THESIS

STABILITY IN THE WEIGHTED ENSEMBLE METHOD

Submitted by

Carter Lyons

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2022

Master's Committee:

Advisor: David Aristoff

Margaret Cheney
Diego Krapf

ABSTRACT

STABILITY IN THE WEIGHTED ENSEMBLE METHOD

In molecular dynamics, a quantity of interest is the mean first passage time, or average transition time, for a molecule to transition from a region A to a different region B. Often, significant potential barriers exist between A and B making the transition from A to B a rare event, which is an event that is highly improbable to occur. Correspondingly, the mean first passage time for a molecule to transition from A to B will be immense. So, using direct Markov chain Monte Carlo techniques to effectively estimate the mean first passage time is computationally infeasible due to the protracted simulations required. Instead, the Markov chain modeling the underlying molecular dynamics is simulated to steady-state and the steady-state flux from A into B is estimated. Then through the Hill relation, the mean first passage time is obtained as the reciprocal of the estimated steady-state flux. Estimating the steady-state flux into B is still a rare event but the difficulty has shifted from lengthy simulation times to a substantial variance on the desired estimate. Therefore, an importance sampling or importance splitting technique that emphasizes reaching B and reduces estimator variance must be used.

Weighted ensemble is one importance sampling Markov chain Monte Carlo method often used to estimate mean first passage times in molecular dynamics. Broadly, weighted ensemble simulates a collection of Markov chain trajectories that are assigned a weight. Periodically, certain trajectories are copied while others are removed, to encourage a transition from A to B, and the trajectory weights are adjusted accordingly. By time-averaging the weighted average of these Markov chain trajectories, weighted ensemble estimates averages with respect to the Markov chain steady-state distribution. We focus on the use of weighted ensemble for estimating the mean first passage time from A to B, through estimating the steady-state flux from A into B, of a Markov chain where upon reaching B is restarted in A according to an initial, or recycle, distribution. First, we give

a mathematical detailing of the weighted ensemble algorithm and provide an unbiased property, ergodic property, and variance formula. The unbiased property gives that the weighted ensemble average of many Markov chain trajectories produces an unbiased estimate for the underlying Markov chain law. Next, the ergodic property states that the weighted ensemble estimator converges almost surely to the desired steady-state average. Lastly, the variance formula provides exact variance of the weighted ensemble estimator.

Next, we analyze the impact of the initial or recycle distribution, in A, on bias and variance of the weighted ensemble estimate and compare against adaptive multilevel splitting. Adaptive multilevel splitting is an importance splitting Markov chain Monte Carlo method also used in molecular dynamics for estimating mean first passage times. It has been studied that adaptive multilevel splitting requires a precise importance sampling of the initial, or recycle, distribution to maintain reasonable variance bounds on the adaptive multilevel splitting estimator. We show that the weighted ensemble estimator is less sensitive to the initial distribution since importance sampling the initial distribution frequently does not reduce the variance of the weighted ensemble estimator significantly. For a generic three state Markov chain and one dimensional overdamped Langevin dynamics, we develop specific conditions which must be satisfied for initial distribution importance sampling to provide a significant variance reduction on the weighted ensemble estimator. Finally, for bias, we develop conditions on A, such that the mean first passage time from A to B is stable with respect to changes in the initial distribution. That is, under a perturbation of the initial distribution the resulting change in the mean first passage time is insignificant. The conditions on A are verified with one dimensional overdamped Langevin dynamics and an example is provided. Furthermore, when the mean first passage time is unstable, we develop bounds, for one dimensional overdamped Langevin dynamics, on the change in the mean first passage time and show the tightness of the bounds with numerical examples.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. David Aristoff and all of his help throughout this project. Furthermore, many thanks to my peers in the graduate mathematics department at Colorado State University for insightful discussions.

DEDICATION

I would like to thank my parents and my wife Sarah for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Many real world problems of physics, chemistry, biology, finance, and engineering require techniques for sampling from a high dimensional distribution, optimization, or numerical integration. These three techniques are examples where Monte Carlo methods have valuable use [1]. A Monte Carlo method, in general, involves drawing random samples from some domain and performing a deterministic calculation on the random samples to achieve a desired result. For complex and high dimensional optimization problems, such as the traveling salesman problem for optimal delivery route design, Monte Carlo methods artificially inject randomness to more efficiently search the objective domain [2]. High dimensional integrals can be estimated through Monte Carlo methods by simulating a random variable or process whose expected value is the integral of interest [3, 4]. Finally, in Bayesian statistics problems, Markov Chain Monte Carlo methods, in particular the Metropolis-Hastings algorithm, are often used to sample from a posterior distribution that may have unknown normalization [5, 6].

Here, we focus on the use of Markov Chain Monte Carlo for numerical integration in biochemical and molecular dynamics systems. Markov chain Monte Carlo samples trajectories of a Markov chain, which has an underlying steady-state distribution of interest. Once the Markov chain has converged to steady-state, drawing samples of the Markov chain is equivalent to drawing samples from the steady-state distribution. Hence, when samples or averages with respect to a certain probability distribution are desired, Markov chain Monte Carlo can be applied by constructing a Markov chain with steady-state distribution equal to the probability distribution of interest. For instance, say we desire samples of a high dimensional probability distribution $p(x)$ but only know $f(x) \propto p(x)$. Then the Metropolis-Hastings algorithm, a Markov Chain Monte Carlo method, constructs a Markov chain with steady-state distribution $p(x)$ by using the ratio $p(x)/p(y) = f(x)/f(y)$ for determining the transition probability between points x and y [5]. In molecular dynamics, we desire an estimate of $\int f d\mu$, where μ is the steady-state distribution of

a Markov chain and f is a bounded function, as certain Markov chains, for example Langevin dynamics, can closely model molecular dynamics [7, 8]. By averaging f applied to samples of a Markov chain, which has converged to steady-state, the integral $\int f d\mu$ can be numerically approximated.

Consider a molecule moving in state space S from $A \subset S$ to $B \subset S$ where A and B are disjoint. We will model the molecular dynamics with a Markov chain $(X_t)_{t\geq 0}$, on state space S, which has an initial distribution, ρ, such that $\text{supp}(\rho) = A$, and upon reaching B the Markov chain is immediately recycled in A according to ρ. We refer to ρ as both the initial distribution and recycle distribution. Important quantities of interest include: the mean first passage time for a molecule to transition from A to B and the probability that a molecule transitions from A to B [9, 10]. We focus on estimating the mean first passage time from A to B given by $\mathbb{E}^{\rho}(\tau_B) = \mathbb{E}(\tau_B \,|\, X_0 \sim \rho)$ where $\tau_B = \inf\{t \geq : X_t \in B\}$ is the first passage time of $(X_t)_{t\geq 0}$ in state B. A naive Markov chain Monte Carlo approach to estimating the mean first passage time involves simulating $(X_t)_{t\geq 0}$ from A to B multiple times and averaging the time to reach B from each trial.

Often, in molecular dynamics, the transition from A to B is a rare event due to the presence of significant potential energy barriers between A and B [10, 11]. A rare event is a highly improbable event, for example a probability on the order of 10^{-8}, to occur. So, the mean first passage time of interest is substantial and applying naive Markov chain Monte Carlo is computationally infeasible due to the simulation time required. Often a Markov chain will converge to steady-state in a faster time than is necessary for multiple passages from A to B, which is required to estimate the mean first passage time directly [12]. Typically then, the mean first passage time is not estimated directly but instead estimated through the Hill relation which states that the steady-state flux from A into B is equal to $1/\mathbb{E}^{\rho}(\tau_B)$ [13]. That is, a Markov chain can be simulated to steady-state where an estimate of the steady-state flux from A into B is determined. Then the mean first passage time is obtained as the reciprocal of the estimated steady-state flux.

Using the Hill relation assists in reducing required computational times but introduces a large variance problem in estimating the steady-state flux. For example, consider estimating $p = \mathbb{P}(X \in$

B) for random variable X. A naive Monte Carlo technique draws samples $X_1, X_2, \ldots, X_N$, which are *i.i.d* to X, and estimates p by averaging the number in B given by $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_B(X_i)$, where $\mathbb{1}_B(x) = 1$ if $x \in B$ and $\mathbb{1}_B(x) = 0$ otherwise. Since $\mathbb{1}_B(X_1), \mathbb{1}_B(X_2), \ldots, \mathbb{1}_B(X_N)$ are Bernoulli random variables the estimate is unbiased $\mathbb{E}(\hat{p}) = p$ with variance Var$(\hat{p}) = p(1-p)/N$. Note, the relative standard deviation, or coefficient of variation, is $\sqrt{Var(\hat{p})}/\mathbb{E}(\hat{p}) = \sqrt{(1-p)/Np}$ which for small p scales as $\mathcal{O}((Np)^{-1/2})$. Implying that the variance of the estimator will be far larger than actual estimate when sampling in B is a rare event. In order to keep the variance within reasonable bounds the number of samples, N, will need to scale as $1/p$, which is computationally infeasible for small p. Intuitively, we expect to require roughly $1/p$ samples of X to sample in B once and have a nonzero estimate for $\hat{p}$. Since the transition from A to B is a rare event then the steady-state flux from A into B will be on the same order as estimating a rare event probability. Thus, applying naive Markov chain Monte Carlo to estimate the steady-state flux requires a computationally infeasible number of independent Markov chain trajectories to keep variance within reasonable bounds and obtain a meaningful, nonzero estimate. Still, using the Hill relation is beneficial as variation reduction techniques exist, such as importance sampling, which make estimating the steady-state flux feasible.

Importance sampling is a variance reduction technique in estimating rare events by allowing for an increased sampling of rare regions while leaving underlying probabilities unchanged [4, 14]. Consider sampling a random variable X, with distribution f, to estimate $\mathbb{E}^f(h(X))$ for a measurable function h. For example, estimating $\mathbb{P}(X \in B) = \mathbb{E}(\mathbb{1}_B(X))$ where where the probability to sample from B, $\int_B f(x)dx$, may be very small. Instead of directly sampling f, importance sampling draw samples of a different random variable Y, with distribution g having the same support as f, and each sample is assigned a weight $w(Y) = f(Y)/g(Y)$ where the function w satisfies $w(y) = 0$ when $f(y) = 0$. Then

$$\mathbb{E}^f(h(X)) = \int f(x)h(x)dx = \int g(y)w(y)h(y)dy = \mathbb{E}^g\left(w(Y)h(Y)\right).$$

So, importance sampling estimates $\mathbb{E}^f(h(X))$ by estimating $\mathbb{E}^g(w(Y)h(Y))$, which can be beneficial when $w(Y)h(Y)$ has less variance than $h(X)$. By taking $g(x) \propto f(x)|h(x)|$, which requires $\text{supp}(h) = \text{supp}(f)$, then

$$\text{Var}_f(h(X)) - \text{Var}_g(w(Y)h(Y)) = \mathbb{E}^f(h^2(X)) - \mathbb{E}^g((w(Y)h(Y))^2) = \text{Var}_f(|h(X)|) \geq 0 \quad (1.1)$$

so the variance of $w(Y)h(Y)$ is less than or equal to the variance of $h(X)$. A particularly used instance of importance sampling is when h is a non-negative function, then $w(Y)h(Y)$ is a constant and hence has zero variance.

One method combining Markov chain Monte Carlo and importance sampling is the weighted ensemble algorithm [12,15]. Weighted ensemble, detailed in Section 2.3, simulates generic Markov chains through an interacting particle scheme that splits, merges, and updates weights of many Markov chain trajectories. This scheme allows for certain regions, such as the rare region of interest B, to be emphasized by splitting Markov chain trajectories near B into many unique copies, known as importance splitting, while merging those further away. A weight is assigned to each trajectory and a simple method of updating weights, by dividing the current weight by the expected number of new copies, provides the desired importance sampling weight. In general, weighted ensemble estimates $\int f d\mu$ which is the average of a measurable function f with respect to the Markov chain steady-state distribution μ. For instance, by taking $f = \mathbb{1}_B$ then $\int f d\mu$ gives the steady-state flux from A into B. The weighted ensemble estimator is produced by time averaging a weighted average of f evaluated at the simulated Markov chain trajectories.

Another common rare event variance reduction method, for Markov chain Monte Carlo in molecular dynamics, is the adaptive multilevel splitting algorithm [11]. Adaptive multilevel splitting estimates the probability of reaching B before A, that is $\mathbb{P}(\tau_B < \tau_A)$, by using importance splitting to estimate a sequence of conditional event probabilities. Each conditional event is no longer a rare event and the product of these conditional event probabilities gives the quantity of interest $\mathbb{P}(\tau_B < \tau_A)$. Once an estimate of the probability of reaching B before A is obtained it can then be used to estimate the mean first passage time from A to B.

Here, we explore the impact of the Markov chain initial and recycle distribution, ρ, on bias and variance of the weighted ensemble steady-state average estimator and compare against adaptive multilevel splitting. First, we provide a detailed, mathematical description of the weighted ensemble algorithm and further details on the adaptive multilevel splitting algorithm. Second, for bias, we develop conditions on A such that the mean first passage time from A to B is stable with respect to changes in the initial distribution. That is, under a perturbation of the initial distribution the resulting change in the mean first passage time is insignificant. The conditions on A are verified with one dimensional overdamped Langevin dynamics. Furthermore, when the mean first passage time is unstable we develop bounds, for one dimensional overdamped Langevin dynamics, on the change in the mean first passage time and show the tightness of the bounds with numerical examples. Next, an explanation for how adaptive multilevel splitting estimates the mean first passage time from A to B along with a discussion on anticipated bias stability is given. Finally, for variance, it has been studied that adaptive multilevel splitting requires a precise importance sampling of the initial, or recycle, distribution to maintain reasonable variance bounds on the adaptive multilevel splitting estimator [16]. We show that the weighted ensemble estimator is less sensitive to the initial distribution since importance sampling the initial distribution frequently does not reduce the variance of the weighted ensemble estimator significantly. For a generic three state Markov chain and one dimensional overdamped Langevin dynamics, we develop conditions that must be satisfied for initial distribution importance sampling to provide a significant variance reduction on the weighted ensemble estimator. In cases where initial distribution importance sampling has an insignificant impact on the variance, we say the weighted ensemble has variance stability and importance sampling of the initial distribution is not required.

Chapter 2

Weighted Ensemble

Weighted ensemble is an importance sampling Markov chain Monte Carlo method, which uses an interacting particle scheme to simulate generic Markov chains [17]. The original design of weighted ensemble was for solving problems in computational chemistry; in particular, weighted ensemble was originally used in simulating protein folding and association times through Brownian motion [12, 15]. One application of particular importance is the calculation of mean first passage times for a molecule to transition from a source region A to a sink region B. Mean first passage times have been used in a variety of biochemical studies including protein folding times, metastatic cancer progression, and polymer translocation [18].

An example mean first passage time problem is shown in Figure 2.1 where we desire the mean first passage time for a molecule to transition between two metastable potential wells, regions where the molecule remains for a significant time, by surpassing a potential barrier. A naive

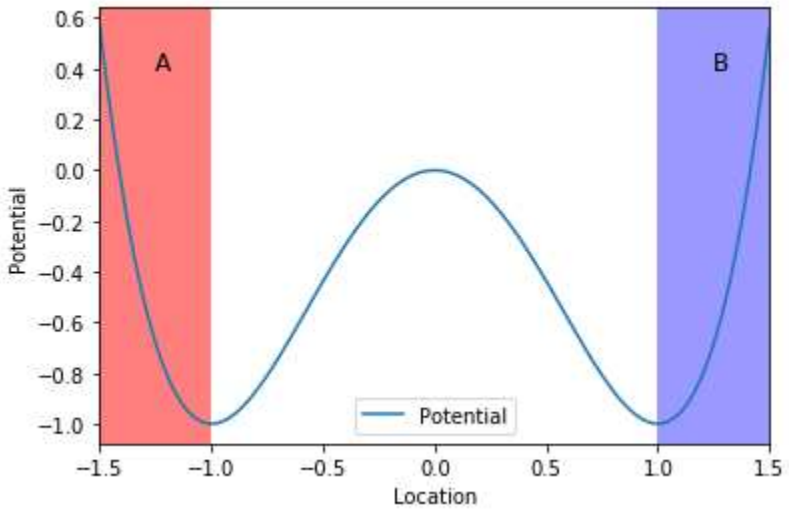


Figure 2.1: Example mean first passage time setup where we desire to know the average time for a molecule to transition from the left region A over the potential barrier to the right region B.

Markov chain Monte Carlo approach to estimating the mean first passage time involves simulating

a Markov chain $(X_t)_{t\geq 0}$, which has dynamics similar to the molecule of interest, from A to B many times and averaging the time to reach B from each trial. Often, the molecular transition of interest involves surpassing a significant potential energy barrier, which is a rare event [10]. So, estimating the mean first passage time by naive Markov chain Monte Carlo is infeasible in the simulation time required. Instead, the steady-state flux from A into B is calculated, which by the Hill relation is the reciprocal of the mean first passage time [13]. Estimating the steady-state flux poses challenges in the amount of variance of the estimator. So, an importance sampling method, such as weighted ensemble, is required to reduce variance and provide computationally feasible estimates of the steady-state flux and thus mean first passage time. Another importance sampling method for mean first passage time problems in molecular dynamics is adaptive multilevel splitting.

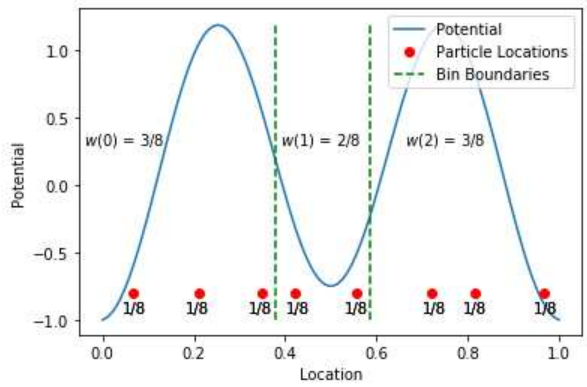
2.1 Weighted Ensemble Description

Here we provide a short, non-technical description of the weighted ensemble algorithm. Initializing weighted ensemble involves generating a predetermined number of Markov chain trajectory starting points in state space and assigning each a non-negative weight such that the total weight is unity. Subsequently, each step in the weighted ensemble algorithm consists of two procedures: a selection process and a mutation. First, the selection process broadly involves the merging, removal, and weighting of many Markov chain trajectories to encourage an overall transition to the desired state space region B. Second, mutation is the transition of each Markov chain trajectory according to the underlying Markov kernel.

