Enhanced Sampling
&
Free Energy Calculations

Methods Based on Equilibrium and Non-equilibrium Simulations

Ye Mei[1]
Pengfei Li[2]
Haohao Fu[3]

April 9, 2024

[1]samuel.y.mei@gmail.com
[2]lipengfei_mail@126.com
[3]fhh2626@mail.nankai.edu.cn

Dedicated to
Dr. Bernard R. Brooks and Dr. Gerhard König.

Contents

List of Figures

List of Tables

Preface

Should we type some words here? Maybe not, 'coz we are terse dudes.

About the companion website

The website[1] for this file contains:

- A link to (freely downloadable) latest version of this document.
- Link to some implementations of WHAM and MBAR.
- Other stuff might appear in the near future (HOPEFULLY!).

Acknowledgements

- YM wants to express his special thanks to Dr. Bernard Brooks[2] and Dr. Gerhard König for helping him toddle in this field.
- We'll also like to thank Dr. Xiangyu Jia[3], Dr. Meiting Wang, Dr. Wei Liu and Dr. Fengjiao Liu for many helpful discussions.

Ye Mei
State Key Laboratory of Precision Spectroscopy
East China Normal University
Shanghai 200062 China
https://qclassic.wordpress.com/

[1] https://github.com/samuelymei/

[2] https://www.lobos.nih.gov/cbs/

[3] https://research.shanghai.nyu.edu/centers-and-institutes/chemistry/people/xiangyu-jia

1

Introduction

"Everything should be made as simple as possible but not simpler."
– Albert Einstein,

Computer simulations of biological systems have made much progress in the past decades. A battery of methods at different levels of sophistication and complexity have been proposed.

However, we are still facing grand difficulties from three aspects, i.e. accuracy of Hamiltonians, efficiency of sampling and reliability of postprocessing methods.[1]

In this booklet, we will not cover the whole spectrum of methods for enhanced samplings and free energy calculations, but only summarize some basic ideas. More complicated implementations of these methods, for instance 2-dimensional replica exchange molecular dynamics simulations, will not be discussed.

Recently, there is one special issue focusing on the methodologies of free energy calculations on Journal of Chemical Theory and Computation (Free Energy Calculations: Three Decades of Adventure in Chemistry and Biophysics, Journal of Chemical Theory and Computation, Volume 10, Issue 7, 2014, `https://pubs.acs.org/toc/jctcce/10/7`). There is also a special issue on the recent development in enhanced sampling methods for molecular systems (Special Topic on Enhanced Sampling for Molecular Systems, Journal of Chemical Physics, Volume 149, Issue 7, 2018, `https://aip.scitation.org/toc/jcp/149/7`).

There are also some good papers for reference

- Andrew Pohorille, Christopher Jarzynski and Christophe Chipot, Good Practices in Free-Energy Calculations, Journal of Physical Chemistry B, 2010, 114 (32), 10235–10253

- Daniel M. Zuckerman, Equilibrium Sampling in Biomolecular Simulations, Annual Review of Biophysics, 2011, 40:41–62

- Jérôme Hénin, Tony Lelièvre, Michael R. Shirts, Omar Valsson, Lucie Delemotte, Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0], Living Journal of Computational Molecular Science, 2022, 4(1), 1583.

There are also two books on this topic you might be interested in:

- Free Energy Calculations: Theory and Applications in Chemistry and Biology, Editors: Christophe Chipot, Andrew Pohorille, ISBN 978-3-540-38448-9, Springer-Verlag Berlin Heidelberg, 2007
- Free Energy Computations: A Mathematical Perspective, Author: Tony Lelievre, Gabriel Stoltz, Mathias Rousset, ISBN-13: 978-1848162471, Imperial College Press, 2010

Before we move into the major content of this booklet, we would like to review some fundamentals that underlie the methods introduced in the following chapters. The first one is the canonical partition function Q for Hamiltonian $H(\mathbf{x}, \mathbf{p}_x)$, which is defined as

$$\begin{aligned} Q(N,V,T) =& \frac{1}{h^{3N}N!}\iint \exp[-\beta H(\mathbf{x},\mathbf{p}_x)]\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{p}_\mathbf{x} \\ =& \frac{1}{\Lambda^{3N}N!}Z(N,V,T), \end{aligned} \tag{1.0.0.1}$$

where $\mathbf{x}$ and $\mathbf{p}_x$ are the coordinates and the conjugate momenta, respectively,

$$Z(N,V,T) = \int \exp\left(-\beta U(\mathbf{x})\right)\mathrm{d}\mathbf{x} \tag{1.0.0.2}$$

is the configurational integral, Λ is the temperature-dependent de Broglie wavelength, and $U(\mathbf{x})$ is the potential energy.

The partition function Q can also be defined in energy space as

$$Q(N,V,T) = \int \exp\left(-\beta E\right)\Omega_{tot}(N,V,E)\,\mathrm{d}E, \tag{1.0.0.3}$$

where

$$\Omega_{tot}(N,V,E) = \frac{1}{h^{3N}N!}\iint_{V^N}\delta(H(\mathbf{x},\mathbf{p}_x)-E)\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{p}_\mathbf{x} \tag{1.0.0.4}$$

is the complete density of states. Correspondingly, we can also define the configurational density of states as

$$\Omega_{con}(E) \propto \frac{1}{N!}\int_{V^N}\delta(U(\mathbf{x})-E)\,\mathrm{d}\mathbf{x}. \tag{1.0.0.5}$$

The Helmholtz free energy is defined in terms of the canonical partition function as

$$A = -\beta^{-1}\ln Q(N,V,T), \tag{1.0.0.6}$$

which connects thermodynamics and statistical mechanics. If we can estimate the value of Q, we can calculate A. However, evaluating Q is very difficult or even impossible in most cases. Fortunately, we are only interested in the free energy differences, ΔA, between two systems or two states of a system denoted by 0 and 1, respectively

$$\Delta A = -\beta^{-1} \ln Q_1/Q_0. \tag{1.0.0.7}$$

For most cases we are dealing with, the masses of particles in systems 0 and 1 are the same, Eq. 1.0.0.7 can be rewritten in terms of the configurational integrals Z_0 and Z_1

$$\Delta A = -\beta^{-1} \ln Z_1/Z_0. \tag{1.0.0.8}$$

The systems 0 and 1 may differ in several ways as you will find in the following chapters. They may have different Hamiltonians, H_0 and H_1.

$$Q_0 = \frac{1}{N!h^{3N}} \int \exp\left[-\beta H_0(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_x \tag{1.0.0.9}$$

$$Q_1 = \frac{1}{N!h^{3N}} \int \exp\left[-\beta H_1(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_x \tag{1.0.0.10}$$

Or they may be characterized by different values of a macroscopic parameter, such as temperature.

$$Q_0 = \frac{1}{N!h^{3N}} \int \exp\left[-\beta_0 H(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_x \tag{1.0.0.11}$$

$$Q_1 = \frac{1}{N!h^{3N}} \int \exp\left[-\beta_1 H(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_x \tag{1.0.0.12}$$

Finally, they may correspond to different regions in the phase space accessible to the system

$$Q_0 = \frac{1}{N!h^{3N}} \int_{\Gamma_0} \exp\left[-\beta H(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_x \tag{1.0.0.13}$$

$$Q_1 = \frac{1}{N!h^{3N}} \int_{\Gamma_1} \exp\left[-\beta H(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_x \tag{1.0.0.14}$$

where Γ_0 and Γ_1 may refer to different conformations of a flexible molecules, or the bound and unbound structures of a protein-ligand complex, etc.

In canonical ensemble (with NVT fixed), the probability of a microstate is

$$\rho(\mathbf{x}) = \frac{1}{Z} \exp\left(-\beta U(\mathbf{x})\right), \tag{1.0.0.15}$$

where $U(\mathbf{x})$ is the potential energy of this microstate. With this probability we can calculate the expectation of any operator $\hat{O}$ on configurations via

$$\langle O \rangle = \frac{\int \hat{O}(\mathbf{x}) \exp\left(-\beta U(\mathbf{x})\right) \mathrm{d}\mathbf{x}}{Z}. \tag{1.0.0.16}$$

Besides state free energies, we may also be interested in free energy profiles along one or several degrees of freedom $\xi(\mathbf{x})$ known as collective variable (CV), reaction coordinate (RC), or order parameter (OP)

$$\begin{aligned} A(\xi) =& -\beta^{-1}\ln Z(\xi) \\ =& -\beta^{-1}\ln\int \exp\left(-\beta U\right)\delta(\xi-\xi(\mathbf{x}))\,\mathrm{d}\mathbf{x} \\ =& -\beta^{-1}\ln\int \exp\left(-\beta U(\xi(\mathbf{x})=\xi)\right)|\mathbf{J}|\,\mathrm{d}q_1\cdots\mathrm{d}q_{N-1}, \end{aligned} \tag{1.0.0.17}$$

where $\mathbf{J}$ is the Jacobian matrix upon changing from Cartesian to some generalized coordinates with its element defined as $[\mathbf{J}(\mathbf{q})]_{ij} = \partial x_i/\partial q_j$ with $q_N = \xi$. $|\mathbf{J}|$ is its determinant. Its gradient over ξ is

$$\begin{aligned} \frac{\partial A(\xi)}{\partial\xi} =& -\beta^{-1}\frac{\int \frac{\partial}{\partial\xi}\left(e^{-\beta U}|\mathbf{J}|\right)\mathrm{d}q_1\cdots\mathrm{d}q_{N-1}}{\int e^{-\beta U}\delta(\xi-\xi(\mathbf{x}))\,\mathrm{d}\mathbf{x}} \\ =& \frac{\int e^{-\beta U}\left[\frac{\partial U}{\partial\xi}-\beta^{-1}\frac{1}{|\mathbf{J}|}\frac{\partial|\mathbf{J}|}{\partial\xi}\right]|\mathbf{J}|\,\mathrm{d}q_1\cdots\mathrm{d}q_{N-1}}{\int e^{-\beta U}\delta(\xi-\xi(\mathbf{x}))\,\mathrm{d}\mathbf{x}} \\ =& \frac{\int e^{-\beta U}\left[\frac{\partial U}{\partial\xi}-\beta^{-1}\frac{\partial\ln|\mathbf{J}|}{\partial\xi}\right]|\mathbf{J}|\,\mathrm{d}q_1\cdots\mathrm{d}q_{N-1}}{\int e^{-\beta U}\delta(\xi-\xi(\mathbf{x}))\,\mathrm{d}\mathbf{x}} \\ =& \frac{\int e^{-\beta U}\left[\frac{\partial U}{\partial\xi}-\beta^{-1}\frac{\partial\ln|\mathbf{J}|}{\partial\xi}\right]\delta(\xi-\xi(\mathbf{x}))\,\mathrm{d}\mathbf{x}}{\int e^{-\beta U}\delta(\xi-\xi(\mathbf{x}))\,\mathrm{d}\mathbf{x}} \\ =& \left\langle\frac{\partial U}{\partial\xi}-\beta^{-1}\frac{\partial\ln|\mathbf{J}|}{\partial\xi}\right\rangle_{\xi}. \end{aligned} \tag{1.0.0.18}$$

Here, $-\frac{\partial U}{\partial\xi}+\beta^{-1}\frac{\partial\ln|\mathbf{J}|}{\partial\xi}$ is the generalized force on ξ to be averaged over the degrees of freedom other than ξ itself. Therefore, $A(\xi)$ is called the potential of mean force. *Note: Some define the potential of mean force as* $\left\langle\frac{\partial U}{\partial\xi}\right\rangle_{\xi}$ *only. But we do not strictly differentiate potential of mean force and free energy profile here.*

Correspondingly, the probability of the CV $\xi(\mathbf{x})$ having a value of ξ is

$$\begin{aligned} \rho(\xi) =& \frac{Z(\xi)}{Z} \\ =& \frac{\int\delta(\xi-\xi(\mathbf{x}))\exp\left(-\beta U\right)\mathrm{d}\mathbf{x}}{\int\exp\left(-\beta U\right)\mathrm{d}\mathbf{x}} \\ =& \left\langle\delta(\xi-\xi(\mathbf{x}))\right\rangle. \end{aligned} \tag{1.0.0.19}$$

Let us take protein-ligand binding

$$\mathrm{P} + \mathrm{L} \rightleftharpoons \mathrm{P\text{-}L}$$

as an example to illustrate how the simulations and free energy methods are used in real problems. The equilibrium constant, K_b, is defined as

$$K_b = \frac{[\text{P-L}]}{[\text{P}][\text{L}]}, \tag{1.0.0.20}$$

where [P-L], [P] and [L] are the equilibrium concentrations of the complex, protein and ligand, respectively. A standard binding free energy can be calculated via $\Delta G_{bind} \equiv -\beta^{-1} \ln \left[K_b C^0\right]$, where C^0 is the standard state concentration of 1 mol/liter ($\equiv 1/1661\text{Å}^{-3}$). K_b can be expressed in terms of a ratio of configurational integrals as

$$K_b = \frac{1}{[L]} \frac{N \int_{site} \mathrm{d}(\mathbf{1}) \int_{bulk} \mathrm{d}(\mathbf{2}) \cdots \int_{bulk} \mathrm{d}(\mathbf{N}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}{\int_{bulk} \mathrm{d}(\mathbf{1}) \int_{bulk} \mathrm{d}(\mathbf{2}) \cdots \int_{bulk} \mathrm{d}(\mathbf{N}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}, \tag{1.0.0.21}$$

where U is the total potential energy of the system, $(\mathbf{1})$, $(\mathbf{2})$, $\cdots$, $(\mathbf{N})$ and $\mathbf{X}$ are the coordinates of the N ligand molecules and the remaining atoms, respectively. For simplicity, we omit the integrals over the $(N-1)$ ligands in bulk, and notice that $\int_{bulk} \mathrm{d}(\mathbf{1}) = V_{bulk} \int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r_1} - \mathbf{r}^*)$. Then, we have

$$\begin{aligned} K_b =& \frac{1}{[L]} \frac{N \int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}{V_{bulk} \int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r_1} - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}} \\ =& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r_1} - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}. \end{aligned} \tag{1.0.0.22}$$

A direct calculation of this ratio is not easy. Practically, we can define a series of intermediate states. Thereupon, the calculation of this ratio can be facilitated by

$$\begin{aligned} K_b = \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}{Z_1} \times \frac{Z_1}{Z_2} \times \cdots \times \frac{Z_{n-1}}{Z_n} \times \\ \frac{Z_n}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r_1} - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}. \end{aligned} \tag{1.0.0.23}$$

There are two categories of methods for computing K_b, i.e. the alchemical strategy[2] and the PMF-based strategy[3]. A comparison of these two strategies can be found in Ref. [4] and [5]. For each ratio in Eq. 1.0.0.23, we shall design a proper simulation and employ a suitable free energy method to calculate the free energy associated with it. Enhanced sampling might be necessary for convergence.

In alchemical strategy, the ligand is decoupled from its environment in the binding pocket and then recouples in water. However, a series of steps with restraints on the conformation, translation and rotation are introduced.

Equation 1.0.0.23 is now realized as

$$
\begin{aligned}
K_b = & \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U_1}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c]}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c+U_t]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c+U_t]}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c+U_t+U_r]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c+U_t+U_r]}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_0+U_c+U_t+U_r]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_0+U_c+U_t+U_r]}}{\int_{bulk} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_0+U_c+U_t]}} \times \\
& \frac{\int_{bulk} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_0+U_c+U_t]}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_0+U_c]}} \times \\
& \frac{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_0+U_c]}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c]}} \times \\
& \frac{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U_1+U_c]}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta U_1}}.
\end{aligned}
\tag{1.0.0.24}
$$

In PMF-based strategy, the ligand is gradually pulled out of the binding pocket into water. Similarly, restraints on the conformation, translation and rotation should also be applied when pulling the ligand molecule. Equation 1.0.0.23 is now realized as

$$
\begin{aligned}
K_b = & \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c]}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c+U_o]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c+U_o]}}{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c+U_o+U_a]}} \times \\
& \frac{\int_{site} \mathrm{d}(\mathbf{1}) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c+U_o+U_a]}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c+U_o]}} \times \\
& \frac{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c+U_o]}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c]}} \times \\
& \frac{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta[U+U_c]}}{\int_{bulk} \mathrm{d}(\mathbf{1})\, \delta(\mathbf{r}_1 - \mathbf{r}^*) \int \mathrm{d}\mathbf{X}\, e^{-\beta U}}.
\end{aligned}
\tag{1.0.0.25}
$$

It is worth emphasizing that this PMF-based strategy does not yield the real entry/escaping pathways nor the free energy barriers.

2

Enhanced Sampling

"Keep the smart guys around you."

– Bernard R. Brooks

From the definition, the free energy of a specific system is dominated by phase space regions with a low potential energy (metastable states). However, these regions might be separated by high energy barriers ($\gg k_B T$). Transitions among these potential energy wells are often hindered by these barriers. According to the Boltzmann's Law, the probability of a sample $\mathbf{R}$ being visited is proportional to the Boltzmann's factor $\exp\left[-\beta E(\mathbf{R})\right]$, where $\beta = 1/k_B T$ is called the inverse temperature. k_B is the Boltzmann constant and T is the temperature. According to some experience, in a $100\,ns$ simulation, the system can overcome a barrier of $10\,k_B T$, which is $6\,kcal/mol$ at room temperature ($300\,K$). If the barrier is $1.5\,kcal/mol$ higher, it takes about $1\,\mu s$ (10 times longer) in average for the system to go over the barrier. If the barrier height reaches $9\,kcal/mol$, it takes $10\,\mu s$. And so on. As a good practice, convergence should be measured after each simulation.[6]

With modern computers, the longest all-atom molecular dynamics simulation for biological systems is probably the one done by D.E. Shaw, which was on a time scale of 1 ms on a special-purpose computer "Anton". For most classical molecular dynamics simulations, the time scales are normally several μs to tens of μs. For simulations using expensive Hamiltonians, such as in QM/MM simulations, the time scales that can be reached are usually three orders shorter. Clearly, molecular dynamics simulations are plagued by a timescale problem. In order to observe abundant transitions among these energy minima, which is required by free energy calculations, enhanced samplings are often indispensable. As shown in the Boltzmann's factor, the essential quantity that determines the rate of transitions is βE. In order to accelerate the phase space sampling, we can either increase the temperature or decrease the energy barrier. All the methods shown below can be classified into these two categories. Some recent review papers might

help.[7–10]

2.1 Replica Exchange Molecular Dynamics

2.1.1 Temperature-Replica Exchange Molecular Dynamics

Temperature replica exchange molecular dynamics (T-REMD) is one class of parallel tempering methods developed by Hansmann, Okamoto and Sugita[11–13] based on many ideas in a category of methods called *generalized-ensemble algorithm*. It is an extension of the well-known simulated annealing method. The basic idea of REMD is schematically summarized in Fig. 2.1. In REMD, the system is replicated into **M** *non-interacting* copies (replicas). Each replica is coupled to a bath at temperature T_m, $(m = 1, \ldots, M)$. At a certain time, the system is at state X, which can be denoted as $X = \left(x_1^{[i(1)]}, \ldots, x_M^{[i(M)]}\right) = \left(x_{m(1)}^{[1]}, \ldots, x_{m(M)}^{[M]}\right)$. Here, we used i and m to label the replica and the temperature respectively. Because the replicas are non-interacting, the weight-factor for a state X in this generalized ensemble is a direct product of the Boltzmann factors for each replica, i.e.

$$W_{REM}(X) = \prod_{m=1}^{M} \exp\left(-\beta_m H\left(q^{[i(m)]}, p^{[i(m)]}\right)\right) = \prod_{i=1}^{M} \exp\left(-\beta_{m(i)} H\left(q^{[i]}, p^{[i]}\right)\right) \tag{2.1.1.1}$$

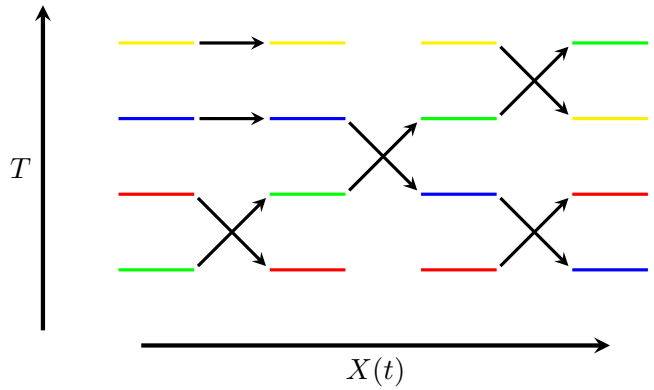

Figure 2.1: A schematic representation of replica exchange molecular dynamics.

Now, we exchange the temperatures of a pair of replicas

$$\begin{cases} x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m \Rightarrow x_n^{[i]'} \equiv \left(q^{[i]}, p^{[i]'}\right)_n \\ x_n^{[j]} \equiv \left(q^{[j]}, p^{[j]}\right)_n \Rightarrow x_m^{[j]'} \equiv \left(q^{[j]}, p^{[j]'}\right)_m \end{cases}, \tag{2.1.1.2}$$

where

$$\begin{cases} p^{[i]'} \equiv \sqrt{\frac{T_n}{T_m}} p^{[i]} \\ p^{[j]'} \equiv \sqrt{\frac{T_m}{T_n}} p^{[j]} \end{cases}. \tag{2.1.1.3}$$

The exchange rule is not trivial. In order for this exchange process to converge towards an equilibrium distribution, it is sufficient to impose the detailed balance condition on the transition probability $w(X \to X')$:

$$W_{REM}(X)w(X \to X') = W_{REM}(X')w(X' \to X). \tag{2.1.1.4}$$

Then we have

$$\begin{aligned}
&\frac{w\left(X \to X'\right)}{w\left(X' \to X\right)} \\
=&\frac{W_{REM}(X')}{W_{REM}(X)} \\
=&\frac{\exp\left(-\beta_m H\left(q^{[j]}, p^{[j]'}\right)\right)\exp\left(-\beta_n H\left(q^{[i]}, p^{[i]'}\right)\right)}{\exp\left(-\beta_m H\left(q^{[i]}, p^{[i]}\right)\right)\exp\left(-\beta_n H\left(q^{[j]}, p^{[j]}\right)\right)} \\
=&\frac{\exp\left\{-\beta_m\left[K\left(p^{[j]'}\right)+U\left(q^{[j]}\right)\right]-\beta_n\left[K\left(p^{[i]'}\right)+U\left(q^{[i]}\right)\right]\right\}}{\exp\left\{-\beta_m\left[K\left(p^{[i]}\right)+U\left(q^{[i]}\right)\right]-\beta_n\left[K\left(p^{[j]}\right)+U\left(q^{[j]}\right)\right]\right\}} \\
=&\frac{\exp\left\{-\beta_m\left[\frac{T_m}{T_n}K\left(p^{[j]}\right)+U\left(q^{[j]}\right)\right]-\beta_n\left[\frac{T_n}{T_m}K\left(p^{[i]}\right)+U\left(q^{[i]}\right)\right]\right\}}{\exp\left\{-\beta_m\left[K\left(p^{[i]}\right)+U\left(q^{[i]}\right)\right]-\beta_n\left[K\left(p^{[j]}\right)+U\left(q^{[j]}\right)\right]\right\}} \\
=&\frac{\exp\left\{-\beta_n K\left(p^{[j]}\right)-\beta_m K\left(p^{[i]}\right)\right\}}{\exp\left\{-\beta_m K\left(p^{[i]}\right)-\beta_n K\left(p^{[j]}\right)\right\}}\frac{\exp\left\{-\beta_m U\left(q^{[j]}\right)-\beta_n U\left(q^{[i]}\right)\right\}}{\exp\left\{-\beta_m U\left(q^{[i]}\right)-\beta_n U\left(q^{[j]}\right)\right\}} \\
=&\exp\{-\Delta\}.
\end{aligned} \tag{2.1.1.5}$$

where $\Delta = [\beta_n - \beta_m]\left[U\left(q^{[i]}\right) - U\left(q^{[j]}\right)\right]$. It can be seen that the kinetic energy terms are fully canceled out. This can be satisfied by the usual Metropolis criterion:

$$w\left(X \to X'\right) \equiv w\left(x_m^{[i]} \middle| x_n^{[j]}\right) = \begin{cases} 1, & \text{if } \Delta \le 0 \\ \exp\left(-\Delta\right), & \text{if } \Delta > 0 \end{cases} \tag{2.1.1.6}$$

The high-temperature replicas and the low-temperature replicas work in a collaborative way, in which the former explore phase space while the latter exploit phase space around local minima. After long time simulations, all the replicas have arrived at a global equilibrium. In order to calculate the free energy or the ensemble average of an operator $\hat{A}$ at T_m, we can extract all the snapshots that have a temperature T_m from M trajectories, if this temperature was among the M chosen temperatures. However, the optimal way is to use Weighted Histogram Analysis Method in Section 3.1.4 or the Multistate Bennett Acceptance Ratio method in Section 3.1.5.

In the above derivation, it only considers exchanges between neighboring states. However, a global permutation is also possible, and sometimes it may improve sampling efficiency.[14]

2.1.2 Hamiltonian-Replica Exchange Molecular Dynamics

Another type of REMD simulation is called Hamiltonian replica exchange molecular dynamics (H-REMD), in which each replicas has its own Hamiltonian, but is coupled to the same temperature.[15] One example is the H-REMD simulation for a torsional angle. The mth replica has a torsional energy term of

$$H_m(\phi) = \lambda(m) \sum_n (V_n/2) \left(1 + \cos\left[n\phi - \delta\right]\right), \tag{2.1.2.1}$$

where λ is a control parameter. $\lambda(0) = 1$ corresponds to the unbiased state and at $\lambda(M)$ (usually $\lambda(M) = 0$) the torsional motion of this dihedral angle has a smaller barrier.

Another example of HREMD is pH-REMD, in which each replica is coupled with different pH of the solution. In other words, the chemical potential of hydronium in each replica is different . Therefore, the protonation states (or probability of being protonated or deprotonated) of titratable residues in each replica may differ from those in other replicas. In the simulations, the protonation states of titratable residues have their protonation states alternated according to the Metropolis criterion

$$P = \begin{cases} 1, & \text{if } \Delta G_{\mathrm{P_A} \to \mathrm{P_A H^+}} \leq 0 \\ \exp\left(-\beta \Delta G_{\mathrm{P_A} \to \mathrm{P_A H^+}}\right), & \text{if } \Delta G_{\mathrm{P_A} \to \mathrm{P_A H^+}} > 0 \end{cases} \tag{2.1.2.2}$$

using Monte Carlo. The derivation of $\Delta G_{\mathrm{P_A} \to \mathrm{P_A H^+}}$ is shown below.

Free energy of molecule A in solution with a concentration [A] can be written as

$$\Delta G_{\mathrm{A}} = \Delta G^0_{\mathrm{A}} + \beta^{-1} \ln \frac{[\mathrm{A}]}{C_0},$$

in which ΔG^0_{A} is the free energy of molecule A at the standard state C_0, i.e. 1 mol/L. The free energy change for a reaction

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C}$$

can be written as

$$\Delta G = \Delta G_{\mathrm{C}} - \Delta G_{\mathrm{A}} - \Delta G_{\mathrm{B}} = \Delta G_0 + \beta^{-1} \ln \frac{[\mathrm{C}]\, C_0}{[\mathrm{A}]\, [\mathrm{B}]}.$$

At equilibrium, the free energy change is zero, we have

$$\Delta G_0 = -\beta^{-1} \ln \frac{[\mathrm{C}]\, C_0}{[\mathrm{A}]\, [\mathrm{B}]}, \tag{2.1.2.3}$$

in which $[\mathrm{A}]\,[\mathrm{B}] \,/\, [\mathrm{C}]\, C_0$ is called the dissociation constant K_a. So,

$$\Delta G_0 = \beta^{-1} \ln K_a. \tag{2.1.2.4}$$

Titration of a residue in a real protein can be written as

$$\mathrm{P_A} + \mathrm{H^+} \rightleftharpoons \mathrm{P_AH^+}$$

with

$$K_a = \frac{[\mathrm{P_A}]\,[\mathrm{H^+}]}{[\mathrm{P_AH^+}]\,C_0}$$

The fraction of the deprotonated species is calculated as

$$\begin{aligned} f_{[\mathrm{P_A}]} &= \frac{[\mathrm{P_A}]}{[\mathrm{P_A}] + [\mathrm{P_AH^+}]} \\ &= \frac{1}{1 + \frac{[\mathrm{P_AH^+}]}{[\mathrm{P_A}]}} \\ &= \frac{1}{1 + \frac{[\mathrm{P_A}][\mathrm{H^+}]}{C_0 K_a [\mathrm{P_A}]}} \\ &= \frac{1}{1 + \frac{1}{C_0 K_a}[\mathrm{H^+}]} \\ &= \frac{1}{1 + \frac{1}{K_a} 10^{-\mathrm{pH}}} \end{aligned} \tag{2.1.2.5}$$

We can check the asymptotic behavior of this equation. At strong acidic condition ($\mathrm{pH} = -\infty$), $f_{[\mathrm{P_A}]} = 0$, indicating that the residue is 100 percent protonated. While at an extremely basic condition ($\mathrm{pH} = \infty$), $f_{[\mathrm{P_A}]} = 1$. This residue is 100 percent deprotonated. From the Henderson–Hasselbalch (HH) equation, the pK_a can be determined by the pH of the state when $[\mathrm{P_A}] / [\mathrm{P_AH^+}] = 1$

$$\begin{aligned} \mathrm{p}K_a &= -\log K_a \\ &= -\log \frac{[\mathrm{P_A}]}{[\mathrm{P_AH^+}]} - \log \frac{[\mathrm{H^+}]}{C_0} \\ &= -\log \frac{[\mathrm{P_A}]}{[\mathrm{P_AH^+}]} + \mathrm{pH}. \end{aligned} \tag{2.1.2.6}$$

The pK_a of each residue in a dipeptide has been determined by experiment. However, when this residue is located in a certain protein, its pK_a is different from that in the dipeptide. The difference is called the pK_a shift. Instead of measuring the pK_a for a residue in a protein, we are more interested in calculating/measuring the titration curve, which is the fraction of the deprotonated state as a function of pH. From Eq. 2.1.2.5, $f_{[P_A]}$ can be easily calculated if we know K_a or equivalently the standard free energy change of protonation in Eq. 2.1.2.4. The standard free energy can be calculated from the partition functions as

$$\begin{aligned}\Delta G_0 &= -\beta^{-1}\ln\frac{Q_{\mathrm{P_AH^+}}}{Q_{\mathrm{P_A}}Q_{\mathrm{H^+}}}\\ &= -\beta^{-1}\ln\frac{\iint\exp(-\beta E_{\mathrm{P_AH^+}})\,\mathrm{d}\mathbf{R}_H\,\mathrm{d}\mathbf{R}_o}{Q_{\mathrm{H^+}}\int\exp(-\beta E_{\mathrm{P_A}})\,\mathrm{d}\mathbf{R}_o},\end{aligned}$$

where $\mathbf{R}_H$ is the coordinates of the specific H atom and the other degrees-of-freedom (DoF) are denoted as $\mathbf{R}_o$. Generally, the absolute value of ΔG_0 is hardly computable. A relative protonation free energy $\Delta\Delta G$ is preferred and is more reliable. Theoretically, the reference state can be any state you like. But the protonation free energy of the dipeptide is often used. The reference protonation process can be written as

$$\mathrm{A} + \mathrm{H^+} \rightleftharpoons \mathrm{AH^+}$$

The free energy change from the reference state is

$$\begin{aligned}&\Delta\Delta G_0\\ =&\Delta G_0 - \Delta G_0^{ref}\\ =&-\beta^{-1}\ln\frac{\iint\exp(-\beta E_{\mathrm{P_AH^+}})\,\mathrm{d}\mathbf{R}_H\,\mathrm{d}\mathbf{R}_o}{Q_{\mathrm{H^+}}\int\exp(-\beta E_{\mathrm{P_A}})\,\mathrm{d}\mathbf{R}_o}\frac{Q_{\mathrm{H^+}}\int\exp(-\beta E_{\mathrm{A}})\,\mathrm{d}\mathbf{R}_o}{\iint\exp(-\beta E_{\mathrm{AH^+}})\,\mathrm{d}\mathbf{R}_H\,\mathrm{d}\mathbf{R}_o}\\ =&-\beta^{-1}\ln\frac{\iint\exp(-\beta E_{\mathrm{P_AH^+}})\,\mathrm{d}\mathbf{R}_H\,\mathrm{d}\mathbf{R}_o\int\exp(-\beta E_{\mathrm{A}})\,\mathrm{d}\mathbf{R}_o}{\int\exp(-\beta E_{\mathrm{P_A}})\,\mathrm{d}\mathbf{R}_o\iint\exp(-\beta E_{\mathrm{AH^+}})\,\mathrm{d}\mathbf{R}_H\,\mathrm{d}\mathbf{R}_o}\\ =&-\beta^{-1}\ln\frac{\iint\exp\left[-\beta\left(E^{bond}_{\mathrm{P_AH^+}}+E^{QM}_{\mathrm{P_AH^+}}+E^{ele}_{\mathrm{P_AH^+}}\right)\right]\mathrm{d}\mathbf{R}_H\exp\left(-\beta E^{other}_{\mathrm{P_AH^+}}\right)\mathrm{d}\mathbf{R}_o}{\iint\exp\left[-\beta\left(\mathrm{E}^{bond}_{\mathrm{AH^+}}+E^{QM}_{\mathrm{AH^+}}+E^{ele}_{\mathrm{AH^+}}\right)\right]\mathrm{d}\mathbf{R}_H\exp\left(-\beta E^{other}_{\mathrm{AH^+}}\right)\mathrm{d}\mathbf{R}_o}\\ &\cdot\frac{\int\exp\left(-\beta E_{\mathrm{A}}\right)\mathrm{d}\mathbf{R}_o}{\int\exp\left(-\beta E_{\mathrm{P_A}}\right)\mathrm{d}\mathbf{R}_o},\end{aligned}\tag{2.1.2.7}$$

where E^{bond} and E^{ele} are the bonded energy and electrostatic interaction energy related to this H atom, respectively. E^{QM} is the energy correction that *may* be required if the molecular mechanical Hamiltonian cannot well capture the energy of the system, such as the missing of charge transfer effect. The sum of the remaining energy term is denoted as E^{other}, which does not explicitly depend on the position of this specific H atom. Eq. 2.1.2.7 is not ready to be computed before some approximations are adopted.

First, we assume that the total energy can be well described by the MM Hamiltonians for both the state interested in and the reference state. Therefore,

$$E^{QM}_{\mathrm{P_AH^+}} = E^{QM}_{\mathrm{AH^+}} = Const,$$

and they can be removed from the integral.

Second, the bonded terms involving hydrogen atoms are usually constrained in the simulations. Therefore, the hydrogen atom in question has only one position and $E^{bond} = 0$. Now, the relative protonation free energy can be simplified as

$$\begin{aligned}\Delta\Delta G_0 = &- \beta^{-1} \ln \frac{\int \exp\left(-\beta E^{ele}_{\mathrm{P_AH^+}}\right) \exp\left(-\beta E^{other}_{\mathrm{P_AH^+}}\right) \mathrm{d}\mathbf{R}_o}{\int \exp\left(-\beta E^{ele}_{\mathrm{AH^+}}\right) \exp\left(-\beta E^{other}_{\mathrm{AH^+}}\right) \mathrm{d}\mathbf{R}_o} \\ &\cdot \frac{\int \exp\left(-\beta E_{\mathrm{A}}\right) \mathrm{d}\mathbf{R}_o}{\int \exp\left(-\beta E_{P_{\mathrm{A}}}\right) \mathrm{d}\mathbf{R}_o}. \end{aligned} \tag{2.1.2.8}$$

Note that $E_{\mathrm{A}} = E^{other}_{\mathrm{AH^+}}$ and $E_{\mathrm{P_A}} = E^{other}_{\mathrm{P_AH^+}}$, we have

$$\Delta\Delta G_0 = - \beta^{-1} \ln \frac{\int \exp\left(-\beta E^{ele}_{\mathrm{P_AH^+}}\right) \exp\left(-\beta E_{\mathrm{P_A}}\right) \mathrm{d}\mathbf{R}_o}{\int \exp\left(-\beta E_{\mathrm{P_A}}\right) \mathrm{d}\mathbf{R}_o} \tag{2.1.2.9}$$

$$\cdot \frac{\int \exp\left(-\beta E_{\mathrm{A}}\right) \mathrm{d}\mathbf{R}_o}{\int \exp\left(-\beta E^{ele}_{\mathrm{AH^+}}\right) \exp\left(-\beta E_{\mathrm{A}}\right) \mathrm{d}\mathbf{R}_o} \tag{2.1.2.10}$$

$$\begin{aligned} =&- \beta^{-1} \ln \left\langle \exp\left(-\beta E^{ele}_{\mathrm{P_AH^+}}\right) \right\rangle_{\mathrm{P_A}} \\ &+ \beta^{-1} \ln \left\langle \exp\left(-\beta E^{ele}_{\mathrm{AH^+}}\right) \right\rangle_{\mathrm{A}} \\ =&\Delta G^{ele}_{\mathrm{P_AH^+}} - \Delta G^{ele}_{\mathrm{AH^+}} \end{aligned} \tag{2.1.2.11}$$

Therefore,

$$-\beta^{-1} \ln 10 \cdot \mathrm{p}K_a = \Delta G^{ele}_{\mathrm{P_AH^+}} - \Delta G^{ele}_{\mathrm{AH^+}} - \beta^{-1} \ln 10 \cdot \mathrm{p}K^{ref}_a.$$

Using Eq. 2.1.2.6, at a certain pH the free energy difference between the deprotonated and the protonated state can be written as

$$\Delta G_{\mathrm{P_A} \to \mathrm{P_AH^+}} = \Delta G^{ele}_{\mathrm{P_AH^+}} + \beta^{-1}(\mathrm{pH} - \mathrm{p}K^{ref}_a) \ln 10 - \Delta G^{ele}_{\mathrm{AH^+}}.$$

In the above equation, $\Delta G^{ele}_{\mathrm{AH^+}}$ can be obtained from a free energy calculation of the model system by alchemically annihilation of the proton. However, $\Delta G^{ele}_{\mathrm{P_AH^+}}$ is unknown. Approximately, it can be replaced with $\Delta H^{ele}_{\mathrm{P_AH^+}}$ averaged over a few snapshots.[16] In order to accelerate the convergence, this pH-REMD is often coupled with other enhanced simulation methods, such as T-REMD[16] and EDS-REMD[17] (see section 2.13).

2.2 Simulated Tempering

Simulated tempering (ST), aka serial tempering, was proposed by Marinari and Parisi[18] and by Lyubartsev et al[19] in 1992, and by Geyer and Thompson[20] in 1995. In ST, there is only one trajectory with controlled jumps in temperature space, which is different from the implementation of T-REMD2.1.1, in which multiple trajectories are running in parallel. In ST, the simulation is carried out in an extended space defined by the configuration variables $\mathbf{X}$ and a new variable m. The latter can take M values $(m = 1, \ldots, M)$. Corresponding to each m, there is an inverse temperature β_m. Let

$$\beta_1 > \beta_2 > \beta_3 > \cdots > \beta_M. \tag{2.2.0.1}$$

The probability distribution $\rho(\mathbf{X}, m)$ will be chosen to be

$$\rho(\mathbf{X}, m) \propto \exp\left[-H(\mathbf{X}, m)\right] \tag{2.2.0.2}$$

with

$$H(\mathbf{X}, m) \equiv \beta_m H(X) - g_m. \tag{2.2.0.3}$$

The total partition function for this extended ensemble is

$$\begin{aligned} Z &= \sum_{m=1}^{M} \int \mathrm{d}\mathbf{X} \, \exp\left[-H(\mathbf{X}, m)\right] \\ &= \sum_{m=1}^{M} \int \mathrm{d}\mathbf{X} \, \exp\left[-\beta_m H(X) + g_m\right] \\ &= \sum_{m=1}^{M} \exp\left(g_m\right) \int \mathrm{d}\mathbf{X} \, \exp\left[-\beta_m H(X)\right] \end{aligned} \tag{2.2.0.4}$$

For each β_m, there is a canonical ensemble with the probability for a configuration $\mathbf{X}$ follows the usual Boltzmann distribution, i.e.

$$\rho(X|m) \propto \exp\left(-\beta_m H(X)\right). \tag{2.2.0.5}$$

and the partition function (with $1/N!$ omitted)

$$Z_m = \int \mathrm{d}\mathbf{X} \, \exp\left[-\beta_m H(\mathbf{X})\right] = \exp\left(-\beta_m f_m\right), \tag{2.2.0.6}$$

where f_m is the associated free energy. With this definition of Z_m, the total partition function can be written as

$$Z = \sum_{m=1}^{M} \exp\left(g_m\right) Z_m, \tag{2.2.0.7}$$

where $\exp\left(g_m\right)$ can be thought as the weight for the mth canonical ensemble in this extended ensemble.

On the other hand, the probability of having a given value of m is simply given by

$$\rho_m = \frac{\int \mathrm{d}\mathbf{X} \, \exp\left[-H(\mathbf{X}, m)\right]}{Z} = \frac{Z_m \exp\left(g_m\right)}{Z} = \frac{1}{Z} \exp\left(-\beta_m f_m + g_m\right). \tag{2.2.0.8}$$

If we make the choice $g_m = \beta_m f_m$, then all the ρ_m become equal. However, f_m is usually unknown beforehand, or it may be never known.

The system may evolve in two types of steps: (1) usual displacements of particles at fixed temperature via molecule dynamics or Monte Carlo and (2) changes of reciprocal temperature with fixed positions of particles. In the first case, detailed balance can be easily maintained. In the second case, in order to maintain detailed balance

$$\rho(\mathbf{X}, m) P(\beta_m \to \beta_{m\pm1} | \mathbf{X}) = \rho(\mathbf{X}, m \pm 1) P(\beta_{m\pm1} \to \beta_m | \mathbf{X}), \tag{2.2.0.9}$$

transition takes place with the probability

$$P(\beta_m \to \beta_{m\pm1} | \mathbf{X}) = \min\left\{1, \exp\left[-(\beta_{m\pm1} - \beta_m) H(\mathbf{X}) + (g_{m\pm1} - g_m)\right]\right\} \tag{2.2.0.10}$$

following the Metropolis criteria.

2.3 Umbrella Sampling

Umbrella Sampling method was proposed by Torrie and Valleau in 1977,[21] and is still widely used nowadays. Suppose we are studying a transition process between two states such as conversion between two dominant conformations or a chemical reaction, and these two states are separated by a high barrier relative to k_BT. Therefore, the transition is a rare event. A schematic representation of the free energy landscape is shown in Fig. 2.2.

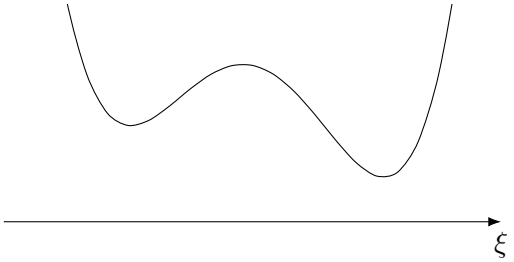

Figure 2.2: A typical free energy surface. Two free energy wells are separated by a barrier higher than k_BT.

Sometimes, we are interested in not only these two dominant states but also the whole pathway. Usually, we define a reaction coordinate ξ, including one or more collective variables (CVs), which can distinguish the two minima and characterize the free-energy pathway. Then we want to know the free-energy change along the reaction coordinate, ξ. It should be noted that choosing a suitable set of CVs is nontrivial for most cases.[22] A CV can be either a real coordinate such as the difference of bond lengths in, for example, an S_N2 reaction, or a thermodynamics coupling parameter (λ) that defines an unphysical path. If we run a simulation with the reaction coordinate set to a local maximum, i.e. the system being the transition state, the system will quickly roll back to the "reactant" or the "product" state in order to reduce the free energy. The consequence is that phase space outside the "reactant" and "product" regions cannot be sampled sufficiently to yield accurate free energy profile in a brute force simulation. In order to enhance the exploration in these regions, we can added a bias potential $\Delta V(\xi)$ into the system, guaranteeing

$$\forall \xi_1 \text{ and } \xi_2, [U(\xi_1) + \Delta V(\xi_1)] - [U(\xi_2) + \Delta V(\xi_2)] < k_BT \qquad (2.3.0.1)$$

then the free-energy surface can be explored within the timescale amenable to MD simulations. $\Delta V(\xi)$ is called the "umbrella potential".

However, as the free-energy surface is usually not known *a priori*, it is difficult to determine $\Delta V(\xi)$. To circumvent this issue, we can stratify the free-energy pathway into multiple *windows*, namely break up the reaction-coordinate space into "parts", by a series of (usually harmonic) restraints,

$\Delta U_i(\xi)$. In other words, the ith simulation, or the simulation of the ith window, is performed on the potential energy surface

$$U_i(\mathbf{R}) = U_0(\mathbf{R}) + \Delta U_i(\xi). \tag{2.3.0.2}$$

The strengths of the biases should be strong enough to maintain the system in the vicinity of where you are interested in, and also should be weak enough that the system can have significant overlap between two adjacent windows. At the same time, the windows should be small enough to guarantee in each window,

$$\forall \xi_1 \text{ and } \xi_2, [U(\xi_1) + \Delta U_i(\xi_1)] - [U(\xi_2) + \Delta U_i(\xi_2)] < k_B T \tag{2.3.0.3}$$

After all the simulations, the (biased) distribution of the samples in the whole region should be as flat as possible. Ensemble average under U_0 can be calculated from the ensembles generated under the biased Hamiltonians U via

$$\begin{aligned}
\langle X(\mathbf{R}) \rangle_0 =& \frac{\int X(\mathbf{R}) \exp\left[-\beta U_0(\mathbf{R})\right] \mathrm{d}\mathbf{R}}{\int \exp\left[-\beta U_0(\mathbf{R})\right] \mathrm{d}\mathbf{R}} \\
=& \frac{\int X(\mathbf{R}) \exp\left[\beta \Delta U_i(\mathbf{R})\right] \exp\left[-\beta U_i(\mathbf{R})\right] \mathrm{d}\mathbf{R}}{\int \exp\left[\beta \Delta U_i(\mathbf{R})\right] \exp\left[-\beta U_i(\mathbf{R})\right] \mathrm{d}\mathbf{R}} \\
=& \frac{\langle X \exp\left(\beta \Delta U_i\right) \rangle_i}{\langle \exp\left(\beta \Delta U_i\right) \rangle_i}.
\end{aligned} \tag{2.3.0.4}$$

Better postprocessing methods are the Weighted Histogram Analysis Method, Umbrella Integration and the Multistate Bennett Acceptance Ratio method (to be discussed in Section 3.1.4, 3.1.6 and 3.1.5).

2.4 Accelerated Molecular Dynamics

Accelerated molecular dynamics, or aMD for short, was proposed by Hamelberg et al in 2004,[23] based on the idea of hyperdynamics developed by Voter[24].

