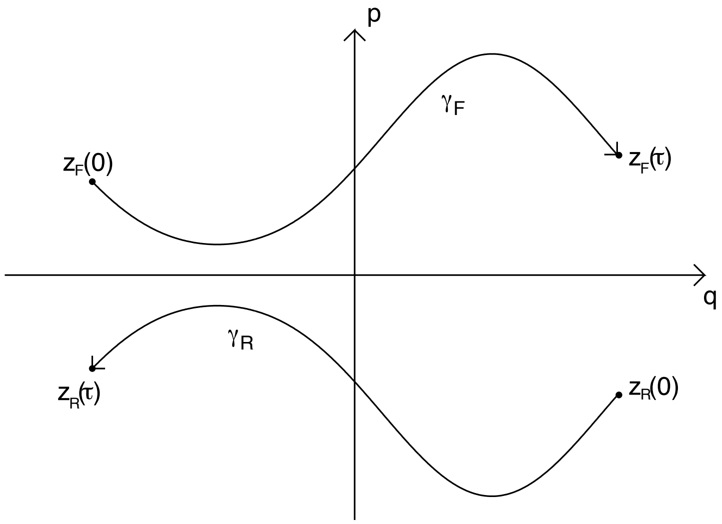

Figure 2.3: An illustration of a pair of conjugate trajectories in phase space. The trajectory γ_R was obtained by reflecting γ_F along the q axis. The arrows indicate the direction of time.

distributions $P_F(\gamma_F)$ and $P_R(\gamma_R)$ are identical. As the process becomes dissipative, $W_F^{typical} - \Delta F \gg \beta^{-1}$, where $W_F^{typical}$ denotes the work performed along a typical trajectory $\gamma_F^{typical}$ of the forward process, and it is difficult to observe the conjugate twin of $\gamma_F^{typical}$ in the reverse process. In other words, the asymmetry between the forward and reverse processes increases with dissipation. We note in passing that the fluctuation theorems of the form Eq. 2.11 reduce to the well known Green-Kubo relations, and Onsager reciprocity relations in the near equilibrium limit (slow rate of driving) and can be viewed as their extensions to processes occurring far from equilibrium [2, 61].

2.3 Other far from equilibrium estimators of ΔF

Eq. 2.11 leads to the following fluctuation theorem for the distribution of work values observed in the forward and reverse processes

$$\frac{P_F(W)}{P_R(-W)} = e^{\beta(W-\Delta F)}. \tag{2.12}$$

This is commonly referred to as the Crooks's fluctuation relation and allows us to construct a number of far from equilibrium estimators of ΔF. In particular, if $f(W)$ is some function of W, Eq 2.12 implies [18]

$$\frac{\langle f(W)\rangle_F}{\langle f(-W)e^{-\beta W}\rangle_R} = e^{-\beta\Delta F}. \tag{2.13}$$

The nonequilibrium work relation Eq 2.5 is now a special case of the theorem Eq 2.13 (with $f(W) = \exp(-\beta W)$). Bennett [6] studied similar generalizations of the FEP identity and solved for the functional form of f which minimizes the variance of the ΔF estimator. The same analysis can be used for Eq. 2.13. In particular, given n_F work values from the forward simulation, and n_R work values from the reverse simulation, Bennett showed that ΔF can be optimally estimated by choosing $f(W) = 1/(1 + \exp(\beta W + K))$

$$e^{-\beta \Delta F} = \frac{\langle 1/(1 + e^{\beta(W+K)}) \rangle_F}{\langle 1/(1 + e^{\beta(W-K)}) \rangle_R} e^{\beta K}, \tag{2.14}$$

where

$$K = -\Delta F + \beta^{-1} \ln n_F / n_R. \tag{2.15}$$

The footnote referenced following Eq. 2.20 explains why Bennett's approach is better than the conventional unidirectional estimators. Eq. 2.14 and Eq. 2.15 need to be solved recursively to obtain an estimate of ΔF [6]. In addition to estimating ΔF from the various identities, it is possible to estimate ΔF graphically from the work distributions. To do so, we follow Bennett's prescription [6], and obtain from Eq 2.12:

$$\left[\ln P_R(-W) + \beta \frac{W}{2} \right] - \left[\ln P_F(W) - \beta \frac{W}{2} \right] = \beta \Delta F \tag{2.16}$$

Hence, by plotting $L_2(W) \equiv [\ln P_R(-W) + \beta W/2]$, and $L_1(W) \equiv [\ln P_F(W) - \beta W/2]$ as functions of W, it is possible to graphically estimate $L_2(W) - L_1(W) = \Delta F$. This is also a useful and stringent consistency check for the fluctuation theorem, as it requires the difference of $L_2(W) - L_1(W)$ to be constant over the range of W values sampled in the simulation.

Nonequilibrium estimators of ΔF based on generalization of the umbrella sampling approach [98] to trajectories have also been developed. In these approaches, transition path sampling [8,20] is used to generate a biased ensemble of trajectories.

The free energy difference ΔF is estimated as an average over this biased trajectory ensemble [62, 95, 106]. The choice of the biasing function determines the efficiency of the estimator.

2.4 Computational Efficiency

While Eq. 2.5, and Eq. 2.14 in principle allow estimation of free energy differences from arbitrarily fast switching simulations, it is often not practically feasible to do so on account of the poor and slow convergence of Eq. 2.5 and Eq. 2.14 [51,58]. In this section, we will attempt to understand the reasons behind the poor efficiency of fast switching simulations.

Consider the estimate of ΔF from the forward (F) process using nonequilibrium work relation Eq. 2.5. The sampling requirements associated with Eq. 2.5 can be studied by rewriting it as follows [51]

$$\begin{aligned} 1 = \langle e^{-\beta(W-\Delta F)}\rangle_F &= \int dW P_F(W) e^{-\beta(W-\Delta F)} \\ &= \int dW P_R(-W) \end{aligned} \tag{2.17}$$

where we have used Eq 2.12. In order to get a reliable estimate of ΔF from the nonequilibrium work relation, it is important to sample work values from the region in which the integrand, $P_F(W)\exp(-\beta[W-\delta F]) = P_R(-W)$, is dominant [51, 58]. In other words, in order to obtain a reliable estimate of ΔF, regions typical to the distribution $P_R(-W)$ should be adequately represented in an ensemble of samples drawn from the distribution $P_F(W)$. Whenever the distribution $P_F(W)$ has a poor overlap with the distribution $P_R(-W)$, the work values that dominate the average

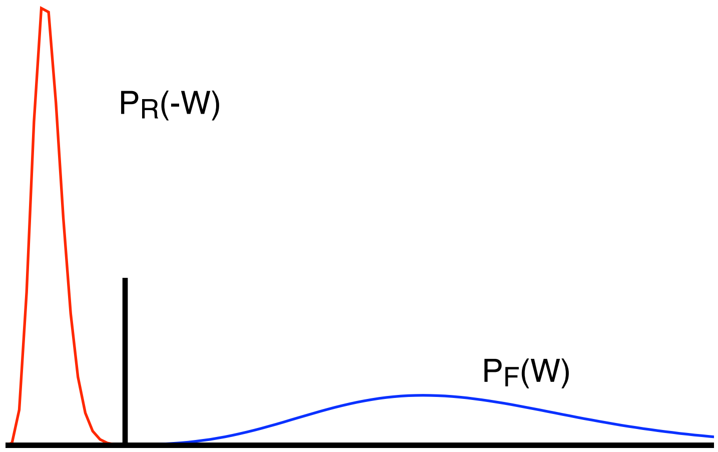

Figure 2.4: A schematic of work distributions observed in a fast switching simulation. The vertical line marks the point where the two distributions intersect ($W = \Delta F$). In the forward (F) simulation, work values, W, are typically sampled from the dominant region of $P_F(W)$. However, in order for the estimate of ΔF to converge, the dominant region of the distribution $P_R(-W)$ needs to adequately sampled. As the two typical regions are far apart, the estimate of ΔF suffers from poor convergence.

in Eq. 2.5 are sampled rarely and consequently it becomes difficult to obtain reliable estimates of ΔF (Fig. 2.4).

The problems with obtaining free energy estimates from fast switching nonequilibrium simulations are now apparent. As the system is driven further from equilibrium (by increasing the rate of switching, $\dot{\lambda(t)}$), the dissipation, $\langle W\rangle_F - \Delta F$, $\langle W\rangle_R + \Delta F$, increases in both the forward and reverse processes. , As we discussed previously, the dissipated work in turn reflects the extent to which realizations in the forward process differ from those obtained in the reverse process (after accounting for time reversal). In fact, the dissipated work can be related to an information theoretic quantification of the extent to which the distributions $P_F(W)\,, P_F(\gamma)$ differ from $P_R(-W), P_R(\tilde{\gamma})$ respectively [51].

$$\langle W\rangle_F - \Delta F = \beta^{-1} D[P_F(W)||P_R(-W)] = \beta^{-1} D[P_F(\gamma)||P_R(\tilde{\gamma})], \tag{2.18}$$

where

$$D[f||g] \equiv \int f \ln(f/g) \tag{2.19}$$

denotes the relative entropy or the Kullback Liebler Divergence between the distributions f and g [15]. The relative entropy between two distributions is non-negative and increases as the distributions become more distinct [15]. Thus, the increase in dissipation with the rate of switching is accompanied by an increase in the "separation" between the distributions $P_F(W)$ and $P_R(-W)$ and it becomes progressively harder to obtain reliable estimates of ΔF from Eq. 2.5. This argument can be made more quantitative [33, 51, 58] and it has been argued that the number of realizations N_s required to obtain a reliable estimate of ΔF from Eq. 2.5 in the forward process grows exponentially with the dissipation accompanying the time reversed process $N_s \sim \exp \beta[\langle W\rangle_R + \Delta F]$.

Other far from equilibrium estimators of ΔF such as Eq. 2.14 also converge poorly when λ is switched rapidly, on account of increasing dissipation and asymmetry between the forward and reverse processes. The sampling requirements associated with Bennett's Acceptance Ratio method (BAR) can be studied by rewriting Eq. 2.14 as [84]

$$\frac{\langle P_H(W)/P_F(W)\rangle_{P_F(W)}}{\langle P_H(W)/P_R(-W)\rangle_{P_R(-W)}} = 1, \tag{2.20}$$

where we have set $n_F = n_R$, $\langle \ldots \rangle_{P_F(W)}$ denotes an average over W values sampled from $P_F(W)$, $\langle \ldots \rangle_{P_R(-W)}$ denotes an average over W values sampled from $P_R(-W)$, $P_H(W) \equiv C^{-1} \frac{P_F(W) P_R(-W)}{P_F(W)+P_R(-W)}$ with $C = \int dW \frac{P_F(W) P_R(-W)}{P_F(W)+P_R(-W)}$ is the normalized harmonic mean distribution. Now, following the reasoning outlined in the paragraph after Eq. 2.17, we can infer that the estimate of ΔF from BAR will converge well if regions typical to the harmonic mean distribution P_H are sampled adequately from the distributions $P_F(W)$ and $P_R(-W)$ [2]. This becomes progressively harder as dissipation and time reversal asymmetry increase. In the umbrella sampling approach, an optimal choice of the biasing function is one for which the biased ensemble has an appreciable overlap with the ensembles corresponding to both $P_F(\gamma_F)$ and $P_R(\gamma_R)$ [62]. Constructing such a biasing function becomes difficult when the distributions $P_F(\gamma_F)$ and $P_R(\gamma_R)$ grow apart.

2.5 Summary

The connection between dissipation and time reversal asymmetry, two attributes of irreversible processes, has been used to argue that nonequilibrium fast switching estimates of ΔF suffer from poor convergence. In the next chapter, we will establish a relation between dissipation and another attribute of irreversible processes, namely the lag (recall Fig 1.1) that develops between the system and the

[2] Since the harmonic mean distribution straddles the two distributions ($P_F(W)$ and $P_R(-W)$), we expect this to be easier than sampling the typical regions of $P_F(W)$ from $P_R(-W)$ and vice versa. The bi-directional (data from both forward and reverse simulations is used) BAR estimator is hence generally more efficient than the so called unidirectional estimator Eq. 2.5.

equilibrium state as the external parameter is varied. This connection will prove useful in the subsequent chapters, Chapters 4, 6, where we will introduce methods that attempt to combat the problem of poor convergence due to dissipation by reducing the lag in nonequilibrium processes.

Chapter 3

Dissipation and lag

[1]Irreversible thermodynamic processes are those that cannot be undone mechanically: the system of interest and its surroundings never return to their original states. There are a number of attributes that we typically associate with such processes. These include (i) *dissipation* – the dispersal of energy among many degrees of freedom; (ii) *time-reversal asymmetry* – the evident directionality of time's arrow; and (iii) *broken equilibrium* – either within the system of interest, or between it and its thermal surroundings. For macroscopic systems these manifestations of irreversibility are related through the strict logic of the second law of thermodynamics.

For microscopic systems the second law must be interpreted statistically, making allowances for fluctuations around the mean behavior. Far from being uninteresting, uninformative "noise", such fluctuations have in recent years been found to

[1] This chapter is based on the publication: S. Vaikuntanathan, C. Jarzynski "Dissipation and Lag in Irreversible Processes", *Euro. Phys. Lett* **87**, *600005*, 2009.

satisfy a number of exact and unexpected relations. [9] These in turn have sharpened our understanding of the second law as it applies at the microscopic scale. (see Ref [53]) Of specific relevance for the present chapter is the discovery of quantitative relations between dissipation and time-reversal asymmetry, two of the above-mentioned manifestations of irreversibility. We have briefly discussed one such relation (Eq. 2.18) in the previous chapter and several such relations have appeared in the literature [26,51,55,65,66]. The central goal of the present chapter is to obtain a general relation (Eq. 3.1) between dissipation and (iii) the loss of equilibrium during an irreversible process.

Consider a process in which a system, initially at a temperature T ($\beta^{-1} = k_B T$), is driven away from equilibrium by varying an external parameter λ from A to B over a time interval $0 \leq t \leq \tau$. Let ρ_t^{eq} denote the equilibrium density corresponding to the value of the external parameter at time t. Although the system begins in equilibrium ($\rho_0 = \rho_0^{\mathrm{eq}}$), at later times $\rho_t \neq \rho_t^{\mathrm{eq}}$. This was illustrated schematically in Fig. 1.1: as λ is varied with time, the system tries to keep pace with – but ultimately lags behind – the continually changing equilibrium distribution. We can use the relative entropy [15], $D[\rho_t||\rho_t^{\mathrm{eq}}] = \int \rho_t \ln \rho_t/\rho_t^{\mathrm{eq}}$, to quantify this lag and measure the extent to which the system is out of equilibrium at time t. As mentioned in Sec 2.4 (see discussion following Eq. 2.18), the relative entropy $D[f||g]$ quantifies the extent to which the distribution f is distinguishable from the distribution g [15].

The central result of this chapter is the inequality

$$W_{\mathrm{diss}}(t) \geq \beta^{-1}\, D[\rho_t||\rho_t^{\mathrm{eq}}], \tag{3.1}$$

where $W_{\text{diss}}(t)$ is the amount of work dissipated up to time t during the process. Thus *the dissipated work dictates the maximum extent to which equilibrium can be broken* – equivalently, the maximum amount of lag – at a given instant during the process.

In this thesis, Eq. 3.1 will become important in the context of estimating free energy differences from nonequilibrium simulations. In particular, we will use Eq. 3.1 in the subsequent chapters to help guide the construction of efficient nonequilibrium estimators of free energy differences. Note that the connection between dissipation and lag has been heuristically well established in free energy estimation simulations [37, 104]. However, the relation derived here, is an exact quantitative relation and not a heuristic one.

We now derive our central result for systems driven arbitrarily far from equilibrium and then illustrate this result with an exactly solvable model system.

3.1 Theory

Following the framework setup in Sec 2.1, we consider a classical system described by a parameter-dependent Hamiltonian $H_\lambda(\mathbf{z})$ where $\mathbf{z}$ denotes a point in the phase space or coordinate space of the system. At fixed parameter value λ and temperature T, the equilibrium state of the system is described by the probability distribution,

$$\rho^{\text{eq}}(\mathbf{z}, \lambda) = \frac{e^{-\beta H_\lambda(\mathbf{z})}}{Z_\lambda}, \tag{3.2}$$

with free energy $F_\lambda = -\beta^{-1} \ln Z_\lambda$.

Imagine a process during which the system is initially brought to thermal equilibrium with a heat bath at temperature β^{-1}, at fixed $\lambda = A$, after which the external parameter is varied from $\lambda(0) = A$ to $\lambda(\tau) = B$ in a time τ. We will again assume that the evolution of the system during this process is governed by dynamics that are Markovian and *balanced*; that is, the equilibrium distribution (Eq. 3.2) is conserved when λ is held fixed. The time-dependent density $\rho_t = \rho(\mathbf{z}, t)$ describes an ensemble of trajectories evolving under these dynamics. This density can be expressed as

$$\rho(\mathbf{z}, t) = \langle \delta(\mathbf{z} - \mathbf{z}_t) \rangle, \tag{3.3}$$

where $\{\mathbf{z}_t\}$ denotes a trajectory, and $\langle \dots \rangle$ denotes an average over the ensemble of trajectories $\{\mathbf{z}_t\}$. An interesting property of such nonequilibrium processes is that if each trajectory $\{\mathbf{z}_t\}$ in the above average is assigned a time dependent statistical weight $\exp[-\beta(W(t) - \Delta F(t)]$ (see equation below) where $W(t)$ denotes the work performed along the trajectory up to a time t, and $\Delta F(t) = F_{\lambda(t)} - F_A$, then the equilibrium distribution $\rho_t^{\mathrm{eq}} \equiv \rho^{\mathrm{eq}}(\mathbf{z}, \lambda(t))$ is reconstructed [18, 42, 48]:

$$\rho^{\mathrm{eq}}(\mathbf{z}, \lambda(t)) \equiv \frac{e^{-\beta H_{\lambda(t)}(\mathbf{z})}}{Z_{\lambda(t)}} = \langle \delta(\mathbf{z} - \mathbf{z}_t) e^{-\beta[W(t) - \Delta F(t)]} \rangle, \tag{3.4}$$

where

$$W(t) \equiv \int_0^t \dot{\lambda} \frac{\partial H_\lambda(\mathbf{z}_{t'})}{\partial \lambda} dt'. \tag{3.5}$$

The above equation can be rewritten as

$$\rho^{\mathrm{eq}}(\mathbf{z}, \lambda(t)) = \frac{e^{-\beta H_{\lambda(t)}(\mathbf{z})}}{Z_{\lambda(t)}} = \rho(\mathbf{z}, t) \langle e^{-\beta[W(t) - \Delta F(t)]} \rangle_{\mathbf{z},t}, \tag{3.6}$$

where $\langle e^{-\beta W(t)} \rangle_{\mathbf{z},t} = \langle \delta(\mathbf{z} - \mathbf{z}_t) e^{-\beta[W(t) - \Delta F(t)]} \rangle / \rho(\mathbf{z}, t)$ and $\langle \dots \rangle_{\mathbf{z},t}$ can be interpreted as an average over all the trajectories that pass through $\mathbf{z}$ at t. Taking the logarithm of both sides of this equation, then invoking Jensen's inequality [15]

$$\langle e^{-\beta[W(t)-\Delta F(t)]}\rangle_{\mathbf{z},t} \geq e^{\langle -\beta[W(t)-\Delta F(t)]\rangle_{\mathbf{z},t}}, \tag{3.7}$$

we get

$$\langle W(t)\rangle_{\mathbf{z},t} - \Delta F(t) \geq \beta^{-1} \ln \frac{\rho(\mathbf{z},t)}{\rho^{\mathrm{eq}}(\mathbf{z},\lambda(t))}. \tag{3.8}$$

Finally, multiplying both sides of Eq 3.8 by ρ_t and integrating with respect to $\mathbf{z}$, we obtain

$$\langle W(t)\rangle - \Delta F(t) \geq \beta^{-1} \int d\mathbf{z}\, \rho_t \ln \frac{\rho_t}{\rho_t^{\mathrm{eq}}} \equiv \beta^{-1} D[\rho_t||\rho_t^{\mathrm{eq}}]. \tag{3.9}$$

Since the left side of this equation represents the work dissipated to time t, and the right side is the relative entropy of ρ_t with respect to ρ_t^{eq}, we have arrived at the central result (Eq. 3.1).

We now comment on a few aspects of this result.

First, the relative entropy $D[f||g]$ is always non-negative, and vanishes only if the distributions f and g are identical. Next, although the relative entropy is not symmetric with respect to the distributions f and g, it is useful to think of the relative entropy as a "distance" between the two distributions [15]. In particular, Stein's lemma [15] relates the value of $D[f||g]$ to the difficulty of statistically distinguishing between two distributions f and g. Thus $D[\rho_t||\rho_t^{\mathrm{eq}}]$ is an information theoretic measure of the lag, that is the deviation of the state of the system from the current equilibrium state at time t. By stipulating that the amount of work dissipated up to time t, $W_{diss}(t)$, must be no less than this measure of lag, Eq. 3.9 makes a statement that is stronger than the second law of thermodynamics, Eq. 2.4

($W_{diss} \geq 0$). In effect, the value $\beta^{-1} D[\rho_t || \rho_t^{\text{eq}}]$ represents a thermodynamic penalty for being out of equilibrium at time t [97].

It is worthwhile to discuss the deviation of the system from the equilibrium state in some detail, for two separate situations.

(a) If the system remains in contact with a heat bath as λ is switched from A to B, then as suggested by Fig. 1.1 we can picture the deviation of ρ_t from ρ_t^{eq} as a *lag* that develops because the system cannot keep pace with the changing equilibrium state [37, 82, 104]. Now Eq. 3.9 tells us that the dissipated work places an upper bound on this lag. In the special case that the parameter is varied quasistatically, and the heat bath is much larger than the system, then on general grounds we expect the system to remain in equilibrium, $\rho_t = \rho_t^{\text{eq}}$; in this case there is no dissipation, since $W(t) = \Delta F(t)$ for a reversible, isothermal process, and both sides of Eq. 3.9 become zero.

(b) If we instead imagine that, after using a heat bath to prepare the system in an initial state of equilibrium, the heat bath is disconnected prior to the actual switching process, then during the interval $0 \leq t \leq \tau$ the now-isolated system evolves under Hamilton's equations. As a result, a unique trajectory passes through any point $\mathbf{z}$ at time t, hence Eq. 3.7 becomes an equality and so does our central result:

$$W_{\text{diss}}(t) = \beta^{-1}\, D[\rho_t || \rho_t^{\text{eq}}]. \tag{3.10}$$

Since the system is not continually attempting to equilibrate with an external heat bath, it is not immediately natural to view the deviation of ρ_t from ρ_t^{eq} in terms of lag. (Indeed, even if λ is varied quasistatically, the distribution ρ_t will deviate from the isothermal, canonical distribution ρ_t^{eq} [19, 48, 76].) However, we can place this scenario within the "lag framework" by considering an isolated system to be a particular, limiting case of a system in contact with an external heat bath, in which the degree of thermal contact is so weak that the effects of the bath are negligible over a time interval of duration τ. If the external parameter is held fixed at $\lambda = B$ for $t > \tau$, then after a very long time the system does relax to a state of thermal equilibrium described by $\rho^{\mathrm{eq}}(\mathbf{z}, B)$. In this chapter we will adopt this perspective, and will view the relative entropy $D[\rho_t||\rho_t^{\mathrm{eq}}]$ as a quantitative measure of lag, even in the case of a thermally isolated system.

We finally note that when $t = \tau$, Eq. 3.1 is equivalent to a result derived by Kawai, Parrondo, and Van den Broeck [55] relating the dissipation to the time-reversal asymmetry.

3.2 Examples

We now illustrate Eq. 3.1 using a model that involves the quasistatic expansion or compression of a dilute gas of particles in d spatial dimensions. The model, shown in Fig. 3.1, is motivated by Refs. [19, 52]. The gas is a two-component mixture, in which component 1 is confined by the piston (open circles in Fig. 3.1), while the particles of component 2 pass freely through the piston (filled circles). Let λ denote the position of the piston, V_λ the volume of space to the left of the piston, V the total volume of the container, and N_1 and N_2 the numbers of particles in each component.

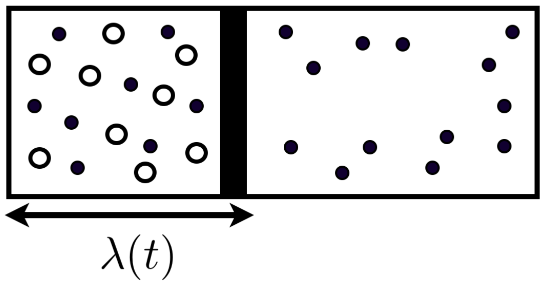

Figure 3.1: A two-component dilute gas, where component 1 (open circles) is confined by the piston, while component 2 (filled circles) is not.

For simplicity, we assume all particles have the same mass, m.

This mixture is initially allowed to come to thermal equilibrium with an external heat bath at temperature β^{-1}, with the piston held fixed at $\lambda = A$; then thermal contact between the gas and the external bath is broken; and finally, from $t = 0$ to $t = \tau$, component 1 undergoes compression or expansion as the piston is manipulated quasistatically according to a protocol $\lambda(t)$. During the latter stage the mixture evolves under Hamilton's equations in $2d(N_1 + N_2)$-dimensional phase space.

This particular model is convenient because it can be used to illustrate both scenarios (a) and (b) discussed in the previous section. If we define our system of interest to be the entire two-component mixture, then this model illustrates a system that is thermally isolated during the switching process, as per scenario (b). Alternatively, if we take our system of interest to be component 1, and view component 2 as part of a heat bath, then the model illustrates scenario (a). We will analyze these two cases below. We will solve explicitly for dissipated work and relative entropy in each case, and will show that our central result is the equality Eq. 3.10, in the case of a thermally isolated system of interest, and an inequality when the system remains in contact with a heat bath as in Eq. 3.1.

3.2.1 Hamiltonian Dynamics

Let $\mathbf{z} \equiv \{\mathbf{z}_1, \mathbf{z}_2\}$ denote a point in the full, $2d(N_1 + N_2)$-dimensional phase space, with $\mathbf{z}_1$ and $\mathbf{z}_2$ denoting the phase coordinates of components 1 and 2, respectively. The Hamiltonian for this system is $H_\lambda(\mathbf{z})$. As in Ref. [19], we take the term "dilute gas" to imply that, while particles do exchange energy via pairwise collisions, the mean free path between collisions is much greater than the characteristic distance between nearby particles. For practical purposes, we take this to mean that the particle-particle interaction terms in $H_\lambda(\mathbf{z})$ can be neglected in the calculations that follow. Thus H_λ is taken to be a sum of kinetic energies and hard-wall potentials that confine the two components to volumes V_λ and V. We also assume that when the piston is held fixed, the Hamiltonian dynamics are ergodic, i.e. the mixture is able to self-equilibrate via particle-particle collisions. Finally, the term "quasistatic" is meant to imply that the compression or expansion proceeds sufficiently slowly for continual self-equilibration to occur.

For fixed λ and positive energy value E, let $\phi_\lambda(E)$ denote the volume of phase space enclosed by the energy shell (i.e. surface of constant energy) $H_\lambda(\mathbf{z}) = E$; and let us think of $g_\lambda(E) \equiv \partial\phi/\partial E$ as the "surface area" of this shell. By explicit calculation, we have

$$\phi_\lambda(E) = \int d\mathbf{z}\,\theta(E - H_\lambda) = \mu^k V_\lambda^{N_1} V^{N_2} \frac{E^k}{k\Gamma(k)} \tag{3.11}$$

$$g_\lambda(E) = \int d\mathbf{z}\,\delta(E - H_\lambda) = \mu^k V_\lambda^{N_1} V^{N_2} \frac{E^{k-1}}{\Gamma(k)} \tag{3.12}$$

where

$$k = \frac{d}{2}(N_1 + N_2) \quad , \quad \mu = 2\pi m \quad , \tag{3.13}$$

and $\Gamma(k)$ is the gamma function. In deriving Eq. 3.11, we have used the well know expression for the volume of a many-dimensional sphere [29]. At temperature β^{-1}, the partition function and free energy of the mixture are:

$$Z_\lambda(\beta) = \int_0^\infty dE\, g_\lambda\, e^{-\beta E} = \mu^k V_\lambda^{N_1} V^{N_2} \beta^{-k} \tag{3.14a}$$

$$F_\lambda(\beta) = -\beta^{-1} \ln Z_\lambda \quad . \tag{3.14b}$$

When the piston is moved quasistatically from $\lambda(0) = A$ to $\lambda(\tau) = B$, the value of $\phi_\lambda(H_\lambda)$ is an adiabatic invariant [19]. By Eq. 3.11, this implies

$$V_A^{N_1} E_0^k = V_{\lambda(t)}^{N_1} E_t^k \tag{3.15}$$

along a trajectory $\{\mathbf{z}_t\}$ with energy $E_t \equiv H_{\lambda(t)}(\mathbf{z}_t)$. The work performed on the mixture is given by net change in its energy,

$$W(t) = E_t - E_0 = \left[\frac{V_A^{N_1/k}}{V_{\lambda(t)}^{N_1/k}} - 1 \right] E_0 \equiv \alpha(t) E_0. \tag{3.16}$$

Since $W(t)$ is determined uniquely by the initial energy, E_0, and initial conditions are sampled from the equilibrium distribution at temperature β^{-1}, we have:

$$\begin{aligned} \langle W(t) \rangle &= \frac{1}{Z_A} \int_0^\infty dE_0 g_A(E_0) e^{-\beta E_0} \alpha(t) E_0 \\ &= \alpha(t) \langle E_0 \rangle = k\beta^{-1} \alpha(t) \quad . \end{aligned} \tag{3.17}$$

Finally, from Eq. 3.14 we get

$$\Delta F(t) = N_1 \beta^{-1} \ln \frac{V_A}{V_{\lambda(t)}} = k\beta^{-1} \ln\left[\alpha(t) + 1\right] \quad . \tag{3.18}$$

From the first expression on the right is is clear that this quantity depends on N_1 but not on N_2; effectively, $\Delta F(t)$ specifies a free energy difference between two equilibrium states of component 1, as the equilibrium state of component 2 is unaffected by the piston.

Combining Eqs. 3.17 and 3.18 yields the following compact expression for the dissipated work:

$$W_{\text{diss}}(t) = k\beta^{-1}\left[\alpha - \ln(\alpha + 1)\right] \quad . \tag{3.19}$$

To compute $D[\rho_t || \rho_t^{\text{eq}}]$, we consider a trajectory $\{\mathbf{z}_t\}$ evolving under Hamilton's equations. By Liouville's theorem, the value of phase space density is conserved along this trajectory, hence

$$\rho(\mathbf{z}_t, t) = \rho(\mathbf{z}_0, 0) = \frac{1}{Z_A(\beta)} e^{-\beta E_0} = \frac{1}{Z_A(\beta)} e^{-\bar{\beta}_t E_t} \quad , \tag{3.20}$$

where $\bar{\beta}_t = \beta/[\alpha(t) + 1]$, and we have made use of Eq. 3.16. With Eq. 3.14a we can confirm that $Z_A(\beta) = Z_{\lambda(t)}(\bar{\beta}_t)$, thus

$$\rho(\mathbf{z}, t) = \frac{1}{Z_{\lambda(t)}(\bar{\beta}_t)} e^{-\bar{\beta}_t H_{\lambda(t)}(\mathbf{z})} \quad . \tag{3.21}$$

In other words, during the quasi-static compression or expansion process the phase space density is a canonical distribution with a slowly time-dependent temperature, $\bar{\beta}_t^{-1}$. By contrast, ρ^{eq} is defined at a constant temperature,

$$\rho^{\text{eq}}(\mathbf{z}, \lambda(t)) = \frac{1}{Z_{\lambda(t)}(\beta)} e^{-\beta H_{\lambda(t)}(\mathbf{z})}. \tag{3.22}$$

We therefore have

$$\ln \frac{\rho_t}{\rho_t^{\mathrm{eq}}} = \left(\beta - \bar{\beta}_t\right) H_{\lambda(t)}(\mathbf{z}) - k \ln \left(\beta / \bar{\beta}_t\right) . \tag{3.23}$$

Multiplying both sides by Eq. 3.21 and integrating, we get

$$\begin{aligned} D[\rho_t || \rho_t^{\mathrm{eq}}] &= \left(\beta - \bar{\beta}_t\right) k \bar{\beta}_t^{-1} - k \ln(\alpha + 1) \\ &= k \left[\alpha - \ln(\alpha + 1)\right] \quad . \end{aligned} \tag{3.24}$$

Comparing with Eq. 3.19, we see that Eq. 3.10 is satisfied.

3.2.2 Stochastic dynamics

Now let us view component 1 of our mixture as the system of interest, and component 2 as part of the heat bath. [2] The phase space of the system of interest is now $2dN_1$-dimensional, and evolution in this space is stochastic rather than deterministic, as the variables $\mathbf{z}_2$ have been projected out. We will use a carat (̂) to denote reduced phase space densities describing the system of interest (component 1):

$$\begin{aligned} \hat{\rho}_t &= \hat{\rho}(\mathbf{z}_1, t) = \int d\mathbf{z}_2 \, \rho(\mathbf{z}, t) \\ \hat{\rho}_t^{\mathrm{eq}} &= \hat{\rho}^{\mathrm{eq}}(\mathbf{z}_1, \lambda(t)) = \int d\mathbf{z}_2 \, \rho^{\mathrm{eq}}(\mathbf{z}, \lambda(t)) \quad . \end{aligned} \tag{3.25}$$

The relative entropy $D \left[\hat{\rho}_t || \hat{\rho}_t^{\mathrm{eq}}\right]$ quantifies the degree to which the system of interest is out of equilibrium (as before, “equilibrium” is defined by the temperature β^{-1}

[2] Thus the entire heat bath is composed of both the external bath used to prepare the initial state of equilibrium, and the particles of component 2, which remain in contact with the system of interest during the process.

and the current value of λ) and we wish to compare this with the dissipated work, $W_{\text{diss}}(t) = \langle W(t) \rangle - \Delta F(t)$.