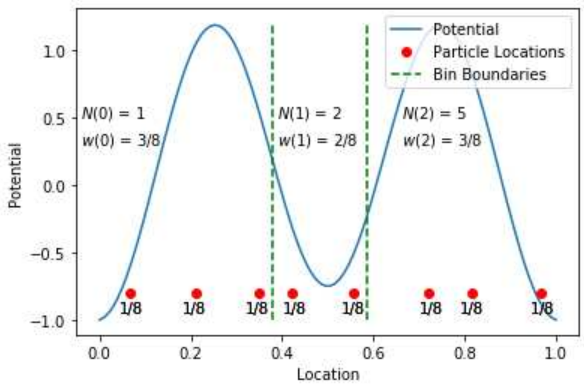
Now we provide further details on each selection process. We call each trajectory of the Markov chain a particle and by genealogical analogy, we call the particles before selection parents and after selection children. Each selection process is broken into the following four steps:

(i) Parents are separated into groups, which we call bins and the total weight of the parents in each bin is calculated. Figure 2.2a shows an example where eight particles, all with equal weight, are separated into three bins. Note that the bins are user chosen parameters that are allowed to vary between different selections.

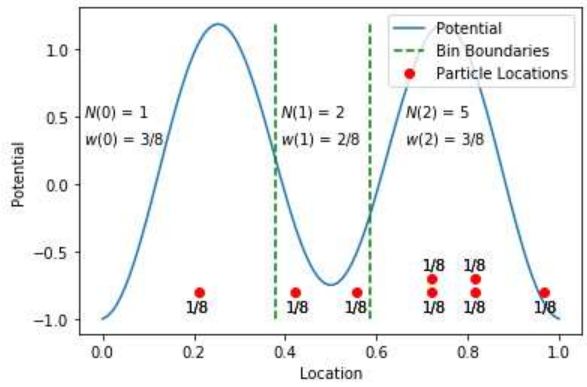
(ii) Each bin is assigned a positive whole number of children, which we call the particle allocation. The total number of children assigned across all bins must be equal to the number of parents. Figure 2.2b shows an example particle allocation where the first bin has a single child, the second bin has two children, and the third bin has five children. Note that the particle allocation is a user chosen parameter which can vary for different selections.



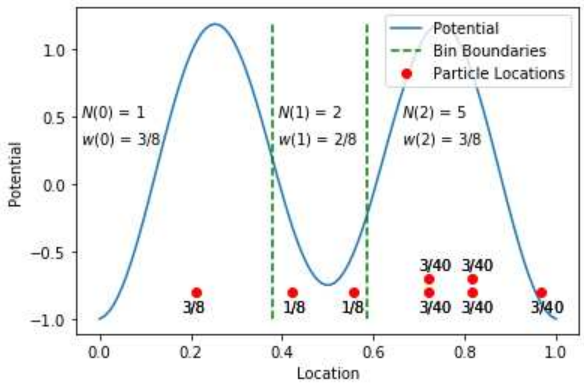
(a) Step (i) in the weighted ensemble selection process. Eight particles, with equal weights, are grouped into three bins and the total bin weight is determined.



(b) Step (ii) in the weighted ensemble selection process. One child is assigned to bin one, two are assigned to bin two, and five children assigned to bin three.



(c) Step (iii) in the weighted ensemble selection process. One sample is drawn from the three particles in bin one, two samples are drawn from bin two, and five samples are drawn from the particles in bin three.



(d) Step (iv) in the weighted ensemble selection process where, in each bin, all particles are assigned a new weight given by the original total bin weight over the number of children assigned to the bin.

Figure 2.2: An example of the weighted ensemble selection process in 1d. The plot on each figure is the potential in which the molecule traverses through. The points on each figure are the x positions of the particles and the number near each point is the particle weight.

(iii) Within each bin, the parents are sampled, proportionally to their weights, a number of times equal to the assigned number of children to the bin. Each child is, initially, an identical copy

of its parent. Figure 2.2c shows an example where in the first bin one sample is drawn, two samples are drawn in the second bin, and five samples are drawn in the third bin.

(iv) After sampling, all children in a given bin are assigned a new, and equivalent, weight calculated by the total bin weight divided by the number of created children. Figure 2.2d shows an example where the eight sampled children are reweighted.

Note that the selection process keeps the total number of particles constant and the total weight at unity. By assigning more children to bins that are closer to the desired state space region B, the weighted ensemble selection process effectively importance samples these regions.

An estimate of the steady-state flux from A into B is given by the time average, over a sufficiently long time, of the weight of the particles entering B. The reciprocal of the time average of weight entering B is the mean first passage time from A to B. Once a particle reaches B it is recycled, or restarted, in A according to a source distribution, which we call the initial distribution while maintaining the same weight.

2.2 Adaptive Multilevel Splitting Description

Next, we provide a description of the adaptive multilevel splitting algorithm to highlight key differences between it and the weighted ensemble algorithm. As with weighted ensemble, adaptive multilevel splitting can be used to estimate mean first passage times in molecular dynamics. Foundationally though, the adaptive multilevel splitting estimator is the probability to transition to B before transitioning to A, given by $\mathbb{P}(\tau_B < \tau_A)$. Once $\mathbb{P}(\tau_B < \tau_A)$ is estimated through adaptive multilevel splitting it can then be used to calculate the mean first passage time from A to B as discussed further in Section 3.1.

To initialize adaptive multilevel splitting, a predetermined number, N, of Markov chain trajectory starting points, $X_0^1, X_0^2, \ldots, X_0^N$, are sampled from an initial distribution. Similar to weighted ensemble, adaptive multilevel splitting consists of a selection process and a mutation. First, mutation is the transition of each Markov chain trajectory, according to the underlying Markov kernel, until A or B is reached. As the transition from A to B is a rare event, it is highly probable a

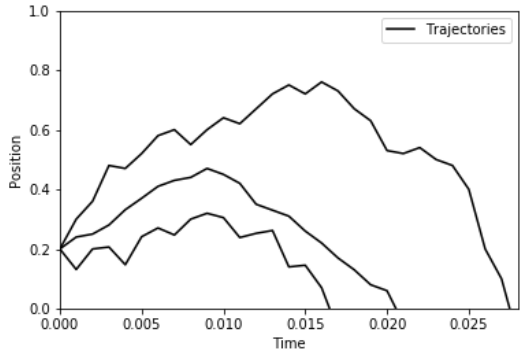
mutation will end with all trajectories reaching A. An example of adaptive multilevel splitting mutation is shown in Figure 2.3a where three 1d trajectories are initialized at $\frac{2}{10}$ and simulated until $A = \{0\}$ or $B = \{1\}$ is reached. Second, the selection process determines which trajectories to keep and copy based on a user chosen reaction coordinate, Φ, and a minimum number of trajectories to discard, K, which are the same across all selections. The reaction coordinate is a real valued function defined on the Markov chain state space, which provides a metric of distance that the input trajectory is from the desired state space region B. Note, the optimal reaction coordinate is the committor function given by $\Phi(x) = \mathbb{P}(\tau_B < \tau_A \,|\, X_0 = x)$. Now the selection process can be summarized in the following steps:

(i) Determine the maximum reaction coordinate value, M_i, of each trajectory, X_t^i, across its simulation time. That is, for set S, if we define $\tau_S^i = \inf\{t \geq 0 : X_t^i \in S\}$ and $\tau^i = \min\{\tau_A^i, \tau_B^i\}$ then $M_i = \max_{0 \leq s \leq \tau^i} \Phi(X_s^i)$. Next, the maximum reaction coordinate values are sorted, in increasing order, forming the order statistics $M_{(1)}, M_{(2)}, \ldots, M_{(N)}$. An example is given in Figure 2.3b with reaction coordinate $\Phi(x) = x$ where the plotted points are the maximum reaction coordinate values.

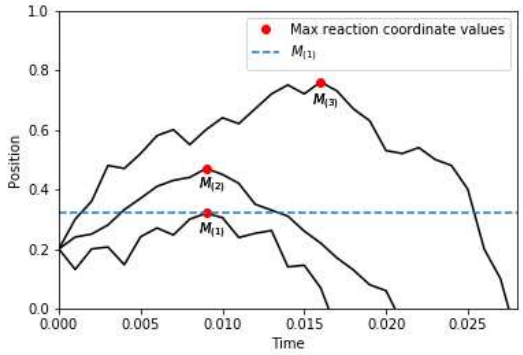
(ii) Discard all D trajectories, X_t^i, satisfying $M_i \leq M_{(K)}$. Randomly sample, with replacement, the $N - D$ stored trajectories D times and copy the sampled trajectory until the first time it enters $\{x : \Phi(x) > M_{(K)}\}$. An example is given in Figure 2.3c with $D = K = 1$ where the second trajectory was sampled once.

We can note that the number of discarded trajectories D will be at least as large as K and may be larger if multiple trajectories have a maximum reaction coordinate value equal to $M_{(K)}$. Following a selection process, the next mutation simulates only the D samples starting from the time each trajectory first surpassed $M_{(K)}$ as shown in Figure 2.3d.

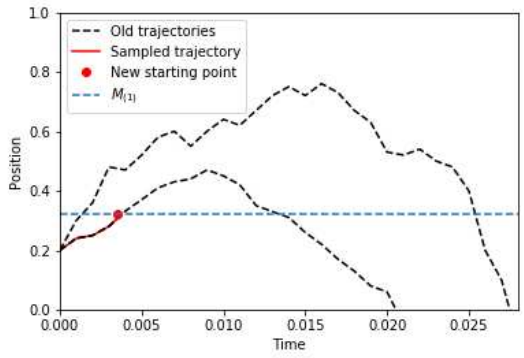
Let $M_{(K)}^t$ and D_t be the Kth order statistic of the maximum reaction coordinate values and the number of discarded trajectories on the tth selection process, respectively. Upon $M_{(K)}^t$ surpassing a user specified parameter $M_{\max}$, one final mutation is performed and then the adaptive multilevel



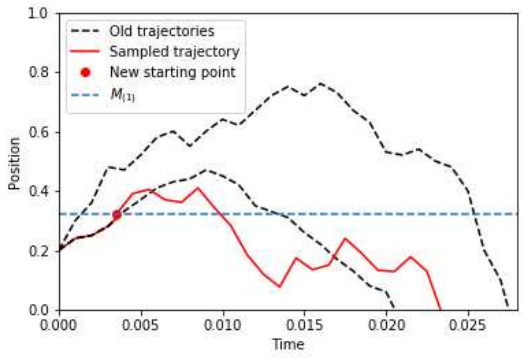
(a) Adaptive multilevel splitting mutation of three trajectories initialized at $X_0^1 = X_0^2 = X_0^3 = 2/10$. Each trajectory is simulated until $A = \{0\}$ or $B = \{1\}$ is reached.



(b) Step (i) in the adaptive multilevel splitting selection process. The maximum reaction coordinate values, the three points on the plot, are identified and sorted to give $M_{(1)}, M_{(2)}$, and $M_{(3)}$.



(c) Step (ii) in the selection process with $K = 1$. The trajectory with maximum reaction coordinate value equal to $M_{(1)}$ is discarded. The trajectory on the right was sampled and the point shows where it first surpasses $M_{(1)}$.



(d) The next mutation in the adaptive multilevel splitting algorithm. The sampled trajectory from step (ii) of the selection process is evolved starting from the point where it first surpassed $M_{(1)}$. Each trajectory is simulated until $A = \{0\}$ or $B = \{1\}$ is reached.

Figure 2.3: An example of the adaptive multilevel splitting process in 1d using $N = 3$ trajectories, discarding at least $K = 1$ trajectories in each selection, and using reaction coordinate $\Phi(x) = x$. The two target sets of state space are $A = \{0\}$ and $B = \{1\}$. Each trajectory is initialized at $2/10$. All trajectories are moving in a potential $V(x)$, which for clarity has not been plotted.

splitting algorithm returns an estimate of $\mathbb{P}(\tau_B < \tau_A)$. Assume $M_{(K)}^t$ first surpasses $M_{\max}$ at step $t = T$ and denote $X_{\tau^1}^1, X_{\tau^2}^2, \ldots, X_{\tau^N}^N$ as the trajectories at the end of the final mutation. Then the adaptive multilevel splitting estimator is given by

$$\widehat{p} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_B(X_{\tau^i}^i) \left(\prod_{t=1}^{T} \frac{N - D_t}{N} \right)$$

which provides an estimate of $\mathbb{P}(\tau_B < \tau_A)$. To understand why $\hat{p}$ estimates $\mathbb{P}(\tau_B < \tau_A)$, define level $L_t = \{x : \Phi(x) = M^t_{(K)}\}$, which is the contour of $\Phi(x)$ at $M^t_{(K)}$. Then, iteratively applying the definition of conditional probability

$$\mathbb{P}(\tau_B < \tau_A) = \mathbb{P}(\tau_B < \tau_A \,|\, \tau_{L_T} < \tau_A)\mathbb{P}(\tau_{L_1} < \tau_A) \prod_{i=2}^{T} \mathbb{P}(\tau_{L_i} < \tau_A \,|\, \tau_{L_{i-1}} < \tau_A).$$

Now, adaptive multilevel splitting implements the approximations

$$\mathbb{P}(\tau_{L_1} < \tau_A) \approx \frac{N - D_1}{N}, \qquad \mathbb{P}(\tau_{L_i} < \tau_A \,|\, \tau_{L_{i-1}} < \tau_A) \approx \frac{N - D_i}{N},$$

and

$$\mathbb{P}(\tau_B < \tau_A \,|\, \tau_{L_T} < \tau_A) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_B(X^i_{\tau^i}),$$

which are asymptotically equal as $N \to \infty$. Thus, $\mathbb{P}(\tau_B < \tau_A)$ is approximated by the adaptive multilevel splitting estimator $\widehat{p}$. Therefore, adaptive multilevel splitting estimates rare event probabilities, by estimating a sequence conditional event probabilities where each conditional event is no longer a rare event.

Lastly, we discuss a few key differences between adaptive multilevel splitting and weighted ensemble. First, adaptive multilevel splitting has only five user chosen parameters: the initial distribution, the number of trajectories, the minimum number of trajectories to discard, the reaction coordinate, and the final level $M_{\max}$. On the other hand, weighted ensemble has far more user chosen parameters as each selection step requires both user chosen bins and particle allocation. More user chosen parameters can be disadvantageous as poor choices of these parameters can lead to poor performance of the algorithm. For instance, the reaction coordinate provides the main source of variance in adaptive multilevel splitting estimator [19]. An advantage of the weighted ensemble parameters is that the bins and particle allocations can easily be chosen for higher dimensional Markov chains as they simply require partitioning the trajectories and assigning a positive number to each set in the partition. Whereas a satisfactory choice for the adaptive multilevel splitting

reaction coordinate is more challenging in higher dimensions. Although, asymptotically in the large number of trajectories limit, adaptive multilevel splitting estimates the committor function, the optimal reaction coordinate given by $\Phi(x) = \mathbb{P}(\tau_B < \tau_A \,|\, X_0 = x)$, which could be used to adaptively update the reaction coordinate across multiple trials [10].

A second difference is that adaptive multilevel splitting has greater storage requirements than weighted ensemble. In weighted ensemble, only the current weights and current trajectory positions need to be stored. On the other hand, adaptive multilevel splitting requires the storage of, potentially large, paths for each trajectory which could be costly for a large number of trajectories or a high dimensional state space. Although, a benefit to storing the trajectory paths is that adaptive multilevel splitting produces both an estimate of a rare event probability and a collection of empirical trajectories reaching it, which can be used to estimate any observable given the rare event [19].

Thirdly, the weighted ensemble estimator relies upon the convergence of the underlying Markov chain to steady-state and the ergodicity of time averaging a weighted average of Markov chains. Instead, the adaptive multilevel splitting estimator relies on the decomposition of a probability into the product of conditional probabilities which are each individually estimated. Fourthly, weighted ensemble is robust in handling intermediate barriers and metastable regions between A and B [20]. As metastable regions and intermediate barriers are often present in practical problems such robustness is desirable. Finally, as we highlight further in Chapter 3 weighted ensemble is at least as stable in bias and variance, with respect to changes in the initial distribution, compared to adaptive multilevel splitting. Adaptive multilevel splitting requires a precise importance sampling on the initial distribution to maintain reasonable estimate variance [16], which is typically not required in weighted ensemble. For bias, adaptive multilevel splitting and weighted ensemble can both produce significantly different estimates under perturbations of the initial distribution.

2.3 Weighted Ensemble Mathematical Foundations

We now provide a mathematical detailing of the weighted ensemble algorithm which is summarized in Algorithm 1. Let $(X_t)_{t\geq 0}$ be a Markov chain, on state space S, governed by Markov kernel K with initial distribution ρ. Assume that K is uniformly, geometrically ergodic with respect to its stationary, or steady-state, distribution μ. Using weighted ensemble we are interested in estimating

$$\int f d\mu$$

the steady-state average of a bounded, real-valued function f on state space, which is called an observable. When estimating the steady-state flux into $B \subset S$, to obtain a mean first passage time estimate through the Hill relation, we take $f = \mathbb{1}_B$.

Let N be the total number of particles used in the algorithm and $T-1$ denote the total number of weighted ensemble steps, that is $T-1$ selection processes and $T-1$ mutations, in the algorithm. The time t denotes the number of weighted ensemble steps which have occurred. We denote the N parent particles, and the corresponding weight of each particle, at time t, by the sequence of tuples

$$\left(\xi_t^1, w_t^1\right), \left(\xi_t^2, w_t^2\right), \ldots, \left(\xi_t^N, w_t^N\right),$$

which are initialized by sampling $\xi_0^1, \xi_0^2, \ldots, \xi_0^N$ from state space, S, and assigned each a positive weight $w_0^1, w_0^2, \ldots, w_0^N$ such that $w_0^1 + w_0^2 + \cdots w_0^N = 1$. For example, we could sample $\xi_0^1, \xi_0^2, \ldots, \xi_0^N$ *i.i.d* from the initial distribution, ρ, or steady-state distribution, μ, and take uniform weights $w_0^1 = \cdots = w_0^N = \frac{1}{N}$. Now the weighted ensemble average of f at time t is given by

$$\langle f \rangle_t \coloneqq \sum_{i=1}^{N} w_t^i f(\xi_t^i). \tag{2.1}$$

Then, we define the weighted ensemble estimate, θ_T, by

$$\theta_T \coloneqq \frac{1}{T}\sum_{t=0}^{T-1}\langle f \rangle_t = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} w_t^i f(\xi_t^i). \tag{2.2}$$

That is θ_T is the time average of the weighted ensemble average of f and provides an estimate of the steady-state average of f for large T. Note, for any initialization except sampling from the steady-state distribution, θ_T is a biased estimate of $\int f d\mu$, for finite T, as the Markov chain must converge to steady-state. So, the steady-state distribution, μ, although typically not known in practice is an optimal initialization.

Each step of the weighted ensemble algorithm evolves every particle through a selection processes and a mutation. That is for each time, $t \geq 0$, and every parent particle and weight, (ξ_t^i, w_t^i), we have

$$\left\{(\xi_t^i,\, w_t^i)\right\}_{i=1}^N \xrightarrow{\text{selection}} \left\{\left(\widehat{\xi}_t^i,\, \widehat{w}_t^i\right)\right\}_{i=1}^N \xrightarrow{\text{mutation}} \left\{(\xi_{t+1}^i,\, w_{t+1}^i)\right\}_{i=1}^N$$

where

$$\left(\widehat{\xi}_t^1,\, \widehat{w}_t^1\right),\, \left(\widehat{\xi}_t^2,\, \widehat{w}_t^2\right), \ldots, \left(\widehat{\xi}_t^N,\, \widehat{w}_t^N\right)$$

denote the children particles and their corresponding weights after the selection process.