In this method, the simulation is performed on the modified potential $V^*(\mathbf{r})$

$$V^*(\mathbf{r}) = \begin{cases} V(\mathbf{r}), & V(\mathbf{r}) \geq E \\ V(\mathbf{r}) + \Delta V(\mathbf{r}), & V(\mathbf{r}) < E \end{cases} \tag{2.4.0.1}$$

in which E is a certain chosen energy, $\Delta V(\mathbf{r})$ is a continuous non-negative boost potential function, and $V(\mathbf{r})$ is the true potential. The bias potential increases the escape rate of the system from potential basins, and the subsequent state to state evolution of the system on the modified potential occurs at an accelerated rate with a nonlinear time scale of Δt_i^*, where

$$\Delta t_i^* = \Delta t_i e^{\beta \Delta V(\mathbf{r}(t_i))}. \tag{2.4.0.2}$$

The ensemble average value of any observable $A(\mathbf{r})$ taken on the modified potential can be written as

$$\begin{aligned} \langle A^* \rangle =& \frac{\int \mathrm{d}\mathbf{r}\, A(\mathbf{r}) e^{-\beta V^*(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\, e^{-\beta V^*(\mathbf{r})}} \\ =& \frac{\int \mathrm{d}\mathbf{r}\, A(\mathbf{r}) e^{-\beta V(\mathbf{r}) - \beta \Delta V(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\, e^{-\beta V(\mathbf{r}) - \beta \Delta V(\mathbf{r})}}. \end{aligned} \tag{2.4.0.3}$$

The correct ensemble average can be written as

$$\begin{aligned} \langle A \rangle =& \frac{\int \mathrm{d}\mathbf{r}\, A(\mathbf{r}) e^{-\beta V(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\, e^{-\beta V(\mathbf{r})}} \\ =& \frac{\int \mathrm{d}\mathbf{r}\, A(\mathbf{r}) e^{-\beta V^*(\mathbf{r}) + \beta \Delta V(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\, e^{-\beta V^*(\mathbf{r}) + \beta \Delta V(\mathbf{r})}} \\ =& \frac{\left\langle A(\mathbf{r}) e^{\beta \Delta V(\mathbf{r})} \right\rangle_{V^*}}{\left\langle e^{\beta \Delta V(\mathbf{r})} \right\rangle_{V^*}}. \end{aligned} \tag{2.4.0.4}$$

Some technique can improve the numerical stability of this reweighting process.[25] The definition of $\Delta V(\mathbf{r})$ is non-unique, and Hamelberg et al proposed a modification of the potential energy surface more akin to snow drifts, which smooths the landscape by filling minima, but maintains the underlying shape of the unmodified potential energy surface and merges smoothly with the original potential at the threshold "boost energy" value E. It is defined as

$$\Delta V(\mathbf{r}) = \frac{(E - V(\mathbf{r}))^2}{\alpha + (E - V(\mathbf{r}))}, \tag{2.4.0.5}$$

where α is a tuning parameter that determines how deep the modified potential energy basin is (when $E - V(\mathbf{r}) = \alpha, \Delta V(\mathbf{r}) = \alpha/2$). With this biasing potential, the derivative of $V^*(\mathbf{r})$ has no discontinuity, and the modified potential reproduces the shape of the minima even at high value of E. Furthermore, E should be carefully chosen, which may require a short trial simulation.

The most recent variant of aMD, the Gaussian accelerated molecular dynamics (GaMD), was developed by Miao et al[26], in which the biasing potential is defined as

$$\Delta V(\mathbf{r}) = \begin{cases} \frac{1}{2}k(E - V(\mathbf{r}))^2, & V(\mathbf{r}) < E \\ 0, & V(\mathbf{r}) \geq E \end{cases} \tag{2.4.0.6}$$

In order to smoothen the potential energy surface for enhanced sampling, the boost potential needs to satisfy the following criteria. First, for any two arbitrary potential values $V(\mathbf{r}_1)$ and $V(\mathbf{r}_2)$ found on the original energy surface, ΔV should be a monotonic function that does no change the relative order of the biased potential values, i.e.

$$\text{sign}\left(V(\mathbf{r}_1) - V(\mathbf{r}_2)\right) = \text{sign}\left(V^*(\mathbf{r}_1) - V^*(\mathbf{r}_2)\right). \tag{2.4.0.7}$$

Without losing generality, let $V(\mathbf{r}_1) < V(\mathbf{r}_2)$, we have

$$V^*(\mathbf{r}_1) < V^*(\mathbf{r}_2), \tag{2.4.0.8}$$

which leads to

$$\begin{aligned} V(\mathbf{r}_1) + \frac{1}{2}k\left[E - V(\mathbf{r}_1)\right]^2 - V(\mathbf{r}_2) - \frac{1}{2}k\left[E - V(\mathbf{r}_2)\right]^2 &< 0 \\ \left[V(\mathbf{r}_1) - V(\mathbf{r}_2)\right] + \frac{1}{2}k\left[(2E - V(\mathbf{r}_1) - V(\mathbf{r}_2))(V(\mathbf{r}_2) - V(\mathbf{r}_1))\right] &< 0 \\ \left[V(\mathbf{r}_1) - V(\mathbf{r}_2)\right]\left[1 - \frac{1}{2}k(2E - V(\mathbf{r}_1) - V(\mathbf{r}_2))\right] &< 0. \end{aligned} \tag{2.4.0.9}$$

Since $V(\mathbf{r}_1) < V(\mathbf{r}_2)$, we have

$$1 - \frac{1}{2}k(2E - V(\mathbf{r}_1) - V(\mathbf{r}_2)) > 0, \tag{2.4.0.10}$$

or equivalently

$$\text{Criterion 1: } E < \frac{1}{k} + \frac{1}{2}\left[V(\mathbf{r}_1) + V(\mathbf{r}_2)\right]. \tag{2.4.0.11}$$

Second, if $V(\mathbf{r}_1) < V(\mathbf{r}_2)$, the potential difference observed on the smoothened energy surface should be smaller than that of the original; i.e., $V^*(\mathbf{r}_2) - V^*(\mathbf{r}_1) < V(\mathbf{r}_2) - V(\mathbf{r}_1)$, which leads to

$$\text{Criterion 2: } E > \frac{1}{2}\left[V(\mathbf{r}_1) + V(\mathbf{r}_2)\right]. \tag{2.4.0.12}$$

Therefore, the threshold energy must satisfy

$$V_{\max} \leq E \leq V_{\min} + \frac{1}{k}, \tag{2.4.0.13}$$

where $V_{\max}$ and $V_{\min}$ are the maximum and minimum of potential energies, and k has to satisfy

$$k \leq \frac{1}{V_{\max} - V_{\min}}. \tag{2.4.0.14}$$

It can be rewritten as

$$k = k_0 \frac{1}{V_{\max} - V_{\min}}, \tag{2.4.0.15}$$

where $k_0 \in (0, 1)$. k_0 determines the magnitude of the applied boost potential. With greater k_0, higher boost potential is added to the potential energy surface. The boost potential is

$$\Delta V(\mathbf{r}) = \frac{1}{2} k_0 \frac{1}{V_{\max} - V_{\min}} (E - V(\mathbf{r}))^2, \quad V(\mathbf{r}) < E. \tag{2.4.0.16}$$

Third, the standard deviation of ΔV needs to be small enough (i.e., narrow distribution) to ensure accurate reweighting using cumulant expansion to the second order:[25]

$$\sigma_{\Delta V} = \sqrt{\left(\frac{\partial \Delta V}{\partial V} \Big|_{V=V_{\text{av}}} \right)^2 \sigma_V^2} = k \left(E - V_{\text{av}} \right) \sigma_V \leq \sigma_0, \tag{2.4.0.17}$$

where V_{av} and σ_V are the average and standard deviation of the system potential energies, and $\sigma_{\Delta V}$ is the standard deviation of ΔV with σ_0 as a user-specified upper limit (e.g., $10 k_B T$) for accurate reweighting. If E is set to the lower bound $E = V_{\max}$, we have

$$k_0 \leq \frac{\sigma_0}{\sigma_V} \frac{V_{\max} - V_{\min}}{V_{\max} - V_{\text{av}}}. \tag{2.4.0.18}$$

For efficient enhanced sampling with the highest possible acceleration, k_0 can then be set to its upper bound as

$$k_0 = \min \left(1.0, \frac{\sigma_0}{\sigma_V} \frac{V_{\max} - V_{\min}}{V_{\max} - V_{\text{av}}} \right). \tag{2.4.0.19}$$

Alternatively, when the threshold energy E is set to its upper bound $E = V_{\min} + (1/k)$, we have

$$k_0 \geq \left(1 - \frac{\sigma_0}{\sigma_V} \right) \frac{V_{\max} - V_{\min}}{V_{\text{av}} - V_{\min}}. \tag{2.4.0.20}$$

Then, we have

$$\text{Criterion 3:} \quad \left(1 - \frac{\sigma_0}{\sigma_V} \right) \frac{V_{\max} - V_{\min}}{V_{\text{av}} - V_{\min}} \leq k_0 \leq \frac{\sigma_0}{\sigma_V} \frac{V_{\max} - V_{\min}}{V_{\max} - V_{\text{av}}}. \tag{2.4.0.21}$$

2.5 Adaptive Biasing Force Method

If the conditional gradient of the free energy with respect to a reaction coordinate (mean force) over the equilibrium distribution of the system *restricted* to the hypersurface where the reaction coordinate is constant can be computed, the free energy profile along this specific reaction coordinate can be readily obtained by thermodynamic integration. In the following, we shall follow the derivation by Ciccotti et al.[27] For a system under molecular constraints, $\sigma_j(x) = 0$, $j = 1, \ldots, M$, the probability density reads

$$\rho(x) = Z_\sigma^{-1} e^{-\beta V(x)} \prod_{j=1}^{M} \delta(\sigma_j(x)), \tag{2.5.0.1}$$

in which

$$Z_\sigma = \int e^{-\beta V(x)} \prod_{j=1}^{M} \delta(\sigma_j(x)) \, \mathrm{d}x \tag{2.5.0.2}$$

is the configuration integral. By definition, the free energy associated with the vectorial reaction coordinate $q(x) = (q_1(x), \ldots, q_N(x))$ is given by

$$F(z) \coloneqq -\beta^{-1} \ln \left[Z_\sigma^{-1} \int e^{-\beta V(x)} \prod_{k=1}^{N} \delta(q_k(x) - z_k) \prod_{j=1}^{M} \delta(\sigma_j(x)) \, \mathrm{d}x \right], \tag{2.5.0.3}$$

where $z = (z_1, \ldots, z_N)$. By differentiating both sides with respect to z_j, we find

$$\frac{\partial F(z)}{\partial z_j} = -\beta^{-1} e^{\beta F(z)} \cdot Z_\sigma^{-1} \int e^{-\beta V(x)} \frac{\partial}{\partial z_j} \prod_{k=1}^{N} \delta(q_k(x) - z_k) \cdot \prod_{j=1}^{M} \delta(\sigma_j(x)) \, \mathrm{d}x. \tag{2.5.0.4}$$

Please note that z_j is a number to which the reaction coordinate is to be constrained. Therefore, $V(x)$ is not a function of z_j.

Notice that

$$\begin{aligned}
&\frac{\partial}{\partial z_j} \prod_{k=1}^{N} \delta(q_k(x) - z_k) \cdot \prod_{j=1}^{M} \delta(\sigma_j(x)) \\
&= -\delta'(q_j(x) - z_j) \prod_{k \neq j} \delta(q_k(x) - z_k) \cdot \prod_{j=1}^{M} \delta(\sigma_j(x)) \\
&= -\left(b_j(x) \cdot \nabla \delta(q_j(x) - z_j)\right) \prod_{k \neq j} \delta(q_k(x) - z_k) \cdot \prod_{j=1}^{M} \delta(\sigma_j(x)) \\
&= -b_j(x) \cdot \nabla \left(\prod_{k=1}^{N} \delta(q_k(x) - z_k) \cdot \prod_{j=1}^{M} \delta(\sigma_j(x)) \right)
\end{aligned} \tag{2.5.0.5}$$

where $b_j(x), j = 1, \ldots, N$ are vector fields satisfying

$$b_j(x) \cdot \nabla \sigma_k(x) = 0, \quad \forall j = 1, \ldots, N,\ k = 1, \ldots, M \tag{2.5.0.6}$$

and

$$b_j(x) \cdot \nabla q_k(x) = \begin{cases} 1 & \text{if } j = k \\ 0 & otherwise \end{cases}. \tag{2.5.0.7}$$

Thereby,

$$\begin{aligned} &\frac{\partial F(z)}{\partial z_j} \\ &= -\beta^{-1} e^{\beta F(z)} \cdot Z_\sigma^{-1} \int e^{-\beta V(x)} b_j(x) \nabla \left(\prod_{k=1}^{N} \delta(q_k(x) - z_k) \prod_{j=1}^{M} \delta(\sigma_j(x)) \right) \mathrm{d}x \\ &= e^{\beta F(z)} \cdot Z_\sigma^{-1} \int e^{-\beta V(x)} \Big(b_j(x) \cdot \nabla V(x) - \beta^{-1} \nabla \cdot b_j(x) \Big) \\ &\cdot \prod_{k=1}^{N} \delta(q_k(x) - z_k) \prod_{j=1}^{M} \delta(\sigma_j(x)) \, \mathrm{d}x \end{aligned} \tag{2.5.0.8}$$

after integration by parts. After rearrangement, the gradient of $F(z)$ (i.e. the mean force) can be expressed as

$$\frac{\partial F}{\partial z_j} = \Big\langle b_j(x) \cdot \nabla V - \beta^{-1} \nabla \cdot b_j(x) \Big\rangle_{q(x)=z, \sigma(x)=0}, \tag{2.5.0.9}$$

where $\langle \cdot \rangle_{q(x)=z,\sigma(x)=0}$ denotes the conditional average under the constraints $q(x) = z$, $\sigma(x) = 0$. For any function $f(x)$

$$\langle f \rangle_{q(x)=z,\sigma(x)=0} = \frac{\int f(x) e^{-\beta V(x)} \prod_{k=1}^{M} \delta(q_k(x) - z_k) \prod_{j=1}^{M} \delta(\sigma_j(x)) \, \mathrm{d}x}{\int e^{-\beta V(x)} \prod_{k=1}^{M} \delta(q_k(x) - z_k) \prod_{j=1}^{M} \delta(\sigma_j(x)) \, \mathrm{d}x}. \tag{2.5.0.10}$$

Being "restricted" here is different from being "constrained". In the latter, there is an additional condition that the velocity of this reaction coordinate must be set to zero. In the standard Blue Moon sampling method developed by Carter et al.[28], constrained molecular dynamics is utilized to compute the conditional expectation in Eq. 2.5.0.10. However, it introduces additional constraints on the momenta, which has to be removed. Therefore, computing the mean force from a constrained ensemble, a correction factor (denoted as $|Z|^{-1/2}$ in Ref. [28]) must be introduced, which arises from performing the momentum integration in the ensemble average. In addition, constrained simulation may cause quasinonergodic effect, in particular when

multiple reaction pathways are present. Therefore, constrained simulation is not recommended.

Alternatively, adaptive biasing force (ABF) method, which was proposed by Darve and Pohorille in 2001[29] and reformulated in 2008[30], can be used for the calculations of free energy profiles. It applies to unconstrained simulations, as well as constrained simulations. In ABF, an external force, $-\left\langle F_\xi|_{\xi^*}\right\rangle \nabla\xi$, that counteracts the mean force is applied. The net result of this procedure is that, after a brief equilibrium, the average force acting on ξ is close to zero and the system undergoes barrierless diffusionlike motion along the order parameter. This means that the sampling of ξ becomes uniform.

We denote by $\boldsymbol{\xi}$ the vector of all order parameters ξ_i, $i = 1, \ldots, N_\xi$. The free energy $A(\boldsymbol{\xi})$ is defined as

$$A(\boldsymbol{\xi}) = -\ln \int e^{-H(\mathbf{x})} \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\, \mathrm{d}\mathbf{x}. \tag{2.5.0.11}$$

β has been absorbed. Now define a thin matrix $\mathbf{W}$ with N_ξ columns, which satisfies

$$\mathbf{J}_{\boldsymbol{\xi}}\mathbf{W} = \mathbf{I}, \tag{2.5.0.12}$$

where the Jacobian $\mathbf{J}_{\boldsymbol{\xi}}$ is a fat matrix with its element defined by

$$[\mathbf{J}_{\boldsymbol{\xi}}]_{ij} = \frac{\partial \xi_i}{\partial x_j}, \tag{2.5.0.13}$$

and $\mathbf{I}$ is a unit matrix. Using the definition of ensemble average and integration by parts, we find

$$\begin{aligned}
\left\langle \mathbf{W}^t\nabla U - (\nabla\cdot\mathbf{W})^t\Big|_{\boldsymbol{\xi}} \right\rangle =& \frac{\int \left(\mathbf{W}^t\nabla U - (\nabla\cdot\mathbf{W})^t\right) e^{-U(\mathbf{x})} \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\, \mathrm{d}\mathbf{x}}{\int e^{-U(\mathbf{x})} \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\, \mathrm{d}\mathbf{x}} \\
=& \frac{-\int (\nabla\cdot(e^{-U(\mathbf{x})}\mathbf{W}))^t \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\, \mathrm{d}\mathbf{x}}{\int e^{-U(\mathbf{x})} \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\, \mathrm{d}\mathbf{x}} \\
=& \frac{\int e^{-U(\mathbf{x})}\mathbf{W}^t\nabla \left(\prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\right) \mathrm{d}\mathbf{x}}{\int e^{-U(\mathbf{x})} \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x}))\, \mathrm{d}\mathbf{x}},
\end{aligned} \tag{2.5.0.14}$$

in which t is the transpose of a vector or matrix.

Let us choose an index i ($1 \leq i \leq N_\xi$) and focus on $\partial A/\partial \xi_i$. Only row i of $\mathbf{W}^t$, w_i, needs to be considered. The gradient can be computed as

$$\nabla \left(\prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x})) \right) = \sum_{k=1}^{N_\xi} \delta'(\xi_k(\mathbf{x}) - \xi_k) \prod_{j \neq k} \delta(\xi_j - \xi_j(\mathbf{x})) \nabla \xi_k. \quad (2.5.0.15)$$

Since we have $\nabla \xi_k w_i = \delta_{ik}$,

$$w_i \cdot \nabla \left(\prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x})) \right) = \delta'(\xi_i(\mathbf{x}) - \xi_i) \prod_{j \neq i} \delta(\xi_j - \xi_j(\mathbf{x})). \quad (2.5.0.16)$$

Therefore, the ith component of $\left\langle \mathbf{W}^t \nabla U - (\nabla \cdot \mathbf{W})^t \big|_{\boldsymbol{\xi}} \right\rangle$ is

$$\frac{\int e^{-U} \delta'(\xi_i(\mathbf{x}) - \xi_i) \prod_{j \neq i} \delta(\xi_j - \xi_j(\mathbf{x})) \, \mathrm{d}\mathbf{x}}{\int e^{-U(\mathbf{x})} \prod_{j=1}^{N_\xi} \delta(\xi_j - \xi_j(\mathbf{x})) \, \mathrm{d}\mathbf{x}} = \frac{\partial A}{\partial \xi_i}, \quad (2.5.0.17)$$

where a property of δ function

$$\int f(x) \delta'(x) \, \mathrm{d}x = - \int f'(x) \delta(x) \, \mathrm{d}x \quad (2.5.0.18)$$

has been used. This proves

$$\nabla_{\boldsymbol{\xi}} A = \left\langle \mathbf{W}^t \nabla U - (\nabla \cdot \mathbf{W})^t \big|_{\boldsymbol{\xi}} \right\rangle. \quad (2.5.0.19)$$

This can be used in conjunction with the calculations of first and second spatial derivatives.

For multiple reaction coordinates, the calculation of $\nabla_\xi A$ can requires only first derivatives by observing that, with $\mathbf{J}(\mathbf{w})_{ij} = \frac{\partial w_i}{\partial x_j}$,

$$\begin{aligned} \left\langle \frac{\mathrm{d}}{\mathrm{d}t}(w_i \cdot \mathbf{p}) \Big|_{\boldsymbol{\xi}} \right\rangle &= \left\langle \mathbf{p}^t \mathbf{M}^{-1} \mathbf{J}(w_i)^t \mathbf{p} - w_i \cdot \nabla U \big|_{\boldsymbol{\xi}} \right\rangle \\ &= \left\langle -w_i \cdot \nabla U + \mathrm{Tr}(\mathbf{J}(w_i)) |_{\boldsymbol{\xi}} \right\rangle \\ &= - \left\langle w_i \cdot \nabla U - \nabla \cdot w_i |_{\boldsymbol{\xi}} \right\rangle \\ &= - \frac{\partial A}{\partial \xi_i}, \end{aligned} \quad (2.5.0.20)$$

where $\mathbf{p}$ is the momenta and $\mathbf{M}$ is the mass matrix. During the deviation, the equality

$$\int \mathbf{u}^t \mathbf{B} \mathbf{u} e^{-\mathbf{u}^t \mathbf{A} \mathbf{u}} \, \mathrm{d}\mathbf{u} = \frac{1}{2} \mathrm{Tr}\,(A^{-1} B) \int e^{-\mathbf{u}^t \mathbf{A} \mathbf{u}} \, \mathrm{d}\mathbf{u} \quad (2.5.0.21)$$

has been used with $\mathbf{u} = \mathbf{p}$, $\mathbf{B} = \mathbf{M}^{-1}\mathbf{J}(\mathbf{W})^t$, and $\mathbf{A} = \mathbf{M}^{-1}$.

For the choice $\mathbf{W}^t = \mathbf{M}_\xi \mathbf{J}_\xi \mathbf{M}^{-1}$, $\mathbf{M}_\xi^{-1} = \mathbf{J}_\xi \mathbf{M}^{-1} \mathbf{J}_\xi^t$, we get

$$\nabla_{\boldsymbol{\xi}} A = -\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{M}_{\boldsymbol{\xi}} \frac{\mathrm{d}\boldsymbol{\xi}}{\mathrm{d}t}\right)\Big|_{\boldsymbol{\xi}} \right\rangle . \tag{2.5.0.22}$$

This equation is much easier to implement numerically than Eq. 2.5.0.19. No second derivatives are involved. This is especially convenient since computing terms like $\partial \mathbf{M}_{\boldsymbol{\xi}} / \partial x_t$ can be quite tedious to implement.

2.6 λ-dynamics and extended-system dynamics

λ-dynamics was developed by Kong and Brooks in 1996.[31] In this method, the coupling parameter λ is treated as a pseudo particle with fictitious mass m_λ. The extended Hamiltonian for the system with a coupling parameter in one dimension can be written as

$$H(\mathbf{R}, \{\lambda_i, i = 1, \ldots, n\}) = H_{Rxn}(\mathbf{R}, \{\lambda_i\}) + \sum_{i=1}^{n} \frac{m_i}{2} \dot{\lambda}_i^2 + U^*(\{\lambda_i\}), \quad (2.6.0.1)$$

where H_{Rxn} is a legitimate mapping provided that $H_{Rxn}(\mathbf{R}, \lambda_i = 0)$ and $H_{Rxn}(\mathbf{R}, \lambda_i = 1)$ correspond to the Hamiltonians for the reactant and product states respectively, and $U^*(\lambda)$ is a restraint that limits the range of λ. The pseudo particles can be coupled to high temperature baths, so it can have strengthened ability to overcome the barrier. However, this might lead to energy transfer between the pseudo degrees of freedom to the configuration degrees of freedom. Therefore, the fictitious mass m_λ should be large enough to make this degree of freedom nearly adiabatic from the rest of the system.[32] λ-dynamics can also be coupled with metadynamics,[33] which will be introduced in Sec. 2.9.

In extended-system dynamics, which can be regarded as a "geometric" version of λ-dynamics, the extended Lagrangian is coupled with the usual Lagrangian. For the one-dimensional case,

$$L(\mathbf{R}) = L_0(\mathbf{R}) + \frac{m_\xi}{2} \dot{\xi}^2 + U^*(\xi) \quad (2.6.0.2)$$

Usually, pseudo springs are used to connect the extended and real CVs, namely

$$U^*(\xi) = \frac{1}{2} k (\xi - \xi_0(\mathbf{R}))^2 \quad (2.6.0.3)$$

where $\xi_0(\mathbf{R})$ is the real CV, ξ is the extended CV and $L_0(\mathbf{R})$ is the usual Lagrangian that drives the dynamics.

The method that makes the pseudo particles, namely ξ in the one-dimensional case, coupled to high-temperature baths, is called temperature accelerated molecular dynamics (TAMD).[34]

In principle, extended-system dynamics can be coupled with many enhanced-sampling algorithms. In such cases, the biases are added on the pseudo particles instead of the real system. The combination of extended-system dynamics and ABF, called extended ABF (eABF)[35], is practically useful, because i) ABF requires the second derivative of the collective variables to calculate the biasing forces, while in eABF, the biasing forces is directly obtained from the pseudo springs and ii) forces are vectors, implying in multidimensional case, biasing forces along different CVs may affect each other when they are not completely decoupled. While in eABF, the extended CVs are always independent.

For any extended-system-based enhanced-sampling algorithm, when the pseudo springs are hard enough, namely, k is sufficiently large for each collective variable, there is approximately

$$A(\xi_0) = A(\xi). \tag{2.6.0.4}$$

This approximation is obvious if the spring is regarded as a two-force member. To estimate the free-energy profile rigorously, an umbrella-integration (UI)[36] or corrected z-averaged restraint (CZAR)[37] estimator can be adopted. For the UI estimator, when the simulation reaches equilibrium, samples that satisfy

$$\xi \in [\xi_i, \xi_{i+1}) \tag{2.6.0.5}$$

namely, ξ of bin i are extracted. Then these samples can be regarded as those from an umbrella sampling simulation, with the restraining center at $\frac{1}{2}(\xi_i+\xi_{i+1})$. Hence, the umbrella integration method can be used to estimate the free-energy profile,

$$\frac{\partial A_i}{\partial \xi_0} = \frac{1}{\beta}\frac{\xi - \langle \xi_0 \rangle_\xi}{(\sigma(\xi_0)_\xi)^2} - k(\xi_0 - \xi) \tag{2.6.0.6}$$

For the CZAR estimator, the extended-system simulation is regarded as an adaptive umbrella-sampling one, and the umbrella potential comes from the spring. Hence,

$$\frac{\partial A}{\partial \xi_0} = \frac{1}{\beta}\frac{\mathrm{d} \ln p(\xi_0)}{\mathrm{d}\xi_0} + k(\langle \xi \rangle_{\xi_0} - \xi_0), \tag{2.6.0.7}$$

where $p(\xi_0)$ is the observed distribution of ξ_0.

2.7 Wang–Landau Algorithm

Wang–Landau algorithm was developed by Wang and Landau in 2001 to accelerate the convergence in calculating the density of states.[38] In conventional Monte Carlo simulation at a certain temperature T, the configurations are generated with a probability proportional to the product of the density of states $g(E)$ and the Boltzmann factor e^{-E/k_BT}. While Wang-Landau algorithm aims to estimate the density of states $g(E)$ via a random walk in energy space to produce a flat histogram. If a random walk in energy space is performed with a probability proportional to the reciprocal of the density of states $1/g(E)$, a flat histogram can be generated for the energy distribution. This is accomplished by simultaneously modifying the estimated density of states in a systematic way to produce a flat histogram over the allowed range of energy and making the density of states converge to the true value. Note that at the beginning of the random walk, the density of state is normally unknown, so we simply set them to one for all the energies, i.e. $g(E) = 1$. Then the random walk in energy space begins by changing the configuration, for instance flipping the spin in Ising model, randomly with a probability

$$p(E_1 \to E_2) = \min\left[\frac{g(E_1)}{g(E_2)}, 1\right], \tag{2.7.0.1}$$

where E_1 and E_2 are energies of the configurations before and after the change. Each time an energy level E is visited, the corresponding density of states is updated by multiplying the existing value by a modification factor $f > 1$, i.e., $g(E) \to g(E)f$. Initially, the modification factor f can be set to a value as large as $f_0 = e^1$, which leads to a crazy exploration in the energy space and the walker can quickly cover all energy levels. This random walk keeps on until we have a "flat" histogram $H(E)$. At this moment, the energy levels have been swept in a coarse manner and the density of states converges to the true value with an accuracy proportional to $\ln(f)$. Now, we reduce the modification factor to a finer one according to some recipe such as $f_1 = \sqrt{f_0}$ (any function that monotonically decrease to 1 will do) and reset the histogram $H(E) = 0$. Then we begin the next round of random walk with a finer modification factor f_1 until the histogram is flat again. This iteration continues with $f_{i+1} = \sqrt{f_i}$ until a pre-selected criterion such as $f_{final} < \exp(10^{-8}) = 1.00000001$ has been reached. In reality, a perfect "flatness" can never be reached. But we can define a "flat" histogram to be the condition that the histograms for all the E level is not less than 80% of the average histogram $\langle H(E)\rangle$. The flowchart of the algorithm is shown below.

This method can be further enhanced by performing, for instance, multiple random walks etc. Besides, the original implementation suffers from convergence difficulty, which originates from the decay scheme of f. Belar-

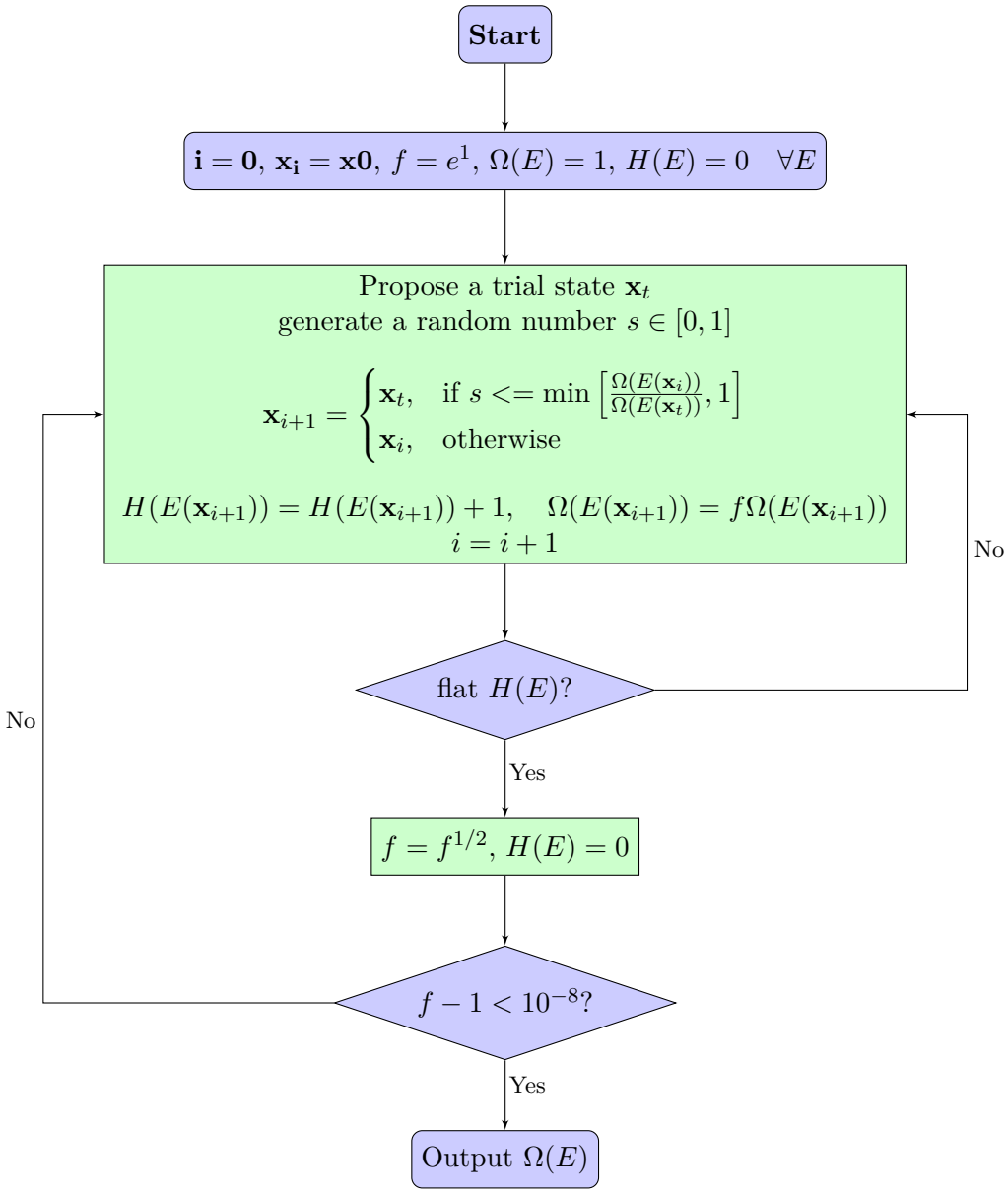

Figure 2.3: The Wang–Landau Algorithm

dinelli and Pereyra proposed a new scheme to solve the difficulty, in which $f = \exp(t^{-1})$.[39]

This algorithm was extended by Atchadé and Liu[40], and by Liang et al[41].

2.8 Accelerated Weight Histogram

The accelerated weight histogram (AWH) method is one of the extended ensemble or generalized ensemble methods, and was proposed by Lidmar in 2012[42] and extended by Lidmar and Hess later[43, 44]. In generalized ensemble methods, the parameters are tuned to ensure each state, corresponding a certain set of values of the parameters, is populated equally in an average sense. However, the best estimate of these parameters are usually unknown before hand. Similar to the Wang-Landau's idea, the accelerated weight histogram method designs an iterative way to update the parameters until a desired distribution of the extended ensemble has been reached.

Consider a model described by a probability distribution $\pi_\lambda(x)$, which depends on parameters λ, and x denotes the configuration of the system. The parameter λ can be either a scalar or a vector, and it can take a discrete set of preselected values $\lambda_m \in \mathcal{M} = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$. In an extended ensemble simulation, states are sampled according to a joint distribution $P(x, \lambda)$, which is expressed as

$$P(x, m) = \frac{1}{\mathcal{Z}} e^{f_m - E_m(x)}. \tag{2.8.0.1}$$

The weights e^{f_m} allow tuning the marginal distribution $P(m)$ to approach any desired form. For any fixed λ_m, we can generate samples, using molecular dynamics or Markov chain Monte Carlo, from the conditional distribution

$$P(x|m) \equiv \pi_m(x) = e^{F_m - E_m(x)}. \tag{2.8.0.2}$$

Usually, this can be done without knowledge of $F_m = -\ln \int e^{-E_m(x)}\, \mathrm{d}x$, the exact (dimensionless) free energy at λ_m.

Complementing the ordinary (MD or MC) moves, transitions in parameter space have to be carried out to make the marginal distribution

$$P(m) = \sum_x P(x, m) = \frac{1}{\mathcal{Z}} e^{f_m - F_m} \tag{2.8.0.3}$$

approximately flat. It indicates that $f_m \approx F_m$. However, F_m is unknown at the beginning of the simulation.

In the AWH algorithm, f_m is updated in an iterative way. The whole procedure can be summarized as follows:

(1) Perform N_x updates of the configurations x at fixed parameter value λ_m,

(2) Perform a parameter move $m \to m'$ using the Gibbs sampler using

$$w_{m'm}(x) = P(m'|x) = \frac{e^{-E_{m'}(x) - f_{m'}}}{\sum\limits_{k \in \mathcal{M}} e^{E_k(x) - f_k}}, \tag{2.8.0.4}$$

which is independent of m. Therefore, the subscript m can be omitted, and we denote it as $w_{m'}$.

(3) Update the weight histogram using

$$W_m \leftarrow W_m + w_m(x) \qquad \forall m. \tag{2.8.0.5}$$

Sample any observables of interest using

$$\langle A \rangle_m = \frac{\sum_t A(x_t, m) w_m(x_t)}{\sum_t w_m(x_t)}, \tag{2.8.0.6}$$

where $\{x_t\}$ denote the time series of visited configurations.

(4) Repeat steps 1-3 until N_I samples have been obtained.

(5) Update the free energy parameters f_m using

$$f_m \leftarrow f_m - \ln\left(\frac{W_m \mathcal{M}}{N}\right) \qquad \forall m \tag{2.8.0.7}$$

and the weight histogram using

$$W_k \leftarrow N/\mathcal{M}. \tag{2.8.0.8}$$

(6) Start a new iteration from step 1 unless the desired accuracy has been reached.

2.9 Metadynamics

Metadynamics, vividly called flooding method, was first suggested by Laio and Parrinello in 2002.[45] Imaging you were standing in a valley and were surrounded by high mountains. In most of the time, you were just wandering near the minimum, because your kinetic energy was not enough to climb the mountains. Suddenly, you realized that you could use metadynamics as a magic to escape from the minimum. You started walking. After each step, you took a bottle of sand out of your miraculous pocket and put the sand under your feet. Then you were lifted up inch-by-inch, and the deposited sand piles discourage you from revisiting where you had visited. And you were finally raised up to the top of the mountain and at that moment you were able to climb over that mountain without much effort and fell into another valley. The magic of sand continued, and at last you smoothed the whole area. Because you kept recording where you had put the sand and how much sand you had put there. You drew the shape the piled sand according to the record and you flipped it. In this way, you got the exact shape of the original free energy landscape up to a constant.

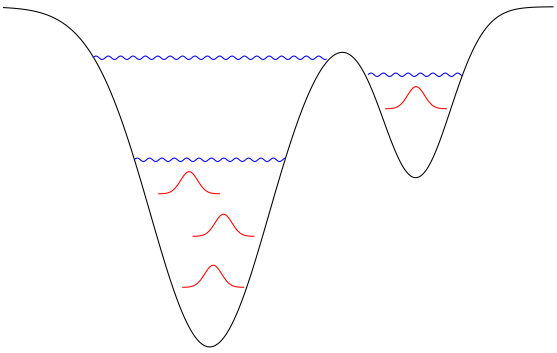

Figure 2.4: A schematic representation of metadynamics. The free energy well is gradually filled up with small Gaussian hills, and a transition is facilitated.

The above texts are merely an informal explanation of metadynamics. Formally, metadynamics belongs to a class of methods in which sampling is facilitated by introducing additional bias potential to pre-selected degrees of freedom, which are often referred as collective variables (CVs). In metadynamics, the bias potential added to the Hamiltonian of the system is history-dependent, and is often written as a sum of Gaussians deposited during the simulation as

$$V_G(\mathbf{s},t) = \int_0^t \mathrm{d}t'\,\omega \exp\left(-\sum_{i=1}^{d}\frac{[\mathbf{s}_i(R)-\mathbf{s}_i(R(t'))]^2}{2\sigma_i^2}\right) \tag{2.9.0.1}$$

on a collective variable $\mathbf{s}$ in d-dimension. σ and ω are two parameters tuning the shape of the Gaussians, which can be time-dependent. Asymptotically,

$$V_G(\mathbf{s}, t \to \infty) = -F(\mathbf{s}) + C. \tag{2.9.0.2}$$

Since the potential energy function for the dynamic variable $\mathbf{s}$ is history-dependent, the dynamics of $\mathbf{s}$ is non-Markovian. However, Bussi et al showed that the extended set of variables including $\mathbf{s}$ and the accumulated field is Markovian.[46]

However, metadynamics suffers from several practical difficulties. First of all, it is often difficult to decide when to stop a metadynamics simulation. Practically, the free energy does not converge. Instead, it fluctuates around the correct values, leading to an average error proportional to the square root of the bias potential deposition rate. Furthermore, the system may be pushed into regions of configurational space not physically relevant in a long run. Recent improvement over the ordinary metadynamics on convergence issue, which is termed well-tempered metadynamics (WTMetaD), can be found in Ref. [47] and is reviewed in Ref. [48]. Instead of completely filling the free energy well, WTMetaD tends to enhance the fluctuation of the CVs in a controllable manner. For instance, we can broaden the ensemble distribution by letting the biased probability distribution

$$P_V(\mathbf{s}) = \frac{[P(\mathbf{s})]^{1/\gamma}}{\int \mathrm{d}\mathbf{s}'[P(\mathbf{s}')]^{1/\gamma}}, \tag{2.9.0.3}$$

with $\gamma > 1$. Taking the definition of the unbiased probability distribution $P(\mathbf{s}) = \frac{\exp[-\beta F(\mathbf{s})]}{\int \exp[-\beta F(\mathbf{s}')]\,\mathrm{d}\mathbf{s}'}$ into the above equation, we find

$$\begin{aligned} P_V(\mathbf{s}) =& \frac{\exp\left[-\beta\frac{1}{\gamma}F(\mathbf{s})\right]}{\int \mathrm{d}\mathbf{s}' \exp\left[-\beta\frac{1}{\gamma}F(\mathbf{s}')\right]} \\ =& \frac{\exp\left\{-\beta\left[F(\mathbf{s}) - \left(1-\frac{1}{\gamma}\right)F(\mathbf{s})\right]\right\}}{\int \mathrm{d}\mathbf{s}' \exp\left\{-\beta\left[F(\mathbf{s}') - \left(1-\frac{1}{\gamma}\right)F(\mathbf{s}')\right]\right\}} \end{aligned} \tag{2.9.0.4}$$

where $-\left(1-\frac{1}{\gamma}\right)F(\mathbf{s})$ is the biasing potential $V(\mathbf{s})$. The limit $\gamma \to 1$ corresponds to the unbiased ensemble.

If $\gamma \to \infty$ limit, then

$$V(\mathbf{s}) = -F(\mathbf{s}) \tag{2.9.0.5}$$

and

$$P_V(\mathbf{s}) = \frac{1}{\int \mathrm{d}\mathbf{s}'} = const, \tag{2.9.0.6}$$

leading to a uniform distribution without any free energy barriers in the CV space, and the plain version of metadynamics is recovered.

Toward this end (Eq. 2.9.0.3), a gradually tempered Gaussian hill

$$V_n(\mathbf{s}) = V_{n-1}(\mathbf{s}) + G(\mathbf{s}, \mathbf{s}_n) \exp\left[-\frac{1}{\gamma - 1}\beta V_{n-1}(\mathbf{s}_n)\right] \tag{2.9.0.7}$$

is accumulated, where $V_0(\mathbf{s}) = 0$ and

$$G(\mathbf{s}, \mathbf{s}') = W \exp\left(-||\mathbf{s} - \mathbf{s}'||^2\right) \tag{2.9.0.8}$$

with $||\mathbf{s} - \mathbf{s}'||^2$ being a distance metric such as

$$||\mathbf{s} - \mathbf{s}'||^2 = \frac{1}{2}\sum_{i,j}(\mathbf{s}_i - \mathbf{s}'_i)\Sigma^{-1}_{i,j}(\mathbf{s}_j - \mathbf{s}'_j). \tag{2.9.0.9}$$

$\Sigma^{-1}_{i,j}$ is the inverse of the covariance matrix $\Sigma_{i,j}$, and the latter is normally diagonal $\Sigma_{i,j} = \delta_{i,j}\sigma_i^2$. W is the height of the Gaussian. $G(\mathbf{s}, \mathbf{s}_n)$ is the biasing kernel centered on the current CV value $\mathbf{s}_n$ and is scaled by $\exp\left[-\frac{1}{\gamma-1}\beta V_{n-1}(\mathbf{s}_n)\right]$ when being accumulated. The scaling factor itself decreases as $1/n$, therefore the change of the biasing potential becomes smaller as the metadynamics simulation progresses.[47, 49]

Practically, the update of the biasing potential is performed every N_G steps. Between any two adjacent updates, the system evolves under the action of the biasing potential $V_n(\mathbf{s}(\mathbf{R}))$. After the nth update, the biasing potential is

$$V(\mathbf{s}, t) = \sum_{k=1}^{n} W \exp\left(-||\mathbf{s} - \mathbf{s}_k||^2\right) \exp\left[-\frac{1}{\gamma - 1}\beta V_{k-1}(\mathbf{s}_k)\right]. \tag{2.9.0.10}$$

The factor $(\gamma - 1)\beta^{-1}$ is sometimes referred to as $k_B\Delta T$, or $\gamma = \frac{\Delta T + T}{T}$.

The remarkable feature of this stochastic update of the biasing potential is that the evolution of the bias can be described asymptotically by an ordinary differential equation (ODE)[49, 50]

$$\frac{\mathrm{d}V(\mathbf{s}, t)}{\mathrm{d}t} = \int \mathrm{d}\mathbf{s}'\, G(\mathbf{s}, \mathbf{s}') \exp\left[-\frac{1}{\gamma - 1}\beta V(\mathbf{s}', t)\right] P_V(\mathbf{s}', t), \tag{2.9.0.11}$$

where

$$P_V(\mathbf{s}, t) = \frac{e^{-\beta[F(\mathbf{s}) + V(\mathbf{s}, t)]}}{\int \mathrm{d}\mathbf{s}'\, e^{-\beta[F(\mathbf{s}') + V(\mathbf{s}', t)]}}. \tag{2.9.0.12}$$

For any $G(\mathbf{s}, \mathbf{s}')$, this ODE has the asymptotic solution

$$V(\mathbf{s}, t) = -\left(1 - \frac{1}{\gamma}\right) F(\mathbf{s}) + c(t), \tag{2.9.0.13}$$

where

$$c(t) = -\frac{1}{\beta}\log\frac{\int \mathrm{d}\mathbf{s}\, e^{-\beta[F(\mathbf{s})+V(\mathbf{s},t)]}}{\int \mathrm{d}\mathbf{s}\, e^{-\beta F(\mathbf{s})}} = -\frac{1}{\beta}\log\frac{\int \mathrm{d}\mathbf{s} P(\mathbf{s}) e^{-\beta V(\mathbf{s},t)}}{\int \mathrm{d}\mathbf{s} P(\mathbf{s})} \quad (2.9.0.14)$$

is independent of $\mathbf{s}$. Metadynamics thus converges to the desired result. Two interesting consequences arise.

First, it can be shown that, by taking the assumption of quasi-equilibrium, a time-dependent estimator for $F(\mathbf{s})$ is given by[51]

$$F(\mathbf{s}) = -\left(\frac{\gamma}{\gamma-1}\right)V(\mathbf{s},t) + \frac{1}{\beta}\log\int \mathrm{d}\mathbf{s}\, \exp\left[\frac{\gamma}{\gamma-1}\beta V(\mathbf{s},t)\right]. \quad (2.9.0.15)$$

Taking this equation into Eq. 2.9.0.14, one obtains

$$c(t) = \frac{1}{\beta}\log\frac{\int \mathrm{d}\mathbf{s}\, \exp\left[\frac{\gamma}{\gamma-1}\beta V(\mathbf{s},t)\right]}{\int \mathrm{d}\mathbf{s}\, \exp\left[\frac{1}{\gamma-1}\beta V(\mathbf{s},t)\right]}. \quad (2.9.0.16)$$

For a brief discussion on the calculations of $c(t)$, please refer to Ref. [52].