Before proceeding further, we note that the stochastic evolution of the system of interest is non-Markovian, thus it is not immediately obvious that the analysis of Section 3.1 can be applied to this situation; see the assumptions stated after Eq. 3.2. To address these concerns, we verify in the following that Eq. 3.6 remains valid for the reduced densities, even though the evolution is non-Markovian. In the full phase space, $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$, of system ($\mathbf{z}_1$) and bath particles ($\mathbf{z}_2$), Eq. 3.6 can be rewritten as

$$\rho^{\text{eq}}(\mathbf{z}, \lambda(t)) e^{-\beta \Delta F(t)} = \rho(\mathbf{z}, t) e^{-\beta W(t)}, \tag{3.26}$$

where $W(t)$ is the work performed along the unique trajectory that passes through $\mathbf{z}$ at time t. Integrating both sides with respect to $\mathbf{z}_2$, we get

$$\hat{\rho}^{\text{eq}}(\mathbf{z}_1, \lambda(t)) e^{-\beta \Delta F(t)} = \hat{\rho}(\mathbf{z}_1, t) \left\langle e^{-\beta W(t)} \right\rangle_{\mathbf{z}_1, t}, \tag{3.27}$$

where we have used the fact that particles of the component 2 pass freely through the piston, and work performed depends only on the $\mathbf{z}_1$ degrees of freedom. Rearranging terms we see that the reduced densities satisfy Eq. 3.6.

Since the particles of component 2 pass freely through the piston, the values of $\langle W(t) \rangle$ and $\Delta F(t)$ are the same as before (see comment following Eq. 3.18). By contrast, since the reduced densities are obtained by projecting from the full phase space to that of component 1, there will be a reduction in the value of the relative entropy [15]: $D\left[\hat{\rho}_t || \hat{\rho}_t^{\text{eq}}\right] < D\left[\rho_t || \rho_t^{\text{eq}}\right]$, as we now confirm by direct evaluation.

Because ρ_t and ρ_t^{eq} are canonical distributions in the full phase space (Eqs. 3.21, 3.22), the reduced densities are also canonical:

$$\begin{aligned} \hat{\rho}_t &= \frac{1}{\hat{Z}_{\lambda(t)}(\bar{\beta}_t)} e^{-\bar{\beta}_t H^{(1)}_{\lambda(t)}(\mathbf{z}_1)} && (3.28) \\ \hat{\rho}^{\rm eq}_t &= \frac{1}{\hat{Z}_{\lambda(t)}(\beta)} e^{-\beta H^{(1)}_{\lambda(t)}(\mathbf{z}_1)}, && (3.29) \end{aligned}$$

where $H^{(1)}$ is the Hamiltonian for component 1, and

$$\hat{Z}_\lambda(\beta) = \mu^{k_1} V_\lambda^{N_1} \beta^{-k_1} \quad , \quad k_1 = dN_1/2. \tag{3.30}$$

We now have

$$\ln \frac{\hat{\rho}_t}{\hat{\rho}^{\rm eq}_t} = \left(\beta - \bar{\beta}_t\right) H^{(1)}_{\lambda(t)}(\mathbf{z}_1) - k_1 \ln\left(\beta/\bar{\beta}_t\right). \tag{3.31}$$

Multiplying by Eq. 3.28 and integrating, we obtain

$$\begin{aligned} D[\hat{\rho}_t||\hat{\rho}^{\rm eq}_t] &= \left(\beta - \bar{\beta}_t\right) k_1 \bar{\beta}_t^{-1} - k_1 \ln(\alpha+1) \\ &= k_1 \left[\alpha - \ln(\alpha+1)\right] && (3.32) \\ &= \frac{N_1}{N} \beta W_{\rm diss}(t) = \frac{N_1}{N} D[\rho_t||\rho^{\rm eq}_t], \end{aligned}$$

where $N = N_1 + N_2$ is the total number of particles in the mixture. [3] As expected, our central result (Eq. 3.1) now holds as a strict inequality.

Finally, let us consider what happens when component 2 is much larger than component 1; formally, $N_2 \to \infty$ with N_1 fixed. By straightforward evaluation we find

$$\begin{aligned} \bar{\beta}_t &= \beta + O(1/N) \\ W_{\rm diss}(t) &\sim 1/N && (3.33) \\ D[\hat{\rho}_t||\hat{\rho}^{\rm eq}_t] &\sim 1/N^2. \end{aligned}$$

[3] Eq. 3.32 is easy to understand: $D[\rho_t||\rho^{\rm eq}_t]$ is a sum of equal contributions from each of the N particles in the mixture, but only N_1 particles contribute to $D[\hat{\rho}_t||\hat{\rho}^{\rm eq}_t]$.

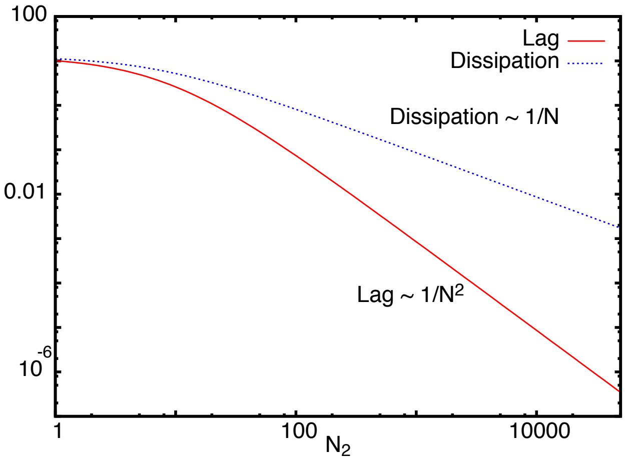

Figure 3.2: Dissipation ($\beta W_{\text{diss}}(t)$) and lag ($D[\hat{\rho}_t||\hat{\rho}_t^{\text{eq}}]$) are plotted as functions of N_2, with $N_1 = 10$, $d = 3$, $V_0/V_{\lambda(t)} = 5$, and $\beta = 1$. The isothermal limit is achieved as $N_2 \to \infty$.

Physically, this limit describes the reversible ($\dot{\lambda} \to 0$) *and isothermal* compression or expansion of component 1, with component 2 playing the role of an infinite heat bath. We see that both $W_{\text{diss}}(t)$ and $D[\hat{\rho}_t||\hat{\rho}_t^{\text{eq}}]$ approach zero, but at different rates, as illustrated in Fig. 3.2.

3.3 Summary

When a system is driven away from equilibrium by the variation of external parameters, the relative entropy $D[\rho_t||\rho_t^{\text{eq}}]$ quantifies the degree to which the current state of the system, $\rho(\mathbf{z}, t)$, lags behind the instantaneous equilibrium state, $\rho^{\text{eq}}(\mathbf{z}, \lambda(t))$. Our central result, Eq. 3.9, shows that the dissipated work, $W_{\text{diss}}(t)$, provides an upper bound on the value of this lag. In the special case that the dynamics of the system are Hamiltonian, the dissipation fully specifies the lag (Eq. 3.10).

These results complement analogous results obtained for the relationship between dissipated work and time-reversal asymmetry [55].

As we saw in the previous chapter, fast switching nonequilibrium estimates of ΔF suffer from poor convergence due to dissipation. Eq. 3.9 in turn relates the dissipation to the lag between the state of the system and the equilibrium state. In the subsequent chapters, we will use Eq. 3.9 to guide the development of methods that seek to improve the efficiency of nonequilibrium estimates of ΔF by reducing the lag in the nonequilibrium process.

Chapter 4

Escorted free energy simulations

4.1 Introduction

[1] In Chapter 2, we saw how estimators of ΔF based on nonequilibrium fluctuation theorems typically suffer from poor convergence whenever the external parameter is switched rapidly. The poor convergence is a consequence of high dissipation which, as we saw in Chapter 3, in turn reflects the lag that develops as the system is unable to keep pace with the equilibrium distribution corresponding to the changing external parameter. This chapter introduces a general strategy, *escorted free energy simulations*, for improving the efficiency of fast switching free energy estimates by reducing the lag. In our approach, the "physical" dynamics ordinarily used during

[1] This chapter is based on the following papers: S. Vaikuntanathan, C. Jarzynski "Escorted Free Energy Simulations: Improving Convergence by Reducing Dissipation", *Phys. Rev. Lett* **100**, *190601*, 2008, and S. Vaikuntanathan, C. Jarzynski "Escorted Free Energy Simulations", *J. Chem. Phys* **134**, *054107*, 2011.

a simulation are modified by the addition of artificial terms that *directly couple the evolution of the system coordinates* $\mathbf{z}$ *to variations in the external parameter,* λ. The central results are identities for ΔF in terms of trajectories generated with the modified dynamics. While these results are valid for an arbitrary choice of artificial dynamics (reducing to the usual nonequilibrium estimators of ΔF discussed in Chapter 2 when no artificial terms are used), the method is particularly effective when these dynamics are constructed so as to escort the system along a near-equilibrium path and reduce the lag. In particular, if the artificial dynamics entirely eliminates the above-mentioned lag, then our method provides a perfect estimator of the free energy difference: $W = \Delta F$ for every realization of the nonequilibrium process.

We begin by describing our strategy for nonequilibrium molecular dynamics simulations and deriving a generalization (Eq. 4.5) of the nonequilibrium work relation, Eq. 2.5. The idea is then extended to nonequilibrium Monte-Carlo simulations (see Eq. 4.39). In Section 4.4, we show that the escorted simulations satisfy a generalized version of Crooks's fluctuation theorem. This in turn allows us to combine our method with Bennett's Acceptance ratio method [6] which provides an optimal asymptotically unbiased estimator of ΔF (Eq. 4.64) [34, 92]. In Section 4.5, we show that while Eqs 4.5, 4.39, 4.64 are identities for all escorted simulations, they are particularly effective as estimators of ΔF when the modified dynamics successfully reduce the lag described above. In particular, we will demonstrate that if these terms eliminate the lag entirely, then Eqs 4.5, 4.39, 4.64 provide perfect (zero variance) estimators: $\mathrm{W} = \Delta F$ for every realization. Finally in Section 4.6, we illustrate the effectiveness of our approach on three model systems.

4.2 Molecular dynamics

As in Section 2.1, we will consider a classical system described by a parameter dependent Hamiltonian $H_\lambda(\mathbf{z})$ and imagine a process in which the system is prepared in a state of equilibrium at $\lambda = A$ and temperature T, $k_B T = \beta^{-1}$, (see Eq. 3.2) after which λ is switched from $\lambda(0) = A$ to $\lambda(\tau) = B$ in a time τ according to a specific protocol $\lambda(t)$. We will again be interested in computing the free energy difference $\Delta F = F_B - F_A$.

Let us suppose that we have a preferred set of equations of motion for simulating the evolution of the system, which we write in the generic form

$$\dot{\mathbf{z}} = \tilde{\mathbf{v}}(\mathbf{z}, \lambda), \tag{4.1}$$

where $\dot{\mathbf{z}} = \mathrm{d}\mathbf{z}/\mathrm{d}t$, and $\tilde{\mathbf{v}}(\mathbf{z}, \lambda)$ typically contains both deterministic and stochastic terms. Eq. 4.1 can be either stationary or explicitly time-dependent, according to whether we hold λ fixed or vary it with time. An ensemble of trajectories evolving under Eq. 4.1 will again be described by a phase space density $\rho(\mathbf{z}, t)$ satisfying a Liouville-type equation,

$$\frac{\partial \rho(\mathbf{z}, t)}{\partial t} = \mathcal{L}_\lambda \cdot \rho(\mathbf{z}, t). \tag{4.2}$$

For example, if Eq. 4.1 represents Hamilton's equations, then $\mathcal{L}_\lambda \rho(\mathbf{z}, t) = \{H_\lambda, \rho\}$ where $\{\,,\,\}$ denotes a Poisson bracket [32], and if Eq. 4.1 represents Langevin dynamics, then $\mathcal{L}_\lambda$ is the Fokker-Planck operator [54].

We will assume that $\mathcal{L}_\lambda \cdot e^{-\beta H_\lambda} = 0$ [42, 48], i.e. the dynamics preserve the equilibrium state when λ is fixed. We will use the term *physical dynamics* to refer to the evolution described by Eq. 4.1 at the single-trajectory level, or Eq. 4.2 at the ensemble level, to emphasize that these dynamics are intended to model, to some degree of realism, the microscopic evolution of our system of interest. For a system evolving according to some physical dynamics, the nonequilibrium work relation, Eq. 2.5, can be used to relate the work performed on the system in the process described above to the free energy difference ΔF.

Let us now suppose that we modify the "physical" equations of motion used to simulate the evolution of the system by adding a term proportional to $\dot{\lambda} = \mathrm{d}\lambda/\mathrm{d}t$:

$$\dot{\mathbf{z}} = \tilde{\mathbf{v}} + \dot{\lambda}\,\mathbf{u}, \tag{4.3}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{z}, \lambda)$ is an arbitrary, continuous vector field on phase space.[2] With this additional, artificial term, every small increment of the work parameter, $\mathrm{d}\lambda$, induces a phase-space displacement, $\mathrm{d}\mathbf{z} = \mathbf{u}\,\mathrm{d}\lambda$. Under these modified dynamics, the phase-space density $\rho(\mathbf{z}, t)$ satisfies

$$\frac{\partial \rho}{\partial t} = \mathcal{L}_\lambda \rho - \dot{\lambda} \nabla \cdot (\mathbf{u}\rho) \equiv \mathcal{L}'_{\lambda,\dot{\lambda}} \rho, \tag{4.4}$$

rather than Eq. 4.2, where the continuity term $-\dot{\lambda}\nabla \cdot (\mathbf{u}\rho)$ accounts for the flow $\dot{\lambda}\mathbf{u}$. Our aim is to use these modified dynamics to reduce lag and dissipation, and ultimately improve the efficiency of the free energy estimate.

When the system evolves under the artificial dynamics, Eq. 4.3, as λ is switched from $\lambda(0) = A$ to $\lambda(\tau) = B$ according to the protocol $\lambda(t)$, we will show that the following equality can be used to estimate ΔF:

$$e^{-\beta \Delta F} = \left\langle e^{-\beta W} \right\rangle_{\mathbf{u}} \quad , \tag{4.5}$$

[2] If $\mathbf{u}$ is not bounded, we must also impose a modest condition "at infinity", namely, $\lim_{z\to\infty} u e^{-\beta H} z^{d-1} = 0$.

where

$$W = \int_0^\tau \dot{\lambda} \left[\frac{\partial H}{\partial \lambda} + \mathbf{u} \cdot \nabla H - \beta^{-1} \nabla \cdot \mathbf{u} \right] \mathrm{d}t, \tag{4.6}$$

is interpreted as the *work* performed on a system evolving under Eq. (4.3), and $\langle \cdots \rangle_{\mathbf{u}}$ indicates an average over an ensemble of trajectories generated in the process, with initial conditions sampled from equilibrium.

It is instructive to derive this result first for the case in which the physical dynamics are Hamiltonian i.e. $\tilde{\mathbf{v}}$ describes Hamilton's equations of motion,

$$\frac{d\mathbf{q}}{dt} = \tilde{\mathbf{v}}_{\mathbf{q}} = \nabla_{\mathbf{p}} H_{\lambda(t)}, \tag{4.7}$$

$$\frac{d\mathbf{p}}{dt} = \tilde{\mathbf{v}}_{\mathbf{p}} = -\nabla_{\mathbf{q}} H_{\lambda(t)}, \tag{4.8}$$

where $\mathbf{z} \equiv (\mathbf{p}, \mathbf{q})$, $\tilde{\mathbf{v}} = (\tilde{\mathbf{v}}_q, \tilde{\mathbf{v}}_p)$, $\mathbf{q} = (\dots q_i \dots)$ is a vector composed of the position degrees of freedom q_i, $\mathbf{p} = (\dots p_i \dots)$ is a vector composed of the momentum degrees of freedom p_i, $\nabla_{\mathbf{p}} = (\dots \partial/\partial p_i \dots)$, and $\nabla_{\mathbf{q}} = (\dots \partial/\partial q_i \dots)$. We present this derivation in Eqs 4.11- 4.17 below, after which we extend the result to other choices of physical dynamics.

Let $\{\mathbf{z}_t\} = \{\mathbf{q}_t, \mathbf{p}_t\}$ denote a trajectory evolving under the modified dynamics,

$$\frac{d\mathbf{q}}{dt} = \nabla_{\mathbf{p}} H_{\lambda(t)} + \dot{\lambda} \mathbf{u}_{\mathbf{q}}, \tag{4.9}$$

$$\frac{d\mathbf{p}}{dt} = -\nabla_{\mathbf{q}} H_{\lambda(t)} + \dot{\lambda} \mathbf{u}_{\mathbf{p}}, \tag{4.10}$$

where $\mathbf{u}_{\mathbf{q}}(\mathbf{q}, \mathbf{p}, \lambda)$ and $\mathbf{u}_{\mathbf{p}}(\mathbf{q}, \mathbf{p}, \lambda)$ specify the components of the flow field $\mathbf{u} = (\mathbf{u}_{\mathbf{q}}, \mathbf{u}_{\mathbf{p}})$ that act on the position and momentum degrees of freedom respectively, as λ is varied from A to B. The modified dynamics are deterministic, allowing us to treat the final conditions as a functions of the initial conditions, $\mathbf{z}_\tau = \mathbf{z}_\tau(\mathbf{z}_0)$.

However, unlike Hamilton's equations, they do not preserve phase space volume: the degree of phase space expansion or compression along a trajectory is given by the Jacobian

$$\left|\frac{\partial \mathbf{z}_\tau}{\partial \mathbf{z}_0}\right| = \exp \int_0^\tau \dot{\lambda}\, \nabla \cdot \mathbf{u}(\mathbf{z}_t, \lambda(t))\, \mathrm{d}t, \tag{4.11}$$

where $\nabla = (\nabla_{\mathbf{q}}, \nabla_{\mathbf{p}})$. This Jacobian need not be unity. The total rate of change in the energy of the system as it evolves along this trajectory is

$$\frac{\mathrm{d}}{\mathrm{dt}} H_\lambda(\mathbf{z}_t) = \dot{\lambda}\frac{\partial H(\mathbf{z}_t)}{\partial \lambda} + (\tilde{\mathbf{v}} + \dot{\lambda}\mathbf{u}) \cdot \nabla H(\mathbf{z}_t) = \dot{\lambda}\left[\frac{\partial H(\mathbf{z}_t)}{\partial \lambda} + \mathbf{u} \cdot \nabla H(\mathbf{z}_t)\right], \tag{4.12}$$

where we have used

$$\tilde{\mathbf{v}} \cdot \nabla H = \tilde{\mathbf{v}}_{\mathbf{u}} \cdot \nabla H_{\mathbf{u}} + \tilde{\mathbf{v}}_{\mathbf{p}} \cdot \nabla H_{\mathbf{p}} = 0 \tag{4.13}$$

when $\tilde{\mathbf{v}}$ describes Hamilton's equations, Eqs. 4.7,4.8 [32]. Integrating Eq. 4.12 along the trajectory $\{\mathbf{z}_t\}$ relates the total change in the energy of the system along that trajectory to the integral of the first two terms in the definition of work in Eq. 4.6 [3].

$$H_{\lambda(\tau)}(\mathbf{z}_\tau) - H_{\lambda(0)}(\mathbf{z}_0) = \int_0^\tau \dot{\lambda}\left[\frac{\partial H(\mathbf{z}_t)}{\partial \lambda} + \mathbf{u} \cdot \nabla H(\mathbf{z}_t)\right] \mathrm{d}t \tag{4.14}$$

Let us now explicitly consider the ensemble average $\langle e^{-\beta W}\rangle_{\mathbf{u}}$ in Eq. 4.5. Since the dynamics are deterministic, and a trajectory can be uniquely specified by its initial point, this can written as

$$\langle e^{-\beta W}\rangle_{\mathbf{u}} = \int d\mathbf{z}_0 \rho(\mathbf{z}_0, 0) e^{-\beta \int_0^\tau \dot{\lambda}\, [\frac{\partial H}{\partial \lambda} + \mathbf{u}\cdot\nabla H - \beta^{-1}\nabla\cdot\mathbf{u}].\, \mathrm{d}t}, \tag{4.15}$$

[3] As the system does not exchange any heat with its surroundings in the process - the physical dynamics are Hamiltonian - this change in energy can be interpreted as the sum of the work done by switching the external parameter $\dot{\lambda}\frac{\partial H}{\partial \lambda}$, and the work done by the artificial dynamics $(\dot{\lambda}\mathbf{u}) \cdot \nabla H(\mathbf{z}_t)$

where the integral in the exponent is performed over the trajectory $\mathbf{z}_t$. Using $\rho(\mathbf{z}_0, 0) = \exp(-\beta H_{\lambda(0)}(\mathbf{z}_0))/Z_0$, and Eq. 4.14, we can rewrite Eq. 4.15 as

$$\langle e^{-\beta W} \rangle_{\mathbf{u}} = \int d\mathbf{z}_0 \frac{e^{-\beta H_{\lambda(\tau)}(\mathbf{z}_\tau)}}{Z_0} e^{\int_0^\tau \dot{\lambda} \nabla \cdot \mathbf{u}.\, \mathrm{d}t}. \tag{4.16}$$

Finally, changing the variable of integration from $\mathbf{z}_0$ to $\mathbf{z}_\tau$ and considering the associated Jacobian factor, Eq. 4.11, we get Eq. 4.5

$$\langle e^{-\beta W} \rangle_{\mathbf{u}} = \int d\mathbf{z}_\tau \frac{e^{-\beta H_{\lambda(\tau)}(\mathbf{z}_\tau)}}{Z_0} = e^{-\beta \Delta F}. \tag{4.17}$$

We will now extend this derivation of our central result, Eq. 4.5, to physical dynamics that satisfy $\mathcal{L}_\lambda \cdot e^{-\beta H_\lambda} = 0$, by generalizing the analysis of Hummer and Szabo [42] to include the $\dot{\lambda}$-dependent terms in Eqs. 4.3 and 4.4. From Eq. 4.4, we have

$$\mathcal{L}'_{\lambda,\dot{\lambda}}\, e^{-\beta H} = \beta \dot{\lambda} \left[\mathbf{u} \cdot (\nabla H) - \beta^{-1} (\nabla \cdot \mathbf{u}) \right] e^{-\beta H}. \tag{4.18}$$

Consider a density $g(\mathbf{z}, t)$ with initial condition $g(\mathbf{z}, 0) = \rho^{eq}(\mathbf{z}, \lambda(0))$, which satisfies the so-called *sink equation*,

$$\frac{\partial g}{\partial t} = \mathcal{L}'_{\lambda,\dot{\lambda}} g - \beta \dot{\lambda} \frac{\not{\partial} H}{\not{\partial} \lambda} g, \tag{4.19}$$

where we have introduced the compact notation

$$\frac{\not{\partial} H}{\not{\partial} \lambda}(\mathbf{z}, \lambda) \equiv \frac{\partial H}{\partial \lambda} + \mathbf{u} \cdot \nabla H - \beta^{-1} \nabla \cdot \mathbf{u}. \tag{4.20}$$

Using Eq. 4.18 we verify by inspection that the function

$$g(\mathbf{z}, t) = \frac{1}{Z_A}\, e^{-\beta H(\mathbf{z}, \lambda(t))} \tag{4.21}$$

is a solution of Eq. 4.19. Independently, sink equations of the kind Eq. 4.19 can be solved using the Feynman-Kac theorem, which provides a path-integral solution for $g(\mathbf{z},t)$ [27,42,43],

$$g(\mathbf{z},t) = \Big\langle \delta(\mathbf{z}-\mathbf{z}_t)\,\exp(-\beta w_t) \Big\rangle_{\mathbf{u}}, \tag{4.22}$$

where $w_t = \int_0^t \mathrm{d}t' \dot{\lambda}\,(\partial H/\partial\lambda)$. Again, $\{\mathbf{z}_t\}$ denotes a trajectory evolving under Eq. 4.3 as λ is varied from A to B; the integrand $\dot{\lambda}\,\partial H/\partial\lambda$ is evaluated along this trajectory. Equating these two solutions, we get

$$\frac{1}{Z_A}\exp\left[-\beta H(\mathbf{z},\lambda(t))\right] = \Big\langle \delta(\mathbf{z}-\mathbf{z}_t)\,\exp(-\beta w_t) \Big\rangle_{\mathbf{u}}, \tag{4.23}$$

Setting $t=\tau$ and integrating Eq. 4.23 over phase space, we obtain Eq. 4.5.

We have derived Eq. 4.5 by equating two solutions of the sink equation (Eq. 4.19): one obtained by inspection (Eq. 4.21), the other via path integration (right side of Eq. 4.23). An alternative derivation proceeds by first *defining* $g(\mathbf{z},t) = \langle\delta(\mathbf{z}-\mathbf{z}_t)\exp(-\beta w_t)\rangle_{\mathbf{u}}$, then showing that this function satisfies Eq. 4.19, whose solution is in turn given by Eq. 4.21. See Refs. [45,48] for analogous derivations of Eq. 2.5.

Eq. 4.5 implies we can estimate ΔF by taking the exponential average of W (Eq. 4.6), over trajectories evolving under the modified dynamics (Eq. 4.3). This generalizes the usual fast switching method: we recover Eq. 2.5 by choosing $\mathbf{u}=\mathbf{0}$. Our approach also contains elements of both the *metric scaling* [67] and *targeted perturbation* [50,63] strategies, reducing to a variant of the former in the case of linear flow fields, $\mathbf{u}=\alpha(\lambda)\,\mathbf{z}$, and to the latter in the limit of instantaneous switching, $\tau\to 0$. In that limit, the term $\tilde{\mathbf{v}}$ in Eq. 4.3 becomes negligible, and the trajectory evolves by integration along the flow field: $\mathrm{d}\mathbf{z}_\lambda/\mathrm{d}\lambda = \mathbf{u}(\mathbf{z}_\lambda,\lambda)$. We will revisit this point in the following section where we will be concerned with Monte-Carlo switching simulations.

4.3 Monte Carlo dynamics

Let us suppose that instead of Eq. 4.1, the evolution of the system is described by a discrete-time Monte Carlo algorithm, parametrized by the value of λ and defined by the transition probability $P_\lambda(\mathbf{z}|\mathbf{z}_0)$: if $\mathbf{z}_0$ represents the microstate of the system at one time step [4] then the next microstate $\mathbf{z}$ is sampled randomly from $P_\lambda(\mathbf{z}|\mathbf{z}_0)$. We assume this algorithm satisfies the conditions of *detailed balance*,

$$\frac{P_\lambda(\mathbf{z}|\mathbf{z}_0)}{P_\lambda(\mathbf{z}_0|\mathbf{z})} = \frac{e^{-\beta H_\lambda(\mathbf{z})}}{e^{-\beta H_\lambda(\mathbf{z}_0)}} \tag{4.24}$$

and *ergodicity* [54]. Routinely used Monte Carlo schemes such as the Metropolis algorithm [24] satisfy these conditions. Eq. 4.24 implies the somewhat weaker condition of *balance*,

$$\int d\mathbf{z}_0\, P_\lambda(\mathbf{z}|\mathbf{z}_0)\, e^{-\beta H_\lambda(\mathbf{z}_0)} = e^{-\beta H_\lambda(\mathbf{z})} \tag{4.25}$$

which we will use in the analysis below. With this Monte Carlo algorithm in place, we first describe a standard procedure for estimating ΔF using nonequilibrium simulations, Eq. 4.26 below, and then we introduce our modified version of this approach.

Imagine a process in which the system is initially prepared in equilibrium, at $\lambda = A$ and temperature β^{-1}, and then the system evolves under the Monte Carlo

[4] In a typical Monte-Carlo simulation, momentum degrees of freedom are not simulated. Hence, in the context of Monte-Carlo simulations, $\mathbf{z}$ will be used to denote a point in configuration space.

dynamics described above, as the value of λ is switched from A to B in N steps according to some pre-determined protocol. This evolution generates a trajectory $\gamma = \{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{N-1}\}$ that can be represented in more detail using the notation

$$[\mathbf{z}_0, \lambda_0] \Rightarrow [\mathbf{z}_0, \lambda_1] \rightarrow [\mathbf{z}_1, \lambda_1] \Rightarrow \cdots \rightarrow [\mathbf{z}_{N-1}, \lambda_{N-1}] \Rightarrow [\mathbf{z}_{N-1}, \lambda_N]. \tag{4.26}$$

Here, the symbol $\Rightarrow$ denotes an update in the value of λ, with the microstate held fixed, while $\rightarrow$ denotes a Monte Carlo step at fixed λ, e.g. the microstate $\mathbf{z}_1$ is sampled from the distribution $P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0)$. Moreover,

$$\lambda_0 \equiv A \qquad , \qquad \lambda_N \equiv B, \tag{4.27}$$

and the initial point $\mathbf{z}_0$ is sampled from $\rho_{eq}(\mathbf{z}_0, A)$.

Because it is specified by the sequence of microstates $\mathbf{z}_0, \cdots \mathbf{z}_{N-1}$, the trajectory γ can be viewed as a point in a dN-dimensional trajectory space, where d is dimensionality of phase (or configuration) space, with $d\gamma = d\mathbf{z}_0 \cdots d\mathbf{z}_{N-1}$. For the process described in the previous paragraph, the probability density for generating this trajectory is

$$p[\gamma] = P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}_{N-2}) \cdots P_{\lambda_2}(\mathbf{z}_2|\mathbf{z}_1)\, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0)\, \rho^{\mathrm{eq}}(\mathbf{z}_0, A) \tag{4.28}$$

where the factors $P_{\lambda_i}(\mathbf{z}_i|\mathbf{z}_{i-1})$ in this equation (read from right to left) correspond to the symbols $\rightarrow$ in Eq. 4.26 (read from left to right). The work performed on the system during this process is the sum of energy changes due to updates in λ, [16, 44, 48, 85]

$$W[\gamma] = \sum_{i=0}^{i=N-1} \delta W_i \equiv \sum_{i=0}^{i=N-1} \left[H_{\lambda_{i+1}}(\mathbf{z}_i) - H_{\lambda_i}(\mathbf{z}_i) \right]. \tag{4.29}$$

Using Eqs. 4.24, 4.28 and 4.29, we arrive at the nonequilibrium work relation for Monte Carlo dynamics, Eq. 2.5 [16, 48]. As mentioned previously, however, this average converges poorly when the process is highly dissipative.

To address the issue of poor convergence, let us now assume that for every integer $0 \leq i < N$, we have a deterministic function $M_i : \mathbf{z} \to \mathbf{z}'$ that takes any point $\mathbf{z}$ in configuration space and maps it to a point $\mathbf{z}'$. We assume that each of these functions is invertible (M_i^{-1} exists), but otherwise the functions are arbitrary. These M_i's then constitute a set of *bijective mappings*, which we use to modify the procedure for generating trajectories, as follows. When the value of the work parameter is switched from λ_i to λ_{i+1}, the configuration space coordinates are simultaneously subjected to the mapping M_i. These deterministic functions play the role of the flow fields introduced in the previous section. This connection is apparent for instance when the consecutive values of the external parameter, λ_i and λ_{i+1} differ by an infinitesimal amount, $\lambda_{i+1} - \lambda_i = \delta\lambda$. Then the mapping M_i can be generically written as

$$\mathbf{z}' = M_i(\mathbf{z}) = \mathbf{z} + \mathbf{u}(\mathbf{z}, \lambda_i)\delta\lambda, \tag{4.30}$$

where $\mathbf{u}$ again denotes a vector flow field. In other words, changes in λ induce a phase-space displacement of $\mathbf{u}\delta\lambda$ just like in the escorted equations of motion, Eq. 4.3.

With the mapping transformations M_i, Eq. 4.26 becomes

$$[\mathbf{z}_0, \lambda_0] \overset{M_0}{\Rightarrow} [\mathbf{z}'_0, \lambda_1] \to [\mathbf{z}_1, \lambda_1] \overset{M_1}{\Rightarrow} \cdots \to [\mathbf{z}_{N-1}, \lambda_{N-1}] \overset{M_{N-1}}{\Rightarrow} [\mathbf{z}'_{N-1}, \lambda_N] \tag{4.31}$$

where

$$\mathbf{z}'_i \equiv M_i(\mathbf{z}_i), \tag{4.32}$$

as indicated by the notation $\overset{M_i}{\Rightarrow}$. (As before, the symbol → denotes a Monte Carlo move at fixed λ.) The bijective maps effectively escort the system by directly coupling increments in λ to changes in the microstate.