First is the weighted ensemble selection process where each time $t \geq 0$ requires user chosen bins, B_t, and particle allocation, N_t. The bins, B_t, are a partition of the particles $\{\xi_t^1, \xi_t^2, \ldots, \xi_t^N\}$. Now the particle allocation is a function $N_t : B_t \to \mathbb{N}$ where $N_t(u)$ gives the number of children assigned to bin $u \in B_t$ and satisfies $\sum_{u \in B_t} N_t(u) = N$ for all $t \geq 0$. Note that the bins, B_t, and particle allocation, N_t, may be informed by current system information such as the particle positions and are unknown random variables for time $s < t$. For each bin $u \in B_t$ we define the total bin weight by

$$w_t(u) = \sum_{i=1}^N w_t^i \mathbb{1}_u(\xi_t^i) \quad \text{where} \quad \mathbb{1}_u(x) = \begin{cases} 1 & x \in u \\ 0 & \text{else.} \end{cases}$$

Then we can define a bin conditional particle weight distribution by

$$p(\xi_t^i \,|\, u) = \frac{w_t^i}{w_t(u)} \mathbb{1}_u(\xi_t^i)$$

which gives the probability to draw ξ_t^i, when sampling from bin u, in a single random sample. After the bins and particle allocation have been assigned, the selection process uses the bin conditional particle weight distribution and, for each bin u, draws $N_t(u)$ samples, with replacement, from the parent particles in u to create the children particles. That is if we define N_t^i to be the number of children of the parent ξ_t^i then for each bin $u \in B_t$

$$\{N_t^i \,:\, \xi_t^i \in u\} \sim \text{Multinomial}\left(N_t(u), \left\{p(\xi_t^i \,|\, u) \,:\, \xi_t^i \in u\right\}\right)$$

where sampling is independent in distinct bins.

Define par $\left(\widehat{\xi}_t^i\right)$ to be the parent of $\widehat{\xi}_t^i$ where par $\left(\widehat{\xi}_t^i\right) = \xi_t^j$ for some $j \in \{1, 2, \ldots, N\}$. Then each child, $\widehat{\xi}_t^i$, is an exact copy of its parent, par $\left(\widehat{\xi}_t^i\right)$, and $\widehat{\xi}_t^i \in u$ if and only if par $\left(\widehat{\xi}_t^i\right) \in u$. After the children have been created, they are reweighted by assigning all children in bin u the same weight

$$\widehat{w}_t^j = \frac{w_t(u)}{N_t(u)} \quad \text{if} \quad \widehat{\xi}_t^j \in u$$

thereby creating children $\widehat{\xi}_t^1$, $\widehat{\xi}_t^2$, $\ldots$, $\widehat{\xi}_t^N$ with weights $\widehat{w}_t^1$, $\widehat{w}_t^2$, $\ldots$, $\widehat{w}_t^N$ and completing the selection process for weighted ensemble. Note that the reweighted process leaves the total weight in each bin $w_t(u)$ fixed so after selection $\widehat{w}_t^1 + \widehat{w}_t^2 + \cdots + \widehat{w}_t^N = 1$.

Second is the weighted ensemble mutation where each child, $\widehat{\xi}_t^j$, is evolved independently according to the Markov kernel K creating a new parent ξ_{t+1}^j. That is the distribution of ξ_{t+1}^j is $K\left(\widehat{\xi}_t^j, \cdot\right)$ for each $j \in \{1, 2, \ldots, N\}$. Particle weights are unchanged during the mutation step so $w_{t+1}^j = \widehat{w}_t^j$ for each $j \in \{1, 2, \ldots, N\}$. Hence, after mutation, a new collection of parents and weights $\left(\xi_{t+1}^1, w_{t+1}^1\right), \left(\xi_{t+1}^2, w_{t+1}^2\right), \ldots, \left(\xi_{t+1}^N, w_{t+1}^N\right)$ have been created for the next weighted ensemble selection process.

Finally, once the maximum number of weighted ensemble steps, $T-1$, is reached the weighted ensemble estimator, θ_T, provides an approximation for the desired steady-state quantity $\int f d\mu$. Note that the weighted ensemble estimator can be continually updated with each step, t, of the

weighted ensemble algorithm and so there is no requirement to store particle trajectories or previous weights throughout the algorithm.

Algorithm 1 Weighted Ensemble Algorithm

Require: Number of particles $N > 0$. Number of iterations $T - 1 \geq 0$. Markov kernel K.

Initialize: Choose initial particles $\xi_0^1, \ldots, \xi_0^N$ and positive weights $w_0^1, \ldots, w_0^N$ which sum to 1.

while $t \leq T - 1$ **do**

Group parents $\xi_t^1, \ldots, \xi_t^N$ in to bins B_t.

Assign each bin $u \in B_t$ a number of children $N_t(u) \geq 1$ such that $\sum_{u \in B_t} N_t(u) = N$.

for bin $u \in B_t$ **do**

Sample the parents $N_t(u)$ times with replacement according to

$$\mathbb{P}(\text{sample } \xi_t^i \text{ from bin } u) = p(\xi_t^i \,|\, u) = \frac{w_t^i}{w_t(u)} \mathbb{1}_u(\xi_t^i), \quad \text{where} \quad w_t(u) = \sum_{i=1}^{N} w_t^i \mathbb{1}_u(\xi_t^i).$$

Assign all $N_t(u)$ children the same weight $\widehat{w}_t^i = w_t(u)/N_t(u)$ if $\widehat{\xi}_t^i \in u$.

end for

Evolve all children $\widehat{\xi}_t^1, \ldots, \widehat{\xi}_t^N$ independently according to the Markov kernel K creating new parents $\xi_{t+1}^1, \ldots, \xi_{t+1}^N$ while keeping the weights fixed $w_{t+1}^j = \widehat{w}_t^j$ for $j = 1, 2, \ldots, N$.

end while

Return: Steady-state estimate $\theta_T = \dfrac{1}{T} \displaystyle\sum_{t=0}^{T-1} \sum_{i=1}^{N} w_t^i f(\xi_t^i)$

Next, we discuss two formulations for mutation. First, when the Markov kernel K is an $n \times n$ stochastic matrix on state space $S = \{1, 2, \ldots, n\}$ then the mutation of each child $\widehat{\xi}_t^j$ is performed by sampling

$$\xi_{t+1}^j \sim \text{Multinomial}\left(1, \left\{K\left(\widehat{\xi}_t^j, 1\right), K\left(\widehat{\xi}_t^j, 2\right), \ldots, K\left(\widehat{\xi}_t^j, n\right)\right\}\right)$$

where $K(i,j)$ is the ijth entry of matrix K. Second, when the Markov chain $(X_t)_{t\geq 0}$ satisfies a stochastic differential equation then we take the Markov kernel, K, to be a Δt skeleton of the underlying dynamics by evaluating the Markov process at resampling times Δt [21]. An example stochastic differential equation for $(X_t)_{t\geq 0}$, on state space $S = \mathbb{R}^d$, is

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t$$

where W_t is standard Brownian motion and $\mu : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^+$ are problem specific functions for $\mathbb{R}^+ = \{x \in \mathbb{R} : x \geq 0\}$. Now we can approximate the Markov process by taking an Euler-Maryuama discretization, over time step δt, of the stochastic differential equation giving

$$X_{t+\delta t} = X_t + \mu(X_t, t)\delta t + \sigma(X_t, t)\sqrt{\delta t}Z_t$$

where Z_t is a standard normal random vector in $\mathbb{R}^d$. Then the mutation for each child, $\widehat{\xi}_t^j$, consists of a user chosen number $k > 0$ discretization update steps

$$\xi_{t+1}^j = \underbrace{m \circ m \circ \cdots \circ m}_{k \text{ times}} \left(\widehat{\xi}_t^j\right) = m^k \left(\widehat{\xi}_t^j\right)$$

where $m(X_t) = X_t + \mu(X_t, t)\delta t + \sigma(X_t, t)\sqrt{\delta t}Z_t$. Such a discretization produces an approximation of both the underlying dynamics and the Markov kernel K with resampling time $\Delta t = k\delta t$. Note that another discretization scheme, such as Milstein or Runge-Kutta, could also be used for approximating the mutation of each child particle.

2.3.1 Weighted Ensemble Properties

Here we give an unbiased property of weighted ensemble, ergodic property of weighted ensemble, and an exact variance formula for the weighted ensemble estimate, θ_T. First, the unbiased property gives that the weighted ensemble average of f, equation (2.1), produces, for each $t \geq 0$,

an unbiased estimate of the law of the underlying Markov chain. That is, if Markov chain $(X_t)_{t\geq 0}$ with kernel K has the weighted ensemble initial distribution $\nu_0(g) = \mathbb{E}\left(\sum_{i=1}^N w_0^i g(\xi_0^i)\right)$, for measurable function g, then $\mathbb{E}(\langle f\rangle_s) = \mathbb{E}(f(X_s))$ for all $s \geq 0$. Note that $K^s g(\xi)$ gives the expectation of $g(X_t)$ when evolved s steps, by kernel K, from $X_0 = \xi$, so $\mathbb{E}(f(X_t)) = \mathbb{E}\left(\sum_{i=1}^N w_0^i K^t f(\xi_0^i)\right)$. Thus, the following proposition gives the unbiased property of weighted ensemble.

Proposition 2.3.1. *For all bounded, measurable functions f and each $t \geq 0$*

$$\mathbb{E}(\langle f\rangle_t) = \mathbb{E}\left(\sum_{i=1}^N w_0^i K^t f(\xi_0^i)\right).$$

Proof. Consider two filtrations

$$\mathcal{F}_t = \sigma\left(\left\{\left(\xi_s^i, w_s^i\right)\right\}_{0\leq s\leq t}^{1\leq i\leq N}, \{B_s, N_s\}_{0\leq s\leq t}, \left\{\left(\widehat{\xi}_s^i, \widehat{w}_s^i\right)\right\}_{0\leq s\leq t-1}^{1\leq i\leq N}, \left\{N_s^i\right\}_{0\leq s\leq t-1}^{1\leq i\leq N}\right) \tag{2.3}$$

$$\widehat{\mathcal{F}}_t = \sigma\left(\left\{\left(\xi_s^i, w_s^i\right)\right\}_{0\leq s\leq t}^{1\leq i\leq N}, \{B_s, N_s\}_{0\leq s\leq t}, \left\{\left(\widehat{\xi}_s^i, \widehat{w}_s^i\right)\right\}_{0\leq s\leq t}^{1\leq i\leq N}, \left\{N_s^i\right\}_{0\leq s\leq t}^{1\leq i\leq N}\right) \tag{2.4}$$

which are the σ-algebras generated by the weighted ensemble algorithm before and after the tth selection processes, respectively. Note that $\mathcal{F}_s \subset \widehat{\mathcal{F}}_s$ for all $s \geq 0$. We now calculate the expected weighted ensemble average of f over each selection and mutation step conditional on the information up to each step. First, for mutation, note that for each $t \geq 0$ and $i \in \{1, 2, \ldots, N\}$ we have $w_{t+1}^i = \widehat{w}_t^i$, which is $\widehat{\mathcal{F}}_t$ measurable, and since ξ_{t+1}^i is distributed according to $K\left(\widehat{\xi}_t^i, \cdot\right)$ then $\mathbb{E}\left(f(\xi_{t+1}^i) \,\middle|\, \widehat{\mathcal{F}}_t\right) = Kf\left(\widehat{\xi}_t^i\right)$. Thus, the expected weighted ensemble average from mutation is

$$\mathbb{E}\left(\langle f\rangle_{t+1} \,\middle|\, \widehat{\mathcal{F}}_t\right) = \sum_{i=1}^N \mathbb{E}\left(w_{t+1}^i f(\xi_{t+1}^i) \,\middle|\, \widehat{\mathcal{F}}_t\right) = \sum_{i=1}^N \widehat{w}_t^i \mathbb{E}\left(f(\xi_{t+1}^i) \,\middle|\, \widehat{\mathcal{F}}_t\right) = \sum_{i=1}^N \widehat{w}_t^i K f\left(\widehat{\xi}_t^i\right). \tag{2.5}$$

Second, for selection, at each $t \geq 0$ we consider each bin $u \in B_t$ of particles. Since the children $\widehat{\xi}_t^i \in u$ are samples, and exact copies, of the parents $\xi_t^j \in u$ where each parent $\xi_t^j \in u$ is sampled N_t^j times we can write $\sum_{i=1}^N f\left(\widehat{\xi}_t^i\right) \mathbb{1}_u\left(\widehat{\xi}_t^i\right) = \sum_{j=1}^N N_t^j f(\xi_t^j) \mathbb{1}_u(\xi_t^j)$. By definition of the reweighting we have, in bin u, $\widehat{w}_t^i = w_t(u)/N_t(u)$, which is $\mathcal{F}_t$ measurable. The number of children N_t^j of

parent ξ_t^j is multinomial distributed so $\mathbb{E}(N_t^j \mid \mathcal{F}_t) = N_t(u)w_t^j/w_t(u)$ for $1 \leq j \leq N$. Thus, since ξ_t^j and B_t are $\mathcal{F}_t$ measurable, the expected weighted ensemble average from selection is

$$\mathbb{E}\left(\sum_{i=1}^{N} \widehat{w}_t^i f\left(\widehat{\xi}_t^i\right) \middle| \mathcal{F}_t\right) = \sum_{u \in B_t} \frac{w_t(u)}{N_t(u)} \sum_{j=1}^{N} f(\xi_t^j) \mathbb{1}_u(\xi_t^j) \mathbb{E}\left(N_t^j \mid \mathcal{F}_t\right) = \sum_{j=1}^{N} w_t^j f(\xi_t^j). \tag{2.6}$$

Now, using the tower property, equation (2.5), and equation (2.6)

$$\mathbb{E}\left(\langle f \rangle_{t+1} \mid \mathcal{F}_t\right) = \mathbb{E}\left[\mathbb{E}\left(\langle f \rangle_{t+1} \mid \widehat{\mathcal{F}}_t\right) \middle| \mathcal{F}_t\right] = \mathbb{E}\left(\sum_{i=1}^{N} \widehat{w}_t^i K f\left(\widehat{\xi}_t^i\right) \middle| \mathcal{F}_t\right) = \sum_{i=1}^{N} w_t^i K f(\xi_t^i). \tag{2.7}$$

Hence, continuing to iterate the tower property by alternatively conditioning on $\widehat{F}_s$ then F_s, for s decreasing from $t-1$ to 0, we have $\mathbb{E}\left(\langle f \rangle_{t+1} \mid \mathcal{F}_0\right) = \sum_{i=1}^{N} w_0^i K^{t+1} f(\xi_0^i)$. Therefore, by the law of total expectation

$$\mathbb{E}\left(\langle f \rangle_{t+1}\right) = \mathbb{E}\left(\sum_{i=1}^{N} w_0^i K^{t+1} f(\xi_0^i)\right). \qquad \boxtimes$$

Second, the ergodic property gives that the weighted ensemble estimate converges to the steady-state estimate as given in the following proposition.

Proposition 2.3.2. *As $T \to \infty$ the estimate θ_T converges almost surely to $\int f d\mu$.*

A formal proof of Proposition 2.3.2 can be found in [22] and follows from the unbiased property, the fact that the weighted ensemble variance, Proposition 2.3.3, scales as $\mathcal{O}(1/T)$, and the Borel-Cantelli lemma.

Finally, to give the variance formula of the weighted ensemble steady-state estimate, θ_T, we define some necessary notation. First, a bin distribution

$$\eta_t^u(x) = \sum_{i=1}^{N} p(\xi_t^i \mid u) \delta(x - \xi_t^t)$$

where $\delta(x)$ is Dirac delta function. Second, an evolution function

$$h_{t,T}(\xi) = \sum_{s=0}^{T-t-1} K^s f(\xi). \tag{2.8}$$

Finally, variance with respect to probability measure η

$$\text{Var}_\eta(g) = \eta(g^2) - \eta(g)^2 \quad \text{where} \quad \eta(g) = \int g d\eta$$

where g is a bounded, measurable function. Since $K(\xi, \cdot)$ is the distribution of a particle evolved from ξ then we can define

$$\text{Var}_{K(\xi,\cdot)} g = Kg^2(\xi) - (Kg)^2(\xi).$$

The next theorem gives the exact variance for the weighted ensemble steady-state estimator (2.2).

Proposition 2.3.3. *The weighted ensemble stead-state estimator variance, for $T > 0$, is given by*

$$\text{Var}\left(\theta_T\right) = \frac{1}{T^2} V_0 + \frac{1}{T^2} \sum_{t=0}^{T-2} V_t^S + \frac{1}{T^2} \sum_{t=0}^{T-2} V_t^M$$

where

$$V_0 = \text{Var}\left(\sum_{i=1}^{N} w_0^i h_{0,T}(\xi_0^i) \right)$$

is the variance from the initialization, and

$$V_t^S = \mathbb{E}\left[\sum_u \frac{w_t(u)^2}{N_t(u)} \text{Var}_{\eta_t^u}(K h_{t+1,T}) \right] \quad \text{and} \quad V_t^M = \mathbb{E}\left[\sum_u \frac{w_t(u)^2}{N_t(u)} \eta_t^u(\text{Var}_K(h_{t+1,T})) \right]$$

are the selection and mutation variances on time t, respectively.

The proof of Proposition 2.3.3 is involved so we simply provide an outline here and leave the intricate details to [22]. First, we define two Doob martingales

$$M_t = \mathbb{E}\left(\theta_T \middle| \mathcal{F}_t\right) \quad \text{and} \quad \widehat{M}_t = \mathbb{E}\left(\theta_T \middle| \widehat{\mathcal{F}}_t\right)$$

using the filtrations in equations (2.3) and (2.4), respectively. Then the variance of the weighted ensemble estimate can be decomposed as

$$\text{Var}\left(\theta_T\right) = \text{Var}(M_0) + \sum_{t=0}^{T-2} \mathbb{E}\left[\left(\widehat{M}_t - M_t\right)^2\right] + \sum_{t=0}^{T-2} \mathbb{E}\left[\left(M_{t+1} - \widehat{M}_t\right)^2\right] \tag{2.9}$$

where the three terms on the right hand side correspond to the initialization variance, selection variance, and mutation variance, respectively. This variance decomposition results from showing the martingale differences are all uncorrelated and then writing M_{T-1} as a telescoping sum of the martingale differences. Note, for $m \in \mathbb{N}$, using equation (2.7) and iteratively applying the tower property by conditioning on $\mathcal{F}_s$, for s decreasing from $t+m-1$ to t, we have

$$\mathbb{E}\left(\langle f\rangle_{t+m} \,|\, \mathcal{F}_t\right) = \mathbb{E}\left(\cdots\left(\mathbb{E}\left(\langle f\rangle_{t+m} \,|\, \mathcal{F}_{t+m-1}\right) \,|\, \mathcal{F}_{t+m-2}\right) \,|\, \cdots\right) |\, \mathcal{F}_t\right) = \sum_{i=1}^{N} w_t^i K^m f(\xi_t^i). \tag{2.10}$$

Since w_s^i and ξ_s^i are $\mathcal{F}_t$ measurable for $s \leq t$ and $i = 1, 2, \ldots, N$, then using equation (2.10) write

$$TM_t = \sum_{s=0}^{t}\sum_{i=1}^{N} w_s^i f(\xi_s^t) + \sum_{s=1}^{T-t-1}\sum_{i=1}^{N} w_t^i K^s f(\xi_t^i) = \sum_{s=0}^{t}\sum_{i=1}^{N} w_s^i f(\xi_s^t) + \sum_{i=1}^{N} w_t^i K h_{t+1,T}(\xi_t^i). \tag{2.11}$$

Similarly, using equation (2.5) and (2.6) with iterative application of the tower property and alternatively conditioning on $\widehat{\mathcal{F}}_s$ then $\mathcal{F}_s$ we have

$$T\widehat{M}_t = \sum_{s=0}^{t}\sum_{i=1}^{N} w_s^i f(\xi_s^t) + \sum_{i=1}^{N} \widehat{w}_t^i K h_{t+1,T}\left(\widehat{\xi}_t^i\right). \tag{2.12}$$

Finally, by substituting equations (2.11) and (2.12) into (2.9) and simplifying gives the desired variance formula.