Second, it offers a practical way of calculating the expectation value of any $\mathbf{R}$-dependent function $O(\mathbf{R})$ as the simulation proceeds. The idea is that at time t the biased probability distribution for $\mathbf{R}$ is given by

$$\begin{aligned}
P_V(\mathbf{R},t) =& \frac{e^{-\beta[U(\mathbf{R})+V(\mathbf{s}(\mathbf{R}),t)]}}{\int \mathrm{d}\mathbf{R}'\, e^{-\beta[U(\mathbf{R}')+V(\mathbf{s}(\mathbf{R}'),t)]}} \\
=& \frac{e^{-\beta U(\mathbf{R})}e^{-\beta V(\mathbf{s}(\mathbf{R}),t)}\int e^{-\beta U(\mathbf{R}')}\,\mathrm{d}\mathbf{R}'}{\int e^{-\beta U(\mathbf{R}')}\,\mathrm{d}\mathbf{R}'\int e^{-\beta[U(\mathbf{R}')+V(\mathbf{s}(\mathbf{R}),t)]}\,\mathrm{d}\mathbf{R}'} \\
=& P(\mathbf{R})\frac{e^{-\beta V(\mathbf{s}(\mathbf{R}),t)}\iint e^{-\beta U(\mathbf{R}')}\delta(\mathbf{s}(\mathbf{R})-s)\,\mathrm{d}\mathbf{R}'\,\mathrm{d}\mathbf{s}}{\iint e^{-\beta[U(\mathbf{R}')+V(\mathbf{s}(\mathbf{R}),t)]}\delta(\mathbf{s}(\mathbf{R})-s)\,\mathrm{d}\mathbf{R}'\,\mathrm{d}\mathbf{s}} \\
=& P(\mathbf{R})\frac{e^{-\beta V(\mathbf{s}(\mathbf{R}),t)}\int e^{-\beta F(\mathbf{s})}\,\mathrm{d}\mathbf{s}}{\int e^{-\beta F(\mathbf{s})}e^{-\beta V(\mathbf{s},t)}\,\mathrm{d}\mathbf{s}} \\
=& P(\mathbf{R})e^{-\beta[V(\mathbf{s}(\mathbf{R}),t)-c(t)]}, \qquad (2.9.0.17)
\end{aligned}$$

where $P(\mathbf{R}) = \frac{e^{-\beta U(\mathbf{R})}}{\int e^{-\beta U(\mathbf{R}')}\,\mathrm{d}\mathbf{R}'}$ is the unbiased Boltzmann distribution, and $e^{\beta[V(\mathbf{s}(\mathbf{R}),t)-c(t)]}$ is the time-dependent unbiasing factor. Here, $e^{-\beta c(t)}$ serves as a normalizing factor. Straightforwardly, the average of $O(\mathbf{R})$ over the unbiased ensemble can be calculated from the metadynamics trajectory as

$$\langle O(\mathbf{R})\rangle = \left\langle O(\mathbf{R})e^{\beta[V(\mathbf{s}(\mathbf{R}),t)-c(t)]}\right\rangle_V. \quad (2.9.0.18)$$

This reweighting can be used to obtain the FES for some set of CVs $\mathbf{s}'$ either biased or unbiased by setting $O(\mathbf{R}) = \delta[\mathbf{s}' - \mathbf{s}'(\mathbf{R})]$. It is also useful

if one chooses $\mathbf{s}'$ as the biased degree of freedom $\mathbf{s}$ and obtain the FES. Disagreement between the FESs obtained directly from the bias potential (Eq. 2.9.0.15) and through reweighting (Eq. 2.9.0.18) is a clear sign that the metadynamics simulation has not converged.

Metadynamics has been implemented in PLUMED (`https://www.plumed.org/doc-v2.9/user-doc/html/index.html`), which can work with major molecular dynamics packages.

2.10 Variationally Enhanced Sampling Method

Variationally Enhanced Sampling (VES) method was developed by Valsson and Parrinello in 2014 an evolution of Metadynamic.[53] It begins with the following functional of a bias potential $V(\mathbf{s})$

$$\Omega[V] = \frac{1}{\beta} \ln \frac{\int \mathrm{d}\mathbf{s}\, e^{-\beta[F(\mathbf{s})+V(\mathbf{s})]}}{\int \mathrm{d}\mathbf{s}\, e^{-\beta F(\mathbf{s})}} + \int \mathrm{d}\mathbf{s}\, p(\mathbf{s}) V(\mathbf{s}), \tag{2.10.0.1}$$

where $p(\mathbf{s})$ is an arbitrary normalized probability distribution, and $F(\mathbf{s})$ is the unbiased free energy surface. This functional is convex and invariant under the addition of an arbitrary constant to $V(\mathbf{s})$, $\Omega[V + k] = \Omega[V]$. The potential that renders $\Omega[V]$ stationary is, with an constant,

$$V(\mathbf{s}) = -F(\mathbf{s}) - (1/\beta) \ln p(\mathbf{s}) \tag{2.10.0.2}$$

for $p(\mathbf{s}) \neq 0$ and $V(\mathbf{s}) = \infty$ otherwise. This stationary point is also the global minimum of $\Omega[V]$ since the functional is convex.

To make use of the variational property of $\Omega[V]$, the bias potential $V(\mathbf{s})$ is expanded as a function of a set of variational parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, and then the function $\Omega(\boldsymbol{\alpha}) = \Omega[V(\boldsymbol{\alpha})]$ is minimized with respect to $\boldsymbol{\alpha}$ until convergence is reached. With the converged potential $V(\mathbf{s}; \boldsymbol{\alpha})$, the free energy surface $F(\mathbf{s})$ can be estimated from Eq. 2.10.0.2.

The gradient $\Omega'(\boldsymbol{\alpha})$

$$\frac{\partial \Omega(\boldsymbol{\alpha})}{\partial \alpha_i} = -\left\langle \frac{\partial V(\mathbf{s};\boldsymbol{\alpha})}{\partial \alpha_i} \right\rangle_{V(\boldsymbol{\alpha})} + \left\langle \frac{\partial V(\mathbf{s};\boldsymbol{\alpha})}{\partial \alpha_i} \right\rangle_{p} \tag{2.10.0.3}$$

and the Hessian $\Omega''(\boldsymbol{\alpha})$

$$\begin{aligned} \frac{\partial^2 \Omega(\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_i} =& \beta \operatorname{Cov}\left[\frac{\partial V(\mathbf{s};\boldsymbol{\alpha})}{\partial \alpha_j}, \frac{\partial V(\mathbf{s};\boldsymbol{\alpha})}{\partial \alpha_i} \right]_{V(\boldsymbol{\alpha})} \\ & - \left\langle \frac{\partial^2 V(\mathbf{s};\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_i} \right\rangle_{V(\boldsymbol{\alpha})} + \left\langle \frac{\partial^2 V(\mathbf{s};\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_i} \right\rangle_{p} \end{aligned} \tag{2.10.0.4}$$

where $\langle \cdots \rangle_{V(\boldsymbol{\alpha})}$ and $\operatorname{Cov}[\cdots]_{V(\boldsymbol{\alpha})}$ are the expectation value and the covariance, respectively, obtained in a biased simulation employing the potential $V(\mathbf{s}; \boldsymbol{\alpha})$, and $\langle \cdots \rangle_p$ is an expectation value in the distribution $p(\mathbf{s})$. A natural approach is to expand $V(\mathbf{s}; \boldsymbol{\alpha})$ in a linear basis set and use the coefficient of this expansion as variational parameters,

$$V(\mathbf{s}; \boldsymbol{\alpha}) = \sum_k \alpha_k G_k(\mathbf{s}). \tag{2.10.0.5}$$

In this case the gradient and the Hessian simplify,

$$\frac{\partial \Omega(\boldsymbol{\alpha})}{\partial \alpha_{\mathrm{i}}} = - \left\langle G_i(\mathbf{s}) \right\rangle_{V(\boldsymbol{\alpha})} + \left\langle G_i(\mathbf{s}) \right\rangle_{p} , \tag{2.10.0.6}$$

$$\frac{\partial^2 \Omega(\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_i} = \beta \, \mathrm{Cov} \left[G_j(\mathbf{s}), G_i(\mathbf{s}) \right]_{V(\boldsymbol{\alpha})} . \tag{2.10.0.7}$$

2.11 On-the-fly Probability Enhanced Sampling

The On-the-fly Probability Enhanced Sampling (OPES) method was developed by Invernizzi and Parrinello in 2020.[54] Different from the well-tempered metadynamics (WTMetaD), in which the biasing potential is defined independent of the probability distribution, the core idea of OPES is to update self-consistently the estimate of the probability distributions and of the biasing potential in an on-the-fly fashion. The probability distribution is constructed via reweighting and is used to estimate the biasing potential.

In order to converge to the flat distribution, the biasing potential at the n^{th} step can be defined as

$$V_n(\mathbf{s}) = \frac{1}{\beta} \log \hat{P}_n(\mathbf{s}), \tag{2.11.0.1}$$

where $\hat{P}_n(\mathbf{s})$ is an estimate of the probability. However, in WTMetaD, we aim at reaching a target distribution $P^{tg}(\mathbf{s})$, which can be obtained if the bias is defined as

$$V(\mathbf{s}) = \frac{1}{\beta} \log \frac{P(\mathbf{s})}{P^{tg}(\mathbf{s})}. \tag{2.11.0.2}$$

In WTMetaD,

$$P^{tg}(\mathbf{s}) \propto [P(\mathbf{s})]^{1/\gamma}, \tag{2.11.0.3}$$

which leads to

$$V(\mathbf{s}) = (1 - 1/\gamma)\frac{1}{\beta} \log P(\mathbf{s}). \tag{2.11.0.4}$$

During the simulation, the estimated probability distribution is expressed as weighted sum of the Gaussian hills according to the previously deposited bias potential as

$$\tilde{P}_n(\mathbf{s}) = \frac{\sum_k^n w_k G(\mathbf{s}, \mathbf{s}_k)}{\sum_k^n w_k}, \tag{2.11.0.5}$$

where the weights $w_k = e^{\beta V_{k-1}(\mathbf{s}_k)}$. The tilde above P indicates that the probability distribution is not normalized. The normalization should be carried out with respect to the CV space, denoted as Ω_n, actually explored up to step n. Thus, the normalization factor is

$$Z_n = \frac{1}{|\Omega_n|} \int_{\Omega_n} \tilde{P}_n(\mathbf{s}) \, \mathrm{d}\mathbf{s}. \tag{2.11.0.6}$$

With an additional positive term $\epsilon \ll 1$, the bias at the n^{th} step becomes

$$V(\mathbf{s}) = (1 - 1/\gamma)\frac{1}{\beta} \log \left(\frac{\tilde{P}_n(\mathbf{s})}{Z_n} + \epsilon \right). \tag{2.11.0.7}$$

ϵ can also allow one to set a limit on the bias, thus providing better control over the desired exploration. It can be chosen to be $\epsilon = e^{-\beta \Delta E/(1-1/\gamma)}$,

where ΔE is the height of the free energy barriers one wishes to overcome during the enhanced sampling. It leads to

$$V(\mathbf{s}) = -\Delta E \tag{2.11.0.8}$$

without depositing biasing potential.

2.12 Orthogonal Space Random Walk

The orthogonal space random walk (OSRW) was developed by Yang in 2008.[55] Phase space sampling is always hindered by free energy barriers. As shown above, several methods have been proposed to accelerate the transition between two states separated by a large free energy barrier, via alchemical process or enhanced conformational switching. In alchemical process, we define a coupling parameter λ. Similarly, in conformational switching we define a reaction coordinate $\mathbf{S}$. Essentially, these two methods are the same, because the coupling parameter λ can be regarded as a coordinate for extended dynamics. Without loss of generality, we can write the free energy difference with the order parameter $\xi = \xi_i$ and $\xi = \xi_f$ as

$$\Delta G(\xi_i \to \xi_f) = \int_{\xi_i}^{\xi_f} \left.\frac{\partial G}{\partial \xi}\right|_{\xi'} \mathrm{d}\xi' = \int_{\xi_i}^{\xi_f} \left\langle \frac{\partial U}{\partial \xi} - \beta^{-1} \frac{\partial \ln |J|}{\partial \xi} \right\rangle_{\xi'} \mathrm{d}\xi', \quad (2.12.0.1)$$

where J is the Jacobian term corresponding to the coordinate transformation between the Cartesian coordinates and the reaction coordinates, and $\frac{\partial U}{\partial \xi} - RT\frac{\partial \ln |J|}{\partial \xi}$ can be regarded as the generalized force F_ξ on ξ. Because the transformation from $\xi = \xi_i$ to $\xi = \xi_f$ is slow, we can either constrain or restrain the system to a series of ξ'. Unfortunately, albeit the acceleration along the reaction coordinate, the relaxation of the other degrees of freedom is usually hindered by some "hidden barriers" and is not able to catch up with the alternation of the reaction coordinate. This is called "Hamiltonian lagging" as identified by Kollman et al.[56] Therefore, acceleration of the space orthogonal to the reaction coordinate is equally important as the acceleration of the reaction coordinate.

Orthogonal space random walk is one of the approaches that can deal with this difficulty. In this method, all the coordinates perpendicular to the reaction coordinate are grouped together into F_ξ. A small two dimensional biasing potential $G(\xi, F_\xi)$, instead of a one-dimensional one as in metadynamics (see Sec. 2.9), is added to the Hamiltonian of the system recursively, which has a functional form like

$$h \exp\left(-\frac{|\xi - \xi(t_i)|^2}{2{w_1}^2}\right) \exp\left(-\frac{|F_\xi - F_\xi(t_i)|^2}{2{w_2}^2}\right). \quad (2.12.0.2)$$

The overall biasing potential

$$G(\xi, F_\xi) = \sum_{t_i} h \exp\left(-\frac{|\xi - \xi(t_i)|^2}{2{w_1}^2}\right) \exp\left(-\frac{|F_\xi - F_\xi(t_i)|^2}{2{w_2}^2}\right). \quad (2.12.0.3)$$

will eventually flatten the underlying free energy surface along the orthogonal space. Application of this biasing potential to conformational free energy calculations is straightforward, while for alchemical free energy calculations

it can be realized by λ-dynamics developed by Charlie Brooks.[31] Similar to metadynamics, the free energy profile along the two-dimensional reaction coordinates $[\xi(t_i), F_\xi]$ can be estimated as $-G\left(\xi, F_\xi\right) + C$, where C is an irrelevant constant. Correspondingly, the generalized force distribution at ξ' should be proportional to $\exp\left[\beta G\left(\xi', F_{\xi'}\right)\right]$, and the free-energy derivative can be obtained via

$$\left.\frac{\partial G}{\partial \xi}\right|_{\xi'} = \langle F_\xi \rangle_{\xi'} = \frac{\sum F_\xi \exp\left[\beta G(\xi, F_\xi)\right] \delta(\xi - \xi')}{\sum \exp\left[\beta G(\xi, F_\xi)\right] \delta(\xi - \xi')}, \tag{2.12.0.4}$$

which can be fed into the thermodynamic integration formula to obtain the free energy change from $\xi = \xi_i$ to any target state with the order parameter ξ as the following

$$\Delta G(\xi) = \int_{\xi_i}^{\xi} \left.\frac{\partial G}{\partial \xi}\right|_{\xi'} \mathrm{d}\xi'. \tag{2.12.0.5}$$

2.13 Enveloping Distribution Sampling

Enveloping distribution sampling method was first proposed by Christ and van Gunsteren in 2007.[57] When calculating the free energy difference between states A and B,

$$\Delta G_{BA} = G_B - G_A = -\beta^{-1} \ln \frac{Q_B}{Q_A}, \tag{2.13.0.1}$$

we may encounter convergence difficulty if the important spaces of these two states are well separated, shown as black lines in Fig. 2.5. Simulation under the Hamiltonian of state A can hardly cover the important region of Hamiltonian B, and then the free energy of state B will be significantly overestimated.

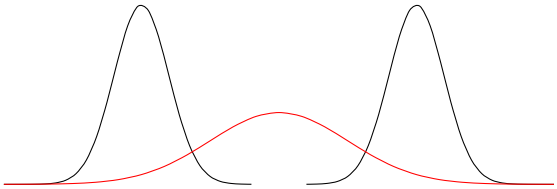

Figure 2.5: The configuration distributions under two Hamiltonians have no visible overlap as shown by solid black curves. A reference state (shown as the red curve) that has remarkable overlap with both states can be introduced to accelerate the convergence of the free energy calculations using, for instance, TP.

A simple solution to this difficulty is "overlap sampling", in which a reference state that can cover the important regions of both Hamiltonians A and B is introduced. We then carry out a simulation for the reference state and the free energy difference between state A and B can be calculated as

$$\Delta G_{BA} = \Delta G_{BR} - \Delta G_{AR} = -\beta^{-1} \ln \frac{\langle e^{-\beta(H_B - H_R)} \rangle_R}{\langle e^{-\beta(H_A - H_R)} \rangle_R}, \tag{2.13.0.2}$$

which is a combination of two thermodynamic perturbation calculations from the reference state to the target states.

However, building the Hamiltonian of the reference state is not trivial. Without knowledge of the Hamiltonians for state A and state B, we cannot generate an effective Hamiltonian, especially in a high dimensional space. Enveloping distribution sampling method provides a natural way to generate the reference state, of which the configurational integral is the sum of the configurational integral of state A and B

$$\begin{aligned} Z_R =& Z_A + Z_B \\ =& \int \left(e^{-\beta H_A(\mathbf{r})} + e^{-\beta H_B(\mathbf{r})} \right) \mathrm{d}\mathbf{r}. \end{aligned} \tag{2.13.0.3}$$

Correspondingly, the Hamiltonian is

$$H_R(\mathbf{r}) = -\beta^{-1} \ln \left(e^{-\beta H_A(\mathbf{r})} + e^{-\beta H_B(\mathbf{r})} \right). \tag{2.13.0.4}$$

A more general form can be written as

$$H_R(\mathbf{r}) = -\left(s\beta\right)^{-1} \ln \left(e^{-s\beta H_A(\mathbf{r})} + e^{-s\beta H_B(\mathbf{r})} \right), \tag{2.13.0.5}$$

where s is a scale factor that modulates the mixing[58] as shown in Fig. 2.6. Increasing s lowers the barrier height separating the two minima in the mixed potential, thereby enhances the transition. Straightforwardly, you may come to the idea that running Hamiltonian-REMD with different s can remarkably increase the efficiency. If you take a close look at Eq. 2.13.0.5, you will find that s appears always with β. In other words, changing s is equivalent to changing the temperature for the simulation. This is one interesting case where H-REMD and T-REMD are coincident with each other.

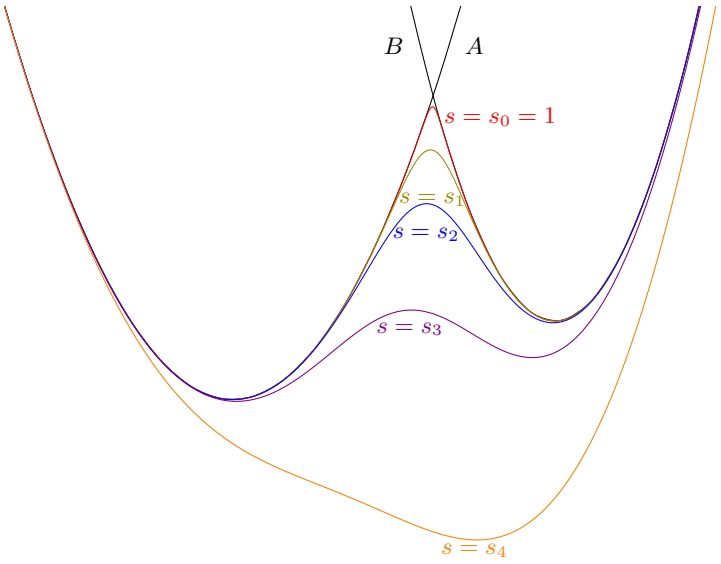

Figure 2.6: State A and state B have only negligible overlap at high energy regions. The reference state generated by the mixing of state A and state B is tuned by s. Decreasing s may lower the barrier between the dominant wells. $s_0, s_1, s_2, s_3, s_4 = 1.0, 0.2, 0.1, 0.05, 0.025$.

The force is also a mixing quantity from two Hamiltonians as

$$\begin{aligned} \mathbf{F}_R^i = -\frac{\partial H_R}{\partial \mathbf{r}^i} =& \frac{e^{-s\beta H_A(\mathbf{r})}}{e^{-s\beta H_A(\mathbf{r})} + e^{-s\beta H_B(\mathbf{r})}} \left(-\frac{\partial H_A(\mathbf{r})}{\partial \mathbf{r}^i} \right) \\ &+ \frac{e^{-s\beta H_B(\mathbf{r})}}{e^{-s\beta H_A(\mathbf{r})} + e^{-s\beta H_B(\mathbf{r})}} \left(-\frac{\partial H_B(\mathbf{r})}{\partial \mathbf{r}^i} \right). \end{aligned} \tag{2.13.0.6}$$

A slight extension of the original EDS implementation reads

$$H_R(\mathbf{r}) = -(s\beta)^{-1} \ln\left(e^{-s\beta(H_A(\mathbf{r})-F_A)} + e^{-s\beta(H_B(\mathbf{r})-F_B)}\right), \qquad (2.13.0.7)$$

where F_A and F_B are the (free) energy offsets. With this new functional form, the mixed state looks like

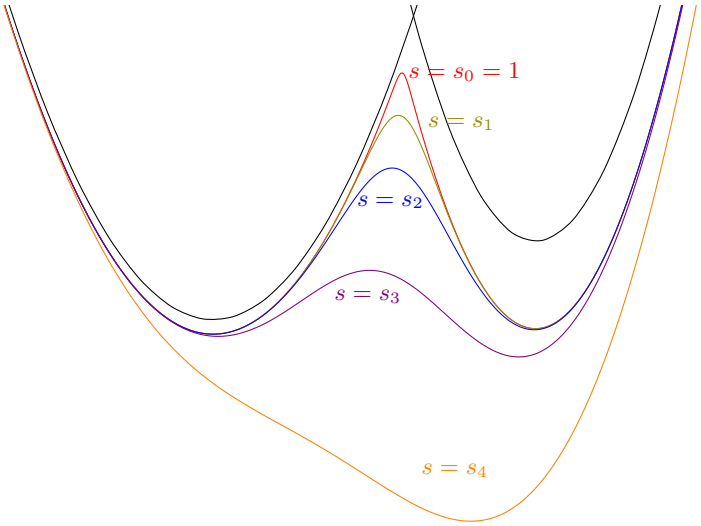

Figure 2.7: The reference state generated by the mixing of state A and state B tuned by s, F_A and F_B. $s_0, s_1, s_2, s_3, s_4 = 1.0, 0.2, 0.1, 0.05, 0.025$.

Interestingly, the idea of EDS can be dated back to the early work by Han[59], of which the best estimate of s is 2. For a more rigorous deviation, please refer to Christ.[60]

Recently, König et al proposed λ-EDS method[61], in which the reference (intermediate) Hamiltonian reads

$$H_{\lambda-EDS} = -[\beta s(\lambda)]^{-1} \ln\left\{(1-\lambda)e^{-\beta s(\lambda)H_A} + \lambda e^{-\beta s(\lambda)[H_B - E(\lambda)]}\right\}. \qquad (2.13.0.8)$$

In this definition, the parameter s is also a function of λ. It returns to EDS with $\lambda = 0.5$, and returns to the minimum variance pathway (MVP) method[62] with $s = 0.5$ and $E = \Delta G_{A\to B}$. With this definition of intermediate states, it can replace soft-core potentials[63] used in alchemical transformations.

2.14 String Method

2.14.1 Zero Temperature String Method

The zero temperature string method was developed by E, Ren and Vanden-Eijnden in 2002.[64] In 2007, they simplified this method by eliminating the projecting the potential force onto the hyperplane perpendicular to the string.[65] The overall algorithm is an iterative application of a simple two-step procedure: (I) evolution of the string by standard ordinary differential equation (ODE) solvers, and (II) the reparametrization of the string by interpolation. The accuracy of this method is determined by the 2nd step, while its efficiency is determined by the 1st step.

The minimum energy path (MEP) is the most probable path that the system will take under the overdamped dynamics to move between two minima at a and b on the potential energy surface $V(x)$, crossing the barriers in-between. Let's denote the MEP by γ, which satisfies

$$(\nabla V)^{\perp}(\gamma) = 0, \tag{2.14.1.1}$$

where $(\nabla V)^{\perp}$ is the component of (∇V) normal to γ,

$$(\nabla V)^{\perp}(\gamma) = \nabla V(\gamma) - (\nabla V(\gamma), \hat{\tau})\,\hat{\tau}. \tag{2.14.1.2}$$

Here $\hat{\tau}$ denotes the unit tangent of the curve γ, and $(\cdot,\cdot)$ denotes the Euclidean inner product.

The string method locates the MEP by evolving a curve connecting a and b, under the potential force field. The simplest dynamics for the evolution is given abstractly by

$$\nu_n = -(\nabla V)^{\perp}, \tag{2.14.1.3}$$

where ν_n denotes the normal velocity of the curve. Only the normal component of the velocity matters for the evolution of a curve, while the tangential velocity only moves points along the curve, changing the parametrization of the curve without changing the curve itself.

First, we take a particular parametrization of the curve $\gamma : \gamma = \{\boldsymbol{\varphi}(\alpha) : \alpha \in [0,1]\}$. Then we have $\hat{\tau}(\alpha) = \boldsymbol{\varphi}_\alpha/|\boldsymbol{\varphi}_\alpha|$, where $\boldsymbol{\varphi}_\alpha$ denotes the derivative of $\boldsymbol{\varphi}$ with respect to α. The simplest parametrization is equal arc-length parametrization, in which α is a constant multiple of the arc length from a to the point $\boldsymbol{\varphi}(\alpha)$. In this case, we also have $|\boldsymbol{\varphi}_\alpha| = \text{const}$ (this constant being the length of the curve γ).

The string method evolves the curve via

$$\dot{\boldsymbol{\varphi}} = \nabla V(\boldsymbol{\varphi}) + \bar{\lambda}\hat{\tau}, \tag{2.14.1.4}$$

where $\bar{\lambda}(\alpha,t)\hat{\tau}(\alpha,t)$ is a Lagrange multiplier term for the purpose of enforcing the particular parametrization of the string. The string is discretized

into a number of images $\{\boldsymbol{\varphi}_i(t), i = 0, 1, \ldots, N\}$. The images along the string are evolved by iterating upon the following two-step procedure based on time splitting of the terms at the right hand side of Eq. 2.14.1.4.

In the first step, the discrete point on the string are evolved over some time interval Δt according to the full potential force,

$$\dot{\boldsymbol{\varphi}}_i = -\nabla V(\boldsymbol{\varphi}_i). \tag{2.14.1.5}$$

This equation can be integrated in time by any suitable ODE solver. If we denote the positions of the images after n iterations of the scheme by $\boldsymbol{\varphi}_i^n$, $i = 0, \ldots, N$, the new set of images are given by

$$\boldsymbol{\varphi}_i^* = \boldsymbol{\varphi}_i^n - \Delta t \nabla V(\boldsymbol{\varphi}_i^n). \tag{2.14.1.6}$$

In the second step, this new set of images are redistributed along the string using a simple interpolation/reparametrization procedure. Two possible schemes for reparametrization can be applied.

Parametrization by equal arc length Given the values $\{\boldsymbol{\varphi}_i^*\}$ on a nonuniform mesh $\{\alpha_i^*\}$, these values are interpolated onto a uniform mesh with the same number of points via two steps:

1. The arc length corresponding to the current images,

$$s_0 = 0, \quad s_i = s_{i-1} + |\boldsymbol{\varphi}_i^* - \boldsymbol{\varphi}_{i-1}^*|, \quad i = 1, 2, \ldots, N. \tag{2.14.1.7}$$

   The mesh $\{\alpha_i^*\}$ is then obtained by normalizing $\{s_i\}$,

$$\alpha_i^* = s_i / s_N. \tag{2.14.1.8}$$

2. The points $\boldsymbol{\varphi}_i^{n+1}$ at the uniform grids points $\alpha_i = i/N$ are obtained by interpolation. This can be done, for example, by using cubic spline interpolation for the data $\{(\alpha_i^*, \boldsymbol{\varphi}_i^*), i = 0, \ldots, N\}$

Parametrization by energy-weighted arc length The energy-weighted arc length parameterization gives finer resolution around the saddle points, and thus better estimate of the energy barrier and also the unstable direction at those points than the equal arc length scheme does. In this scheme, the energy-weighted arc length corresponding to the current images are computed,

$$s_0^w = 0, \quad s_i^w = s_{i-1}^w + W_{i-(1/2)}|\boldsymbol{\varphi}_i^* - \boldsymbol{\varphi}_{i-1}^*|, \quad i = 1, 2, \ldots, N. \tag{2.14.1.9}$$

Here $W_{i-(1/2)} = W(V_{i+1/2})$ and $V_{i+1/2}$ is the average of the potential energy at $\boldsymbol{\varphi}_{i-1}^*$ and $\boldsymbol{\varphi}_i^*$. The weight function $W(z)$ is some positive, increasing function of $z \in \mathbb{R}$. The mesh $\{\alpha_i^*\}$ is obtained by normalizing $\{s_i^w\} : \alpha_i^* = s_i^w / s_N^w$. The new points $\boldsymbol{\varphi}_i$ on $\alpha_i = i/N$ are then calculated by cubic spline interpolation across the points $\{\boldsymbol{\varphi}_i^*, i = 0, \ldots, N\}$.

Once the new points $\{\boldsymbol{\varphi}_i^{n+1}, i = 0, \ldots, N\}$ are calculated, the algorithm goes back to step 1 and iterates until convergence.

2.14.2 Finite Temperature String Method

Finite temperature string method is an extension of the zero temperature string method. Let $\boldsymbol{\varphi}(\alpha)$ be a curve in configuration space parametrized by $\alpha \in [0,1]$ whose end points, $\boldsymbol{\varphi}(0)$ and $\boldsymbol{\varphi}(1)$, belong to the two metastable sets. In order to have $\boldsymbol{\varphi}(\alpha)$ converges towards the center of the effective reaction tube, an ensemble of realizations $\{\boldsymbol{\varphi}^{\omega}(\alpha)\}$ are introduced and their mean is defined to be the string, i.e., $\langle \boldsymbol{\varphi}^{\omega}(\alpha) \rangle \equiv \boldsymbol{\varphi}(\alpha)$. Each realization evolves by

$$\boldsymbol{\varphi}_t^{\omega} = -\left(\nabla V(\boldsymbol{\varphi}^{\omega})\right)^{\perp} + (\eta^{\omega})^{\perp} + r\hat{t}. \tag{2.14.2.1}$$

Here $\hat{t} = \boldsymbol{\varphi}_{\alpha}/|\boldsymbol{\varphi}_{\alpha}|$ is the unit tangent vector along $\boldsymbol{\varphi}$ and $a^{\perp} = a - (\hat{t} \cdot a)\hat{t}$ is the projection of the vector a in the hyperplane normal to the string $\boldsymbol{\varphi}(\alpha)$ denoted here by $S(\alpha)$. η^{ω} is a white noise which covariance

$$\left\langle \eta_j^{\omega}(\alpha,t)\eta_k^{\omega}(\alpha',0) \right\rangle = \begin{cases} 2k_B T\delta_{jk}\delta t & \text{if } \alpha = \alpha' \\ 0 & \text{otherwise} \end{cases} \tag{2.14.2.2}$$

The scalar field $r \equiv r(\alpha,t)$ is a Lagrange multiplier term to preserve some particular parametrization of the string $\boldsymbol{\varphi}$ chosen beforehand, for instance the equal arc length or equal energy-weighted arc length.

The equilibrium density function for Eq. 2.14.2.1 is given by

$$\rho(\mathbf{q},\alpha) = Z^{-1}(\alpha)e^{-\beta V(\mathbf{q})}\delta_{S(\alpha)}(\mathbf{q}) \tag{2.14.2.3}$$

where $\delta_{S(\alpha)}(\mathbf{q})$ is the Dirac distribution concentrated on $S(\alpha)$, and $Z(\alpha) = \int_{S(\alpha)} e^{-\beta V(\mathbf{q})}\,\mathrm{d}\sigma$ is the normalization constant. By definition, the center of the transition tube is given by

$$\boldsymbol{\varphi}(\alpha) = Z^{-1}(\alpha)\int_{S(\alpha)} \mathbf{q}e^{-\beta V(\mathbf{q})}\,\mathrm{d}\sigma. \tag{2.14.2.4}$$

The width of the effective transition tube itself can be characterized by a few times the variance of $\mathbf{q}$ around $\boldsymbol{\varphi}(\alpha)$; i.e., its local radius square can be defined as

$$R^2(\alpha) = \lambda Z^{-1}(\alpha)\int_{S(\alpha)} |\mathbf{q} - \boldsymbol{\varphi}(\alpha)|^2 e^{-\beta V(\mathbf{q})}\,\mathrm{d}\sigma, \tag{2.14.2.5}$$

where λ is a number of order unity. The integral of $\rho(\mathbf{q},\alpha)$ in the ball of radius $R(\alpha)$ centered around $\rho(\alpha)$ gives the probability that a dynamical trajectory involved in the transition event crosses the plane $S(\alpha)$ within this ball.

The free energy is given by

$$F(\alpha) = -k_B T \ln \int_{S(\alpha)} e^{-\beta V(\mathbf{q})}\,\mathrm{d}\sigma. \tag{2.14.2.6}$$

Using thermodynamic integration, the free energy difference between α_1 and α_2 becomes

$$F(\alpha_2) - F(\alpha_1) = \int_{\alpha_1}^{\alpha_2} \left\langle (\hat{t} \cdot \nabla V)((\hat{t} \cdot \boldsymbol{\varphi})_\alpha - \hat{t} \cdot \boldsymbol{\varphi}) \right\rangle \mathrm{d}\alpha. \tag{2.14.2.7}$$

In this implementation, constrained molecular dynamics must be carried out for the calculations of ensemble averages over S_α, which is usually difficult. In 2009, Vanden-Eijnden and Venturoli updated this FTS algorithm by replacing the hyperplanes perpendicular to the string with Voronoi cells.[66] This new algorithm begins with an initial set of images, $\boldsymbol{\varphi}_\alpha^0$ with $\alpha = 0, \ldots, N$ with equal arc length, i.e. $|\boldsymbol{\varphi}_{\alpha+1}^0 - \boldsymbol{\varphi}_\alpha^0| = |\boldsymbol{\varphi}_\alpha^0 - \boldsymbol{\varphi}_{\alpha-1}^0|$ for all $\alpha = 1, \ldots, N-1$. Each image is associated with a replica of the original system, $\mathbf{x}_\alpha^0$, $(\boldsymbol{\varphi}(\mathbf{x}_\alpha^0) = \boldsymbol{\varphi}_\alpha^0)$. Then the positions of these systems are updated iteratively for $n \geq 0$ upon the following steps:

1. Update $\mathbf{x}_\alpha^n$ with a reflecting boundary condition at the boundary of the Voronio cell associated with the image $\boldsymbol{\varphi}_\alpha^n$ via, for example,

$$\mathbf{x}_\alpha^* = \mathbf{x}_\alpha^n - \Delta t \nabla V(\mathbf{x}_\alpha^n) + \sqrt{2\beta^{-1}\Delta t}\boldsymbol{\xi}_\alpha^n \tag{2.14.2.8}$$

and set

$$\mathbf{x}_\alpha^{n+1} = \begin{cases} \mathbf{x}_\alpha^*, & \text{if } \mathbf{x}_\alpha^* \in B_\alpha^n \\ \mathbf{x}_\alpha^n, & \text{otherwise} \end{cases} \tag{2.14.2.9}$$

where

$$B_\alpha^n = \{\mathbf{x} \text{ such that } |\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}_\alpha^n| < |\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}_{\alpha'}^n| \text{ for all } \alpha' \neq \alpha\} \tag{2.14.2.10}$$

Δt denotes the time step used for numerical integration and $\boldsymbol{\xi}_\alpha^n$ are independent Gaussian variables with mean 0 and variance 1.

2. Compute the running average of each $\mathbf{x}_\alpha^n$,

$$\bar{\mathbf{x}}_\alpha^n = \frac{1}{n} \sum_{n'=0}^{n} \mathbf{x}_\alpha^{n'}. \tag{2.14.2.11}$$

3. Update each image along the string toward the running average $\bar{\mathbf{x}}_\alpha^n$ while keeping the string smooth. To do so use

$$\boldsymbol{\varphi}_\alpha^* = \boldsymbol{\varphi}_\alpha^n - \Delta\tau(\boldsymbol{\varphi}_\alpha^n - \boldsymbol{\varphi}(\bar{\mathbf{x}}_\alpha^n)) + \mathbf{r}_\alpha^*, \tag{2.14.2.12}$$

where $\Delta\tau > 0$, $\mathbf{r}_0^* = \mathbf{r}_N^* = 0$, and for $\alpha = 1, \ldots, N-1$,

$$\mathbf{r}_\alpha^* = k^n(\boldsymbol{\varphi}_{\alpha+1}^* + \boldsymbol{\varphi}_{\alpha-1}^* - 2\boldsymbol{\varphi}_\alpha^*). \tag{2.14.2.13}$$

Here $k^n > 0$ is an adjustable parameter, whose value controls how aggressive the smoothing is.

4. Enforcing the equal arc-length parametrization by interpolating a piecewise linear curve through $\{\boldsymbol{\varphi}^*_\alpha\}_{\alpha=0,\dots,N}$ and redistributing points at equal distance along this curve to obtain $\{\boldsymbol{\varphi}^{n+1}_\alpha\}_{\alpha=0,\dots,N}$.

5. If $\mathbf{x}^{n+1}_\alpha \in B^{n+1}_\alpha$ go to step 1, otherwise set $\mathbf{x}^{n+1}_\alpha = \boldsymbol{\varphi}^{n+1}_\alpha$ and go to step 1. Repeat until convergence of $\{\boldsymbol{\varphi}^{n+1}_\alpha\}_{\alpha=0,\dots,N}$.

2.15 Optimally Adjusted Mixed Sampling

Optimally Adjusted Mixed Sampling method was proposed by Tan in 2017[67], which accommodates both adaptive serial tempering and a generalized Wang–Landau algorithm.

2.15.1 Labeled Mixture Sampling

Let us begin with the non-adaptive version of self-adjusted mixture sampling, which is termed labeled mixture sampling. With m simulation conditions in total, we define the joint distribution on the space $\{1, \ldots, m\} \times \chi$,

$$(L, X) \sim p(j, x; \xi) \propto \frac{\pi_j}{e^{-\xi_j}} q_j(x), \tag{2.15.1.1}$$

where $\pi = (\pi_1, \ldots, \pi_m)^T$ are the fixed mixture weights with $\sum_{j=1}^m \pi_j = 1$, $q_j(x)$ is the unnormalized density function for state j, and $\xi = (\xi_1, \ldots, \xi_m)^T$ with $\xi_1 = 0$ are the *hypothesized* values of the true ratios of normalizing constants $\xi^* = (\xi_1^*, \ldots, \xi_m^*)^T$ with

$$\xi_j^* = -\ln(Z_j/Z_1). \tag{2.15.1.2}$$

Note that there is a minus sign in this equation, following the convention of chemists. In a serial tempering method, m is the total number of temperature states. By normalization

$$C \sum_{j=1}^m \int \frac{\pi_j}{e^{-\xi_j}} q_j(x)\, \mathrm{d}x = 1, \tag{2.15.1.3}$$

we find

$$C = \frac{1}{\sum_{l=1}^m \pi_l e^{\xi_l - \xi_l^*}}. \tag{2.15.1.4}$$

Therefore,

$$p(j, x; \xi) = \frac{\pi_j e^{\xi_j} q_j(x)}{\sum_{l=1}^m \pi_l e^{\xi_l - \xi_l^*}}. \tag{2.15.1.5}$$

The marginal distribution of L is

$$p(L = j; \xi) = \sum_{j=1}^m p(j, x; \xi) = \frac{\pi_j e^{\xi_j - \xi_j^*}}{\sum_{l=1}^m \pi_l e^{\xi_l - \xi_l^*}}, \quad j = 1, \ldots, m. \tag{2.15.1.6}$$

The marginal distribution of X is

$$p(x; \xi) = \sum_{j=1}^m p(j, x; \xi) = \frac{\sum_{j=1}^m \pi_j e^{\xi_j} q_j(x)}{\sum_{l=1}^m \pi_l e^{\xi_l - \xi_l^*}}, \quad x \in \chi. \tag{2.15.1.7}$$

Given any fixed choice of ξ and π, this mixed ensemble can be sampled by standard Markov Chain Monte Carlo (MCMC). Starting from some initial values (L_0, X_0), the Metropolis-Hasting (MH) algorithm is as follows

1. Generate (j, x) from a proposal distribution $Q\{(L_{t-1}, X_{t-1}), \cdot; \xi\}$.

2. Set $(L_t, X_t) = (j, x)$ with probability

$$\min\left[1, \frac{Q\{(j,x),(L_{t-1},X_{t-1});\xi\}}{Q\{(L_{t-1},X_{t-1}),(j,x);\xi\}} \frac{p(j,x;\xi)}{p(L_{t-1},X_{t-1};\xi)}\right],$$

and with the remaining probability, set $(L_t, X_t) = (L_{t-1}, X_{t-1})$.

For overlap-based sampling, assume that a Markov transition kernel $\Psi_j(x;\cdot)$ is constructed by MCMC with $P_j = q_j/Z_j$ being the stationary distribution. Then, a particular choice of $Q(\cdot,\cdot;\xi)$ is to update L_t and X_t one at a time, leading to a two-block MH algorithm using $\Psi = (\Psi_1, \ldots, \Psi_m)$. The conditional distributions are

$$p(x|L=j) \propto q_j(x), \tag{2.15.1.8}$$

$$p(L=j|x;\xi) = \frac{\pi_j e^{\xi_j} q_j(x)}{\sum_{l=1}^m \pi_l e^{\xi_l} q_l(x)} \propto \frac{\pi_j}{e^{-\xi_j}} q_j(x). \tag{2.15.1.9}$$

Specifically, two sampling schemes can be proposed. One is global-jump algorithm by sampling directly from $p(L = j|x;\xi)$.

Global-jump labeled mixture sampling:

1. *Global jump*: Generate $L_t \sim p(L = \cdot|X_{t-1};\xi)$.

2. *Markov move*: Generate $X_t \sim \Psi_{L_t}(X_{t-1}, \cdot)$.

The other is local-jump algorithm, which is exactly serial tempering method.

Local-jump labeled mixture sampling:

1. *Local jump*: Generate $j \sim \Gamma(L_{t-1}, \cdot)$ and then set $L_t = j$ with probability

$$\min\left[1, \frac{\Gamma(j, L_{t-1})}{\Gamma(L_{t-1}, j)} \frac{p(j|X_{t-1};\xi)}{p(L_{t-1}|X_{t-1};\xi)}\right],$$

and, with the remaining probability, set $L_t = L_{t-1}$.

2. *Markov move*: Generate $X_t \sim \Psi_{L_t}(X_{t-1}, \cdot)$.

2.15.2 Self-Adjusted Mixture Sampling

It can be seen from Eq. 2.15.1.5 that if we choose $\xi = \xi^*$, we have

$$P(L = j;\xi) = \pi_j, \tag{2.15.2.1}$$

which is the optimum condition. However, ξ^* is unknown *a priori*. If we are unlucky by choosing ξ not close to ξ^*, then the marginal probability of one label j may be orders of magnitude smaller than those of other labels,

indicating that P_j is not adequately sampled. Motivated by the Wang–Landau algorithm, the self-adjusted mixture sampling method was proposed. Let $\xi^{(0)}$ be some initial choice of ξ, for example, the m vector of zeros. Denote by $\xi^{(t)} = (\xi_1^{(t)}, \ldots, \xi_m^{(t)})^T$ a choice of ξ at iteration t.

Stochastic approximation Monte Carlo (SAMC):

1. *Local jump and Markov move*: same as local-jump labeled mixture sampling.

2. *Free energy update*: set $\delta^{(t)} = (1\{L_t = 1\}, \ldots, q\{L_t = m\})^T$ and

$$\xi^{t-\frac{1}{2}} = \xi^{t-1} + \gamma_t(\delta^{(t)} - \pi), \quad \xi^{(t)} = \xi^{(t-\frac{1}{2})} - \xi_1^{t-\frac{1}{2}},$$

   where $\xi_1^{(t-\frac{1}{2})}$ is the first element of $\xi^{(t-\frac{1}{2})}$ and $\gamma_t = t_0/\max(t_0, t)$ for some fixed value $t_0 > 1$.

The free energy update in the SAMC algorithm is an application of stochastic approximation to find ξ^* as a unique solution to $p(L = j; \xi) = \pi_j$ $(j = 1, \ldots, m)$. Informally, there is a self-adjusting mechanism as follows. If the jth element ξ_j^{t-1} is greater (or smaller) than ξ_j^*, then the label L_t will take value j more likely (or less likely) than with probability π_j.

TO BE CHECKED

3

Postprocessing

"The source of mistake is always between the keyboard and the chair. So, check, double check and check again."

– Gerhard König

3.1 Rigorous Methods

In the following, we will introduce some widely used post-processing methods. Each method has its pros-and-cons. For comparison of these methods, please refer to some benchmark papers.[68–70]

3.1.1 Thermodynamic Perturbation

Thermodynamic Perturbation (TP), also known as Free Energy Perturbation (FEP), exponential average, or Zwanzig equation, was developed by Zwanzig,[71] and Landau and Lifshitz, independently, and probably by Peierls[72].

A reference system containing N-particles can be described by Hamiltonian $H_0(\mathbf{x}, \mathbf{p}_x)$, which is a function of $3N$ Cartesian coordinates, $\mathbf{x}$, and their conjugated momenta, $\mathbf{p}_x$. Similarly, the target system can be described by Hamiltonian $H_1(\mathbf{x}, \mathbf{p}_x)$. These two systems are related by

$$H_1(\mathbf{x}, \mathbf{p}_x) = H_0(\mathbf{x}, \mathbf{p}_x) + \Delta H(\mathbf{x}, \mathbf{p}_x) \tag{3.1.1.1}$$

The Helmholtz free energy difference between the target and the reference systems, ΔA, can be given in terms of the ratio of the corresponding partition functions, Q_1 and Q_0:

$$\Delta A = -\frac{1}{\beta} \ln \frac{Q_1}{Q_0}, \tag{3.1.1.2}$$

where $\beta = (k_B T)^{-1}$, and

$$Q_i = \frac{1}{h^{3N} N!} \iint \exp\left[-\beta H_i(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}. \tag{3.1.1.3}$$

Taking Eq. 3.1.1.3 into Eq. 3.1.1.2, we obtain

$$\Delta A = -\frac{1}{\beta} \ln \frac{\iint \exp\left[-\beta H_1(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}}{\iint \exp\left[-\beta H_0(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}} \tag{3.1.1.4}$$

$$= -\frac{1}{\beta} \ln \frac{\iint \exp\left[-\beta \Delta H(\mathbf{x}, \mathbf{p}_x)\right] \exp\left[-\beta H_0(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}}{\iint \exp\left[-\beta H_0(\mathbf{x}, \mathbf{p}_x)\right] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}}, \tag{3.1.1.5}$$

The probability density function of finding the reference system in a state defined by positions $\mathbf{x}$ and momenta $\mathbf{p}_x$ is

$$P_0(\mathbf{x}, \mathbf{p}_x) = \frac{\exp[-\beta H_0(\mathbf{x}, \mathbf{p}_x)]}{\iint \exp[-\beta H_0(\mathbf{x}, \mathbf{p}_x)] \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}} \tag{3.1.1.6}$$

If the probability density function is used, Eq. 3.1.1.5 becomes

$$\Delta A = -\frac{1}{\beta} \ln \iint \exp[-\beta \Delta H(\mathbf{x}, \mathbf{p}_x)] P_0(\mathbf{x}, \mathbf{p}_x)\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}, \tag{3.1.1.7}$$

or, equivalently,

$$\Delta A = -\frac{1}{\beta} \ln \left\langle \exp\left[-\beta \Delta H(\mathbf{x}, \mathbf{p}_x)\right] \right\rangle_0, \tag{3.1.1.8}$$

Here, $\langle \cdots \rangle_0$ denotes an ensemble average over configurations sampled from the reference state. Equation 3.1.1.8 is the basic equation of TP. It states that ΔA can be estimated by sampling only equilibrium configurations of the reference state.