In the escorted trajectory (Eq. 4.31), the system visits a sequence of $2N$ points in configuration space: the N "primary" microstates $\mathbf{z}_0, \cdots \mathbf{z}_{N-1}$, alternating with the N "secondary" microstates $\mathbf{z}'_0, \cdots \mathbf{z}'_{N-1}$. Since each $\mathbf{z}'_i$ is uniquely determined from $\mathbf{z}_i$ (Eq. 4.32), the sequence of primary microstates $\gamma = \{\mathbf{z}_0, \cdots \mathbf{z}_{N-1}\}$ fully specifies the trajectory; that is, trajectory space remains dN-dimensional, with $d\gamma = d\mathbf{z}_0 \cdots d\mathbf{z}_{N-1}$. The probability density for generating a trajectory γ is given by the following modification of Eq. 4.28:

$$p[\gamma] = P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}'_{N-2}) \cdots P_{\lambda_2}(\mathbf{z}_2|\mathbf{z}'_1)\, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0)\, \rho^{\mathrm{eq}}(\mathbf{z}_0, A) \tag{4.33}$$

Taking a cue from Refs [50,67], and following Eq. 4.6, we define the work performed on the system as it evolves along the escorted trajectory Eq. 4.33 as

$$W[\gamma] = \sum_{i=0}^{N-1} \delta W_i \equiv \sum_{i=0}^{N-1} \left[H_{\lambda_{i+1}}(\mathbf{z}'_i) - H_{\lambda_i}(\mathbf{z}_i) - \beta^{-1} \ln J_i(\mathbf{z}_i) \right], \tag{4.34}$$

where $J_i(\mathbf{z}) = |\partial \mathbf{z}'/\partial \mathbf{z}|$ is the Jacobian associated with the map $M_i : \mathbf{z} \to \mathbf{z}'$. Averaging $\exp(-\beta W[\gamma])$ over the ensemble of trajectories, we have

$$\begin{aligned} \langle e^{-\beta W} \rangle \quad &= \quad \int d\gamma\, p[\gamma]\, e^{-\beta W[\gamma]} \\ = \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1} \quad \cdots \quad &\int d\mathbf{z}_0\, e^{-\beta \sum_{i=0}^{N-1} \delta W_i}\, P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}'_{N-2}) \times \ldots \\ \times \ldots \quad &P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0)\, e^{-\beta H_{\lambda_0}(\mathbf{z}_0)} \end{aligned}$$

To evaluate this expression, we first identify all factors in the integrand that do not depend on $\mathbf{z}_0$ or $\mathbf{z}_0'$, and we pull these outside the innermost integral, $\int d\mathbf{z}_0$, which gives us (for that integral):

$$\begin{aligned} & \int d\mathbf{z}_0 \, e^{-\beta\delta W_0} \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0') \, e^{-\beta H_{\lambda_0}(\mathbf{z}_0)} && (4.35) \\ = & \int d\mathbf{z}_0 \, J_0(\mathbf{z}_0) \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0') \, e^{-\beta H_{\lambda_1}(\mathbf{z}_0')} && (4.36) \\ = & \int d\mathbf{z}_0' \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0') \, e^{-\beta H_{\lambda_1}(\mathbf{z}_0')} = e^{-\beta H_{\lambda_1}(\mathbf{z}_1)} && (4.37) \end{aligned}$$

We have used Eq. 4.34 to get to the second line, followed by a change in the variables of integration to get to the third line, $d\mathbf{z}_0 \, J_0(\mathbf{z}_0) \to d\mathbf{z}_0'$, and we have invoked Eq. 4.25 to arrive at the final result. This process can be repeated for the integrals $\int d\mathbf{z}_1$ to $\int d\mathbf{z}_{N-2}$, which brings us to:

$$\begin{aligned} \langle e^{-\beta W} \rangle &= \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1} \, e^{-\beta\delta W_{N-1}} \, e^{-\beta H_{\lambda_{N-1}}(\mathbf{z}_{N-1})} \\ &= \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1} \, J_{N-1}(\mathbf{z}_{N-1}) \, e^{-\beta H_{\lambda_N}(\mathbf{z}_{N-1}')} \qquad (4.38) \\ &= \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1}' \, e^{-\beta H_{\lambda_N}(\mathbf{z}_{N-1}')} = \frac{Z_{\lambda_N}}{Z_{\lambda_0}}, \end{aligned}$$

and therefore

$$\langle e^{-\beta W} \rangle = e^{-\beta \Delta F}. \qquad (4.39)$$

This equation is an identity for ΔF in terms of escorted trajectories, generated as per Eq. 4.31. For the special case in which each mapping is the identity, $M_i = I$, we recover the usual scheme, Eq. 4.26, and then Eq. 4.39 reduces to the nonequilibrium work relation, Eq. 2.5.

When $N = 1$, i.e. when the external parameter is switched in one step (sudden switching), the escorted trajectory described in Eq. 4.31 reduces to

$$[\mathbf{z}_0, \lambda_0] \overset{M_0}{\Rightarrow} [\mathbf{z}_0', \lambda_1] \tag{4.40}$$

where $\lambda_0 = A$ and $\lambda_1 = B$. Since $\mathbf{z}_0'$ is uniquely determined by $\mathbf{z}_0$, the average in Eq. 4.39 when $N = 1$ is simply an average over the initial points $\mathbf{z}_0$ which are sampled from the equilibrium ensemble A,

$$\langle \exp\left[-\beta\left(H_B(\mathbf{z}_0') - H_A(\mathbf{z}_0) - \beta^{-1}\ln J_0(\mathbf{z}_0)\right)\right]\rangle_A = e^{-\beta\Delta F}, \tag{4.41}$$

where as before $\langle\ldots\rangle_A$ denotes an average over the equilibrium state A. The above relation was first derived by Jarzynski [50] and is referred to as the Targeted Free Energy Perturbation identity (TFEP). This identity is a generalization of the FEP identity much in the same way as Eq. 4.39 (and Eq. 4.5) is a generalization of the nonequilibrium work relation, Eq. 2.5. Indeed, when $M_0 = I$, the TFEP identity, Eq. 4.41, reduces to the FEP identity.

4.4 Fluctuation Theorem

Let us now consider not only the switching process described by Eq. 4.3, which we will henceforth designate the *forward* process, but also its time-reversed analogue, the *reverse* process. In the reverse process, the system is prepared in equilibrium at $\lambda = B$ and temperature β^{-1}. The work parameter is then switched to $\lambda = A$ according to the time reversed protocol, $\tilde{\lambda}(t) = \lambda(\tau - t)$. The equations of motion now read

$$\dot{\mathbf{z}} = \tilde{\mathbf{v}} + \dot{\tilde{\lambda}}\,\mathbf{u} = \tilde{\mathbf{v}} - \dot{\lambda}\,\mathbf{u}, \tag{4.42}$$

where we have used $\dot{\tilde{\lambda}} = -\dot{\lambda}$. [5] In this section, we will compare distributions of W in the forward and reverse processes and show that they also satisfy the Crooks's fluctuation theorem (Eq. 2.12). Again, we will begin by deriving the result for the case when the physical dynamics are Hamiltonian (Eq. 4.43-4.49 below).

We start by considering the work performed on the system as it evolves along a trajectory $\tilde{\gamma} \equiv \{\tilde{\mathbf{z}}_t\}$ in the reverse process:

$$W_R(\tilde{\mathbf{z}}_0) = \int_0^\tau \dot{\tilde{\lambda}} \left[\frac{\partial H}{\partial \lambda} + \mathbf{u} \cdot \nabla H - \beta^{-1} \nabla \cdot \mathbf{u} \right] \mathrm{d}t, \tag{4.43}$$

where the integral is along the trajectory $\{\tilde{\mathbf{z}}_t\}$. In general, the work performed on the system will depend on the entire trajectory. However, since the dynamics are deterministic, specifying the initial conditions is sufficient to describe the entire trajectory.

Let us now construct the density $P_F(W)$, where just as before (see Section 2.3), $P_F(W)$ denotes the probability distribution of work values in the forward process,

$$P_F(W) = \int d\,\mathbf{z}_0 \rho^{eq}(\mathbf{z}_0, \lambda(0)) \delta(W_F(\mathbf{z}_0) - W), \tag{4.44}$$

where $\mathbf{z}_0$ is the initial point of the trajectory $\{\mathbf{z}_t\}$. Using Eq. 4.11 to change the variables of integraion, we can rewrite Eq. 4.44 as

$$P_F(W) = \int d\,\mathbf{z}_\tau \rho^{eq}(\mathbf{z}_0, \lambda(0)) \delta(W_F(\mathbf{z}_0) - W) e^{-\int_0^\tau \dot{\lambda} \nabla \cdot \mathbf{u}(\mathbf{z}_t, \lambda(t))\, \mathrm{d}t}. \tag{4.45}$$

[5] We will assume that the Hamiltonian is invariant under time reversal. This invariance is broken for instance when the system evolves in the presence of a magnetic field. In such cases, the appropriate time reversed process is one where the signs of both $\dot{\lambda}$ and the magnetic field are inverted.

Using $\rho^{eq}(\mathbf{z}_0, \lambda(0))/\rho^{eq}(\mathbf{z}_\tau, \lambda(\tau)) = e^{\beta(H_{\lambda(\tau)}(\mathbf{z}_\tau) - H_{\lambda(0)}(\mathbf{z}_0) - \Delta F)}$ and Eq. 4.14 we obtain

$$P_F(W) = \int d\,\mathbf{z}_\tau \rho^{eq}(\mathbf{z}_\tau, \lambda(\tau))\delta(W_F(\mathbf{z}_0) - W)e^{\beta(W_F(\mathbf{z}_0) - \Delta F)}. \tag{4.46}$$

If a trajectory γ is a solution of the equations of motion in the forward process, its conjugate twin $\gamma^* = \{\mathbf{z}^*_{\tau-t}\}$ is solution of the time reversed equations of motion in the reverse process [6].The work performed along the conjugate trajectory $\gamma^* = \{\mathbf{z}^*_{\tau-t}\}$ in the reverse process, $W_R(\mathbf{z}^*_\tau)$, is related to the work performed along the trajectory γ in the forward process by

$$W_F(\mathbf{z}_0) = -W_R(\mathbf{z}^*_\tau). \tag{4.47}$$

Finally, changing the variables of integration in Eq. 4.46 from $\mathbf{z}_\tau$ to $\mathbf{z}^*_\tau$ (the Jacobian for this transformation is unity) we obtain

$$P_F(W) = e^{\beta(W - \Delta F)} \int d\,\mathbf{z}^*_\tau \rho^{eq}(\mathbf{z}^*_\tau, \lambda(\tau))\delta(-W_R(\mathbf{z}^*_\tau) - W) \tag{4.48}$$

The integral in the above equation is simply $P_R(-W)$, where $P_R(W)$ denotes the distribution of work values in the reverse process. Rearranging the terms, we obtain the Crooks's fluctuation relation for escorted simulations,

$$\frac{P_F(W)}{P_R(-W)} = e^{\beta(W - \Delta F)}. \tag{4.49}$$

In Appendix A we sketch a general derivation of this fluctuation relation for escorted simulations.

[6] The notation γ^* refers to the conjugate trajectory of γ and is obtained by both reversing the order of phase-space points visited in γ and inverting the momentum degrees of freedom in each of these phase-space points. The notation $\mathbf{z}^*_t$ refers to the phase-space point obtained by inverting the momentum degrees of freedom in the phase-space point $\mathbf{z}_t$.

In the case of Monte-Carlo simulations, the work parameter is switched to $\lambda = A$ from $\lambda = B$ in N steps in the reverse process, following a sequence $\{\tilde{\lambda}_0, \tilde{\lambda}_1, \cdots, \tilde{\lambda}_N\}$ that is the reversal of the protocol used during the forward process:

$$\tilde{\lambda}_i \equiv \lambda_{N-i} \tag{4.50}$$

During the reverse process, changes in λ are coupled to the system's evolution through the inverse mapping functions, $\tilde{M}_i \equiv M^{-1}_{N-1-i}$, generating a trajectory

$$[\tilde{\mathbf{z}}'_{N-1}, \tilde{\lambda}_N] \overset{\tilde{M}_{N-1}}{\Leftarrow} [\tilde{\mathbf{z}}_{N-1}, \tilde{\lambda}_{N-1}] \leftarrow \cdots \overset{\tilde{M}_1}{\Leftarrow} [\tilde{\mathbf{z}}_1, \tilde{\lambda}_1] \leftarrow [\tilde{\mathbf{z}}'_0, \tilde{\lambda}_1] \overset{\tilde{M}_0}{\Leftarrow} [\tilde{\mathbf{z}}_0, \tilde{\lambda}_0] \tag{4.51}$$

where $\tilde{\mathbf{z}}'_i \equiv \tilde{M}_i(\tilde{\mathbf{z}}_i)$, and the initial state $\tilde{\mathbf{z}}_0$ is sampled from ρ^B_{eq}. The direction of the arrows indicates the progression of time. The probability density for obtaining a trajectory $\tilde{\gamma} = \{\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{N-1}\}$ is

$$p[\tilde{\gamma}] = P_{\tilde{\lambda}_{N-1}}(\tilde{\mathbf{z}}_{N-1}|\tilde{\mathbf{z}}'_{N-2}), \cdots P_{\tilde{\lambda}_2}(\tilde{\mathbf{z}}_2|\tilde{\mathbf{z}}'_1)\, P_{\tilde{\lambda}_1}(\tilde{\mathbf{z}}_1|\tilde{\mathbf{z}}'_0)\, \rho^{\mathrm{eq}}(\tilde{\mathbf{z}}_0, B) \tag{4.52}$$

with $d\tilde{\gamma} = d\tilde{\mathbf{z}}_0 \cdots d\tilde{\mathbf{z}}_{N-1}$. Following Eq. 4.34, the work performed during this process is

$$W_R[\tilde{\gamma}] = \sum_{i=0}^{N-1} \left[H_{\tilde{\lambda}_{i+1}}(\tilde{\mathbf{z}}'_i) - H_{\tilde{\lambda}_i}(\tilde{\mathbf{z}}_i) - \beta^{-1} \ln \tilde{J}_i(\tilde{\mathbf{z}}_i) \right], \tag{4.53}$$

where $\tilde{J}_i(\tilde{\mathbf{z}}) = |\partial \tilde{\mathbf{z}}' / \partial \tilde{\mathbf{z}}|$ is the Jacobian for the mapping $\tilde{M}_i$.

To establish Eq. 4.49 for Monte-Carlo escorted simulations, we will again need to consider a *conjugate pair* of trajectories, γ and γ^* [7], related by time-reversal.

[7] In Monte-Carlo simulations, $\mathbf{z}_i$ represents a point in configuration space which is invariant under time reversal and not full phase space. Hence, we have not used the notation $\mathbf{z}^*_i$ in the conjugate trajectory.

Specifically, if $\gamma = \{\mathbf{z}_0, \cdots \mathbf{z}_{N-1}\}_F$ is a trajectory generated during the forward process, that visits the sequence of microstates

$$\mathbf{z}_0 \overset{M_0}{\Rightarrow} \mathbf{z}'_0 \rightarrow \mathbf{z}_1 \overset{M_1}{\Rightarrow} \mathbf{z}'_1 \rightarrow \cdots \rightarrow \mathbf{z}_{N-1} \overset{M_{N-1}}{\Rightarrow} \mathbf{z}'_{N-1} \quad , \tag{4.54}$$

then its conjugate twin, $\gamma^* = \{\mathbf{z}'_{N-1}, \cdots \mathbf{z}'_0\}_R$, generated during the reverse process, visits the same microstates, in reverse order:

$$\mathbf{z}_0 \overset{\tilde{M}_{N-1}}{\Leftarrow} \mathbf{z}'_0 \leftarrow \mathbf{z}_1 \overset{\tilde{M}_{N-2}}{\Leftarrow} \mathbf{z}'_1 \leftarrow \cdots \leftarrow \mathbf{z}_{N-1} \overset{\tilde{M}_0}{\Leftarrow} \mathbf{z}'_{N-1} \tag{4.55}$$

that is $\tilde{\mathbf{z}}_i = \mathbf{z}'_{N-1-i}$ and $\tilde{\mathbf{z}}'_i = \mathbf{z}_{N-1-i}$ (see Eq. 4.51). Note that the primary microstates of γ are the secondary microstates of γ^*, and vice-versa, and the work function is odd under time-reversal:

$$W_F[\gamma] = -W_R[\gamma^*]. \tag{4.56}$$

We wish to evaluate the quantity

$$P_F(W)\, e^{-\beta(W-\Delta F)} = \int d\gamma\, p_F[\gamma]\, e^{-\beta(W_F[\gamma]-\Delta F)}\, \delta(W - W_F[\gamma]) \tag{4.57}$$

with $p_F[\gamma]$ given by Eq. 4.33. To this end, we first decompose $W_F[\gamma]$ as follows:

$$W_F[\gamma] = \Delta E_F[\gamma] - Q_F[\gamma] - \beta^{-1} S_F[\gamma], \tag{4.58}$$

where

$$\Delta E_F[\gamma] \equiv H_{\lambda_N}(\mathbf{z}'_{N-1}) - H_{\lambda_0}(\mathbf{z}_0) \tag{4.59a}$$

$$Q_F[\gamma] \equiv \sum_{i=1}^{N-1} \left[H_{\lambda_i}(\mathbf{z}_i) - H_{\lambda_i}(\mathbf{z}'_{i-1}) \right] \tag{4.59b}$$

$$S_F[\gamma] \equiv \sum_{i=0}^{N-1} \ln J_{\lambda_i}(\mathbf{z}_i) = \ln \prod_{i=0}^{N-1} \left| \frac{\partial \mathbf{z}'_i}{\partial \mathbf{z}_i} \right| = \ln \left| \frac{\partial \gamma^*}{\partial \gamma} \right| \tag{4.59c}$$

Here $\Delta E_F[\gamma]$ is the total change in the energy of the system as it evolves along the trajectory γ, $Q_F[\gamma]$ can be interpreted as the heat transfered to the system from the reservoir [67], and $S_F[\gamma]$ is an entropy-like term, which arises because the mappings M_i need not preserve volume. The quantities defined in Eq. 4.59 satisfy the properties

$$\begin{aligned}
P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}'_{N-2})\cdots P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0) &= P_{\lambda_{N-1}}(\mathbf{z}'_{N-2}|\mathbf{z}_{N-1})\cdots \times \\
&\times \;\; \cdots \; P_{\lambda_1}(\mathbf{z}'_0|\mathbf{z}_1)\, e^{-\beta Q_F[\gamma]} \\
\rho^{eq}(\mathbf{z}_0,\lambda_0) &= \rho^{eq}(\mathbf{z}'_{N-1},\lambda_N)\, e^{\beta(\Delta E_F[\gamma]-\Delta F)}
\end{aligned}$$

where we have used the detailed balance condition Eq. 4.24. These properties then give us

$$\begin{aligned}
p_F[\gamma] &= P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}'_{N-2})\cdots P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0)\,\rho^{\mathrm{eq}}(\mathbf{z}_0,\lambda_0) \\
&= P_{\lambda_{N-1}}(\mathbf{z}'_{N-2}|\mathbf{z}_{N-1})\cdots P_{\lambda_1}(\mathbf{z}'_0|\mathbf{z}_1)\, e^{-\beta Q_F[\gamma]} \\
&\quad \times\; \rho^{eq}(\mathbf{z}'_{N-1},\lambda_N)\, e^{\beta(\Delta E_F[\gamma]-\Delta F)} \\
&= p_R[\gamma^*]\, e^{\beta(W_F[\gamma]-\Delta F)}\, e^{S_F[\gamma]}
\end{aligned} \tag{4.61}$$

hence

$$p_F[\gamma]\, e^{-\beta(W_F[\gamma]-\Delta F)} = p_R[\gamma^*]\left|\frac{\partial\gamma^*}{\partial\gamma}\right| \tag{4.62}$$

Substituting this result into the integrand on the right side of Eq. 4.57, then changing the variables of integration from $d\gamma$ to $d\gamma^*$, and invoking Eq. 4.56, we finally arrive at the result we set out to establish.

$$P_F(W)\, e^{-\beta(W-\Delta F)} = P_R(-W) \tag{4.63}$$

We have explicitly used the stronger detailed balance condition in this proof. When the Monte-Carlo dynamics only satisfy the weaker balance condition, Eq. 4.25, the steady state at a particular value of λ supports a nonzero current [54]. In such cases, the time reversed process should be performed with Monte-Carlo dynamics that support a steady state current with the same magnitude but opposite sign [17]. The fluctuation theorem, Eq. 4.63, then remains valid [17].

As we have seen in Section 2.3, the fluctuation theorems allow us to construct a number of far-from-equilibrium estimators of ΔF. In particular, given n_F work values from the forward escorted simulation, and n_R work values from the reverse escorted simulation, we can optimally estimate ΔF using Bennett's Acceptance Ratio (BAR) method (Eq. 2.14) just as we would in the case of the usual nonequilibrium simulations. Also, we can use Eq. 2.16 as a means to both graphically estimate ΔF and as a consistency check for the fluctuation theorem. Since we repeatedly use Eq. 2.14 and Eq. 2.16 in next sections, we have reproduced the equations (Eq. 4.64 is the BAR estimator and Eq. 4.66 describes the procedure to graphically estimate ΔF) below for convenience.

$$e^{-\beta \Delta F} = \frac{\langle 1/(1+e^{\beta(W+K)}) \rangle_F}{\langle 1/(1+e^{\beta(W-K)}) \rangle_R} e^{\beta K}, \tag{4.64}$$

where

$$K = -\Delta F + \beta^{-1} \ln n_F/n_R. \tag{4.65}$$

Eq. 4.64 and Eq. 4.65 need to be solved recursively to obtain an estimate of ΔF. The free energy difference can be graphically estimated using

$$L_2(W) - L_1(W) = \beta \Delta F, \tag{4.66}$$

where $L_2(W) \equiv \{\ln P_R(-W) + \beta W/2\}$, and $L_1(W) \equiv \{\ln P_F(W) - \beta W/2\}$.

4.5 Computational efficiency and figures of merit

While Eqs 4.5, 4.39, Eq. 4.64 are valid for any set of flow fields or bijective mapping functions (depending on whether the simulation uses continuous time molecular dynamics or discrete time Monte-Carlo) the efficiency of using escorted simulations to estimate ΔF depends strongly on the choice of these functions. In the previous chapter, we established a relation between dissipation and lag for systems driven away from equilibrium (Eq. 3.9). Starting from Eq. 4.23 (and its analogous version in escorted Monte-Carlo simulations), that result can be derived even for escorted free energy simulations. We reproduce the result below (in the context of the forward process) for convenience,

$$\langle W \rangle_F - \Delta F \geq \beta^{-1} D[\rho_f || \rho_B^{eq}] \tag{4.67}$$

where $\langle W \rangle_F - \Delta F$ measures the total dissipation in the forward escorted simulation, ρ_f denotes the density of the system at the end of the switching process, and ρ_B^{eq} denotes the equilibrium density corresponding to the value of λ at the end of the switching process. The relative entropy $D[\rho_f || \rho_B^{eq}]$ quantifies the lag between the state of the system and equilibrium state at the end of the process. Eq. 4.67 is an equality if the dynamics are deterministic or if lag is eliminated and the system is in equilibrium throughout (see below).

Since the convergence of exponential averages such as Eq. 4.5 and Eq. 4.39, deteriorates rapidly with $\langle W \rangle$ [33, 51, 58], which as a result of Eq. 4.67 can be correlated to the lag, it is reasonable to speculate that a choice of dynamics that decreases the lag will improve the convergence of estimator of ΔF.

To pursue this idea, let us imagine for a moment that we are able to construct a *perfect* flow field, $\mathbf{u}^*$, that eliminates the lag entirely. In this case the distribution $\rho(\mathbf{z}, t) = \rho^{\mathrm{eq}}(\mathbf{z}, \lambda_t)$ is a solution of Eq. (4.4). Substituting this solution into Eq. (4.4), we get, using $\mathcal{L}_\lambda \cdot \rho^{\mathrm{eq}} = 0$,

$$\frac{\partial \rho^{\mathrm{eq}}}{\partial \lambda} + \nabla \cdot (\mathbf{u}^* \rho^{\mathrm{eq}}) = 0. \tag{4.68}$$

Setting $\rho^{\mathrm{eq}} = e^{\beta(F-H)}$, we obtain

$$\frac{\mathrm{d}F}{\mathrm{d}\lambda}(\lambda) = \frac{\partial H}{\partial \lambda} + \mathbf{u}^* \cdot \nabla H - \beta^{-1} \nabla \cdot \mathbf{u}^* \equiv \frac{\not{\partial} H}{\not{\partial} \lambda}, \tag{4.69}$$

therefore

$$W_F = \int_0^\tau \dot{\lambda} \frac{\not{\partial} H}{\not{\partial} \lambda} \, \mathrm{d}t = \int_0^\tau \dot{\lambda} \frac{\mathrm{d}F}{\mathrm{d}\lambda} \, \mathrm{d}t = \Delta F \tag{4.70}$$

for every trajectory $\mathbf{z}_t$. Thus, for a perfect flow field $\mathbf{u}^*$, there is no dissipation ($W_{\mathrm{diss}} = 0$) and *a single trajectory provides the correct free energy difference.*

In the case of Monte-Carlo simulations, a perfect set of mappings $\{M_i^*\}$ that eliminate the lag also eliminate dissipation. Under this set of mappings, the equilibrium distribution $\rho^{eq}(\mathbf{z}, \lambda_i)$ transforms to the distribution $\rho^{eq}(\mathbf{z}', \lambda_{i+1})$ [50], in other words

$$\rho^{eq}(\mathbf{z}', \lambda_{i+1}) = \frac{\rho^{eq}(\mathbf{z}, \lambda_i)}{J^*_{\lambda_i}(\mathbf{z})} \tag{4.71}$$

[Under a bijective map $M : \mathbf{x} \to \mathbf{y}$, a distribution $f(\mathbf{x})$ is transformed to the distribution $\eta(\mathbf{y}) = f(\mathbf{x})/J(\mathbf{x})$, where $J(\mathbf{x}) = |\partial \mathbf{y}/\partial \mathbf{x}|$.] Using $\rho^{eq}(\mathbf{z}, \lambda) = e^{\beta(F_\lambda - H_\lambda(\mathbf{z}))}$, and taking the logarithm of both sides of Eq. 4.71, we obtain (for a perfect set of mappings)

$$\delta W_i \equiv H_{\lambda_{i+1}}(\mathbf{z}') - H_{\lambda_i}(\mathbf{z}) - \beta^{-1} \ln J^*_{\lambda_i}(\mathbf{z}) = F_{\lambda_{i+1}} - F_{\lambda_i}, \tag{4.72}$$

hence $W_F[\gamma] = \Delta F$ *for every trajectory* γ (Eq. 4.34) and dissipation is eliminated [8].

Although on general grounds we expect that perfect flow fields and a perfect set of mapping functions typically exist,[9] it seems unlikely we will be able to solve for $\mathbf{u}^*$ or $\{M_i^*\}$ analytically, apart from a few simple systems. Indeed, Eq. 4.69 (Eq. 4.71) suggests that an expression for $\mathrm{d}F/\mathrm{d}\lambda$ $(F_{\lambda_{i+1}} - F_{\lambda_i})$ is required to obtain $\mathbf{u}^*$ $(\{M_i^*\})$. However, by revealing that *elimination* of the lag results in a zero-variance estimator of ΔF, Eq. 4.70 and Eq. 4.72 support our earlier speculation: if we can construct artificial dynamics that *reduce* the lag, then we should expect improved convergence of the exponential average. In such cases the dissipation accompanying the escorted simulations is less than that for the unescorted simulations, leading to improved convergence of the free energy estimate.

As an example of a strategy that can be used to construct good flow fields and mappings, consider a system of identical, mutually interacting particles, in an external potential $U_\lambda(\mathbf{r})$:

$$H_\lambda(\mathbf{z}) = \sum_k U_\lambda(\mathbf{r}_k) + \sum_{k<l} V(\mathbf{r}_k, \mathbf{r}_l) \tag{4.73}$$

The probability distribution of a single, tagged particle is then given by the single-

[8] It is straightforward to show that when escorted dynamics eliminate the lag and dissipation in the forward process, the lag and dissipation are also eliminated in the reverse escorted processes.

[9] Since Eq. (4.68) is of the form $\nabla \cdot \mathbf{A} = q(\mathbf{z}, \lambda)$, a formal solution can be constructed using Green's functions.

particle density

$$\rho_\lambda^{(1)}(\mathbf{r}) = \frac{1}{Z_\lambda} \int d\mathbf{z}\, \delta[\mathbf{r}_k(\mathbf{z}) - \mathbf{r}]\, e^{-\beta H_\lambda(\mathbf{z})} \tag{4.74}$$

where $\mathbf{r}_k(\mathbf{z})$ specifies the coordinates of the tagged particle as a function of the microstate $\mathbf{z}$. Now consider a reference system of non-interacting particles, described by a Hamiltonian

$$\bar{H}_\lambda(\mathbf{z}) = \sum_k \bar{U}_\lambda(\mathbf{r}_k) \tag{4.75}$$

with a similarly defined single-particle density $\bar{\rho}_\lambda^{(1)}(\mathbf{r})$; and imagine that $\bar{U}_\lambda$ is chosen so that these single-particle densities are identical or nearly identical: $\rho_\lambda^{(1)}(\mathbf{r}) \approx \bar{\rho}_\lambda^{(1)}(\mathbf{r})$. In this case a set of mappings $\{M_i\}$ or flow fields that are perfect or near-perfect for the reference system $(\bar{H}_\lambda)$, might be quite effective in reducing lag in the original system (H_λ). We will illustrate this mean-field-like approach in Section 4.6.3, and we note that a similar strategy was explored by Hahn and Then in the context of targeted free energy perturbation [34].

It will be useful to develop a figure of merit, allowing us to compare the efficiency of our method for different sets of mappings or flow fields. One approach would be simply to compare the error bars associated with the statistical fluctuations in the respective free energy estimates. Unfortunately, estimates of ΔF obtained from convex nonlinear averages such as the one obtained from Eq. 4.39, are systematically biased for any finite number of realizations [33,104,107]. Following [104,107], consider for example the estimate of ΔF, $\Delta\mathcal{F}_{N_s}$, obtained from a particular set of N_s simulations,

$$\Delta\mathcal{F}_{N_s} = -\beta^{-1} \ln \frac{1}{N_s} \sum_{i=1}^{N_s} e^{-\beta W_i}. \tag{4.76}$$

The average $\Delta F_{N_s} \equiv \langle \Delta \mathcal{F}_{N_s} \rangle$ over all such sets of N_s simulations is systematically biased for any finite N_s whenever the simulation is performed irreversibly, $\Delta F_{N_s} > \Delta F$. This can be easily verified by applying Jensen's inequality to Eq. 4.76 above. This bias can be large, and as a result the statistical error bars by themselves might not be sufficiently reliable to quantify the efficiency of the mapping. In the following paragraphs we discuss alternative figures of merit.

We begin by noting that when the unidirectional estimator, Eqs. 4.5, 4.39 is used in conjunction with simulations of the forward process, then the number of realizations (N_s) required to obtain a reliable estimate of ΔF is roughly given by [51, 58]

$$N_s \sim e^{\beta(\langle W \rangle_R + \Delta F)} \tag{4.77}$$

where $\langle W \rangle_R + \Delta F$ is the dissipation accompanying the reverse process. While this provides some intuition for the convergence of Eqs. 4.5, 4.39, its usefulness as a figure of merit is somewhat limited as it requires simulations of both the forward and the reverse processes, and in that case we are better off using a bidirectional estimator such as Eq. 4.64.

When we do have simulations of both processes, then an easily computed figure of merit is the hysteresis, $\langle W_{diss} \rangle_F + \langle W_{diss} \rangle_R = \langle W \rangle_F + \langle W \rangle_R$. The value of this quantity is zero if the mappings or flow fields are perfect, otherwise it is positive. It is interesting to note that the hysteresis can be related to an information-theoretic measure of overlap between the forward and reverse work distributions $P_F(W)$ and

$P_R(-W)$: [22]

$$D[P_F||P_R] + D[P_R||P_F] = \beta(\langle W\rangle_F + \langle W\rangle_R). \tag{4.78}$$

Here $D[p||q] \equiv \int p\ln(p/q) \geq 0$ denotes the relative entropy between the distributions p and q, and the symmetrized quantity $D[p||q] + D[q||p]$ (also known as the *Jeffreys divergence* [15]) provides a measure of the difference, or more precisely the lack of overlap, between the distributions. The right side of Eq. 4.78 can be estimated from a modest sample of forward and reverse simulations. If the artificial dynamics reduce the hysteresis, $\langle W\rangle_F + \langle W\rangle_R$, then this indicates increased overlap between the work distributions, and therefore improved convergence [51].

When $n_F = n_R = N_s \gg 1$, the mean square error of the Bennett estimator is [6, 34, 35, 92]

$$\langle (F_{BAR}^{est} - \Delta F)^2\rangle = \frac{2}{\beta^2 N_s}\left(\frac{1}{2C} - 1\right). \tag{4.79}$$

Here F_{BAR}^{est} denotes the estimate of ΔF obtained from Eq. 4.64, and

$$C \equiv \int dW \frac{P_F(W)P_R(-W)}{P_F(W) + P_R(-W)} = \left\langle \frac{1}{1+\exp[\beta(W-\Delta F)]}\right\rangle_F = \left\langle \frac{1}{1+\exp[\beta(W+\Delta F)]}\right\rangle_R \tag{4.80}$$

(This result can be generalized to the case $n_F \neq n_R$ [34].) As discussed by Bennett [6] and Hahn and Then [34, 35], the value of C measures the overlap between $P_F(W)$ and $P_R(-W)$, and provides a rough figure of merit for the Bennett estimator. When lag is eliminated and the two distributions coincide, then C attains its maximum value, $C = 1/2$, whereas when there is poor overlap, $C \approx 0$. Thus we expect that the higher the value of the overlap function C, the smaller the number of realizations N_s required to estimate ΔF from Eq. 4.64 with a prescribed accuracy. Indeed, Eq. 4.79

suggests a lower bound on the number of realizations needed to achieve a mean square error less than β^{-2}: $N_s > 1/C$. Note that since C is an ensemble average (Eq. 4.80), it can readily be estimated from available simulation data.

In Appendix B, we derive an upper bound on the number of realizations needed to obtain a reliable estimate of ΔF using Bennett's method, N_s (Eq. B.7). Combining these bounds gives us

$$\frac{1}{C} < N_s < \frac{1}{C^2} \tag{4.81}$$

While Eq. 4.81 does not provide a good estimate for N_s [10], it does allow us to argue heuristically that whenever the artificial dynamics succeed in increasing the value of C, the convergence of the Bennett estimator is improved. We will illustrate this point in the following section.