2.3.2 Exact Calculations for 1d Overdamped Langevin Dynamics

We now consider a particular Markov chain $(X_t)_{t\geq 0}$, on $S \subseteq \mathbb{R}$, which is governed by overdamped Langevin dynamics through the stochastic differential equation

$$dX_t = -V'(X_t)dt + \sqrt{2\beta^{-1}}dW_t \tag{2.13}$$

where $V : \mathbb{R} \to \mathbb{R}$ is a smooth function and $\beta \in \mathbb{R}$ is a positive parameter. Overdamped Langevin dynamics is often used as a process for simulating certain molecular dynamics. The function V represents a potential that the molecule is moving through and β is an inverse temperature parameter of the molecule. We are interested in the Markov chain $(X_t)_{t\geq 0}$ transitioning from region A to disjoint region B where, upon reaching B, we restart X_t in A according to an initial or recycle distribution ρ.

Recall that we assume the Markov kernel K is a Δt skeleton of the continuous time overdamped Langevin dynamics by evaluating the Markov process at resampling times Δt [21]. Consequently, when X_t reaches B we enforce that it remains there until the end of the current Δt resampling time interval at which point it is recycled according to ρ. By recycling only at the end of a Δt resampling time, we ensure that no trajectories reaching B are missed as in practice the weighted ensemble estimator (2.2) is updated at the end of each Δt interval. Now we assume, for analysis, the resampling time limit of $\Delta t \to 0$. That is, we observe the Markov chain, X_t, at all time $t \geq 0$. Although, in practice, a positive resampling time is required and we must use some time discretization to approximate the stochastic differential equation and thus approximate the Markov kernel K.

In the resampling time limit, $\Delta t \to 0$, we are interested in the infinitesimal generator, L, of the Markov process instead of a Δt skeleton kernel K. The infinitesimal generator is defined, for

general Markov chain $(Y_t)_{t\geq 0}$, by

$$Lg(x) = \lim_{\Delta t \to 0^+} \frac{\mathbb{E}(g(Y_{\Delta t})|Y_0 = x) - g(x)}{\Delta t} \tag{2.14}$$

where g is a function such that the limit exists and is finite [23] . For 1d overdamped Langevin dynamics the infinitesimal generator is

$$Lg(x) = -V'(x)g'(x) + \beta^{-1}g''(x) \tag{2.15}$$

for twice continuously differentiable functions $g : \mathbb{R} \to \mathbb{R}$ [23]. Another useful operator is the Fokker-Planck operator, L^*, which is the adjoint of the infinitesimal generator [24]. For 1d overdamped Langevin dynamics, the Fokker-Planck operator is

$$L^*g = (V'g)' + \beta^{-1}g'' \tag{2.16}$$

for twice continuously differentiable functions $g : \mathbb{R} \to \mathbb{R}$ [24].

Now recall $\tau_B = \inf\{t \geq 0 : X_t \in B\}$ is the first time X_t is in the target set B. Define $u(x) \coloneqq \mathbb{E}(\tau_B \,|\, X_0 = x) = \mathbb{E}^x(\tau_B)$ to be the mean first passage time for X_t to reach B starting at $x \in S$. We are interested in the mean first passage time when X_t starts in A according to the initial distribution ρ given by

$$\mathbb{E}(\tau_B \,|\, X_0 \sim \rho) = \mathbb{E}^\rho(\tau_B) = \int_A u(x)\rho(x)dx.$$

We now derive the mean first passage time $u(x)$ when X_t has state space $S = (a, b)$ where a is a reflecting boundary and define $B = \{b\}$. The mean first passage time $u(x)$ satisfies [23]

$$\begin{cases} Lu = -1 & \text{on } (a, b) \\ u(b) = 0 \quad u'(a) = 0. \end{cases} \tag{2.17}$$

Using equation (2.15) and integrating factor $e^{-\beta V(x)}$ then equation (2.17) gives

$$u'(x) = -\beta e^{\beta V(x)} \int_a^x e^{-\beta V(y)} dy + C e^{\beta V(x)}.$$

The boundary condition $u'(a) = 0$ implies $C = 0$. Hence, integrating $u'(x)$ and applying the second boundary condition $u(b) = 0$ gives

$$u(x) = \beta \int_x^b \left(e^{\beta V(z)} \int_a^z e^{-\beta V(y)} dy \right) dz. \tag{2.18}$$

Next, we derive the steady-state distribution $\mu(x)$ for $(X_t)_{t \geq 0}$ on state space $S = (a, b)$ where a is a reflecting boundary, $B = \{b\}$, and when X_t reaches B it is immediately recycled according to ρ. The steady-state distribution can be readily calculated using the Fokker-Planck operator, L^*, as it satisfies [25]

$$\begin{cases} L^* \mu = -\rho / \mathbb{E}^\rho(\tau_B) & \text{on } (a, b) \\ \mu(b) = 0 & \int_a^b \mu(x) dx = 1. \end{cases} \tag{2.19}$$

Applying equation (2.16) to equation (2.19) and integrating gives

$$V' \mu(x) + \beta^{-1} \mu'(x) = -\frac{1}{\mathbb{E}^\rho(\tau_B)} \int_a^x \rho(y) dy + C.$$

The boundary condition $\mu(b) = 0$ and Hill relation $1/\mathbb{E}^\rho(\tau_B) = -\beta^{-1} \mu'(b)$ imply that $C = 0$. Then using integrating factor $e^{\beta V(x)}$, the condition $\mu(b) = 0$, and $\int_a^b \mu(x) = 1$ gives

$$\mu(x) = \frac{\int_x^b e^{\beta(V(z) - V(x))} \left(\int_a^z \rho(y) dy \right) dz}{\int_a^b \left(\int_x^b e^{\beta(V(z) - V(x))} \left(\int_a^z \rho(y) dy \right) dz \right) dx}. \tag{2.20}$$

Finally, we derive explicit and computable formulas for a flux discrepancy function h and variance function v^2 which are useful in understanding the weighted ensemble variance of continuous time Markov chains. First, for positive resampling time Δt and K a Δt skeleton of the underlying

Markov process, we use $h_{t,T}$ given in (2.8) to define function $h_{\Delta t} : S \to \mathbb{R}$ by

$$h_{\Delta t}(\xi) \coloneqq \lim_{T\to\infty} \left[h_{t,T}(\xi) - (T-t) \int f d\mu \right] = \sum_{s=0}^{\infty} \left[\mathbb{E}^{\xi}(f(X_{s\Delta t})) - \mathbb{E}^{\mu}(f(X_{s\Delta t})) \right]$$

Note, $h_{\Delta t}(\xi)$ exists by the assumed geometric ergodicity of K and can be understood as the mean discrepancy between $f(X_t)$ starting from $X_0 = \xi$ versus starting from steady-state $X_0 \sim \mu$ [26]. Now the flux discrepancy function is given in the resampling time limit, $\Delta t \to 0$, by

$$h(\xi) = \lim_{\Delta t\to 0} (\Delta t) h_{\Delta t}(\xi) = \int_0^{\infty} \left[\mathbb{E}^{\xi}(f(X_t)) - \mathbb{E}^{\mu}(f(X_t)) \right] dt.$$

The flux discrepancy function, h, can be interpreted in the same manner as $h_{\Delta t}$, but we now have observed the Markov chain, X_t, at all time $t \geq 0$ instead of at discrete Δt intervals.

When $f = \mathbb{1}_B$, as used for estimating the steady-state flux into B, then $h(\xi)$ produces the discrepancy between the steady-state flux into B and the flux into B starting from ξ. Also, for $f = \mathbb{1}_B$, we can write

$$h(\xi) = \lim_{T\to\infty} \left(\mathbb{E}^{\xi}(C_T) - \mathbb{E}^{\mu}(C_T) \right) \tag{2.21}$$

where C_s is a random variable giving the number of times the Markov chain, X_t, crosses into B over the interval $[0, s)$. Hence, $h(\xi)$ can alternatively be understood as the discrepancy in the number of crossing into B starting from ξ versus starting from μ. Applying the infinitesimal generator (2.14) to h in (2.21), and using the fact that $\mathbb{E}^x(C_{t+\Delta t}) = \mathbb{E}^x(C_{\Delta t} + \mathbb{E}^{X_{\Delta t}}(C_t))$ along with the Hill relation, shows that h satisfies $Lh = 1/\mathbb{E}^{\rho}(\tau_B)$ outside of B. Since, upon reaching B, the Markov chain X_t is recycled according to ρ then starting in B and starting at ρ should be equivalent except we have one more crossing to count when starting in B. Thus, the flux discrepancy function satisfies $h(x) = \rho(h) + 1$ in B. Lastly, note that h is invariant under μ since

$$\mu(h) = \lim_{T\to\infty} \left(\int \mathbb{E}^{\xi}(C_T) d\mu - \mathbb{E}^{\mu}(C_T) \right) = \lim_{T\to\infty} \left(\mathbb{E}^{\mu}(C_T) - \mathbb{E}^{\mu}(C_T) \right) = 0.$$

Hence, when estimating the steady-state flux into B from ρ the flux discrepancy function satisfies

$$\begin{cases} Lh = 1/\mathbb{E}^{\rho}(\tau_B) & \text{on } (a,b) \\ h(b) = \rho(h) + 1 & \mu(h) = 0 \end{cases}$$

which has the solution $h(x) = (\mathbb{E}^{\mu}(\tau_B) - \mathbb{E}^{x}(\tau_B))/\mathbb{E}^{\rho}(\tau_B)$ [27]. For 1d overdamped Langevin dynamics we can use equations (2.18) and (2.20) to write

$$h(x) = \frac{1}{\mathbb{E}^{\rho}(\tau_B)} \left(\beta \int_a^x \int_a^z e^{\beta(V(z)-V(y))} dy dz + \mathbb{E}^{\mu}(\tau_B) - u(a) \right). \tag{2.22}$$

Now for the variance function v^2 we again start by assuming a positive resampling time Δt and letting K be a Δt skeleton of the underlying Markov process. We define function $v^2_{\Delta t} : S \to \mathbb{R}$ by

$$v^2_{\Delta t}(\xi) \coloneqq \text{Var}_{K(\xi,\cdot)}(h) = Kh^2(\xi) - (Kh(\xi))^2 = \mathbb{E}^{\xi}(h^2(X_{\Delta t})) - [\mathbb{E}^{\xi}(h(X_{\Delta t}))]^2.$$

Then the variance function, v^2, is given in the resampling time limit, $\Delta t \to 0$, by

$$\begin{aligned} v^2(\xi) \coloneqq \lim_{\Delta t \to 0} \frac{v^2_{\Delta t}(\xi)}{\Delta t} &= \lim_{\Delta t \to 0} \left(\frac{\mathbb{E}^{\xi}(h^2(X_{\Delta t})) - h^2(\xi)}{\Delta t} - \left(\mathbb{E}^{\xi}(h(X_{\Delta t})) + h(\xi)\right) \frac{\mathbb{E}^{\xi}(h(X_{\Delta t})) - h(\xi)}{\Delta t} \right) \\ &= Lh^2(\xi) - 2h(\xi)Lh(\xi) \end{aligned}$$

where we used the definition of the infinitesimal generator (2.14). Hence, using equations (2.15) and (2.22) the variance function for 1d overdamped Langevin dynamics is

$$v^2(\xi) = 2\beta^{-1} \left| \frac{d}{d\xi} h(\xi) \right|^2 = \frac{2\beta}{\left(\mathbb{E}^{\rho}(\tau_B)\right)^2} \left(\int_a^{\xi} e^{\beta(V(\xi)-V(y))} dy \right)^2. \tag{2.23}$$

We can note that the variance function is also given by $v^2 = \lim_{\Delta t \to 0} \Delta t \left(\lim_{T \to \infty} \text{Var}_K(h_{t+1,T}) \right)$ and so under certain assumptions, discussed further in Section 3.2.2, we will approximate the mutation variance, in Proposition 2.3.3, using the flux discrepancy function, h, and variance function, v^2.

Chapter 3

Stability

In this chapter we discuss two types of stability: mean first passage time bias, or simply bias, stability and estimator variance, or simply variance, stability. By bias stability, we mean that perturbing the Markov chain initial distribution does not significantly change the resulting mean first passage time. Instead for variance stability, we mean that importance sampling the Markov chain initial distribution does not drastically decrease the variance of the desired estimator.

Here we analyze the bias and variance stability, with respect to the Markov chain initial distribution, for the weighted ensemble algorithm. In particular, we develop conditions such that the mean first passage time, from weighted ensemble, is stable and unstable with respect to perturbations in the initial distribution. Furthermore, we develop conditions such that initial condition importance sampling appreciably reduces the variance of the weighted ensemble estimator. These weighted ensemble stability results are contrasted against similar stability results for adaptive multilevel splitting, which is another rare event simulation technique used in molecular dynamics to estimate mean first passage times. It has been studied that adaptive multilevel splitting is often unstable with respect to the Markov chain initial distribution [16, 28, 29]. One example is that a certain importance sampling on the initial distribution is required for adaptive multilevel splitting to maintain useful variance bounds on the desired estimator [16].

Throughout the following sections we will make use of Laplace's Method, which provides an, asymptotically equal, approximation to integrals of functions of the form $h(x)e^{\beta g(x)}$ for large β [30]. We motivate Laplace's Method by deriving the approximation of $\int_a^b e^{\beta g(x)}$ for large β when $g(x)$ obtains a global maximum at $x_{\max} \in (a, b)$. As $x_{\max}$ is an interior point on the interval (a, b) it must be a critical point of $g(x)$ and $g''(x) < 0$. Then the Taylor series of $g(x)$ about $x_{\max}$ is given by

$$g(x) = g(x_{\max}) - \frac{1}{2} \left| g''(x_{\max}) \right| (x - x_{\max})^2 + \mathcal{O}((x - x_{\max})^3).$$

So we can approximate

$$\int_a^b e^{\beta g(x)dx} \approx e^{\beta g(x_{\max})} \int_a^b \exp\left(-\frac{1}{2}\beta|g''(x_{\max})|(x-x_{\max})^2\right) dx.$$

Now, $\exp\left(-\frac{1}{2}\beta|g''(x_{\max})|(x-x_{\max})^2\right)$ is a Gaussian function which, for large β, decays rapidly away from $x_{\max}$. Therefore

$$\int_a^b e^{\beta g(x)dx} \approx e^{\beta g(x_{\max})} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\beta|g''(x_{\max})|(x-x_{\max})^2\right) dx = \sqrt{\frac{2\pi}{\beta|g''(x_{\max})|}} e^{\beta g(x_{\max})}$$

where the final equality holds from the normalization of a Gaussian distribution. Note the above approximations are exact in the limit of $\beta \to \infty$.

Next, we give the general form for Laplace's Method. On interval (a, b) let $h(x)$ be a positive function and $g(x)$ be a twice differentiable function which obtains a global max at $x_{\max}$. We consider three cases depending on the location of $x_{\max}$ and properties of $g(x)$ at $x_{\max}$. First, let $x_{\max} \in (a, b)$ then $x_{\max}$ is a critical point of $g(x)$ and $g''(x_{\max}) < 0$. Then

$$\int_a^b h(x)e^{\beta g(x)}dx = \sqrt{\frac{2\pi}{\beta|g''(x_{\max})|}} h(x_{\max})e^{\beta g(x_{\max})}(1 + \mathcal{O}(\beta^{-1})) \tag{3.1}$$

in the large β limit [30]. So, for large β, we can approximate

$$\int_a^b h(x)e^{\beta g(x)}dx \sim \sqrt{\frac{2\pi}{\beta|g''(x_{\max})|}} h(x_{\max})e^{\beta g(x_{\max})}. \tag{3.2}$$

Second, let $x_{\max}$ be a limit of integration and a critical point of $g(x)$. Then, in the Gaussian approximation of $e^{\beta g(x)}$ we will only integrate over half of a Gaussian. So, the resulting Laplace's Method approximations are simply $\frac{1}{2}$ of (3.1) and (3.2). Third, let $x_{\max}$ be a limit of integration but not a critical point of $g(x)$. Now we Taylor expand $g(x)$ to first order instead of second order

and approximate, for large β,

$$\int_a^b h(x)e^{\beta g(x)}dx \sim \frac{h(x_{\max})}{\beta|g'(x_{\max})|}e^{\beta g(x_{\max})}. \tag{3.3}$$

3.1 Bias Stability in the Mean First Passage Time

We now analyze the mean first passage time, from an initial distribution, of a Markov chain governed by overdamped Langevin dynamics. Overdamped Langevin dynamics (2.13) models a molecule with constant temperature, β^{-1}, moving in a potential energy field, $V(x)$. We aim to understand the stability of the mean first passage time from A to B, with respect to perturbations of the Markov chain initial distribution, in the large β, or equivalently small temperature, limit. We find that if A does not contain significant internal potential energy barriers, at least as large as the largest forward barrier between A and B, then the mean first passage time is stable to changes in the initial distribution. On the other hand, if A contains a significant internal barrier then the mean first passage time can become unstable to changes in the initial distribution. These two ideas are formulated and verified by analyzing a one dimensional Markov chain.