Note that integration over the kinetic term in the partition function, Eq. 3.1.1.3, can be carried out analytically. Thus, it cancels out in Eq. 3.1.1.2, and Eq. 3.1.1.8 becomes

$$\Delta A_f = -\frac{1}{\beta} \ln \left\langle \exp(-\beta \Delta U) \right\rangle_0, \tag{3.1.1.9}$$

where ΔU is the difference in the potential energy between the target and the reference states. The subscript f is an indication of a forward ($0 \to 1$) TP calculation. The integration implied by the statistical average is now carried out over particle coordinates only. The variance of ΔA is

$$\delta^2 \Delta A_f = \frac{1}{N_0 \beta^2} \left(\frac{\left\langle (\exp(-\beta \Delta U))^2 \right\rangle_0}{\left(\left\langle \exp(-\beta \Delta U) \right\rangle_0 \right)^2} - 1 \right). \tag{3.1.1.10}$$

If we exchange the reference and the target systems, and repeat the same derivation, using the same convention for ΔA and ΔU as before, we have a backward TP expression for the free energy difference

$$\Delta A_b = \frac{1}{\beta} \ln \langle \exp(\beta \Delta U) \rangle_1, \tag{3.1.1.11}$$

and the variance is

$$\delta^2 \Delta A_b = \frac{1}{N_1 \beta^2} \left(\frac{\langle (\exp(\beta \Delta U))^2 \rangle_1}{(\langle \exp(\beta \Delta U) \rangle_1)^2} - 1 \right). \tag{3.1.1.12}$$

Although expressions Eq. 3.1.1.9 and Eq. 3.1.1.11 are formally equivalent, their convergence properties may be quite different. This means that there is a preferred direction to carry out the required transformation between the two states. One should start the perturbation from the state having larger important region in phase space. In other words, the reference system should be the one with higher entropy, and the transformation should proceed in the direction in which the entropy decreases. If we have the free energy differences from both the forward and backward TP calculations, we can compute the "best estimate" of ΔA as

$$\Delta A = \frac{(\delta^2 \Delta A_f)^{-1} \Delta A_f + (\delta^2 \Delta A_b)^{-1} \Delta A_b}{(\delta^2 \Delta A_f)^{-1} + (\delta^2 \Delta A_b)^{-1}}, \tag{3.1.1.13}$$

with variance

$$\delta^2 \Delta A = \frac{1}{(\delta^2 \Delta A_f)^{-1} + (\delta^2 \Delta A_b)^{-1}}. \tag{3.1.1.14}$$

Since ΔA is calculated as the average over a quantity that depends only on ΔU, this average can be computed over probability distribution $P_0(\Delta U)$ instead of $P_0(\mathbf{x}, \mathbf{p}_x)$. Then, ΔA in Eq. 3.1.1.7 can be expressed as a one-dimensional integral over energy difference

$$\Delta A = -\frac{1}{\beta} \ln \int \exp(-\beta \Delta U) P_0(\Delta U) \, \mathrm{d}\Delta U, \tag{3.1.1.15}$$

If U_0 and U_1 were functions of a sufficient number of identically distributed random variable, ΔU would follow a Gaussian distribution, which is a consequence of the central limit theorem. In practice, the probability distribution $P_0(\Delta U)$ deviates somewhat from an ideal Gaussian case, but still has a "Gaussian-like" shape. This indicates that the value of the integral in Eq. 3.1.1.15 depends on the low-energy tail of the distribution. Using the language of Jarzynski for the nonequilibrium work[73], the maximum of $P_0(\Delta U)$ is the typical energy difference, and the peak value of $P_0(\Delta U) \cdot \exp(-\beta \Delta U)$ is the dominant realization that contributes the most to ΔA. It clearly shows in Fig. 3.1 that the dominant realization lies to the left of the typical one.

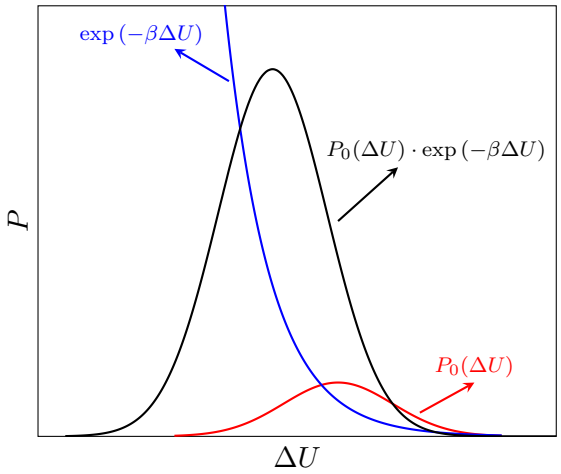

Figure 3.1: $P_0(\Delta U)$, the Boltzmann factor $\exp(-\beta\Delta U)$ and their product, which is the integrand in Eq. 3.1.1.15. The low-ΔU tail of the integrand is poorly sampled with $P_0(\Delta U)$ and, therefore, is known with low statistical accuracy. However, it provides an important contribution to the integral.

Even though $P_0(\Delta U)$ is only rarely an exact Gaussian, it is instructive to consider this case in more detail. If we substitute

$$P_0(\Delta U) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\Delta U - \langle \Delta U \rangle_0)^2}{2\sigma^2}\right] \tag{3.1.1.16}$$

where

$$\sigma^2 = \left\langle \Delta U^2 \right\rangle_0 - \langle \Delta U \rangle_0^2 \tag{3.1.1.17}$$

to Eq. 3.1.1.15, we obtain

$$\exp(-\beta\Delta A) = \frac{C}{\sqrt{2\pi}\sigma} \int \exp\left[-\frac{\left[\Delta U - \left(\langle \Delta U \rangle_0 - \beta\sigma^2\right)\right]^2}{2\sigma^2}\right] \mathrm{d}\Delta U \tag{3.1.1.18}$$

Here,

$$C = \exp\left[-\beta\left(\langle \Delta U \rangle_0 - \frac{1}{2}\beta\sigma^2\right)\right] \tag{3.1.1.19}$$

is independent of ΔU. It can be seen from Eq. 3.1.1.18 that from $P_0(\Delta U)$ to $\exp(-\beta\Delta U)P_0(\Delta U)$, the maximum shifts to the left by $\beta\sigma^2$, which is proportional to the variance of the Gaussian distribution $P_0(\Delta U)$, σ^2. In other words, the wider the distribution $P_0(\Delta U)$, the slower the convergence of TP calculation is. For the number of samples required to ensure the convergence of the TP calculation, the readers may refer to Ref. [74].

If $P_0(\Delta U)$ is Gaussian, the integral in Eq. 3.1.1.18 can be evaluated analytically using cumulant expansion (see appendix D)

$$\Delta A = \langle \Delta U \rangle_0 - \frac{1}{2}\beta\sigma^2. \tag{3.1.1.20}$$

If the distribution of ΔU deviates from Gaussian, there will be extra terms measuring the skewness of Gaussian. With the leading term, ΔA becomes

$$\Delta A = \langle \Delta U \rangle_0 - \frac{1}{2}\beta\sigma^2 + \frac{\beta^2}{6}\left(\left\langle \Delta U^3 \right\rangle_0 - 3\left\langle \Delta U^2 \right\rangle_0 \langle \Delta U \rangle_0 + 2\langle \Delta U \rangle_0^3\right). \tag{3.1.1.21}$$

In 2002, Jarzynski proposed a generalized free energy perturbation method termed "targeted free energy perturbation".[75] This method generalizes the free energy perturbation method between two states, A and B, by introducing an intermediate state $A'(\mathbf{y})$ by mapping $\mathcal{M}$ from the microstate of the system $\mathbf{x}$ to another one $\mathbf{y}$ in the configuration space or phase space

$$\mathcal{M} : \mathbf{x} \to \mathbf{y}(\mathbf{x}), \tag{3.1.1.22}$$

The potential energy function of the mapped state $A'(\mathbf{y})$ is

$$U_{A'}(\mathbf{y}) = U_A(\mathbf{x}) + \beta^{-1} \ln J(\mathbf{x}), \tag{3.1.1.23}$$

where $J(\mathbf{x}) = |\partial \mathbf{y}/\partial \mathbf{x}|$ is the Jacobian of the mapping $\mathcal{M}$. Zhu et al also used this transformation, but for enhancing conformational sampling.[76] The Helmholtz free energy difference between state A and state A' is zero by noticing that

$$\begin{aligned}
\Delta F_{A'A} =& -\beta^{-1} \ln \frac{Z_{A'}}{Z_A} \\
=& -\beta^{-1} \ln \frac{\int e^{-\beta U_{A'}(\mathbf{y})}\, \mathrm{d}y}{Z_A} \\
=& -\beta^{-1} \ln \frac{\int e^{-\beta(U_{A'}(\mathbf{y}) - U_A(\mathbf{x}))} e^{-\beta U_A(\mathbf{x})}\, \mathrm{d}y}{Z_A} \\
=& -\beta^{-1} \ln \frac{\int J^{-1}(\mathbf{x}) e^{-\beta U_A(\mathbf{x})}\, \mathrm{d}y}{Z_A} \\
=& -\beta^{-1} \ln \frac{\int e^{-\beta U_A(\mathbf{x})}\, \mathrm{d}x}{Z_A} \\
=& 0.
\end{aligned}$$

The calculation of the free energy difference between state A and state B can be replaced by the calculation of free energy difference between state A' and state B. The latter can be written as

$$\begin{aligned}
\Delta F_{BA'} =& -\beta^{-1} \ln \int e^{-\beta(U_B(\mathbf{y}) - U_{A'}(\mathbf{y}))} \rho_{A'}(\mathbf{y})\, \mathrm{d}\mathbf{y} \\
=& -\beta^{-1} \ln \int e^{-\beta\Phi(\mathbf{x})} \rho_A(\mathbf{x})\, \mathrm{d}\mathbf{x}
\end{aligned} \tag{3.1.1.24}$$

in which

$$\begin{aligned}\Phi(\mathbf{x}) \equiv & U_B(\mathbf{y}) - U_{A'}(\mathbf{y}) \\ = & U_B(\mathbf{y}) - U_A(\mathbf{x}) - \beta^{-1} \ln J(\mathbf{x}),\end{aligned} \tag{3.1.1.25}$$

and the relationship between the distributions

$$\rho_{A'}(\mathbf{y}) = \rho_A(\mathbf{x})/J(\mathbf{x}) \tag{3.1.1.26}$$

has been invoked.

As a special case $\mathcal{M} : \mathbf{x} \to \mathbf{x}$,

$$\Phi(\mathbf{x}) \equiv U_B(\mathbf{x}) - U_A(\mathbf{x}), \tag{3.1.1.27}$$

and it reduces to the traditional free energy perturbation. It implies that there may exist an invertible mapping $\mathcal{M}$ for which the average of $\exp\{-\beta\Phi\}$ converges more rapidly than the average of $\exp\{-\beta(U_B - U_A)\}$. It can also be asserted that

$$\begin{aligned}e^{-\beta\Delta F} &= \left\langle e^{-\beta\Phi} \right\rangle_A \\ &= \int \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) e^{-\beta\Phi(\mathbf{x})} \\ &= \int \mathrm{d}\phi \int \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) e^{-\beta\Phi(\mathbf{x})} \delta(\Phi(\mathbf{x}) - \phi) \\ &= \int \mathrm{d}\phi\, e^{-\beta\phi} \int \rho(\mathbf{x}) \delta(\Phi(\mathbf{x}) - \phi)\, \mathrm{d}\mathbf{x} \\ &= \int \mathrm{d}\phi\, e^{-\beta\phi} p(\phi|\mathcal{M}),\end{aligned} \tag{3.1.1.28}$$

where $p(\phi|\mathcal{M})$ is the distribution of values of $\phi = \Phi(\mathbf{x})$ for $\mathbf{x}$ sampled from A. In practice, ΔF can be estimated by averaging $\exp(-\beta\Phi)$ over a finite number of sampled microstates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

$$\Delta F = -\beta^{-1} \ln \frac{1}{N} \sum_{n=1}^{N} e^{-\beta\Phi(\mathbf{x}_n)}. \tag{3.1.1.29}$$

The rate of convergence depends on the choice of $\mathcal{M}$. A perfect mapping $\mathcal{M}^*$ that transforms A exactly onto B, we will find from Eq. 3.1.1.26 that

$$\frac{1}{Z_B} e^{-\beta E_B(\mathbf{y})} = \frac{1}{Z_A} e^{-\beta E_A(\mathbf{x})}/J(\mathbf{x}), \tag{3.1.1.30}$$

which leads to

$$p(\phi|\mathcal{M}^*) = \delta(\phi - \Delta F). \tag{3.1.1.31}$$

Then the convergence of the finite sampling is immediate: $\Phi(\mathbf{x}) = \Delta F$ for every sampled $\mathbf{x}$. Although constructing a perfect transformation is usually

impossible for real-world problems, a transformation that keeps a narrow distribution of ϕ's, which implies good overlap between the transformed state and state B, can accelerate the convergence of the generalized free energy perturbation calculations.

In 2009, Hahn and Then extended this idea and proposed a bidirectional formulation as a generalized Bennett Acceptance Ratio method[77]. They began with the identity

$$\frac{\rho_{A'}(\mathbf{y}(\mathbf{x}))}{\rho_B(\mathbf{y}(\mathbf{x}))} = e^{\beta[\Phi(\mathbf{x})-\Delta F]} \qquad \forall \mathbf{x} \in \Gamma_0. \tag{3.1.1.32}$$

Multiplying both sides with $\delta[W - \Phi(\mathbf{x})]\rho_B(\mathbf{y}(\mathbf{x}))$ and integrating with respect to $\mathbf{y}$, the left-hand side yields

$$\begin{aligned}\int_{\mathbf{y}(\Gamma_A)} &\delta[W - \Phi(\mathbf{x})]\rho_{A'}(\mathbf{y}(\mathbf{x}))\,\mathrm{d}\mathbf{y} \\ &= \int_{\Gamma_0} \delta[W - \Phi(\mathbf{x})]\rho_A(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= p(W|A;\mathcal{M}),\end{aligned} \tag{3.1.1.33}$$

and the right-hand side yields

$$\begin{aligned}\int_{\mathbf{y}(\Gamma_A)} &e^{\beta[\Phi(\mathbf{x})-\Delta F]}\delta[W - \Phi(\mathbf{x})]\rho_B(\mathbf{y})\,\mathrm{d}\mathbf{y} \\ &= e^{\beta[W-\Delta F]} \int_{\mathbf{y}(\Gamma_A)} \delta[W - \Phi(\mathbf{x})]\rho_B(\mathbf{y})\,\mathrm{d}\mathbf{y} \\ &= e^{\beta[W-\Delta F]} p(W|B;\mathcal{M}),\end{aligned} \tag{3.1.1.34}$$

where $p(W|A;\mathcal{M})$ ($p(W|B;\mathcal{M})$) is the probability density for the outcome of a specific value of the generalized work W in forward (reverse) direction subject to the map $\mathcal{M}$ when sampled from ρ_A (ρ_B). Therefore,

$$\frac{p(W|A;\mathcal{M})}{p(W|B;\mathcal{M})} = e^{\beta[W-\Delta F]}. \tag{3.1.1.35}$$

Bayes theorem,

$$p(W|Y)p_Y = p(Y|W)p(W), \quad \text{with } Y = A \text{ or } B \tag{3.1.1.36}$$

implies the "balance" equation

$$\begin{aligned}p_B \int p(A|W)p(W|B)\,\mathrm{d}W &= \int p(A|W)p(B|W)p(W)\,\mathrm{d}W \\ &= \int p(B|W)p(W|A)p_A\,\mathrm{d}W \\ &= p_A \int p(B|W)p(W|A)\,\mathrm{d}W,\end{aligned} \tag{3.1.1.37}$$

or $p_B \langle p(A|W)\rangle_B = p_A \langle p(B|W)\rangle_A$.

Combining Eq. 3.1.1.35 and Eq. 3.1.1.36 leads to

$$\frac{p(A|W)}{p(B|W)} = e^{\beta(W-C)} \tag{3.1.1.38}$$

with

$$C = \Delta F + \frac{1}{\beta} \ln \frac{p_B}{p_A}. \tag{3.1.1.39}$$

With the normalization condition $p(A|W) + p(B|W) = 1$, it yields

$$p(A|W) = \frac{1}{1 + e^{\beta(C-W)}} \tag{3.1.1.40}$$

and

$$p(B|W) = \frac{e^{\beta(C-W)}}{1 + e^{\beta(C-W)}} = \frac{1}{1 + e^{\beta(-C+W)}}. \tag{3.1.1.41}$$

Replacing both, the ensemble averages by sample averages and the ratio $\frac{p_B}{p_A}$ by $\frac{n_B}{n_A}$, the balance equation results in

$$\sum_{j=1}^{n_B} \frac{1}{1 + \frac{n_B}{n_A} \exp\left(\beta \widehat{\Delta F}_{AB} - \beta W_B^j\right)} = \sum_{i=1}^{n_A} \frac{1}{1 + \frac{n_B}{n_A} \exp\left(-\beta \widehat{\Delta F}_{AB} + \beta W_A^i\right)}. \tag{3.1.1.42}$$

They also suggested using Kullback-Leibler divergence to characterize the similarity between the mapped state A' and state B.

Ding and Zhang further extended this idea and integrated BAR with deep generative model and developed an efficient free energy method DeepBAR.[78]

3.1.2 Thermodynamic Integration

Thermodynamic Integration (TI) method was proposed by Kirkwood.[79]. If the free energy, A, is a continuous function of λ, the free energy difference between two states corresponding to $\lambda = 0$ and $\lambda = 1$ can be computed via

$$\Delta A = \int_0^1 \frac{\partial A(\lambda)}{\partial \lambda}\, \mathrm{d}\lambda. \tag{3.1.2.1}$$

With

$$A(\lambda) = -\beta^{-1} \ln Q(\lambda), \tag{3.1.2.2}$$

the partial derivative can be expressed as

$$\begin{aligned} \frac{\partial A(\lambda)}{\partial \lambda} =& -\beta^{-1} \left[\frac{\partial \ln Q(\lambda)}{\partial \lambda} \right] \\ =& -\frac{\beta^{-1}}{Q(\lambda)} \frac{\partial Q(\lambda)}{\partial \lambda}. \end{aligned} \tag{3.1.2.3}$$

From the definition of Q

$$Q_{NVT}(\lambda) = \frac{1}{h^{3N} N!} \iint \exp[-\beta H(\mathbf{x}, \mathbf{p}_x, \lambda)]\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}, \tag{3.1.2.4}$$

we have

$$\begin{aligned} \frac{\partial Q(\lambda)}{\partial \lambda} =& \frac{1}{h^{3N} N!} \iint \frac{\partial}{\partial \lambda} \exp[-\beta H(\mathbf{x}, \mathbf{p}_x, \lambda)]\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x} \\ =& -\frac{\beta}{h^{3N} N!} \iint \frac{\partial H(\mathbf{x}, \mathbf{p}_x, \lambda)}{\partial \lambda} \exp[-\beta H(\mathbf{x}, \mathbf{p}_x, \lambda)]\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}. \end{aligned} \tag{3.1.2.5}$$

Substituting back into the expression for $\partial A/\partial \lambda$ yields

$$\begin{aligned} \frac{\partial A(\lambda)}{\partial \lambda} =& \frac{1}{h^{3N} N!} \frac{1}{Q(\lambda)} \iint \frac{\partial H(\mathbf{x}, \mathbf{p}_x, \lambda)}{\partial \lambda} \exp[-\beta H(\mathbf{x}, \mathbf{p}_x, \lambda)]\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}, \\ =& \frac{1}{h^{3N} N!} \iint \frac{\partial H(\mathbf{x}, \mathbf{p}_x, \lambda)}{\partial \lambda} \cdot \frac{\exp[-\beta H(\mathbf{x}, \mathbf{p}_x, \lambda)]}{Q(\lambda)}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{p}_\mathbf{x}, \\ =& \left\langle \frac{\partial H(\mathbf{x}, \mathbf{p}_x, \lambda)}{\partial \lambda} \right\rangle_\lambda \end{aligned} \tag{3.1.2.6}$$

Thus, the basic TI formula is

$$\Delta A = \int_0^1 \left\langle \frac{\partial H(\mathbf{x}, \mathbf{p}_x, \lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{3.1.2.7}$$

where $\langle \cdots \rangle_\lambda$ corresponds to the ensemble average obtained using the Hamiltonian $H(\lambda)$. In practice, the ensemble of configurations can be obtained by molecular dynamics or Monte Carlo simulations. It is common practice in free energy calculations to use the coupling parameter λ for defining the

transformation from the initial state A with Hamiltonian H_A to the final state B with Hamiltonian H_B. The simplest coupling is linear transformation as

$$H(\lambda) = (1 - \lambda)H_A + \lambda H_B, \tag{3.1.2.8}$$

for which

$$\frac{\partial H(\lambda)}{\partial \lambda} = H_B - H_A. \tag{3.1.2.9}$$

However, this simple mixing scheme does not provide the optimal alchemical transformation pathway. Better schemes are possible, for instance the minimum variance path (MVP)[62] and the optimal transport pathway[80]. The accuracy of TI integral formula depends on the exactness of the numerical integration method.[70] Practically, the integrand in Eq. 3.1.2.7 needs to be evaluated over a number of discrete points λ_i, and then be summed up to give the free energy difference between $\lambda = 0$ and 1, for instance via the trapezoidal rule

$$\Delta A = \sum_{i=0}^{N-1} \frac{1}{2} \left(\left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda_i} + \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda_{i+1}} \right) (\lambda_{i+1} - \lambda_i). \tag{3.1.2.10}$$

A finite number of λ_i values between 0 and 1 are chosen and for each of them a complete molecular dynamics simulation is carried out resulting in an ensemble of configurations generated with $H(\lambda_i)$. The ensemble average of the derivative of the Hamiltonian with respect to λ is then calculated for each λ_i.

In addition to summation method, the simplest numerical integration is to evaluate the integrand at the midpoint:

$$\Delta A \simeq \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda=\frac{1}{2}} \tag{3.1.2.11}$$

This might be a good first thing to do to get some impression of what is going on, but is only accurate for very smooth or small changes.

3.1.3 Bennett Acceptance Ratio

"The thing that differentiates scientists is purely an artistic ability to discern what is a good idea, what is a beautiful idea, what is worth spending time on, and most importantly, what is a problem that is sufficiently interesting, yet sufficiently difficult, that is hasn't yet been solved, but the time for solving it has come now."

– Professor Savas Dimopoulos, Stanford University

Bennett acceptance ratio was developed by Bennett in 1976,[81] and was re-discovered by Crooks[82] for Markovian and balanced dynamics and by Shirts et al[83] using maximum likelihood over 20 years later. The Metropolis function is defined as

$$M(x) = min\{1, \exp(-x)\}, \tag{3.1.3.1}$$

which has the property

$$M(x)/M(-x) = \exp(-x). \tag{3.1.3.2}$$

If we make a trial move that keeps the same configuration $(q_1, \cdots, q_N)$ but switches the potential function from U_0 to U_1 or vice-versa, the acceptance probabilities for such a pair of trial moves must satisfy the detailed balance

$$M(U_1 - U_0)\exp(-U_0) = M(U_0 - U_1)\exp(-U_1). \tag{3.1.3.3}$$

Integrating this identity over all of configuration space and multiplying by the trivial factors Q_0/Q_0 and Q_1/Q_1, one obtains:

$$Q_0 \frac{\int M(U_1 - U_0)\exp(-U_0)\,\mathrm{d}\mathbf{q}}{Q_0} = Q_1 \frac{\int M(U_0 - U_1)\exp(-U_1)\,\mathrm{d}\mathbf{q}}{Q_1}, \tag{3.1.3.4}$$

or simply

$$\frac{Q_0}{Q_1} = \frac{\langle M(U_0 - U_1)\rangle_1}{\langle M(U_1 - U_0)\rangle_0}. \tag{3.1.3.5}$$

The physical meaning of this formula is that a Monte Carlo calculation that includes potential-switching trial moves would distribute configurations between U_1 and U_0 in the ratio of their configurational integrals.

A formula more general than Eq. 3.1.3.5 can be written as

$$\frac{Q_0}{Q_1} = \frac{Q_0}{Q_1}\frac{\int W\exp(-U_0 - U_1)\,\mathrm{d}\mathbf{q}}{\int W\exp(-U_1 - U_0)\,\mathrm{d}\mathbf{q}} = \frac{\langle W\exp(-U_0)\rangle_1}{\langle W\exp(-U_1)\rangle_0}, \tag{3.1.3.6}$$

where W is an arbitrary weighting function.

Optimization of the free energy estimate is most easily carried out in the limit of large sample sizes. Let the available data consist of n_0 statistically independent configurations from the U_0 ensemble and n_1 from the U_1 ensemble, and let the data be used in Eq. 3.1.3.6 to obtain a finite-sample estimate of the reduced free energy difference $\Delta A = A_1 - A_0 = \ln(Q_0/Q_1)$. Using the error propagation law of uncorrelated variables ($covar(x_1, x_2) = 0$),[84]

$$\delta^2 [y(x_1, x_2)] = \left(\frac{\partial y}{\partial x_1}\right)^2 \delta^2(x_1) + \left(\frac{\partial y}{\partial x_2}\right)^2 \delta^2(x_2). \tag{3.1.3.7}$$

Thus we have the variance of ΔA

$$\begin{aligned} \delta^2(\Delta A) &= \left(\frac{\partial \Delta A}{\partial Q_0}\right)^2 \delta^2 Q_0 + \left(\frac{\partial \Delta A}{\partial Q_1}\right)^2 \delta^2 Q_1 \\ &= \left(\frac{1}{Q_0}\right)^2 \delta^2 Q_0 + \left(-\frac{1}{Q_1}\right)^2 \delta^2 Q_1 \\ &= \left(\frac{1}{Q_0}\right)^2 \delta^2 Q_0 + \left(\frac{1}{Q_1}\right)^2 \delta^2 Q_1. \end{aligned} \tag{3.1.3.8}$$

With the definition of variance $\delta^2 X = \langle X^2 \rangle - \langle X \rangle^2$, we have

$$\begin{aligned} \delta^2 Q_0 &= \delta^2 \langle W \exp(-U_0) \rangle_1 \\ &= \delta^2 \left(\frac{1}{n_1} \sum_{i=1}^{n_1} W_i \exp(-U_0(i)) \right) \\ &= \sum_{i=1}^{n_1} \left(\frac{1}{n_1}\right)^2 \delta^2 (W_i \exp(-U_0(i))) \\ &= \frac{1}{n_1} \delta^2 (W_i \exp(-U_0(i))) \\ &= \frac{1}{n_1} \left\{ \left\langle [W \exp(-U_0)]^2 \right\rangle_1 - [\langle W \exp(-U_0) \rangle_1]^2 \right\} \\ &= \frac{1}{n_1} \left\{ \left\langle W^2 \exp(-2U_0) \right\rangle_1 - [\langle W \exp(-U_0) \rangle_1]^2 \right\}, \end{aligned} \tag{3.1.3.9}$$

which shows that the variance of the mean of the samples equals to the variance of the samples divided by the number of samples.

With sufficiently large sample sizes, the error of this estimate will be

nearly Gaussian, and its expected square is exactly the variance of ΔA

$$\begin{aligned}
&\delta^2(\Delta A_{est} - \Delta A) \\
\approx&\frac{\langle W^2 \exp(-2U_1)\rangle_0}{n_0[\langle W \exp(-U_1)\rangle_0]^2} + \frac{\langle W^2 \exp(-2U_0)\rangle_1}{n_1[\langle W \exp(-U_0)\rangle_1]^2} - \frac{1}{n_0} - \frac{1}{n_1} \\
=&\frac{\int \left[\frac{Q_0}{n_0}\exp(-U_1) + \frac{Q_1}{n_1}\exp(-U_0)\right] W^2 \exp(-U_0 - U_1)\,\mathrm{d}\mathbf{q}}{\left[\int W \exp(-U_0 - U_1)\,\mathrm{d}\mathbf{q}\right]^2} \\
&- \frac{1}{n_0} - \frac{1}{n_1}.
\end{aligned} \tag{3.1.3.10}$$

To minimize it with respect to W, we have

$$W = const \times \left(\frac{Q_0}{n_0}\exp(-U_1) + \frac{Q_1}{n_1}\exp(-U_0)\right)^{-1}. \tag{3.1.3.11}$$

Substituting this into Eq. 3.1.3.6 yields

$$\frac{Q_0}{Q_1} = \frac{\langle f(U_0 - U_1 + C)\rangle_1}{\langle f(U_1 - U_0 - C)\rangle_0}\exp(+C), \tag{3.1.3.12}$$

where

$$C = \ln\frac{Q_0 n_1}{Q_1 n_0}, \tag{3.1.3.13}$$

and f denotes the Fermi function

$$f(x) = \frac{1}{1 + \exp(+x)}. \tag{3.1.3.14}$$

It can also be expressed as[83]

$$n_1\langle f(U_0 - U_1 + C)\rangle_1 = n_0\langle f(U_1 - U_0 - C)\rangle_0. \tag{3.1.3.15}$$

It should be noted that Eq. 3.1.3.12 is true for any C, which is actually a shift for one of the potential function. But the particular value specified in Eq. 3.1.3.13 minimizes the expected square error given the finite numbers (n_0 and n_1) of samples.

The variance of ΔA can be obtained by substituting Eq. 3.1.3.11 into Eq. 3.1.3.10, and is

$$\begin{aligned}
\delta^2\Delta A =&\frac{\langle f^2(U_1 - U_0 - C)\rangle_0}{n_0\langle f(U_1 - U_0 - C)\rangle_0^2} + \frac{\langle f^2(U_0 - U_1 + C)\rangle_1}{n_1\langle f(U_0 - U_1 + C)\rangle_1^2} - \frac{1}{n_0} - \frac{1}{n_1} \\
=&\left(\int \frac{n_0 n_1 \rho_0 \rho_1}{n_0\rho_0 + n_1\rho_1}\,\mathrm{d}\mathbf{q}\right)^{-1} - \frac{n_0 + n_1}{n_0 n_1},
\end{aligned} \tag{3.1.3.16}$$

in which $\rho_i = \exp(-U_i)/Q_i$ is the probability.

It is worth emphasizing that Bennett acceptance ratio is asymptotically unbiased, and no other asymptotically unbiased estimator has lower asymptotic mean-squared error. However, it is not clear whether its behavior is always better than other estimators with finite sample sizes.

Shirts et al showed that BAR can be interpreted in terms of the maximum likelihood estimate of the free energy difference given a set of work values in the forward and reverse directions.[83] Starting from (refer to Appendix C for the proof)

$$\ln\left[\frac{P(W|F)}{P(-W|R)}\right] = \beta(W - \Delta F), \tag{3.1.3.17}$$

where $P(W|F)$ ($P(W|R)$) is probability distribution for the work from the two states in the forward (reverse) direction, which can also be thought as the conditional probability of a work value given that it is a forward (reverse) measurement. To simplify the notation, the work W from the reverse direction will be replaced by $-W$. Using the fact that $P(F|W) + P(R|W) = 1$, the ratio can be rewritten as

$$\frac{P(W|F)}{P(R|R)} = \frac{P(F|W)P(R)}{P(R|W)P(F)} = \frac{P(F|W)}{1 - P(F|W)}\frac{P(R)}{P(F)}. \tag{3.1.3.18}$$

It can be realized that $P(R)/P(F) = n_R/n_F$, where n_F (n_R) is the number of forward (reverse) measurement. Defining $M = \ln n_F/n_R$, Eq. 3.1.3.17 can be rewritten as

$$\ln\frac{P(F|W)}{1 - P(F|W)} = (M + W - \Delta F), \tag{3.1.3.19}$$

which leads to

$$P(F|W_i) = \frac{1}{1 + \exp[-(M - W_i - \Delta F)]}, \tag{3.1.3.20}$$

$$P(R|W_i) = \frac{1}{1 + \exp[M - W_i - \Delta F]} \tag{3.1.3.21}$$

for a single work measurement W_i, given a value of the free energy difference ΔF.

Given a value for ΔF, the overall likelihood L becomes

$$L(\Delta F) = \prod_{i=1}^{n_F} P(F|W_i) \prod_{i=1}^{n_R} P(R|W_i). \tag{3.1.3.22}$$

The most likely value of ΔF is the value that maximizes the (log) likelihood, therefore we have

$$\begin{aligned} 0 = \frac{\partial \log L(\Delta F)}{\partial \Delta F} = &- \sum_{i=1}^{n_F} \frac{1}{1 + \exp\left[M + W_i - \Delta F\right]} \\ &+ \sum_{j=1}^{n_R} \frac{1}{1 + \exp\left[-(M + W_j - \Delta F)\right]}, \end{aligned} \tag{3.1.3.23}$$

which equivalent to the BAR method.

3.1.4 Weighted Histogram Analysis Method

The weighted histogram analysis method is a generalization of the histogram method developed by Ferrenberg and Swendsen.[85]

Weighted Histogram Analysis Method for Parallel Tempering

The following derivation quite follows Ref. [86]. One of the central quantities in statistical mechanics is configurational integral Z, which in textbook is often written as

$$Z = \int \exp\left(-\beta U(\mathbf{R})\right) \mathrm{d}\mathbf{R}. \tag{3.1.4.1}$$

This is an integral in coordinate space. It also can be written as an integral in energy space

$$Z = \int \Omega(U) \exp\left(-\beta U\right) \mathrm{d}U, \tag{3.1.4.2}$$

where $\Omega(U)$ is density of states and $\Omega(U)\Delta U$ is the number of states in the region $U - \Delta U/2 < U < U + \Delta U/2$. Accordingly, the statistical expectation of an operator $\mathbf{A}$ can be calculated by

$$\langle \mathbf{A} \rangle = \frac{\int \mathbf{A}(U)\Omega(U) \exp\left(-\beta U\right) \mathrm{d}U}{\int \Omega(U) \exp\left(-\beta U\right) \mathrm{d}U}, \tag{3.1.4.3}$$

where

$$\mathbf{A}(U') = \frac{\int \delta(U(\mathbf{R}) - U')\mathbf{A}(\mathbf{R}) \, \mathrm{d}\mathbf{R}}{\int \delta(U(\mathbf{R}) - U') \, \mathrm{d}\mathbf{R}}, \tag{3.1.4.4}$$

is the average of $\mathbf{A}$ over all the samples with an energy U'. Therefore, the core objective is to calculate $\Omega(U)$.

Suppose we have one trajectory with N snapshots denoted as $\{\mathbf{R}_n\}$. We then discretize the energy space into M bins with width ΔU, and count the number of snapshots into each bin. For convenience, we define $\psi_m(U)$ as

$$\psi_m(U) = \begin{cases} 1 & \text{if } U \in [U_m - \Delta U/2, U_m + \Delta U/2) \\ 0 & \text{otherwise} \end{cases} \tag{3.1.4.5}$$

Then the histogram for the mth energy bin is

$$H_m = \sum_{n=1}^{N} \psi_m(U(\mathbf{R}_n)) = N \cdot \frac{1}{N} \sum_{n=1}^{N} \psi_m(U(\mathbf{R}_n)) = N \cdot \langle \psi_m \rangle, \tag{3.1.4.6}$$

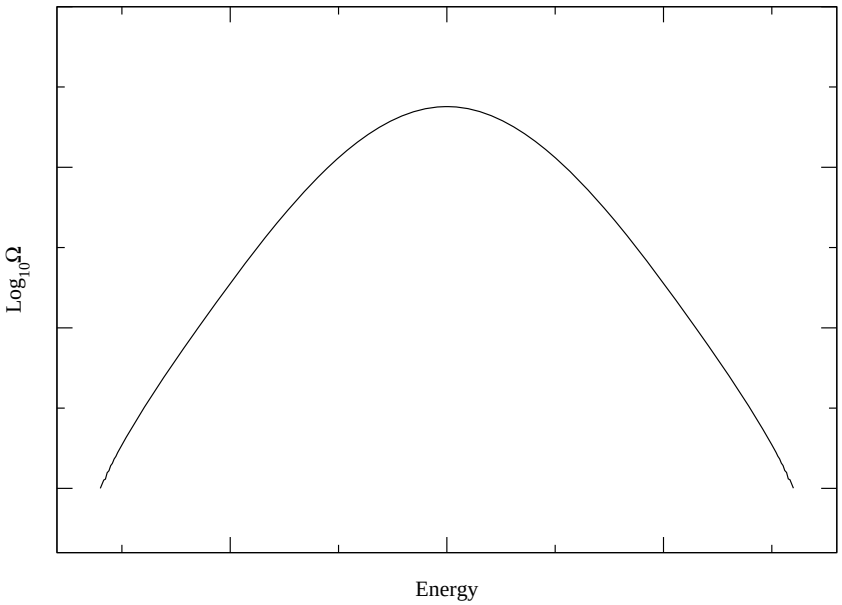

Figure 3.2: Density of states of a 2D periodic Ising model

with variances (see Appendix A)

$$\begin{aligned}\delta^2 H_m =& N^2 \delta^2(\langle \psi_m \rangle) \\ =& g_m N \left(\left\langle {\psi_m}^2 \right\rangle - \langle \psi_m \rangle^2 \right) \\ =& g_m N \left(\langle \psi_m \rangle - \langle \psi_m \rangle^2 \right) \\ =& g_m H_m \left(1 - \frac{H_m}{N} \right). \end{aligned} \tag{3.1.4.7}$$

A sample histogram is shown in Fig. 3.2.

The ratio of the histogram H_m to the total number of snapshots N divided by the bin width ΔU can be approximately taken as the probability of states in this bin, i.e.,

$$\frac{\Omega_m \exp\left(-\beta U_m\right)}{Z} \approx \frac{H_m}{N \Delta U}. \tag{3.1.4.8}$$

Therefore,

$$\begin{aligned}\Omega_m =& \frac{1}{\Delta U} \cdot \frac{H_m}{N} \cdot \frac{Z(\beta)}{\exp\left(-\beta U_m\right)} \\ =& \frac{H_m}{N \Delta U \exp\left[f - \beta U_m\right]}, \end{aligned} \tag{3.1.4.9}$$

and variances

$$\delta^2 \Omega_m = \frac{\delta^2 H_m}{\left(N \Delta U \exp\left[f - \beta U_m\right]\right)^2}, \tag{3.1.4.10}$$

in which we have defined a dimensionless free energy $f = -\ln Z(\beta)$.

Practically, we may run multiple (K) trajectories using, for example, replica exchange molecular dynamics simulations. For each trajectory (index k), we have unique estimators for the histogram H_{mk}, the density of states Ω_{mk} and their variances $\delta^2 H_{mk}$ and $\delta^2 \Omega_{mk}$ being

$$H_{mk} = \sum_{n=1}^{N_k} \psi_m(U(\mathbf{R}_{kn})), \tag{3.1.4.11}$$

$$\delta^2 H_{mk} = g_{mk} H_{mk} \left(1 - \frac{H_{mk}}{N_k}\right), \tag{3.1.4.12}$$

$$\Omega_{mk} = \frac{H_{mk}}{N_k \Delta U \exp\left[f_k - \beta_k U_m\right]}, \tag{3.1.4.13}$$

and

$$\delta^2 \Omega_{mk} = \frac{\delta^2 H_{mk}}{\left(N_k \Delta U \exp\left[f_k - \beta_k U_m\right]\right)^2}, \tag{3.1.4.14}$$

The optimum estimator of the density of states from all the simulations is

$$\Omega_m = \frac{\sum\limits_{k=1}^{K} \left[\delta^2 \Omega_{mk}\right]^{-1} \Omega_{mk}}{\sum\limits_{k=1}^{K} \left[\delta^2 \Omega_{mk}\right]^{-1}}, \tag{3.1.4.15}$$

which is the weighted average of density of states of all the trajectories with the weight reversely proportional to the variances (see Appendix B).

To make the expression simpler, here we take some approximations. First, normally the energy space is split into a large number of bins. The histogram in each bin is much smaller than the total number of snapshots, i.e. $H_{mk} \ll N_k$. With this approximation, we have

$$\delta^2 H_{mk} \approx g_{mk} H_{mk}. \tag{3.1.4.16}$$

The expectation of H_{mk} can be related to the optimum estimator of the density of states, i.e.

$$\overline{H_{mk}} = N_k \Delta U \Omega_m \exp\left(f_k - \beta_k U_m\right). \tag{3.1.4.17}$$

Then we have

$$\delta^2 H_{mk} = g_{mk} N_k \Delta U \Omega_m \exp\left(f_k - \beta_k U_m\right) \tag{3.1.4.18}$$

and

$$\delta^2 \Omega_{mk} = \frac{\Omega_m}{{g_{mk}}^{-1} N_k \Delta U \exp\left(f_k - \beta_k U_m\right)}. \tag{3.1.4.19}$$

Taking Eq. 3.1.4.13 and Eq. 3.1.4.19 into Eq. 3.1.4.15, we find

$$\Omega_m = \frac{\sum_{k=1}^{K} {g_{mk}}^{-1} H_{mk}}{\sum_{k=1}^{K} {g_{mk}}^{-1} N_k \Delta U \exp\left(f_k - \beta_k U_m\right)}, \tag{3.1.4.20}$$

in which

$$f_k = -\ln \int \Omega(U) \exp\left(-\beta_k U\right) \mathrm{d}U = -\ln \sum_{m=1}^{M} \Omega_m \Delta U \exp\left(-\beta_k U_m\right). \tag{3.1.4.21}$$

Obviously, Eq. 3.1.4.20 and Eq. 3.1.4.21 must be solved iteratively. Applying the error propagation rule to Eq. 3.1.4.20 and using Eq. 3.1.4.18, the variance of Ω_m is given by

$$\delta^2 \Omega_m = \frac{\Omega_m}{\sum_{k=1}^{K} {g_{mk}}^{-1} N_k \Delta U \exp\left(f_k - \beta_k U_m\right)}. \tag{3.1.4.22}$$

Using the density of states and its variance, we can estimate the expectation of any configuration function $A(\mathbf{R})$ at any inverse temperature β

$$\langle A \rangle_\beta \approx \frac{\sum_{m=1}^{M} \Omega_m \Delta U \exp\left(-\beta U_m\right) A_m}{\sum_{m=1}^{M} \Omega_m \Delta U \exp\left(-\beta U_m\right)}, \tag{3.1.4.23}$$

where

$$A_m = \frac{\int \mathrm{d}\mathbf{R}\, A(\mathbf{R}) \psi_m(U(\mathbf{R}))}{\int \mathrm{d}\mathbf{R}\, \psi_m(U(\mathbf{R}))}. \tag{3.1.4.24}$$

Using histograms of bin m from all the simulations and defining $H_m = \sum_{k=1}^{K} H_{mk}$, an estimator of A_m denoted as $\hat{A}_m$ can be calculated as

$$\hat{A}_m = {H_m}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_m(U(\mathbf{R}_{kn})) A(\mathbf{R}_{kn}). \tag{3.1.4.25}$$

Taking Eq. 3.1.4.25 into Eq. 3.1.4.23, we obtain an estimator of $\hat{A}(\beta)$

$$\hat{A}(\beta) = \frac{\sum_{m=1}^{M} \Omega_m \Delta U \exp(-\beta U_m) {H_m}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_m(U(\mathbf{R}_{kn})) A(\mathbf{R}_{kn})}{\sum_{m=1}^{M} \Omega_m \Delta U \exp(-\beta U_m)} \tag{3.1.4.26}$$

$$= \frac{\sum_{m=1}^{M} \Omega_m \Delta U \exp(-\beta U_m) {H_m}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_m(U(\mathbf{R}_{kn})) A(\mathbf{R}_{kn})}{\sum_{m=1}^{M} \Omega_m \Delta U \exp(-\beta U_m) {H_m}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_m(U(\mathbf{R}_{kn}))} \tag{3.1.4.27}$$

$$= \frac{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta) A_{kn}}{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta)}, \tag{3.1.4.28}$$

where the per-configuration weights $w_{kn}(\beta)$ is given by

$$w_{kn}(\beta) = \sum_{m=1}^{M} {H_m}^{-1} \psi_m(U(\mathbf{R}_{kn})) \Omega_m \exp(-\beta U_m) \tag{3.1.4.29}$$

Weighted Histogram Analysis Method From Minimizing Statistical Error

In this section, the "traditional" derivation method of WHAM are briefly reviewed.[87] In the WHAM, the goal is to get an optimal unbiased probability distribution $\rho_0(\eta)$, where η is a series of discretized histogram bins indexed by $j = 1, 2, 3, ..., M$ along a certain reaction coordinate. WHAM can be used to analyze the Umbrella Sampling (US) simulations, where a set of simulations indexed by i or $k = 1, 2, 3, ..., S$ are performed with a series of biasing potentials added on the reaction coordinate η. To consider a reference molecular system with the potential energy $U_0(\mathbf{x})$, where $\mathbf{x}$ is the set of atomic coordinates. The reaction coordinate η is a function of the atomic coordinates, i.e. $\eta(\mathbf{x})$. To suppose that the ith molecular simulation has been performed using potential energy function

$$U_i^{(b)}(\eta) = U_0(\mathbf{x}) + W_i(\eta(\mathbf{x})), \tag{3.1.4.30}$$

where $W_i(\eta(\mathbf{x}))$ is the biasing potential added on the reaction coordinate η, e.g. $W_i(\eta) = \frac{1}{2} k_i (\eta - \eta_i)^2$ in a harmonic form. From these simulations a set of normalized biased probability distributions $\rho_i^{(b)}(\eta)$ can be obtained.

$$\rho_i^{(b)}(\eta) = \frac{e^{-\beta U_i^{(b)}(\eta)}}{Q_i^{(b)}}, \tag{3.1.4.31}$$

where $Q_i^{(b)} = \int e^{-\beta U_i^{(b)}(\eta)}\,\mathrm{d}\eta = e^{-\beta f_i^{(b)}}$ and $f_i^{(b)}$ is the biased free energy. The corresponding unnormalized unbiased probability distribution $\rho_i^{(u)}(\eta)$ from the ith simulation is defined as,

$$\rho_i^{(u)}(\eta) = e^{\beta[W_i(\eta) - f_i^{(b)}]}\rho_i^{(b)}(\eta) \tag{3.1.4.32}$$

In the following, the free energy $f_i^{(b)}$ is assumed to be known. It has been shown that in the WHAM method, the total normalized unbiased probability distribution $\rho_0(\eta)$ can be obtained by a linear η-dependent combination of the unbiased histograms $\rho_i^{(u)}(\eta)$

$$\rho_0(\eta) = C\sum_{i=1}^{S} p_i(\eta)\rho_i^{(u)}(\eta), \tag{3.1.4.33}$$

where C is the normalization factor. p_i is the weight to be optimized, which is under a constraint that

$$\sum_{i=1}^{S} p_i(\eta) = 1. \tag{3.1.4.34}$$

These weights are chosen so as to minimize the statistical error made on the total unbiased probability distribution $\rho_0(\eta)$, that is, for any given value of η,

$$\frac{\partial(\sigma^2[\rho_0(\eta)])}{\partial p_i(\eta)} = 0. \tag{3.1.4.35}$$

It can be easily found that $\rho_0(\eta)$ satisfies

$$\begin{aligned}
\rho_0(\eta) =& C\sum_{i=1}^{S} \frac{N_i e^{-\beta[W_i(\eta) - f_i^{(b)}]}}{\sum_{k=1}^{S} N_k e^{-\beta[W_k(\eta) - f_k^{(b)}]}}\rho_i^{(u)}(\eta) \\
=& C\sum_{i=1}^{S} \frac{N_i}{\sum_{k=1}^{S} N_k e^{-\beta[W_k(\eta) - f_k^{(b)}]}}\rho_i^{(b)}(\eta) \\
=& C\frac{\sum_{i=1}^{S} N_i\rho_i^{(b)}(\eta)}{\sum_{k=1}^{S} N_k e^{-\beta[W_k(\eta) - f_k^{(b)}]}},
\end{aligned} \tag{3.1.4.36}$$

where $\rho_i^{(b)}(\eta)$ can be written as a δ function,

$$\rho_i^{(b)}(\eta) \equiv \frac{1}{N_i}\sum_{l=1}^{N_i}\delta(\eta - \eta_{i,l}), \tag{3.1.4.37}$$

where $\eta_{i,l}$ is the reaction coordinates of the lth configuration in the ith biased simulation .