4.6 Examples

4.6.1 One dimensional model system

Consider Sun's one-dimensional model system [95],

$$H(p,q,\lambda) = \frac{p^2}{2m} + q^4 - 16(1-\lambda)q^2 = \frac{p^2}{2m} + V(q,\lambda). \tag{4.82}$$

For $A \equiv 0 \leq \lambda < 1 \equiv B$, the potential energy profile $V(q,\lambda)$ is a double well, with minima at $\pm q_0(\lambda) \equiv \pm\sqrt{8(1-\lambda)}$ separated by a barrier of height $64(1-\lambda)^2$

[10] For $C << 1$, the upper and lower bounds in Eq. 4.81 can be orders of magnitude apart. Nevertheless, Eq. 4.81 can serve as a good consistency check for the quality of the estimates. For example an estimate of ΔF using Bennett's method from a data set of size $N_s \sim 10^6$ is reliable if $C \sim 0.001$.

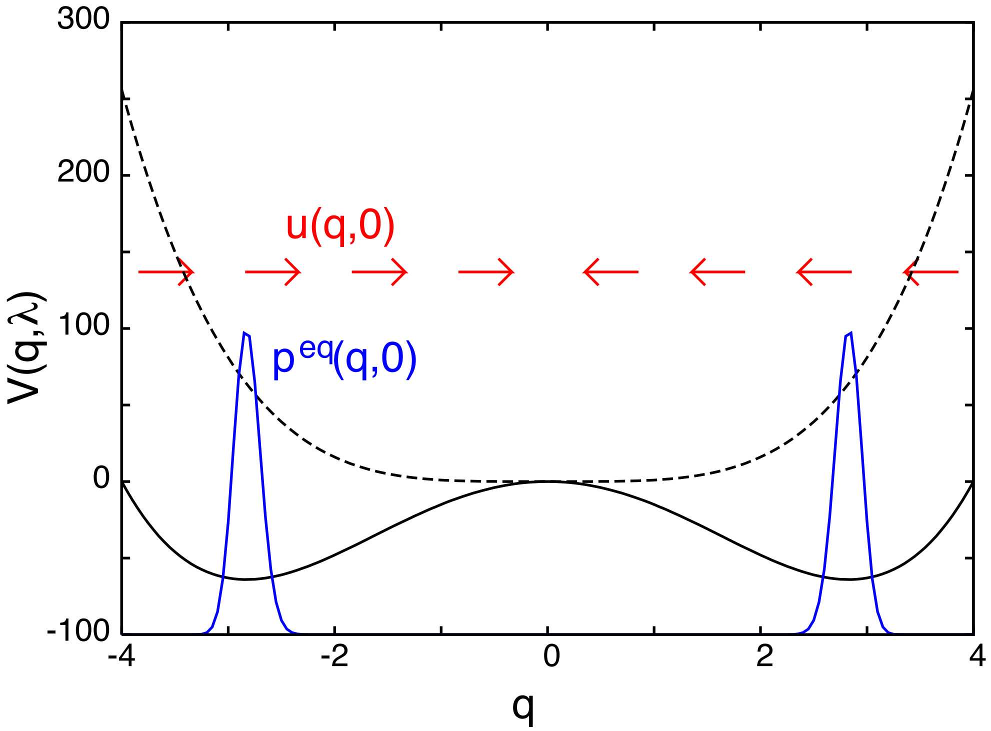

Figure 4.1: The potential energy landscape for $\lambda = 0$ (solid line) and $\lambda = 1$ (dashed line). Also depicted are the equilibrium distribution and the flow field, at $\lambda = 0$.

(Fig. 4.1). Setting $\beta = 1$, the equilibrium distribution is bimodal and sharply peaked around $\pm q_0$; as $\lambda \to 1$ the two peaks coalesce as V becomes a single, quartic well. Analytical evaluation of the partition functions gives $\Delta F = F_B - F_A = 62.94...$ [63].

The direct application of nonequilibrium work relation, Eq. 2.5, to this model gives poor results when the switching is performed rapidly [76, 95]. A typical simulation begins with the system near $\pm q_0(0)$; then, as λ is varied from 0 to 1, the two minima at $\pm q_0(\lambda)$ approach one another, but the system lags behind, resulting in large dissipation and poor free energy estimates. This is illustrated by the open circles in Fig. 4.2, obtained from simulations during which the system evolved under Hamilton's equations, integrated using the velocity Verlet algorithm. Only for $\tau = 1$ does Eq. (2.5) provide an accurate estimate of ΔF. (The *systematic* error evident in Fig. 4.2 arises after taking the logarithm of both sides of Eq. (2.5) [107]; see the discussion following Eq. 4.76.)

To illustrate the application of Eq. (4.5), let us take

$$u(q,\lambda) = \frac{\mathrm{d}q_0}{\mathrm{d}\lambda} \tanh\left[64(1-\lambda)q_0 q\right], \tag{4.83}$$

with $q_0 = q_0(\lambda)$ as given above. This field acts only on the coordinate q, and not on the momentum p. We arrived at Eq. (4.83) by using crude approximations to estimate the solution of Eq. (4.69), modeling p^{eq} as a pair of Gaussians. Omitting the details of this calculation, we note that near either peak of p^{eq}, $u(q,\lambda)$ displaces the system toward the origin at a speed $\dot{\lambda}|u| \approx \dot{\lambda}\,\mathrm{d}q_0/\mathrm{d}\lambda$ (see Fig. 4.1). This is the speed at which the two minima of $V(q,\lambda)$ approach the origin. Intuitively, we expect this flow to reduce the lag between ρ and p^{eq}. Moreover, the dynamics are deterministic and the connection between dissipation and lag in Eq. 4.67 is expressed as an equality. Consequently any reduction in lag will directly lead to a reduction in dissipation.

We repeated the simulations described above, now adding the term $\dot{\lambda}\,u$ to the dynamics. The resulting estimates of ΔF, obtained using Eq. (4.5) and depicted as filled circles in Fig. 4.2, are remarkably accurate over the entire range of switching times. Indeed, for all $\tau = 0.01, \cdots, 1.0$, the work values W were sharply peaked

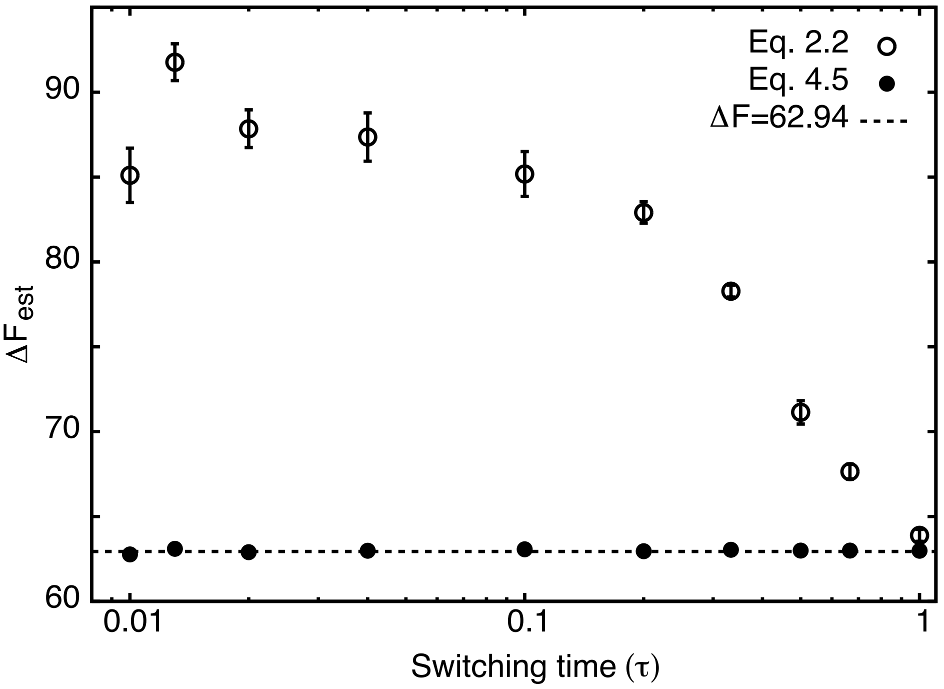

Figure 4.2: Comparison of estimates of ΔF using Eqs. 2.5 and 4.5. We performed simulations for switching times ranging from $\tau = 0.01$ to $\tau = 1.0$. Each $\Delta F_{\rm est}$ was obtained using 10^6 trajectories, evolving under either Eq. 4.1 (open circles) or Eqs. 4.3, 4.83 (filled circles). Error bars were computed using the bootstrap method [21]; for the filled circles these were smaller than the symbols, and are not shown.

around ΔF (data not shown), confirming that dissipation is greatly reduced and that the flow field escorts the system through a sequence of near-equilibrium states, even when λ is switched rapidly. We stress, however, that this choice of flow field is neither perfect ($u \neq u^*$) nor unique. In particular, we expect it could be improved near $\lambda = 1$, where the approximations made on the way to Eq. (4.83) break down.

4.6.2 Cavity Expansion

As a second example, we estimate the free energy cost associated with growing a hard-sphere solute in a fluid. Consider a system composed of n_p point particles inside a cubic container of volume L^3 (L is the length of a side of the cube), centered at the origin with periodic boundaries. The particles are excluded from a spherical region of radius R, also centered at the origin. The particles interact with one another via the WCA pairwise interaction potential [24] which is denoted by $V(\mathbf{r}_k, \mathbf{r}_l)$. The energy of the system at a microstate $\mathbf{z} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{n_p})$ is given by

$$H_R(\mathbf{z}) = \Theta(\mathbf{z}, R) + \sum_{k=1}^{n_p-1} \sum_{l>k}^{n_p} V(\mathbf{r}_k, \mathbf{r}_l) \tag{4.84}$$

where $\Theta(\mathbf{z}, R) = 0$ whenever $|\mathbf{r}_k| > R$ for all $k = 1, \cdots n_p$, that is when there are no particles inside the spherical cavity; and $\Theta(\mathbf{z}, R) = \infty$ otherwise. The function $\Theta(\mathbf{z}, R)$ ensures that particles are excluded from the spherical region around the origin. We wish to compute the free energy cost, ΔF, associated with increasing the radius of the cavity from R_A to R_B.

A hypothetical estimate of ΔF using *unescorted* nonequilibrium simulations (Eq. 4.26) involves "growing out" the spherical cavity in discrete increments, as

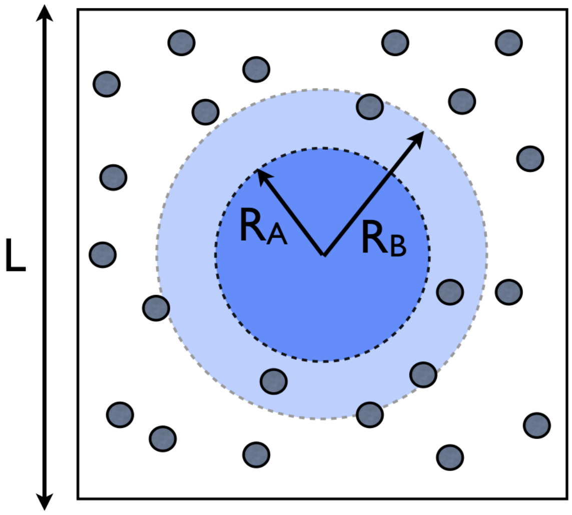

Figure 4.3: A schematic of the cavity expansion problem

follows. Starting with a microstate $\mathbf{z}_0$ sampled from equilibrium at $R = R_A$, the radius of the sphere is increased by an amount δR_0. If all n_p fluid particles remain outside the enlarged sphere, then $\delta W_0 = 0$; but if one or more particles now finds itself inside the sphere ($r_k < R_A + \delta R_0$) then $\delta W_0 = \infty$. One or more Monte Carlo steps are then taken, after which the radius is again increased by some amount, δR_1, and δW_1 is determined in the same fashion as δW_0. In principle this continues until the radius of the sphere is R_B, and then the work is tallied for the entire trajectory: $W = \sum_i \delta W_i$. In practice the trajectory can be terminated as soon as $\delta W_i = \infty$ at some step i, since this implies $W = \infty$. For this procedure, Eq. 2.5 can be rewritten as

$$P(W = 0) = e^{-\beta \Delta F}, \tag{4.85}$$

where $P(W = 0)$ is the probability of generating a trajectory for which $W = 0$; that is, a trajectory in which the sphere is successfully grown out to radius R_B, without overtaking any fluid particles along the way. The quantity $P(W = 0)$ is estimated directly, by generating a number of trajectories and counting the "successes" ($W = 0$). For a sufficiently dense fluid, however, a successful trajectory is a rare event, $P(W = 0) \ll 1$, and this approach converges poorly. Note also that this approach does not give the correct free energy difference in the reverse case of a shrinking sphere (from $R = R_B$ to $R = R_A$), since $W = 0$ for every trajectory in that situation.

For the hypothetical procedure just described, Eq. 4.85 implies that the probability to generate a successful trajectory does not depend on the number of increments used to grow the cavity from R_A to R_B. Therefore the most computationally efficient implementation is to grow the sphere out in a single step, which corresponds to the free energy perturbation method (FEP) [13, 24]. In this case $P(W = 0)$ is just the probability to observe no particles in the region $R_A < r < R_B$, for an equilibrium simulation at cavity radius R_A. Since we are interested in the probability that the region $R_A < r < R_B$ is vacant, we will use the more suggestive notation $P(n = 0)$ instead of $P(W = 0)$.

To improve convergence by means of escorted simulations (Eq. 4.31), we constructed mapping functions M_i that move the fluid particles out of the way of the growing sphere, to prevent infinite values of δW_i. Specifically, as the cavity radius R is increased from R_i to R_{i+1}, the location of the n^{th} particle, $\mathbf{r}_n$, is mapped to $\mathbf{r}'_n = m_i(\mathbf{r}_n)$, where [50]

$$m_i(\mathbf{r}_n) = \left[1 + \frac{(R_{i+1}^3 - R_i^3)(L^3 - 8r_n^3)}{(L^3 - 8R_i^3)r_n^3}\right]^{1/3} \mathbf{r}_n \qquad \text{if } r_n \le L/2 \tag{4.86}$$

and $m_i(\mathbf{r}_n) = \mathbf{r}_n$ if $r_n > L/2$. The notation $m_i : \mathbf{r}_n \to \mathbf{r}'_n$ denotes a single-particle mapping; the full mapping $M_i : \mathbf{z} \to \mathbf{z}'$ is obtained by applying m_i to all n_p fluid particles. To picture the effect of this mapping, let $\mathcal{S}_i$ denote the region of space defined by the conditions $R_i \leq r \leq L/2$, that is a spherical shell of inner radius R_i and outer radius $L/2$ (just touching the sides of the cubic container). Under the mapping $m_i : \mathbf{r} \to \mathbf{r}'$, the shell $\mathcal{S}_i$ is compressed uniformly onto the shell $\mathcal{S}_{i+1}$, leaving the eight corners of the box $r > L/2$ untouched. [11] In this manner, the particles that would otherwise have found themselves inside the enlarged sphere are pushed outside of it, resulting in a finite contribution to the work (Eq. 4.34),

$$\delta W_i = \sum_{k=1}^{n_p-1} \sum_{l>k}^{n_p} [V(\mathbf{r}'_k, \mathbf{r}'_l) - V(\mathbf{r}_k, \mathbf{r}_l)] - n_0 \beta^{-1} \ln \gamma \tag{4.87}$$

where $n_0 = n_0(\mathbf{z})$ is the number of particles found within the shell $R_i \leq r \leq L/2$ (before the mapping is applied), and $\gamma = (L^3 - 8R_{i+1}^3)/(L^3 - 8R_i^3) < 1$ is the ratio of shell volumes, $|\mathcal{S}_{i+1}|/|\mathcal{S}_i|$. The first term on the right side of Eq. 4.87 gives the net change in the energy of the system associated with the escorted switch $[\mathbf{z}_i, R_i] \overset{M_i}{\Rightarrow} [\mathbf{z}'_i, R_{i+1}]$, while the second is the Jacobian term $-\beta^{-1} \ln J_i(\mathbf{z}_i)$.

Unlike the unescorted approach or free energy perturbation, the escorted approach with the mapping given by Eq. 4.86 is applicable in both the forward (growing spherical cavity) and reverse (shrinking cavity) directions. In the reverse direction, as the solute radius is decreased from R_{i+1} to R_i, the shell $\mathcal{S}_{i+1}$ is uniformly

[11] An even better mapping would uniformly compress the entire region $r > R_i$, including the eight corners, onto the region $r > R_{i+1}$. However, due to the geometric mismatch between the spherical inner surface and cubic outer surface of these regions, such a mapping is not represented by a simple formula such as Eq. 4.86, and would need to be constructed numerically.

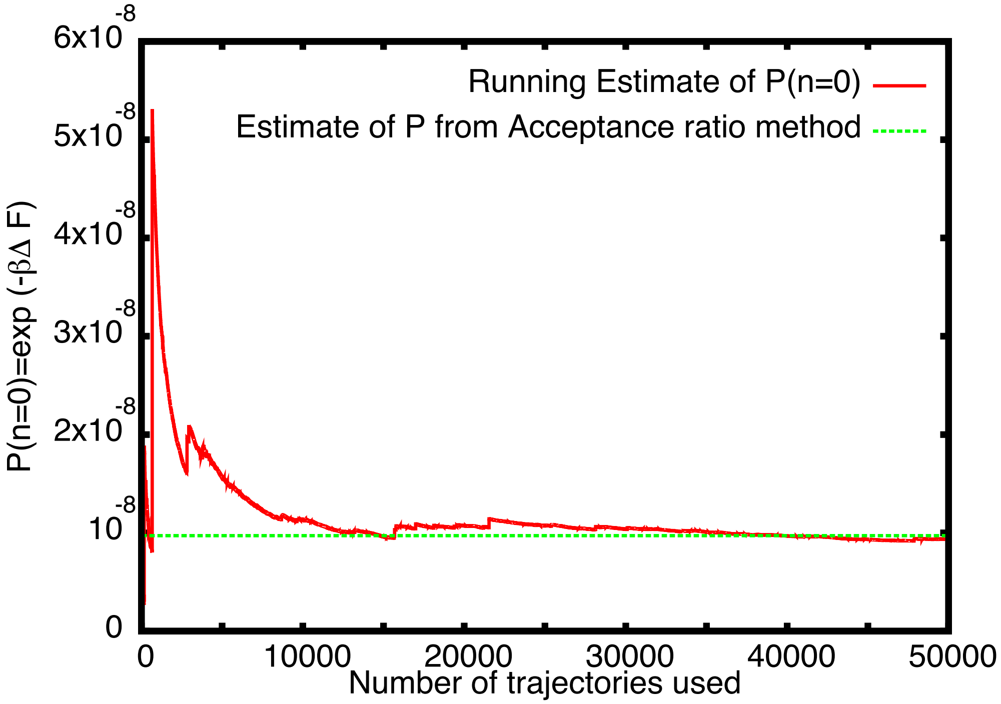

Figure 4.4: Running estimate of the probability that the region $R_A \leq r \leq R_B$ is devoid of fluid particles, $P(n = 0) = \exp(-\beta\Delta F)$, from escorted free energy simulations in which R is switched from R_A to R_B, plotted as a function of the number of trajectories used to obtain the estimate. The (green) horizontal line is the estimate of $\exp(-\beta\Delta F)$ obtained using Bennett's Acceptance ratio (BAR) method with $n_F = n_R = 50000$ trajectories. Observe that the running estimate converges to the BAR estimate in 50000 trajectories.

$\langle W \rangle_F$	22.288± 0.012
$\langle W \rangle_R$	-14.458± 0.013
$\langle W \rangle_F + \langle W \rangle_R$	7.830± 0.018
ΔF_F^{est}	18.487 ± 0.085
ΔF_R^{est}	-18.334 ± 0.078
ΔF_{BAR}^{est}	18.456 ± 0.011
C	0.120 ± 0.001

Table 4.1: Estimates and figures of merit. Here ΔF_F^{est} denotes the estimate of $\Delta F \equiv F_B - F_A$ from the forward process ($R_A \rightarrow R_B$) and ΔF_R^{est} denotes the estimate of $-\Delta F$ from the reverse process ($R_A \leftarrow R_B$). ΔF_{BAR}^{est} denotes the estimate of ΔF obtained from Bennett's Acceptance Ratio method.

expanded onto the shell $\mathcal{S}_i$. The corresponding increment in work is given by a formula similar to Eq. 4.87. As a result, one can combine work values from forward and reverse escorted simulations using Bennett's Acceptance Ratio (BAR), Eq. 4.64.

We have performed both forward and reverse simulations of this system using $N_p = 1000$ WCA particles, with $L = 10.42\sigma$, $R_A = 2.0\sigma$, $R_B = 2.05\sigma$, and at a reduced temperature $T^* \equiv k_B T/\epsilon = 1$, where the WCA parameters σ and ϵ set the units of length and energy, respectively. Minimum image convention and periodic boundary conditions were used [24].

Fig. 4.4 shows a running estimate of $\exp(-\beta\Delta F)$ obtained from escorted simulations in which the solute radius was switched from R_A to R_B in $N = 10$ steps, with each increment in R alternating with one Monte Carlo sweep [12]. The horizontal line denotes the final estimate of $\exp(-\beta\Delta F)$ obtained using Bennett's Acceptance Ratio (BAR) method with $n_F = n_R = 50000$ escorted trajectories. Fig. 4.4 clearly shows that the running estimate of $\exp(-\beta\Delta F)$ converges to the final BAR estimate. Using a total of $N_s = 50000$ independent escorted trajectories, estimates of ΔF and the figures of merit were obtained, and are summarized in Table 4.1 (The value of C and ΔF_{BAR}^{est} were estimated using $n_F = n_R = N_s = 50000$ trajectories). Statistical error bars were computed using the bootstrap method [21]. While an analytical expression for ΔF is not available for this example, the agreement between the estimates obtained by growing the solute (F), shrinking it (R), and applying BAR gives us confidence in the result, $\Delta F \approx 18.4\ k_BT$.

As an additional consistency check, in Fig. 4.5 we verify that the escorted simulations satisfy the fluctuation theorem Eq. 4.63 using Eq. 4.66. The flatness of the difference $L_2 - L_1$ over the region for which we have good statistics is in agreement with Eq. 4.66, and provides a useful and stringent consistency check [13, 24], which gives us further confidence in our estimates.

While the highly accurate estimates listed in Table 4.1 were generated using $N_s = 50000$ escorted trajectories, we found that we were able to obtain estimates of ΔF with error bars around 1 k_BT using only $N_s = 100$ realizations for the

[12] Because the quantity $\exp(-\beta\Delta F)$ has a particularly simple interpretation in this context - it is the probability $P(n = 0)$ to find no particles in the region between R_A and R_B - it is convenient to plot the running estimate of $\exp(-\beta\Delta F)$ rather than ΔF itself.

unidirectional estimators, and $N_s = 10$ realizations for the bidirectional estimator (data not shown).

To compare the escorted method with unescorted free energy perturbation (FEP), we first sampled $N_s = 100000$ independent configurations from the canonical ensemble with cavity radius $R = R_A$, by generating a single, long equilibrium Monte Carlo trajectory and sampling one configuration per 10 Monte Carlo sweeps. This involved a total computational time approximately equal to that of generating 50000 escorted trajectories. Among these 10^5 configurations we did not observe a single one in which the region $R_A \leq r \leq R_B$ was spontaneously devoid of particles ($W = 0$), in other words we were unable to obtain an estimate of ΔF using free energy perturbation. This is consistent with the result $P(n = 0) \approx e^{-18.4} \approx 10^{-8}$ (Fig. 4.4, Table 4.1), which suggests that roughly 10^8 independent configurations are needed to observe one for which $W = 0$.

For a more efficient implementation of FEP, we divided the interval $[R_A, R_B]$ into ten stages (sub-intervals), and then used FEP to estimate the free energy change for each stage, keeping the total computational time fixed. This provided a final estimate of ΔF with error bars comparable to those of the unidirectional escorted estimators in Table 4.1, but still considerably larger than those of the bidirectional estimates (data not shown). [13]

[13] Of course, even after dividing the problem into stages, one can apply escorting by separately treating each stage as a switching simulation with one step, $N = 1$, and using the mappings given by Eq. 4.86. We found that this further reduces the error bars by nearly a factor of six.

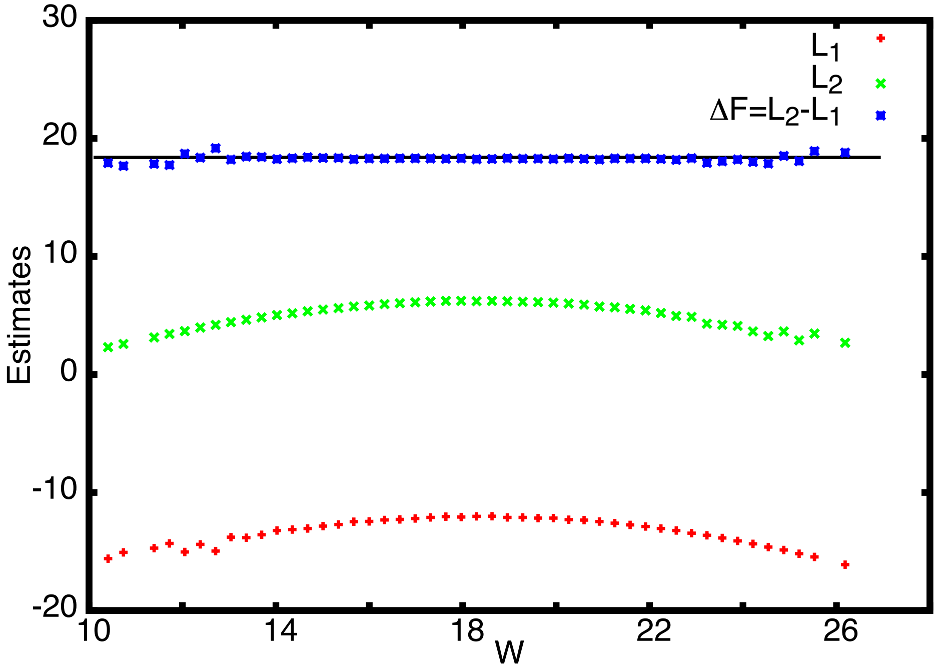

Figure 4.5: Graphical verification of the fluctuation theorem and estimation of ΔF. The horizontal line indicates the estimate of ΔF obtained from the acceptance ratio method (Table 4.1).

4.6.3 Dipole Fluid

As our third example, we consider n_p point Lennard-Jones dipoles in a cubic container of size L with periodic boundaries, and we compute the free energy cost associated with introducing a uniform electric field in the container. The energy of the system in an external electric field $\mathbf{E} = E\hat{\mathbf{e}}_z$, where $\hat{\mathbf{e}}_z$ denotes a unit vector along the z-axis, is given by

$$H_{E,\gamma}(\mathbf{z}) = -\sum_{k=1}^{n_p} \mathbf{p}_k \cdot \mathbf{E} + \sum_{k=1}^{n_p-1} \sum_{l>k}^{n_p} V_{LJ}(\mathbf{r}_k, \mathbf{r}_l) - \gamma \frac{\mathbf{p}_k \cdot \mathbf{p}_l}{|\mathbf{r}_k - \mathbf{r}_l|^4} \tag{4.88}$$

where $\mathbf{z} = \{\mathbf{r}_1, \mathbf{p}_1, \ldots \mathbf{r}_{n_p}, \mathbf{p}_{n_p}\}$, $\mathbf{p}_k$ denotes the dipole moment vector of the k^{th} particle, and $V_{LJ}(\mathbf{r}_k, \mathbf{r}_l)$ denotes the Lennard-Jones pairwise interaction potential. The parameter γ controls the strength of the dipole-dipole interaction. We set $|\mathbf{p}_k| = 1$ for all k. In spherical polar coordinates, $\mathbf{p}_k = (1, \theta_k, \phi_k)$, and the measure on $\mathbf{z}$ space is hence $d\mathbf{z} = \Pi_{k=1}^{n_p} d\mathbf{r}_k d\cos(\theta_k) d\phi_k$.

Taking the electric field to be the external parameter, we wish to compute the free energy difference between the ensembles corresponding to $E = 0$ and $E = E_f$ at some temperature β^{-1} by performing nonequilibrium switching simulations. Our first task is to construct a mapping function that escorts the system along a near equilibrium path as E is switched. Following Eq. 4.75, we consider the energy function $\bar{H}_E(\mathbf{z}) \equiv H_{E,0}(\mathbf{z})$ (i.e. $\gamma = 0$ in Eq. 4.88), which describes a reference system of non-interacting Lennard-Jones dipoles in a field of strength E. The change in free energy as the field is switched from E_i to E_{i+1} can be solved analytically and is given by

$$\bar{F}_{E_{i+1}} - \bar{F}_{E_i} = -n_p \frac{1}{\beta} \ln \left[\frac{\sinh(\beta E_{i+1})}{\sinh(\beta E_i)} \frac{E_i}{E_{i+1}} \right] \tag{4.89}$$

We now use this result to solve for a perfect set of mappings for this system of non-interacting dipoles.

Let $m_i : \zeta \equiv \cos(\theta) \to \zeta'$ denote a mapping that acts on the $\zeta = \cos(\theta)$ degree of freedom of a dipole when the external field is switched from E_i to E_{i+1}. The full mapping M_i is obtained by applying the mapping m_i to all n_p particles. We look for the perfect mapping M_i that transforms the canonical distribution corresponding to $\bar{H}_{E_i}(\mathbf{z})$ to the canonical distribution corresponding to $\bar{H}_{E_{i+1}}(\mathbf{z}')$. The following equation for the perfect single particle mapping m_i can be obtained from Eq. 4.72 by using Eqs. 4.88 and 4.89 and by noting that $\mathbf{p}_k \cdot \mathbf{E} = E\zeta_k$:

$$E_{i+1} m_i(\zeta) - E_i \zeta - \frac{1}{\beta} \ln \frac{dm_i(\zeta)}{d\zeta} = -\frac{1}{\beta} \ln \frac{\sinh(\beta E_{i+1})}{\sinh(\beta E_i)} \frac{E_i}{E_{i+1}} \tag{4.90}$$

This differential equation has the solution

$$m_i(\zeta) = \frac{1}{\beta E_{i+1}} \ln \left[\frac{\sinh(\beta E_{i+1})}{\sinh(\beta E_i)} (e^{\beta E_i \zeta} - e^{\beta E_i}) + e^{\beta E_{i+1}} \right] \tag{4.91}$$

While Eq. 4.91 is a perfect mapping only when there are no dipole-dipole interactions ($\gamma = 0$) we expect this mapping to work reasonably well for small values of γ. We will use the term *simple mapping* in reference to Eq. 4.91.

We also constructed a set of mapping functions using mean field [10] arguments as follows. In the absence of long range order, mean field theory suggests that the interacting dipole-fluid system ($\gamma \neq 0$) in an electric field of strength E can be approximated by a system of non-interacting dipoles ($\gamma = 0$) in an effective field of strength E'. We obtained approximate values for this effective electric field by first simulating a fluid of interacting dipoles ($\gamma \neq 0$), and numerically evaluating the single-dipole distribution $P(\zeta)\,, \zeta = \cos(\theta)$, at $E = E_f$. The thermal distribution of ζ for a non-interacting dipole in a field of strength E'_f, is $P_0(\zeta) \propto \exp(\beta E'_f \zeta)$. Hence E'_f can be estimated by fitting P_0 to the numerically obtained distribution $P(\zeta)$. For all other values of E, we calculate the effective fields by linear scaling, $E' = EE'_f/E_f$. Again, using Eq. 4.72 with $\bar{H}_E(\mathbf{z}) = H_{E',0}(\mathbf{z})$ we obtain a new set of mapping functions. In particular, when the E field is switched from E_i to E_{i+1}, the $\zeta_k = \cos(\theta_k)$ degree of freedom of the k^{th} dipole is transformed according to Eq. 4.92

$$m_i(\zeta_k) = \frac{1}{\beta E'_{i+1}} \ln \left[\frac{\sinh(\beta E'_{i+1})}{\sinh(\beta E'_i)} (e^{\beta E'_i \zeta_k} - e^{\beta E'_i}) + e^{\beta E'_{i+1}} \right] \tag{4.92}$$

We will refer to Eq. 4.92 as a *mean field mapping*. Since the single-dipole distributions for the interacting system at field strength E are (by construction) closely approximated by the single-particle distributions for the non-interacting system at E', we expect the mean field mappings to perform better than the simple mappings of Eq. 4.91.

We performed numerical simulations of the fully interacting dipole fluid with $n_p = 800$ particles. The parameters σ, ϵ of the Lennard-Jones potential set the length and the energy scale of the system, and we took $L = 10\sigma$ and $T^* = k_B T/\epsilon = 1$. Minimum image convention and periodic boundary conditions [24] were used. We performed $N_s = 10^4$ forward and reverse simulations to estimate the free energy difference between the ensembles corresponding to $E = 0$ and $E = 1$, switching the field strength in $N = 10$ equal increments. Ten Monte Carlo sweeps were performed between these updates in E. We obtained estimates of ΔF using: (1) unescorted switching simulations (Eq. 2.5), (2) escorted simulations with the simple mappings (Eq. 4.91), and (3) escorted simulations with the mean field mappings (Eq. 4.92). For the latter, the effective fields were obtained as described in the previous paragraph. In particular, we found $E'_f \approx 1.5E_f$ and therefore we took $E'_i = 1.5E_i$ in Eq. 4.92.