Assume Markov chain $(X_t)_{t\geq 0}$ is governed by overdamped Langevin dynamics (2.13) on state space $S = (a, b)$ where a is a reflecting boundary and b is an absorbing boundary. Let the initial distribution $X_0 \sim \rho$ where $\text{supp}(\rho) = A \subseteq (a, b)$ and when X_t reaches $B = \{b\}$ it is immediately recycled, or restarted, in A according to ρ. Define a related process $(\widetilde{X}_t)_{t\geq 0}$ which obeys the same dynamics as $(X_t)_{t\geq 0}$ except the initial and recycle distribution are given by $\widetilde{\rho}$, with $\text{supp}(\widetilde{\rho}) = \text{supp}(\rho)$, which satisfies

$$|\rho(x) - \widetilde{\rho}(x)| \leq \epsilon$$

for $\epsilon > 0$ a perturbation. Now define

$$\Delta V \coloneqq \sup_{x\in(a,b)} \inf_{y\in(a,x)} (V(x) - V(y)) \tag{3.4}$$

which is the largest forward potential energy barrier of $V(x)$ on (a,b) and let $\Delta V = V(x^*) - V(y^*)$ for $x^* \in (a,b]$ and $y^* \in [a,x^*)$.

The following proposition shows that the mean first passage time is stable so long as ΔV is not in the interior of A; that is, $\text{supp}(\rho) \subseteq (a,x^*)$.

Proposition 3.1.1. *If* $\text{supp}(\rho) = \text{supp}(\widetilde{\rho}) \subseteq (a,x^*)$ *then* $|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)| \sim 0$ *in the large* β *limit.*

Proof. For $x^* \in (a,b]$ and $y^* \in [a,x^*)$ assume $\Delta V = V(x^*) - V(y^*)$ is the largest forward potential energy barrier. Let ρ and $\widetilde{\rho}$ be any two densities satisfying $\text{supp}(\rho) = \text{supp}(\widetilde{\rho}) \subseteq (a,x^*)$. Note, for $x \in (a,x^*)$ and large β, using Laplace's Method and equation (2.18)

$$u(x) = \beta \int_x^b \int_a^z e^{\beta(V(z)-V(y))} dy dz \sim C_{y^*} C_{x^*} e^{\beta(V(x^*)-V(y^*))} = C_{y^*} C_{x^*} e^{\beta \Delta V}$$

where C_{y^*} and C_{x^*} are two constants from Laplace's Method dependent on y^* and x^*, respectively. Therefore, in the large β limit

$$\left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right| = \left| \int_a^{x^*} u(x) \left(\rho(x) - \widetilde{\rho}(x)\right) dx \right| \sim C_{y^*} C_{x^*} e^{\beta \Delta V} \left| \int_a^{x^*} (\rho(x) - \widetilde{\rho}(x)) dx \right| = 0$$

since $\int_a^{x_{\max}} \rho(x) dx = \int_a^{x_{\max}} \widetilde{\rho}(x) dx = 1$. ⊠

Proposition 3.1.1 gives that the mean first passage time from A to B is stable so long as the largest forward barrier ΔV is not in the interior of A. As $A = \text{supp}(\rho) \subseteq (a,x^*)$ then the largest forward barrier may be on the boundary of A but is not completely contained within A. In fact, the requirement that $|\rho(x) - \widetilde{\rho}(x)| < \epsilon$ is not necessary nor assumed for Proposition 3.1.1 and we only require that the initial density has support left of x^*, the maximum in the largest forward potential barrier. Proposition 3.1.1 is supported by, and follows from, Eyring-Kramers Law which states that the mean first passage time over a potential barrier scales as a constant time $e^{\beta \Delta V}$ where ΔV is the size of the barrier to be surpassed [31–33]. That is $u(x)$ scales as a constant for all starting

points left of x^* since the main potential barrier still must be surpassed. Hence, the choice of initial distribution, ρ, with support left x^*, is inconsequential.

Consider the following example where $S = (0, 1)$, $\rho = \epsilon \mathbb{1}_{(0,\frac{1}{4})}(x) + (4 - \epsilon)\mathbb{1}_{[\frac{1}{4},\frac{1}{2})}$,

$$\widetilde{\rho}(x) = 2\epsilon \mathbb{1}_{(0,\frac{1}{4})}(x) + (4 - 2\epsilon)\mathbb{1}_{[\frac{1}{4},\frac{1}{2})}(x),$$

and we have a single well, single barrier sinusoidal potential $V(x) = -\sin(2\pi x)$ which is shown in Figure 3.1. Note that the maximum forward potential barrier to be surpassed is $\Delta V = 2$. Numerical calculations for the relative difference in the mean first passage times, at various values of β, are given in Table 3.1. We note that for large β the relative difference is zero as the mean first passage times are equal, which is as expected from Proposition 3.1.1. Note that we use the relative

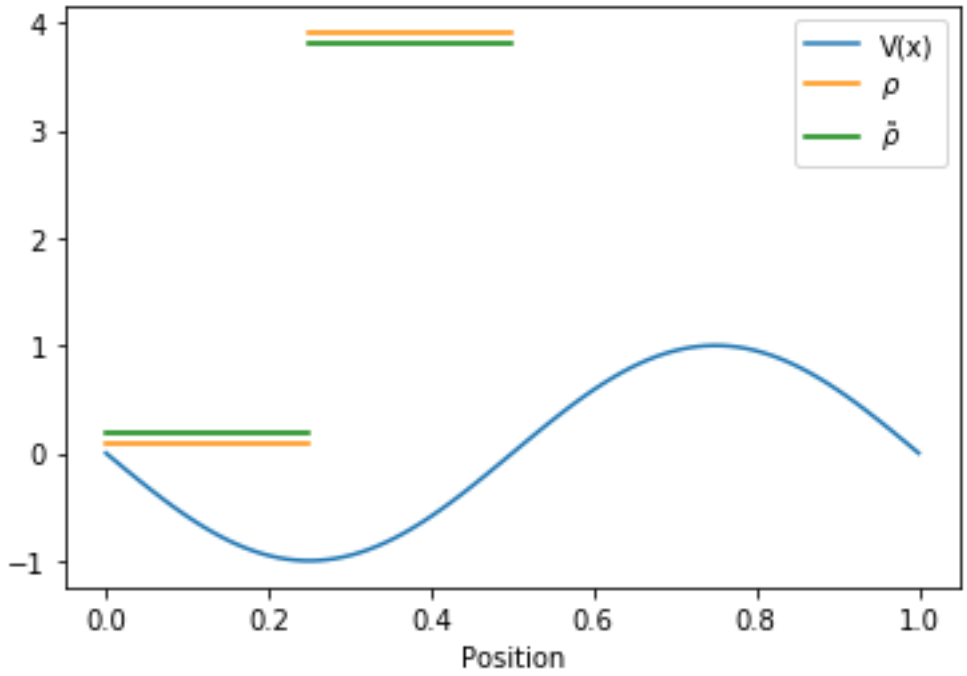


Figure 3.1: Sinusoidal potential with two initial distributions.

difference in the mean first passage since, for small β, it is possible for there to be a seemingly significant absolute difference in the mean first passage times especially when A contains an internal barrier. Such a difference is actually inconsequential compared to the order of magnitude of

Table 3.1: Numerical calculations for the relative difference in the mean first passage times, for various β, when using a sinusoidal potential and the maximum difference in the recycle distributions is $\epsilon = 10^{-2}$.

ϵ	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
β	1	5	15	100	150
$\frac{\left\|\mathbb{E}^{\rho}(\tau_B)-\mathbb{E}^{\widetilde{\rho}}(\tau_B)\right\|}{\mathbb{E}^{\rho}(\tau_B)}$	1.2×10^{-4}	4.6×10^{-7}	3.4×10^{-12}	1.2×10^{-16}	0

each mean first passage time. Define ΔV to be the largest potential barrier between A and B and ΔV_A to be the largest potential barrier in A. Then $\left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right|$ will, at a maximum, scale as $e^{\beta \Delta V_A}$ where as $\mathbb{E}^{\rho}(\tau_B)$ and $\mathbb{E}^{\widetilde{\rho}}(\tau_B)$ both will scale as $e^{\beta \Delta V}$. Since $\Delta V_A < \Delta V$ then $\mathbb{E}^{\rho}(\tau_B)$ and $\mathbb{E}^{\widetilde{\rho}}(\tau_B)$ will grow to exponential order faster than $\left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right|$.

Now consider when the support initial distribution includes the maximum of the largest forward barrier. That is the initial set A contains a significant internal barrier. In such a case, the mean first passage time can be unstable with respect to the initial distribution. This is reasonable as the mean first passage time $u(x)$ can be significantly larger for $x < x^*$ than for $x \geq x^*$ as $x \geq x^*$ skips the necessity to overcome the largest potential barrier. Applying Eyring-Kramers we expect that $|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)|$ will, at a maximum, scale as a constant times $\epsilon e^{\beta \Delta V}$ when $|\widetilde{\rho}(x) - \rho(x)| \leq \epsilon$. We develop a bound on the absolute difference $|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)|$, which show consistency with Eyring-Kramers Law, and provide examples to display the tightness of each bound.

Recall the largest forward potential energy barrier is $\Delta V = V(x^*) - V(y^*)$ for $x^* \in (a, b]$ and $y^* \in [a, x^*)$. Assume that x^* does not occur on the boundary b and $\text{supp}(\rho) = (x_1, x_2) \subseteq (a, b)$ such that $x^* \in (x_1, x_2)$. Then, for $x \in (x_1, x^*)$ and large β, using equation (2.18) and Laplace's Method

$$u(x) = \beta \int_x^b \int_a^z e^{\beta(V(z)-V(y))} dy dz \sim C_{y^*} \sqrt{\frac{2\pi\beta}{|V''(x^*)|}} e^{\beta(V(x^*)-V(y^*))} = C_{y^*} \sqrt{\frac{2\pi\beta}{|V''(x^*)|}} e^{\beta \Delta V}$$

where

$$C_{y^*} = \begin{cases} \sqrt{\frac{2\pi}{\beta|V''(y^*)|}} & y^* \in (a,b) \\ \sqrt{\frac{\pi}{2\beta|V''(y^*)|}} & y^* = a \text{ and } V'(y^*) = 0 \\ \frac{1}{\beta|V'(y^*)|} & y^* = a \text{ and } V'(y^*) \neq 0. \end{cases}$$

Therefore

$$\begin{aligned} \left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right| &= \left| \int_{x_1}^{x_2} u(x)\rho(x)dx - \int_{x_1}^{x_2} u(x)\widetilde{\rho}(x)dx \right| \\ &\sim C_{y^*}\sqrt{\frac{2\pi\beta}{|V''(x^*)|}}e^{\beta\Delta V}\left| \int_{x_1}^{x^*} (\rho(x) - \tilde{\rho}(x))dx \right| \\ &\leq \epsilon(x^* - x_1)C_{y^*}\sqrt{\frac{2\pi\beta}{|V''(x^*)|}}e^{\beta\Delta V}. \end{aligned} \tag{3.5}$$

As expected from Eyring-Kramers Law, the bound on the difference in the mean first passage times scales as a constant times $\epsilon e^{\beta\Delta V}$.

Consider the following example where $S = (0,1)$,

$$\rho = \epsilon\mathbb{1}_{\left(0,\frac{1}{4}\right)}(x) + (4-\epsilon)\mathbb{1}_{\left[\frac{1}{4},\frac{1}{2}\right)}, \qquad \widetilde{\rho}(x) = 2\epsilon\mathbb{1}_{\left(0,\frac{1}{4}\right)}(x) + (4-2\epsilon)\mathbb{1}_{\left[\frac{1}{4},\frac{1}{2}\right)}(x),$$

and we have a single barrier Gaussian potential, which is shown in Figure 3.2,

$$V(x) = 4\exp\left(-\frac{1}{2}\left(\frac{x-\frac{1}{4}}{\frac{1}{8}}\right)^2\right).$$

Note, the maximum potential barrier to be surpassed is $\Delta V = 4(1-e^{-2}) \approx 3.46$, $C_{y^*} = 1/\beta|V'(0)|$, and $x^* = 1/4$. Thus the difference in the mean first passage times (3.5) is bounded by

$$\left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right| \lesssim \frac{\epsilon}{4|V'(0)|}\sqrt{\frac{2\pi}{\beta\left|V''\left(\frac{1}{4}\right)\right|}}e^{\beta\Delta V}.$$

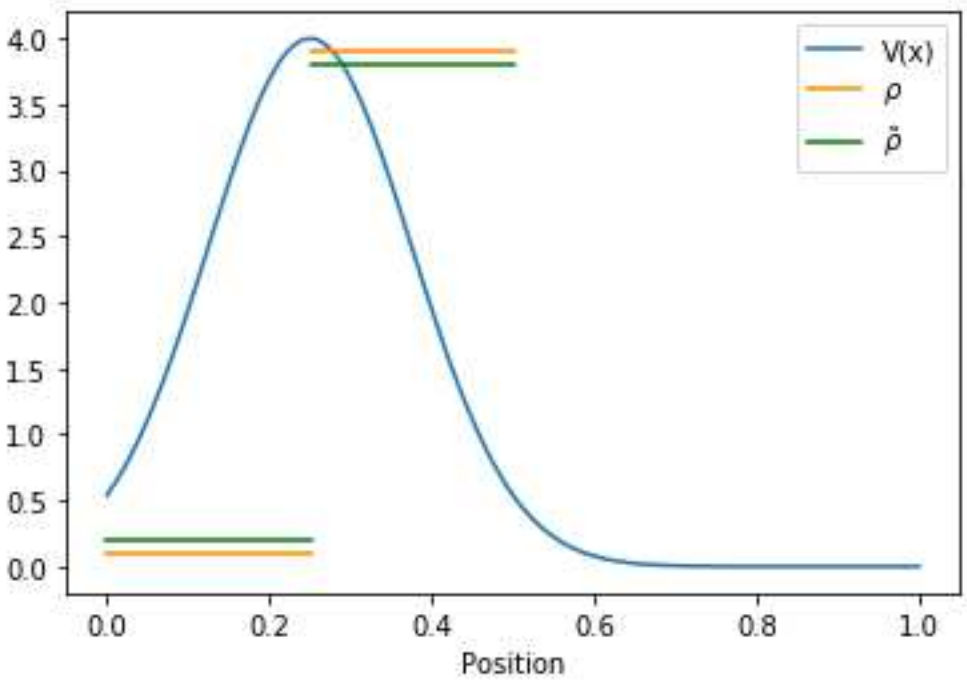


Figure 3.2: Gaussian potential with two recycle distributions.

A numerical calculation for the difference in the mean first passage times, at various values of β, along with the bound (3.5) and Eyring-Kramers Law bound $\epsilon e^{\beta \Delta V}$ is given in Table 3.2.

Table 3.2: Numerical calculations, for various β, of the relative difference in the mean first passage time, the absolute difference in the mean first passage time, Eyring-Kramers Law bound, and bound (3.5) when using a Gaussian potential and the maximum difference in the recycle distributions is $\epsilon = 10^{-2}$.

ϵ	10^{-2}	10^{-2}	10^{-2}	10^{-2}
β	5	15	100	150
$\frac{\left\|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\tilde{\rho}}(\tau_B)\right\|}{\mathbb{E}^{\rho}(\tau_B)} \times 100\%$	4.7%	8.3%	19.7%	23.2%
$\left\|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\tilde{\rho}}(\tau_B)\right\|$	5.0×10^{2}	3.5×10^{17}	7.1×10^{144}	7.4×10^{219}
$\epsilon e^{\beta \Delta V}$	3.2×10^{5}	3.4×10^{20}	1.6×10^{148}	2.0×10^{223}
$\frac{\epsilon}{4\lvert V'(0)\rvert}\sqrt{\frac{2\pi}{\beta\lvert V''\left(\frac{1}{4}\right)\rvert}}e^{\beta \Delta V}$	6.55×10^{2}	3.97×10^{17}	7.29×10^{144}	7.56×10^{219}

Table 3.2 shows that for large β we have a significant discrepancy in the mean first passage times even though the initial conditions differ by a perturbation. The relative differences show that the absolute differences $\left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\tilde{\rho}}(\tau_B)\right|$ are significant in comparison to the original mean first passage $\mathbb{E}^{\rho}(\tau_B)$. So, when A contains the largest forward potential barrier there can be instability in the mean first passage time when changing the initial distribution. We also see from Table 3.2 that the bound (3.5) is reasonable as it differs from the actual absolute difference in the mean first passage times by less than a factor of 1.14 for each case of β.

Consider a second example using the same state space, S, the same recycle densities, ρ and $\tilde{\rho}$, and we have a single barrier, single well double Gaussian potential

$$V(x) = 3\exp\left(-\frac{1}{2}\left(\frac{x-\frac{1}{5}}{\frac{1}{11}}\right)^2\right) - 3\exp\left(-\frac{1}{2}\left(\frac{x-\frac{1}{9}}{\frac{1}{11}}\right)^2\right).$$

Note that the maximum barrier to be surpassed is $\Delta V \approx 3.29$, $y^* \approx 0.061$, $C_{y^*} = \sqrt{\frac{2\pi}{\beta|V''(y^*)|}}$, and

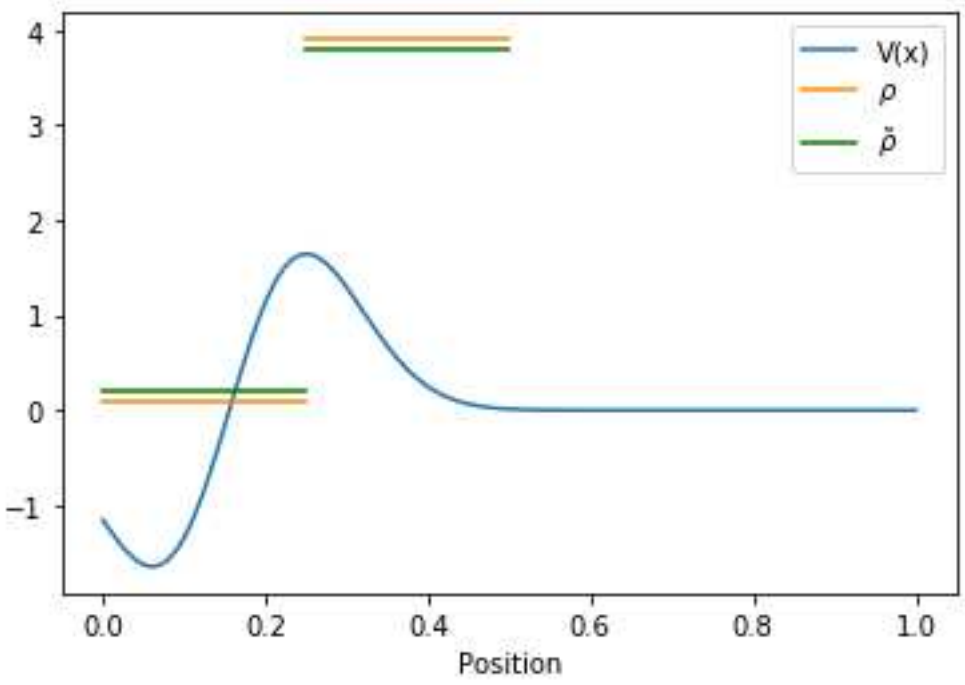


Figure 3.3: Single barrier, single well double Gaussian potential with two recycle distributions.