Until now, the treatment assumes that the free energy parameters $f_i^{(b)}$ are known. In fact, these parameters can be obtained self-consistently. Indeed, the definition of the free energy $f_i^{(b)}$ is,

$$\begin{aligned} e^{-\beta f_i^{(b)}} &= \int e^{-\beta U_i^{(b)}(\eta)}\,\mathrm{d}\eta \\ &= \int \rho_0(\eta) e^{-\beta W_i(\eta)}\,\mathrm{d}\eta \\ &= C \int \frac{\sum_{i=1}^{S} N_i \rho_i^{(b)}(\eta)}{\sum_{k=1}^{S} N_k e^{-\beta[W_k(\eta)-f_k^{(b)}]}} e^{-\beta W_i(\eta)}\,\mathrm{d}\eta \end{aligned} \tag{3.1.4.38}$$

The set of parameters $f_i^{(b)}$ appear on both sides of Eq. 3.1.4.38, which must be solved iteratively with an initial guess of $f_i^{(b)}$ until convergence is reached. The unbiased free energy corresponding to the histogram can be calculated by

$$f_0(\eta) = -\beta^{-1} \ln \rho_0(\eta) \tag{3.1.4.39}$$

with W in Eq. 3.1.4.38 being 0. The constant C in Eq. 3.1.4.36 is irrelevant, which only causes a constant shift to the free energy profiles. To get rid of it, one may subtract the offset constant $f_0(\eta_1)$ from all the $f_0(\eta_j)$.

Weighted Histogram Analysis Method From Maximum Likelihood

The following derivation quite follows Ref. [88], in which maximum likelihood principle is utilized. Suppose we have performed K simulations, each at a different inverse temperature β_k and possibly with different biasing potential $w_k(\mathbf{R})$. We then discretize the 2D plane spanned by the coordinate and unbiased potential energy into bins, each characterized by $\mathbf{R}_j$ and E_h. To make the following derivation cleaner, we map the 2D bins to one dimensional series with index $l, l = 1, \ldots, L$. Next, we construct histograms for bins using all the samples from the simulations. The probability of finding the system in bin l during the kth simulation can be written as

$$p_{k,l} = f_k c_{k,l} p_l^0, \tag{3.1.4.40}$$

in which p_l^0 is the (simulation-independent) unbiased probability,

$$\begin{aligned} c_{k,l} &= \exp\left[-\beta_k \left(E_l + w_{k,l}\right) + \beta_0 E_l\right] \\ &= \exp\left[-\left(\beta_k - \beta_0\right) E_l\right] \exp\left(-\beta_k w_{k,l}\right) \end{aligned} \tag{3.1.4.41}$$

is the bias factor, E_l is the unbiased energy of bin l, $f_k = \left\{\sum_l c_{k,l} p_l^0\right\}^{-1}$ is the normalization factor. If we expand the expressions for $c_{k,l}$ and p_l^0, we

find

$$f_k \approx \left(\sum_l \exp\left[-\beta_k \left(E_l + w_{k,l}\right) + \beta_0 E_l\right] \frac{\exp\left(-\beta_0 E_l\right)}{\sum\limits_j \exp\left(-\beta_0 E_j\right)} \right)^{-1}$$
$$= \left(\frac{\sum\limits_l \exp\left[-\beta_k (E_l + w_{k,l})\right]}{\sum\limits_j \exp\left(-\beta_0 E_j\right)} \right)^{-1}$$
$$\approx \frac{Z_0}{Z_k}, \tag{3.1.4.42}$$

which is approximately the ratio of two configurational integrals.

It is worth emphasizing that the biasing potential can be multiple dimensional as, for instance, in a two-dimensional umbrella sampling. If the biasing is only in temperature-space as in replica exchange molecular dynamics

$$c_{k,l} = \exp\left[-\left(\beta_k - \beta_0\right) E_l\right], \tag{3.1.4.43}$$

while if the biasing is only in potential energy as in umbrella sampling

$$c_{k,l} = \exp\left(-\beta_0 w_{k,l}\right). \tag{3.1.4.44}$$

If we assume that each count in each histogram is independent, then the likelihood of observing the kth histogram distribution is given by the multinomial distribution

$$P(n_{k,1}, n_{k,2}, \ldots, n_{k,L} | p_{k,1}, p_{k,2}, \ldots, p_{k,L}) =$$
$$\frac{\left(\sum\limits_l n_{k,l}\right)!}{\prod\limits_l n_{k,l}!} \prod_{l=1}^{L} \left(p_{k,l}\right)^{n_{k,l}} \propto \prod_{l=1}^{L} \left(f_k c_{k,l} p_l^0\right)^{n_{k,l}}. \tag{3.1.4.45}$$

For all K simulations, the likelihood is the product of multinomial

$$P(n_{1,1}, \ldots, n_{1,L}; \ldots; n_{K,1}, \ldots, n_{K,L} | p_1^0, \ldots, p_L^0) \propto$$
$$\prod_{k=1}^{K} \prod_{l=1}^{L} \left(f_k c_{k,l} p_l^0\right)^{n_{k,l}}, \tag{3.1.4.46}$$

where the likelihood is conditional only on the unbiased probabilities p_l^0, since the bias factors $c_{k,l}$ are known parameters, and the normalization constants f_k are known conditional on p_l^0. The maximum likelihood estimate of the unbiased probabilities can be found by maximizing P in Eq. 3.1.4.46 with respect to $p_1^0, \ldots, p_L^0$ and are given by solutions of the simultaneous nonlinear equations

$$p_l^0 = \frac{\sum\limits_{k=1}^{K} n_{k,l}}{\sum\limits_{k=1}^{K} N_k f_k c_{k,l}} \text{ (for each } l) \tag{3.1.4.47}$$

and

$$f_k = \left\{ \sum_l c_{k,l} p_l^0 \right\}^{-1}, \tag{3.1.4.48}$$

where N_k is the total number of counts in the kth histogram. This is equivalent to the maximum *a posteriori* (MAP) estimation with a uniform *prior* probability for p_l^0[89].

Binless Weighted Histogram Analysis Method

The following derivation quite follows Ref. [90]. Let us start with the definition of a generalized energy function u and its corresponding coefficient θ. For instance, for canonical ensemble, $u = U(x)$ is the potential energy function, and $\theta = \beta$ is the inverse temperature. For isothermal grand-canonical ensemble, $u = (U(x), N)$, and $\theta = (\beta, \beta\mu)$, in which N is the number of particles and μ is the chemical potential. For temperature replica exchange molecular dynamics, $u = U(x)$ and $\theta_k = \beta_k$ for the kth replica. For umbrella sampling, $u = (U_0(x), \omega_1(x), \omega_2(x), \ldots, \omega_d(x))$, where $U_0(x)$ is the unbiased Hamiltonian and $\omega_k(x)$ is the biasing potential in window k. Correspondingly, $\theta_k = (\beta, 0, \ldots, 0, \beta, 0, \ldots, 0)$, in which all the elements are zero except for the first and the $(k+1)$th element.

Assume that simulations are conducted at m coefficient vectors θ_r $(r = 1, \ldots, m)$ and with the same energy vector $u(x)$. (Note that in this notation the dimensionality, d, of the θ and u vectors and the number of simulations, m, are, in general, distinct. For instance, for temperature replica exchange molecular dynamics as shown above, $d = 1$, while m is the number of replicas.) Denoted by $\{x_{ji} : i = 1, \ldots, n_j\}$ the set of configurations of size n_j from the the jth simulation, and denoted by $u_{ji} = u(x_{ji})$ the corresponding generalized energy vectors. The total sample size is $n = \sum_{j=1}^m n_j$. Now, consider a generalized ensemble whose Boltzmann probability density function is

$$\frac{1}{Z_\theta} e^{-\theta^{\mathrm{T}} u(x)}, \tag{3.1.4.49}$$

where

$$Z_\theta = \int e^{-\theta^{\mathrm{T}} u(x)} \, \mathrm{d}x \tag{3.1.4.50}$$

is the generalized configurational integral in physics or the normalization constant in statistics, and the superscript T is the transpose operator. The induced probability density function of $u(x)$ at θ is

$$\frac{1}{Z_\theta} \Omega(u) e^{-\theta^{\mathrm{T}} u}, \tag{3.1.4.51}$$

where $\Omega(u)$, formally defined as

$$\Omega(u) = \int \delta(u(x) - u) \, \mathrm{d}x, \tag{3.1.4.52}$$

is a generalized density of states, which does not depend on θ. The generalized configurational integral can also be determined from $\Omega(u)$ as

$$Z_\theta = \int \Omega(u) e^{-\theta^{\mathrm{T}} u} \, \mathrm{d}u. \tag{3.1.4.53}$$

As we have shown that the WHAM method involves constructing a histogram, $H_r(u)$, from each sample $\{u_{ri} : i = 1, \ldots, n_r\}$, which $H_r(u)$ indicates the number of observations falling into a bin about u, for example, an interval or a rectangle if $u(x)$ is one or two-dimensional. Then $\Omega(u)$ is estimated by

$$\hat{\Omega}(u)\Delta u = \frac{\sum_{r=1}^{m} H_r(u)}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{-\theta_r^{\mathrm{T}} u}}, \tag{3.1.4.54}$$

where the configurational integrals $(Z_{\theta_1}, \ldots, Z_{\theta_m})$ are defined by self-consistency according to Eq. 3.1.4.54

$$\hat{Z}_{\theta_k} = \sum_u \frac{\sum_{r=1}^{m} H_r(u)}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta_k - \theta_r)^{\mathrm{T}} u}} \quad (k = 1, \ldots, m), \tag{3.1.4.55}$$

where the summation $\sum_u$ is taken over all possible bins centered at u of size Δu. Also, the configurational integral Z_θ at any other parameter value is estimated by

$$\hat{Z}_\theta = \sum_u \frac{\sum_{r=1}^{m} H_r(u)}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta - \theta_r)^{\mathrm{T}} u}}. \tag{3.1.4.56}$$

We can take

$$\frac{1}{\hat{Z}_\theta} \frac{\sum_{r=1}^{m} H_r(u)}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta - \theta_r)^{\mathrm{T}} u}} \tag{3.1.4.57}$$

as the weight of bin u under condition θ.

Now, let $h(u)$ be a function of u, and denoted by $\langle h \rangle_\theta$ the expectation of $h(u)$. The WHAM estimate $\hat{h}_\theta$ for $\langle h \rangle_\theta$ is

$$\hat{h}_\theta = \frac{1}{\hat{Z}_\theta} \sum_u h(u) \frac{\sum_{r=1}^{m} H_r(u)}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta - \theta_r)^{\mathrm{T}} u}}. \tag{3.1.4.58}$$

It is interesting to note that the summation over bins in Eq. 3.1.4.58 can be equivalently expressed in terms of a weighted average over observations

$$\hat{h}_\theta = \sum_{ji} h(u_{ji}^b) F_{ji}(\theta), \tag{3.1.4.59}$$

where u_{ji}^b is a representative generalized energy of the bin containing u_{ji}, F_{ji} is the "WHAM weight" of u_{ji} that, by comparing Eqs. 3.1.4.58 and 3.1.4.59, is defined as

$$F_{ji}(\theta) = \frac{\hat{Z}_\theta^{-1}}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta - \theta_r)^{\mathrm{T}} u_{ji}^b}} = \frac{1}{\hat{Z}_\theta} e^{-\theta^T u_{ji}^b} G_{ji} \tag{3.1.4.60}$$

and

$$G_{ji} = \frac{1}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{-\theta_r^{\mathrm{T}} u_{ji}^b}} \tag{3.1.4.61}$$

is the θ-dependent component of the WHAM weight $F_{ji}(\theta)$ for each observation.

Equation 3.1.4.59 states that the expectation value of any observable can be obtained by attaching a statistical weight $F_{ji}(\theta)$ to each observation u_{ji} which depends on the bin to which it is assigned. An obvious simplification is to express the WHAM estimate of $\langle h \rangle_\theta$ and the WHAM weights (Eq. 3.1.4.60) in terms of the actual observation u_{ji} rather than their closest bin representative u_{ji}^b.

To understand binless WHAM, it is useful to introduce the concept of the measure G defined by

$$\mathrm{d}G = \Omega(u)\,\mathrm{d}u, \tag{3.1.4.62}$$

that is, $G(A) = \int_A \Omega(u)\,\mathrm{d}u$ for every measurable set A of u. Informally, this equation says that for an infinitesimal bin about u of size $\mathrm{d}u$, the weight assigned under G is $\Omega(u)\,\mathrm{d}u$. Thereafter G is called the measure of states. The probability distribution of $u(x)$, F_θ, is related to G as

$$\mathrm{d}F_\theta = \frac{1}{Z_\theta} e^{-\theta^{\mathrm{T}} u} \Omega(u)\,\mathrm{d}u = \frac{1}{Z_\theta} e^{-\theta^{\mathrm{T}} u}\,\mathrm{d}G, \tag{3.1.4.63}$$

that is

$$F_\theta(A) = \frac{1}{Z_\theta} \int_A e^{-\theta^{\mathrm{T}} u}\,\mathrm{d}G \tag{3.1.4.64}$$

for every measurable set A of u. The configurational integral can then be expressed as

$$Z_\theta = \int e^{-\theta^{\mathrm{T}} u}\,\mathrm{d}G. \tag{3.1.4.65}$$

The pooled data $\{u_{ji} : i = 1, \ldots, n_j, j = 1, \ldots, m\}$ can be regarded as an approximate sample from the mixture distribution, F_*, whose components are $(F_{\theta_1}, \ldots, F_{\theta_m})$ with proportions $(n_1/n, \ldots, n_m/n)$. F_* is related to G as

$$\mathrm{d}F_* = \left\{ \sum_{r=1}^{m} \frac{n_r}{n} Z_{\theta_r}^{-1} e^{-\theta_r^{\mathrm{T}} u} \right\} \Omega(u)\,\mathrm{d}u = \left\{ \sum_{r=1}^{m} \frac{n_r}{n} Z_{\theta_r}^{-1} e^{-\theta_r^{\mathrm{T}} u} \right\} \mathrm{d}G \tag{3.1.4.66}$$

For an infinitesimal bin about u of size $\mathrm{d}u$, the probability assigned under F_* is the expression in the curly brackets times the weight assigned under G. Dividing both sides of Eq. 3.1.4.66 by the quantity in the curly brackets gives

$$\mathrm{d}G = \left\{ \sum_{r=1}^{m} \frac{n_r}{n} Z_{\theta_r}^{-1} e^{-\theta_r^{\mathrm{T}} u} \right\}^{-1} \mathrm{d}F_*. \tag{3.1.4.67}$$

For an infinitesimal bin about u of size $\mathrm{d}u$, the weight assigned under G is the inverse of the quantity in the curly brackets times the probability assigned under F_*.

Relationship (3.1.4.67) can be used for estimating G from the pooled data by importance weighting. Recall that the pooled data form an approximate sample from F_*. Then F_* can be estimated by the empirical distribution $\hat{F}_*$ for which each observation u_{ji} is assigned the probability n^{-1}. By Eq. 3.1.4.67, the resulting estimator $\hat{G}$ is a discrete measure for which each observation u_{ji} is assigned the weight

$$\hat{G}(u_{ji}) = \frac{1}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{-\theta_r^{\mathrm{T}} u_{ji}}}, \tag{3.1.4.68}$$

where

$$\begin{aligned}\hat{Z}_{\theta_k} =& \sum_{j=1}^{m}\sum_{i=1}^{n_j} e^{-\theta_k^{\mathrm{T}} u_{ji}} \hat{G}(u_{ji}) \\ =& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{1}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta_k-\theta_r)^{\mathrm{T}} u_{ji}}} \quad (k = 1, \ldots, m).\end{aligned} \tag{3.1.4.69}$$

Formulas 3.1.4.68 and 3.1.4.69 provide a binless extension of Eq. 3.1.4.54 and 3.1.4.55 in WHAM.

Again, the configurational integral Z_θ at any other parameter value is estimated by

$$\begin{aligned}\hat{Z}_\theta =& \sum_{j=1}^{m}\sum_{i=1}^{n_j} e^{-\theta^{\mathrm{T}} u_{ji}} \hat{G}(u_{ji}) \\ =& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{1}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta-\theta_r)^{\mathrm{T}} u_{ji}}}.\end{aligned} \tag{3.1.4.70}$$

The expectation $\langle h \rangle_\theta$ is by definition $Z_\theta^{-1} \int h(u) e^{-\theta^{\mathrm{T}} u} \, \mathrm{d}G$ and hence estimated by

$$\begin{aligned}&\frac{1}{\hat{Z}_\theta} \sum_{j=1}^{m}\sum_{i=1}^{n_j} h(u_{ji}) e^{-\theta^{\mathrm{T}} u_{ji}} \hat{G}(u_{ji}) \\ =& \frac{1}{\hat{Z}_\theta} \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{h(u_{ji})}{\sum_{r=1}^{m} n_r \hat{Z}_{\theta_r}^{-1} e^{(\theta-\theta_r)^{\mathrm{T}} u_{ji}}}.\end{aligned} \tag{3.1.4.71}$$

Formulas 3.1.4.70 and 3.1.4.71 provide a binless extension of Eqs. 3.1.4.56 and 3.1.4.58 in WHAM.

Bayesian Reconstruction of the Density of States

Habeck reformulated the WHAM equation for the density of states from Bayesian view in 2007.[91]

To make this section self-explained, we begin with the definition of the density of states

$$g(E) = \int \mathrm{d}x\, \delta[E - E(x)], \tag{3.1.4.72}$$

from which the configurational integral can be calculated

$$Z(\beta) = \int \mathrm{d}E\, g(E) e^{-\beta E} \tag{3.1.4.73}$$

for a continuous system or

$$Z(\beta) = \sum_i g(E_i) e^{-\beta E_i} \tag{3.1.4.74}$$

for a discrete representation.

Collected in M simulations at inverse temperatures $\beta_1, \ldots, \beta_M$ with sample sizes $N_1, \ldots, N_M$, respectively, the complete data set can be written as $D = \{(\beta_i, E_{i1}, \ldots, E_{iN}), i = 1, \ldots, M\}$. The jth sample in the ith simulation has a probability

$$p(E(x_{ij})|\beta_i) = g(E(x_{ij})) e^{-\beta_i E(x_{ij})} / Z(\beta_i). \tag{3.1.4.75}$$

If all the samples are statistically independent, the probability of the data set is

$$\begin{aligned} L(g) &= \prod_{i=1}^{M} \prod_{j=1}^{N_i} \left\{ g(E(x_{ij})) e^{-\beta_i E(x_{ij})} / Z(\beta_i) \right\} \\ &= \prod_{i=1}^{M} \prod_{j=1}^{N_i} \left\{ g(E(x_{ij})) e^{-\beta_i E(x_{ij})} \right\} / \left\{ \prod_{i=1}^{M} \prod_{j=1}^{N_i} Z(\beta_i) \right\} \\ &= \prod_{i=1}^{M} \prod_{j=1}^{N_i} g(E(x_{ij})) e^{-\beta_i E(x_{ij})} / \prod_{i=1}^{M} [Z(\beta_i)]^{N_i}. \end{aligned} \tag{3.1.4.76}$$

By introducing $H_i(E) = \sum_j \delta(E - E(x_{ij}))$ and $H(E) = \sum_i H_i(E)$, the

likelihood function can be rewritten as

$$
\begin{aligned}
L(g) &= \exp\left\{\log \frac{\prod_{i=1}^{M}\prod_{j=1}^{N_i} g(E(x_{ij}))e^{-\beta_i E(x_{ij})}}{\prod_{i=1}^{M}[Z(\beta_i)]^{N_i}}\right\} \\
&= \exp\left\{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\log\left[g(E(x_{ij}))e^{-\beta_i E(x_{ij})}\right] - \sum_{i=1}^{M} N_i \log Z(\beta_i)\right\} \\
&= \exp\left\{\int \mathrm{d}E \sum_{i=1}^{M}\sum_{j=1}^{N_i}\delta(E - E(x_{ij}))\log\left[g(E)e^{-\beta_i E}\right] - \sum_{i=1}^{M} N_i \log Z(\beta_i)\right\} \\
&= \exp\left\{\int \mathrm{d}E H(E)\log[g(E)] - \sum_{i=1}^{M} N_i \log Z(\beta_i) - \sum_{i=1}^{M}\beta_i \int \mathrm{d}E H_i(E)E\right\} \\
&\propto \exp\left\{\int \mathrm{d}E H(E)\log[g(E)] - \sum_{i=1}^{M} N_i \log Z(\beta_i)\right\} \qquad (3.1.4.77)
\end{aligned}
$$

Please note that the constant term $e^{-\beta_i E}$ has been omitted, which does not affect the *a posteriori* estimate of $g(E)$.

Maximizing the log likelihood with respect to $g(E)$ leads to

$$g(E) = \frac{H(E)}{\sum_i N_i e^{-\beta_i E}/Z(\beta_i)}, \qquad (3.1.4.78)$$

or in the discrete form

$$g(E_k) = \frac{H(E_k)}{\sum_i N_i e^{-\beta_i E_k}/Z(\beta_i)}. \qquad (3.1.4.79)$$

This estimate suffers from overfitting because it only places nonzero probabilities at exactly the observed energies. Even for discrete systems, overfitting can arise if the data are of poor quality or sparse. To control the complexity of g using the idea in Bayesian statistics, we assign a prior probability $\pi(g)$ to $g(E)$, and the posterior probability of the density of states becomes

$$p(g) \propto L(g)\pi(g). \qquad (3.1.4.80)$$

For instance, with the Dirichlet prior $\pi(g) = \prod_k g_k^{n_k - 1}$, it becomes (the Boltzmann factors are also omitted)

$$p(g) \propto \prod_{k=1} g_k^{H_k + n_k - 1} / \prod_{i=1}^{M}[Z(\beta_i)]^{N_i} \qquad (3.1.4.81)$$

in a discrete form and maximizing $\log p(g)$ leads to

$$g(E_k) = \frac{H(E_k) + n_k - 1}{\sum_i N_i e^{-\beta_i E_k}/Z(\beta_i)}. \qquad (3.1.4.82)$$

Intuitively, n_k can be thought as the "pseudo count".

In the nonparametric setting, one does not assume a specific functional form but tries to determine g entirely from the data. Alternatively, one could model g parametrically using a family of functions involving parameters θ. Then the posterior distribution is a probability density over the θ parameter space. For instance, the density of state can be expanded by a finite mixture of Gaussian functions

$$g(E) = \sum_{c=1}^{C} \pi_c G(E; \epsilon_c, \sigma_c), \tag{3.1.4.83}$$

where $\theta = \{\pi_c, \epsilon_c, \sigma_c\}$ are parameters to be estimated under the constraint

$$\sum_c \pi_c = 1, \quad \pi_c \in [0, 1]. \tag{3.1.4.84}$$

Habeck extended this idea and proposed a Gibbs sampling method to determine the free energies and density of states as well as their uncertainties in 2012.[92]

Using a homogeneous Dirichlet prior, Eq. 3.1.4.81 becomes

$$p(g|D, \alpha) \propto \prod_{k=1}^{K} g_k^{H_k + \alpha/K - 1} / \prod_{i=1}^{M} \left[\sum_k g_k q_{ki} \right]^{N_i}, \tag{3.1.4.85}$$

where K is the total number of the discrete energy levels, and for simplicity we have defined $q_{ki} = e^{-\beta_i E_k}$. Using the integral $x^{-n} = \frac{1}{(n-1)!} \int_0^\infty t^{n-1} e^{-tx} \, \mathrm{d}t$, the posterior probability $p(g|D, \alpha)$ can be interpreted as the marginal distribution of the augmented posterior probability

$$p(g, t|D, \alpha) \propto \left(\prod_{k=1}^{K} g_k^{H_k + \alpha/K - 1} \right) \left(\prod_i t_i^{N_i - 1} \right) \prod_{k,i} e^{-g_k t_i q_{ki}} \tag{3.1.4.86}$$

by noting that

$$\begin{aligned}
&\int_0^\infty \cdots \int_0^\infty \mathrm{d}t_1 \cdots \mathrm{d}t_M \prod_i t_i^{N_i - 1} \prod_{k,i} e^{-g_k t_i q_{ki}} \\
=& \int_0^\infty \cdots \int_0^\infty \mathrm{d}t_1 \cdots \mathrm{d}t_M \prod_i t_i^{N_i - 1} \prod_i e^{-t_i \sum_k g_k q_{ki}} \\
=& \prod_i \int_0^\infty \mathrm{d}t_i t_i^{N_i - 1} e^{-t_i \sum_k g_k q_{ki}} \\
=& \prod_i (N_i - 1)! \left(\sum_k g_k q_{ki} \right)^{N_i}.
\end{aligned} \tag{3.1.4.87}$$

The conditional posterior probabilities for g and t can be drawn from Gibbs sampling iteratively via

$$p(t_i|g, D, \alpha) = G(N_i, \sum_k g_k q_{ki}) \tag{3.1.4.88a}$$

$$p(g_k|t, D, \alpha) = G(H_k + \alpha/K, \sum_i t_i q_{ki}) \tag{3.1.4.88b}$$

where $G(a, b)$ denotes the Gamma distribution with shape parameter a and scale parameter b. After each iteration, the density of states is normalized. The conditional expectations are

$$\langle t_i|g\rangle = \frac{N_i}{\sum_k g_k q_{ki}} = \frac{N_i}{Z_i} \quad \text{and} \quad \langle g_k|t\rangle = \frac{H_k + \alpha/K}{\sum_i t_i q_{ki}}. \tag{3.1.4.89}$$

The auxiliary function t_i is inversely proportional to the partition function of the corresponding ensemble. Therefore, $f_i = \log t_i$ is an analog of the free energy. Replacing t_i with f_i results in the augmented distribution of the density of states and free energies $p(g, f|D, \alpha)$, which follows

$$p(g, f|D, \alpha) \propto \left(\prod_{k=1}^{K} g_k^{H_k+\alpha/K-1}\right)\left(\prod_i e^{N_i f_i - f_i}\right)\prod_{k,i} e^{-g_k q_{ki} e^{f_i}}. \tag{3.1.4.90}$$

By integrating over the density of states, the posterior probability of the free energies satisfies

$$p(f|D, \alpha) \propto \prod_i e^{N_i f_i} / \prod_k \left(\sum_j q_{kj} e^{f_j}\right)^{H_k+\alpha/K}. \tag{3.1.4.91}$$

3.1.5 Multistate Bennett Acceptance Ratio

"An alleged scientific discovery has no merit unless it can be explained to a barmaid."

– Ernest Rutherford

"So, you can never be a good scientist unless you go to bar regularly."

– Yihan Shao

The Multistate Bennett Acceptance Ratio (MBAR) method was developed by Shirts and Chodera in 2008.[93] The following derivation quite follows Ref. [94], which is based on Ref. [95]. Imaging you have carried out a series of simulations such as umbrella sampling, or replica exchange molecular dynamics simulations. Now you have K trajectories in total and each trajectory is characterized by Hamiltonian H_k and inverse temperature β_k. The trajectories unnecessarily have the same number of conformations. Instead, the number of conformations in trajectory k is N_k. Now, you mix all the samples and randomly pick one sample out of them. The probability for this sample to have coordinates $\mathbf{R}$ is

$$p_m(\mathbf{R}) = \frac{1}{N}\sum_{k=1}^{K} N_k p_k(\mathbf{R}), \tag{3.1.5.1}$$

in which $N = \sum_{k=1}^{K} N_k$ and the subscript m means mixed ensemble. $p_k(\mathbf{R})$ is the probability of finding this snapshot in trajectory k, which satisfies

$$p_k(\mathbf{R}) = {c_k}^{-1} q_k(\mathbf{R}). \tag{3.1.5.2}$$

c_k is the normalization constant. You can see that this mixed ensemble does not follow Boltzmann statistics, even if q_k does. It can be proved that if p_k is normalized, then p_m is also normalized.

The expectation of any operator $\hat{O}$ averaged over this mixed ensemble can be calculated by

$$\langle O \rangle_m = \int O(\mathbf{R}) p_m(\mathbf{R})\, \mathrm{d}\mathbf{R} \approx \frac{1}{N}\sum_{n=1}^{N} O(\mathbf{R}_n). \tag{3.1.5.3}$$

Using energy reweighting[21], we can calculate the expectation of this operator under *any* other Hamiltonian H_i and probability p_i, which can be

expressed as

$$\begin{aligned}\langle O\rangle_i &= \int O(\mathbf{R})p_i(\mathbf{R})\,\mathrm{d}\mathbf{R}\\ &= \int O(\mathbf{R})\frac{p_i(\mathbf{R})}{p_m(\mathbf{R})}p_m(\mathbf{R})\,\mathrm{d}\mathbf{R}\\ &\approx \frac{1}{N}\sum_{n=1}^{N} O(\mathbf{R}_n)\frac{p_i(\mathbf{R}_n)}{p_m(\mathbf{R}_n)}\\ &= \frac{1}{N}\sum_{n=1}^{N} O(\mathbf{R}_n){c_i}^{-1}\frac{q_i(\mathbf{R}_n)}{p_m(\mathbf{R}_n)}\\ &= \sum_{n=1}^{N} O(\mathbf{R}_n){c_i}^{-1}\frac{q_i(\mathbf{R}_n)}{\sum_{k=1}^{K} N_k p_k(\mathbf{R}_n)} \end{aligned} \tag{3.1.5.4}$$

Let $O = 1$, we find

$$1 = \sum_{n=1}^{N} {c_i}^{-1}\frac{q_i(\mathbf{R}_n)}{\sum_{k=1}^{K} N_k p_k(\mathbf{R}_n)}. \tag{3.1.5.5}$$

Since c_i does not depend on n,

$$\begin{aligned} c_i &= \sum_{n=1}^{N}\frac{q_i(\mathbf{R}_n)}{\sum_{k=1}^{K} N_k p_k(\mathbf{R}_n)}\\ &= \sum_{n=1}^{N}\frac{q_i(\mathbf{R}_n)}{\sum_{k=1}^{K} N_k {c_k}^{-1} q_k(\mathbf{R}_n)} \end{aligned} \tag{3.1.5.6}$$

In Boltzmann statistics, $q_k(\mathbf{R}) = \exp\left[-\beta_k U_k(\mathbf{R})\right]$ and $c_k = \int q_k(\mathbf{R})d\mathbf{R}$ is the partition function or the normalization constant. *Note that we have not assumed anything about the statistics of ensemble k and i. Besides, i is unnecessarily within $\{k\}$. For instance, if we run replica exchange molecular dynamics simulations at K inverse temperatures $\beta_1, \ldots, \beta_K$, β_i can be either one of these inverse temperatures or any other inverse temperature between β_1 and β_K. But extrapolation to inverse temperatures outside the range of $[\beta_K, \beta_1]$ is not recommended.*

If q_k and q_i follow Boltzmann statistics, and we define free energy $f_i = -{\beta_i}^{-1}\ln c_i$, Eq. 3.1.5.6 becomes

$$f_i = -{\beta_i}^{-1}\ln\sum_{n=1}^{N}\frac{\exp\left[-\beta_i U_i(\mathbf{R}_n)\right]}{\sum_{k=1}^{K} N_k \exp\left[\beta_k f_k - \beta_k U_k(\mathbf{R}_n)\right]}, \tag{3.1.5.7}$$

which must be solved self-consistently and can be determined up to a constant. We can fix f_1 (to 0 usually).

Again, from Eq. 3.1.5.4, we can define

$$W_{in} = {c_i}^{-1}\frac{q_i(\mathbf{R}_n)}{\sum_{k=1}^{K} N_k {c_k}^{-1} q_k(\mathbf{R}_n)}, \tag{3.1.5.8}$$

which is the weight of snapshot n in ensemble i determined by Hamiltonian H_i and temperature β_i. Specifically, for the Boltzmann statistics,

$$W_{in} = \frac{e^{-\beta_i[U_i(\mathbf{R}_n)-f_i]}}{\sum_{k=1}^{K} N_k e^{-\beta_k[U_k(\mathbf{R}_n)-f_k]}}. \tag{3.1.5.9}$$

Here, $e^{\beta_i f_i}$ in the numerator serves as a normalization factor for W_{in}, which is unknown beforehand. Practically, we can define an unnormalized weight function

$$W'_{in} = \frac{e^{-\beta_i U_i(\mathbf{R}_n)}}{\sum_{k=1}^{K} N_k e^{-\beta_k[U_k(\mathbf{R}_n)-f_k]}}. \tag{3.1.5.10}$$

As a special case, suppose we have performed a single simulation with an inverse temperature β and potential energy function $U_0(\mathbf{R})$. The expectation of an operator $\hat{X}$ under an inverse temperature β and potential energy function $U_1(\mathbf{R})$ is

$$\begin{aligned}
\langle X \rangle_1 =& \frac{\sum_n X(\mathbf{R}_n) W'_{1n}}{\sum_n W'_{1n}} \\
=& \frac{\sum_n X(\mathbf{R}_n) \frac{e^{-\beta U_1(\mathbf{R}_n)}}{N e^{-\beta[U_0(\mathbf{R}_n)-f_0]}}}{\sum_n \frac{e^{-\beta U_1(\mathbf{R}_n)}}{N e^{-\beta[U_0(\mathbf{R}_n)-f_0]}}} \\
=& \frac{\frac{1}{N}\sum_n X(\mathbf{R}_n) e^{-\beta \Delta U}}{\frac{1}{N}\sum_n e^{-\beta \Delta U}} \\
=& \frac{\left\langle X e^{-\beta \Delta U} \right\rangle_0}{\left\langle e^{-\beta \Delta U} \right\rangle_0},
\end{aligned} \tag{3.1.5.11}$$

which returns to Eq. 2.3.0.4. Similarly, Eq. 3.1.5.7 becomes

$$\exp(-\beta f_1) = \sum_{n=1}^{N} \frac{\exp(-\beta U_1(\mathbf{R}_n))}{N \exp(\beta f_0 - \beta U_0(\mathbf{R}_n))}. \tag{3.1.5.12}$$

By moving the term $\exp(-\beta f_0)$ to left hand side of this equation, we have

$$\Delta f = -\beta^{-1} \ln \sum_{n=1}^{N} \frac{1}{N} \exp(-\beta \Delta U) = -\beta^{-1} \ln \langle \exp(-\beta \Delta U) \rangle_0, \tag{3.1.5.13}$$

which is Eq. 3.1.1.9. It shows that TP is a special case of MBAR when only one thermodynamic state is sampled and used for extrapolation.

Equation 3.1.5.7 can also be solved using the Newton-Raphson algorithm by defining the residual $\mathbf{r}$ with elements

$$r_i = 1 - \sum_{n=1}^{N} \frac{\exp[\beta_i f_i - \beta_i U_i(\mathbf{R}_n)]}{\sum_{k=1}^{K} N_k \exp[\beta_k f_k - \beta_k U_k(\mathbf{R}_n)]}. \tag{3.1.5.14}$$

With the current value of $\mathbf{f}^\nu = [f_1^\nu, f_2^\nu, \ldots, f_K^\nu]^T$ and $\mathbf{r}(\mathbf{f}^\nu) \neq \mathbf{0}$, we look for $\Delta\mathbf{f}^\nu$ so that $\mathbf{r}(\mathbf{f}^\nu + \Delta\mathbf{f}^\nu) = \mathbf{0}$. Truncating after the first-order term of the Taylor expansion, $\mathbf{r}(\mathbf{f}^\nu + \Delta\mathbf{f}^\nu)$ can be approximately expressed as

$$\mathbf{r}(\mathbf{f}^\nu + \Delta\mathbf{f}^\nu) \approx \mathbf{r}(\mathbf{f}^\nu) + \mathbf{J}^\nu \Delta\mathbf{f}^\nu, \tag{3.1.5.15}$$

where $\mathbf{J}^\nu$ is the $K \times K$ Jacobian:

$$\mathbf{J}^\nu = \begin{bmatrix} \frac{\partial r_1(\mathbf{f}^\nu)}{\partial f_1^\nu} & \frac{\partial r_1(\mathbf{f}^\nu)}{\partial f_2^\nu} & \cdots & \frac{\partial r_1(\mathbf{f}^\nu)}{\partial f_K^\nu} \\ \frac{\partial r_2(\mathbf{f}^\nu)}{\partial f_1^\nu} & \frac{\partial r_2(\mathbf{f}^\nu)}{\partial f_2^\nu} & \cdots & \frac{\partial r_2(\mathbf{f}^\nu)}{\partial f_K^\nu} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_K(\mathbf{f}^\nu)}{\partial f_1^\nu} & \frac{\partial r_K(\mathbf{f}^\nu)}{\partial f_2^\nu} & \cdots & \frac{\partial r_K(\mathbf{f}^\nu)}{\partial f_K^\nu} \end{bmatrix} \tag{3.1.5.16}$$

with

$$\begin{aligned} \mathrm{J}_{ij}^\nu =& \frac{\partial r_i(\mathbf{f}^\nu)}{\partial f_j^\nu} \\ =& \sum_{n=1}^{N} \frac{N_j \beta_j \exp\left[\beta_i f_i - \beta_i U_i(\mathbf{R}_n)\right] \exp\left[\beta_j f_j - \beta_j U_j(\mathbf{R}_n)\right]}{\left\{\sum_{k=1}^{K} N_k \exp\left[\beta_k f_k - \beta_k U_k(\mathbf{R}_n)\right]\right\}^2} \\ & - \sum_{n=1}^{N} \frac{\beta_i \exp\left[\beta_i f_i - \beta_i U_i(\mathbf{R}_n)\right]}{\sum_{k=1}^{K} N_k \exp\left[\beta_k f_k - \beta_k U_k(\mathbf{R}_n)\right]} \delta_{ij}. \end{aligned} \tag{3.1.5.17}$$

Since $\mathbf{r}(\mathbf{f}^\nu + \Delta\mathbf{f}^\nu) = \mathbf{0}$, we find

$$\Delta\mathbf{f}^\nu \approx -\left[\mathbf{J}^\nu\right]^{-1} \mathbf{r}(\mathbf{f}^\nu). \tag{3.1.5.18}$$

and

$$\mathbf{f}^{\nu+1} = \mathbf{f}^\nu + \Delta\mathbf{f}^\nu \tag{3.1.5.19}$$

This iteration continues until this convergence has been reached by checking if $|\Delta\mathbf{f}| < \boldsymbol{\epsilon}$.

Ding et al[96] showed that solving Eq. 3.1.5.7 is equivalent to minimizing

$$f(b_1, b_2, \ldots, b_K) = \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{N_k} \ln \left\{ \sum_{l=1}^{K} \exp\left[-u_l(\mathbf{R}_j^k) - b_l\right] \right\} + \sum_{k=1}^{K} \frac{N_k}{N} b_k \tag{3.1.5.20}$$

with respect to b_k, where

$$b_k = -\ln \frac{N_k}{N} - f_k. \tag{3.1.5.21}$$

3.1.6 Umbrella Integration

Umbrella Integration (UI) was developed by Kästner and Thiel in 2005.[97] Usually, the Weighted Histogram Analysis Method (discussed in Sec. 3.1.4) or Multistate Bennett Acceptance Ratio (discussed in Sec. 3.1.5) are used for the postprocessing of the trajectories from umbrella sampling. There is an assumption in these methods that a global equilibrium has been reached among all the states. However, you can imagine that in umbrella sampling the explored configurational phase space is confined in a narrow region around the center of the restraining potential of each window. Each simulation sees only the landscape locally. Therefore the assumption of a global equilibrium is not valid. Amendments have been proposed. Among them is the umbrella integration method.

It can be seen from Eq. 3.1.4.40, the unbiased free energy $A_i^u(\xi)$ from the ith biased simulation is related to the biased probability $P_i^b(\xi)$ via

$$A_i^u(\xi) = -\beta^{-1} \ln P_i^b(\xi) - w_i(\xi) + F_i, \tag{3.1.6.1}$$

where $w_i(\xi) = \frac{1}{2}K_i(\xi - \xi_i)^2$ is the biasing potential of this window, and F_i is a constant to be solved for by WHAM. Instead of searching for the optimum $\{F_i\}$ for all the windows, in UI the free energy change is computed using the central idea of thermodynamic integration (discussed in Sec. 3.1.2), i.e. via

$$\Delta A^u_{a\to b} = \int_{\xi_a}^{\xi_b} \frac{\partial A^u}{\partial \xi}\, \mathrm{d}\xi. \tag{3.1.6.2}$$

Taking partial derivative on both sides of Eq. 3.1.6.1 with respect to ξ, we have

$$\frac{\partial A_i^u}{\partial \xi} = -\beta^{-1} \frac{\partial \ln P_i^b(\xi)}{\partial \xi} - \frac{\mathrm{d}w_i}{\mathrm{d}\xi}. \tag{3.1.6.3}$$

By assuming that the biased probability follows a normal distribution

$$P_i^b(\xi) = \frac{1}{\sqrt{2\pi}\sigma_i^b} \exp\left[-\frac{1}{2}\left(\frac{\xi - \overline{\xi_i^b}}{\sigma_i^b}\right)^2\right], \tag{3.1.6.4}$$

where $\overline{\xi_i^b}$ and σ_i^b of each window are calculated from the trajectory, Eq. 3.1.6.3 becomes

$$\frac{\partial A_i^u}{\partial \xi} = \beta^{-1} \frac{\xi - \overline{\xi_i^b}}{(\sigma_i^b)^2} - K_i(\xi - xi_i), \tag{3.1.6.5}$$

with a variance[98]

$$\mathrm{var}\left(\frac{\partial A_i^u}{\partial \xi}\right) = \frac{2(\xi - \overline{\xi_i^b})^2 + (\sigma_i^b)^2}{N_i \beta^2 (\sigma_i^b)^4} \tag{3.1.6.6}$$

To combine the different windows, the reaction coordinate is divided into bins that span the whole range of ξ and are independent of the windows. For

each bin centered at ξ_{bin}, the windows are combined by a weighted average as

$$\left.\frac{\partial A(\xi)}{\partial \xi}\right|_{\xi_{bin}} = \sum_{i}^{windows} p_i(\xi_{bin}) \left(\frac{\partial A_i^u(\xi)}{\partial \xi}\right)_{\xi_{bin}}, \tag{3.1.6.7}$$

with the normalized weight

$$p_i(\xi) = N_i P_i^b(\xi) / \sum_i N_i P_i^b(\xi). \tag{3.1.6.8}$$

N_i is the total number of steps sampled for window i. This ensures a high weight at the center of the distribution. The variance is

$$\mathrm{var}\left(\frac{\partial A^u}{\partial \xi}\right) = \sum_{i}^{windows} {p_i}^2 \mathrm{var}\left(\frac{\partial A_i^u}{\partial \xi}\right). \tag{3.1.6.9}$$

High-order correction to the biased distribution can be found in Ref. [99].

3.1.7 Non-Equilibrium Work

Non-Equilibrium Work for Free Energy Difference between Two States

Non-Equilibrium Work (NEW) method for equilibrium free energy calculations was proposed by Jarzynski.[100]. In 1997, Jarzynski showed

$$\langle \exp\left[-\beta W(\tau)\right]\rangle = \exp\left(-\beta \Delta A\right), \tag{3.1.7.1}$$

which is now called the Jarzynski equality. Here, W is the accumulated work along a path $\lambda(t)$ connecting the initial and final states, with $\lambda(0) = 0$ and $\lambda(\tau) = 1$, and $\Delta A = A(1) - A(0)$ the free energy difference between these two states. $\langle \cdots \rangle$ in Eq. 3.1.7.1 is an average over a series of trajectories with the initial conditions chosen according to the equilibrium Boltzmann probability in state $\lambda(0)$. The path average samples all the realizations of dynamic paths weighted by their respective path actions under the time evolution of the system with an explicitly time-dependent Hamiltonian. This equality was also obtained by Crooks for Markovian and microscopically reversible dynamics[101] and by Jarzynski via a master-equation approach[102].

Now, we consider creating an equilibrium configuration for the state $\lambda = 0$ and then slowly changing λ from 0 to 1. As the coupling parameter advances, the system continues to sample phase space by molecular dynamics or Monte Carlo simulations, but under an explicitly time-dependent Hamiltonian. In the limit of a very slow transformation, the system will remain close to the equilibrium. The free energy difference can then be evaluated by changing λ continuously

$$\Delta A = \lim_{\tau \to \infty} \int_0^\tau \left. \frac{\partial H\left[\mathbf{x}(t); \lambda\right]}{\partial \lambda} \right|_{\lambda = \lambda(t)} \dot{\lambda}(t)\, \mathrm{d}t, \tag{3.1.7.2}$$

where $\dot{\lambda}(t)$ is the time derivative of the coupling parameter λ. In Eq. 3.1.7.2, the limit of $\tau \to \infty$ ensures that the transformation is performed infinitely slowly, and thus reversibly. The right-hand side of Eq. 3.1.7.2 is the "reversible work" done to the system during the transformation.

If the system is instead transformed between the initial and final states over a finite time interval τ, the system will not be able to sample the phase space exhaustively at each value of λ, making this transformation irreversible. As the transformation proceeds, the system will be gradually driven out of equilibrium, causing hysteresis effects. From the second law of thermodynamic, it is expected that the work $W(\tau)$ performed during the nonequilibrium transformation is on average larger than or equal to the free energy difference between the two states

$$\langle W(\tau)\rangle \geq \Delta A, \tag{3.1.7.3}$$

and the difference accounts for heat-dissipation effect. The work $W(\tau)$ performed on the system is the accumulated energy cost required to change the system

$$W(\tau) = \int_0^\tau \frac{\partial H[\mathbf{x}(t);\lambda]}{\partial \lambda}\Big|_{\lambda=\lambda(t)} \dot{\lambda}(t)\,\mathrm{d}t \tag{3.1.7.4}$$

The equality in Eq. 3.1.7.3 will normally be achieved only if the transformation is infinitely slow, $\tau \to \infty$. For paths of finite length, the amount of dissipated work, $\langle W(\tau)\rangle - \Delta A \geq 0$, depends on the chosen transformation path $\lambda(t)$.