Fig. 4.6 shows the work distributions $P_F(W)$ and $P_R(-W)$ for these sets of simulations, and reveals a progression from virtually no overlap for the unescorted simulations, to some overlap for the simulations with the simple mappings, to nearly perfect overlap when using the mean field mappings. This trend is in agreement with the expectations mentioned above, and provides direct evidence that the mappings we have constructed substantially reduce the lag and dissipation. The first three rows of Table 4.2 quantify these observations. In particular, row 3 gives the distance between the means of $P_F(W)$ and $P_R(-W)$, and shows that this measure of hysteresis proceeds from nearly $250k_BT$ to about $24k_BT$ to less than $1k_BT$ in the three cases. Rows 4 to 6 illustrate the effect of this trend on the efficiency and accuracy of the free energy estimates. The estimates of ΔF (that is, ΔF_F^{est}, $-\Delta F_R^{est}$, and ΔF_{BAR}^{est}) obtained from the unescorted simulations differ substantially from one another, indicating a high degree of bias. The estimates corresponding to the simple mappings are markedly better, though they still suggest a degree of bias on the order of $1k_BT$. Finally, the simulations with the mean field mappings are in agreement to within about $0.05k_BT$, indicating excellent accuracy and efficiency. These findings

	No mapping	Mapping	Mean field mapping
$\langle W \rangle_F$	-60.409 ± 0.126	-177.074 ± 0.039	-189.079 ± 0.010
$\langle W \rangle_R$	302.958 ± 0.132	200.607 ± 0.045	189.971 ± 0.010
$\langle W \rangle_F + \langle W \rangle_R$	242.549 ± 0.182	23.533 ± 0.060	0.892 ± 0.014
ΔF_F^{est}	-114.189 ± 3.913	-187.612 ± 0.405	-189.552 ± 0.011
ΔF_R^{est}	262.232 ± 0.711	191.877 ± 0.310	189.502 ± 0.0140
ΔF_{BAR}^{est}	-128.215 ± 3.324	-189.599 ± 0.110	-189.530 ± 0.008
C	~ 0	0.011 ± 0.001	0.407 ± 0.001

Table 4.2: Estimates and Figures of Merit for $\gamma = 0.1$. Note that the simulations with the mapping are much more efficient than those without. The forward and reverse work histograms obtained from the simulations without any mappings were so far apart that a reliable estimate of C could not be obtained.

are also in agreement with the values of the overlap integral C, shown in row 7. This was too low to be estimated using the unescorted simulations, and approaches its maximal value of 1/2 when using the mean field mappings. Using escorted simulations with the mean field mappings, with the acceptance ratio method (BAR), we found that we were able to generate estimates of ΔF with error bars on the order of $0.2k_BT$, with about $N_s \sim 1/C^2 \sim 10$ (data not shown).

4.7 Summary

Nonequilibrium fast switching estimates of free energy differences often perform poorly due to dissipation (see Fig 1.1). The strategy developed here seeks to address this issue. By modifying the dynamics with additional terms that serve to escort the system along a near equilibrium trajectory and consequently reduce dissipation, we obtain efficient fast switching estimators (Eq. 4.5, Eq. 4.39, Eq. 4.64) for the free energy difference. The success of the strategy depends crucially on the choice of the escorting functions: the more effectively these reduce the dissipation, the more efficient the resulting estimator of ΔF.

The examples presented in Section 4.6 illustrate this point. In the example of a particle in the one dimensional Sun potential [95], the key to success with our method is a flow field $\mathbf{u}$ that reduces lag, and therefore dissipation, by mimicking the effect of a variation of λ on the distribution p^{eq}. For the hard sphere solute, we used a simple mapping function that uniformly compresses the solvent, vacating the region into which the hard sphere expands (Eq. 4.86). With this escorting function we were able to estimate ΔF directly from single-stage switching simulations, which would not have been feasible without escorting. In the example of the Lennard-Jones dipole fluid, we used a reference system of non-interacting dipoles to construct a reasonable set of mapping functions (Eq.4.91), and then we further refined these mappings using mean field arguments (Eq. 4.92). Figure 4.6 and Table 4.2 illustrate the correlation between reduced dissipation and increased computational efficiency. Because mean field theory often provides a reasonable description of many-body systems, we speculate that this approach will prove effective for more complex problems of physical interest.

We have also discussed figures of merit, specifically the dissipation in the forward and reverse processes, and the overlap integral C (Eqs. 4.77, 4.78, 4.80). For the two examples in Section 4.6, we found that these quantities indeed track the effectiveness of the mapping functions. This suggests that these figures of merit might be useful to iteratively improve the performance of the mapping functions.

Our method might also be combined with *steered molecular dynamics* [47, 80], in which a constraining potential is used to drag a coordinate ξ along a desired path $\tilde{\xi}_t$. By adding a flow field that acts on this coordinate and others coupled to it, one might be able to reduce the lag between ξ and $\tilde{\xi}_t$. For free energy calculations along a reaction path for which we do not have good intuition, *transition path sampling* [8] could provide information useful for designing an effective flow field.

The method we propose is distinct from path-space sampling schemes [28, 95, 105, 106], in which the convergence of Eq. (2.5) is improved by modifying the *probabilities* with which physical trajectories are generated, for instance by biasing in favor of small work values. In our approach, by contrast, we modify the equations of motion themselves, thereby sampling from an entirely different set of trajectories. (For example in the one dimensional example, we generated non-Hamiltonian trajectories, rather than a statistically re-weighted sampling of Hamiltonian trajectories.) The distinction is particularly evident in the case of a perfect flow field $\mathbf{u}^*$, when *every* trajectory gives $W = \Delta F$.

Finally, it would be interesting to combine our approach with Hummer and

Szabo's approach for computing the potentials of mean force [42], and the *large time step* [75] and *optimal protocol* [89] strategies, recently proposed for improving the efficiency of free energy estimates.

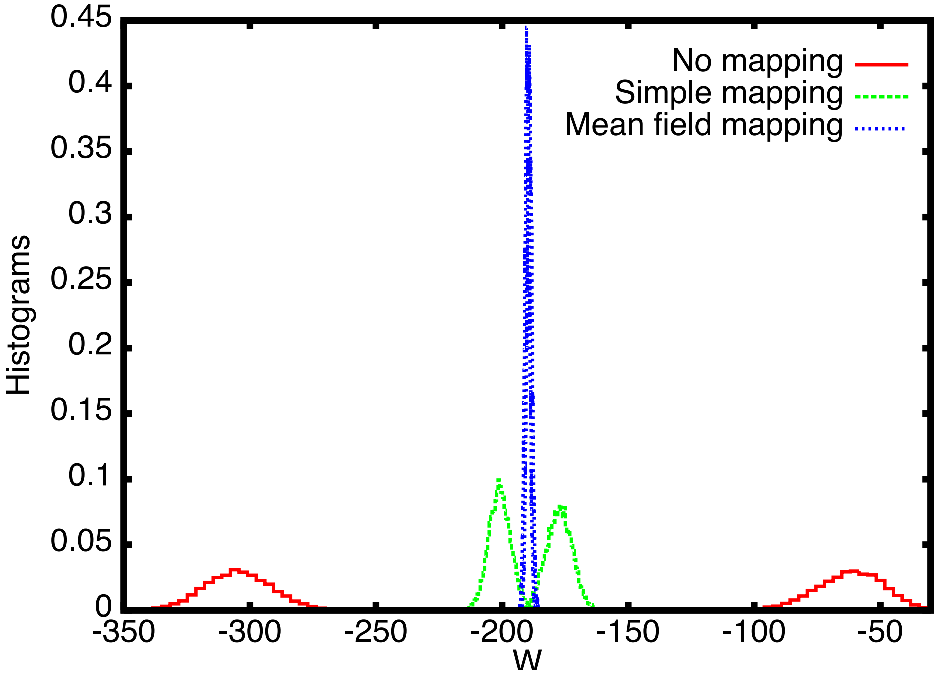

Figure 4.6: Work histograms obtained from forward and reverse simulations performed at $\gamma = 0.1$. The degree of overlap between $P_F(W)$ (right) and $P_R(-W)$ (left) provides an indication of the efficiency of the free energy estimate. For unescorted simulations (no mapping) we see no overlap, reflecting considerable dissipation and poor efficiency (Table 4.2). With the mapping given by Eq. 4.91 the overlap is much improved, and with the mean field mapping, Eq. 4.92 the forward and reverse distributions are nearly identical.

Chapter 5

Estimating solvation free energies using escorted free energy simulations

5.1 Introduction

Solvation free energies, i.e. the free energy differences associated with introducing a solute molecule into a solvent, are important quantities in computational thermodynamics, especially in the context of computer studies of phase equilibria [24, 103], and the hydrophobic effect [11, 30, 73, 81, 101]. To set up this free energy estimation problem, we again consider a system composed of N_p solvent molecules. Let us suppose that we are interested in computing the solvation free energy of a solute particle which interacts with a solvent molecule centered at $\mathbf{r}_k$

according to the spherically symmetric potential $V_{\mathbf{r}_s}(|\mathbf{r}_k - \mathbf{r}_s|)$, where $\mathbf{r}_s$ describes the position of center of the solute particle. The solvation free energy, ΔF_{sol}, is the free energy difference between the equilibrium state with N_p solvent molecules, and the equilibrium state with N_p solvent molecules and one solute molecule [1].

This solvation free energy can be written as a sum of two components, an *ideal* component, ΔF_{sol}^{id}, that describes the free energy difference associated with introducing the solute in an ideal gas under the same conditions (which can be evaluated analytically), and an *excess* component, ΔF_{sol}^{ex} [24]. The free energy difference ΔF_{sol}^{ex} can be evaluated from computer simulations by imagining a process in which a point $\mathbf{r}$ inside the simulation box (with the solvent fluid) is chosen randomly after which the potential $V_{\mathbf{r}}$ is gradually turned on. It is useful to think of this as a process in which the "size" of a solute particle centered at $\mathbf{r}$ is gradually increased as in the cavity expansion example discussed in Chapter 4. The work performed in this process can be used to estimate ΔF_{sol}^{ex} using the nonequilibrium work relation Eq. 2.5 [2] [24]. Such free energy calculations can be time consuming and inefficient due to the high dissipation and lag which will result if the the solvent molecules are not given sufficient time to re-equilibrate around the solute as it is grown out.

One possible strategy to alleviate this problem is to use the escorting functions, Eq. 4.86, introduced in the cavity expansion example. As we saw in Section 4.6.2,

[1] As usual, we have assumed that the bulk properties of the solvent molecules are the same in the two equilibrium states.

[2] In constant pressure simulations, the volume of the simulation box should also be included in the calculation, see for example Section 7.2.2 of Ref [24].

these escorting functions can be quite effective in simulations in which a hard sphere solute is grown in a WCA fluid. Given the success of the escorting functions in that example problem, it is interesting to investigate the effectiveness of the escorting functions in other general settings, such as growing a solute in a fluid of Lennard-Jones or water molecules. To this end, in this chapter we describe simulations of a solvent fluid in the presence of a radially symmetric potential with a hard repulsive core and short ranged dispersive interactions, and compute the free energy cost associated with increasing the size of the hard core excluded volume region. This is meant to model a calculation in which the size of a solute centered at the origin is increased.

We compute this free energy difference by *suddenly* switching the size of the hard core region, and compare the efficiency of free energy estimates obtained without the escorting functions to those obtained with the escorting functions. The nonequilibrium work relation and its escorted generalization reduce to the Free Energy Perturbation (FEP) identity and the Targeted Free Energy Perturbation (TFEP) identity [50] (see Eq. 4.41) respectively in the limit of sudden switching. It is easier to investigate the effectiveness of the escorting functions in this limit and hence we refrained from performing the usual switching simulations in which the size of the hard core region would have been grown at a finite rate. The use of the FEP identity to compute solvation free energies (and excess chemical potentials) is commonly referred to as Widom's particle insertion method [103].

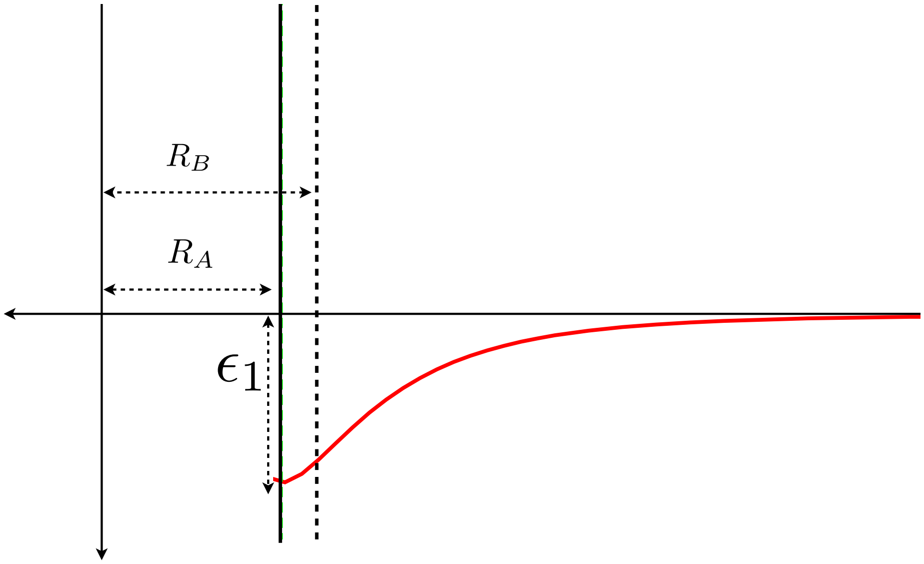

Figure 5.1: A sketch of the potential $V_O^R(r)$ at $R = R_A$. As mentioned in the text, we have set $\sigma_2 = R_A - 2^{(1/6)}\sigma_1$. The potential has an excluded volume interaction for $r < R$. A positive value of ϵ_1 sets the strength of the short-ranged solute-solvent attractions. We are interested in estimating the free energy difference associated with changing R, the radius of the excluded volume interaction, from R_A to R_B while keeping the short-ranged attractions constant.

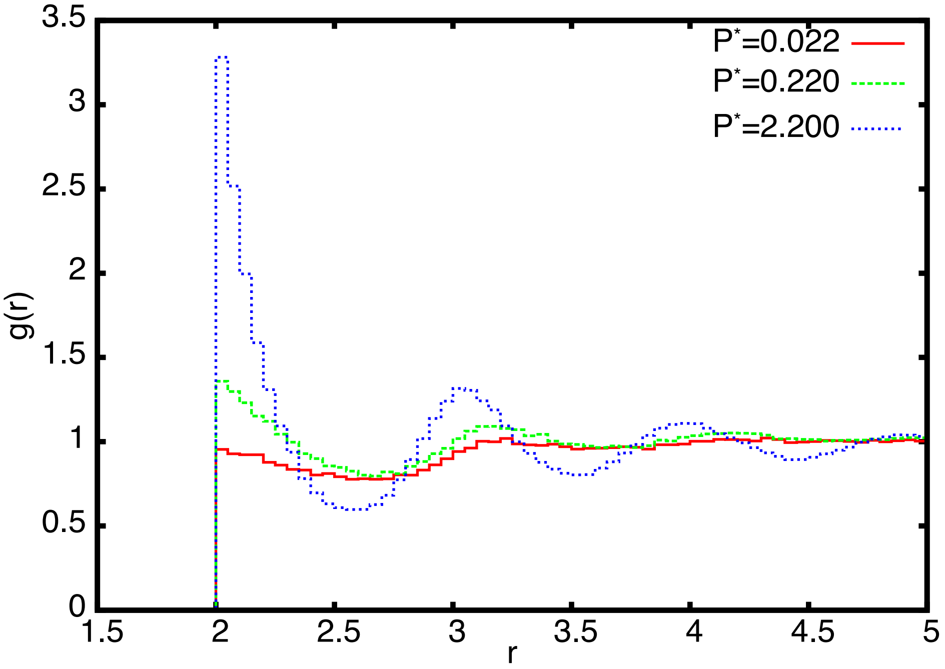

Figure 5.2: The solute-solvent radial distribution function $g(r)$ as a function of the distance r from the center of the solute at different values of P^*. Notice the occurrence of drying at $P^* = 0.022$. The density of solvent at the point of contact with solute increases with the reduced pressure P^*. The distances are in units of σ.

5.2 Model and simulations

As we mentioned in the previous section, we will simulate a system of solvent molecules (either particles interacting according to the Lennard-Jones potential in Sec 5.2.1 or SPC/E water molecules in Sec 5.2.2) in the presence of a solute particle placed in the origin. The interactions between the solute and the solvent particles are modeled using the potential in Eq. 5.1 below. In particular, a solvent molecule whose center is at a distance r from the origin interacts with the solute via the potential

$$V_O^R(r) = \begin{cases} 4\epsilon_1((\frac{\sigma_1}{r-\sigma_2})^{12} - (\frac{\sigma_1}{r-\sigma_2})^6) & \text{if } r \geq R \\ \\ \infty & \text{if } r < R \end{cases} \tag{5.1}$$

where $R > \sigma_2 > \sigma_1 > 0$ and $\epsilon_1 \geq 0$. The potential described in Eq. 5.1 has a hard sphere excluded volume interaction for $r \leq R$ and models a solute with a highly repulsive core. We will be interested in computing the free energy cost (ΔF) associated with increasing R from R_A to R_B. The value of σ_2 is set to $\sigma_2 = R_A - 2^{1/6}\sigma_1$ in both the ensembles. The potential is illustrated in Fig 5.1. A positive value of ϵ_1 sets the strength of a short-ranged solute-solvent attraction, and σ_1 determines the range of these attractive interactions. We will compute ΔF using both TFEP and FEP and compare the effectiveness of the two estimates in various settings. We will use the escorting transformation in Eq. 4.86 for the TFEP calculations.

Since we are growing the radius of the hard sphere excluded volume and are not changing the short-ranged solute-solvent attractions, the FEP calculations will only involve estimating the probability $P(n = 0)$ that the region $R_A \leq r \leq R_B$ is devoid of particles in ensemble A. The free energy difference ΔF is then estimated using the relation $\Delta F = -\beta^{-1} \ln P(n = 0)$. Also recall that the free energy perturbation method can only be applied in switching simulations where the radius of the hard sphere increases and hence we cannot use FEP in the reverse simulations (where R is switched from R_B to R_A). TFEP on the other hand has no such limitation and hence we will obtain estimates of ΔF using both forward and reverse TFEP simulations.

5.2.1 Simulations with a Lennard-Jones Fluid

We first consider the free energy cost associated with growing a hard solute ($\epsilon_1 = 0$ in Eq. 5.1 above) in a Lennard-Jones fluid (i.e. the solvent molecules interact according to the Lennard-Jones potential). The length and the energy scales respectively of the Lennard-Jones fluid is set by ϵ, σ.

We performed Monte-Carlo simulations with $N_p = 864$ Lennard-Jones particles in a cubic box with minimum image periodic boundary conditions. The solute modeled by the potential Eq. 5.1 was placed in the center of the simulation box at the origin. We used this setting in all our simulatons. We performed simulations at three different bulk (reduced) pressures, $P^* = P\sigma^3/\epsilon = 0.022, 0.22, 2.2$, and a (reduced) temperature $T^* = k_B T/\epsilon = 0.85$. The two hard sphere radii were chosen to be $R_A = 2.0\sigma$ and $R_B = 2.05\sigma$ and we set $\sigma_1 = \sigma$. The length of the simulation box in a particular realization was used in the mapping transformation Eq. 4.86 for that realization. We followed this procedure in all the NPT simulations.

At the lowest value of the pressure, $P^* = 0.022$, and at $T^* = 0.85$, the Lennard-Jones fluid is close to liquid-vapor coexistence. These bulk conditions were chosen as they can cause the solvent to recede from the surface of the solute (drying) [40] [3]. The free energy calculations described below were performed by simulating a long equilibrium trajectory (at the appropriate equilibrium state) and sampling points every 200 Monte-Carlo sweeps. We used $n_F = n_R = N_s = 1000$ such configurations in the calculation. Fig 5.2, shows the solute-solvent radial distribution function, $g(\mathbf{r}) = g(r) = \langle \sum_{k=1}^{N_p} \delta(\mathbf{r}_k - \mathbf{r}) \rangle / \rho$,where ρ denotes the bulk density of the solvent, and we have used the fact that g is spherically symmetric in our case and only depends on the distance r between the centers of the solvent and the solute, at the three different values of pressure and with $R_A = 2.0\sigma$. At $P^* = 0.022$, the solvent begins to recede from surface of the solute indicating the onset of drying. Drying is not favored at the higher values of pressure, and the density of solvent molecules at the surface of the solute, the contact density, increases with pressure. As the contact density increases, sampling a realization from the state A in which the region $R_A \leq r \leq R_B$ is devoid of particles becomes increasingly difficult. Consequently, we anticipate that it will be tougher to estimate ΔF using FEP. Indeed, the error bars on the estimate of ΔF from FEP in Table 5.1 bear out this trend. In fact, with $N_p = 1000$ equilibrium samples, we were not able to obtain an estimate of ΔF using FEP at $P^* = 2.2$.

The estimates of ΔF from TFEP in both the forward and reverse process, the

[3] This simulation was suggested by Prof. John D. Weeks

estimate using Bennett's Acceptance ratio (BAR), and the average work performed in the forward and reverse TFEP processes are also tabulated in Table 5.1 [4].Unlike in the example discussed in the previous chapter where the escorted approach is significantly better than the usual method, we note that FEP performs slightly better than TFEP at the lowest value of pressure. However, the efficiency of the TFEP calculation decreases only modestly with pressure and TFEP starts to outperform FEP as the pressure (and the contact density) is increased. The improvement in the efficiency of the TFEP calculation with respect to the FEP approach is reflected in the error bars and also in the estimates of average work performed in the TFEP calculation. For example, at $P^* = 2.2$, the dissipation in the reverse process is $\langle W_{diss} \rangle_R \sim 3.4$. Recall that the convergence of the forward TFEP estimator is controlled by the dissipation in the reverse process, and the number of realizations required to obtain a reliable estimate of ΔF grows exponentially with $\langle W_{diss} \rangle_R$, $N_s \sim \exp(\beta \langle W_{diss} \rangle_R)$. This rough consideration tells us that $N_s \sim 100$ realizations are sufficient to obtain a reliable estimate of ΔF. On the other hand, the probability that the region $R_A \leq R \leq R_B$ is vacant at this value of P^* is $P(n = 0) = \exp(-\beta \Delta F) \sim 2.5 \times 10^{-4}$. This implies that we will need at least $N_s \sim 1/P(n = 0) \sim 4000$ realizations to observe a realization in which the aforementioned region is vacant and obtain an estimate of ΔF from FEP. This simple

[4] The estimates of ΔF from forward and reverse simulations are mutually consistent (and comparable to the estimate of ΔF from FEP in the first two instances). This gives us confidence in our ΔF estimates. Moreover, our data satisfies the graphical test of the fluctuation theorem, Eq. 4.66. This gives us further confidence in our ΔF estimates.

P^*	FEP	TFEP(F)	TFEP(R)	$\langle W\rangle_F$	-$\langle W\rangle_R$	BAR
0.022	1.473 ± 0.060	1.338 ± 0.079	1.466 ± 0.099	2.477	0.048	1.422 ± 0.031
0.220	2.278 ± 0.072	2.213 ± 0.095	1.945 ± 0.099	3.576	0.742	2.109 ± 0.025
2.200	$--$	7.355 ± 0.169	6.628 ± 0.224	10.616	3.621	7.025 ± 0.071

Table 5.1: Estimates of ΔF obtained using FEP, and the forward (F) and reverse (R) TFEP calculations along with error bars at different values of P^* and at $T^* = 0.85$ for a hard sphere solute. The BAR column has estimates of ΔF obtained using Bennett's acceptance ratio method. Observe that the TFEP becomes more efficient than FEP as P^* and the contact density increase. At the highest value of P^*, we did not observe a single realization where the region between R_A and R_B is vacant. All estimates of ΔF and $\langle W\rangle$ are in units of ϵ.

analysis clearly shows that the TFEP estimator becomes more efficient than the FEP estimator as the pressure and consequently the contact density is increased.

The trends observed in this example, i.e. the improvement in the relative efficiency of the TFEP estimator with contact density, will be observed in the other free energy calculations described below where we will increase the contact density by increasing ϵ_1. In Section 5.3, we will analyze these trends.

For our next set of simulations, we considered a system of $N_p = 864$ Lennard-Jones particles at $P^* = 0.022$ and $T^* = 0.85$ with $\epsilon_1^* = \epsilon_1/\epsilon = 0, 1, 2, 4$ respectively (as the value of ϵ_1 is increased, the attraction between the solute and solvent in-

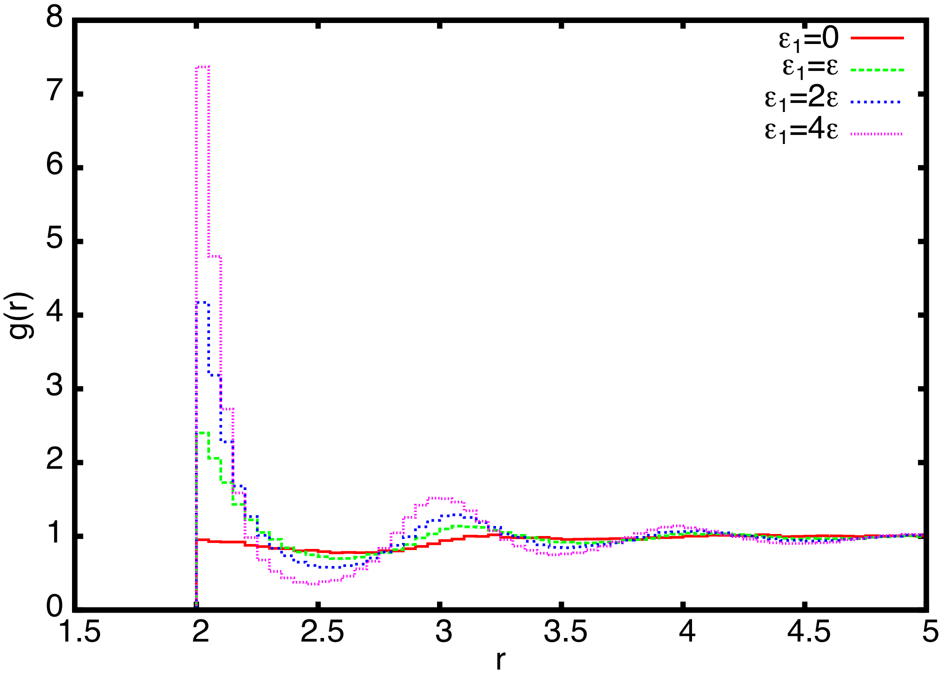

Figure 5.3: The solute-solvent radial distribution function $g(r)$ as a function of the distance r from the center of the solute at different values of ϵ_1. Notice the onset of drying at $\epsilon_1 = 0$. The density of solvent at the point of contact with solute increases with ϵ_1. The distances are in units of σ.

creases and the contact density increases). R was set to $R_A = 2.0\sigma$ in ensemble A and to $R_B = 2.05\sigma$ in ensemble B. The free energy calculations described below were again performed by simulating a long equilibrium trajectory (at the appropriate equilibrium state) and sampling points every 200 Monte-Carlo sweeps. We used $n_F = n_R = N_s = 1000$ such configurations in this calculation.

The solute solvent radial distribution functions $g(r)$ at various values of ϵ_1 are plotted in Fig 5.3. As the value of ϵ_1 is increased the contact density increases. We computed the free energy difference using both FEP and TFEP (in the forward and reverse directions). The results are tabulated in Table 5.2. We again observe that the FEP method performs well at the lower values of ϵ_1 but quickly becomes inefficient as ϵ_1 is increased. On the other hand, just as in the previous example, the efficiency of the TFEP calculation decreases only by a modest amount. In particular, for the highest value of ϵ_1, we would have required $N_s = 1/P(n = 0) \sim 30,000$ realizations to obtain an estimate using FEP while the TFEP method provides a reasonably accurate estimate with just $N_s = 1000$ points.

In the next section, Sec 5.2.2, we report results from similar simulations performed with a model of water as the solvent fluid. We will find that the general characteristics observed in the simulations with the Lennard-Jones fluid hold for water also.

ϵ_1/ϵ	FEP	TFEP(F)	TFEP(R)	$\langle W \rangle_F$	-$\langle W \rangle_R$	BAR
0	1.473 ± 0.060	1.338 ± 0.079	1.466 ± 0.099	2.477	0.048	1.422 ± 0.031
1	3.740 ± 0.169	3.859 ± 0.115	3.769 ± 0.078	5.437	2.472	3.882 ± 0.034
2	$--$	7.324 ± 0.164	7.108 ± 0.096	9.426	5.252	7.276 ± 0.044
4	$--$	14.609 ± 0.174	15.099 ± 0.210	17.511	11.932	14.726 ± 0.062

Table 5.2: Estimates of ΔF obtained using FEP, and the forward (F) and reverse (R) TFEP calculations along with error bars at different values of ϵ_1 and at $P^* = 0.022, T^* = 0.85$. The BAR column has estimates of ΔF obtained using Bennett's acceptance ratio method. Observe that the TFEP becomes more efficient than FEP as ϵ_1 increases. At the highest values of ϵ_1, we did not observe a single realization where the region between R_A and R_B is vacant. All estimates of ΔF and $\langle W \rangle$ are in units ϵ.

5.2.2 Water Simulations

In this section, we describe results from simulations in which the solute described by Eq. 5.1 is solvated in a fluid of water molecules. We used the popular SPC/E [7] model to simulate the water molecules. The SPC/E model is a tetrahedral water model with an oxygen-hydrogen bond distance of $1\AA$ and with point charges of $+0.4238\,\mathrm{e}$ and $-0.8476\,\mathrm{e}$ (e denotes electronic charge units) on the hydrogen and oxygen respectively. The oxygen atoms of water molecules interact according to a Lennard-Jones potential with $\epsilon = 0.650\,\mathrm{KJ/mol}$ and $\sigma = 3.166\AA$.

In the simulations with the Lennard-Jones fluid, the solvent-solvent interactions are short-ranged, and can be safely truncated [5] beyond a cutoff distance R_c. Consequently, minimum-image [24] periodic boundary conditions can be used to simulate the fluid. In simulations with water however, long-ranged Coulomb $(1/r)$ interactions have to be considered. A naive truncation scheme in which the Coulomb interactions are truncated beyond a cutoff distance R_c is not reasonable [24]. Hence, either more careful truncation schemes need to be used (see for example Chapter 4 of Ref [86]), or the long-ranged interactions need to be accounted, for example by summing over all the periodic images using Ewald sums [24] or using mean field methods such as Local Molecular Field theory [12, 87, 101, 102]. The simulations described in this section were performed using the Gaussian truncation scheme [86, 87]. We briefly discuss this scheme and its limitations below before proceeding to describe the free energy simulations.

[5] The effects of the neglected components of the potential (after truncation) can be accounted for perturbatively.

The Gaussian truncation scheme involves splitting the $1/r$ potential as follows [87]

$$\frac{1}{r} = v_0(r) + v_1(r) = \frac{\mathrm{erfc}(\frac{\mathrm{r}}{\sigma_\mathrm{c}})}{r} + \frac{\mathrm{erf}(\frac{\mathrm{r}}{\sigma_\mathrm{c}})}{r}, \tag{5.2}$$

where erf(r), and erfc(r) denote the error function and the complementary error function respectively. In this decompostion, $v_0(r)$ captures the rapidly varying short-ranged component of the Coulomb potential and $v_1(r)$ captures the slowly varying long range component. The range of the potential $v_0(r)$ is set by σ_c. In the Gaussian truncation scheme, simulations are performed only with $v_0(r)$. The long-ranged effects due to $v_1(r)$ are ignored. This truncation works rather well in homogenous liquids provided σ_c is large enough to capture all the rapidly varying short-ranged forces. In particular, the structural properties (e.g. pair correlation functions) obtained from simulations with this truncation compare reasonably [87] [6]to those obtained from simulations in which the long-ranged components $v_1(r)$ are explicitly considered using Ewald sums [24].

While the fluid in our simulations is inhomogenous due to the presence of the solute, the solutes we consider are relatively small in size: $R_A = 6\mathring{A}$ is the largest solute simulated. Under these conditions, we expect the trends observed in the free energy calculations with Gaussian truncated water, to hold even in simulations that explicitly include the long-ranged components. To justify this assertion, we used Local Molecular Field (LMF) theory which is a method developed by Weeks and co-wokers [12,87,101,102] as an alternative to methods such as Ewald summation to

[6] However, the thermodynamic properties of spherically truncated water (P, $\langle H \rangle$, and the free energy F) differ from those of water with long-ranged interactions.

account for long-ranged interactions. The central idea of the LMF theory involves splitting the inter-particle potential into short-range and long-range components. Simulations are then performed using only the short-ranged potential. The effects of the long-ranged components are taken into account through the imposition of a mean-field external potential, $\psi(r)$. The LMF theory provides a method to compute self-consistently the potential $\psi(r)$. As the effects of the long-ranged forces are only taken into account in a mean-field fashion, simulations using the LMF method tend to be much faster than those in which the long-ranged components are considered explicitly.

For Coulomb interactions, the decomposition in Eq. 5.2 can again be used to separate the potential into short and long-ranged parts [12, 86, 87]. We computed the self-consistent field [7], $\psi(r)$, using the recently introduced perturbation method of Hu and Weeks [39]. While the introduction of these fields does indeed modify the properties of the system, they do not significantly alter the trends in the free energy calculations for the solute sizes considered in this thesis. Hence, in subsequent discussions, we simply report results from simulations with Gaussian truncated water.

We performed Monte-Carlo simulations with $N_p = 1000$ SPC/E water molecules at $T = 300K$ and $P = 1$atm and with $\sigma_c = 4.25\mathring{A}$, $\sigma_2 = 6\mathring{A}$, $\sigma_1 = 1.5\mathring{A}$. The ensemble A was simulated with $R_A = 6.0\mathring{A}$ and ensemble B with $R_B = 6.05\mathring{A}$. The simulations were performed at four different values of ϵ_1, $\epsilon_1/(k_BT) = 0, 1, 2, 4$. We computed ΔF using both FEP and TFEP simulations and compared their efficiency

[7] We gratefully acknowledge help from Rick Remsing and Prof. John D. Weeks in performing the LMF calculations.