$x^* \approx \frac{1}{4}$. Thus, the bound on the absolute difference in the mean first passage times (3.5) is

$$\left|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right| \lesssim \frac{\epsilon}{2} \frac{\pi}{\sqrt{|V''(y^*)V''(x^*)|}} e^{\beta \Delta V}.$$

A numerical calculation for the difference in the mean first passage times, at various values of β, along with the bound (3.5) and Eyring-Kramers Law bound $\epsilon e^{\beta \Delta V}$ is given in Table 3.3.

Table 3.3: Numerical calculations, for various β, of the relative difference in the mean first passage time, absolute difference in the mean first passage time, Eyring-Kramers Law bound, and bound (3.5) when using a single barrier, single well double Gaussian potential and the maximum difference in the recycle distributions is $\epsilon = 10^{-2}$.

ϵ	10^{-2}	10^{-2}	10^{-2}	10^{-2}
β	5	15	100	150
$\frac{\left\|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right\|}{\mathbb{E}^{\rho}(\tau_B)} \times 100\%$	4.3%	8.9%	21.2%	24.7%
$\left\|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)\right\|$	5.82×10^2	1.13×10^{17}	3.26×10^{138}	9.08×10^{209}
$\epsilon e^{\beta \Delta V}$	1.4×10^5	2.7×10^{19}	7.7×10^{140}	2.2×10^{212}
$\frac{\epsilon}{2} \frac{\pi}{\sqrt{\lvert V''(y^*)V''(x^*)\rvert}} e^{\beta \Delta V}$	5.95×10^2	1.16×10^{17}	3.29×10^{138}	9.14×10^{209}

From Table 3.3 we again have a significant absolute difference and relative difference in the mean first passage times, for large β, when the initial distributions differ by only a perturbation. This implies that A containing a significant potential barrier can cause instability in the mean first passage time even when A also contains a potential well. Lastly, from Table 3.3, we can note that the bound (3.5) differs, for each β, by less than a factor of 1.03 to the actual absolute difference in the mean first passage times.

Even though Table 3.2 and Table 3.3 highlight examples where the mean first passage time is unstable with respect to the initial distribution, instability is not guaranteed. For instance, so long as $\int_a^{x^*} \rho(x)dx = \int_a^{x^*} \widetilde{\rho}(x)dx$ then we can still apply Proposition 3.1.1 and approximate

$\left|\mathbb{E}^{\rho}(\tau) - \mathbb{E}^{\widetilde{\rho}}(\tau)\right| \sim 0$ in the large β limit. Also, when A contains the largest forward barrier, then by Laplace's Method the relative error in the mean first passage time is

$$\frac{|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)|}{\mathbb{E}^{\rho}(\tau_B)} \sim \frac{\left|\int_{x_1}^{x^*}(\rho(x) - \widetilde{\rho}(x))dx\right|}{\int_{x_1}^{x^*}\rho(x)dx} = \left|1 - \frac{\int_{x_1}^{x^*}\widetilde{\rho}(x)dx}{\int_{x_1}^{x^*}\rho(x)dx}\right|. \tag{3.6}$$

Thus, to have a significant relative difference in the mean first passage times we require that ρ and $\widetilde{\rho}$ have a sufficient relative integral difference over (x_1, x^*). For an example where we can bound the relative mean first passage time, take $\rho \sim U\left(0, \frac{1}{2}\right)$, $\widetilde{\rho}(x) = (2-\epsilon)\mathbb{1}_{\left(0,\frac{1}{4}\right)}(x) + (2+\epsilon)\mathbb{1}_{\left[\frac{1}{4},\frac{1}{2}\right)}(x)$, and $V(x)$ any potential such that $x^* = \frac{1}{4}$ such as the potential in Figure 3.3. Then

$$\frac{|\mathbb{E}^{\rho}(\tau_B) - \mathbb{E}^{\widetilde{\rho}}(\tau_B)|}{\mathbb{E}^{\rho}(\tau_B)} \approx \left|1 - \frac{\int_{x_1}^{x^*}\widetilde{\rho}(x)dx}{\int_{x_1}^{x^*}\rho(x)dx}\right| = \frac{\left|\int_0^{1/4}\epsilon dx\right|}{\int_0^{1/4}2dx} = \frac{\epsilon}{2} < \epsilon.$$

Hence, when A contains a significant internal barrier it is possible, but not guaranteed, for the mean first passage time to be unstable to perturbations in the initial distribution.

Even though the bias stability analysis was for 1d overdamped Langevin dynamics we expect the general stability conditions on A to hold in higher dimensions. That is, when A does or does not contain a significant internal barrier then the mean first passage time is unstable or stable to changes in the initial distribution, respectively. While explicit formulas may not exist for higher dimensions, throughout our analysis we have shown that our results are consistent with and a consequence of Eyring-Kramers Law, which does generalize to higher dimensions. Hence, we expect the developed bias stability requirements on A to generalize to higher dimensions by applying Eyring-Kramers Law.

Recall adaptive multilevel splitting, Section 2.2, is another algorithm commonly used in molecular dynamics to estimate mean first passage times. In the calculation of mean first passage times, adaptive multilevel splitting relies on the committor function $p(x) = \mathbb{P}(\tau_B < \tau_A \,|\, X_0 = x)$, which is the probability a Markov chain X_t reaches B before A from starting at x [29]. It is known that

the committor function, $p(x)$, satisfies [23]

$$\begin{cases} Lp(x) = 0 \;\; x \in S \setminus (A \cup B) \\ p(x) = 0 \;\; x \in \partial A \quad p(x) = 1 \;\; x \in \partial B, \end{cases}$$

which for 1d overdamped Langevin dynamics on (a, b) produces

$$p(x) = \frac{\int_a^x e^{\beta V(y)} dy}{\int_a^b \int_a^z e^{\beta V(y)} dy dz}.$$

Next, we explain how adaptive multilevel splitting estimates a mean first passage time using the committor function. Recall from Section 2.2 that adaptive multilevel splitting relies on the contours of the reaction coordinate, $L_z = \{x : \Phi(x) = z\}$, which we call the level at z. To compute the mean first passage time, $\mathbb{E}(\tau_B)$, we consider the point $z_{\min}$, typically chosen so that $L_{z_{\min}}$ is close to A, and slice trajectories which end in B into components. Starting from A, a trajectory will touch $L_{z_{\min}}$ within time T_1^1 then two events can happen:

1. From $L_{z_{\min}}$ the trajectory goes to A before reaching B, which occurs with a probability $1-p$ and with time T_2^1

2. From $L_{z_{\min}}$ the trajectory goes to B before reaching A, which occurs with probability p and with time T_3.

As the transition from A to B is a rare event it is far more likely that the first case, the trajectory goes to A before reaching B, will happen. Define M to be a random variable giving the number of trials that a trajectory must be started in A until the second case, the trajectory goes to B before reaching A, occurs. For each trial define $T_1^1, T_1^2, \ldots, T_1^M$ and $T_2^1, T_2^2, \ldots, T_2^{M-1}$ to be the first passage times from A to $L_{z_{\min}}$ and from $L_{z_{\min}}$ back to A before reaching B, respectively. Then by the Markov property

$$\tau_B = T_1^M + T_3 + \sum_{i=1}^{M-1} (T_1^i + T_2^i).$$

We can note that $T_1^1, T_1^2, \ldots, T_1^M$ and $T_2^1, T_2^2, \ldots, T_2^{M-1}$ are two collections of identically distributed random variables and M is a geometric random variable, with success probability p, which

is independent from each T_1^i, T_2^i, and T_3. Since $\mathbb{E}(M) = 1/p$ then

$$\begin{aligned}\mathbb{E}(\tau_B) &= \mathbb{E}\left(T_1^M + T_3 + \sum_{i=1}^{M-1}(T_1^i + T_2^i)\right) \\ &= \mathbb{E}(T_1^1 + T_3) + (\mathbb{E}(M) - 1)\,\mathbb{E}(T_1^1 + T_2^1) \\ &= \mathbb{E}(T_1^1 + T_3) + \left(\frac{1}{p} - 1\right)\mathbb{E}(T_1^1 + T_2^1).\end{aligned}$$

Now to estimate the mean first passage time, $\mathbb{E}(\tau_B)$, we require estimates for three quantities of interest. First, $\mathbb{E}(T_1^1 + T_3)$ which is the mean time for a trajectory starting from A to reach $L_{z_{\min}}$ and then go to B before A. Estimating $\mathbb{E}(T_1^1 + T_3)$, which is generally still the mean first passage time corresponding to a rare event transition, can be done through a slight modification of the adaptive multilevel splitting algorithm to where the total transition time is calculated by accounting for all the recrossing into A [29]. Second, we require $\mathbb{E}(T_1^1 + T_2^1)$ which is the mean time for a trajectory starting from A to reach $L_{z_{\min}}$ then return to A before reaching B. As $L_{z_{\min}}$ is typically chosen to be close to A, estimating $\mathbb{E}(T_1^1 + T_2^1)$ can reliable be done with naive Markov chain Monte Carlo simulations. Finally, we require p which is the probability a trajectory starting from $L_{z_{\min}}$ reaches B before A. The probability p, given as $p(z_{\min})$ for 1d overdamped Langevin dynamics, can be directly estimated from the adaptive multilevel splitting algorithm where the initial distribution is chosen to have support in $L_{z_{\min}}$.

So, adaptive multilevel splitting has multiple sources of initialization that can affect the mean first passage time. First, is the initial distribution in A which affects the quantity T_1^1. Second, the choice of $z_{\min}$ and the choice of initial distribution in $L_{z_{\min}}$, which affects T_2^1, T_3, and p. Similar to weighted ensemble, if A contains significant internal barriers then we could expect large changes in $\mathbb{E}(T_1^1)$ under small changes in the initial distribution. Large changes in the mean first passage time could be expected as the first passage time to $L_{z_{\min}}$ of a trajectory needing to cross the large forward barrier in A will be significantly higher than the first passage time to $L_{z_{\min}}$ of a trajectory which was initialized past the barrier. Thus, altering the chance of initializing past a significant barrier can produce large changes in the mean first passage time. Likewise, if

$L_{z_{\min}}$ contains a significant potential barrier then slight changes in the initial distribution, which increase or decrease the chance of sampling before or after the barrier, could produce impactful changes in the mean first passage times $\mathbb{E}(T_2^1)$ and $\mathbb{E}(T_3)$ or produce an impactful change on the probability p. Accordingly, a minor adjustment to $z_{\min}$ may result in $L_{z_{\min}}$ containing significant potential barriers where it did not before although this also depends on the choice for reaction coordinate. Even when A does not contain significant internal barriers the choice of $z_{\min}$ and initial distribution in $L_{z_{\min}}$ may affect the stability of the mean first passage time estimate. Thus, we anticipate that weighted ensemble should have greater mean first passage time bias stability than adaptive multilevel splitting as adaptive multilevel splitting has more factors affecting the mean first passage time, which could make it unstable in cases where weighted ensemble is stable.

3.2 Variance Stability

As stated in Section 2.3 the weighted ensemble estimate, θ_T, converges almost surely to $\int f d\mu$ at $T \to \infty$. So, a main point of open work is in reducing the variance of the weighted ensemble estimate θ_T. Note, for a sufficient number of particles we can optimally choose the particle allocation $N_t(u)$, based on chosen bins B_t, to minimize the mutation variance. Recall that the mutation variance, Proposition 2.3.3, at time t is given by

$$\mathbb{E}\left[\sum_u \frac{w_t(u)^2}{N_t(u)} \eta_t^u(\text{Var}_K(h_{t+1,T}))\right].$$

We want to minimize $\sum_u \frac{w_t(u)^2}{N_t(u)} \eta_t^u(\text{Var}_K(h_{t+1,T}))$ subject to the constraint $\sum_u N_t(u) = N$. Define $C_u^2 = w_t(u)^2 \eta_t^u(\text{Var}_K(h_{t+1,T}))$ and $\boldsymbol{N} = [N_t(u)]_{u \in B_t}$ where B_t are the bins at time t. Then we can form a Lagrange multiplier function

$$\mathcal{L}(\boldsymbol{N}) = \sum_u \frac{C_u^2}{N_t(u)} + \lambda \left(\sum_u N_t(u) - N\right) \quad \text{with gradient} \quad \nabla \mathcal{L}(\boldsymbol{N}) = \left[\lambda - \frac{C_u^2}{N_t(u)^2}\right]_{u \in B_t}.$$

Solving were the gradient is zero gives $N_t(u) = C_u/\sqrt{\lambda}$ for each $u \in B_t$. Applying constraint $\sum_u = N_t(u) = N$ produces $\lambda = \left(\sum_u C_u\right)^2 / N^2$. So the optimal particle allocation is $N_t = NC_u/(\sum_u C_u)$ which produces an optimal mutation variance of

$$\mathbb{E}\left[\sum_u \frac{C_u^2\left(\sum_u C_u\right)}{NC_u}\right] = \frac{1}{N}\mathbb{E}\left[\left(\sum_u C_u\right)^2\right] = \frac{1}{N}\mathbb{E}\left[\left(\sum_u w_t(u)\sqrt{\eta_t^u(\mathrm{Var}_K(h_{t+1,T}))}\right)^2\right]. \quad (3.7)$$

It is also known how to optimally choose the bins, B_t, to minimize the selection variance with a sufficient number of bins [26].

Here we consider a Markov chain $(X_t)_{t\geq 0}$ on state space S which is initialized in $A \subseteq S$ according to a source distribution ρ and upon reaching $B \subset S$ the Markov chain X_t immediately restarts in A according to ρ. The source, ρ, will be termed as both the initial distribution and recycle distribution. We explore the impact, on the weighted ensemble variance, of importance sampling the initial and recycle distribution, ρ, which we refer to as initial condition importance sampling. In particular, during the weighted ensemble selection process we define all particles in B to be a bin u_B. In the subsequent mutation step, each particle in B will be recycled according to ρ and we can reduce the mutation variance from u_B, which we call the recycle variance, by initial condition importance sampling. When the reduction in the recycle variance, from initial condition importance sampling, is impactful on the total variance from the mutation step then we say that weighted ensemble has variance instability. Instead, we say that weighted ensemble has variance stability when initial condition importance sampling is not impactful on the total mutation step variance. Similar to bias stability, Section 3.1, we find that if A does not contain a significant internal potential energy barrier, at least as large as the largest forward potential barrier, then weighted ensemble has variance stability. On the other hand, when A does contain a significant internal barrier then variance instability is possible and we develop conditions where an impactful reduction in variance is gained by initial condition importance sampling.

Intuitively, for variance instability, we expect that if $R \subset A$ is rare, with respect to ρ, and the steady-state flux of particles from R to B is significant then importance sampling ρ to emphasize

R should help reduce variance in the weighted ensemble estimate. That is placing more particles in a region R which contributes significantly to the number of particles that reach B then the total number of particles reaching B should increase. Thereby increasing the number of samples that contribute to the weighted ensemble estimate and thus decreasing the weighted ensemble variance. By analyzing a three state Markov chain and 1d overdamped Langevin dynamics we find this is not the case. Even when A contains a significant internal potential barrier, initial condition importance sampling is often not useful for variance reduction. Counter intuitively, we need a set $R \subset A$ which is not only rare with respect to ρ but also rare with respect to K and the steady-state flux from R to B must be insignificant; in particular, there should exist at least one other region of the initial region, A, which is more probable to transition to B from than transitioning from R to B is. By rare with respect to K we signify that is it highly improbable that the Markov chain (X_t) transitions to R.

A reason that initial condition importance sampling is ineffective is that reaching B from A is a rare event. Otherwise, standard Monte Carlo could simply be used instead of weighted ensemble. Thus, the total particle weight reaching B, $w_t(u_B)$, is diminutive. Since the mutation variance from each bin depends on the square of the total bin weight, $w_t(u)^2$, then the recycle variance from u_B is also diminutive. Hence, we could expect initial condition importance sampling to produce no impactful reduction in variance as a majority of the mutation variance is not from recycling.

Variance stability in the weighted ensemble algorithm is a benefit of this algorithm over adaptive multilevel splitting. Through each selection process in adaptive multilevel splitting, only a few samples are kept while all other trajectories become copies of the few. So, only a few samples, which are often improbable to sample, from the initial distribution end up contributing a lot to the committor probability or the adaptive multilevel splitting estimator [16]. This phenomenon leads to high variance of the adaptive multilevel splitting estimator and importance sampling on the initial condition is required to manage the estimator variance. This is not desirable as importance sampling the initial condition for adaptive multilevel splitting is computationally expensive [28] resulting in slower simulations and a requirement for more computational resources. Even for simple

problems when the initialization occurs over a metastable region that does not include a significant internal barrier, such as the example in Figure 2.1, importance sampling is still necessary to maintain reasonable adaptive multilevel splitting estimator variances [16].

3.2.1 Three State Markov Model

Consider a Markov chain $(X_t)_{t\in\mathbb{N}}$ on discrete state space $S = \{1, 2, 3\}$ with transition matrix

$$K = \begin{pmatrix} 1-\epsilon^p-\epsilon^q & \epsilon^p & \epsilon^q \\ \epsilon^r & 1-\epsilon^r-\epsilon^s & \epsilon^s \\ 1-\epsilon^t & \epsilon^t & 0 \end{pmatrix}$$

where ϵ is a small positive real number and $p, q, r, s,$ and t are all non-negative real numbers. The initial and recycle distribution is $\boldsymbol{\rho} = K(3, \cdot) = [1-\epsilon^t \ \ \epsilon^t \ \ 0]$. We develop conditions on the powers $p, q, r, s,$ and t such that initial condition importance sampling provides a significant variance reduction for the weighted ensemble estimator. Let $\boldsymbol{\mu}$ be the steady-state distribution of K and define $\boldsymbol{f} = [0\ 0\ 1]^T$ since we are interested in state $B = \{3\}$. Define $\boldsymbol{v}^2 \coloneqq \lim_{T\to\infty} \text{Var}_K(h_{t+1,T}) = \text{Var}_K(\boldsymbol{h})$ where $\boldsymbol{h} = [h_1 \ h_2 \ h_3]^T$ is the solution to the Poisson equation $(I - K)\boldsymbol{h} = \boldsymbol{f} - \boldsymbol{\mu}(\boldsymbol{f})$ satisfying $\boldsymbol{\mu}(\boldsymbol{h}) = 0$ [34].

Now assume for all time, $t \geq 0$, that all particles in state $i \in S$ form a bin denoted by u_i. Since the initialization variance, $\frac{1}{T^2}V_0$, scales as $\mathcal{O}(1/T^2)$ and we are concerned with long time horizons, T, then we can approximate $\frac{1}{T^2}V_0 \approx 0$. Also, the selection variance, V_t^S, is zero for all $t \geq 0$ as $Kh_{t+1,T}$ is a constant since all particles in a given bin u_i are at the exact same state i. Hence, the only variance is from the mutation step which we can approximate, for each bin and time t, by

$$\boldsymbol{\sigma}_t^2 = \left[\sigma_1^2(t) \ \sigma_2^2(t) \ \sigma_3^2(t)\right]^T \coloneqq \left[\frac{\mu_1^2}{N_t(1)}\text{Var}_{K(1,\cdot)}(\boldsymbol{h}) \ \ \frac{\mu_2^2}{N_t(2)}\text{Var}_{K(2,\cdot)}(\boldsymbol{h}) \ \ \frac{\mu_3^2}{N_t(3)}\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})\right]^T$$

where $\boldsymbol{N} = [N_t(1) \ N_t(2) \ N_t(3)]^T$ is the particle allocation at time t. The approximations used replaces $w_t(u_i)$ with μ_i and $\text{Var}_K(\boldsymbol{h}_{t+1,T})$ with $\text{Var}_K(\boldsymbol{h})$, which are valid for large T and a large number of particles. Since we are interested in long time horizons and are not concerned with lim-

iting the number of particles the approximations are reasonable. Also, $\eta_t^u\left(\text{Var}_K(\boldsymbol{h})\right) = \text{Var}_K(\boldsymbol{h})$ was used, which holds since all particles in a given bin are at the exact same state.