Jarzynski equality, Eq. 3.1.7.1, immediately leads to the second law in the form of Eq. 3.1.7.3 because of the Jensen's inequality, $\langle e^{-x}\rangle \geq e^{-\langle x\rangle}$. Moreover, TI and TP can be thought as the limiting cases of the nonequilibrium process. When $\tau \to \infty$ or $\dot{\lambda}(t) \to 0$, this is an infinitely slow transformation and the Eq. 3.1.7.2 is the formula of TI

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial H(\mathbf{x}, \mathbf{p}_x, \lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{3.1.7.5}$$

When $\tau \to 0$ or $\dot{\lambda}(t) \to \infty$, this is an infinitely fast transformation where the configurations will not relax and the work is simply the change in the Hamiltonian when going from the initial to the final state,

$$\lim_{\tau\to 0} W(\tau) = H(\mathbf{x}(0);\lambda=1) - H(\mathbf{x}(0);\lambda=0) \tag{3.1.7.6}$$

Substituting the Eq. 3.1.7.6 into the Eq. 3.1.7.1, the formula of TP can be recovered

$$\Delta A = -\frac{1}{\beta} \ln \langle \exp[-\beta \Delta H(\mathbf{x}, \mathbf{p}_x)]\rangle_0\,, \tag{3.1.7.7}$$

In Ref. [101], Crooks showed that the distributions of work values from the forward and the backward paths satisfy a relation that is central to the histogram methods in free energy calculations

$$\frac{p_f[w = W(\tau)]}{p_b[w = -\underline{W}(\tau)]} = \exp[\beta(w - \Delta A)], \tag{3.1.7.8}$$

where $p_f[w = W(\tau)]$ and $p_b[w = -\underline{W}(\tau)]$ are the probability densities of the work values in the forward and the reverse transformations (with a sign change for the work in the reverse path). Both are normalized, i.e., $\int p_f(w)\,\mathrm{d}w = \int p_b(w)\,\mathrm{d}w = 1$. It is noted that Jarzynski equality Eq. 3.1.7.1 follows from Eq. 3.1.7.8 simply by integration over w because the probability densities are normalized to 1:

$$\int p_f(w) e^{-\beta w}\,\mathrm{d}w = \int p_b(w) e^{-\beta \Delta A}\,\mathrm{d}w, \tag{3.1.7.9}$$

Because of the normalization condition, the right-hand side is equal to $\exp(-\beta\Delta A)$, and Jarzynski equality follows. The bias, variance and mean square error of the Jarzynski estimator were studied by Gore et al.[74]

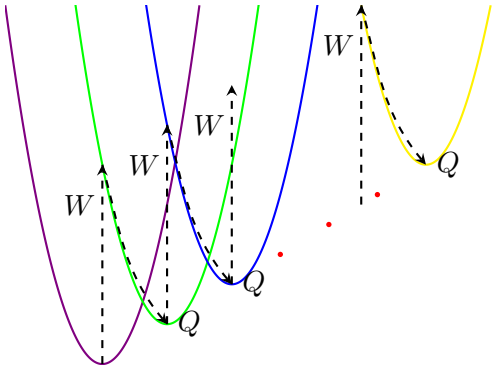

Figure 3.3: The accumulation of work and heat along a nonequilibrium trajectory. The work is defined as the energy change when the coupling parameter switches from λ_i to λ_{i+1} with the coordinates fixed, while the dissipated heat is defined as the energy relaxation when the coordinate change with the coupling parameter fixed.

Following the Crooks Fluctuation Theorem (CFT),[101] Bennett acceptance ratio can be applicable to nonequilibrium calculations. This approach was combined with a maximum likelihood estimate, and accurate free energy differences were obtained.[83] In this approach, ΔA is calculated via

$$\sum_{i=1}^{n_F} \frac{1}{1+\exp\left[\beta(M+W_i-\Delta A)\right]} = \sum_{j=1}^{n_R} \frac{1}{1+\exp\left[-\beta(M+W_j-\Delta A)\right]}, \tag{3.1.7.10}$$

where n_F and n_R are the numbers of the forward and reverse transformations respectively, W_i and W_j are the work of forward and reverse measurements respectively, and $M = \beta^{-1}\ln(n_F/n_R)$. The corresponding statistical variance of ΔA, σ^2, is calculated using Eq. 10 in Ref. [83].

Non-Equilibrium Work for Free Energy Profiles

When calculating the free energy profile in a pulling experiment, the Jarzynski equality is no longer straightforwardly applicable, because it relates the nonequilibrium work to free energy differences at different times, not positions along a predefined reaction coordinate. In order to surmount this difficulty, Hummer and Szabo extended the Jarzynski equality by measuring force/extension along pulling.[103]

Let us begin with a system of which the phase-space density evolves according to a Liouville-type equation:

$$\frac{\partial f(\mathbf{x},t)}{\partial t} = \mathscr{L}_t f(\mathbf{x},t). \tag{3.1.7.11}$$

$\mathscr{L}_t$ is an explicitly time-dependent evolution operator that has the Boltzmann distribution as a stationary solution, $\mathscr{L}_t e^{-\beta\mathscr{H}(\mathbf{x},t)} = 0$, where $\mathscr{H}(\mathbf{x},t)$ is a time-dependent Hamiltonian. For diffusive dynamics on a potential $V(\mathbf{x},t)$, the time evolution is governed by $\mathscr{L}_t = D\nabla e^{-\beta V(\mathbf{x},t)}\nabla e^{\beta V(\mathbf{x},t)}$, where D is the diffusion coefficient and $\nabla = \partial/\partial\mathbf{x}$. Now consider the unnormalized Boltzmann distribution at time t,

$$p(\mathbf{x},t) = \frac{e^{-\beta\mathscr{H}(\mathbf{x},t)}}{\int e^{-\beta\mathscr{H}(\mathbf{x}',0)}\,\mathrm{d}\mathbf{x}'}. \tag{3.1.7.12}$$

Because this distribution is stationary ($\mathscr{L}_t p = 0$), and because $\partial p/\partial t = -\beta(\partial\mathscr{H}/\partial t)p$, it follows that the above $p(\mathbf{x},t)$ is a solution of the sink equation

$$\frac{\partial p}{\partial t} = \mathscr{L}_t p - \beta\frac{\partial\mathscr{H}}{\partial t}p, \tag{3.1.7.13}$$

of which the solution, starting from an equilibrium distribution at time $t = 0$, can be expressed as a path integral by using the Feynman-Kac theorem. Equating these two differential solutions immediately gives:

$$\frac{e^{-\beta\mathscr{H}(\mathbf{x},t)}}{\int e^{-\beta\mathscr{H}(\mathbf{x}',0)}\,\mathrm{d}\mathbf{x}'} = \left\langle \delta(\mathbf{x}-\mathbf{x}_t)\exp\left[-\beta\int_0^t \frac{\partial\mathscr{H}}{\partial t'}(\mathbf{x}_{t'},t')\,\mathrm{d}t'\right]\right\rangle. \tag{3.1.7.14}$$

The average $\langle\cdots\rangle$ is over an ensemble of trajectories starting from the equilibrium distribution at $t = 0$ and evolving according to Eq. 3.1.7.11. Each trajectory is weighted with the Boltzmann factor of the external work w_t done to the system,

$$w_t = \int_0^t \frac{\partial\mathscr{H}}{\partial t'}(\mathbf{x}_{t'},t')\,\mathrm{d}t'. \tag{3.1.7.15}$$

Integrating on both sides of Eq. 3.1.7.14 with respect to $\mathbf{x}$, we obtain Jarzynski equality

$$e^{-\beta\Delta G(t)} \equiv \frac{\int e^{-\beta\mathscr{H}(\mathbf{x},t)}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta\mathscr{H}(\mathbf{x},0)}\,\mathrm{d}\mathbf{x}} = \left\langle e^{-\beta w_t}\right\rangle \tag{3.1.7.16}$$

between the Boltzmann-averaged work w_t and the equilibrium free energy difference $\Delta G(t)$ between times t and 0.

In a single-molecule pulling experiment, e.g. using atomic force microscope (AFM), the sample is moved at a constant speed v relative to the cantilever with spring constant k. The position $z_t = vt + \delta z_t$ of the cantilever tip with respect to the sample is recorded, where δz_t is the displacement of the cantilever tip. It can be described by a Hamiltonian

$\mathscr{H}(\mathbf{x},t) = \mathscr{H}_0(\mathbf{x}) + k(z(\mathbf{x}) - vt)^2/2$, where $\mathscr{H}_0(\mathbf{x})$ is the Hamiltonian of the resting, unperturbed system. *It is worth noting that $\mathscr{H}(\mathbf{x},t)$ is the Hamiltonian for the total system, including the cantilever.* Substituting this Hamiltonian into Eq. 3.1.7.14, multiplying both sides by $\delta[z - z(\mathbf{x})]$, and integrating over all $\mathbf{x}$, we have

$$\begin{aligned}
\frac{\int \delta[z - z(\mathbf{x})] e^{-\beta\left\{\mathscr{H}_0(\mathbf{x}) + k[z(\mathbf{x}) - vt]^2/2\right\}}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta\mathscr{H}(\mathbf{x},0)}\,\mathrm{d}\mathbf{x}} &= \\
&\int \delta[z - z(\mathbf{x})] \left\langle \delta(\mathbf{x} - \mathbf{x}_t) e^{-\beta\int_0^t -kv[z(\mathbf{x}_{t'}) - vt')]\,\mathrm{d}t'} \right\rangle \mathrm{d}\mathbf{x} \\
\frac{\int \delta[z - z(\mathbf{x})] e^{-\beta\mathscr{H}_0(\mathbf{x})}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta\mathscr{H}(\mathbf{x},0)}\,\mathrm{d}\mathbf{x}} e^{-\beta k(z-vt)^2/2} &= \\
&\left\langle \delta[z - z(\mathbf{x_t})] e^{-\beta\left[-kv\int_0^t z(\mathbf{x}_{t'})\,\mathrm{d}t' + kv^2t^2/2\right]} \right\rangle \\
\frac{\int \delta[z - z(\mathbf{x})] e^{-\beta\mathscr{H}_0(\mathbf{x})}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta\mathscr{H}(\mathbf{x},0)}\,\mathrm{d}\mathbf{x}} &= \\
&\left\langle \delta[z - z(\mathbf{x_t})] e^{-\beta\left[kv^2t^2/2 - kv\int_0^t z(\mathbf{x}_{t'})\,\mathrm{d}t' - k(z(\mathbf{x}_t) - vt)^2/2\right]} \right\rangle
\end{aligned} \tag{3.1.7.17}$$

and finally taking the logarithm, we have:

$$\begin{aligned}
G_0(z) \equiv &-\beta^{-1} \ln \frac{\int \delta[z - z(\mathbf{x})] e^{-\beta\mathscr{H}_0(\mathbf{x})}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta\mathscr{H}_0(\mathbf{x})}\,\mathrm{d}\mathbf{x}} \\
&= -\beta^{-1} \ln \left\langle \delta(z - z_t) e^{-\beta\Delta w_t} \right\rangle - \beta^{-1} \ln \frac{\int e^{-\beta\mathscr{H}_0(\mathbf{x}) + k[z(\mathbf{x})]^2/2}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta\mathscr{H}_0(\mathbf{x})}\,\mathrm{d}\mathbf{x}} \\
&= -\beta^{-1} \ln \left\langle \delta(z - z_t) e^{-\beta\Delta w_t} \right\rangle + \delta G,
\end{aligned} \tag{3.1.7.18}$$

where $G_0(z)$ is the unperturbed free energy profile along the pulling coordinate z, and Δw_t is the external work minus the instantaneous biasing potential, $\Delta w_t = w_t - k(z_t - vt)^2/2 = kv(vt^2/2 - \int_0^t z_{t'}\,\mathrm{d}t') - k(z_t - vt)^2/2$. δG is independent of time. At time $t = 0$, the trajectories are started from points z_0 drawn from a Boltzmann distribution corresponding to Hamiltonian $\mathscr{H}(\mathbf{x},0) = \mathscr{H}_0(\mathbf{x}) + kz^2/2$, which is *NOT* $\mathscr{H}_0(\mathbf{x})$.

At each time slice t, one can in principle obtain an estimate of the whole free energy surface. In practice with finite number of trajectories, at any

given time t, only a narrow region around the equilibrium position $z = vt$ will be sampled adequately. Thus, an average over several time slices and repeated trajectories is required to obtain an optimal estimate of the free energy profile. At every time slice t, one obtains an ensemble of positions z_t and corresponding w_ts. The position z_t are binned, and the corresponding histogram values are incremented by $e^{-\beta w_t}$. The complete free energy profile $G_0(z)$ can be reconstructed by adapting the weighted histogram method:

$$G_0(z) = -\beta^{-1} \ln \frac{\sum_t \frac{\langle \delta(z-z_t) \exp(-\beta w_t) \rangle}{\langle \exp(-\beta w_t) \rangle}}{\sum_t \frac{\exp[-\beta u(z,t)]}{\langle \exp(-\beta w_t) \rangle}}, \tag{3.1.7.19}$$

where the sum is over time slices t and $u(z,t) = k(z - vt)^2/2$ is the time dependent biasing potential. As in the weighted histogram method, this procedure can be refined by making Eq. 3.1.7.19 self-consistent through replacement of $\langle \exp(-\beta w_t) \rangle$ with

$$\exp[-\beta \Delta G(t)] = \frac{\int \exp\{-\beta [u(z,t) + G_0(z)]\} \, \mathrm{d}z}{\int \exp\{-\beta [u(z,0) + G_0(z)]\} \, \mathrm{d}z}, \tag{3.1.7.20}$$

thus requiring an iterative solution for $\Delta G(t)$. Note that Eq. 3.1.7.19 can be rewritten as

$$G_0(z) = -\beta^{-1} \ln \frac{\sum_t \frac{\langle \delta(z-z_t) \exp(-\beta \Delta w_t) \rangle \exp[-\beta u(z,t)]}{\langle \exp(-\beta w_t) \rangle}}{\sum_t \frac{\exp[-\beta u(z,t)]}{\langle \exp(-\beta w_t) \rangle}}, \tag{3.1.7.21}$$

which is the natural logarithm of the weighted average of $\langle \delta(z - z_t) \exp(-\beta \Delta w_t) \rangle$ over all the time slices with $\frac{\exp[-\beta u(z,t)]}{\langle \exp(-\beta w_t) \rangle}$ being the weight.

3.1.8 Transition-Based Reweighting Analysis Method

Transition-Based Reweighting Analysis Method (TRAM), which relies on the maximum likelihood analysis of the thermodynamic and kinetic information, was developed by Frank Noé and coworkers in 2014.[104]. It incorporates WHAM with Markov state model, and avoids the weakness in both methods. In WHAM, a global equilibrium among all the thermodynamic states must be reached. For the Markov state model (MSM), the kinetic information can be extracted from only one thermodynamic state. In contrast, TRAM is a class of estimators that (1) take the statistical weights of samples at different thermodynamic states into account, and (2) exploit transitions observed in the sampled trajectories, without assuming that these trajectories are sampled from equilibrium.

Let us assume that there are K molecular dynamics (MD) or Markov chain Monte Carlo (MCMC) simulations have been performed, each in a specific thermodynamic state (Hamiltonian, temperature, etc) indexed by $k \in \{1, \ldots, K\}$. For simulations with varying thermodynamic state in a single trajectory, with replica-exchange simulation being a typical example, each contiguous sequence is treated as a separated trajectory at one of the K thermodynamic states. We further assume the configuration space (that has been visited by the simulations) is discretized into cells indexed by $i, j \in \{1, \ldots, n\}$.

Similar to the WHAM analysis, the unbiased probability, π_i, and the biased probability under thermodynamic state k, $\pi_i^{(k)}$ are related by a known and constant bias factor $\gamma_i^{(k)}$

$$\pi_i^{(k)} = f^{(k)} \pi_i \gamma_i^{(k)}, \tag{3.1.8.1}$$

$$f^{(k)} = \frac{1}{\sum_l \pi_l \gamma_l^{(k)}}, \tag{3.1.8.2}$$

where $f^{(k)}$ is a normalization constant. Thus, the bias is multiplicative in probability or additive in the potential. As we have shown in section 3.1.4, the WHAM estimator can be derived by maximizing the likelihood

$$L_{\mathrm{WHAM}} = \prod_k \prod_i (\pi_i^{(k)})^{N_i^{(k)}} \tag{3.1.8.3}$$

with an implied assumption that every count $N_i^{(k)}$ is independently drawn from the biased distribution $\pi_i^{(k)}$.

The maximum likelihood Markov model is the transition matrix $\mathbf{P} = (p_{ij})$ between n discrete configuration states, that maximizes the likelihood of the observed transitions between these states. The likelihood of a Markov model is a product of all transition probabilities corresponding to the observed trajectory. To obtain a reversible Markov state model, this likelihood

is maximized under the constraints of detailed balance with respect to the equilibrium distribution $\boldsymbol{\pi}$

$$L_{\mathrm{MSM}} = \prod_i \prod_j p_{ij}^{c_{ij}}, \tag{3.1.8.4}$$

$$s.t. \quad \pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i, j, \tag{3.1.8.5}$$

where c_{ij} is the number of times the trajectories were observed in state i at time t and in state j at a later time $t + \tau$, where τ is the lag time at which the Markov model is estimated.

In TRAM, WHAM and MSM are combined as follows: every trajectory at thermodynamic state k is treated as a Markov chain with the configuration-state transition counts $c_{ij}^{(k)}$, without assuming that every count is sampled from global equilibrium. In contrast to Markov models, we exploit the fact that equilibrium probabilities can be reweighted between different thermodynamic states via Eqs. 3.1.8.1 and 3.1.8.2. The resulting likelihood of all $\mathbf{P}^{(k)}$ and $\boldsymbol{\pi}$, based on simulations at all thermodynamic states can be formulated as

$$L_{\mathrm{TRAM}} = \prod_k \prod_i \prod_j (p_{ij}^{(k)})^{c_{ij}^{(k)}}, \tag{3.1.8.6}$$

$$s.t. \quad \pi_i^{(k)} p_{ij}^{(k)} = \pi_j^{(k)} p_{ji}^{(k)} \quad \text{for all } i, j, k. \tag{3.1.8.7}$$

Here, $\mathbf{P}^{(k)} = (p_{ij}^{(k)})$ is the Markov transition matrix at thermodynamic state k, and $c_{ij}^{(k)}$ are the number of transitions observed at that simulation condition. $\boldsymbol{\pi}^{(k)}$ is the vector of equilibrium probabilities of discrete states at each thermodynamic state. *Because each Markov model $\mathbf{P}^{(k)}$ must have the distribution $\boldsymbol{\pi}^{(k)}$ as a stationary distribution, all Markov models are coupled too, which makes the maximization of the TRAM likelihood Eqs. 3.1.8.6 and 3.1.8.7 difficult, and it can neither be achieved by WHAM, nor by existing MSM estimators.*

Taking natural logarithm on the TRAM likelihood, we find

$$\ln L_{\mathrm{TRAM}} = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^{(k)} \ln p_{ij}^{(k)}, \tag{3.1.8.8}$$

with constraints

$$\pi_i \gamma_i^{(k)} p_{ij}^{(k)} = \pi_j \gamma_j^{(k)} p_{ji}^{(k)} \quad \text{for all } i, j, k, \tag{3.1.8.9}$$

which is from Eqs. 3.1.8.1 and 3.1.8.7 with $f^{(k)}$ canceled. In addition, $\mathbf{P}^{(k)}$ and $\boldsymbol{\pi}$ should satisfy the normalization conditions

$$\sum_j p_{ij}^{(k)} = 1 \quad \forall i, k, \tag{3.1.8.10}$$

$$\sum_j \pi_j = 1. \tag{3.1.8.11}$$

The normalization of $\boldsymbol{\pi}^{(k)}$ is naturally satisfied due to Eqs. 3.1.8.1 and 3.1.8.2. Due to the existence of constraints, the numbers of free variables are $n-1$ for $\boldsymbol{\pi}$ and $n(n-1)/2$ for $\mathbf{P}^{(k)}$.

Using the Lagrange duality theory, it can be shown that the optimal solution of the discrete TRAM problem above fulfills the following two conditions

$$\sum_k \sum_j \frac{\left(c_{ij}^{(k)} + c_{ji}^{(k)}\right) \gamma_i^{(k)} \pi_i \nu_j^{(k)}}{\gamma_i^{(k)} \pi_i \nu_j^{(k)} + \gamma_j^{(k)} \pi_j \nu_i^{(k)}} = \sum_k \sum_j c_{ji}^{(k)}, \tag{3.1.8.12}$$

$$\sum_j \frac{\left(c_{ij}^{(k)} + c_{ji}^{(k)}\right) \gamma_j^{(k)} \pi_j}{\gamma_i^{(k)} \pi_i \nu_j^{(k)} + \gamma_j^{(k)} \pi_j \nu_i^{(k)}} = 1 \tag{3.1.8.13}$$

where $\nu_i^{(k)}$ are unknown Lagrange multipliers. To numerically solve the discrete TRAM problem, an initial guess for $\boldsymbol{\pi}$ and $\mathbf{v}^{(k)}$ can be made as

$$\pi_i^{init} := 1/n, \qquad v_i^{(k),init} := \sum_j c_{ij}^{(k)}, \tag{3.1.8.14}$$

and the following equations must be solved iteratively until $\boldsymbol{\pi}$ is converged:

$$v_i^{(k),new} := v_i^{(k)} \sum_j \frac{\left(c_{ij}^{(k)} + c_{ji}^{(k)}\right) \gamma_i^{(k)} \pi_j}{\gamma_i^{(k)} \pi_i \nu_j^{(k)} + \gamma_j^{(k)} \pi_j \nu_i^{(k)}}, \tag{3.1.8.15}$$

$$\pi_i^{new} := \frac{\sum_{k,j} c_{ji}^{(k)}}{\sum_{k,j} \frac{\left(c_{ij}^{(k)} + c_{ji}^{(k)}\right) \gamma_i^{(k)} \nu_j^{(k)}}{\gamma_i^{(k)} \pi_i \nu_j^{(k)} + \gamma_j^{(k)} \pi_j \nu_i^{(k)}}}. \tag{3.1.8.16}$$

3.2 Approximate Methods

3.2.1 Molecular Mechanics/Poisson-Boltzmann Surface Area

The following derivation follows Ref. [105]. The Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) method is often used in the calculations of binding free energy of a substrate to a receptor. The standard binding free energy for a reaction between a receptor (A) and a substrate (B)

$$\mathrm{A} \quad + \quad \mathrm{B} \quad \rightleftharpoons \quad \mathrm{AB}$$

is expressed as a ratio of configuration integrals

$$\begin{aligned}\Delta G^0_{AB} =& -\beta^{-1} \ln \left(\frac{C^0}{8\pi^2} \cdot \frac{Z_{N,AB} Z_{N,0}}{Z_{N,A} Z_{N,B}} \right) + P^0 \langle \Delta V_{AB} \rangle \\ =& -\beta^{-1} \ln \left(\frac{C^0}{8\pi^2} \cdot \frac{\frac{Z_{N,AB}}{Z_{N,0}}}{\frac{Z_{N,A}}{Z_{N,0}} \frac{Z_{N,B}}{Z_{N,0}}} \right) + P^0 \langle \Delta V_{AB} \rangle ,\end{aligned} \tag{3.2.1.1}$$

where R is the gas constant, T is the temperature, C^0 is the standard state concentration (1 M), N is the number of solvent molecules, and $P^0 \langle \Delta V_{AB} \rangle$ is the pressure-volume work associated with changing the system size after the association of two species into one complex. For water solution at 1 atm, the last term is negligibly small. There are no mass-dependent terms in Eq. 3.2.1.1, which is a direct result of equal kinetic contribution to the partition function of the bound and the free species. The configuration integral of the receptor, A, in solution is

$$Z_{N,A} = \int e^{-\beta U(r_A, r_S)} \, \mathrm{d}r_A \, \mathrm{d}r_S, \tag{3.2.1.2}$$

where $U(r_A, r_S)$ is the potential energy as a function of all solute coordinates, r_A, and solvent coordinates, r_S, and β is the reciprocal of the product of the Boltzmann constant and temperature. The total potential energy can be decomposed into $U(r_A) + U(r_S) + \Delta U(r_A, r_S)$. Similar for B, the substrate. For pure solvent, the configuration integral is

$$Z_{N,0} = \int e^{-\beta U(r_S)} \, \mathrm{d}r_S. \tag{3.2.1.3}$$

The ratio of configuration integrals in Eq. 3.2.1.1 can be simplified with an implicit solvent approximation, as

$$\begin{aligned}\frac{Z_{N,A}}{Z_{N,0}} = Z_A =& \frac{\int e^{-\beta U(r_A)} \left\{ \int e^{-\beta \Delta U(r_A, r_S)} e^{-\beta U(r_S)} \, \mathrm{d}r_S \right\} \mathrm{d}r_A}{\int e^{-\beta U(r_S)} \, \mathrm{d}r_S} \\ =& \int e^{-\beta [U(r_A) + W(r_A)]} \, \mathrm{d}r_A,\end{aligned} \tag{3.2.1.4}$$

where

$$W(r_A) = -\beta^{-1} \ln \left(\frac{\int e^{-\beta \Delta U(r_A, r_S)} e^{-\beta U(r_S)} \, \mathrm{d}r_S}{\int e^{-\beta U(r_S)} \, \mathrm{d}r_S} \right) \tag{3.2.1.5}$$

is the solvation free energy of the receptor A at fixed coordinate r_A. Analogous equations hold for the complex and substrate.

For the complex, we define the position (translational degrees of freedom) and orientation (rotational degrees of freedom) of the substrate with respective to the receptor as $\delta_B \equiv (x_1, x_2, x_3, \xi_1, \xi_2, \xi_3)$. Generally, these degree-of-freedom is very limited. The complex configuration integral is

$$Z_{AB} = \int e^{-\beta [U(r_A, r_{B'}, \delta_B) + W(r_A, r_{B'}, \delta_B)]} \, \mathrm{d}r_A \, \mathrm{d}r_{B'} \, \mathrm{d}\delta_B, \tag{3.2.1.6}$$

where $r_{B'}$ represents the remaining internal degrees of freedom of the bound substrate and δ_B spans conformations where A and B form a complex. Then we assume that the translational and rotational motions of the substrate in the bound state are not strongly coupled with the other degrees of freedom, and we decompose the potential and solvation energies as (*so weird!*)

$$\begin{aligned} &U\left(r_A, r_{B'}, \delta_B\right) + W\left(r_A, r_{B'}, \delta_B\right) \\ &\quad \approx U_1(\delta_B) + W_1(\delta_B) + U_2(r_A, r_{B'}) + W_2(r_A, r_{B'}). \end{aligned} \tag{3.2.1.7}$$

We further assume that the residual translational and rotational motions of the substrate are uncorrelated. Therefore

$$U_1(\delta_B) \approx U(x_1, x_2, x_3) + U(\xi_1, \xi_2, \xi_3), \tag{3.2.1.8}$$

and

$$W_1(\delta_B) \approx W(x_1, x_2, x_3) + W(\xi_1, \xi_2, \xi_3). \tag{3.2.1.9}$$

Now, Eq. 3.2.1.1 can bre written as

$$\Delta G^0_{AB} = -\beta^{-1} \ln \left[\frac{C^0 Z^{trans}_{B'} Z^{rot}_{B'} Z_{AB'}}{8\pi^2 Z_A Z_B} \right], \tag{3.2.1.10}$$

where

$$Z^{trans}_{B'} = \int e^{-\beta [U(x_1, x_2, x_3) + W(x_1, x_2, x_3)]} \, \mathrm{d}x_1 \, \mathrm{d}x_2 \, \mathrm{d}x_3 \tag{3.2.1.11}$$

and

$$Z^{rot}_{B'} = \int e^{-\beta [U(\xi_1, \xi_2, \xi_3) + W(\xi_1, \xi_2, \xi_3)]} \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 \, \mathrm{d}\xi_3. \tag{3.2.1.12}$$

As a first-order approximation, we assume that the energetic landscape of each species has an energy and a volume (entropy),

$$Z_A = \int e^{-\beta [U(r_A) + W(r_A)]} \, \mathrm{d}r_A \approx Z^{int}_A e^{-\beta \langle E_A \rangle}, \tag{3.2.1.13}$$

where $\langle E_A \rangle = \langle U(r_A) + W(r_A) \rangle$. We further assume (*how many approximations we have taken!*) that $Z_A^{int} Z_B^{int} \approx Z_{AB}^{int}$, then

$$\Delta G_{AB}^0 = -\beta^{-1} \ln \left(\frac{C^0 Z_{B'}^{trans} Z_{B'}^{rot}}{8\pi^2} \right) + \left(\langle E_{AB'} \rangle - \langle E_A \rangle - \langle E_B \rangle \right). \quad (3.2.1.14)$$

The bound substrate's translational configuration integral, $Z_{B'}^{trans}$, can be conceptually linked to the volume of space that its center of mass occupies through the simulation. The effective volume was measured based on the assumption that the translational motion is restrained by three harmonic potential. By solving eigenstates of the center-of-mass covariance matrix, the eigenvalues describe the variance Δx_i^2 along each principal axis. Thus, the translational configuration integral can be calculated as

$$\begin{aligned} Z_{B'}^{trans} =& \int e^{\left(-k_1 \Delta x_1^2 / 2k_B T\right)} \mathrm{d}x_1 \int e^{\left(-k_2 \Delta x_2^2 / 2k_B T\right)} \mathrm{d}x_2 \int e^{\left(-k_3 \Delta x_3^2 / 2k_B T\right)} \mathrm{d}x_3 \\ =& (2\pi)^{3/2} \left(\left\langle \Delta x_1^2 \right\rangle \left\langle \Delta x_2^2 \right\rangle \left\langle \Delta x_3^2 \right\rangle \right)^{1/2} \end{aligned} \quad (3.2.1.15)$$

where

$$k_i = \frac{k_B T}{\left\langle \Delta x_i^2 \right\rangle}. \quad (3.2.1.16)$$

The rotational configuration integral can be accounted in a similar manner.

4

Evaluation of Reliability

"A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it."

– Albert Einstein

4.1 Overlap Matrix

Overlap matrix proposed by Mobley et. al.,[106] can be used to essentially measures the magnitude of the phase space overlap. For example, after MBAR method is used to analyze the US simulations or a series of alchemical window simulations, the overlap matrix can be used to examine the reliabilities of MBAR calculations. The formula about the overlap matrix is shown as follows:

For the US simulations, with the weight of the lth configuration in the ith biased simulation appearing in the tth simulation defined as

$$w_t(\mathbf{x}_{i,l}) = \frac{e^{-\beta\left[W_t(\mathbf{x}_{i,l}) - f_t^{(b)}\right]}}{\sum_{k=1}^{S} N_k e^{-\beta\left[W_k(\mathbf{x}_{i,l}) - f_k^{(b)}\right]}}, \tag{4.1.0.1}$$

the elements of the $S \times S$ overlap matrix are[106]

$$\begin{aligned} O_{tt'} &= \sum_{i=1}^{S}\sum_{l=1}^{N_i} N_t w_t(\mathbf{x}_{i,l}) w_{t'}(\mathbf{x}_{i,l}) \\ &= \sum_{i=1}^{S}\sum_{l=1}^{N_i} \frac{N_t e^{-\beta\left[W_t(\mathbf{x}_{i,l}) - f_t^{(b)}\right]} e^{-\beta\left[W_{t'}(\mathbf{x}_{i,l}) - f_{t'}^{(b)}\right]}}{\left\{\sum_{k=1}^{S} N_k e^{-\beta\left[W_k(\mathbf{x}_{i,l}) - f_k^{(b)}\right]}\right\}^2}. \end{aligned} \tag{4.1.0.2}$$

Consecutive windows should have substantial overlap with the diagonal and the first off-diagonal elements no smaller than 0.03 as recommended[106].

For a series of alchemical window simulations, it is a $K \times K$ matrix with entries

$$O_{ij} = \sum_{n=1}^{N} \frac{N_i p_i(x_n)}{\sum_{k=1}^{K} N_k p_k(x_n)} \cdot \frac{p_j(x_n)}{\sum_{k=1}^{K} N_k p_k(x_n)}, \tag{4.1.0.3}$$

where $p_i(x_n) = e^{\beta G_i - \beta U_i(x_n)}$ is the probability of sample x_n occurring when simulation state i and N samples are collected with N_1 samples from $p_1(x)$ distribution, N_2 samples from $p_2(x)$ distribution, and so on. K is the total number of the states. O_{ij} can be interpreted as the average probability of a sample generated in state j being observed in the ith state. The average is computed over samples collected from all the K states, not just the samples from state j. Therefore O_{ij} is a measure of the overlap in the phase space of state i and j. The larger the better. The largest eigenvalue is 1. Similarly, consecutive windows should have substantial overlap with the diagonal and the first off-diagonal elements no smaller than 0.03 as recommended[106].

4.2 Π Metric for Neglected-tail Bias Model

The neglected-tail bias model is developed by Kofke and coworkers for the estimate of bias in free energy via thermodynamic perturbation or the Jarzynski equality[107, 108]. Here, we will take the latter as an example. Let $p_A(W)$ and $p_B(W)$ be the distributions/probabilities of work samples W in a forward ($A \to B$) and a backward ($B \to A$) nonequilibrium conversions. The Jarzynski equality discussed in 3.1.7 shows that the free energy difference is given by

$$\exp\left(-\beta \Delta A\right) = \int_{-\infty}^{\infty} p_A(W) e^{-\beta W} \, \mathrm{d}W, \tag{4.2.0.1}$$

or

$$\exp\left(+\beta \Delta A\right) = \int_{-\infty}^{\infty} p_B(W) e^{+\beta W} \, \mathrm{d}W. \tag{4.2.0.2}$$

These two distributions are related

$$p_A(W) e^{-\beta W} = p_B(W) e^{-\beta \Delta A}. \tag{4.2.0.3}$$

The neglected-tail bias model asserts that *all* of the bias is due the neglect of contributions below a particular value of W, designated as W^*, such that no sampling is contributed below this value, and *perfect sampling* is achieved for $W > W^*$.

It can be easily imagined that W^* depends on the sample size. When more sampling is performed, more likely a more negative value of W^* will

be encountered. Given $p_A(W)$, the probability distribution for W^* being observed once within M samples is

$$P_A^*(W^*) = M p_A(W^*)[1 - C_A(W^*)]^{M-1}, \tag{4.2.0.4}$$

where $C_A(W)$ is the cumulative distribution function defined as

$$C_A(W^*) = \int_{-\infty}^{W^*} p_A(W)\,\mathrm{d}W \tag{4.2.0.5}$$

and

$$1 - C_A(W^*) = \int_{W^*}^{\infty} p_A(W)\,\mathrm{d}W. \tag{4.2.0.6}$$

The bias can be written as

$$\begin{aligned} B(M) &= \langle \Delta A(M) \rangle - \Delta A \\ &= -\beta^{-1} \ln \left[\frac{1}{M} \left(e^{-\beta W^*} + (M-1) \cdot \frac{\int_{W^*}^{\infty} \mathrm{d}W e^{-\beta W} p_A(W)}{1 - C_A(W^*)} \right) \right] - \Delta A \\ &= -\beta^{-1} \ln \left[\frac{1}{M} \left(e^{-\beta W_{dis}^*} + (M-1) \cdot \frac{1 - C_B(W^*)}{1 - C_A(W^*)} \right) \right], \end{aligned} \tag{4.2.0.7}$$

where

$$W_{dis} = W - \Delta A \tag{4.2.0.8}$$

is the dissipated work. For the third equality, we have employed Eq. 4.2.0.3. The bias can be evaluated as the average of the inaccuracy over the distribution of W^*

$$B(M) = -\beta^{-1} \int_{-\infty}^{\infty} \mathrm{d}W^* P_A^*(W^*) \ln \left[\frac{1}{M} \left(e^{-\beta W_{dis}^*} + (M-1) \cdot \frac{1 - C_B(W^*)}{1 - C_A(W^*)} \right) \right]. \tag{4.2.0.9}$$

Alternatively, it can be estimated for a single value of W^*, for instance the mode $\hat{W}^*$ of P_A^*. It follows from

$$\left. \frac{\mathrm{d} \ln P_A(W^*)}{\mathrm{d}W^*} \right|_{W=\hat{W}^*} = 0 \tag{4.2.0.10}$$

and is given by the solution of

$$\left. \frac{\mathrm{d} \ln p_A(W)}{\mathrm{d}W} \right|_{W=\hat{W}^*} = \frac{(M-1) p_A(\hat{W}^*)}{1 - C_A(\hat{W}^*)}. \tag{4.2.0.11}$$

or, with a small approximation

$$\left. \frac{\mathrm{d} \ln p_A(W)}{\mathrm{d}W} \right|_{W=\hat{W}^*} = (M-1) p_A(\hat{W}^*). \tag{4.2.0.12}$$

Assuming W^* has a Gaussian distribution centered at $\hat{W}^*$ as

$$P_A^*(W^*) \approx \frac{1}{\sqrt{2\pi}\sigma^*} \exp\left[-(W^* - \hat{W}^*)^2/2(\sigma^*)^2\right] \tag{4.2.0.13}$$

in which, using Eq. 4.2.0.4 and the property in Eq. 4.2.0.11,

$$\frac{1}{(\sigma^*)^2} = -\frac{\mathrm{d}^2 \ln P_A^*}{\mathrm{d}\,(W^*)^2} = -\left[\frac{\partial^2 \ln p_A}{\partial W^2} - \frac{M}{M-1}\left(\frac{\partial \ln p_A}{\partial W}\right)^2\right]\Bigg|_{W=\hat{W}^*}. \tag{4.2.0.14}$$

Corresponding relations can be given for the bias of the $B \to A$ work process. Thus from the work distributions the expected performance of a simulation of given sampling length can be predicted.

Next, let us look at the work distributions. We assume a Gaussian-work distribution

$$p_A(W) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(W - \bar{W})^2/2\sigma^2\right]. \tag{4.2.0.15}$$

From Eq. 4.2.0.3, it can be found that $p_B(W)$ is also a Gaussian with the same variance but shifted by $\beta\sigma^2$

$$p_B(W) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(W - \bar{W} + \beta\sigma^2)^2/2\sigma^2\right]. \tag{4.2.0.16}$$

The free energy difference is

$$\Delta A = \bar{W} - \beta\sigma^2. \tag{4.2.0.17}$$

The cumulative distribution functions are

$$C_A(W) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\bar{W} - W}{\sqrt{2}\sigma}\right), \tag{4.2.0.18}$$

$$C_B(W) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\bar{W} - \beta\sigma^2 - W}{\sqrt{2}\sigma}\right), \tag{4.2.0.19}$$

where $\mathrm{erfc}(x)$ is the complementary error function.

By taking Eq. 4.2.0.15 into Eq. 4.2.0.12, the mode of the least-work distribution is approximately

$$\hat{W}^* = \bar{W} - \sigma\sqrt{\mathbf{W}_L\left[\frac{1}{2\pi}(M-1)^2\right]}, \tag{4.2.0.20}$$

where $\mathbf{W}_L(x)$ is the Lambert W function, defined as the solution to $x = we^w$. With a slight approximation to Eq. 4.2.0.7 as

$$B(M) = -\beta^{-1} \ln\left[1 - C_B(W^*)\right] \tag{4.2.0.21}$$

and this estimated mode used for W^*, this Gaussian-work bias becomes

$$B(M) = -\beta^{-1} \ln \left\{ \frac{1}{2} \operatorname{erfc} \left[-\frac{1}{\sqrt{2}} \left(\sqrt{\mathbf{W}_L \left[\frac{1}{2\pi} (M-1)^2 \right]} - \beta\sigma \right) \right] \right\}, \tag{4.2.0.22}$$

and depends solely on

$$\Pi = \sqrt{\mathbf{W}_L \left[\frac{1}{2\pi} (M-1)^2 \right]} - \beta\sigma. \tag{4.2.0.23}$$

By using Eq. 4.2.0.17, it can be rewritten as

$$\Pi = \sqrt{\mathbf{W}_L \left[\frac{1}{2\pi} (M-1)^2 \right]} - \sqrt{2\beta(\bar{W} - \Delta A)}. \tag{4.2.0.24}$$

Wu and Kofke suggested that the number of work samples (non-equilibrium trajectories) should be sufficient to ensure $\Pi > 0.5$.[108] This idea has also been applied to the estimate of the bias in bridge estimators such as BAR.[109]

4.3 Kullback–Leibler divergence

The convergence rate of thermodynamic perturbation heavily depends on the similarity between the simulated Hamiltonian (A) and the target one (B). Kullback–Leibler divergence offers a quantitative way to measure the similarity by

$$\begin{aligned} KL(p_A||p_B) &= \int_\Gamma \mathrm{d}\mathbf{x}\, p_A(\mathbf{x}) \ln \left[\frac{p_A(\mathbf{x})}{p_B(\mathbf{x})} \right] \\ &= -\left(-\int_\Gamma \mathrm{d}\mathbf{x}\, p_A(\mathbf{x}) \ln p_A(\mathbf{x}) \right) - \int_\Gamma \mathrm{d}\mathbf{x}\, p_A(\mathbf{x}) \ln p_B(\mathbf{x}), \end{aligned} \tag{4.3.0.1}$$

where $H(p_A) = -\int_\Gamma \mathrm{d}\mathbf{x}\, p_A(\mathbf{x}) \ln p_A(\mathbf{x})$ is the entropy of the distribution under Hamiltonian A, and $-\int_\Gamma \mathrm{d}\mathbf{x}\, p_A(\mathbf{x}) \ln p_B(\mathbf{x})$ is the cross-entropy between the distribution under Hamiltonian A and that under Hamiltonian B. $KL(p_A||p_B)$ is also known as the relative entropy in Information theory. Similarly, we can also have

$$KL(p_B||p_A) = \int_\Gamma \mathrm{d}x\, p_B(\mathbf{x}) \ln \left[\frac{p_B(\mathbf{x})}{p_A(\mathbf{x})} \right]. \tag{4.3.0.2}$$

With the Boltzmann statistics,

$$p_A(\mathbf{x}) = e^{-\beta U(\mathbf{x})}/Q_A. \tag{4.3.0.3}$$

The relative entropies can be written as

$$\begin{aligned} KL(p_A||p_B) &= \int_\Gamma \mathrm{dx}\, \frac{e^{-\beta U_A(\mathbf{x})}}{Q_A} \ln\left[\frac{e^{-\beta U_A(\mathbf{x})}}{Q_A}\frac{Q_B}{e^{-\beta U_B(\mathbf{x})}}\right] \\ &= \int_\Gamma \mathrm{dx}\, \frac{e^{-\beta U_A(\mathbf{x})}}{Q_A} \left\{\beta\left[U_B(\mathbf{x}) - U_A(\mathbf{x})\right] - \beta\Delta A\right\} \qquad (4.3.0.4) \\ &= \beta\left\langle \Delta U\right\rangle_A - \beta\Delta A \qquad (4.3.0.5) \end{aligned}$$

and

$$KL(p_B||p_A) = -\beta\left\langle \Delta U\right\rangle_B + \beta\Delta A, \qquad (4.3.0.6)$$

which are exactly the dissipated work[110].

4.4 Mutual information

The definition of entropy

$$H(p) = -\int \mathrm{d}\mathbf{x}\, p(\mathbf{x}) \ln p(\mathbf{x}) \qquad (4.4.0.1)$$

can be extended to joint distributions as

$$H(p) = \iint \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}\, p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x},\mathbf{y}) \qquad (4.4.0.2)$$

The conditional entropy $H(\mathbf{Y}|\mathbf{X})$ is the amount of information needed to determine $\mathbf{Y}$ when $\mathbf{X}$ is known

$$\begin{aligned} H(\mathbf{Y}|\mathbf{X}) &= \int \mathrm{d}\mathbf{x}\, p(\mathbf{x}) H(\mathbf{Y}|\mathbf{X}=\mathbf{x}) \\ &= -\iint \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}\, p(\mathbf{x},\mathbf{y}) \log p(\mathbf{y}|\mathbf{x}). \qquad (4.4.0.3) \end{aligned}$$

The chain rule

$$H(\mathbf{X}_{1:n}) = H(\mathbf{X}_1) + H(\mathbf{X}_2|\mathbf{X}_1) + \cdots + H(\mathbf{X}_n|\mathbf{X}_{1:n-1}) \qquad (4.4.0.4)$$

can be easily derived from above. In general, $H(\mathbf{Y}|\mathbf{X}) \neq H(\mathbf{X}|\mathbf{Y})$, and $H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{Y})$ when $\mathbf{X}$ and $\mathbf{Y}$ are independent.

From the chain rule above, it can be seen that

$$H(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{Y}) + H(\mathbf{X}|\mathbf{Y}). \qquad (4.4.0.5)$$

By switching the terms, it can be found that

$$H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}), \qquad (4.4.0.6)$$

which can be interpreted as the entropy reduction in $\mathbf{X}$ when $\mathbf{Y}$ is known, or vice versa. The mutual information is thus defined

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) =& H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) \\ =& H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ =& \iint \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}\, p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}. \end{aligned} \tag{4.4.0.7}$$

It can also be written as

$$I(\mathbf{X};\mathbf{Y}) = KL\left(p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})\right). \tag{4.4.0.8}$$

5

Dimension Reduction

"Success is stumbling from failure to failure with no loss of enthusiasm."

– Winston Churchill

Natural data is often represented in high-dimensional spaces, leading to the "curse of dimensionality" challenge in analysis. This complexity makes it difficult for humans to visualize and interpret such data. However, the generation of these data usually occurs within a limited number of degrees of freedom. Identifying slowly varying order parameters, or collective variables (CVs), is crucial in the field of physical chemistry, especially for complex systems. These CVs are typically expected to exist on a low-dimensional manifold, capturing the slow dynamics of rare events amid a range of faster occurrences. Identifying effective CVs is challenging, often relying on intuition. According to Peters[111], an ideal CV should meet three criteria: (i) it should be a function of the instantaneous configuration space, excluding velocities; (ii) its value should change monotonically between two states, with corresponding isosurfaces creating non-intersecting dividing surfaces in the configuration space; (iii) it allows for projecting a free energy profile along it, ensuring that the reduced dynamics along the CV are consistent with those in the full phase space.

For a brief introduction to the modern dimension reduction methods or manifold learning methods, please refer to Ref. [112]. Specially, the generic problem addressed by dimension reduction, or manifold learning, is as follows: given a set of k observations $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^D$, find a set of points $\mathbf{y}_1, \ldots, \mathbf{y}_k$ in $\mathbb{R}^d$ ($d \ll D$) that serve as the best representation of $\mathbf{x}_i$. Most dimensionality reduction algorithms fit into either one of two broad categories: Matrix factorization (such as PCA) or Graph layout (such as t-SNE and UMAP).

But, how can one gauge a dimensionality reduction algorithm's performance effectively? The answer may not be unique, and some metrics should be considered.

- **Data reconstruction error** is the difference between the original data and the reconstructed data, which is obtained by applying the inverse transformation of the dimensionality reduction algorithm. The lower the data reconstruction error, the better the algorithm is at retaining the essential features of the original data. The data reconstruction error can be quantified in different ways, such as mean squared error, root mean squared error, or mean absolute error.

- **Data compression ratio** is the ratio of the size of the original data to the size of the reduced data. The higher the data compression ratio, the more efficient the algorithm is at reducing the dimensionality of the data. However, large data compression ratio often leads to high data reconstructure error. Therefore, one should always balance the trade-off between data compression and data reconstruction when choosing a dimensionality reduction algorithm.

- **Data visualization quality** can be used when the dimensionality of the data is reduced to two or three dimensions, which can be easily plotted and visualized with bare eyes to find how well the reduced data captures the patterns, clusters, and outliers in the original data.