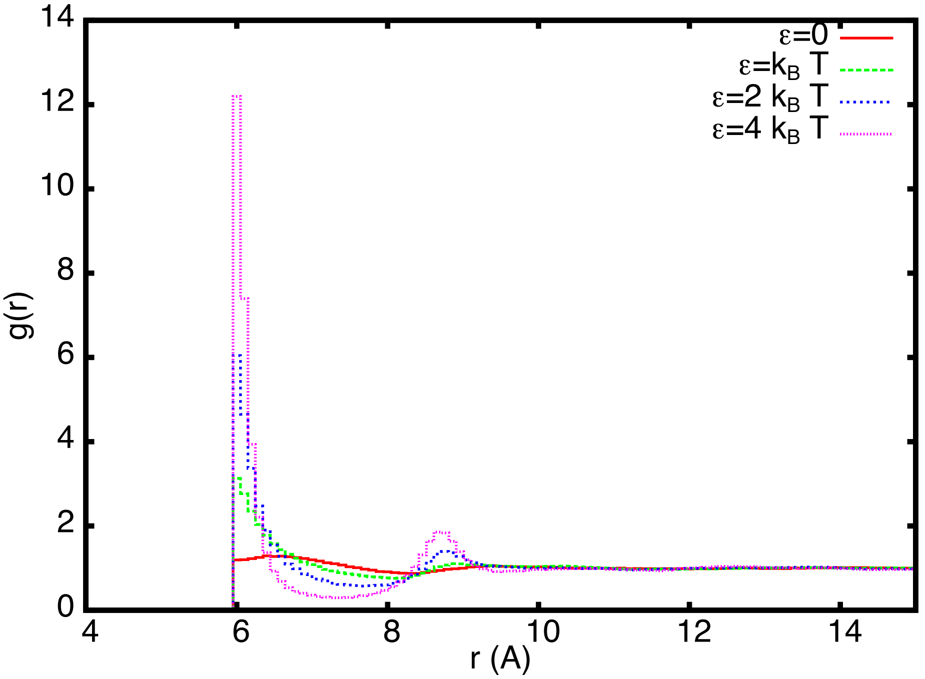

Figure 5.4: The solute-oxygen radial distribution function $g(r)$ as a function of the distance r of center of the water molecule (oxygen atom) from the center of the solute at different values of ϵ_1. The density of solvent at the point of contact with solute increases with ϵ_1. The length of the simulation box was around $L \sim 30\mathring{A}$.

as the value of ϵ_1 is increased. In the TFEP simulations, the center of each water molecule is subject to the mapping transformation in Eq. 4.86. The solute-oxygen radial distribution functions, $g(r)$, are plotted in Fig 5.4. Just as in the Lennard-Jones simulations, we find that the contact density increases as ϵ_1 is increased and the solute is made more hydrophilic.

The results from the free energy calculations are tabulated in Table 5.3 and the

$\beta\epsilon_1$	FEP	TFEP (F)	TFEP(R)	BAR	C
0	2.492 ± 0.070	2.54 ± 0.056	2.22 ± 0.09	2.359 ± 0.054	0.345
1	5.845 ± 0.121	5.883 ± 0.121	6.00 ± 0.112	5.971 ± 0.048	0.325
2	12.050 ± 0.561	12.09 ± 0.129	11.98 ± 0.107	12.081 ± 0.058	0.294
4	$--$	27.252 ± 0.207	27.29 ± 0.194	27.250 ± 0.073	0.239

Table 5.3: Estimates of ΔF along with error bars at different values of ϵ_1 at $R_A = 6.0\mathring{A}$, $R_B = 6.05\mathring{A}$, $P = 1$atm and $T = 300K$ in water . Observe that TFEP becomes more efficient than FEP as ϵ_1 increases. At the highest values of ϵ_1, we did not observe a single realization where the region between R_A and R_B is vacant. All estimates of ΔF are in units kJ/mol. At $T = 300K$, $\beta^{-1} = 2.5\,$kJ/mol. The symbol C has been defined in Eq. 4.80 in Sec 4.5.

trends are identical to what was observed in the previous simulations. ΔF increases with ϵ_1 and the efficiency of the FEP method decreases rapidly. Consequently, while FEP is more efficient than TFEP for $\epsilon_1 = 0$, it becomes more efficient to use TFEP at higher values of ϵ_1. Indeed, at the highest value of ϵ_1, we were not able to obtain an estimate of ΔF using FEP from $N_s = 3000$ realizations [8] ($P(n = 0) \sim 10^{-5}$), while TFEP provides a rather good estimate of ΔF from the same number of equilibrium samples.

We also performed the same set of simulations at a lower value of the initial solute radius, $R_A = 4\AA$. The results from these simulations are given in Table 5.4. Here too we find that the TFEP method starts to become more efficient (relatively) as the value of ϵ_1 is increased.

5.3 Discussion

In the previous sections we observed that the relative effectiveness of the TFEP approach (in comparison to FEP) increases as contact density increases [9]. This can be attributed to the fact that the efficiency of the FEP approach depends sensitively on the contact density. In particular, recall that in the FEP approach we seek to compute the probability $P(n = 0)$ that the region between $R_A \leq r \leq R_B$ is vacant. This probability decreases dramatically with increase in contact density, and conse-

[8] Again, a long equilibrium trajectory was generated and points were sampled every 200 Monte-Carlo sweeps.

[9] Unless explicitly stated, we will only be concerned with free energy calculations of the kind described in the previous sections where the radius of the excluded volume interaction is increased.

$\beta\epsilon_1$	FEP	TFEP
0	1.628 ± 0.022	1.658 ± 0.039
1	3.189 ± 0.045	3.232 ± 0.047
2	5.601 ± 0.060	5.548 ± 0.050
4	12.048 ± 0.259	11.801 ± 0.073

Table 5.4: Estimates of ΔF along with error bars at different values of ϵ_1 at $R_A = 4.0\mathring{A}$, $R_B = 4.05\mathring{A}$, $P = 1\,\text{atm}$ and $T = 300K$ in water . Observe that the TFEP becomes more efficient that FEP as ϵ_1 increases. All estimates of ΔF are in units kJ/mol. At $T = 300K$, $\beta^{-1} = 2.5$kJ/mol.

quently the number of realizations required to obtain an estimate of $P(n = 0)$ (and ΔF) from FEP, $N_s \sim 1/P(n = 0)$, becomes rather large. On the other hand, in the TFEP calculation, the mapping transformation compresses the fluid particles in the region $R_A \leq r \leq L/2$ into the region $R_B \leq r \leq L/2$. Thus the solvent molecules do not encounter the hard sphere component of the solute as it is grown out and the work values are never infinite. While there is a penalty for this mapping transformation - the solvent particles might be compressed into energetically unfavorable configurations after the mapping transformation, thus resulting in high W values - both fluid particles in the region close to the solute and in the bulk contribute to this penalty. Thus one can argue that the efficiency of the TFEP approach will not depend as sensitively on the contact density as that of FEP, and using TFEP can become beneficial as the contact density is increased. In the following, we will present calculations that support this speculation. In particular, we will compare the efficiencies of the TFEP and FEP approaches by analytically deriving an expression for the variance of the free energy estimators in the two approaches.

5.3.1 Comparing the effectiveness of the FEP and TFEP approaches

We begin by considering the easier of the two, FEP. In the limit of a large of equilibrium samples, N_s, the variance of the free energy estimate obtained from FEP, ΔF_{N_s} is [33]

$$\langle(\Delta F_{N_s} - \langle\Delta F_{N_s}\rangle)^2\rangle = \frac{1/(P(n=0))}{\beta^2 N_s} \tag{5.3}$$

where $P(n=0) = e^{-\beta\Delta F}$.

We will derive an estimate for the variance of the TFEP estimator in the limit that R_A and R_B differ by some infinitesimal amount, $R_B = R_A + \delta R$ (see Eq. 5.10). In this limit, the mapping transformation Eq. 4.86 can be rewritten as

$$m(\mathbf{r}) = \mathbf{r} + \bar{\mathbf{u}}(\mathbf{r})\delta R, \tag{5.4}$$

where

$$\bar{\mathbf{u}}(\mathbf{r}) = \begin{cases} R_A^2 \cdot (L^3 - 8r^3)/(r^2 \cdot (L^3 - 8R_A^3))\hat{\mathbf{e}}_\mathbf{r} & \text{if } r \leq L/2 \\ 0 & \text{if } r > L/2 \end{cases} \tag{5.5}$$

where L denotes the length of the simulation box (in a particular realization in case of constant pressure simulations). Using Eq. 5.4, the expression for the work performed in the TFEP simulations (see Eq. 4.87) can be rewritten as

$$W = \tfrac{1}{2}\textstyle\sum_{k\neq l}^{N_p}\left[V(\mathbf{r}_k + \bar{\mathbf{u}}(\mathbf{r}_k)\delta R, \mathbf{r}_l + \bar{\mathbf{u}}(\mathbf{r}_l)\delta R) - V(\mathbf{r}_k, \mathbf{r}_l)\right]$$

$$-\beta^{-1}\ln(1 + \textstyle\sum_{k=1}^{N_p}\nabla\cdot\bar{\mathbf{u}}(\mathbf{r}_k)\delta R)$$

Performing a Taylor expansion to first order in δR, we get

$$W = \delta R \sum_{k=1}^{N_p}\left[-\bar{\mathbf{u}}(\mathbf{r}_k)\cdot\mathbf{F}_k(\mathbf{r}_k) - \beta^{-1}\nabla\cdot\bar{\mathbf{u}}(\mathbf{r}_k)\right] \tag{5.6}$$

where

$$\nabla\cdot\bar{\mathbf{u}} = \begin{cases} -\frac{24R_A^2}{L^3-8R_A^3} & \text{if } r \leq L/2 \\ \\ 0 & r > L/2 \end{cases} \tag{5.7}$$

and $\mathbf{F}_k(\mathbf{r}_k)$ denotes the force on the k^{th} solvent particle due to the solute and all the other solvent particle. In the case of molecular solvents, $\mathbf{F}_k$ denotes the force on the k^{th} molecule (sum of the forces on all the atoms in the molecule) and $\mathbf{r}_k$ denotes the position of the center of the k^{th} molecule. We have suppressed the dependence of the force on the positions of all the other fluid particle for convenience. Finally in this limit, the TFEP identity for ΔF is equivalent to the following equation (again using a Taylor expansion to first order in δR)

$$\langle-\sum_{k=1}^{N_p}\left[\bar{\mathbf{u}}(\mathbf{r}_k)\cdot\mathbf{F}_k(\mathbf{r}_k) + \beta^{-1}\nabla\cdot\bar{\mathbf{u}}(\mathbf{r}_k)\right]\rangle\delta R = \Delta F. \tag{5.8}$$

We note in passing the similarity [1] between the expression for ΔF on the L.H.S of Eq. 5.6 and the virial expression for pressure [10] (for a discussion of other such "hyper virial" expression, see [1]). In fact, in the limit $R_A \to \infty$, growing out the solute is locally equivalent to moving a hard wall. The free energy cost ΔF_{HW} associated with displacing a hard wall with surface area A by δR is related to the bulk pressure P, $P = \Delta F_{HW}/(\delta R A) = 1/A(\partial F_{HW}/\partial R)$. This connection between P and ΔF_{HW} becomes apparent if we consider the free energy cost associated with a process in which one of the faces of the simulation box (assume constant volume simulation for now), say the face on the y-z plane at $x = -L/2$, is displaced by δR to $x = -L/2 + \delta R$. ΔF_{HW} can be computed using the TFEP method by constructing a mapping transformation like Eq. 5.4, with $\bar{\mathbf{u}}(\mathbf{r}) = \bar{\mathbf{u}}_{HW}(\mathbf{r}) = (L/2 - x)\hat{\mathbf{e}}_x/L$ acting on all particles inside the simulation box. Substituting this expression into Eq. 5.8, and after some algebra, we obtain

$$\frac{1}{L^3}\langle\sum_{k=1}^{N_p}\left[x_k\hat{\mathbf{e}}_x \cdot \mathbf{F}_k(\mathbf{r}_k)\right]\rangle + \beta^{-1}\frac{N_p}{L^3} = \frac{\Delta F_{HW}}{\delta R L^2}. \tag{5.9}$$

The L.H.S of Eq. 5.9 the usual virial expression for the pressure.

The variance of the estimate of ΔF obtained from N_s equilibrium samples using Eq. 5.8 is given by

$$\sigma^2(N_s) = \frac{\langle(-\sum_{k=1}^{N_p}\left[\bar{\mathbf{u}}(\mathbf{r}_k)\cdot\mathbf{F}_k(\mathbf{r}_k) + \beta^{-1}\nabla\cdot\bar{\mathbf{u}}(\mathbf{r}_k)\right]\delta R - \Delta F)^2\rangle}{N_s}. \tag{5.10}$$

Using $\Delta F = (\partial F/\partial R)\delta R$, we rewrite the above equation as

$$\sigma^2(N_s) = \Delta F^2 \frac{\langle\left[\sum_{k=1}^{N_p}\bar{\mathbf{u}}(\mathbf{r}_k)\cdot\mathbf{F}_k(\mathbf{r}_k) + \beta^{-1}\nabla\cdot\bar{\mathbf{u}}(\mathbf{r}_k)\right]^2\rangle - (\frac{\partial F}{\partial R})^2}{(\frac{\partial F}{\partial R})^2 N_s}. \tag{5.11}$$

In the following, we will separately consider the contributions to the sum $(\sum_{k=1}^{N_p}[\ldots])$ in the numerator of the equation above from the molecules in the bulk, which we denote will denote by x_b, and the molecules close to the solute, which we will denote by x_s. This allows us to write

$$\sum_{k=1}^{N_p}\left[\bar{\mathbf{u}}(\mathbf{r}_k)\cdot\mathbf{F}_k(\mathbf{r}_k) + \beta^{-1}\nabla\cdot\bar{\mathbf{u}}(\mathbf{r}_k)\right] = x_s + x_b. \tag{5.12}$$

The bulk is defined to be the region in which the influence of the solute is not felt and the fluid is homogenous and uniform. The R.H.S of Eq. 5.11 can then be written as

$$\sigma^2(N_s) = \Delta F^2 \frac{\langle x_s^2 \rangle + \langle x_b^2 \rangle + 2\langle x_s x_b \rangle - (\frac{\partial F}{\partial R})^2}{(\frac{\partial F}{\partial R})^2 N_s} \tag{5.13}$$

Ignoring for the moment correlations between x_s and x_b, we will rewrite the product $\langle x_s x_b \rangle$ as $\langle x_s \rangle \langle x_b \rangle$.

In the thermodynamic limit, $N_p \to \infty$, $L \to \infty$, $N_p/L^3 = \rho$, where ρ is the bulk density of the solvent, the average force on any particle in the bulk is zero from symmetry considerations. Hence, the terms $\bar{\mathbf{u}}(\mathbf{r}_k) \cdot \mathbf{F}_k(\mathbf{r}_k)$ do not contribute to the bulk average. Further, from Eq. 5.7, we get

$$\langle \left[\sum_{bulk} \beta^{-1} \nabla \cdot \bar{\mathbf{u}}(\mathbf{r}_k) \right] \rangle = -\beta^{-1} 24 \langle N_{bulk}/(L^3 - 8R_A^3) \rangle, \tag{5.14}$$

where N_{bulk} denotes the number of solvent particles in the bulk region in a particular realization. In the thermodynamic limit, this average can be be written in a simpler form using

$$\langle N_{bulk}/(4/3\pi(L/2)^3) \rangle = \rho$$

$$\langle [\textstyle\sum_{bulk} \beta^{-1} \nabla \cdot \bar{\mathbf{u}}(\mathbf{r}_k)] \rangle = -\beta^{-1} 4 R_A^2 \pi \rho.$$

Hence,

$$\langle x_b \rangle = -\beta^{-1} 4 R_A^2 \pi \rho. \tag{5.15}$$

Eq. 5.15 then implies that $\langle x_s \rangle \equiv -\partial F/\partial R - \langle x_b \rangle = -\partial F/\partial R + \beta^{-1} 4 \pi R_A^2 \rho$. These relations allow us to rewrite Eq. 5.13 as [10]

$$\sigma^2(N_s) + \frac{\Delta F^2}{N_s} = \Delta F^2 \frac{\langle x_s^2 \rangle + \langle x_b^2 \rangle + 2(\beta^{-1} 4 R_A^2 \pi \rho)^2 + 2\frac{\partial F}{\partial R}\beta^{-1} 4 R_A^2 \pi \rho}{(\frac{\partial F}{\partial R})^2 N_s}. \qquad (5.16)$$

Let us now consider estimating the free energy difference, $\Delta F = F_B - F_A$, with many solutes at different values of $\epsilon_1 \geq 0$. Assume that we have adjusted $R_B - R_A = \delta R$ such the value of ΔF is the same in each case and therefore the variance of the estimate of ΔF obtained from FEP remains the same in all the cases (see Eq. 5.3). Next, consider the variance of the estimate obtained from TFEP (Eq. 5.11) as the value of ϵ_1 is increased. Scaled particle theory [94] tells us that $\partial F / \partial R$ is proportional to the the density of solvent at the point of contact with solute, ρ_c, $\partial F/\partial R = \beta^{-1} 4 \pi R^2 \rho_c$. As ρ_c increases with ϵ_1, so does $\partial F/\partial R$.

Since the quantities $\langle x_b^2 \rangle$ and $2(\beta^{-1} 4 R_A^2 \pi \rho)^2$ are bulk properties and will not change (appreciably) when ϵ_1 is increased, the increase in contact density with ϵ_1 implies that the ratio $(\langle x_b^2 \rangle + 2(\beta^{-1} 4 R_A^2 \pi \rho)^2)/(N_s(\partial F/\partial R)^2)$ in Eq. 5.16 decreases with ϵ_1. It is easy to see that the last ratio in the R.H.S of Eq. 5.16 also decreases with ϵ_1. We are then only left with the quantity $\langle x_s^2 \rangle / (N_s(\partial F/\partial R)^2)$.

To study the behavior of this quantity at various of ϵ_1, we analyzed the statistics of x_s [11] and computed estimates of $\langle x_s^2 \rangle / \langle x_s \rangle^2$. The results are plotted in Fig 5.5. In these, we see a clear drop in the value of this quantity as ϵ_1 is increased. These results empirically suggest that $\langle x_s^2 \rangle / (N_s(\partial F/\partial R)^2)$ decreases with increasing

[10] We have obtained expressions for $\langle x_s^2 \rangle$ and $\langle x_b^2 \rangle$ in the thermodynamic limit. However, they are not central to the analysis presented below and hence have been omitted.

[11] Only the solvent particles in the shell $R_A \leq r \leq R_A + 2.5\sigma_1$ were considered for the x_s calculation.

ϵ_1. Hence it is reasonable to speculate that in general, the variance of the TFEP estimate of ΔF will decrease as ϵ_1 is increased, and as we observed in the previous section, at some point it might be more beneficial to use TFEP rather than FEP to estimate the same free energy difference. Even in the case of simulations at increasing values of bulk pressure (as with the first set of simulations performed in this chapter), we find that the relative fluctuations in x_s decrease with increasing contact density. In such instances, while the bulk quantities in numerator of Eq. 5.16 also grow with pressure, we do not expect them to grow faster than ρ_c^2. Consequently, we expect our hypothesis to hold true even here.

The decrease in the relative fluctuations of x_s with increasing contact density ρ_c is interesting in the light of recent results obtained by Hummer, Chandler, Garde and co-workers [30, 73, 81]. In their studies, they considered various model solutes, both hydrophobic and hydrophilic, solvated in water, and studied the properties of water-solute interface. In particular they observed that the interfacial region between a large hydrophobic solute and water is wide, highly compressible and resembles a liquid vapor interface [94]. As the solute is made more hydrophilic, the interface becomes well defined (the interfacial width decreases), the relative fluctuations in the particle numbers in the interfacial region decrease, and the interfacial region begins to resemble the bulk liquid in its properties. In our simulations, we make the solute more hydrophilic (we use this term in an extended sense to include fluids like the Lennard-Jones fluid) by increasing ϵ_1 or by increasing the pressure and as we observed in Fig 5.5, this reduces the relative fluctuations in x_s. It will be interesting to see if this decrease can be directly related to the change in the nature of the

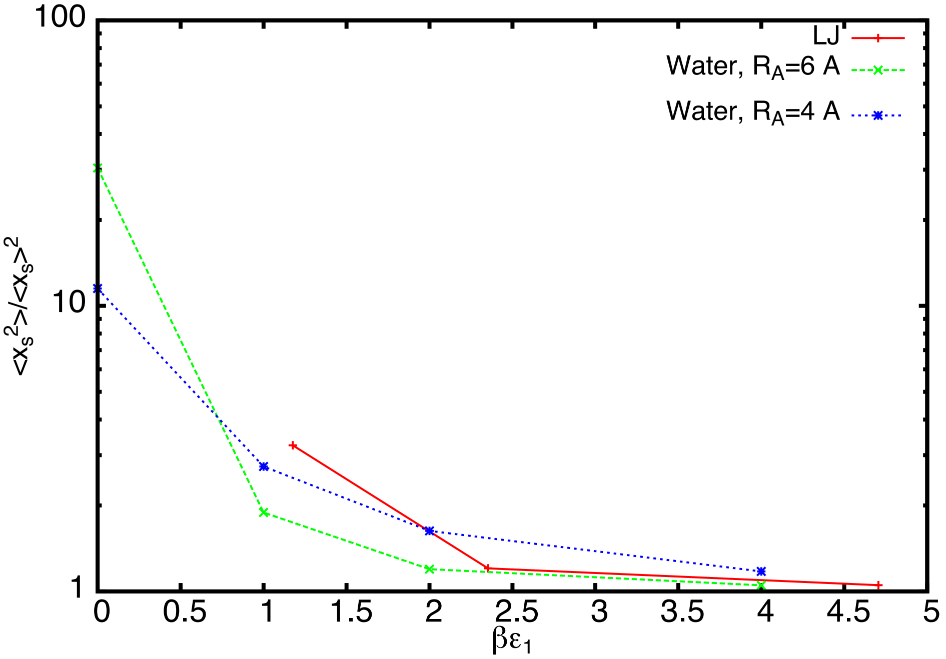

Figure 5.5: Plots of $\langle x_s^2 \rangle / \langle x_s \rangle^2$ at different values of $\beta\epsilon_1$ for the solute described by Eq. 5.1 in a) Lennard-Jones (LJ) solvent with $R_A = 2\sigma$, b) spherically truncated SPC/E water with $R_A = 6\AA$, and c) spherically truncated SPC/E water with $R_A = 4\AA$. We did not plot the value corresponding to $\epsilon_1 = 0$ in case (a) because x_s was negative in this instance. When x_s becomes negative, the trends in $\langle x_s^2 \rangle / \langle x_s \rangle^2$ do not accurately represent the trends in $\langle x_s^2 \rangle / (N_s (\partial F / \partial R)^2)$.

solute-solvent interface as the solute becomes more hydrophilic.

We have not investigated the effectiveness of the TFEP estimator as a function of the size of the solute in this chapter. We can however make some rough predictions in this direction using an analytical expression for the variance of the TFEP estimator (obtained from Eq. 5.16). In particular, we find that the variance is proportional to $1/R_A$. This suggests that the TFEP method will be more efficient for larger solutes. In the limit of an infinite solute, $R_A \to \infty$, the variance of the TFEP vanishes implying that TFEP will always be more efficient [12]. This result is not very surprising given the connection between the virial expression for pressure and the TFEP estimator when $R_A \to \infty$. As we discussed previously, the pressure is related to the $\partial F/\partial R$ in this limit. Hence the pressure can either be computed by estimating the contact density which is equivalent to the FEP approach or by using the virial expression which is equivalent to the TFEP estimator. It is a well established fact that it is a beneficial to estimate P using the virial expression rather than by estimating the contact density directly.

The simulations performed in this chapter involved sudden switching of the solute radius. These simulations allowed us to analyze the effectiveness of the mappings in Eq. 4.86. We believe that the conclusions of this analysis will also be valid for escorted simulations in which the radius of the solute is grown gradually.

The analysis presented here (and the results discussed in this chapter) assume

[12] We have performed some preliminary simulations in which we compared the effectiveness of FEP and TFEP for hard sphere solutes of various sizes in a WCA fluid. We find that the TFEP approach becomes more efficient as the size of the hard sphere solute increases.

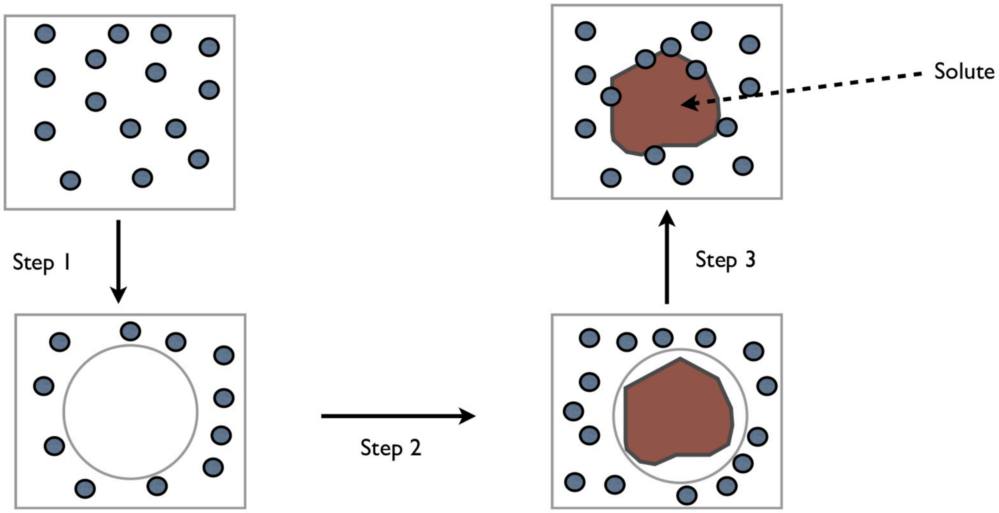

Figure 5.6: Schematic of the QCT calculaiton.

a hard solute with attractive interactions. A more realistic solute will not have hard sphere repulsions and will be better modeled by a potential with soft repulsions (say for example Eq. 5.1 without the hard repulsion). We have performed some preliminary simulations with such potentials and we find that the escorted method begins to become more efficient than usual nonequilibrium work relation as the solute becomes more hydrophilic.

5.3.2 Quasi Chemical Theory

The model calculations considered here can also be of use in Quasi Chemical Theory (QCT) developed by Pratt and coworkers [96]. In the QCT approach, a solute molecule with radius R is introduced into the solvent in three steps. First, in step (1) a spherical cavity with radius γ is created in the solvent and free energy cost of creating a spherical excluded volume of radius γ with the solvent in state

A is computed. This is equivalent to computing the solvation free energy of a hard sphere solute with radius γ in the fluid. Next, in step (2) the solute particle is placed inside the hard sphere (the solute particle and the hard sphere solute do not interact) and the interactions between the solute and the solvent are switched on. Finally, in step (3), the hard sphere solute is removed. In the final equilibrium state, the particle is solvated in the solvent fluid. The solvation free energy is calculated as the sum of the free energy differences in the three intermediate steps (see illustration in Fig 5.6).

This result does not immediately alleviate the problems associated with estimating solvation free energies using Eq. 2.5. If the radius of the hard sphere solute (γ) is much lesser than that of the actual solute, the free energy differences in Steps 1 and 3 contribute negligibly to the sum. Using Eq. 2.5 to compute the free energy difference in Step 2 when the interactions between the solute and the solvent are switched on will invariably be tough as we will encounter the same problems that plague the usual estimator of of the solvation free energy. If the size of the hard sphere solute is comparable or greater than the actual solute, computing the free energy difference associated with switching on interactions between the solvent and the solute in Step 2 will be relatively easy as overlaps between solute and solvent cores are avoided. Computing the free energy differences in Steps 1, 3 will still be hard. However, the free energy difference in Step 1 is independent of the interactions between the actual solute and solvent as we are simply solvating a hard sphere particle in the fluid. These can presumably be computed once and tabulated for future reference. The problem then reduces to computing the free energy difference in Step

3, or equivalently computing the free energy difference associated with creating a region of size γ around the solute from which the solvent molecules are excluded. In the ΔF calculation with the potential in Eq. 5.1, the radius of the excluded volume region is changed while holding other aspects of the potential fixed. This is analogous to the calculations one would perform in Step 3, and as we saw the mappings can be effective in such calculations.

5.4 Summary

In this chapter we compared the effectiveness of the FEP and TFEP methods in providing estimates of the free energy cost associated with growing a hard sphere solute with short range dispersive interactions (Eq. 5.1) in solvents composed of (a) Lennard-Jones particles (Sec. 5.2.1) and (b) water molecules (Sec. 5.2.2). The FEP identity is a limiting case of the nonequilibrium work relation, Eq. 2.5, when the external parameter is switched infinitely fast while the TFEP identity (see Eq. 4.41) is a limiting case of the escorted generalization of the nonequilibrium work relation, Eq. 4.39. Comparing the effectiveness of the FEP and TFEP approaches allows us to study the effectiveness of the escorting transformations, Eq. 4.86, in this free energy estimation problem. In our analysis, we found that the TFEP approach starts to outperform the FEP approach as the number of solvent molecules in contact with the solute particle increases. We expect this trend to hold in general and anticipate that it will be beneficial to use the escorted approach in such regimes.

Chapter 6

Protocol Postprocessing

[1] The previous chapters described a strategy, escorted free energy simulations, to improve the convergence of $\Delta F = F_B - F_A$ estimates obtained from nonequilibrium simulations in which a system of interest is driven irreversibly between two equilibrium states A, B by varying an external paramter λ at a finite rate $\dot{\lambda}$ using a protocol $\lambda(t)$. This method involved generating artificial trajectories with reduced lag and dissipation. In this chapter, we consider an alternative strategy, *protocol postprocessing*. This strategy involves introducing a function $\lambda^*(t)$ with $\lambda^*(0) = \lambda(0)$, which we will refer to as the *analysis protocol*, and constructing an estimator (see Eq 6.7 below) for the free energy difference $\Delta F^*(t) \equiv F_{\lambda^*(t)} - F_{\lambda(0)}$

[1] This chapter is based on the publication: D. D. L. Minh, S. Vaikuntanathan "Density-Dependent Analysis of Nonequilibrium Paths Improves Free Energy Estimates II. A Feynman-Kac Formalism ", *J. Chem. Phys* **134**, *034117*, 2011. The paper was jointly written by Minh and Vaikuntanathan. The central result Eq. 6.7 was derived by Vaikuntanathan. The simulation codes were written by Minh and the results were jointly analyzed by Minh and Vaikuntanathan.

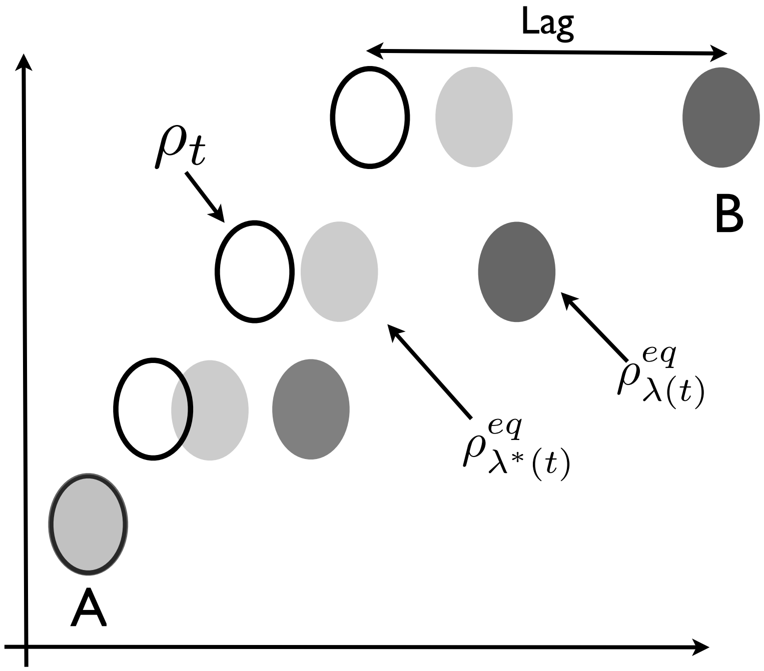

Figure 6.1: Lag in driven nonequilibrium processes. Consider a system driven from state A to state B in a finite-time process. In the above schematic, the ovals represent regions of phase space. The darkly shaded ovals are regions of phase space containing most of the density $\rho^{eq}_{\lambda(t)}$ of the equilibrium state corresponding to the value of the external parameter at time t. The unshaded ovals denote the phase space regions containing most of the density ρ_t actually accessed by the system during the process. In a reversible process, the two would be indistinguishable. Since the system is driven out of equilibrium, however, a lag builds up between ρ_t and $\rho^{eq}_{\lambda(t)}$. This lag is correlated to dissipation and is ultimately responsible for the poor convergence of free energy estimates based on nonequilibrium processes. If one is able to obtain a function $\lambda^*(t)$ with $\lambda^*(0) = A$ such that the equilibrium states $\rho^{eq}_{\lambda^*(t)}$ are closer to the ρ_t (e.g. the lightly shaded ovals), then the convergence of free energy estimates may be improved using Eq. 6.7.

using trajectories generated in the original, i.e. unescorted, process. While this result is valid for any choice of $\lambda^*(t)$ and reduces to the nonequilibrium work relation Eq 2.5 for $\lambda^*(t) = \lambda(t)$, we will argue that Eq. 6.7 provides efficient estimates of the free energy difference $\Delta F^*(t)$ whenever the equilibrium densities corresponding to the analysis protocol $\lambda^*(t)$ have a high degree of overlap with density of the system (see Fig 6.1).

Protocol postprocessing was previously introduced by Minh [68] in the context of importance sampling in path-space. [3–5, 74, 76, 95, 106] In the present work, we utilize an alternative mathematical formalism, the Feynman-Kac theorem. [27, 42, 90]. The new formalism has at least two advantages over the previous method: first, in certain special cases, it is *analytically* a zero-variance estimator. Secondly, for a few simple model systems, we find that the bias and variance of free energy estimates are substantially reduced.