Now, when we importance sample from the distribution $\boldsymbol{\nu}$ the variance from bin u_3 is given by

$$\frac{\mu_3^2}{N_t(3)}\text{Var}_\nu\left(\left(\boldsymbol{\rho}^T \odot \boldsymbol{h}\right) \oslash \boldsymbol{\nu}\right)$$

where $\odot$ and $\oslash$ are componentwise multiplication and division respectively. Note that we require that $\left[\left(\boldsymbol{\rho}^T \odot \boldsymbol{h}\right) \oslash \boldsymbol{\nu}\right]_j = 0$ when $\left[\boldsymbol{\rho}^T \odot \boldsymbol{h}\right]_j = 0$. To see a gain from initial condition importance sampling we require the conditions

$$\text{Var}_\nu\left(\left(\boldsymbol{\rho}^T \odot \boldsymbol{h}\right) \oslash \boldsymbol{\nu}\right) < \text{Var}_\rho(\boldsymbol{h}) \tag{3.8}$$

and

$$\mu_1^2\text{Var}_{K(1,\cdot)}(\boldsymbol{h}) + \mu_2^2\text{Var}_{K(2,\cdot)}(\boldsymbol{h}) \leq \mu_3^2\text{Var}_\rho(\boldsymbol{h}). \tag{3.9}$$

The first condition, equation (3.8), states that importance sampling does reduce the recycle variance from bin u_3. If we take $\boldsymbol{\nu} = (\boldsymbol{\rho}^T \odot |\boldsymbol{h}|)/(\boldsymbol{\rho}|\boldsymbol{h}|)$, where $|\boldsymbol{h}|$ is the elementwise absolute value of $\boldsymbol{h}$, then by (1.1)

$$\text{Var}_\rho(\boldsymbol{h}) - \text{Var}_\nu\left(\left(\boldsymbol{\rho}^T \odot \boldsymbol{h}\right) \oslash \boldsymbol{\nu}\right) = \text{Var}_\rho(|\boldsymbol{h}|)$$

which is positive since variance is non-negative and $|\boldsymbol{h}|$ will not be a constant. Hence, condition (3.8) is guaranteed to hold.

Next, the second condition, equation (3.9), states that the recycle variance from bin u_3 must be at least as large as the mutation variance from all other bins when we have a uniform particle allocation $N_t(1) = N_t(2) = N_t(3)$. Note, assuming a uniform particle allocation is reasonable as often such an allocation is used in practice. We require the second condition since when the variance from bin u_3 is small, relative to all the other mutation variances, then reducing the variance through initial condition importance sampling may not produce a significant impact on the total variance. For example, if $\sigma_1^2 = 1$, $\sigma_2^2 = 2 \times 10^{-4}$, and $\sigma_3^2 = 10^{-6}$ then even if by importance

sampling we make the variance from u_3 equal to 0 the impact on the total variance $\sigma_1^2 + \sigma_2^2 + \sigma_3^2$ is negligible. We can further highlight the necessity of equation (3.9) by bounding the optimal variance improvement factor (VIF) given by

$$\text{Optimal VIF} = \frac{\frac{1}{N}\left(\mu_1\sqrt{\text{Var}_{K(1,\cdot)}(\boldsymbol{h})} + \mu_2\sqrt{\text{Var}_{K(2,\cdot)}(\boldsymbol{h})} + \mu_3\sqrt{\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})}\right)^2}{\frac{1}{N}\left(\mu_1\sqrt{\text{Var}_{K(1,\cdot)}(\boldsymbol{h})} + \mu_2\sqrt{\text{Var}_{K(2,\cdot)}(\boldsymbol{h})} + \mu_3\sqrt{\text{Var}_{\nu}((\boldsymbol{\rho}^T \odot \boldsymbol{h}) \oslash \boldsymbol{\nu})}\right)^2}, \tag{3.10}$$

which, from equation (3.7), is the optimal weighted ensemble variance over the optimal weighted ensemble plus initial condition importance sampling variance. When (3.9) does not hold then

$$\text{Optimal VIF} \leq \left(\frac{\mu_1\sqrt{\text{Var}_{K(1,\cdot)}(\boldsymbol{h})} + \mu_2\sqrt{\text{Var}_{K(2,\cdot)}(\boldsymbol{h})} + \mu_3\sqrt{\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})}}{\mu_1\sqrt{\text{Var}_{K(1,\cdot)}(\boldsymbol{h})} + \mu_2\sqrt{\text{Var}_{K(2,\cdot)}(\boldsymbol{h})}}\right)^2 \leq 4.$$

This implies a maximum standard deviation improvement of 2, which in general is not significant especially for the computational cost of implementing the initial condition importance sampling.

Now we establish conditions on powers p, q, r, s, and t such that the requirement (3.9) holds. These conditions are established by finding cases of powers p, q, r, s, and t where

$$\frac{\mu_1^2\text{Var}_{K(1,\cdot)}(\boldsymbol{h})}{\mu_3^2\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})} = \mathcal{O}(\epsilon^{k_1}) \quad \text{and} \quad \frac{\mu_2^2\text{Var}_{K(2,\cdot)}(\boldsymbol{h})}{\mu_3^2\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})} = \mathcal{O}(\epsilon^{k_2}) \tag{3.11}$$

for $k_1, k_2 > 0$. When (3.11) holds then

$$\frac{\mu_1^2\text{Var}_{K(1,\cdot)}(\boldsymbol{h}) + \mu_2^2\text{Var}_{K(2,\cdot)}(\boldsymbol{h})}{\mu_3^2\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})} = \mathcal{O}(\epsilon^{\min\{k_1,k_2\}}) \leq 1$$

for small enough ϵ. Hence, the powers p, q, r, s, and t which satisfy (3.11) will also satisfy (3.9), for small enough ϵ, and are captured in the four cases below

$$\begin{cases} p > t + 2q, \;\; t > s \;\mid\; r \geq s, \;\; q < s, \;\; p + q \leq 2s, \text{ and } 2q \leq r + s \\ 2s > t + 3q, \; t > s \;\mid\; r \geq s, \;\; q < s, \;\; s \leq p, \;\; 2s \leq p + q, \;\; p \geq q + t, \text{ and } 2q \leq r + s \\ p > t + 2q, \;\; t > r \;\mid\; r \leq s, \;\; q < s, \;\; p + q \leq 2r, \text{ and } 2q \leq r + s \\ 2r > t + 3q, \; t > r \;\mid\; r \leq s, \;\; q < s, \;\; r \leq p, \;\; 2r \leq p + q, \text{ and } p \geq q + t. \end{cases} \tag{3.12}$$

Further details establishing the conditions in (3.12) are lengthy and mundane so they are omitted here and instead provided in Appendix A. Note, the first two inequalities in each case of equation (3.12) determine the scaling powers k_1 and k_2 on ϵ in (3.11). For instance, $p > t + 2q$ and $t > s$ are the first two inequalities of case one of equation (3.12) so $k_1 = p - t - 2q$ and $k_2 = t - s$.

Table 3.4 lists all natural number powers of p, q, r, s, and t which are less than ten and satisfy at least one of the four cases in equation (3.12). We note that the nine combinations of powers listed in Table 3.4 represents less than a tenth of a percent of the total number of possible natural number choices of powers. That is cases where initial condition importance sampling provides a significant variance reduction are uncommon or fringe cases. So, stability of the weighted ensemble variance can typically be expected for a three state Markov model.

Table 3.4: Natural number powers less than ten on small parameter ϵ such that initial condition importance sampling in a three state Markov model provides a significant reduction in the weighted ensemble estimator variance.

p	9	9	9	9	9	9	9	9	9
q	1	1	1	1	1	1	1	1	1
r	5	6	7	8	9	5	5	5	5
s	5	5	5	5	5	6	7	8	9
t	6	6	6	6	6	6	6	6	6

We consider, as an example, the powers in first column of Table 3.4, that is $(p, q, r, s, t) = (9, 1, 5, 5, 6)$, which satisfies case one of equation (3.12). From the transition matrix, K, we can see that state 1 will almost never transition to state 2, leaving state 2 is highly improbable, and recycling to state 2 is highly improbable. Also, we have $\mu_1^2 \text{Var}_{K(1,\cdot)}(\boldsymbol{h})/\mu_3^2 \text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h}) = \mathcal{O}(\epsilon)$ and $\mu_2^2 \text{Var}_{K(2,\cdot)}(\boldsymbol{h})/\mu_3^2 \text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h}) = \mathcal{O}(\epsilon)$ so the variance reduction from initial condition importance sampling is expected to scale as $\mathcal{O}(\epsilon^{-1})$. For various ϵ, Table 3.5 lists the variance improvement factor (VIF) from initial condition importance sampling given by

$$\text{VIF} = \frac{\mu_1^2 \text{Var}_{K(1,\cdot)}(\boldsymbol{h}) + \mu_2^2 \text{Var}_{K(2,\cdot)}(\boldsymbol{h}) + \mu_3^2 \text{Var}_{\rho}(\boldsymbol{h})}{\mu_1^2 \text{Var}_{K(1,\cdot)}(\boldsymbol{h}) + \mu_2^2 \text{Var}_{K(2,\cdot)}(\boldsymbol{h}) + \mu_3^2 \text{Var}_{\nu}((\boldsymbol{\rho}^T \odot \boldsymbol{h}) \oslash \boldsymbol{\nu})}. \tag{3.13}$$

Table 3.5 also lists the optimal VIF (3.10), along with the mean first passage time to $B = \{3\}$ from ρ, from state 1, and from state 2. As expected, the mean first passage time from state 2 to $B = \{3\}$ is large as leaving state 2 is highly improbable, and consequently reaching B from state 2 will be highly improbable. Also, the mean first passage time from state 1 to B is approximately $1/\epsilon$, which is expected as transitioning from state 1 to B occurs with probability ϵ. Finally, as ϵ decreases we see both the VIF and optimal VIF increase approximately as $\mathcal{O}(\epsilon^{-1})$.

Table 3.5: Variance improvement factor (VIF), optimal VIF, and mean first passage times for powers $(p, q, r, s, t) = (9, 1, 5, 5, 6)$ on ϵ in a three state Markov transition matrix.

ϵ	VIF	Optimal VIF	$\mathbb{E}^1(\tau_B)$	$\mathbb{E}^2(\tau_B)$	$\mathbb{E}^\rho(\tau_B)$
10^{-1}	2.86	3.36	10.0	5.0×10^4	10.05
10^{-2}	19.25	17.69	100	5.0×10^9	100.005
10^{-3}	182.88	133.91	1.0×10^3	5.0×10^{14}	1.0×10^3
10^{-4}	1.82×10^3	1.21×10^3	1.0×10^4	5.0×10^{19}	1.0×10^4

One explanation for why initial condition importance sampling significantly reduces variance in the cases listed in equation (3.12) is: Excluding the sink there exists a region of state space, R, that is very rare (potentially even impossible) for a particle to transition to, but it is possible to recycle into. The probability of recycling to R is also rare but more probable than a particle transitioning there. Region R is also sufficiently difficult to escape from, at least as difficult as it to recycle to R, implying a large mean first passage time from R to B. Furthermore, there must exist another region of state space that is far more probable to transition to B from than it is to transition from R to B. In such a case, importance sampling to this rare region R reduces the variance in the weighted ensemble steady-state estimator by reducing the impact, on variance, that a Markov chain trajectory has when it escapes R and reaches B. A three state transition matrix, K, which satisfies one of the four cases of inequalities in equation (3.12) has rare region $R = \{2\}$ as shown in Table 3.5.

We can further support this idea by considering the transition matrix identical to K except the recycling density is $K(3, \cdot) = \rho = [\epsilon^t \; 1 - \epsilon^t \; 0]$. Now it is rare to recycle to state 1 instead of state 2. Again by comparing cases on how the mutation variance from bins u_1, u_2, and u_3 scale as

powers of ϵ we get the following four cases of inequalities for which (3.9) holds

$$\begin{cases} r > t + 2s, \quad t > q \quad | \quad p \geq q, \;\; s < q, \;\; r + s \leq 2q, \text{ and } 2s \leq p + q \\ 2q > t + 3s, \;\; t > q \quad | \quad p \geq q, \;\; s < q, \;\; q \leq r, \;\; 2q \leq r + s, \;\; r \geq s + t, \text{ and } 2s \leq p + q \\ r > t + 2s, \quad t > p \quad | \quad p \leq q, \;\; s < q, \;\; r + s \leq 2p, \text{ and } 2s \leq p + q \\ 2p > t + 3s, \;\; t > p \quad | \quad p \leq q, \;\; s < q, \;\; p \leq r, \;\; 2p \leq r + s, \text{ and } r \geq s + t. \end{cases}$$

Comparing to the original cases of inequality (3.12) we have exchanged p and r and exchanged q and s, which makes $R = \{1\}$ the rare region.

3.2.2 1D Overdamped Langevin Dynamics

Consider a Markov chain $(X_t)_{t \geq 0}$ which is governed by overdamped Langevin dynamics (2.13) on state space $S = (a, b)$ where a is a reflecting boundary and b is an absorbing boundary. Let the initial distribution $X_0 \sim \rho$ where $\text{supp}(\rho) = A \subseteq (a, b)$ and when X_t reaches $B = \{b\}$ it is immediately recycled in A according to ρ.

In steady-state, $T \to \infty$, and with a sufficiently large number of particles and bins, we can approximate a bin u_x over every point x in state space and $w_t(u_x) \approx \mu(x)$. Further including the resampling time limit, $\Delta t \to 0$, we can approximate $\eta_t^u \left(\text{Var}_K \left(h_{t+1,T} \right) \right) \approx v^2(x)$ where $v^2(x)$ is the variance function given in equation (2.23). Since all particles in a bin u_x are considered to be at the same point x the selection variance, V_t^S, is zero. The initialization variance contributes only at time $t = 0$ so for large T we can approximate $\frac{1}{T^2} V_0 \approx 0$. Hence, we only have mutation variance, which includes the recycle variance and can be approximated, at time t, by

$$V_t^M = \mathbb{E} \left[\sum_u \frac{w_t(u)^2}{N_t(u)} \eta_t^u (\text{Var}_K(h_{t+1,T})) \right] \approx \int_a^b \frac{\mu(x)^2}{N_t(x)} v(x)^2 dx + \frac{(\mathcal{J}\mu(b))^2}{N_t(b)} \text{Var}_\rho(h) \tag{3.14}$$

where $N_t(x) = N_t(u_x)$ is the particle allocation at time t and $\mathcal{J}\mu(b)$ is the steady-state flux into $B = \{b\}$. Note, $\mathcal{J}$ is a flux operator which for 1d overdamped Langevin dynamics given by

$$\mathcal{J} g = -V' g - \beta^{-1} g'$$

for continuously differentiable functions $g : \mathbb{R} \to \mathbb{R}$ [23]. The second term on the right hand side of equation (3.14) is the initial condition, or recycle, variance which results from the fact that $w_t(u_b) \approx \mathcal{J}\mu(b)$. Note we do not approximate $w_t(u_b)$ with $\mu(b)$ since X_t recycles immediately upon reaching $B = \{b\}$. Thus, $\mu(b)$ is zero whereas $w_t(u_b)$ is not and will be equal to the flux of weight into b at time t, which we approximate with the steady-state flux into b, $\mathcal{J}\mu(b)$. When importance sampling from distribution ν, the recycle variance is given by

$$\frac{(\mathcal{J}\mu(b))^2}{N_t(b)}\text{Var}_\nu\left(\frac{\rho h}{\nu}\right)$$

where we require $\rho h/\nu = 0$ when $\nu = 0$.

As in the three state model, Section 3.2.1, we require two conditions

$$\text{Var}_\nu\left(\frac{\rho h}{\nu}\right) < \text{Var}_\rho(h) \tag{3.15}$$

and

$$\int_a^b \mu(x)^2 v(x)^2 dx \leq (\mathcal{J}\mu(b))^2 \text{Var}_\rho(h) \tag{3.16}$$

to have a significant variance reduction from initial condition importance sampling. Requirement (3.15) states that importance sampling does reduce the recycle variance. By choosing $\nu = \rho|h|/\int_a^b \rho(x)|h(x)|dx$ then using (1.1)

$$\text{Var}_\rho(h) - \text{Var}_\nu\left(\frac{\rho h}{\nu}\right) = \text{Var}_\rho(|h|)$$

which is positive since variance is non-negative and $|h|$ will not a constant outside of trivial choices for $V(x)$. Hence, the first requirement (3.15) is guaranteed to hold when the optimal importance sampling distribution, $\nu \propto \rho|h|$, is chosen. Now the second requirement (3.16) states the recycle variance is at least as large as all other mutation variances when assuming a uniform particle allocation. Again, as in Section 3.2.1, assuming a uniform particle allocation is reasonable as often such an allocation is used in practice and requirement (3.16) ensures the recycle variance is large

enough to have an impact on the total mutation variance. Furthermore, when (3.16) does not hold we can bound the optimal variance improvement factor (VIF)

$$\text{Optimal VIF} = \frac{\frac{1}{N}\left(\int_a^b \mu(x)^2 v(x)^2 dx + (\mathcal{J}\mu(b))^2 \text{Var}_\rho(h)\right)^2}{\frac{1}{N}\left(\int_a^b \mu(x)^2 v(x)^2 dx + (\mathcal{J}\mu(b))^2 \text{Var}_\nu\left(\frac{\rho h}{\nu}\right)\right)^2} \leq \left(1 + \frac{(\mathcal{J}\mu(b))^2 \text{Var}_\rho(h)}{\int_a^b \mu(x)^2 v(x)^2 dx}\right)^2 \leq 4.$$

Now, let $\Delta V = V(x^*) - V(y^*)$ be the largest forward barrier (3.4) of $V(x)$ on (a, b). The following proposition shows that weighted ensemble has variance stability so long as ΔV is not in the interior of A; that is $\text{supp}(\rho) \subseteq (a, x^*)$.