- **Data classification accuracy** is the proportion of correctly predicted labels out of the total number of labels for a supervised learning problem. Different classifiers, such as logistic regression, k-nearest neighbors, or support vector machines, can be used to test the data classification accuracy. The higher the data classification accuracy, the better the algorithm is at preserving the discriminative power of the data.

- **Algorithm complexity and scalability**, such as time complexity, space complexity, or iteration complexity, refers to how fast and how well the algorithm can handle large and high-dimensional datasets. The algorithm complexity and scalability depend on factors such as the computational cost, the memory usage, and the convergence rate of the algorithm.

- **Algorithm suitability and robustness** means how well the algorithm fits the characteristics and the objectives of the data and the problem. The algorithm suitability and robustness depend on various aspects such as the data type, the data distribution, the data noise, and the data interpretation. One can use various methods to test the algorithm suitability and robustness, such as cross-validation, sensitivity analysis, or parameter tuning.

5.1 Principal Component Analysis

Principal component analysis (PCA) was first developed by Hotelling in 1933,[113] which is one of the dimension reduction methods that **linearly** transforms a data set consisting of a large number of interrelated variables to a new set of uncorrelated variables, the principal components (PCs), while retaining as much as possible of the variation present in the data set. The output PCs are ordered so that the first few retain most of the variation present in all of the original variables. There have been many excellent review and tutorial of this method. For a (probably) most recent one, please refer to Ref. [114]

Suppose that $\mathbf{x}$ is a vector of p random variables, of which the covariance matrix is $\mathbf{\Sigma}$. When $\mathbf{\Sigma}$ is unknown, it is often replaced by a sample variance $\mathbf{S}$. Let $\boldsymbol{\alpha}_k$ (for $k = 1, \ldots, p$) be the kth eigenvector of $\mathbf{\Sigma}$ corresponding to its kth largest eigenvalue of λ_k. The coordinate on the kth PC can be written as

$$z_k = \boldsymbol{\alpha}_k{}^{\mathrm{T}}\mathbf{x} = \sum_{j=1}^{p} \alpha_{kj} x_j, \tag{5.1.0.1}$$

where T denotes transpose. Normally, $\boldsymbol{\alpha}_k$ is chosen to have a unit length (*i.e.* $\boldsymbol{\alpha}_k{}^{\mathrm{T}}\boldsymbol{\alpha}_k = 1$). Then the variance of z_k, $\mathrm{var}(z_k)$, equals to λ_k.

To derive the form of the PCs, first consider $\boldsymbol{\alpha}_1$, which maximizes $\mathrm{var}\left[\boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{x}\right] = \boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_1 = 1$. Using the technique of Lagrange multipliers, it becomes to maximize

$$\boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_1 - 1), \tag{5.1.0.2}$$

where λ is a Lagrange multiplier. Differentiation with respect to $\boldsymbol{\alpha}_1$ gives

$$\mathbf{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = 0. \tag{5.1.0.3}$$

Thus, λ is an eigenvalue of $\mathbf{\Sigma}$, and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector. Also note that

$$\boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1{}^{\mathrm{T}}\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_1 = \lambda. \tag{5.1.0.4}$$

Therefore, in order to maximize $\boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_1$, λ must be the largest eigenvalue of $\mathbf{\Sigma}$.

Now, let us look at the second PC, $\boldsymbol{\alpha}_2\mathbf{x}$, which maximizes $\boldsymbol{\alpha}_2{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_2$ subject to being uncorrelated with the first PC, $\boldsymbol{\alpha}_1\mathbf{x}$, *i.e.* $\mathrm{cov}\left[\boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathrm{x}, \boldsymbol{\alpha}_2{}^{\mathrm{T}}\mathrm{x}\right] = 0$. Since

$$\mathrm{cov}\left[\boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathrm{x}, \boldsymbol{\alpha}_2{}^{\mathrm{T}}\mathrm{x}\right] = \boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{\alpha}_1, \tag{5.1.0.5}$$

any one of the following equations

$$\begin{aligned} \boldsymbol{\alpha}_1{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_2 = 0, \quad \boldsymbol{\alpha}_2{}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\alpha}_1 = 0 \\ \boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_2 = 0, \quad \boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0 \end{aligned} \tag{5.1.0.6}$$

could be used to specify the constraint. Using, for instance, the last one, as well as the normalization condition, the quantity to be maximized is

$$\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\left(\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{\alpha}_2 - 1\right) - \phi\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{\alpha}_1, \tag{5.1.0.7}$$

where λ and ϕ are Lagrange multipliers. Differentiation with respect to $\boldsymbol{\alpha}_2$ gives

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = \mathbf{0}. \tag{5.1.0.8}$$

Multiplying on the left by $\boldsymbol{\alpha}_1{}^{\mathrm{T}}$ gives

$$\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0. \tag{5.1.0.9}$$

Since the first two terms are zero and $\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{\alpha}_1 = 1$, it leads to $\phi = 0$. Therefore,

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = \mathbf{0}, \tag{5.1.0.10}$$

indicating that λ is an eigenvalue of $\boldsymbol{\Sigma}$ again, and $\boldsymbol{\alpha}_2$ the corresponding eigenvector. We can keep on doing this analysis for the third, fourth, ..., pth PCs, and show that λ_3, λ_4, ..., λ_p are the third, fourth largest, ..., and the smallest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$, ..., $\boldsymbol{\alpha}_p$ are the corresponding eigenvectors. Furthermore,

$$\mathrm{var}\left[\boldsymbol{\alpha}_k{}^{\mathrm{T}}\mathbf{x}\right] = \lambda_k \quad \text{for } k = 1, 2, \ldots, p. \tag{5.1.0.11}$$

Kernel Principal Component Analysis[115] (Kernel PCA): Let us suppose all the data have being mapped into feature space, $\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_l)$, and are centered, i.e. $\sum_{k=l}^{l} \boldsymbol{\phi}(\mathbf{x}_k) = 0$. PCA for the covariance matrix

$$\bar{C} = \frac{1}{l}\sum_{j=1}^{l} \boldsymbol{\phi}(\mathbf{x}_j)\boldsymbol{\phi}(\mathbf{x}_j)^{\mathrm{T}} \tag{5.1.0.12}$$

gives eigenvalues $\lambda > 0$ and eigenvectors $\mathbf{V}$ satisfying $\lambda\mathbf{V} = \bar{C}\mathbf{V}$. $\mathbf{V}$ lies in the span of $\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_l)$. Therefore, the eigen equation can be written as

$$\lambda(\boldsymbol{\phi}(\mathbf{x}_k) \cdot \mathbf{V}) = (\boldsymbol{\phi}(\mathbf{x}_k) \cdot \bar{C}\mathbf{V}) \quad \text{for all} \quad k = 1, \ldots, l, \tag{5.1.0.13}$$

and there exist coefficients $\alpha_1, \ldots, \alpha_l$ such that

$$\mathbf{V} = \sum_{i=1}^{l} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i). \tag{5.1.0.14}$$

Now let us define an $l \times l$ matrix K by

$$K_{ij} := (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j)), \tag{5.1.0.15}$$

and take Eq. 5.1.0.12 and 5.1.0.14 into Eq. 5.1.0.13, it leads to

$$l\lambda K\boldsymbol{\alpha} = K^2\boldsymbol{\alpha}, \tag{5.1.0.16}$$

where $\boldsymbol{\alpha}$ is the column vector with elements $\alpha_1, \ldots, \alpha_l$. Its solutions can be found by solving the eigenvalue problem

$$l\lambda\boldsymbol{\alpha} = K\boldsymbol{\alpha} \tag{5.1.0.17}$$

for nonzero eigenvalues. Enforcing the eigenvectors corresponding to nonzero eigenvalues to be normalized, i.e. $(\mathbf{V}^k \cdot \mathbf{V}^k) = 1$, the solution $\boldsymbol{\alpha}^k$ must satisfy

$$1 = \sum_{i,j=1}^{l} \alpha_i^k \alpha_j^k (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j)) = (\boldsymbol{\alpha}^k \cdot K\boldsymbol{\alpha}^k) = l\lambda_k(\boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^k). \tag{5.1.0.18}$$

Projections of the data in the feature space $\boldsymbol{\phi}(\mathbf{x})$ onto the eigenvectors $\mathbf{V}^k$ can be computed via

$$(\mathbf{V}^k \cdot \boldsymbol{\phi}(\mathbf{x})) = \sum_{i=1}^{l} \alpha_i^k (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x})), \tag{5.1.0.19}$$

which gives the principal components.

The algorithm for kernel PCA:

1. Computing the dot product matrix $K_{ij} = (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j))$
2. Calculating the eigenvalue and eigenvectors of K (Eq. 5.1.0.17), and normalizing the eigenvector expansion coefficients $\boldsymbol{\alpha}^k$ (Eq. 5.1.0.18)
3. Extracting principal components in correspondence to the kernel K for any point $\mathbf{x}$ by projection onto the eigenvectors (Eq. 5.1.0.19) components

5.2 Multidimensional Scaling

Multidimensional scaling (MDS) is a method that represents measurements of similarity (or dissimilarity) among pairs of objects as distances between points of a low-dimensional multidimensional space. Since MDS is often used for data visualization, the mapped space usually has a very low dimension, for instance 2.

There are two main variations of MDS: metric MDS and non-metric MDS. Metric MDS aims to preserve the actual distances or dissimilarities between objects as accurately as possible in the lower-dimensional space. Metric MDS assumes that the pairwise distances or dissimilarities are metric, meaning they satisfy the triangle inequality. It uses techniques such as eigenvalue decomposition or optimization algorithms to find the configuration of points that minimizes the difference between the original distances and the distances in the lower-dimensional space. Non-metric MDS, also known as ordinal MDS, does not assume that the pairwise distances or dissimilarities are metric. Instead, it focuses on preserving the ordinal relationships between objects, meaning it tries to maintain the rank order of distances or dissimilarities rather than their actual values.

Here, we only look at metric MDS. For two vectors $\mathbf{a}$ and $\mathbf{b}$, the distance between them can be written as

$$d^2(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})^{\mathrm{T}}(\mathbf{a} - \mathbf{b}) = \mathbf{a}^{\mathrm{T}}\mathbf{a} + \mathbf{b}^{\mathrm{T}}\mathbf{b} - 2 \times (\mathbf{a}^{\mathrm{T}}\mathbf{b}), \tag{5.2.0.1}$$

where $\mathbf{a}^{\mathrm{T}}\mathbf{b}$ is the scalar product between $\mathbf{a}$ and $\mathbf{b}$. For n observations in a m-dimensional space ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^m$) stored in a matrix denoted by $\mathbf{X}$, which has a shape of $m \times n$, the cross product matrix $\mathbf{S}$ is then obtained as

$$\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}, \tag{5.2.0.2}$$

which has a shape of $n \times n$. The squared Euclidean distance matrix

$$\mathbf{D} = \mathbf{s} \cdot \mathbf{1}^{\mathrm{T}} + \mathbf{1} \cdot \mathbf{s}^{\mathrm{T}} - 2\mathbf{S}, \tag{5.2.0.3}$$

where $\mathbf{s} = \mathrm{diag}(S_{ii})$ is the $n \times 1$ vector of the diagonal element of $\mathbf{S}$ and $\mathbf{1}$ is a $n \times 1$ vector of 1s. It shows that the cross product matrix $\mathbf{S}$ can be computed from the distance matrix $\mathbf{D}$, which is the basic idea of metric MDS. Clearly, the solutions are not unique, since distances are invariant with respect to any change of origin. Therefore, constraints must be imposed on the calculations of $\mathbf{X}$. An obvious choice is to choose the origin of the distance as the center of gravity of the dimensions.

Defining a $n \times n$ centering matrix

$$\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1} \cdot \mathbf{1}^{\mathrm{T}}, \tag{5.2.0.4}$$

the cross-product matrix is obtained from matrix $\mathbf{D}$ as

$$\tilde{\mathbf{S}} = -\frac{1}{2}\mathbf{H}\mathbf{D}\mathbf{H}^{\mathrm{T}}. \tag{5.2.0.5}$$

To show the relationship between $\tilde{\mathbf{S}}$ and $\mathbf{S}$, let us take Eq. 5.2.0.3 into Eq. 5.2.0.5, we find

$$\tilde{\mathbf{S}} = -\frac{1}{2}\mathbf{H}\left(\mathbf{s}\cdot\mathbf{1}^{\mathrm{T}} + \mathbf{1}\cdot\mathbf{s}^{\mathrm{T}} - 2\mathbf{S}\right)\mathbf{H}^{\mathrm{T}} \tag{5.2.0.6}$$

Note that $\mathbf{1}^{\mathrm{T}}\cdot\mathbf{H}^{\mathrm{T}} = \mathbf{1}^{\mathrm{T}}\cdot\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^{\mathrm{T}}\right) = 0$, it yields

$$\tilde{\mathbf{S}} = \mathbf{H}\mathbf{S}\mathbf{H}^{\mathrm{T}}. \tag{5.2.0.7}$$

The eigen-decomposition of this matrix gives

$$\tilde{\mathbf{S}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}, \tag{5.2.0.8}$$

where $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. The best Euclidean approximation of the original distance matrix is thus obtained as

$$\tilde{\mathbf{Y}} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^{\mathrm{T}}. \tag{5.2.0.9}$$

In practice, one often chooses the top k nonzero eigenvalues of $\tilde{\mathbf{S}}$ and build a k-dimensional Euclidean embedding of data $\tilde{\mathbf{Y}}_k = {\mathbf{\Lambda}_k}^{1/2}{\mathbf{U}_k}^{\mathrm{T}}$, where

$$\begin{aligned} {\mathbf{U}_k}^{\mathrm{T}} &= [u_1, \ldots, u_k]^{\mathrm{T}}, \quad u_k \in \mathbb{R}^n, \\ \mathbf{\Lambda}_k &= \mathrm{diag}(\lambda_1, \ldots, \lambda_k) \end{aligned}$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$.

5.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA), aka Fisher linear discriminant analysis, was originally developed in 1936 by Ronald A. Fisher[116]. It is a dimensionality reduction method for classification problems that preserves as much of the class discriminatory information as possible by maximizing the ratio of the between-class variance to the within-class variance. Closely related to PCA, LDA is also based on linear transformations.

Given the original data set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ represents the i^{th} sample with M features ($\mathbf{x}_i \in \mathbb{R}^M$), and N is the total number of samples. Assume the data samples are catagorized into C classes, $\mathbf{X} = [\boldsymbol{\omega_1}, \boldsymbol{\omega_2}, \cdots, \boldsymbol{\omega_C}]$, and class j contains n_j samples. The sum of n_j equals to the total number of samples.

$$N = \sum_{j=1}^{C} n_j. \tag{5.3.0.1}$$

LDA seeks to obtain a transformation of $\mathbf{X}$ to $\mathbf{Y}$ through projecting the samples in $\mathbf{X}$ onto a hyperplane with dimension $C-1$. The sample mean for class $\boldsymbol{\mu_j}$ is calculated as

$$\boldsymbol{\mu_j} = \frac{1}{n_j} \sum_{\mathbf{x}_i \in \boldsymbol{\omega_j}} \mathbf{x}_i, \tag{5.3.0.2}$$

and the sample mean of all the classes is computed as

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i = \sum_{j=1}^{C} \frac{n_j}{N} \boldsymbol{\mu}_j. \tag{5.3.0.3}$$

Their projections, $\mathbf{m}_i$ and $\mathbf{m}$, are computed via

$$\mathbf{m}_j = \mathbf{W}^{\mathrm{T}} \boldsymbol{\mu}_j \tag{5.3.0.4}$$

and

$$\mathbf{m} = \mathbf{W}^{\mathrm{T}} \boldsymbol{\mu}, \tag{5.3.0.5}$$

where $\mathbf{W}$ is the transformation matrix of LDA.

To calculate the between-class variance (scatter) $\mathbf{S}_B$, the separation distance between different classes will be calculated by

$$\mathbf{S}_B = \sum_{j=1}^{C} n_j \mathbf{S}_{Bj}, \tag{5.3.0.6}$$

where

$$\mathbf{S}_{Bj} = (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}}. \tag{5.3.0.7}$$

In the projected space,

$$(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^{\mathrm{T}} = \mathbf{W}^{\mathrm{T}}\mathbf{S}_{Bj}\mathbf{W}. \tag{5.3.0.8}$$

The within-class variance represents the difference between the mean and the samples within each class. The within-class variance (scatter) of each class $\mathbf{S}_{Wj}$ is calculated as

$$\begin{aligned}
&\sum_{j=1}^{C} \sum_{\mathbf{x}_i \in \boldsymbol{\omega_j}} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^{\mathrm{T}} \\
&\quad = \sum_{j=1}^{C} \sum_{\mathbf{x}_i \in \boldsymbol{\omega_j}} \left(\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_j\right)\left(\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_j\right)^{\mathrm{T}} \\
&\quad = \sum_{j=1}^{C} \sum_{\mathbf{x}_i \in \boldsymbol{\omega_j}} \mathbf{W}^{\mathrm{T}}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^{\mathrm{T}}\mathbf{W} \\
&\quad = \sum_{j=1}^{C} \mathbf{W}^{\mathrm{T}}\mathbf{S}_{Wj}\mathbf{W} \\
&\quad = \mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}
\end{aligned} \tag{5.3.0.9}$$

The transformation matrix $\mathbf{W}$ can be calculated by maximizing the ratio of the determinant of $\mathbf{S}_B$ to the determinant of $\mathbf{S}_W$ in the projected space (known as Fishers criterion)

$$\arg\max_{\mathbf{W}} \frac{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}|}. \tag{5.3.0.10}$$

Note that the determinant of the co-variance matrix tells us how much variance a class has, and it has the same value under any ortho-normal projection. So Fishers criterion tries to find the projection that maximizes the variance of the class means and minimzses the variance of the individual classes.

This optimization problem can have infinitely number of solutions with the same objective function value, due that for a solution $\mathbf{W}$ all the vectors $c \cdot \mathbf{W}$ give exactly the same value for the objective function. Without loss of generality, the constraint

$$\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W} = 1 \tag{5.3.0.11}$$

can be applied. Then the problem becomes

$$\arg\max_{\mathbf{W}} \mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W} \tag{5.3.0.12}$$

s.t.

$$\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W} = 1. \tag{5.3.0.13}$$

The Lagrangian for this optimization is

$$\mathcal{L}_{LDA} = \mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W} - \lambda(\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W} - 1). \tag{5.3.0.14}$$

Minimizing $\mathcal{L}_{LDA}$ with respect to $\mathbf{W}$ leads to

$$\mathbf{S}_B\mathbf{W} = \mathbf{S}_W\mathbf{W}\Lambda. \tag{5.3.0.15}$$

Multiplying on both sides by the inverse of $\mathbf{S}_W$ (if $\mathbf{S}_W$ is a non-singular), it becomes

$$\mathbf{S}_W{}^{-1}\mathbf{S_B}\mathbf{W} = \mathbf{W}\Lambda. \tag{5.3.0.16}$$

Then the Fisher's criterion is maximized when the projection matrix $\mathbf{W}$ is composed of the eigenvectors of $\mathbf{S}_W{}^{-1}\mathbf{S_B}$, and Λ are the associated eigenvalues. Notice that there will be at most $C-1$ eigenvectors with non-zero real corresponding eigenvalues. Each eigenvectors represents one axis of the LDA space, and the corresponding eigenvalues represents the discriminatory ability between different classes. Thus, the eigenvectors with the k highest eigenvalues are used to construct a lower dimensional space.

5.4 CUR Decomposition

CUR decomposition, developed by Mahoney and Drineas[117], finds a low-rank approximation of matrix A as the product of three matrices C, U, and R, where C is a matrix consisting of selected columns of the original matrix, R is a matrix consisting of selected rows of the original matrix, and U is a matrix that ideally reconstructs the original matrix from C and R. Usually the CUR is designed to be a rank-k approximation, which requires that C contains k columns of A, R contains k rows of A, and U is a k-by-k matrix. The CUR matrix decomposition technique was developed to provide more interpretable and computationally efficient alternatives to SVD in principal component analysis (PCA), despite the fact that CUR is usually less accurate than SVD.

The fundamental questions of the CUR decomposition methods are: 1) Which columns of A should be used to build C? Which rows should be used for R? 2) How to obtain the best U given C and R?

5.5 Independent Component Analysis

TODO: to check the equations.

Independent component analysis (ICA) is a special case of blind source separation, which is used for separating a multivariate signal into additive subcomponents. ICA is considered as an extension of the principal component analysis (PCA, see section 5.1) technique. However, PCA searches uncorrelated components while ICA looks for independent components. For a not very recent review, please refer to Ref. [118].

ICA is built based on three assumptions.

1. **Independence**: The source signals are independent of each other.
2. **Non-Gaussianity**: The mixed signals are Gaussian, but the values in each source signal have non-Gaussian distributions.
3. **Complexity**: Mixed signals are more complex than source signals.

The *signals* must be preprocessed before they can be projected to find the unmixing matrix and the *sources*. The preprocessing steps include centering, whitening and dimensionality reduction. Suppose a random r-dimensional vector $\mathbf{X} = (X_1, \ldots, X_r)^{\mathrm{T}}$ has been detected, of which the mean and the covariance matrix are $\mathrm{E}\{\mathbf{X}\} = \boldsymbol{\mu}$ and $\mathrm{cov}\{\mathbf{X}\} = \boldsymbol{\Sigma}_{\mathbf{XX}}$, respectively. These *signals* should be preprocessed by first centering so that they have zero mean, and then by sphering (or whitening) so that the uncorrelated components have unit variances. Using spectral decomposition, we have $\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$, where the columns of the unitary matrix $\mathbf{U}$ are the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{XX}}$, and the diagonal elements of the diagonal matrix $\boldsymbol{\Lambda}$ are the corresponding eigenvalues. The centered and sphered version of $\mathbf{X}$ can be given by

$$\mathbf{Z} \leftarrow \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu}). \tag{5.5.0.1}$$

In the above, we have assumed that both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{\mathbf{XX}}$ are known. When they are unknown as in most cases in practice, we take n observations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, on $\mathbf{X}$ to compute $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^{\mathrm{T}}$. To reduce the dimensionality of the data, only the first $J < r$ sphered variables (corresponding to the eigenvectors with the largest magnitudes of the eigenvalues) are retained, where J is chosen to explain a certain (high) proportion of the total variance as we do in PCA.

The observed data set $\mathbf{X} = (X_1, \ldots, X_r)^{\mathrm{T}}$ are generated by

$$\mathbf{X} = f(\mathbf{S}) + \mathbf{e}, \tag{5.5.0.2}$$

where $\mathbf{S} = (S_1, \ldots, S_m)^{\mathrm{T}}$ is an unknown vector of source whose components are independent latent variables, $f : \mathbb{R}^m \to \mathbb{R}^r$ is an unknowing mixing function, and $\mathbf{e}$ represents measurement noise with a zero mean. We assume that $\mathbf{E}(\mathbf{S}) = \mathbf{0}$ and $\mathbf{cov}(\mathbf{S}) = \mathbf{I}_m$.

Nomenclature: If f is taken to be a linear (nonlinear) function, Eq. 5.5.0.2 is described as a *linear* (*nonlinear*) ICA model. In most applications of ICA, it is assumed that the additive noise $\mathbf{e}$ is zero, and all noise in the model is to be associated with the components of the source. Such a model is referred to as *noiseless* ICA. Otherwise, it is referred to as *noisy* ICA. In most ICA applications, $\mathbf{X}$ is regarded as a stochastic process $\mathbf{X}(t) = (X_1(t), \ldots, X_r(t))^{\mathrm{T}}$, where t is a time or index parameter. In the linear noiseless ICA model with temporally structured sources and *static* (time-independent) mixing, the model is written as $\mathbf{X}(t) = \mathbf{AS}(t)$, where $\mathbf{S}(t)$ is assumed to be a *stationary sources*. If $\mathbf{A}$ is also time dependent, this model is referred to as *dynamic mixing*. In the following, we only consider *static* mixing.

With the observed data set $\mathbf{X} = (X_1, \ldots, X_r)^{\mathrm{T}}$ as an input, the task of ICA is to transform $\mathbf{X}$ into a vector of source with maximally independent components $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$ using a linear static transformation $\mathbf{W}$ as $\mathbf{Y} = \mathbf{WX}$. The independence is measured by some function $F(\mathbf{Y})$. ICA finds the independent components (also called factors, latent variables or sources) by maximizing the statistical independence of the estimated components. There are many ways to define a proxy for independence, and the choice governs the form of the ICA algorithm, such as by 1) minimization of mutual information, which used measures like Kullback-Leibler Divergence and maximum entropy, 2) maximization of non-Gaussianity, which uses kurtosis and negentropy, and 3) using maximum likelihood estimation method.

5.5.1 Maximizing the non-Gaussianity

Non-Gaussianity can be measured by kurtosis and negative entropy.

Kurtosis

The signal (sources) can be extracted by finding the orientation of the weight vectors which maximizes the kurtosis of the extracted signal. *Although it is simple to calculate, it is sensitive to outliers. Therefore, it is not a robust way to measure the non-Gaussianity.* The kurtosis (K) for any probability density function is defined as

$$K(\mathbf{X}) = \mathrm{E}\left[\mathbf{X}^4\right] - 3\left[\mathrm{E}\left[\mathbf{X}^2\right]\right]^2. \tag{5.5.1.1}$$

The normalized kurtosis ($\hat{K}$) is the ratio between the fourth and second central moments, and it is given by

$$\hat{K} = \frac{\mathrm{E}\left[\mathbf{X}^4\right]}{\left[\mathrm{E}\left[\mathbf{X}^2\right]\right]^2} - 3 \approx \frac{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2\right)^2} - 3 \tag{5.5.1.2}$$

For whitened data $\mathbf{Z}$ with a unit variance and zero mean,

$$K(\mathbf{Z}) = \hat{K}(\mathbf{Z}) = \mathrm{E}\left[\mathbf{Z}^4\right] - 3. \tag{5.5.1.3}$$

The ICs can be found by maximizing kurtosis of extracted signals $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{Z}$, which can be written as

$$K(\mathbf{Y}) = \mathrm{E}\left[\left(\mathbf{W}^{\mathrm{T}}\mathbf{Z}\right)^4\right] \tag{5.5.1.4}$$

with the gradient being

$$\frac{\partial K(\mathbf{W}^{\mathrm{T}}\mathbf{Z})}{\partial \mathbf{W}} = c\mathrm{E}\left[\mathbf{Z}\left(\mathbf{W}^{\mathrm{T}}\mathbf{Z}\right)^3\right]. \tag{5.5.1.5}$$

The weight vector is updated iteratively via

$$\mathbf{W}_{new} = \mathbf{W}_{old} + \eta\mathrm{E}\left[\mathbf{Z}\left(\mathbf{W}^{\mathrm{T}}\mathbf{Z}\right)^3\right], \tag{5.5.1.6}$$

and

$$\mathbf{W}_{new} = \frac{\mathbf{W}_{new}}{\|\mathbf{W}_{new}\|}, \tag{5.5.1.7}$$

since $\|\mathbf{W}\| = 1$.

Negative entropy

Negative entropy, or negentropy for short, is defined as

$$J(\mathbf{Y}) = H(\mathbf{Y}_{Gaussian}) - H(\mathbf{Y}), \tag{5.5.1.8}$$

where $H(\mathbf{Y}_{Gaussian})$ is the entropy of a Gaussian random variable whose covariance matrix is equal to the covariance matrix of $\mathbf{Y}$. The entropy of a random variable $\mathbf{Y}$ which has N possible outcomes is

$$H(\mathbf{Y}) = -\mathrm{E}\left[\log p_y(y)\right] = -\sum_i^N p_y(y^i)\log p_y(y^i), \tag{5.5.1.9}$$

where $p_y(y^i)$ is the probability of the event y^i, $i = 1, \ldots, N$. The negentropy is always nonnegative because the entropy of a Gaussian distribution is the maximum among all other random distributions with the same variance. It is zero only when all variables are Gaussian distributed, i.e., $H(\mathbf{Y}_{Gaussian}) = H(\mathbf{Y})$. Moreover, it is invariant for invertible linear transformation and scale-invariant. However, calculating the entropy from a finite data is computationally difficult. Hence, different approximations have been introduced. For example,

$$J(y) \approx \sum_{i=1}^{p} k_i \left(\mathrm{E}\left[G_i(y)\right] - \mathrm{E}\left[G_i(v)\right]\right)^2, \tag{5.5.1.10}$$

where k_i are some positive constants, v indicates a Gaussian variable with zero mean and unit variance, G_i represent some quadratic function. The function G has different choices such as

$$G_1(y) = \frac{1}{a}\log\cosh(a_1 y) \quad G_2(y) = -\exp(-y^2/2), \tag{5.5.1.11}$$

where $1 \leq a_1 \leq 2$.

5.5.2 Minimization of mutual information

The amount of mutual information between the m components of $\mathbf{Y}$ can be written as

$$I(\mathbf{Y}) = c - \sum_{j=1}^{m} J(Y_j), \tag{5.5.2.1}$$

where $c = mH(\mathbf{Y}_{Gaussian}) - H(\mathbf{X})$ does not depend on the unmixing matrix $\mathbf{W}$ and is a constant. Therefore, minimizing the mutual information between the components of $\mathbf{Y}$ is equivalent to maximizing the sum of the negentropies of the independent components of $\mathbf{Y}$.

5.5.3 Maximum likelihood

For a noise-free ICA model, $\mathbf{X} = \mathbf{AS}$. Hence,

$$P_{\mathbf{X}}(\mathbf{X}) = \frac{P_{\mathbf{S}}(\mathbf{S})}{|\det\mathbf{A}|} = |\det\mathbf{W}| P_{\mathbf{S}}(\mathbf{S}). \tag{5.5.3.1}$$

For independent source signals, $P_{\mathbf{S}}(\mathbf{S}) = \prod_i p_i(\mathbf{s}_i)$, $P_{\mathbf{X}}(\mathbf{X})$ becomes

$$P_{\mathbf{X}}(\mathbf{X}) = |\det\mathbf{W}| \prod_i p_i(\mathbf{s}_i) = |\det\mathbf{W}| \prod_i p_i(\mathbf{w}_i^T\mathbf{X}). \tag{5.5.3.2}$$

Given T observations of $\mathbf{X}$, the likelihood of $\mathbf{W}$ is given by

$$\mathcal{L}(\mathbf{W}) = \prod_t^T |\det\mathbf{W}| \prod_i^p p_i(\mathbf{w}_i^T\mathbf{x}(t)). \tag{5.5.3.3}$$

Usually, a log-likelihood is preferred:

$$\log\mathcal{L}(\mathbf{W}) = \sum_t^T \sum_i^p \log p_i(\mathbf{w}_i^T\mathbf{x}(t)) + T\log|\det\mathbf{W}|, \tag{5.5.3.4}$$

or

$$\begin{aligned}\frac{1}{T}\log\mathcal{L}(\mathbf{W}) =& \mathrm{E}\left[\sum_t^T \sum_i^p \log p_i(\mathbf{w}_i^T\mathbf{x}(t))\right] + \log|\det\mathbf{W}|, \\ =& -\sum_i H(\mathbf{w}_i^T\mathbf{X}) + \log|\det\mathbf{W}|.\end{aligned} \tag{5.5.3.5}$$

Therefore, the likelihood and mutual information are approximately equal, and they only differ by the sign and an additive constant.

5.6 Isometric Feature Mapping (Isomap)

Isomap, introduced in 2000 for the first time, is a nonlinear generalization of the MDS algorithm in which Euclidean distances are replaced by geodesic distances.[119] Isomap seeks a mapping such that the Euclidean distance in the transformed space match the corresponding geodesic distance between data points. However, the geometric structure of the given data is unknown usually. In order to obtain the geodesic distance between the points, it has been assumed that, in a small neighborhood, the Euclidean distance is a good approximation for the geodesic distance. While for the points far apart, the geodesic distance is approximated as the sum of Euclidean distances along the shortest connecting path.

The first step is to build a weighted neighborhood graph $G(\mathcal{V}, \mathcal{E})$ from the given data by connecting only nearby points, where the vertices or nodes, $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, are the input data and the edges, $\mathcal{E} = \{e_{ij}\}$, indicate the neighborhood relationship between the points. The weight w_{ij} of edge e_{ij} equals to the distance d_{ij} between those points if they are close to each, or 0 otherwise. Closeness is defined either by the ϵ-approach, if $\|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon$ where $\epsilon > 0$, or by the K-nearest neighbors. Then, Dijkstra's algorithm or Floyd's algorithm is applied with the nearest neighbor graph G to find the shortest-path distances ($d_G(i, j)$) for all pairs of data points. Finally, MDS is applied to the distance matrix $d_G(i, j)$ to find a k-dimensional representation $\mathbf{Y}$ of the original data.

Isomap algorithm is sensitive to noise in the data.

5.7 Locally Linear Embedding (LLE)

Locally linear embedding was introduced by Roweis and Saul in 2000.[120] It differs from Isomap by eliminating the need to estimate pairwise distances between widely separated data points.

Suppose sufficient data have been sampled from some underlying manifold, which are denoted as $\{\mathbf{x}_i\} \in \mathbb{R}^D$, for $i \in [1, N]$. It can be expected that each data point and its neighbors lie on or close to a locally linear patch of the manifold. Then each data point is reconstructed linearly from its neighbors, and the reconstruction errors measured by the cost function

$$\epsilon(\mathbf{W}) = \sum_i \|\mathbf{x}_i - \sum_j W_{ij}\mathbf{x}_j\|^2. \tag{5.7.0.1}$$

The weights $\mathbf{W}$ can be obtained by minimizing the cost function subject to two constraints: first, $W_{ij} = 0$ if $\mathbf{x}_j$ does not belong to the set of neighbors of $\mathbf{x}_i$; second, the rows of $\mathbf{W}$ sum to one: $\sum_j W_{ij} = 1$. Formally, we minimize the Lagrangian function

$$f(\mathbf{x}_i) = \mathbf{w}_i^{\mathrm{T}} G_i \mathbf{w}_i - \lambda\left(\mathbf{1}_n^{\mathrm{T}}\mathbf{w}_i - 1\right) \tag{5.7.0.2}$$

with respect to $\mathbf{w}_i$, where $G_i = (G_{i,jk})$ is an $(n \times n)$ Gram matrix with $G_{i,jk} = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_k)$. Minimizing $f(\mathbf{x}_i)$ with respect to $\mathbf{w}_i$ yields

$$\widehat{\mathbf{w}}_i = \frac{\lambda}{2} G_i^{-1}\mathbf{1}_n. \tag{5.7.0.3}$$

Multiplying both sides of this equation from the left by $\mathbf{1}_n^{\mathrm{T}}$ and using the normalization condition $\mathbf{1}_n^{\mathrm{T}}\mathbf{w}_i = 1$, we find

$$\frac{\lambda}{2} = \frac{1}{\mathbf{1}_n^{\mathrm{T}}\mathbf{G}_i^{-1}\mathbf{1}_n}. \tag{5.7.0.4}$$

Then, the optimal weights can be rewritten as

$$\widehat{\mathbf{w}}_i = \frac{\mathbf{G}_i^{-1}\mathbf{1}_n}{\mathbf{1}_n^{\mathrm{T}}\mathbf{G}_i^{-1}\mathbf{1}_n}. \tag{5.7.0.5}$$

In the final step of this algorithm, each high-dimensional data points $\mathbf{x}_i$ is mapped to a low-dimensional observation $\mathbf{y}_i$ by minimizing the embedding cost function

$$\phi(\mathbf{Y}) = \sum_i \|\mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j\|^2 \tag{5.7.0.6}$$

with fixed $\mathbf{W}$. It can be converted into a $N \times N$ eigenvalue problem by the transformation

$$\phi(\mathbf{Y}) = \sum_{ij}(\delta_{ij} - W_{ij} - W_{ji} + \sum_k W_{ki}W_{kj})\mathbf{y}_i \cdot \mathbf{y}_j, \tag{5.7.0.7}$$

in which we have assumed/forced

$$\sum_i \mathbf{y}_i = 0 \quad \text{and} \quad \frac{1}{N}\sum_i \mathbf{y}_i \otimes \mathbf{y}_i = \mathbf{I}. \tag{5.7.0.8}$$

The smallest eigenvalue is zero with corresponding eigenvector $\mathbf{v}_n = n^{-1/2}\mathbf{1}_n$. The next d smallest eigenvectors define the embedding coordinates.

5.8 Laplacian Eigenmaps

Laplacian eigenmaps was developed by Belkin and Niyogi in 2001.[121]

Given k points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ in $\mathbb{R}^t$, a weighted graph $G = (V, E)$ with k nodes, one for each point, and a set of edges connecting neighboring points to each other are constructed. Whether two points are neighbors of each other is determined by their closeness in either of the two ways:

- ϵ-neighborhoods: With a prechosen parameter ϵ, nodes i and j are connected by an edge if $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \epsilon$.
- n-nearest neighbors: Given parameter n, Nodes i and j are connected by an edge if i is among n nearest neighbors of j or j is among n nearest neighbors of i.

Now the weight for each edge can be calculated in either of the two ways:

- Heat kernel: If nodes i and j are connected, set

$$W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, \tag{5.8.0.1}$$

  where σ is the kernel bandwidth.

- Simple-minded: $W_{ij} = 1$ if and only if vertices i and j are connected by an edge.

Compute eigenvalues and eigenvectors for the generalized eigenvector problem:

$$L\mathbf{y} = \lambda D\mathbf{y} \tag{5.8.0.2}$$

where D is a diagonal weight matrix with $D_{ii} = \sum_j W_{ji}$, and $L = D - W$ is the Laplacian matrix. Let $\mathbf{y}_0, \ldots, \mathbf{y}_{k-1}$ be the solutions of the equation above, ordered by their eigenvalues with $\mathbf{y}_0$ having the smallest eigenvalue (in fact 0). The image of $\mathbf{x}_i$ under the embedding into the lower dimensional space $\mathbf{R}^m$ is given by $(\mathbf{y}_1(i), \ldots, \mathbf{y}_m(i))$.

Justification for the process above is that the points connected on the graph should stay as close as possible after embedding, which means we should minimize the objective function

$$\sum_{i,j \in E} (\mathbf{y_i} - \mathbf{y}_j)^2 W_{ij} \tag{5.8.0.3}$$

with respect to $\mathbf{y}_1, \ldots, \mathbf{y}_n$ subject to appropriate constraints. It means that when W_{ij} is large, $\mathbf{x}_i$ and $\mathbf{x}_j$ are very close to each other. Then, $\mathbf{y}_i$ and $\mathbf{y}_j$ must still be close. On the contrary, when W_{ij} is small, meaning that $\mathbf{x}_i$ and

$\mathbf{x}_j$ are far away from each other, then there is much flexibility in putting $\mathbf{y}_i$ and $\mathbf{y}_j$ on the line. For any $\mathbf{y}$, we have

$$\begin{aligned}
\frac{1}{2}\sum_{i,j\in E}(\mathbf{y}_i-\mathbf{y}_j)^2W_{ij} =& \frac{1}{2}\sum_{i,j\in E}(\mathbf{y}_i^2+\mathbf{y}_j^2-2\mathbf{y}_i\mathbf{y}_j)W_{ij}\\
=& \frac{1}{2}\sum_i \mathbf{y}_i^2\sum_j W_{ij}+\frac{1}{2}\sum_j \mathbf{y}_j^2\sum_i W_{ij}-\sum_{i,j\in E}\mathbf{y}_i\mathbf{y}_jW_{ij}\\
=& \frac{1}{2}\sum_i \mathbf{y}_i^2D_{ii}+\frac{1}{2}\sum_j \mathbf{y}_j^2D_{jj}-\sum_{i,j\in E}\mathbf{y}_i\mathbf{y}_jW_{ij}\\
=& \sum_i \mathbf{y}_i^2D_{ii}-\sum_{i,j\in E}\mathbf{y}_i\mathbf{y}_jW_{ij}\\
=& \sum_{i,j\in E}\mathbf{y}_iD_{i,j}\mathbf{y}_j-\sum_{i,j\in E}\mathbf{y}_i\mathbf{y}_jW_{ij}\\
=& \sum_{i,j\in E}\mathbf{y}_i(D_{ij}-W_{ij})\mathbf{y}_j\\
=& \mathbf{y}^{\mathrm{T}}L\mathbf{y}. \qquad (5.8.0.4)
\end{aligned}$$

Therefore, the minimization problem reduces to $\arg\min_{\substack{\mathbf{y}\\ \mathbf{y}^{\mathrm{T}}D\mathbf{y}=1}} \mathbf{y}^{\mathrm{T}}L\mathbf{y}$. The constraint $\mathbf{y}^{\mathrm{T}}D\mathbf{y}=1$ removes an arbitrary scaling factor in the embedding. L is semi-definite defined, and the vector $\mathbf{y}$ that minimizes the objective function is given by the minimum eigenvalue solution to the generalized eigenvalue problem $L\mathbf{y}=\lambda D\mathbf{y}$. Alternatively, using the Lagrangian multiplier and minimization with respect to $\mathbf{y}$, we have

$$\frac{\partial}{\partial \mathbf{y}}\left[\mathbf{y}^{\mathrm{T}}L\mathbf{y}+\lambda\left(\mathbf{y}^{\mathrm{T}}D\mathbf{y}-1\right)\right]=0, \qquad (5.8.0.5)$$

and it leads to

$$L\mathbf{y}=\lambda D\mathbf{y}. \qquad (5.8.0.6)$$

Since All the rows (and columns) sum to 0, it is easy to see that $\mathbf{y}=\mathbf{1}$ (all 1s) is an eigenvector with eigenvalue 0. To eliminate this trivial solution which collapses all vertices of G onto the real number 1, an additional constraint of orthogonality must be imposed to obtain

$$\mathbf{y}_{opt}=\arg\min_{\substack{\mathbf{y}\\ \mathbf{y}^{\mathrm{T}}D\mathbf{y}=1\\ \mathbf{y}^{\mathrm{T}}D\mathbf{1}=0}} \mathbf{y}^{\mathrm{T}}L\mathbf{y}. \qquad (5.8.0.7)$$

Thus, the solution $\mathbf{y}_{opt}$ is now given by the eigenvector with smallest non-zero eigenvalue. More generally, the embedding of the graph into $\mathbb{R}^m$ $(m>1)$ is given by the $m\times k$ matrix $Y=[\mathbf{y}_1\,\mathbf{y}_2\cdots\mathbf{y}_k]$. It reduces to

$$Y_{opt}=\arg\min_{\substack{Y\\ YDY^{\mathrm{T}}=I}} \mathrm{tr}\left(YLY^{\mathrm{T}}\right). \qquad (5.8.0.8)$$

The constraint is used to prevent a collapse onto a subspace of fewer than $m - 1$ dimensions.

5.9 Diffusion Map

A diffusion map was developed by Coifman et al. in 2005[122–124], which computes a family of embeddings of a data set into Euclidean space (often low-dimensional) whose coordinates can be computed from the eigenvectors and eigenvalues of a diffusion operator on the data, while ensuring that the diffusion distance in the original space between points is well approximated by the Euclidean distance in the reduced-dimensional space. Diffusion maps are part of the family of nonlinear dimensionality reduction methods which focus on discovering the underlying manifold that the data has been sampled from. By integrating local similarities at different scales, diffusion maps give a global description of the data-set. Compared with other methods, the diffusion map algorithm is robust to noise perturbation and computationally inexpensive.

Given a set of N data points $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, the connectivity between data points x_i and x_j is defined by the transition probability between these two points $p(x_i, x_j)$, which is measured by their distance

$$p(x_i, x_j) \propto k(x_i, x_j). \tag{5.9.0.1}$$

The kernel $k(x, y)$ defines a local measure of similarity within a certain neighborhood. Since a given kernel will capture a specific feature of the data set, it should be well-designed to match the requirement of the application. A Gaussian kernel

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\alpha}\right) \tag{5.9.0.2}$$

is frequently used, where α tunes the size of the neighborhood. Transition probability matrix $\mathbf{P}$ must be row-normalized, which leads to

$$p(x_i, x_j) = k(x_i, x_j) / \sum_{x_j \in \mathbf{x}} k(x_i, x_j), \tag{5.9.0.3}$$

or in matrix form

$$P = D^{-1}K, \tag{5.9.0.4}$$

where D is the diagonal matrix consisting of the row-sums of K. *Note*: P is non-symmetric, which has a leading eigenvalue $\lambda_1 = 1$ with multiplicity 1.

Suppose there are three data points $\{x_1, x_2, x_3\}$ and the single-step transition probability matrix is

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 \\ p_{21} & p_{22} & p_{23} \\ 0 & p_{32} & p_{33} \end{bmatrix}, \tag{5.9.0.5}$$

where we have assumed that the transition probability from x_1 to x_3 is zero and vice versa. It can be easily found that

$$P^2 = \begin{bmatrix} p_{11}p_{11} + p_{12}p_{21} & p_{11}p_{12} + p_{12}p_{22} & p_{12}p_{23} \\ p_{21}p_{11} + p_{22}p_{21} & p_{21}p_{12} + p_{22}p_{22} + p_{23}p_{32} & p_{22}p_{23} + p_{23}p_{33} \\ p_{32}p_{21} & p_{32}p_{22} + p_{33}p_{32} & p_{32}p_{23} + p_{33}p_{33} \end{bmatrix}. \tag{5.9.0.6}$$

It shows that after two steps of propagation, the probability of transition from x_1 to x_3 is no longer zero, instead it equals to the probability of transition from x_1 to x_2, p_{12}, in the first step multiplied by the probability of transition from x_2 to x_3, p_{23}, in the second step. Similar observation can also be found for other transitions among the data points.

It can be seen from the above that P_{ij}^t sum all paths of length t from point x_i to point x_j. With increasing value of t, different scales of the data structure can be visualized. This is the diffusion process, during which the local connectivity is integrated to present the global connectivity of a data set. Besides, pathways built by long, low probability jumps are gradually replaced by short, high probability jumps.

With the global geometric structure of a data set uncovered by the diffusion process described above, a diffusion metric can be defined for this structure. The metric measures the similarity of two points in the observed space as the connectivity between them, and is defined as

$$\begin{aligned} D_t(x_i, x_j)^2 &= \sum_{u \in X} |p_t(x_i, u) - p_t(x_j, u)|^2 \\ &= \sum_k |P_{ik}^t - P_{kj}^t|^2. \end{aligned} \tag{5.9.0.7}$$

$D_t(x_i, x_j)^2$ is known as the diffusion distance between point x_i and x_j. In order to have a small diffusion distance between two points, there should be many high probability paths of length t linking these two points. Since it sums over all possible paths, this algorithm is robust to noise perturbation. Besides, the path probabilities between x_i, u and u, x_j must be roughly equal. This happens when both x_i and x_j are well connected via u. However, calculating diffusion distances is computational expensive. Therefore, it is convenient to map data points into a Euclidean space, in which the diffusion distance becomes the Euclidean distance in this new diffusion space.