6.1 Protocol Postprocessing strategy

As usual, we are interested in driven nonequilibrium processes in which the system is first prepared in equilibrium with $\lambda = \lambda(0)$ and temperature β^{-1}, after which the external parameters are switched according to the protocol $\lambda(t)$. Just as in Chapter 2, we will assume that the dynamics of the system preserve the canonical distribution when λ is held fixed. Each realization of the nonequilibrium process will again be described by the trajectory $\{\mathbf{z}_t\}$. The phase space density $\rho(\mathbf{z}, t)$ of an ensemble of such trajectories evolves according to the Liouville-type equation, Eq 4.2 which we reproduce here for convenience,

$$\frac{\partial \rho(\mathbf{z},t)}{\partial t} = \mathcal{L}_{\lambda(t)} \cdot \rho(\mathbf{z},t), \tag{6.1}$$

where the operator $\mathcal{L}_\lambda$ has the property $\mathcal{L}_\lambda \cdot e^{-\beta H_\lambda(\mathbf{z})} = 0$ [42,48], i.e. the dynamics preserve the equilibrium distribution.

In the protocol postprocessing strategy, trajectories are first generated according to the *sampling* protocol $\lambda(t)$. Next, a potentially distinct *analysis* protocol $\lambda^*(t)$, with $\lambda^*(0) = \lambda(0)$, is introduced. This analysis protocol is not used to generate any new trajectories. Rather, the previously generated trajectories are used as samples for estimating the free energy difference $\Delta F^*(t) \equiv F_{\lambda^*(t)} - F_{\lambda^*(0)}$. The standard form of nonequilibrium work relation can be seen as a special case where the sampling and analysis protocols are identical. While the formalism described below is valid for any $\lambda^*(t)$, it will not always be advantageous. In Section 6.3, however, we will describe how to choose a protocol $\lambda^*(t)$ that leads to an efficient free energy estimate.

We begin the derivation by formally separating the evolution operator into two terms,

$$\mathcal{L}_{\lambda(t)} = \mathcal{L}_{\lambda^*(t)} + \mathcal{A}(t), \tag{6.2}$$

where the auxiliary operator $\mathcal{A}(t)$ represents the difference between the evolution operators given the sampling and analysis protocols.

Following Hummer and Szabo's approach [42], consider a density $g(\mathbf{z},t)$ that satisfies a "sink" equation analogous to Eq. 4.19,

$$\frac{\partial g(\mathbf{z},t)}{\partial t} = \mathcal{L}_{\lambda(t)} \cdot g(\mathbf{z},t) + w^*(\mathbf{z},t) g(\mathbf{z},t), \tag{6.3}$$

where the function $w^*(\mathbf{z},t)$ includes not only a time-derivative of the Hamiltonian, but also a term containing the operator $\mathcal{A}(t)$, explicitly,

$$w^*(\mathbf{z},t) = -\beta\left(\frac{\partial H_{\lambda^*(t)}(\mathbf{z})}{\partial t} + \frac{\mathcal{A}(t)\cdot e^{-\beta H_{\lambda^*(t)}(\mathbf{z})}}{\beta e^{-\beta H_{\lambda^*(t)}(\mathbf{z})}}\right). \tag{6.4}$$

Here, the operator $\mathcal{A}(t)$ only acts on the factor $e^{-\beta H_{\lambda^*(t)}(\mathbf{z})}$ in the numerator. One solution to Eq. 6.3 is $g(\mathbf{z},t) = Z^{-1}_{\lambda(0)} e^{-\beta H_{\lambda^*(t)}(\mathbf{z})}$ as verified by explicit substitution.

By equating this solution to the path integral solution obtained from the Feynman-Kac theorem [27,42], we obtain an equation analogous to Eq. 4.23, namely

$$\frac{e^{-\beta H_{\lambda^*(t)}(\mathbf{z})}}{Z_{\lambda(0)}} = \left\langle \delta(\mathbf{z}-\mathbf{z}_t) e^{-\beta \mathcal{W}^*_t}\right\rangle_\Lambda . \tag{6.5}$$

where the angled brackets $\langle ...\rangle_\Lambda$ denote a path-ensemble average, or expectation, over all possible realizations of the driven nonequilibrium process with the protocol $\lambda(t)$; the protocol $\lambda^*(t)$ has nothing to do with sampling and the work $\mathcal{W}^*_t$ has the modified form,

$$\mathcal{W}^*_t = \int_0^t ds\ \left(\frac{\partial H_{\lambda^*(s)}(\mathbf{z}_s)}{\partial s} + \frac{\mathcal{A}(s)\cdot e^{-\beta H_{\lambda^*(s)}(\mathbf{z}_s)}}{\beta e^{-\beta H_{\lambda^*(s)}(\mathbf{z}_s)}}\right). \tag{6.6}$$

Integrating over $\mathbf{z}$, we obtain a protocol postprocessing form of nonequilibrium work relation, namely,

$$e^{-\beta \Delta F^*(t)} = \left\langle e^{-\beta \mathcal{W}^*_t}\right\rangle_\Lambda . \tag{6.7}$$

As a specific example, let us consider a system moving with overdamped Langevin (Brownian) dynamics in a one-dimensional potential $U_{\lambda(t)}(q)$. The density $\rho(q,t)$ evolves according to the Smoluchowski equation,

$$\frac{\partial \rho}{\partial t} = \mathcal{L}_{\lambda(t)}\rho = \frac{1}{\zeta}\frac{\partial}{\partial q}\left(U'_{\lambda(t)}(q)\rho\right) + D\frac{\partial^2}{\partial q^2}\rho, \tag{6.8}$$

where $D^{-1} = \beta\zeta$ is the diffusion coefficient and the prime symbol represents a derivative with respect to q.

Given an analysis protocol $\lambda^*(t)$, the auxiliary operator $\mathcal{A}(t)$ for this example system is defined as,

$$\mathcal{A}(t) \cdot f \equiv -\beta D \frac{\partial}{\partial q} \left(\Delta U'(q,t) f\right), \tag{6.9}$$

where

$$\Delta U(q,t) \equiv U_{\lambda^*(t)}(q) - U_{\lambda(t)}(q). \tag{6.10}$$

Substituting this expression into Eq. 6.6, we obtain a modified form of the work,

$$\begin{aligned} \mathcal{W}_t^* &= \int_0^t ds \left(\frac{\partial U_{\lambda^*(s)}(q_s)}{\partial s} - \frac{\beta D \frac{\partial}{\partial q} \left(\Delta U'(q_s, s) e^{-\beta U_{\lambda^*(s)}(q_s)}\right)}{e^{-\beta U_{\lambda^*(s)}(q_s)}} \right) \\ &= \int_0^t ds \left(\frac{\partial U_{\lambda^*(s)}(q_s)}{\partial s} + \beta^2 D \Delta U'(q_s, s) U'_{\lambda^*(s)}(q_s) - \beta D \Delta U''(q_s, s) \right) \end{aligned} \tag{6.11}$$

Using this expression for W_t^* in Eq. 6.7, we can now estimate the free energy difference $F_{\lambda^*(t)} - F_{\lambda(0)}$ from trajectories generated in the process in which external parameter is switched according to the protocol $\lambda(t)$.

6.2 Importance Sampling Formalism

Section 6.1 is not the first description of protocol postprocessing; it was preceded by a formalism based on importance sampling by Minh [68]. In this section, we describe the previous formalism in the current notation and compare it with the present results.

Explicitly in terms of path integrals, we may rewrite Eq. 2.5 as,

$$e^{-\beta \Delta F^*(t)} = \left\langle e^{-\beta W_t^*} \right\rangle_{\Lambda^*} \equiv \frac{\int d\gamma \; e^{-\beta W_t^*} P_{\lambda^*(t)}[\gamma]}{\int d\gamma \; P_{\lambda^*(t)}[\gamma]} \tag{6.12}$$

where $\gamma \equiv \{\mathbf{z}_t\}$ denotes a trajectory, $W_t^* \equiv \int_0^t ds \left(\frac{\partial H_{\lambda^*(s)}(\mathbf{z}_s)}{\partial s}\right)$ denotes the work performed on the system as it evolves along a particular trajectory in which the external parameter is changed according to the protocol $\lambda^*(t)$, $P_{\lambda^*(t)}[\gamma]$ is the probability density associated with the trajectory γ, when the external parameter is switched according to the protocol $\lambda^*(t)$, and $d\gamma$ is a measure over paths.

Now suppose that the external parameter is changed according to the protocol $\lambda(t)$ for which the associated probability density of a trajectory γ is $P_{\lambda(t)}[\gamma]$. The same free energy difference may be computed by estimating different path integrals, [68, 106]

$$e^{-\beta \Delta F^*(t)} = \frac{\int d\gamma \; e^{-\beta W_t^*} \left(\frac{P_{\lambda^*(t)}[\gamma]}{P_{\lambda(t)}[\gamma]}\right) P_{\lambda(t)}[\gamma]}{\int d\gamma \; \left(\frac{P_{\lambda^*(t)}[\gamma]}{P_{\lambda(t)}[\gamma]}\right) P_{\lambda(t)}[\gamma]} \equiv \frac{\left\langle r e^{-\beta W_t^*}\right\rangle_\lambda}{\langle r \rangle_\lambda} \tag{6.13}$$

where $r = P_{\lambda^*(t)}[\gamma]/P_{\lambda(t)}[\gamma]$ is the ratio of densities. If the two protocols sampling are identical, then $r = 1$.

This expression differs from Eq. 6.7 in that it includes two expectations, the definitions of work are different, and it requires a ratio of probabilities, r. The ratio is different from a "modification" to the work term. For example, in overdamped Langevin dynamics, this ratio is, [68, 70]

$$r = \exp\left[-\frac{\beta}{2}\left(\Delta U(q_t, t) + \int_0^t ds \; \left(\frac{\beta D \Delta U'(q_s^2)}{2} - D\Delta U''(q_s, s) - \frac{\partial \Delta U(q_s, s)}{\partial s}\right)\right)\right]$$

Now suppose that we break down $\mathcal{W}_t^*$ in Eq. 6.11, into one term with W_t^* and a "modification" term. If we multiply this modification term by $-\beta$ and take the exponent, we obtain a term which is used similarly to r,

$$\exp\left[-\beta\left(\int_0^t ds \left(\beta D \Delta U'(q_s, s) U'_{\lambda^*(s)}(q_s) - D\Delta U''(q_s, s)\right)\right)\right], \tag{6.14}$$

but is quite distinct.

In later sections, we will describe several advantages of the new formalism.

6.3 Dissipation and Lag

As protocol postprocessing is merely another mathematical formalism for computing free energies, there is no *a priori* reason to expect that it will perform any better or worse than the usual nonequilibrium work estimator, Eq. 2.5. For clever choices of the analysis protocol, however, we can show that Eq. 6.7 leads to a highly efficient estimator for $\Delta F^*(t)$. In this section, we will follow the approaches outlined in Section 4.5 to establish this result.

6.3.1 Exactly solved models

Suppose that we construct a "perfect" analysis protocol $\lambda^*(t)$ whose instantaneous equilibrium density is equivalent to the nonequilibrium density, so that $\rho(\mathbf{z},t) = \rho^{eq}(\mathbf{z},\lambda^*(t))$, where $\rho^{eq}(\mathbf{z},\lambda) = F_\lambda^{-1} e^{-\beta H_\lambda(\mathbf{z})} = e^{-\beta(H_\lambda(\mathbf{z})-F_\lambda)}$ denotes the equilibrium distribution corresponding to β^{-1} and λ. When a perfect analysis protocol is used, then

$$\mathcal{W}_t^* = \Delta F^*(t) \tag{6.15}$$

for *every* trajectory. This may be seen by first substituting $\rho(\mathbf{z},t) = e^{-\beta(H_{\lambda^*(t)}(\mathbf{z})-F_{\lambda_t^*})}$ in the evolution equation,

$$\frac{\partial\rho(\mathbf{z},t)}{\partial t} = \mathcal{L}_\lambda(t)\cdot\rho(\mathbf{z},t) = \mathcal{L}_{\lambda^*(t)}\cdot\rho(\mathbf{z},t) + \mathcal{A}(t)\cdot\rho(\mathbf{z},t) \tag{6.16}$$

where we have used Eq. 6.2. Since $\mathcal{L}_{\lambda^*(t)} \cdot \rho(\mathbf{z}, t) = 0$ for this $\rho(\mathbf{z}, t)$, we obtain,

$$\begin{aligned} -\beta \left(\frac{\partial H_{\lambda^*(t)}(\mathbf{z})}{\partial t} - \frac{\partial F_{\lambda_t^*}}{\partial t} \right) e^{-\beta H_{\lambda^*(t)}(\mathbf{z}) - F_{\lambda_t^*}} &= \mathcal{A}(t) \cdot e^{-\beta (H_{\lambda^*(t)}(\mathbf{z}) - F_{\lambda_t^*})} \\ \frac{\partial F_{\lambda_t^*}}{\partial t} &= \frac{\partial H_{\lambda^*(t)}(\mathbf{z})}{\partial t} + \frac{\mathcal{A}(t) \cdot e^{-\beta (H_{\lambda^*(t)}(\mathbf{z}))}}{\beta e^{-\beta H_{\lambda^*(t)}(\mathbf{z})}} \end{aligned}$$

By substituting this into the modified work, Eq. 6.6 and integrating, we obtain Eq. 6.15. As this equation is valid for every trajectory, Eq. 6.7 is a *zero variance* estimator of $\Delta F^*(t)$.

As a demonstration of this principle, consider two exactly solved [68] models: a Brownian particle in a one dimensional harmonic oscillator that either (i) has its center moving at a constant velocity, or (ii) has a time-dependent natural frequency. In both cases, the potential has the general time-dependent form $U_{\lambda(t)}(q) = \frac{k(t)}{2}(q - \bar{q}(t))^2$ where the vector $\lambda(t) = \{k(t), \bar{q}(t)\}$ denotes the set of external parameters. The Smoluchowski equation describing the evolution of the phase space density $\rho(q, t)$ can be solved to give [68, 70]

$$\rho(q, t) = \sqrt{\frac{\beta k_T(t)}{2\pi}} e^{\frac{-\beta k_T(t)}{2}(q - q_T(t))^2}, \tag{6.17}$$

where

$$\begin{aligned} q_T &= \langle q \rangle, \\ k_T(t) &= 1/(\langle q^2 - \langle q \rangle^2), \end{aligned}$$

where $\langle \ldots \rangle$ denotes an average over the distribution $\rho(q, t)$. In case (i), the spring coefficient $k(t)$ is held fixed at k while $\bar{q}(t)$ is switched according to $\bar{q}(t) = vt$ ($\lambda(t) = \{k, vt\}$). In this case, the free energy difference is always zero and $k_T(t)$ is a constant, k, and,

$$q_T(t) = vt - \frac{v}{\beta Dk}(1 - e^{-\beta Dkt}). \tag{6.18}$$

In case (ii), $\bar{q}(t)$ is held fixed at $\bar{q}(t) = 0$ and the spring coefficient $k(t)$ is switched according to $k(t) = vt$ ($\lambda(t) = \{vt, 0\}$). In this case, $q_T(t) = 0$, and

$$k_T(t) = \frac{k(0)e^{2\beta D \int_0^t ds\ k(s)}}{1 + 2\beta Dk(0)\left[\int_0^t du\ e^{2\beta D \int_0^u ds\ k(s)}\right]}. \tag{6.19}$$

In either case, we may choose the analysis protocol $\lambda^*(t) \equiv \{k_T(t), q_T(t)\}$ such that $U_{\lambda^*(t)}(q) = \frac{k_T(t)}{2}(q - q_T(t))^2$. With this choice, the Boltzmann distribution corresponding to the analysis protocol is equal to $\rho(q, t)$. Hence, the modified work calculated from Eq. 6.11 is always equal to the free energy difference $\Delta F^*(t)$. In contrast, the importance sampling form of protocol postprocessing yields different work values for each trajectory.

6.3.2 Dissipation Bounds Lag

In general, it is not feasible to find a perfect analysis protocol. Indeed, in most cases, the nonequilibrium densities ρ will not belong to the family of equilibrium distributions indexed by λ, $\rho^{eq}(\mathbf{z}, \lambda)$. However, Eq. 6.15 suggests that efficient estimators of free energy energies can be obtained if we can find an analysis protocol $\lambda^*(t)$ such that $\rho^{eq}(\mathbf{z}, \lambda^*(t))$ closely resembles the nonequilibrium density $\rho(\mathbf{z}, t)$. In the following paragraphs, we will make this argument more rigorous.

The convergence of the protocol postprocessing form of nonequilibrium work relation will depend on a criterion analogous to that in the original form [51, 58]. To see this, consider the distribution associated with $\mathcal{W}_t^*$, $P(\mathcal{W}_t^*)$ and construct the distribution $Q(\mathcal{W}_t^*) = P(\mathcal{W}_t^*)\exp[-\beta(\mathcal{W}_t^* - \Delta F^*(t))]$. The distribution Q is normalized thanks to Eq. 6.7. If we now study the convergence requirements of Eq 6.7, (just as we studied the convergence requirements of Eq 2.5), we will find that in order for the estimate of $\Delta F^*(t)$ to converge reliably, values of $\mathcal{W}_t^*$ near the peak of Q need to adequately sampled from the distribution P. Thus, when the distributions P and Q are far apart, the estimate of $F_{\Lambda^*(t)}$ converges poorly. We can use the relative entropy $D[P||Q] = \int P\ln(P/Q)$ [15] to quantify the extent to which P and Q differ from each other. The "average dissipation" in the protocol processing, $\mathcal{W}_d^* \equiv \langle\mathcal{W}_t^*\rangle_\Lambda - \Delta F^*(t)$ is related to this relative entropy,

$$\langle\mathcal{W}_t^*\rangle_\Lambda - \Delta F^*(t) = \beta^{-1}D[P||Q]. \tag{6.20}$$

Hence, whenever the dissipation is lowered, the convergence of the free energy estimate is improved. This dissipation can in turn be related to the relative entropy between the distributions $\rho(\mathbf{z},t)$ describing the state of the system and the equilibrium state corresponding to $\lambda^*(t)$, $\rho^{eq}(\mathbf{z},\lambda^*(t))$ [2]. This relative entropy can be interpreted as a measure of the "lag" in the protocol postprocessing formalism,

$$\langle\mathcal{W}_t^*\rangle_\Lambda - \Delta F^*(t) \quad \geq \quad \beta^{-1}D[\rho(\mathbf{z},t)||\rho^{eq}(\mathbf{z},\lambda^*(t))]. \tag{6.21}$$

Eq. 6.21 suggests, but does not prove (the inequality goes the wrong way), that a reasonable strategy for reducing dissipation and improving the convergence of the free energy estimator is to choose an analysis protocol in which the "analysis" density closely resembles the evolving state of the system.

[2] This relation can be derived starting from Eq 6.5 and following the procedure outlined in Chapter 3

6.4 General Case

Based on the results in Section 6.3, we speculate that a reasonable strategy for minimizing dissipation and improving the efficiency of the free energy estimator is to choose an analysis protocol $\lambda^*(t)$ so that the Kullback-Leibler divergence $D[\rho(\mathbf{z},t)||\rho^{eq}(\mathbf{z},\lambda^*(t)))]$ is small for all t. Obtaining such a protocol will usually entail a search over the space of λ to find an equilibrium distribution $\rho^{eq}(\mathbf{z},\lambda^*(t)))$ that is similar to $\rho(\mathbf{z},t)$. While the nonequilibrium distribution is not analytically tractable for most systems, it is possible to use sampled trajectories to compare the relative entropy between $\rho(\mathbf{z},t)$ and $\rho^{eq}(\mathbf{z},\lambda)$ for different values of λ. Specifically, given a set of trajectories $\{\gamma_1, \gamma_2, ..., , \gamma_{N_s}\}$ and several candidate values of λ, the relative entropy $D[\rho(\mathbf{z},t)||\rho^{eq}(\mathbf{z},\lambda^*(t)))]$ is minimized by the parameter vector λ that minimizes $\langle H_\lambda(\mathbf{z})\rangle_{\rho(\mathbf{z},t)} - F_\lambda$, which may be estimated by the sample average, [68]

$$D_{Test}(\gamma,t) = \frac{1}{N_s}\left[\sum_{n=1}^{N_s} H_\lambda(\mathbf{z}_{nt})\right] - F_\lambda. \tag{6.22}$$

where $\mathbf{z}_{nt}$ denotes the state of system in phase space at time t as it evolves along the trajectory γ_n. It is sufficient to minimize $D_{Test}(\gamma,t)$ as the other integral in $D[\rho(\mathbf{z},t)||\rho^{eq}(\mathbf{z},\lambda^*(t)))]$, $\int d\mathbf{z}\ \rho(\mathbf{z},t)\ln\rho(\mathbf{z},t)$ does not depend on λ. A reasonable choice for the search space of λ is the range of the sampling protocol $\lambda(t)$. This choice has the advantage that $F_{\lambda(t)} - F_{\lambda(0)}$ may be estimated via nonequilibrium work relation; for distributions that are not accessed during the sampling protocol, it may be more difficult to estimate corresponding free energies.

As noted in Section 6.1, the flexibility in choosing λ^* means that the free energy $\Delta F^*(t)$ may be different from $\Delta F(t)$. Indeed, unless there is no lag, an analysis protocol that minimizes the lag will *always* have different states than the sampling protocol. Since we are typically interested in free energies between the end states of the sampling protocol ($A \equiv \lambda(0)$ and $B \equiv \lambda(T)$), this discrepancy was addressed by introducing an adaptive algorithm, nonequilibrium density-dependent sampling (NEDDS). [68] NEDDS is equally applicable to the current formalism.

In brief, NEDDS entails running all N_s desired simulations of the nonequilibrium process simultaneously. The sampling protocol initially involves an interpolation between the desired end states A and B. After reaching state B, the protocol extrapolates past it until an adaptively determined stopping time. (While such an extrapolation may not always be physically meaningful, it is nearly always computationally feasible.) Without loss of generality, let us assume that $A < B$. The stopping time is decided by performing the following calculations while the simulations are in progress:

1. The free energy difference, $F_{\lambda(t)} - F_{\lambda(0)}$, between the initial and instantaneous state at the current time step, t, is estimated using the nonequilibrium work relation.

2. D_{Test} is evaluated with λ values from the current state and all preceding states using Eq. 6.22.

3. If the choice of λ that minimizes D_{Test}, λ^{min}, is between A and B, $A < \lambda^{min} < B$, then it is appended to the analysis protocol, $\lambda^*(t) = \lambda^{min}$. Otherwise, if it is at or beyond B, $\lambda^{min} \geq B$, then the final value of the analysis protocol is set to B, $\lambda^*(t) = B$.

4. Lastly, $\mathcal{W}_t^*$ is incremented and $\Delta F^*(t)$ is evaluated by protocol postprocessing.

This procedure ensures that protocol postprocessing estimators can compute the free energy difference between the states A and B.

6.5 Model Systems

We now demonstrate NEDDS with protocol postprocessing (both importance sampling and Feynman-Kac) formalisms and compare its efficiency to standard sample mean estimates from nonequilibrium work relation, Eq. 2.5, on three model systems. First, consider an overdamped Brownian particle evolving on the one-dimensional surface,

$$U(q, \lambda) = q^4 - 16\lambda q^2, \tag{6.23}$$

as studied by Sun [95]. In this system, the free energy difference between the states $\lambda = 0$ and $\lambda = 1$ at $\beta = 1$ was analytically found to be $F_{\lambda=1} - F_{\lambda=0} = -62.9407$ [76]. Recall that we studied this model system in Sec 4.6.1.

Simulations of nonequilibrium driven processes were performed in which λ was switched between 0 and 1 according to the discretized equation of motion,

$$q_{j+1} = q_j - D\Delta t U_j' + \sqrt{2D\Delta t} R_j, \tag{6.24}$$

where q_j is the position at the j^{th} time step (or at time $j\Delta t$), $D = 1$ is the diffusion coefficient, $\Delta t = 0.0001$ is the time step, and R_j is a standard normal random variable. λ was incremented at each time step by $v\Delta t$. NEDDS was used to obtain the analysis protocol $\lambda^*(t)$ concluding at $\lambda^*(t) = 1$, and the free energy difference

$F_{\lambda=1} - F_{\lambda=0}$ was computed using either Eq. 6.7 or Eq. 6.13. For comparison, the standard nonequilibrium work relation estimate was applied to two types of simulations taking the same amount of simulation time as the analysis protocol obtained from NEDDS: either (i) λ was switched between 0 and 1 at a slower velocity, or (ii) the NEDDS analysis protocol was used as a new sampling protocol.

While the importance sampling formalism [68] was found to be an improvement over the standard form of nonequilibrium work relation, [68] we find that the estimator based on Eq. 6.7 is even better (Fig. 6.2). Even for the fastest switching rates, where dissipation is expected to be high, the systematic bias [107] is largely eliminated. No benefit was found from using the analysis protocol from NEDDS as a new sampling protocol; in fact, the bias was worse than with the constant velocity protocol.

We also performed similar tests on another one-dimensional surface,

$$U(q, \lambda) = (5q^3 - 10q + 3)q + \frac{15}{2}(q - \lambda(t))^2, \tag{6.25}$$

first described by Hummer [41]. Hummer's surface, a double well potential that includes a harmonic bias, mimics the setup of a single-molecule pulling experiment, and hence has been used to demonstrate estimators of free energies [69,71] and other quantities [72] in the context of these experiments. The simulations were performed using the same equation of motion, diffusion coefficient, and time step as described above for Sun's system and λ was switched from -1.5 to 1.5.

The performance trends with Hummer's system are similar to those with Sun's (Fig. 6.3). The exact free energy difference was calculated numerically [69] and is

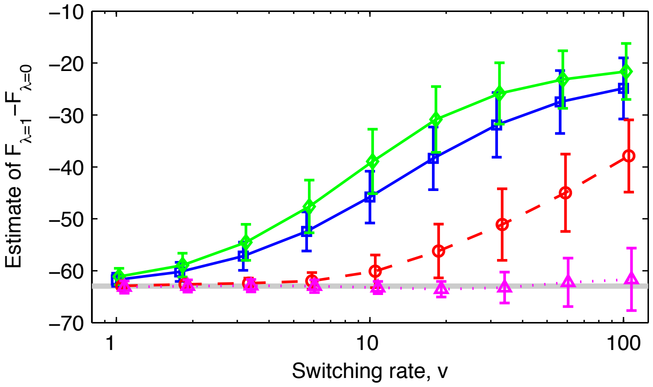

Figure 6.2: Comparison of free energy estimates for Sun's system: NEDDS simulations were analyzed with importance sampling, Eq. 6.13 (circles), or the Feynman-Kac formalism, Eq. 6.7 (triangles). Standard nonequilibrium work relation estimates, Eq. 2.5 (squares), were performed on slower simulations with the same total time as the NEDDS simulations or by using the analysis protocol as a new sampling protocol (diamonds). The symbols indicate the mean and error bars indicate the standard deviation of 10000 estimates, each based on 50 trajectories. The simulation time step was $\Delta t = 0.001$ and the rate v indicates that λ was incremented by $v\Delta t$ at each time step of the NEDDS simulations. While the switching rates are equivalent, some points are given a small horizontal offset to prevent error bar overlap. The exact free energy is shown as a shaded line.

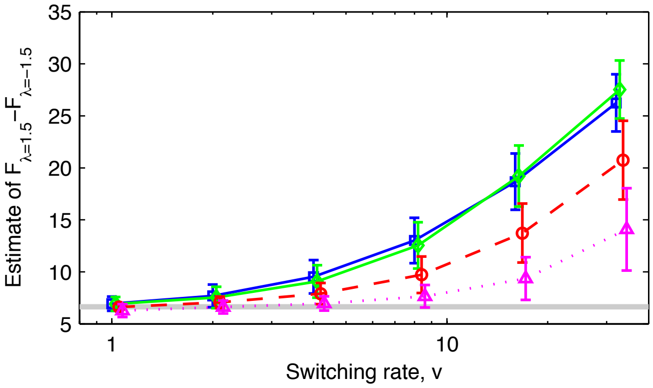

Figure 6.3: Comparison of free energy estimates for Hummer's system. The caption for Fig. 6.2 applies here, except that the potential is Hummer's rather than Sun's and each estimate is based on 250 trajectories.

shown as the shaded line. Results from the standard form of the nonequilibrium work relation are more biased than with NEDDS and the importance sampling formalism, which in turn is more biased than the Feynman-Kac formalism. In contrast to Sun's system, however, the estimates from Eq. 6.7 are noticeably biased at the fastest switching rates. Another distinction between the trends from the two systems is that results obtained using a constant velocity protocol and using the analysis protocol as a new sampling protocol are rather similar.

As a final demonstration, we consider a two-dimensional surface,

$$U(x, y, \lambda) = 5(x^2 - 1)^2 + 5(x - y)^2 + \frac{15}{2}(x + \cos(\pi\lambda))^2 + \frac{15}{2}(y + 1 - \sin(2\pi\lambda) - 2\lambda)^2, \tag{6.26}$$

in which λ dictates the progress of a harmonic bias along a curve (Fig. 6.4). Simulations were performed as with the 1D potentials, using Eq. 6.24 along each di-

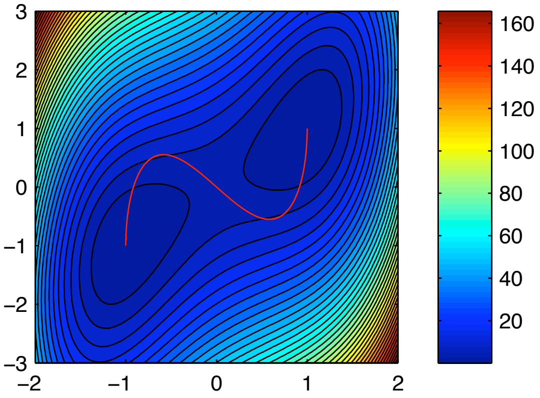

Figure 6.4: Potential energy surface for a 2D system. The contour plot is of $5(x^2-1)^2+5(x-y)^2$ and the red line traces the equilibrium position of the harmonic bias $\frac{15}{2}(x+\cos(\pi\lambda))^2+\frac{15}{2}(y+1-\sin(2\pi\lambda)-2\lambda)^2$ as λ goes from 0 (left) to 1 (right).

mension, as well as the same diffusion coefficient and time step and λ was switched from 0 to 1. The exact free energy difference is zero from symmetry arguments. The performance trends in this system are the same as in Hummer's system (Fig. 6.5).

6.6 Discussion and Conclusion

We have presented a method for analyzing nonequilibrium trajectories which borrows from a similar philosophy as previous work by Minh [68] but is based on a distinct mathematical formalism. The new formalism has the advantages that it analytically is a zero-variance estimator if a "perfect" analysis protocol is obtained, and it improves the convergence of free energy estimates in all our tested model systems. Further tests on more complex multidimensional systems are a potential future research direction.

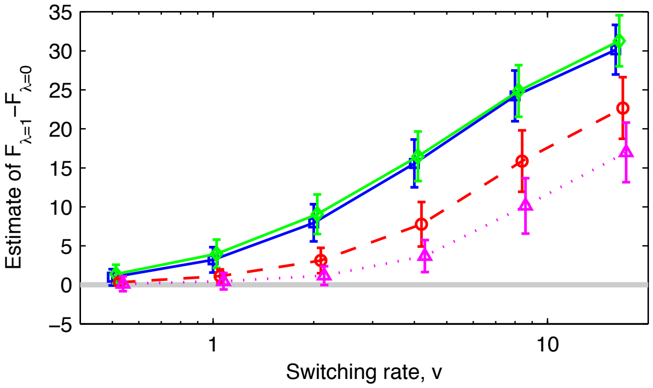

Figure 6.5: Comparison of free energy estimates for a 2D system with $\Delta F = 0$. The caption for Fig. 6.2 applies here, except that the potential is Eq. 6.26 rather than Sun's system, each estimate is based on 250 trajectories, and multidimensional versions of the importance sampling and Feynman-Kac formalisms were used.

We expect that protocol postprocessing will be most useful when (i) there is little phase space overlap between the end states of interest (otherwise free energy differences can be computed without nonequilibrium work identities), (ii) estimates of ΔF from the nonequilibrium work relation suffer from poor convergence for a given nonequilbrium process in which the system is driven between the end states of interest and (iii) it is reasonable to speculate that the nonequilibrium driven process has a nonequilibrium density $\rho(\mathbf{z}, t)$ that always resembles an equilibrium density $\rho^{eq}(\mathbf{z}, \lambda)$ parameterized by a λ vector along the protocol. Exact convergence properties, of course, will depend on the system.

Finally, we note that the protocol postprocessing method can readily be combined with the escorted free energy simulation approach described previously and it might be possible to construct efficient hybrid estimators of ΔF.

Chapter 7

Summary and future outlook

Nonequilibrium estimates of equilibrium free energy differences, ΔF, typically suffer from poor convergence due to dissipation. This thesis has developed methods to improve the efficiency of such estimates by reducing dissipation. The development of the methods was guided by an exact relation between the dissipation in a nonequilibrium process and the "lag", i.e. the extent to which the system deviates from the true equilbrium state in a nonequilibrium process.

The first strategy developed, "escorted" free energy simulations, involved modifying the dynamics ordinarily used to simulate the evolution of the system by adding artificial terms that couple the evolution to changes in the external parameter, and constructing estimators for ΔF in terms of these artificial trajectories. Whenever the artificial terms manage to reduce the lag and dissipation, our method provides an improved estimator of ΔF. We illustrated this method on a few model systems. In particular, we demonstrated how prior intuition for the problem, and mean-field arguments can be used to construct effective escorting dynamics.

A natural next step in this research will be to apply this approach to other free energy estimation problems. In the following, we list four free energy estimation problems which we think will be good test cases for the escorted free energy simulations approach. (1) In the dipole fluid example in Section 4.6.3, the particles interact via a simple dipole-dipole interaction that favors the alignment of dipoles. This model system is especially amenable to mean field treatment. In particular, when the parameter γ controlling the dipole-dipole coupling (see Eq. 4.88) is large, mean field theory can provide a rather accurate description of the system [31]. We used this property of the system to construct effective escorting dynamics. It will be interesting to see whether this approach works when the dipole interactions are modeled more realistically, say using the Stockmayer model (see [24]), and where mean field arguments might not be as effective in describing the system. (2) Another interesting free energy problem that combines elements of both the cavity expansion example (Section 4.6.2), and the dipole fluid example is computing the free energy cost associated with introducing a charged particle in a fluid. Developing methods to efficiently estimate this free energy is a long standing problem in computational chemistry [1]. In both problems (1) and (2), it might become necessary to improve on the escorted dynamics introduced in this thesis. For example, the escorted transformation developed for the cavity expansion example is near perfect for a reference system of ideal gas particles. The hard sphere fluid is another reference system whose thermodynamic properties are well established, and it will

[1] This problem was suggested by Attila Szabo and Gerhard Hummer

be worth investigating whether such reference systems can be used to develop better escorting transformation. (3) A technique used to compute the free energies of solids involves constructing a protocol in which the solid crystal is transformed into a Einstein crystal with the same lattice structure. The Einstein crystal is an analytically tractable system in which mutually non-interacting atoms are tethered to their respective lattice points by harmonic potentials. The protocol connecting the actual crystal state to the Einstein crystal state is one where the harmonic potential is gradually switched on by increasing the spring constant. At the end of the protocol, the harmonic potential is sufficiently strong that the inter-atomic interactions can be ignored and the final state can be treated as an Einstein crystal [23]. One approach to constructing escorting dynamics for this process is to consider the Einstein crystal as a reference system and construct *perfect* escorting dynamics for it. Such dynamics might be effective in providing efficient estimates of free energies of solids. (4) The escorted free energy approach could also be potentially useful in obtaining estimates of free energies (and potentials of mean force) from simulations of single molecule pulling experiments. This problem will also be an ideal test for the other approach introduced in this thesis, "protocol postprocessing", in which the trajectories ordinarily generated in the course of a nonequlibrium simulation are reprocessed to obtain efficient estimates of ΔF. This approach requires the construction of an *analysis* protocol that describes a sequence of equilibrium states that resemble the states visited by the system in the nonequilibrium process (see Fig. 6.1). In a single molecule pulling simulation where one end of the molecule of interest is stretched at some speed v (see the one dimensional analogue, Eq. 6.26), analysis protocols corresponding to a lower speed v^* might help improve the efficiency of the free energy estimate. One might also like to investigate a hybrid free energy estimation approach in which the two methods developed here, escorted free energy simulations and "protocol postprocessing" are used in conjunction to obtain efficient ΔF estimates.

Finally, there might be some interesting connections between the escorted free energy formalism developed here, and the recently discovered nonequilibrium work information fluctuation relations for thermodynamic processes evolving under feedback [38, 83, 88]. By accounting for the amount of information gained about the state of the system in a feedback process, these recent results have shown that nonequilibrium processes evolving under feedback also satisfy fluctuation theorems of the kind discussed in this thesis. Since the artificial escorted dynamics discussed in Chapter 4 in some sense model a continual feedback process [56], it will be useful to clarify the relationship between escorted fluctuation theorems, and fluctuation theorems for processes with feedback.

Appendix A

Fluctuation theorem for stochastic escorted simulations

Chapter 4 presents a fluctuation theorem for escorted Hamiltonian dynamics (Eq. 4.43-4.49), and for discrete-time Monte-Carlo dynamics (Eq. 4.51-4.63). In this appendix, we will first sketch a general derivation of the fluctuation theorem for continuous time escorted stochastic dynamics using the proof presented in the discrete-time Monte-Carlo case. We then present a derivation in the specific case of a one-dimensional system with physical dynamics described by an over-damped Langevin equation [54].

Let us imagine a pair of forward and reverse escorted process of duration τ (as usual, λ is switched from A to B according to $\lambda(t)$ in the forward process) with an escorting flow field $\mathbf{u}(\mathbf{z}, \lambda)$, and consider a discretization scheme in which the system is allowed to evolve under the physical dynamics for a time δt at fixed λ, after which

λ is switched by $\dot{\lambda(t)}\delta t$. Updates in λ are accompanied by a displacement of $\mathbf{u}\dot{\lambda(t)}\delta t$ in the phase-space coordinates of the system. In the limit $\delta t \to 0$, the discretization scheme described above is equivalent to escorted equations of motion, Eq. 4.3.

It helps to visualize this discretization scheme using Eq. 4.31 (reproduced below for convenience) with $N = \tau/\delta t$, and with the mapping functions $M_i : \mathbf{z} \to \mathbf{z}'$,

$$[\mathbf{z}_0, \lambda_0] \overset{M_0}{\Rightarrow} [\mathbf{z}'_0, \lambda_1] \to [\mathbf{z}_1, \lambda_1] \overset{M_1}{\Rightarrow} \cdots \to [\mathbf{z}_{N-1}, \lambda_{N-1}] \overset{M_{N-1}}{\Rightarrow} [\mathbf{z}'_{N-1}, \lambda_N], \tag{A.1}$$

where

$$\mathbf{z}' = \mathbf{z} + \mathbf{u}(\mathbf{z}, \lambda_i)\dot{\lambda(t)}\delta t, \tag{A.2}$$

and λ_i denotes the value of λ after the i^{th} time step, $\lambda_i = \lambda(0) + i\dot{\lambda(t)}\delta t$.

The evolution of the system in the i^{th} time interval from $\mathbf{z}'_{i-1}$ to $\mathbf{z}_i$ at fixed λ_i can be described by the transition probability $P_{\lambda_i}(\mathbf{z}_i|\mathbf{z}'_{i-1})$. As in Eq. 4.24, we will assume that this transition probability is detailed balanced. Commonly used equations of motion (such as Langevin, over-damped Langevin) satisfy this criterion. Following the proof of the fluctuation theorem for Monte-Carlo dynamics in Sec. 4.4, the trajectories generated according to this discretized scheme satisfy the fluctuation relation

$$\frac{P_F(W)}{P_R(-W)} = e^{\beta(W-\Delta F)}, \tag{A.3}$$

where

$$W[\{\mathbf{z}_i\}] = \sum_{i=0}^{N-1} \left[H_{\lambda_{i+1}}(\mathbf{z}'_i) - H_{\lambda_i}(\mathbf{z}_i) - \beta^{-1} \ln J_i(\mathbf{z}_i)\right]. \tag{A.4}$$

Taking the limit $\delta t \to 0$, and using $\ln J_i(\mathbf{z}_i) = \nabla \cdot \mathbf{u}(\mathbf{z}_i, \lambda_i)\dot{\lambda(t)}\delta t$, we can rewrite

Eq. A.5 as an integral,

$$W[\{\mathbf{z}_t\}] = \int_0^\tau \dot{\lambda} \left(\frac{\partial H}{\partial \lambda} + \mathbf{u} \cdot \nabla H - \beta^{-1} \nabla \cdot \mathbf{u} \right). \tag{A.5}$$

Eq. A.3 is now a fluctuation theorem for escorted stochastic molecular dynamics simulations.

We will now derive the fluctuation theorem for a system with one degree of freedom, x, when the physical dynamics of the system are over-damped Langevin equations of motion (Eq. A.6 below) without using the discretization scheme described above. The physical dynamics of the system are given by

$$\dot{x} = -\mu \frac{\partial V_{\lambda(t)}(x)}{\partial x} + \sqrt{2D}\zeta(t), \tag{A.6}$$

where $V_\lambda(x)$ denotes a one dimensional potential surface, $\mu\,, D$ denote the friction and diffusion constants respectively, and $\zeta(t)$ denotes a Gaussian white noise process with $\langle \zeta(t) \rangle = 0$, and $\langle \zeta(t)\zeta(t') \rangle = \delta(t - t')$. The physical dynamics satisfy the fluctuation dissipation relation, $D/\mu = k_B T = \beta^{-1}$.

Let us suppose we modify the equations of motion by adding an extra term $\dot{\lambda} u(x, \lambda)$,

$$\dot{x} = -\mu \frac{\partial V_{\lambda(t)}(x)}{\partial x} + \sqrt{2D}\zeta(t) + \dot{\lambda} u(x, \lambda). \tag{A.7}$$

We will discretize Eq. A.7 using the Ito convention [25] with a time step δt,

$$x_{i+1} - x_i = -\delta t \mu \frac{\partial V_{\lambda_i}(x_i)}{\partial x} + \sqrt{2D}\eta_i + \dot{\lambda} u(x_i, \lambda_i)\delta t. \tag{A.8}$$

where x_0 denotes the initial state of the system and is sampled from the equilibrium state A, $x_i\,, \lambda_i$ denote the location of the system and the value of the external parameter respectively at time $i\delta t$, and $\eta_i = \zeta(t_i)\delta t$ is a gaussian random variable with the properties $\langle \eta_i \rangle = 0$ and $\langle \eta_i \eta_j \rangle = \delta t \delta_{i,j}$ [25]. For convenience, we will define $f(x_i, \lambda_i, \dot{\lambda}) = \mu F_{\lambda_i}(x_i) + \dot{\lambda} u(x_i, \lambda_i)$, where $F_\lambda = -\partial V_\lambda / \partial x$. We wish to construct the conditional probability density associated with the path $\{x_0, x_1, \ldots, x_N\}$ given the initial point, x_0, $P(x_1, x_2, \ldots, x_N | x_0)$. To do so, following Seifert [91], we use the fact the distribution of $\{\eta_i\}$ values is known and perform a transformation of coordinates from $\{\eta_i\}$ to $\{x_i\}$ using Eq. A.8. The two densities are related according to

$$P(x_1, x_2, \ldots, x_N | x_0) = P(\eta_0, \eta_1, \ldots \eta_{N-1})/J = \frac{1}{J(\sqrt{2\pi\delta t})^N} \exp - \left[\sum_{i=0}^{N-1} \eta_i^2 / (2\delta t) \right], \tag{A.9}$$

where J denotes the Jacobian for the transformation. For this transformation, $J = (\sqrt{2D})^N$. Substituting $\eta_i = \left[x_{i+1} - x_i - f(x_i, \lambda_i, \dot{\lambda}) \delta t \right] / \sqrt{2D}$ (from Eq. A.8) in Eq. A.9, we get

$$P(x_1, x_2, \ldots, x_N | x_0) = \frac{1}{\left(\sqrt{4D\pi\delta t}\right)^N} \exp - \frac{1}{4D} \left[\sum_{i=0}^{N-1} \left(\frac{x_{i+1} - x_i}{\delta t} - f(x_i, \lambda_i, \dot{\lambda}) \right)^2 \delta t \right], \tag{A.10}$$

To obtain the fluctuation theorem, let us consider the conjugate trajectory $\{x_N, \ldots, x_0\}$ in the reverse process. We will again use the subscripts F and R to denote quantities corresponding to the forward and reverse processes. The density conditional probability density associated with this conjugate trajectory in the reverse process is given by

$$P_R(x_{N-1}, \ldots, x_0 | x_N) = \frac{1}{\left(\sqrt{4D\pi\delta t}\right)^N} \exp - \left[\frac{1}{4D} \sum_{i=0}^{N-1} \left(\frac{x_i - x_{i+1}}{\delta t} - f(x_{i+1}, \lambda_{i+1}, -\dot{\lambda}) \right)^2 \delta t \right]. \tag{A.11}$$

Taking the ratio of Eq. A.11 and Eq. A.10, we get

$$\frac{P_F(x_1, x_2, \ldots, x_N|x_0)}{P_R(x_{N-1}, \ldots, x_0|x_N)} = \exp -\frac{1}{D}\sum_{i=0}^{N-1}\Big[K_i(x_{i+1}-x_i) + \mu u(x_i, \lambda_i)F_{\lambda_i}(x_i)\dot{\lambda}\delta t\Big], \tag{A.12}$$

where $K_i \equiv -2\mu\Big[F_{\lambda_i}(x_i) + F_{\lambda_{i+1}}(x_{i+1}) + \dot{\lambda}(u(x_{i+1}, \lambda_{i+1}) - u(x_i, \lambda_i))\Big]$. Since we have used the Ito discretization, the following formula substitutes the normal rules of differential calculus [25],

$$g(x_{i+1}, \lambda_{i+1}) - g(x_i, \lambda_i) = \dot{\lambda}\frac{\partial g}{\partial \lambda}\delta t + D\frac{\partial^2 g}{\partial x^2}\delta t + \frac{\partial g}{\partial x}(x_{i+1} - x_i), \tag{A.13}$$

where $g(x, \lambda)$ is some continuous differentiable function. This formula is commonly referred to as the Ito formula [25]. When $D = 0$, we recover the normal rules of calculus. Using Eq. A.13, and $(x_{i+1} - x_i)^2 \to 2D\delta t$ as $\delta t \to 0$, repeatedly, we obtain the following result,

$$\frac{P_F(x_1, x_2, \ldots, x_N|x_0)}{P_R(x_{N-1}, \ldots, x_0|x_N)} = \exp\left[-\beta\Delta V + \beta\dot{\lambda}\sum_{i=0}^{N-1}\frac{\not\partial V_{\lambda_i}(x_i)}{\not\partial\lambda}\delta t\right], \tag{A.14}$$

where $\Delta V = V_{\lambda_N}(x_N) - V_{\lambda_0}(x_0)$,

$$\frac{\not\partial V_{\lambda_i}(x_i)}{\not\partial\lambda} = \left[\frac{\partial V_{\lambda_i}}{\partial\lambda}(x_i) + u(x_i, \lambda_i)\frac{\partial V_{\lambda_i}}{\partial x} - \beta^{-1}\frac{\partial u(x_i, \lambda_i)}{\partial x}\right], \tag{A.15}$$

and we have used $\mu/D = \beta$. Using Eq. A.14, we have

$$\begin{aligned}
&P_F(x_0, x_1, x_2, \ldots, x_N) = P_F(x_1, x_2, \ldots, x_N|x_0)\rho^{\mathrm{eq}}(x_0, \lambda_0)\\
=\quad & P_R(x_{N-1}, \ldots, x_0|x_N)\rho^{\mathrm{eq}}(x_N, \lambda_N)\exp\beta\left[\left(\dot{\lambda}\sum_{i=0}^{N-1}\frac{\not\partial V_{\lambda_i}(x_i)}{\not\partial\lambda}\delta t\right) - \Delta F\right]\\
=\quad & P_R(x_N, x_{N-1}, \ldots, x_0)\exp\beta\left[\left(\dot{\lambda}\sum_{i=0}^{N-1}\frac{\not\partial V_{\lambda_i}(x_i)}{\not\partial\lambda}\delta t\right) - \Delta F\right].
\end{aligned}$$

Finally, taking the limit $\delta t \to 0$ and writing the expression in the exponent above as an integral, we obtain the fluctuation theorem for a conjugate pair of escorted trajectories, $\{x_t\}$ and $\{x_{\tau-t}\}$,

$$\frac{P_F[\{x_t\}]}{P_R[\{x_{\tau-t}\}}] = \exp\beta\left[\int_0^\tau \dot{\lambda}\frac{\not\partial V_{\lambda_t}(x_t)}{\not\partial\lambda}\,dt - \Delta F\right]. \tag{A.16}$$

This completes the proof.

Appendix B

Figures of Merit

Here we derive a relation between N_s and C for the bidirectional Bennett estimator defined in Eq. 4.64. The Bennett estimator, Eq. 4.64 can be rewritten as a ratio of two free energy perturbation identities [84]

$$\frac{\langle P_H(W)/P_F(W)\rangle_{P_F(W)}}{\langle P_H(W)/P_R(-W)\rangle_{P_R(-W)}} = 1, \tag{B.1}$$

where $\langle \ldots \rangle_{P_F(W)}$ denotes an average over W values sampled from $P_F(W)$, $\langle \ldots \rangle_{P_R(-W)}$ denotes an average over W values sampled from $P_R(-W)$,

$$P_H(W) \equiv C^{-1} \frac{P_F(W) P_R(-W)}{P_F(W) + P_R(-W)}, \tag{B.2}$$

with

$$C = \int dW \, \frac{P_F(W) P_R(-W)}{P_F(W) + P_R(-W)}, \tag{B.3}$$

is the normalized harmonic mean distribution. As the averages in the numerator and the denominator are over different ensembles, let us separately consider the number of the realizations required for each to converge.

The dominant contributions to the average in the numerator come from work values that are typically sampled from the harmonic mean distribution P_H [51]. The probability that these dominant values are observed in the forward process can be given by

$$P = \int_{Typical} dW P_F(W) = \int_{Typical} dW P_H P_F(W)/P_H, \tag{B.4}$$

where $\int_{Typical}$ denotes that the integration is performed over the range of W values that are typically sampled from the harmonic mean distribution, $P_H(W)$.

Following Ref [51], we now write

$$P \sim \int_{Typical} dW P_H e^{\ln \frac{P_F}{P_H}} \sim e^{\langle \ln \frac{P_F}{P_H} \rangle_H} \int_{Typical} dW P_H \sim e^{\langle \ln \frac{P_F}{P_H} \rangle_H} \tag{B.5}$$

The number of realizations N_s required for adequate sampling can be roughly given by $N_s \sim P^{-1} \sim \exp D[P_H||P_F]$, where we have used $-\langle \ln \frac{P_F}{P_H} \rangle_H = D[P_H||P_F]$. The relative entropy $D[P_H||P_F]$ satisfies the following inequality

$$\begin{aligned} D[P_H||P_F] &= \int \frac{1}{C} \frac{P_R P_F}{P_R + P_F} \ln \frac{P_R}{C(P_F + P_R)} \\ &\leq \ln \int \frac{1}{4C^2} \frac{4P_R^2 P_F}{(P_R + P_F)^2} \\ &\leq \ln \frac{1}{4C^2} \int P_R \\ &= -2 \ln 2C \end{aligned} \tag{B.6}$$

where we have used the Jensen's inequality [15] for concave functions together with the identity $4P_F P_R \leq (P_F + P_R)^2$. Finally, using Eq. B.6, the number of realizations required to obtain a reliable estimate of ΔF using Bennett's method is bounded by

$$N_s \leq \frac{1}{C^2} \tag{B.7}$$

We have omitted numerical factors and constants in the above relation as it is already an approximate equation.

Bibliography

[1] A. B. Adib and C. Jarzynski. Unbiased estimators for spatial distribution functions of classical fluids. *J. Chem. Phys*, 122(1):014114, 2005.

[2] D. Andrieux and P. Gaspard. Fluctuation theorem for currents and schnakenberg network theory. *J. Stat. Phys.*, 127:107–131, 2007. 10.1007/s10955-006-9233-5.

[3] M. Athenes. Computation of a chemical potential using a residence weight algorithm. *Phys. Rev. E*, 66:046705, 2002.

[4] M. Athenes. A path-sampling scheme for computing thermodynamic properties of a many-body system in a generalized ensemble. *Eur. Phys. J. B*, 38:651–663, 2004.

[5] E. Atilgan and S. X. Sun. Equilibrium free energy estimates based on nonequilibrium work relations and extended dynamics. *J. Chem. Phys.*, 121(21):10392–10400, 2004.

[6] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys*, 22:245–268, 1976.

[7] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91(24):6269–6271, 1987.

[8] P. G. Bolhius, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: throwing ropes over dark mountain passes. *Ann. Rev. Phys. Chem*, 53:291, 2002.

[9] Carlos Bustamante, Jan Liphardt, and Felix Ritort. The nonequilibrium thermodynamics of small systems. *Physics Today*, 58:43–48, July 2005.

[10] D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, New York, 1987.

[11] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437:640–647, 2005.

[12] Y.-G. Chen and J. D. Weeks. Local molecular field theory for effective attractions between like charged objects in systems with strong coulomb interactions. *Proc. Natl. Acad. Sci. U.S.A*, 103(20):7560–7565, 2006.

[13] C. Chipot and A. Pohorille. *Free Energy Calculations*. Springer, Berlin, 2007.

[14] D. Collin, F. Ritort, C. Jarzynski, C. B. Smith, I. Tinoco Jr, and C. Bustamante. Verification of the crooks fluctuation theorem and recovery of rna folding free energies. *Nature*, 437:231–243, 2005.

[15] T. M. Cover. and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[16] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *J. Stat. Phys.*, 90(5/6):1481 – 1487, 1998.

[17] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721 – 2726, 1999.

[18] G.E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61:2361, 2000.

[19] G.E. Crooks and C. Jarzynski. Work distribution for the adiabatic compression of a dilute and interacting classical gas. *Phys. Rev. E*, 75:021116, 2007.

[20] C. Dellago, P. G. Bolhius, F. S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *J. Chem. Phys*, 108:1964, 1998.

[21] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Applied Mathematics, Montpelier, Vermont, 1982.

[22] E. H. Feng and G. E. Crooks. Length of time's arrow. *Phys. Rev. Lett.*, 101(9):090602, 2008.

[23] D. Frenkel and A. J. C. Ladd. New Monte Carlo method to compute the free energy of arbitrary solids. application to the fcc and hcp phases of hard spheres. *J. Chem. Phys*, 81:3188, 1984.

[24] D. Frenkel and B. Smit. *Understanding Molecular Simulation*. Academic Press, San Diego, 2nd edition, 2002.

[25] C.W. Gardiner. *Handbook of stochastic methods for physics, chemistry, and the natural sciences*. Springer series in synergetics. Springer-Verlag, 1985.

[26] Pierre Gaspard. Time-reversed dynamical entropy and irreversibility in markovian random processes. *J. Stat. Phys.*, 117(3/4):599 – 615, November 2004.

[27] H. Ge and D.-Q. Jiang. Generalized Jarzynski's equality of inhomogeneous multidimensional diffusion processes. *J. Stat. Phys.*, 131:675 – 689, 2008.

[28] P. L. Geissler and C. Dellago. Equilibrium time correlation functions from irreversible transformations in trajectory space. *J. Phys. Chem. B*, 108:6667 – 6672, 2004.

[29] J. W. Gibbs. *Elementary principles in statistical mechanics: Development with especial reference to the rational foundation of thermodynamics.* New York: C Scribner, 1902.

[30] R. Godawat, S. N. Jamadagni, and S. Garde. Characterizing hydrophobicity of interfaces by using cavity formation, solute binding, and water correlations. *Proc. Natl. Acad. Sci. U.S.A*, 106(36):15119–15124, 2009.

[31] N. Goldenfeld. *Lectures on phase transitions and the renormalization group.* Frontiers in physics. Addison-Wesley, Advanced Book Program, 1992.

[32] H. Goldstein, C.P. Poole, and J.L. Safko. *Classical mechanics.* Addison Wesley, 2002.

[33] J. Gore, F. Ritort, and C. Bustamante. Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proc. Natl. Acad. Sci. U.S.A*, 100:12564, 2003.

[34] A. M. Hahn and H. Then. Using bijective maps to improve free-energy estimates. *Phys. Rev. E*, 79:011113, 2009.

[35] A. M. Hahn and H. Then. Measuring the convergence of Monte Carlo free-energy calculations. *Phys. Rev. E*, 81(4):041117, 2010.

[36] D.A. Hendrix and C. Jarzynski. A "fast growth" method of computing free energy differences. *J. Chem. Phys.*, 114(14):5974–5981, 2001.

[37] J. Hermans. Simple analysis of noise and hysteresis in (slow-growth) free energy simulations. *J. Phys. Chem.*, 95:9029–9032, 1991.

[38] J. M. Horowitz and S. Vaikuntanathan. Nonequilibrium detailed fluctuation theorem for repeated discrete feedback. *Phys. Rev. E*, 82(6):061120, 2010.

[39] Z. Hu and J. D. Weeks. Efficient solutions of self-consistent mean field equations for dewetting and electrostatics in nonuniform liquids. *Phys. Rev. Lett.*, 105(14):140602, 2010.

[40] D. M. Huang, P. L. Geissler, and D. Chandler. Scaling of hydrophobic solvation free energies. *J. Phys. Chem B*, 105(28):6704–6709, 2001.

[41] G Hummer. *Free Energy Calculations*, Volume 86. Springer, Berlin, 2007.

[42] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci. U.S.A*, 98:3658, 2001.

[43] G. Hummer and A. Szabo. Free energy surfaces from single-molecule force spectroscopy. *Acc. Chem. Res.*, 38:504 – 513, 2005.

[44] J. E. Hunter III, W. P. Reinhardt, and T. F. Davis. A finite-time variational method for determining optimal paths and obtaining bounds on free energy changes from computer simulations. *J. Chem. Phys*, 99:6856 – 6864, 1993.

[45] A. Imparato and L. Peliti. Work-probability distribution in systems driven out of equilibrium. *Phys. Rev. E*, 72:046114, 2005.

[46] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opinion Struct. Biology*, 11:224 – 230, 2001.

[47] S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys. J*, 72:1568 – 1581, 1997.

[48] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56:5018, 1997.

[49] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690, 1997.

[50] C. Jarzynski. Targeted free energy perturbation. *Phys. Rev. E*, 65:046122, 2002.

[51] C. Jarzynski. Rare events and the convergence of exponentially averaged work values. *Phys. Rev. E*, 73:046105, 2006.

[52] C. Jarzynski. *Systems driven away from equilibrium.* http://outofeq2007.ihp.free.fr/ness-home.html, 2007.

[53] C. Jarzynski. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Ann. Rev. Cond. Matt. Phys.*, 2(329-351), 2011.

[54] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry.* Elsevier, New York, 2007.

[55] R. Kawai, J.M.R. Parrondo, and C. Van den Broeck. Dissipation: The phase-space perspective. *Phys. Rev. Lett.*, 98:080602, 2007.

[56] K. H. Kim and H. Qian. Fluctuation theorems for a molecular refrigerator. *Phys. Rev. E*, 75(2):022102, Feb 2007.

[57] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys*, 3(300), 1935.

[58] D. A. Kofke. On the sampling requirements for exponential-work free-energy calculations. *Mol. Phys.*, 104:3701, 2006. And references therein.

[59] D. P. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics.* Cambridge University Press, Cambridge, 2000.

[60] L.D. Landau and E.M. Lifshitz. *Statistical Physics.* Pergamon Press, Oxford, 3rd edition, 1990.

[61] J. L. Lebowitz and H. Spohn. A gallavotti-cohen type symmetry in the large deviation functional for stochastic dynamics. *J. Stat. Phys*, 95:333–365, 1999.

[62] W. Lechner and C. Dellago. On the efficiency of path sampling methods for the calculation of free energies from non-equilibrium simulations. *J. Stat. Phys*, 2007(04):P04001, 2007.

[63] W. Lechner, H. Oberhofer, C. Dellago, and P.L. Geissler. Equilibrium free energies from fast-switching trajectories with large time steps. *J. Chem. Phys*, 124:044113, 2006.

[64] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science*, 296(5574):1832–1835, 2002.

[65] C. Maes. The fluctuation theorem as a gibbs property. *J. Stat. Phys.*, 95(1-2):367–392.

[66] C. Maes and K. Netočný. Time-reversal and entropy. *J. Stat. Phys.*, 110(1-2):269–310.

[67] M. A. Miller and W. P. Reinhardt. Efficient free energy calculations by variationally optimized metric scaling: Concepts and applications to the volume dependence of cluster free energies and to solid–solid phase transitions. *J. Chem. Phys*, 113(17):7035–7046, 2000.

[68] D. D. L. Minh. Density-dependent analysis of nonequilibrium paths improves free energy estimates. *J. Chem. Phys.*, 130:204102, 2009.

[69] D. D. L. Minh and A. B. Adib. Optimized free energies from bidirectional single-molecule force spectroscopy. *Phys. Rev. Lett.*, 100:180602, 2008.

[70] D. D. L. Minh and A. B. Adib. Path integral analysis of jarzynski's equality: Analytical results. *Phys. Rev. E*, 79:021122, 2009.

[71] D. D. L. Minh and J. D. Chodera. Optimal estimators and asymptotic variances for nonequilibrium path-ensemble averages. *J. Chem. Phys.*, 131(13):134110, 2009.

[72] 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. *J. Chem. Phys*, 134(2):024111, 2011.

[73] J. Mittal and G. Hummer. Static and dynamic correlations in water at hydrophobic interfaces. *Proc. Natl. Acad. Sci. U.S.A*, 105(51):20130–20135, 2008.

[74] H Oberhofer and C Dellago. Optimum bias for fast-switching free energy calculations. *Comput. Phys. Commun.*, 179:41–45, 2008.

[75] H. Oberhofer, C. Dellago, and S. Boresch. Single molecule pulling with large time steps. *Phys. Rev. E*, 75:061106, 2007.

[76] H. Oberhofer, C. Dellago, and P.L. Geissler. Biased sampling of nonequilibrium trajectories: Can fast switching simulations outperform conventional free energy calculation methods? *J. Phys. Chem B*, 109:6902, 2005.

[77] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Ann. Rev. Phys. Chem.*, 48(1):545–600, 1997.

[78] A. Z. Panagiotopoulous. Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble. *Mol. Phys.*, 61:813–826, 1987.

[79] V. S. Pande, A. Y.. Grosberg, and T. Tanaka. Heteropolymer freezing and design: Towards physical models of protein folding. *Rev. Mod. Phys.*, 72(1):259–314, 2000.

[80] S. Park and K. Schulten. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.*, 120:5946 – 5961, 2004.

[81] A. J. Patel, P. Varilly, and D. Chandler. Fluctuations of water near extended hydrophobic and hydrophilic surfaces. *The Journal of Physical Chemistry B*, 114(4):1632–1637, 2010. PMID: 20058869.

[82] D. A. Pearlman and P.A. Kollman. The lag between the hamiltonian and the system configuration in free energy perturbation calculations. *J. Chem. Phys*, 91:7831, 1989.

[83] M. Ponmurugan. Generalized detailed fluctuation theorem under nonequilibrium feedback control. *Phys. Rev. E*, 82(3):031129, 2010.

[84] R. J. Radmer and P.A. Kollman. Free energy calculation methods: A theoretical and empirical comparison of numerical errors and a new method for qualitative estimates of free energy changes. *J. Comput. Chem*, 18:902, 1997.

[85] W. P. Reinhardt and J. E. Hunter III. Variational path optimization and upper and lower bounds to free energy changes via finite time minimization of external work. *J. Chem. Phys*, 97:1599 – 1601, 1992.

[86] J. M. Rodgers. *Statistical mechanical theory and simulations of charged fluids and water*. University of Maryland, College Park, 2008.

[87] J. M. Rodgers and J. D. Weeks. Interplay of local hydrogen-bonding and long-ranged dipolar forces in simulations of confined water. *Proc. Natl. Acad. Sci. U.S.A*, 105(49):19136, 2008.

[88] T. Sagawa and M. Ueda. Generalized Jarzynski equality under nonequilibrium feedback control. *Phys. Rev. Lett.*, 104(9):090602, 2010.

[89] T. Schmiedl and U. Seifert. Optimal finite-time processes in stochastic thermodynamics. *Phys. Rev. Lett.*, 98:108301, 2007.

[90] Z. Schuss. *Theory and Applications of Stochastic Differential Equations*. Wiley, New York, 1980.

[91] U. Seifert. Stochastic thermodynamics: principles and perspectives. *Euro. Phys. J. B*. 64:423–431, 2008.

[92] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande. Equilibrium free energies from nonequilibrium measurements using maximum likelihood methods. *Phys. Rev. Lett.*, 91:140601, 2003.

[93] T. Speck and U. Seifert. Distribution of work in isothermal nonequilibrium processes. *Phys. Rev. E*, 70(6):066112, 2004.

[94] F. H. Stillinger. Structure in aqueous solutions of nonpolar solutes from the standpoint of scaled-particle theory. *J. Sol. Chem.*, 2:141–158, 1973. 10.1007/BF00651970.

[95] S. X. Sun. Equilibrium free energies from path sampling of non-equilibrium trajectories. *J. Chem. Phys*, 118:5759, 2003.

[96] M. E. Paulaitis T. L. Beck and L. R. Pratt. *The Potential Distribution Theorem and Models of Molecular Solutions*. Cambridge University Press, 2006.

[97] K. Takara, H.-H. Hasegawa, and D. J. Driebe. Generalization of the second law for a transition between nonequilibrium states. *Phys. Lett. A.*, 375:88–92, 2010.

[98] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187 – 199, 1977.

[99] S. Vaikuntanathan and C. Jarzynski. Dissipation and lag in irreversible processes. *EPL (Europhysics Letters)*, 87(6):60005 (6pp), 2009.

[100] M. Watanabe and W. P. Reinhardt. Direct dynamical calculation of entropy and free energy by adiabatic switching. *Phys. Rev. Lett.*, 65(26):3301–3304, 1990.

[101] J. D. Weeks. Connecting local structure to interface formation: A molecular scale van der Waals theory of nonuniform liquids. *Ann. Rev. Phys. Chem.*, 53(1):533–562, 2002.

[102] J. D. Weeks, K. Katsov, and K. Vollmayr. Roles of repulsive and attractive forces in determining the structure of nonuniform liquids: Generalized mean field theory. *Phys. Rev. Lett.*, 81(20):4400–4403, 1998.

[103] B. Widom. Some topics in the theory of fluids. *J. Chem. Phys*, 39:2808, 1963.

[104] R.H. Wood. Estimation of errors in free energy calculations due to the lag between the hamiltonian and the system configuration. *J. Phys. Chem*, 95:4838, 1991.

[105] D. Wu and D. A. Kofke. Rosenbluth sampled nonequilibrium work method for calculation of free energies in molecular simulation. *J. Chem. Phys.*, 122:204104, 2005.

[106] F. M. Ytreberg and D. M. Zuckerman. Single-ensemble nonequilibrium path-sampling estimates of free energy differences. *J. Chem. Phys*, 120:10876, 2004.

[107] D. M. Zuckerman and T. B. Woolf. Theory of a systematic computational error in free energy differences. *Phys. Rev. Lett.*, 89:180602, 2002.

[108] R. W. Zwanzig. High temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys*, 22:1420, 1954.