A diffusion map maps coordinates between data and diffusion space by reorganizing data according to the diffusion metric. It preserves a data set's intrinsic geometry with data points mapped to a lower-dimensional structure. With the mapping

$$y_i := \begin{bmatrix} p_t(x_i, x_1) \\ p_t(x_i, x_2) \\ \vdots \\ p_t(x_i, x_N) \end{bmatrix} = P_{i*}^{\mathrm{T}}, \tag{5.9.0.8}$$

the Euclidean distance between two mapped points, y_i and y_j, is

$$\begin{aligned}\|y_i - y_j\|_E^2 =& \sum_{u\in X} |p_t(x_i,u) - p_t(x_j,u)|^2 \\ =& \sum_k |P_{ik}^t - P_{kj}^t|^2 \\ =& D_t(x_i,x_j)^2. \end{aligned} \tag{5.9.0.9}$$

However, this is not a good mapping for dimension reduction, and we must transform this mapping into a new coordinate system. As said above, P is non-symmetric. Let its left and right eigenvectors be $\{e_k\}$ and $\{v_k\}$ and the eigenvalues $\{\lambda_k\}$. The eigen decomposition

$$P = \sum_k \lambda_k v_k e_k^{\mathrm{T}} \tag{5.9.0.10}$$

indicates that each row of the diffusion matrix P can be expressed in terms of a new basis $\{e_k\}$, the left eigenvectors of P. In this new coordinate system, each row of P is represented by a point

$$y_i' = \begin{bmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{bmatrix}, \tag{5.9.0.11}$$

where $\phi_k(s)$ indicates the sth element of the kth eigenvector of P. The Euclidean distance between mapped points y_i' and y_j' is the diffusion distance. In most cases, λ_k decays very fast. Therefore, dimension reduction can be achieved by retaining the m dimensions associated with the dominant eigenvectors.

5.10 t-Distributed Stochastic Neighbor Embedding Algorithm (t-SNE)

The t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm, proposed by Laurens van der Maaten and Geoffrey Hinton in 2008,[125] was an improved version of the SNE algorithm developed by Hinton and Roweis in 2002.[126]

In SNE, the high-dimensional Euclidean distances between data points are converted by conditional probabilities that represent similarities. The similarity of data point x_j to data point x_i is the conditional probability, $p_{j|i}$, that x_i would take x_j as its neighbor, which is proportional to their probability density under a Gaussian centered at x_i

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}. \tag{5.10.0.1}$$

The magnitudes of $\{\sigma_i\}$ tune the structure of the connections. Similarly, for the low-dimensional mapped data points y_i and y_j, the conditional probability, denoted as $q_{j|i}$, is written also in a Gaussian form

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|y_i - y_j\|^2\right)} \tag{5.10.0.2}$$

with variance set to $1/\sqrt{2}$.

If the mapped data points y_i and y_j faithfully model the similarity between the data points x_i and x_j, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. A natural way to measure the faithfulness is the Kullback–Leibler divergence. SNE minimizes the sum of Kullback–Leibler divergence over all data points

$$C = \sum_i KL(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}} \tag{5.10.0.3}$$

using a gradient descent method. Because the Kullback–Leibler is not symmetric, the cost function of SNE focuses on retaining the local structure of the data point in the map.

The minimization of the cost function C is performed using a gradient descent method with the gradient

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j). \tag{5.10.0.4}$$

To escape from poor local minima, a relatively large momentum term is added to the gradient, and the update of $\mathcal{Y}$ is written as

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta\frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}). \tag{5.10.0.5}$$

Despite this optimization strategy, the method still does not ensure that the global best is obtained. Therefore, it is common to run the optimization several times on the data set with different initial condition and random numbers.

Usually, it is unlikely that a single value of σ_i will be optimal for all the data points. Instead, it is related to the distribution density of the data points, that is varying in the high-dimensional space. In SNE, the value of σ_i is determined by a fixed perplexity specified by the user. The perplexity can be interpreted as a smooth measure of the effective number of neighbors, which is defined as

$$perp(P_i) = 2^{H(P_i)}, \tag{5.10.0.6}$$

where $H(P_i)$ is the Shannon entropy of P_i measured in bits

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}. \tag{5.10.0.7}$$

Because of the asymmetry of the Kullback–Leibler divergence in the cost function, SNE often suffers from the "crowding problem". In order to solve this problem, t-SNE algorithm utilizes symmetrized version of the cost function and a heavily-tailed Student-t distribution rather than a Gaussian to compute the similarity between two points *in the low-dimensional space.*

As an alternative to minimizing the sum of the Kullback–Leibler divergences between the conditional probabilities $p_{j|i}$ and $q_{j|i}$, it is also possible to minimizes a single Kullback-Leibler divergence between a joint probability distribution, P, in the high-dimensional space and a joint probability distribution, Q, in the low-dimensional space:

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{5.10.0.8}$$

in which

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k\neq l} \exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)} \tag{5.10.0.9}$$

and

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq l} \exp\left(-\|y_k - y_l\|^2\right)} \tag{5.10.0.10}$$

with $p_{ii} = q_{ii} = 0$. However, t-SNE circumvent this problem by forcing the joint probabilities p_{ij} in the high-dimensional space to be symmetric as $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ and using Eq. 5.10.0.10 for q_{ij} in the low-dimensional space. The gradient of the cost function with respect to the mapped points becomes

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j). \tag{5.10.0.11}$$

5.11 Uniform Manifold Approximation and Projection (UMAP)

author killed by math

5.12 Spectral Gap Optimization of Order Parameters

Spectral Gap Optimization of Order Parameters (SGOOP) was proposed by Tiwary and Berne in 2016.[127] The idea of this method is that the best CV is one with the maximum separation of timescales between visible slow and hidden fast processes, and the timescale separation is calculated as the spectral gap between the slow and fast eigenvalues of the transition probability matrix. Input to this method is any available information about the static and dynamic properties of the system, accumulated through (i) a biased simulation performed along a suboptimal trial CV, and/or (ii) short unbiased simulations or experimental observations. This information is then processed utilizing the principle of maximum caliber to set up an unbiased master equation for the dynamics of various trial CV, and the best CV is optimized by maximizing the spectral gap of the associated transfer matrix.

With a set of order parameters $\{\Theta\}$, a trial CV is defined as $f(\Theta)$. This CV can be multidimensional. The CV is then discretized in grids labeled by n. For a fixed Δt, the instantaneous probability $p_n(t)$ follows the master equation

$$\frac{\Delta p_n(t)}{\Delta t} = \sum_m k_{mn} p_m(t) - \sum_m k_{nm} p_n(t) = \sum_m \mathbf{K}_{nm} p_m(t), \tag{5.12.0.1}$$

where k_{nm} is the rate of transition from grid n to m per unit time, $\mathbf{K}_{nm} = k_{mn}$ and

$$\mathbf{K}_{nn} = -\sum_{m \neq n} k_{nm} = k_{nn} - 1. \tag{5.12.0.2}$$

If the dynamics of $f(\Theta)$ is Markovian, the transition probability matrix $\mathbf{\Omega}$ is given for small Δt by the following

$$\mathbf{\Omega} = \exp(\mathbf{K}\Delta t) \approx \mathbf{I} - \mathbf{K}\Delta t. \tag{5.12.0.3}$$

Maximum caliber approach defines an entropy S as a functional of the probabilities of micropaths as

$$S = -\sum_{ab} p_a \omega_{ab} \log \omega_{ab}. \tag{5.12.0.4}$$

Path ensemble averages of time-dependent quantities A_{ab} can be calculated via

$$\langle A \rangle = \sum_{ab} p_a \omega_{ab} A_{ab}. \tag{5.12.0.5}$$

The path entropy S, constraints on the observables $\langle A^n \rangle$, and some others constraints such as detailed balance is collectively called caliber. Maximizing the caliber leads to

$$\omega_{ab} = \sqrt{\frac{p_b}{p_a}} e^{-\sum_i \rho_i A^i_{ab}}, \tag{5.12.0.6}$$

where ρ_i is the Lagrange multiplier for the associated constraint. When the only observable available is the mean number of transition $\langle N \rangle$ in observation interval Δt over the entire gridded CV, the above equation takes a particularly simple form

$$\omega_{ab} = \sqrt{\frac{p_b}{p_a}} e^{-\rho}. \tag{5.12.0.7}$$

Let $\{\lambda\}$ denote the set of eigenvalue of $\mathbf{\Omega}$ with $\lambda_0 = 1 > \lambda_1 > \lambda_2 \cdots$. The spectral gap is defined as $\lambda_s - \lambda_{s+1}$, where s is the number of barriers apparent from the free-energy estimate projected on the CV at hand, that are higher than a user-defined threshold. The optimal CV is obtained by maximizing the spectral gap.

Appendix A

Statistical Uncertainty in the Estimator for Correlated Time Series Data

"It is not the estimate or the forecast that matters so much as the degree of confidence with the opinion."

– Nassim Nicholas Taleb

Suppose we have a time series of correlated sequential observations of the randomly sampled variable X denoted as $\{x_n\}, n = 1, \ldots, N$ that come from a stationary, time-reversible stochastic process. The expectation of X can be estimated as the time average of the samples

$$\hat{X} = \frac{1}{N} \sum_{n=1}^{N} x_n. \tag{A.0.0.1}$$

Because of the existence of correlation among the samples, the variance for the expectation, which is defined as

$$\delta^2 \hat{X} \equiv \left\langle \left(\hat{X} - \left\langle \hat{X} \right\rangle \right)^2 \right\rangle = \left\langle \hat{X}^2 \right\rangle - \left\langle \hat{X} \right\rangle^2, \tag{A.0.0.2}$$

is complicated. We first take Eq. A.0.0.1 into Eq. A.0.0.2, and split the sum into one term capturing the variance in the observations and a remaining term capturing the correlation between the observations as

$$\begin{aligned} \delta^2 \hat{X} =& \frac{1}{N^2} \sum_{n,n'=1}^{N} \left[\langle x_n x_{n'} \rangle - \langle x_n \rangle \langle x_{n'} \rangle \right] \\ =& \frac{1}{N^2} \sum_{n=1}^{N} \left[\left\langle x_n^2 \right\rangle - \langle x_n \rangle^2 \right] + \frac{1}{N^2} \sum_{n \neq n'=1}^{N} \left[\langle x_n x_{n'} \rangle - \langle x_n \rangle \langle x_{n'} \rangle \right] \end{aligned} \tag{A.0.0.3}$$

Because of the stationarity, it becomes

$$\begin{aligned}\delta^2\hat{X} =& \frac{1}{N}\left[\left\langle x_n^2\right\rangle - \langle x_n\rangle^2\right] \\ &+ \frac{1}{N^2}\sum_{t=1}^{N-1}(N-t)\left[\langle x_n x_{n+t}\rangle + \langle x_{n+t}x_n\rangle - \langle x_n\rangle\langle x_{n+t}\rangle - \langle x_{n+t}\rangle\langle x_n\rangle\right]\end{aligned} \tag{A.0.0.4}$$

and because of the time-reversibility, it can be further simplified to

$$\begin{aligned}\delta^2\hat{X} =& \frac{1}{N}\left[\left\langle x_n^2\right\rangle - \langle x_n\rangle^2\right] \\ &+ \frac{2}{N}\sum_{t=1}^{N-1}\left(\frac{N-t}{N}\right)\left[\langle x_n x_{n+t}\rangle - \langle x_n\rangle\langle x_{n+t}\rangle\right] \\ \equiv& \frac{{\sigma_x}^2}{N}(1+2\tau) = \frac{{\sigma_x}^2}{N/g},\end{aligned} \tag{A.0.0.5}$$

where ${\sigma_x}^2$, statistical inefficiency g, and autocorrelation time τ (in units of the sampling interval) are given by

$${\sigma_x}^2 \equiv \left\langle x_n^2\right\rangle - \langle x_n\rangle^2 \tag{A.0.0.6}$$

$$\tau \equiv \sum_{t=1}^{N-1}\left(\frac{N-t}{N}\right)C_t \tag{A.0.0.7}$$

$$C_t = \frac{\langle x_n x_{n+t}\rangle - \langle x_n\rangle\langle x_n\rangle}{{\sigma_x}^2} \tag{A.0.0.8}$$

$$g \equiv 1 + 2\tau \tag{A.0.0.9}$$

The quantity $g \equiv 1 + 2\tau > 1$ can be regarded as a statistical inefficiency, in that N/g gives the effective number of *uncorrelated* configurations contained in the time series.

Appendix B

The Optimal Mean of Independent Measurements with Uncertainties

Suppose we have N measurements of a quantity x, which are denoted as $\{x_i\}$, with i = 1, ..., N. Each measurement has a variance $\delta^2 x_i$. To find the optimal mean of this data set, we first write the mean of $\{x_i\}$ as a weighted average of them

$$\bar{x} = \sum_{i=1}^{N} a_i x_i, \tag{B.0.0.1}$$

in which a_i are the normalized weights, i.e.

$$\sum_{i=1}^{N} a_i = 1. \tag{B.0.0.2}$$

According to the error propagation rule, if the measurements are independent, the variance of the mean $\bar{x}$ can be written as

$$\delta^2 \bar{x} = \sum_{i=1}^{N} {a_i}^2 \delta^2 x_i. \tag{B.0.0.3}$$

Minimizing $\delta^2 \bar{x}$ with respect to a_i under the constraint of Eq. B.0.0.2 using the Lagrange multiplier λ, we find

$$\begin{aligned} \frac{\partial L}{\partial a_j} =& \frac{\partial}{\partial a_j} \left[\sum_{i=1}^{N} {a_i}^2 \delta^2 x_i + \lambda \left(1 - \sum_{i=1}^{N} a_i \right) \right] \\ =& 2 a_j \delta^2 x_j - \lambda \\ =& 0 \end{aligned} \tag{B.0.0.4}$$

for all x_j. It can be easily identified that a_j is inversely proportional to $\delta^2 x_j$, i.e.

$$a_j = \frac{\left(\delta^2 x_j\right)^{-1}}{\sum\limits_{i=1}^{N} \left(\delta^2 x_i\right)^{-1}}, \tag{B.0.0.5}$$

and

$$\bar{x} = \sum_{j=1}^{N} \frac{\left(\delta^2 x_j\right)^{-1}}{\sum\limits_{i=1}^{N} \left(\delta^2 x_i\right)^{-1}} x_j, \tag{B.0.0.6}$$

with

$$\delta^2 \bar{x} = \sum_{j=1}^{N} \frac{\left(\delta^2 x_j\right)^{-1}}{\left(\sum\limits_{i=1}^{N} \left(\delta^2 x_i\right)^{-1}\right)^2} = \frac{1}{\sum\limits_{i=1}^{N} \left(\delta^2 x_i\right)^{-1}}. \tag{B.0.0.7}$$

Appendix C

The Relationship between the ΔU Distributions in Forward and Backward TP

In a forward TP calculation between H_0 and H_1, the distribution of ΔU is

$$f(\Delta U) = \frac{\int e^{-\beta H_0} \delta(H_1 - H_0 - \Delta U)\, \mathrm{d}x}{\int e^{-\beta H_0}\, \mathrm{d}x}. \qquad \text{(C.0.0.1)}$$

While in a backward TP, the distribution is

$$\begin{aligned}
g(\Delta U) =& \frac{\int e^{-\beta H_1}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x}{\int e^{-\beta H_1}\,\mathrm{d}x} \\
=& \frac{f(\Delta U)\int e^{-\beta H_1}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x \int e^{-\beta H_0}\,\mathrm{d}x}{\int e^{-\beta H_1}\,\mathrm{d}x \int e^{-\beta H_0}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x} \\
=& f(\Delta U)\frac{\int e^{-\beta H_0}\,\mathrm{d}x}{\int e^{-\beta H_1}\,\mathrm{d}x}\frac{\int e^{-\beta H_1}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x}{\int e^{-\beta H_0}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x} \\
=& f(\Delta U)e^{\beta\Delta A}\frac{\int e^{-\beta\Delta H}e^{-\beta H_0}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x}{\int e^{-\beta H_0}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x} \\
=& f(\Delta U)e^{\beta\Delta A}\frac{e^{-\beta\Delta U}\int e^{-\beta H_0}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x}{\int e^{-\beta H_0}\delta(H_1 - H_0 - \Delta U)\,\mathrm{d}x} \\
=& f(\Delta U)e^{\beta\Delta A}e^{-\beta\Delta U} \qquad \text{(C.0.0.2)}
\end{aligned}$$

Or it can be written as

$$g(\Delta U)e^{\beta\Delta U} = f(\Delta U)e^{\beta\Delta A}, \qquad \text{(C.0.0.3)}$$

which was first shown by Shing and Gubbins for Widom insertion and deletion[128].

By looking at this equation and the integral in energy space for TP (Eq. 3.1.1.15),

$$\Delta A = -\frac{1}{\beta}\ln\int \exp(-\beta\Delta U)f(\Delta U)\,\mathrm{d}\Delta U, \qquad \text{(C.0.0.4)}$$

it can be easily realized that the integrand is proportional to the probability of energy difference sampled from state 1 ($g(\Delta U)$) and the free energy difference ΔA can be estimated reliably only when the sampling at state 0 covers the representative configurations of state 1.

Appendix D

Cumulant Expansion for the Free Energy Difference in Thermodynamic Perturbation Calculations

In Thermodynamic Perturbation, the free energy difference between states are expressed as

$$\Delta A = -\beta^{-1} \ln \langle \exp\left[-\beta \Delta U\right] \rangle_0. \tag{D.0.0.1}$$

Due to the exponential term on the right-hand-side of this equation, the convergence is usually slow. This expression can be expanded into Taylor series, and the leading terms can converge much faster than the complete sum.

Using the Taylor expansion for $e^x = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} x^n$ first, we have

$$\exp\left[-\beta \Delta U\right] = 1 + (-\beta \Delta U) + \frac{1}{2!}(-\beta \delta U)^2 + \frac{1}{3!}(-\beta \Delta U)^3 + \cdots . \tag{D.0.0.2}$$

Then Eq. D.0.0.1 becomes

$$\Delta A = -\beta^{-1} \ln \left(1 - \beta \left\langle \Delta U \right\rangle_0 + \frac{1}{2}\beta^2 \left\langle \Delta U^2 \right\rangle_0 - \frac{1}{6}\beta^3 \left\langle \Delta U^3 \right\rangle_0 + \cdots \right). \tag{D.0.0.3}$$

Using the Taylor expansion for $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{2}{3!}x^3 - \frac{6}{4!}x^4 + \cdots$, we

have

$$\begin{aligned}
\Delta A = & \langle \Delta U \rangle_0 - \frac{1}{2}\beta \left\langle \Delta U^2 \right\rangle_0 + \frac{1}{6}\beta^2 \left\langle \Delta U^3 \right\rangle_0 + \cdots \\
& + \frac{1}{2}\left(\beta \langle \Delta U \rangle_0^2 - \beta^2 \langle \Delta U \rangle_0 \left\langle \Delta U^2 \right\rangle_0 + \frac{1}{4}\beta^3 \left\langle \Delta U^2 \right\rangle_0^2 + \cdots \right) \\
& - \frac{1}{3}\left(-\beta^2 \langle \Delta U \rangle_0^3 + \cdots \right) \\
& + \cdots \\
= & \langle \Delta U \rangle_0 - \frac{\beta}{2}\left(\left\langle \Delta U^2 \right\rangle_0 - \langle \Delta U \rangle_0^2 \right) \\
& + \frac{\beta^2}{6}\left(\left\langle \Delta U^3 \right\rangle_0 - 3 \left\langle \Delta U^2 \right\rangle_0 \langle \Delta U \rangle_0 + 2 \langle \Delta U \rangle_0^3 \right) + \cdots \qquad \text{(D.0.0.4)}
\end{aligned}$$

Appendix E

MBAR Returns to BAR When Only Two States Are Considered

When there are only two states, the free energy in Eq. 3.1.5.7 for the 1st state in MBAR becomes

$$\begin{aligned} f_1 =& -\beta_1{}^{-1} \ln \sum_{n=1}^{N} \frac{\exp\left(-\beta_1 U_1(\mathbf{R}_n)\right)}{\sum_{k=1}^{2} N_k \exp\left(\beta_k f_k - \beta_k U_k(\mathbf{R}_n)\right)} \\ =& -\beta_1{}^{-1} \ln \sum_{j=1}^{2} \sum_{n=1}^{N_j} \frac{\exp\left(-\beta_1 U_1(\mathbf{R}_{jn})\right)}{\sum_{k=1}^{2} N_k \exp\left(\beta_k f_k - \beta_k U_k(\mathbf{R}_{jn})\right)}, \end{aligned} \tag{E.0.0.1}$$

or equivalently we have

$$1 = \sum_{n=1}^{N} \frac{\exp\left(\beta_1 f_1 - \beta_1 U_1(\mathbf{R}_n)\right)}{N_1 \exp\left(\beta_1 f_1 - \beta_1 U_1(\mathbf{R}_n)\right) + N_2 \exp\left(\beta_2 f_2 - \beta_2 U_2(\mathbf{R}_n)\right)}, \tag{E.0.0.2}$$

$$\begin{aligned} N_1 =& \sum_{n=1}^{N_1} \frac{1}{1 + \frac{N_2}{N_1} \exp\left(\Delta f - \Delta U(\mathbf{R}_{1n})\right)} \\ &+ \sum_{n=1}^{N_2} \frac{1}{1 + \frac{N_2}{N_1} \exp\left(\Delta f - \Delta U(\mathbf{R}_{2n})\right)} \end{aligned} \tag{E.0.0.3}$$

where $\Delta f = \beta_2 f_2 - \beta_1 f_1$ and $\Delta U = \beta_2 U_2 - \beta_1 U_1$. We further define $M = -\ln \frac{N_2}{N_1}$, then

$$\begin{aligned} N_1 &= \sum_{n=1}^{N_1} \frac{1}{1+\exp\left(\Delta f - \Delta U(\mathbf{R}_{1n}) - M\right)} \\ &\quad + \sum_{n=1}^{N_2} \frac{1}{1+\exp\left(\Delta f - \Delta U(\mathbf{R}_{2n}) - M\right)} \\ 0 &= \sum_{n=1}^{N_1} \left[\frac{1}{1+\exp\left(\Delta f - \Delta U(\mathbf{R}_{1n}) - M\right)} - 1\right] \\ &\quad + \sum_{n=1}^{N_2} \frac{1}{1+\exp\left(\Delta f - \Delta U(\mathbf{R}_{2n}) - M\right)} \end{aligned}$$

$$\sum_{n=1}^{N_1} \frac{1}{1+\exp\left(-\Delta f + \Delta U(\mathbf{R}_{1n}) + M\right)} = \sum_{n=1}^{N_2} \frac{1}{1+\exp\left(\Delta f - \Delta U(\mathbf{R}_{2n}) - M\right)}$$

$$\sum_{n=1}^{N_1} f\left(-\Delta f + \Delta U(\mathbf{R}_{1n}) + M\right) = \sum_{n=1}^{N_2} f\left(\Delta f - \Delta U(\mathbf{R}_{2n}) - M\right)$$

$$N_1 \left\langle f\left(-\Delta f + \Delta U(\mathbf{R}_{1n}) + M\right)\right\rangle_1 = N_2 \left\langle f\left(\Delta f - \Delta U(\mathbf{R}_{2n}) - M\right)\right\rangle_2$$

$$\frac{\left\langle f\left(\Delta f - \Delta U(\mathbf{R}_{2n}) - M\right)\right\rangle_2}{\left\langle f\left(-\Delta f + \Delta U(\mathbf{R}_{1n}) + M\right)\right\rangle_1} = \frac{N_1}{N_2},$$

which is Eq. 3.1.3.12.

Appendix F

MBAR is a binless form of WHAM

Maybe you have already noticed that MBAR and WHAM have very similar forms for the free energy. So you may want to ask if there is any connection between MBAR and WHAM. The answer is YES. MBAR is a binless form of WHAM.[90] Let us follow Zhang et al[129] and rewrite Eq. 3.1.4.21 into an integral form

$$f_i = -\ln \int \Omega \exp\left(-\beta_i U\right) \mathrm{d}U. \tag{F.0.0.1}$$

Taking Eq. 3.1.4.20 into Eq. F.0.0.1, we find

$$f_i = -\ln \int \frac{\sum_{k=1}^{K} H_k(U) \exp\left(-\beta_i U\right)}{\sum_{k=1}^{K} N_k \exp\left(f_k - \beta_k U\right)} \mathrm{d}U, \tag{F.0.0.2}$$

where ${g_{mk}}^{-1}$ has been omitted and H_{mk} has been changed to continuous form $H_k(U)$. From the definition,

$$H_k(U) = \sum_{\mathbf{R}}^{(k)} \delta(U(\mathbf{R}) - U). \tag{F.0.0.3}$$

Taking Eq. F.0.0.3 into Eq. F.0.0.2, we have

$$\begin{aligned} f_i &= -\ln \sum_{k=1}^{K} \sum_{\mathbf{R}}^{(k)} \frac{\exp\left(-\beta_i U(\mathbf{R})\right)}{\sum_{k=1}^{K} N_k \exp\left(f_k - \beta_k U(\mathbf{R})\right)} \\ &= -\ln \sum_{n=1}^{N} \frac{\exp\left(-\beta_i U_i(\mathbf{R_n})\right)}{\sum_{k=1}^{K} N_k \exp\left(f_k - \beta_k U_k(\mathbf{R_n})\right)}, \end{aligned} \tag{F.0.0.4}$$

which is Eq. 3.1.5.7.

Appendix G

Jensen's inequality

$$e^{\langle X \rangle} \leq \left\langle e^{X} \right\rangle \tag{G.0.0.1}$$

6

Bias-variance decomposition

Given a set of samples, the mean squared error (M.S.E.) of an estimator $\hat{\theta}$ for some parameter θ is

$$\text{M.S.E.} = \mathbb{E}\Big[\Big(\hat{\theta} - \theta\Big)^2\Big]. \tag{6.0.0.1}$$

It can be decomposed into a bias term and a variance term as following

$$\begin{aligned}
\text{M.S.E.} =& \mathbb{E}\Big[\Big(\hat{\theta} - \theta\Big)^2\Big] \\
=& \mathbb{E}\Big[\Big(\hat{\theta} - \mathbb{E}\Big(\hat{\theta}\Big) + \mathbb{E}\Big(\hat{\theta}\Big) - \theta\Big)^2\Big] \\
=& \mathbb{E}\Big[\Big(\hat{\theta} - \mathbb{E}\Big(\hat{\theta}\Big)\Big)^2 + 2\Big(\hat{\theta} - \mathbb{E}\Big(\hat{\theta}\Big)\Big)\Big(\mathbb{E}\Big(\hat{\theta}\Big) - \theta\Big) + \Big(\mathbb{E}\Big(\hat{\theta}\Big) - \theta\Big)^2\Big] \\
=& \mathbb{E}\Big[\Big(\hat{\theta} - \mathbb{E}\Big(\hat{\theta}\Big)\Big)^2\Big] + \mathbb{E}\Big[\Big(\mathbb{E}\Big(\hat{\theta}\Big) - \theta\Big)^2\Big] \\
=& \text{Var} + \text{Bias}^2,
\end{aligned} \tag{6.0.0.2}$$

where we have utilized

$$\mathbb{E}\Big[\Big(\hat{\theta} - \mathbb{E}\Big(\hat{\theta}\Big)\Big)\Big(\mathbb{E}\Big(\hat{\theta}\Big) - \theta\Big)\Big] = \Big(\mathbb{E}\Big(\hat{\theta}\Big) - \mathbb{E}\Big(\hat{\theta}\Big)\Big)\Big(\mathbb{E}\Big(\hat{\theta}\Big) - \theta\Big) = 0, \tag{6.0.0.3}$$

since $\mathbb{E}(\hat{\theta})$ and θ are constants.

Bibliography

[1] Niels Hansen and Wilfred F. van Gunsteren. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.*, 10(7):2632–2647, 2014.

[2] Yuqing Deng and Benoît Roux. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.*, 2(5):1255–1273, 2006.

[3] Hyung-June Woo and Benoît Roux. Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 102(19):6825–6830, 2005.

[4] Yuqing Deng and Benoît Roux. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B*, 113(8):2234–2246, 2009.

[5] James C. Gumbart, Benoît Roux, and Christophe Chipot. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.*, 9(1):794–802, 2013.

[6] Alan Grossfield, Paul N. Patrone, Daniel R. Roe, Andrew J. Schultz, Daniel Siderius, and Daniel M. Zuckerman. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living J. Computat. Mol. Sci.*, 1:5067, 2018.

[7] Daniel M. Zuckerman. Equilibrium Sampling in Biomolecular Simulations. *Annu. Rev. Biophys.*, 40(1):41–62, 2011.

[8] Rafael C. Bernardi, Marcelo C.R. Melo, and Klaus Schulten. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta*, 1850(5):872–877, 2015.

[9] Anna S. Kamenik, Stephanie M. Linker, and Sereina Riniker. Enhanced Sampling without Borders: On Global Biasing Functions and How to Reweight Them. *Phys. Chem. Chem. Phys.*, 24:1225–1236, 2022.

[10] Haochuan Chen and Christophe Chipot. Enhancing Sampling with Free-energy Calculations. *Curr. Opin. Struct. Biol.*, 77:102497, 2022.

[11] Ulrich H. E. Hansmann and Yuko Okamoto. Prediction of Peptide Conformation by Multicanonical Algorithm: New Approach to the Multiple-Minima Problem. *J. Comput. Chem.*, 14(11):1333–1338, 1993.

[12] Ulrich H.E. Hansmann. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.*, 281(1):140–150, 1997.

[13] Yuji Sugita and Yuko Okamoto. Replica-exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.*, 314(1):141–151, 1999.

[14] John D. Chodera and Michael R. Shirts. Replica Exchange and Expanded Ensemble Simulations as Gibbs Sampling: Simple Improvements for Enhanced Mixing. *J. Chem. Phys.*, 135(19):194110, 2011.

[15] Soonmin Jang, Seokmin Shin, and Youngshang Pak. Replica-exchange Method Using the Generalized Effective Potential. *Phys. Rev. Lett.*, 91:058305, 2003.

[16] Yilin Meng and Adrian E. Roitberg. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model. *J. Chem. Theory Comput.*, 6(4):1401–1412, 2010.

[17] Juyong Lee, Benjamin T. Miller, Ana Damjanović, and Bernard R. Brooks. Constant pH Molecular Dynamics in Explicit Solvent with Enveloping Distribution Sampling and Hamiltonian Exchange. *J. Chem. Theory Comput.*, 10(7):2738–2750, 2014.

[18] E. Marinari and G. Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhys. Lett.*, 19(6):451–458, 1992.

[19] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New Approach to Monte Carlo Calculation of the Free Energy: Method of Expanded Ensembles. *J. Chem. Phys.*, 96(3):1776–1783, 1992.

[20] Charles J. Geyer and Elizabeth A. Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *J. Am. Stat. Assoc.*, 90(431):909–920, 1995.

[21] 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.

[22] Christian Leitold, Christopher J. Mundy, Marcel D. Baer, Gregory K. Schenter, and Baron Peters. Solvent Reaction Coordinate for an S_N2 Reaction. *J. Chem. Phys.*, 153(2):024103, 2020.

[23] Donald Hamelberg, John Mongan, and J. Andrew McCammon. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.*, 120(24):11919–11929, 2004.

[24] Arthur F. Voter. Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Phys. Rev. Lett.*, 78:3908–3911, 1997.

[25] Yinglong Miao, William Sinko, Levi Pierce, Denis Bucher, Ross C. Walker, and J. Andrew McCammon. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J. Chem. Theory Comput.*, 10(7):2677–2689, 2014.

[26] Yinglong Miao, Victoria A. Feher, and J. Andrew McCammon. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.*, 11(8):3584–3595, 2015.

[27] Giovanni Ciccotti, Raymond Kapral, and Eric Vanden-Eijnden. Blue Moon Sampling, Vectorial Reaction Coordinates, and Unbiased Constrained Dynamics. *ChemPhysChem*, 6(9):1809–1814, 2005.

[28] E. A. Carter, Giovanni Ciccotti, James T. Hynes, and Raymond Kapral. Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events. *Chem. Phys. Lett.*, 156(5):472–477, 1989.

[29] Eric Darve and Andrew Pohorille. Calculating Free Energies Using Average Force. *J. Chem. Phys.*, 115(20):9169–9183, 2001.

[30] Eric Darve, David Rodríguez-Gómez, and Andrew Pohorille. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.*, 128(14):144120, 2008.

[31] Xianjun Kong and Charles L. Brooks III. λ-dynamics: A New Approach to Free Energy Calculations. *J. Chem. Phys.*, 105(6):2414–2423, 1996.

[32] Jerry B. Abrams, Lula Rosso, and Mark E. Tuckerman. Efficient and Precise Solvation Free Energies via Alchemical Adiabatic Molecular Dynamics. *J. Chem. Phys.*, 125(7):074115, 2006.

[33] Pan Wu, Xiangqian Hu, and Weitao Yang. λ-Metadynamics Approach To Compute Absolute Solvation Free Energy. *J. Phys. Chem. Lett.*, 2(17):2099–2103, 2011.

[34] Luca Maragliano and Eric Vanden-Eijnden. A Temperature Accelerated Method for Sampling Free Energy and Determining Reaction Pathways in Rare Events Simulations. *Chem. Phys. Lett.*, 426(1):168–175, 2006.

[35] Haohao Fu, Xueguang Shao, Christophe Chipot, and Wensheng Cai. Extended Adaptive Biasing Force Algorithm. An On-the-Fly Implementation for Accurate Free-Energy Calculations. *J. Chem. Theory Comput.*, 12(8):3506–3513, 2016.

[36] Lianqing Zheng and Wei Yang. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *J. Chem. Theory Comput.*, 8(3):810–823, 2012.

[37] Adrien Lesage, Tony Lelièvre, Gabriel Stoltz, and Jérôme Hénin. Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method. *J. Phys. Chem. B*, 121(15):3676–3685, 2017.

[38] Fugao Wang and D. P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.*, 86:2050–2053, 2001.

[39] R. E. Belardinelli and V. D. Pereyra. Fast Algorithm to Calculate Density of States. *Phys. Rev. E*, 75:046701, 2007.

[40] Yves F. Atchadé and Jun S. Liu. The Wang-Landau Algorithm in General State Spaces: Applications and Convergence Analysis. *Stat. Sinica*, 20(1):209–233, 2010.

[41] Faming Liang, Chuanhai Liu, and Raymond J. Carroll. Stochastic Approximation in Monte Carlo Computation. *J. Am. Stat. Assoc.*, 102(477):305–320, 2007.

[42] Jack Lidmar. Improving the Efficiency of Extended Ensemble Simulations: The Accelerated Weight Histogram Method. *Phys. Rev. E*, 85:056708, 2012.

[43] V. Lindahl, J. Lidmar, and B. Hess. Accelerated Weight Histogram Method for Exploring Free Energy Landscapes. *J. Chem. Phys.*, 141(4):044110, 2014.

[44] M. Lundborg, J. Lidmar, and B. Hess. The Accelerated Weight Histogram Method for Alchemical Free Energy Calculations. *J. Chem. Phys.*, 154(20):204103, 2021.

[45] Alessandro Laio and Michele Parrinello. Escaping Free-energy Minima. *Proc. Natl. Acad. Sci. U. S. A.*, 99(20):12562–12566, 2002.

[46] Giovanni Bussi, Alessandro Laio, and Michele Parrinello. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Phys. Rev. Lett.*, 96:090601, 2006.

[47] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.*, 100:020603, 2008.

[48] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.*, 67(1):159–184, 2016.

[49] James F. Dama, Michele Parrinello, and Gregory A. Voth. Well-Tempered Metadynamics Converges Asymptotically. *Phys. Rev. Lett.*, 112:240602, 2014.

[50] Pratyush Tiwary, James F. Dama, and Michele Parrinello. A Perturbative Solution to Metadynamics Ordinary Differential Equation. *J. Chem. Phys.*, 143(23):234112, 2015.

[51] Pratyush Tiwary and Michele Parrinello. A Time-Independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B*, 119(3):736–742, 2015.

[52] F. Giberti, B. Cheng, G. A. Tribello, and M. Ceriotti. Iterative Unbiasing of Quasi-Equilibrium Sampling. *J. Chem. Theory Comput.*, 16(1):100–107, 2020.

[53] Omar Valsson and Michele Parrinello. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.*, 113:090601, 2014.

[54] Michele Invernizzi and Michele Parrinello. Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *J. Phys. Chem. Lett.*, 11(7):2731–2736, 2020.

[55] Lianqing Zheng, Mengen Chen, and Wei Yang. Random Walk in Orthogonal Space to Achieve Efficient Free-energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U. S. A.*, 105(51):20227–20232, 2008.

[56] David A. Pearlman and Peter A. Kollman. The Lag between the Hamiltonian and the System Configuration in Free Energy Perturbation Calculations. *J. Chem. Phys.*, 91(12):7831–7839, 1989.

[57] Clara D. Christ and Wilfred F. van Gunsteren. Enveloping Distribution Sampling: A Method to Calculate Free Energy Differences from a Single Simulation. *J. Chem. Phys.*, 126(18):184110, 2007.

[58] Clara D. Christ and Wilfred F. van Gunsteren. Simple, Efficient, and Reliable Computation of Multiple Free Energy Differences from a Single Simulation: A Reference Hamiltonian Parameter Update Scheme for Enveloping Distribution Sampling (EDS). *J. Chem. Theory Comput.*, 5(2):276–286, 2009.

[59] Kyu-Kwang Han. A New Monte Carlo Method for Estimating Free Energy and Chemical Potential. *Phys. Lett. A*, 165(1):28–32, 1992.

[60] Clara D. Christ and Wilfred F. van Gunsteren. Multiple Free Energies from a Single Simulation: Extending Enveloping Distribution Sampling to Nonoverlapping Phase-space Distributions. *J. Chem. Phys.*, 128(17):174112, 2008.

[61] Gerhard König, Nina Glaser, Benjamin Schroeder, Alžbeta Kubincová, Philippe H. Hünenberger, and Sereina Riniker. An Alternative to Conventional λ-Intermediate States in Alchemical Free Energy Calculations: λ-Enveloping Distribution Sampling. *J. Chem. Inf. Model.*, 60(11):5407–5423, 2020.

[62] Arnaud Blondel. Ensemble Variance in Free Energy Calculations by Thermodynamic Integration: Theory, Optimal "Alchemical" Path, and Practical Solutions. *J. Comput. Chem.*, 25(7):985–993, 2004.

[63] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation-shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. *J. Chem. Phys.*, 100(12):9025–9031, 1994.

[64] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String Method for the Study of Rare Events. *Phys. Rev. B*, 66:052301, 2002.

[65] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Simplified and Improved String Method for Computing the Minimum Energy Paths in Barrier-Crossing Events. *J. Chem. Phys.*, 126(16):164103, 2007.

[66] Eric Vanden-Eijnden and Maddalena Venturoli. Revisiting the Finite Temperature String Method for the Calculation of Reaction Tubes and Free Energies. *J. Chem. Phys.*, 130(19):194103, 2009.

[67] Zhiqiang Tan. Optimally Adjusted Mixture Sampling and Locally Weighted Histogram Analysis. *J. Comput. Graph. Stat.*, 26(1):54–65, 2017.

[68] Michael R. Shirts and Vijay S. Pande. Comparison of Efficiency and Bias of Free Energies Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. *J. Chem. Phys.*, 122(14):144107, 2005.

[69] F. Marty Ytreberg, Robert H. Swendsen, and Daniel M. Zuckerman. Comparison of Free Energy Methods for Molecular Systems. *J. Chem. Phys.*, 125(18):184114, 2006.

[70] Himanshu Paliwal and Michael R. Shirts. A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. *J. Chem. Theory Comput.*, 7(12):4115–4134, 2011.

[71] Robert W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22(8):1420, 1954.

[72] William L. Jorgensen and Laura L. Thomas. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.*, 4(6):869–876, 2008.

[73] Christopher Jarzynski. Rare Events and the Convergence of Exponentially Averaged Work Values. *Phys. Rev. E*, 73:046105, 2006.

[74] Jeff Gore, Felix Ritort, and Carlos Bustamante. Bias and Error in Estimates of Equilibrium Free-energy Differences from Nonequilibrium Measurements. *Proc. Natl. Acad. Sci. U. S. A.*, 100(22):12564–12569, 2003.

[75] Christopher Jarzynski. Targeted Free Energy Perturbation. *Phys. Rev. E*, 65:046122, 2002.

[76] Zhongwei Zhu, Mark E. Tuckerman, Shane O. Samuelson, and Glenn J. Martyna. Using Novel Variable Transformations to Enhance Conformational Sampling in Molecular Dynamics. *Phys. Rev. Lett.*, 88:100201, 2002.

[77] A. M. Hahn and H. Then. Using Bijective Maps to Improve Free-energy Estimates. *Phys. Rev. E*, 79:011113, 2009.

[78] Xinqiang Ding and Bin Zhang. DeepBAR: A Fast and Exact Method for Binding Free Energy Computation. *J. Phys. Chem. Lett.*, 12:2509–2515, 2021.

[79] John G. Kirkwood. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.*, 3(5):300–313, 1935.

[80] S. Decherchi and A. Cavalli. Optimal Transport for Free Energy Estimation. *J. Phys. Chem. Lett.*, 14(6):1618–1625, 2023.

[81] Charles H. Bennett. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.*, 22(2):245–268, 1976.

[82] Gavin E. Crooks. Path-ensemble Averages in Systems Driven Far From Equilibrium. *Phys. Rev. E*, 61:2361–2366, 2000.

[83] Michael R. Shirts, Eric Bair, Giles Hooker, and Vijay S. Pande. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Phys. Rev. Lett.*, 91:140601, 2003.

[84] Herman J. C. Berendsen. *A Student's Guide to Data and Error Analysis*. Cambridge University Press, Cambridge, 2011.

[85] Alan M. Ferrenberg and Robert H. Swendsen. Optimized Monte Carlo Data Analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[86] John D. Chodera, William C. Swope, Jed W. Pitera, Chaok Seok, and Ken A. Dill. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.*, 3(1):26–41, 2007.

[87] Marc Souaille and Benoît Roux. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.*, 135:40–57, 2001.

[88] Emilio Gallicchio, Michael Andrec, Anthony K. Felts, and Ronald M. Levy. Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths. *J. Phys. Chem. B*, 109(14):6722–6731, 2005.

[89] Andrew L. Ferguson. BayesWHAM: A Bayesian Approach for Free Energy Estimation, Reweighting, and Uncertainty Quantification in the Weighted Histogram Analysis Method. *J. Comput. Chem.*, 38(18):1583–1605, 2017.

[90] Zhiqiang Tan, Emilio Gallicchio, Mauro Lapelosa, and Ronald M. Levy. Theory of Binless Multi-state Free Energy Estimation with Applications to Protein–Ligand Binding. *J. Chem. Phys.*, 136(14):144102, 2012.

[91] Michael Habeck. Bayesian Reconstruction of the Density of States. *Phys. Rev. Lett.*, 98:200601, 2007.

[92] Michael Habeck. Bayesian Estimation of Free Energies From Equilibrium Simulations. *Phys. Rev. Lett.*, 109:100601, 2012.

[93] Michael R. Shirts and John D. Chodera. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.*, 129(12):124105, 2008.

[94] Michael R. Shirts. Reweighting from the Mixture Distribution as a Better Way to Describe the Multistate Bennett Acceptance Ratio. arXiv.org, 2017. https://arxiv.org/abs/1704.00891.

[95] Charles J. Geyer. Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo. Tech. Rep. (University of Minnesota), 1994.

[96] Xinqiang Ding, Jonah Z. Vilseck, and Charles L. Brooks. Fast Solver for Large Scale Multistate Bennett Acceptance Ratio Equations. *J. Chem. Theory Comput.*, 15(2):799–802, 2019.

[97] Johannes Kästner and Walter Thiel. Bridging the Gap between Thermodynamic Integration and Umbrella Sampling Provides a Novel Analysis Method: "Umbrella Integration". *J. Chem. Phys.*, 123(14):144104, 2005.

[98] Johannes Kästner and Walter Thiel. Analysis of the Statistical Error in Umbrella Sampling Simulations by Umbrella Integration. *J. Chem. Phys.*, 124(23):234106, 2006.

[99] Johannes Kästner. Umbrella Integration with Higher-Order Correction Terms. *J. Chem. Phys.*, 136(23):234102, 2012.

[100] Christopher Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.*, 78:2690, 1997.

[101] Gavin E. Crooks. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Statis. Phys.*, 90:1481, 1998.

[102] C. Jarzynski. Equilibrium Free-energy Differences from Nonequilibrium Measurements: A Master-equation Approach. *Phys. Rev. E*, 56:5018–5035, Nov 1997.

[103] Gerhard Hummer and Attila Szabo. Free Energy Reconstruction from Nonequilibrium Single-molecule Pulling Experiments. *Proc. Natl. Acad. Sci. U. S. A.*, 98(7):3658–3661, 2001.

[104] Hao Wu, Antonia S. J. S. Mey, Edina Rosta, and Frank Noé. Statistically Optimal Analysis of State-discretized Trajectory Data from Multiple Thermodynamic States. *J. Chem. Phys.*, 141(21):214106, 2014.

[105] Jessica M.J. Swanson, Richard H. Henchman, and J. Andrew McCammon. Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophys. J.*, 86(1):67–74, 2004.

[106] Pavel V. Klimovich, Michael R. Shirts, and David L. Mobley. Guidelines for the Analysis of Free Energy Calculations. *J. Comput. Aided Mol. Des.*, 29(5):397–411, 2015.

[107] Nandou Lu and David A. Kofke. Accuracy of Free-Energy Perturbation Calculations in Molecular Simulation. I. Modeling. *J. Chem. Phys.*, 114(17):7303–7311, 2001.

[108] Di Wu and David A. Kofke. Model for Small-sample Bias of Free-Energy Calculations Applied to Gaussian-Distributed Nonequilibrium Work Measurements. *J. Chem. Phys.*, 121(18):8742–8747, 2004.

[109] Brian K. Radak. Finite-Sample Bias in Free Energy Bridge Estimators. *J. Chem. Phys.*, 151(3):034105, 2019.

[110] Di Wu and David A. Kofke. Phase-Space Overlap Measures. I. Fail-Safe Bias Detection in Free Energies Calculated by Molecular Simulation. *J. Chem. Phys.*, 123(5):054103, 2005.

[111] Baron Peters. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. Phys. Chem.*, 67(1):669–690, 2016.

[112] Alan Julian Izenman. Introduction to Manifold Learning. *Wiley Interdiscip. Rev. Comput. Stat.*, 4(5):439–446, 2012.

[113] Harold Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.*, 24:498–520, 1933.

[114] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina. Principal Component Analysis. *Nat. Rev. Methods Primers*, 2:100, 2022.

[115] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.*, 10(5):1299–1319, 1998.

[116] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.*, 7(2):179–188, 1936.

[117] Michael W. Mahoney and Petros Drineas. CUR Matrix Decompositions for Improved Data Analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 106(3):697–702, 2009.

[118] A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Netw.*, 13(4):411–430, 2000.

[119] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[120] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[121] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[122] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U. S. A.*, 102(21):7426–7431, 2005.

[123] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Multiscale Methods. *Proc. Natl. Acad. Sci. U. S. A.*, 102(21):7432–7437, 2005.

[124] Ronald R. Coifman and Stéphane Lafon. Diffusion Maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[125] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008.

[126] Geoffrey E Hinton and Sam Roweis. Stochastic Neighbor Embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

[127] Pratyush Tiwary and B. J. Berne. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. U. S. A.*, 113(11):2839–2844, 2016.

[128] K. S. Shing and K. E. Gubbins. The Chemical Potential in Dense Fluids and Fluid Mixtures via Computer Simulation. *Mol. Phys.*, 46(5):1109–1128, 1982.

[129] Cheng Zhang, Chun-Liang Lai, and B. Montgomery Pettitt. Accelerating the Weighted Histogram Analysis Method by Direct Inversion in the Iterative Subspace. *Mol. Simulat.*, 42(13):1079–1089, 2016.