Proposition 3.2.1. *If* $\text{supp}(\rho) \subseteq (a, x^*)$ *then* $\text{Var}_\rho(h) \sim 0$ *in the large* β *limit.*

Proof. For $x^* \in (a, b]$ and $y^* \in [a, x^*)$ assume $\Delta V = V(x^*) - V(y^*)$ is the largest forward potential energy barrier and let ρ be any density satisfying $\text{supp}(\rho) \subseteq (a, x^*)$. From Proposition 3.1.1, when $x \in (a, x^*)$ then for large β, by Laplace's Method, the mean first passage time can be approximated by $u(x) \sim C_{y^*} C_{x^*} e^{\beta \Delta V}$, where C_{y^*} and C_{x^*} are constants dependent on y^* and x^*, respectively. Therefore

$$\text{Var}_\rho(h) = \frac{\text{Var}_\rho(u)}{\left(\mathbb{E}^\rho(\tau_B)\right)^2} \sim \left(\frac{C_{y^*} C_{x^*}}{\mathbb{E}^\rho(\tau_B)}\right)^2 e^{2\beta \Delta V} \left[\int_a^{x^*} \rho(x) dx - \left(\int_a^{x^*} \rho(x) dx\right)^2\right] = 0$$

as $\int_a^{x^*} \rho(x) = 1$. ⊠

Proposition 3.2.1 gives that initial condition importance sampling will not be beneficial, in the large β limit, when A does not contain significant internal barriers as condition (3.15) will not hold. Furthermore, when A does not contain significant internal barriers, Proposition 3.2.1 shows that the variance from recycling is zero in the large β limit and thus could not be reduced further from initial condition importance sampling.

Now, when A contains a significant internal potential barrier, we will develop a condition using (3.16) such that initial condition importance sampling has an impactful reduction on the variance of the weighted ensemble estimator. Let C be the normalization of the steady-state distribution μ

(2.20). Then from equations (2.20) and (2.23) we have

$$\int_a^b \mu(x)^2 v(x)^2 dx = \frac{2\beta}{C^2(\mathbb{E}^\rho(\tau_B))^2} \int_a^b \left(\int_x^b e^{\beta V(z)} \left(\int_a^z \rho(y) dy \right) dz \right)^2 \left(\int_a^x e^{-\beta V(z)} dz \right)^2 dx.$$

Since $\mu(b) = 0$ then $\mathcal{J}\mu(b) = -\beta^{-1}\mu'(b) = \beta^{-1}C^{-1} = 1/\mathbb{E}^\rho(\tau_B)$ where the final equality results from the Hill relation. Using properties of variance if $\kappa^2 = \text{Var}_\eta(g)$ then $\text{Var}_\eta(sg + c) = s^2\kappa^2$ for η a probability measure, g a measurable function, and s and c real constants. Thus, using equation (2.22) we have

$$(\mathcal{J}\mu(b))^2 \text{Var}_\rho(h) = \frac{1}{\beta^2 C^2 (\mathbb{E}^\rho(\tau_B))^2} \text{Var}_\rho(u)$$

where u is the mean first passage time given in (2.18). Hence, inequality (3.16) is equivalent to

$$\text{Var}_\rho(u) \geq 2\beta^3 \int_a^b \left(\int_x^b e^{\beta V(z)} \left(\int_a^z \rho(y) dy \right) dz \right)^2 \left(\int_a^x e^{-\beta V(z)} dz \right)^2 dx \coloneqq 2\beta^3 I \tag{3.17}$$

where we define $I = \int_a^b \left(\int_x^b e^{\beta V(z)} \left(\int_a^z \rho(y) dy \right) dz \right)^2 \left(\int_a^x e^{-\beta V(z)} dz \right)^2 dx$.

Next, we further simplify (3.17), for particular potential V, using Laplace's Method. Assume that the potential $V(x)$ obtains two local minimums at y_1, y_2 and two local maximums at x_1, x_2 such that $a < y_1 < x_1 < y_2 < x_2 < b$. Define ρ as the mixture of δ-functions

$$\rho(x) = \lambda\delta(x - y_1) + (1 - \lambda)\delta(x - y_2).$$

That is, upon reaching B the Markov chain recycles at one of two points y_1 and y_2 with probability λ and $1-\lambda$, respectively. We further assume that $\lambda = e^{-\beta d}$ for some positive constant d. The choice of potential, V, and initial density, ρ, are specific to draw an analogy to the three state Markov model where we consider $(a, x_1), [x_1, x_2)$, and $[x_2, b)$ as the three states from the 1d model. Then transition probabilities between bins, for large β, is then determined by the size of the barriers between the bins, namely $\Delta V_1 = V(x_1) - V(y_1)$ and $\Delta V_2 = V(x_2) - V(y_2)$, and the recycle

density ρ. Hence, we desire a condition from (3.17) which is dependent on the barrier sizes as well as ρ.

Define points

$$\chi_1 \coloneqq \inf\{x \in (x_1, y_2) : V(x) < V(x_2)\}, \quad \chi_2 \coloneqq \inf\{x \in (x_1, y_2) : V(x) < V(y_1)\},$$

and define Laplace function

$$L(x, y) \coloneqq \frac{e^{\beta(V(x)-V(y))}}{\sqrt{|V''(x)|\,|V''(y)|}}.$$

Note, when $V(x_1) < V(x_2)$ then $\{x \in (x_1, y_2) : V(x) < V(x_2)\} = \emptyset$ and so we take $\chi_1 = x_1$ in such a case. Similarly, when $V(y_1) < V(y_2)$ then $\{x \in (x_1, y_2) : V(x) < V(y_1)\} = \emptyset$ and thus we will take $\chi_2 = y_2$.

Now, in the large β limit, using Laplace's Method

$$\begin{aligned} I \sim &\left(\frac{2\pi}{\beta}\right)^2 \left(\lambda^2(y_1 - x_1)(L(x_1, y_1) + L(x_2, y_1))^2 + (\chi_2 - \chi_1)L^2(x_2, y_1)\mathbb{1}_{\chi_1<\chi_2}\right) \\ &+ \left(\frac{2\pi}{\beta}\right)^2 (x_2 - y_2)\left(L(x_2, y_1) + L(x_2, y_2)\right)^2 \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} \mathrm{Var}_\rho(u) \sim{}& \lambda\left[2\pi L(x_1, y_1) + 2\pi\left(L(x_2, y_1) + L(x_2, y_2)\right)\right]^2 + (1 - \lambda)\left[2\pi\left(L(x_2, y_1) + L(x_2, y_2)\right)\right]^2 \\ &- \left(\lambda\left[2\pi L(x_1, y_1) + 2\pi\left(L(x_2, y_1) + L(x_2, y_2)\right)\right] + (1 - \lambda)\left[2\pi\left(L(x_2, y_1) + L(x_2, y_2)\right)\right]\right)^2 \\ ={}& (2\pi)^2\lambda(1 - \lambda)L^2(x_1, y_1). \end{aligned} \tag{3.19}$$

Using (3.18) and (3.19), then (3.17) gives

$$\begin{aligned} \lambda(1 - \lambda)L^2(x_1, y_1) \geq{}& 2\beta\left(\lambda^2(y_1 - x_1)(L(x_1, y_1) + L(x_2, y_1))^2 + (\chi_2 - \chi_1)L^2(x_2, y_1)\mathbb{1}_{\chi_1<\chi_2}\right) \\ &+ 2\beta(x_2 - y_2)\left(L(x_2, y_1) + L(x_2, y_2)\right)^2. \end{aligned} \tag{3.20}$$

From (3.20) we can note that $\Delta V_1 = V(x_1) - V(y_1)$ must be the largest forward barrier. That is ΔV_1 must be a larger barrier than $\Delta V_2 = V(x_2) - V(y_2)$ and $\Delta V_{12} = V(x_2) - V(y_1)$. This is required as the left hand side of (3.20) scales no greater than $\mathcal{O}(e^{2\beta\Delta V_1})$ where as the right hand side of (3.20) scales no less than $\mathcal{O}(e^{2\beta \max\{\Delta V_2, \Delta V_{12}\}})$. Thus, if $\Delta V_2 > \Delta V_1$ or $\Delta V_{12} > \Delta V_1$ then (3.20) cannot hold. Consequently, as we desire that (3.20) depends on both potential barriers then $d \geq \Delta V_1 - \max\{\Delta V_2, \Delta V_{12}\}$ since $\lambda^2 L^2(x_1, y_1) = \mathcal{O}(e^{2\beta(\Delta V_1 - d)})$ and $(L(x_2, y_1 + L^2(x_2, y_2))^2 = \mathcal{O}(e^{2\beta \max\{\Delta V_2, \Delta V_{12}\}})$. Similarly, we require that $d \leq 2(\Delta V_1 - \max\{\Delta V_2, \Delta V_{12}\})$ otherwise the left hand side of (3.20) will be a lower exponential order than the right hand side of (3.20). Also, as $\Delta V_{12} > \Delta V_1$ then x_1 must be the global maximizer of $V(x)$ on (a, b).

Consider an example potential, $V(x)$, shown in Figure 3.4 which is a linear combination of a sinusoidal and multiple Gaussian functions. Here $y_1 \approx .876$, $V(y_1) \approx 1$, $|V''(y_1)| \approx 2.94$,

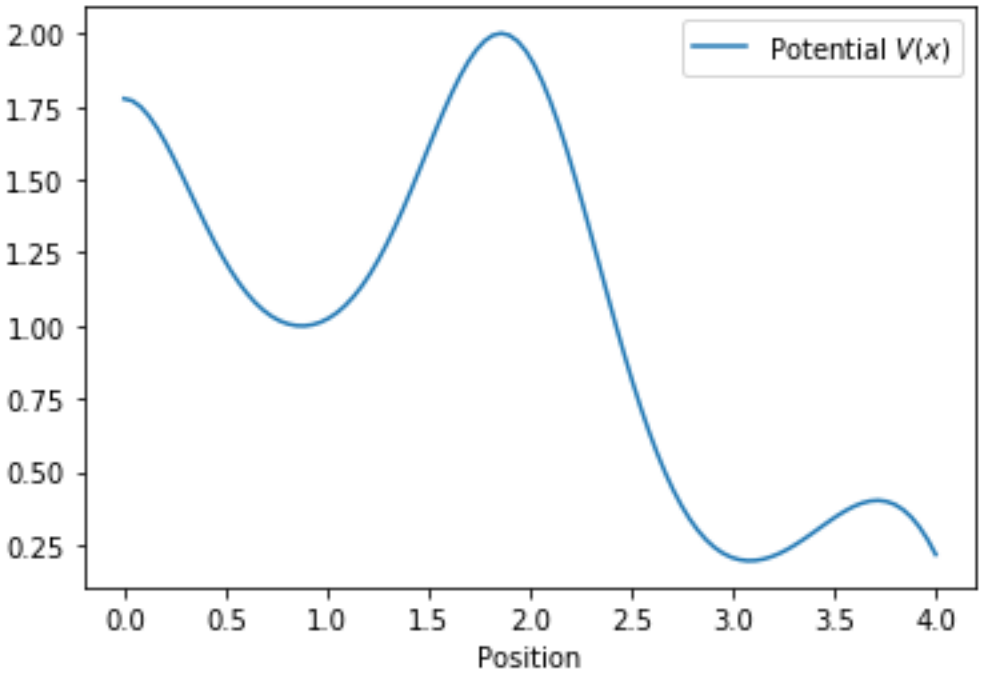


Figure 3.4: Two barrier potential formed as a linear combination of sinusoidal and multiple Gaussian functions.

$x_1 \approx 1.858$, $V(x_1) \approx 2$, $|V''(x_1)| \approx 8$, $y_2 \approx 3.084$, $V(y_2) \approx .2$, $|V''(y_2)| \approx 2.67$, $x_2 \approx 3.713$, $V(x_2) \approx .4$, and $|V''(x_2)| \approx 3.5$. For the recycle density we choose $d = 6/5$. Then the

left hand side of (3.20) scales as $\lambda(1-\lambda)L^2(x_1,y_1) = \mathcal{O}(e^{4\beta/5})$ whereas the right hand side of (3.20) scales as $2\beta\left(\lambda^2(y_1-x_1)L^2(x_1,y_1)+(x_2-y_2)L^2(x_2,y_2)\right) = \mathcal{O}(\beta e^{2\beta/5})$. Hence, (3.20) and consequently (3.16) holds for large β. We now map the 1d continuous overdamped Langevin dynamics onto a three state Markov model on states $(0,x_1)$, $[x_1,x_2)$, and $[x_2,4)$. Let Δt be a time step then, in the large β limit, we get the following three state transition matrix [35]

$$P = \begin{pmatrix} 1-\frac{\Delta t}{\pi L(x_1,y_1)} & \frac{\Delta t}{2\pi L(x_1,y_1)} & \frac{\Delta t}{2\pi L(x_1,y_1)} \\ \frac{\Delta t}{2\pi L(x_1,y_2)} & 1-\frac{\Delta t}{2\pi}\left(\frac{1}{L(x_1,y_2)}+\frac{1}{L(x_2,y_2)}\right) & \frac{\Delta t}{2\pi L(x_2,y_2)} \\ e^{-\beta d} & 1-e^{-\beta d} & 0 \end{pmatrix}.$$

Table 3.6 gives a numerically calculated variance improvement factor (VIF) (3.13) and optimal VIF (3.10), for various β, of the three state Markov model generated by P. We can see that both the VIF

Table 3.6: Variance improvement factor (VIF) and optimal VIF for the three state representation of 1d overdamped Langevin on potential in Figure 3.4 with different choices for β.

β	5	10	15	20	25	30	35
VIF	1.47	2.37	4.70	10.99	28.05	74.41	200.97
Optimal VIF	1.65	2.80	5.30	10.96	24.53	58.46	148.07

and optimal VIF increase exponentially in β as further highlighted in Figure 3.5. An exponential relationship is anticipated as the ratio of left hand side and right hand side of (3.20) will grow exponentially in β. Thus, for a sufficiently large β, we have multiple orders of magnitude reduction in the weighted ensemble estimator variance by importance sampling the initial distribution, ρ.

First, We note that the choice of recycle density, ρ, as a mixture of δ-functions is not strictly necessary. Instead, if we assume $\text{supp}(\rho) = (a,x_2)$ and define $\lambda = \int_a^{x_1}\rho(x)dx$ the same analysis and resulting condition (3.20) should hold. Although, care would need to be taken in handling cases where the all mass of ρ in (a,x_1) approaches x_1 as this may slightly alter some of the underlying Laplace's Method approximations used. Second, we note that for 1d overdamped Langevin dynamics the required rare region R is (a,x_1) since recycling in (a,x_1) must be sufficiently difficult and transitioning to (x_2,b) is far more likely from (x_1,x_2) than from (a,x_1), for large β,

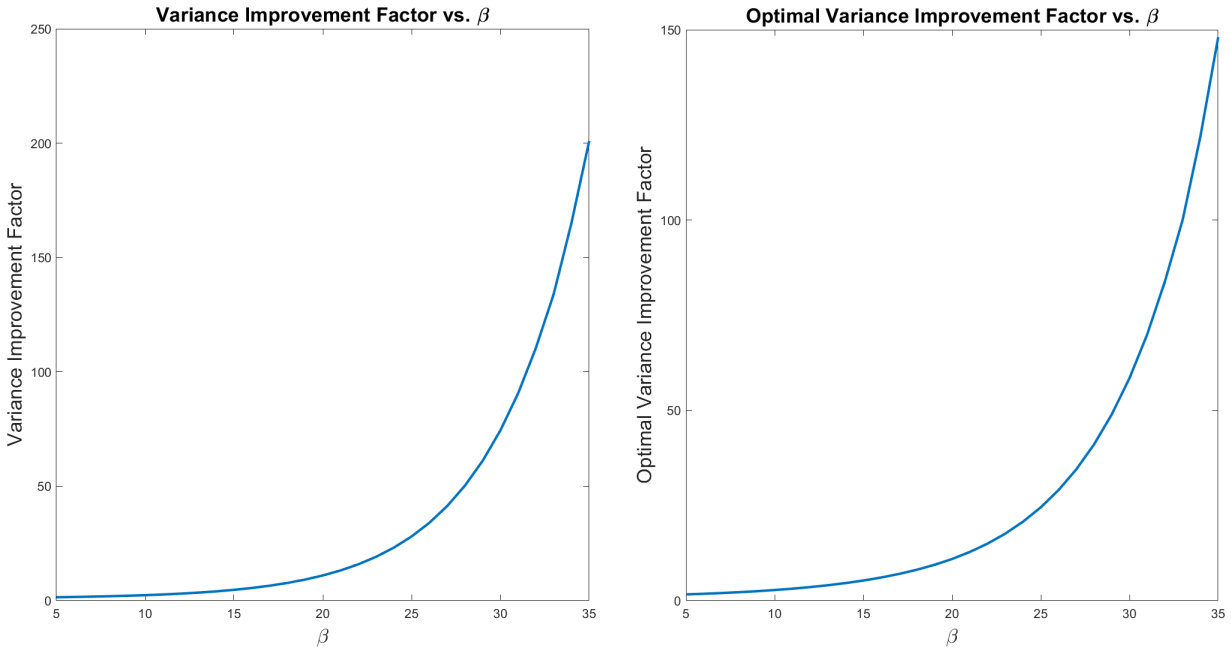


Figure 3.5: Variance improvement factor (left) and optimal variance improvement factor (right) versus β. Both improvement factors are exponentially dependent on β.

as $\Delta V_1 > \max\{\Delta V_2, \Delta V_{12}\}$. Furthermore, define $\mu_1 = \int_a^{x_1} \mu(x)dx$ and $\mu_2 = \int_{x_1}^{x_2} \mu(x)dx$ to be the mass of the steady-steady distribution in (a, x_1) and (x_1, x_2), respectively. Then, applying Laplace's Method to the recycle variance when importance sampling, $\text{Var}_\nu(\rho h/\nu)$, using optimal importance sampling distribution, $\nu \propto \rho|h|$, we get that $\text{Var}_\nu(\rho h/\nu)$ is proportional μ_1. Hence, to obtain near zero recycle variance with importance sampling we desire μ_1 to be near zero. Note that applying Laplace's Method, for large β, we have $\mu_1 \propto \lambda L(x_1, y_1) + L(x_2, y_1)$ and $\mu_2 \propto L(x_2, y_2)$. Thus, we require that y_2 is the global minimizer of $V(x)$ on (a, b) so that in the large β limit $\mu_1 \sim 0$, $\mu_2 \sim 1$, and $\text{Var}_\nu(\rho h/\nu) \sim 0$. Therefore, transitioning to (a, x_1) is also incredibly rare as it involves overcoming the largest potential barrier $V(x_1) - V(y_1)$.

In conclusion, when A does not contain a significant internal barrier then weighted ensemble has variance stability. Alternatively, when A does contain a significant internal barrier then a condition on the barrier sizes and recycle density must be satisfied for there to be instability in the weighted ensemble variance. As in the three state model, variance instability in the 1d overdamped Langevin model requires a region $R \subset A$ which is rare to recycle and transition to and has an insignificant contribution to the steady-state flux into B compared to another region

of A. Therefore, weighted ensemble often has variance stability in initial condition importance sampling whereas adaptive multilevel splitting does not, which is a benefit of weighted ensemble over adaptive multilevel splitting.

Bibliography

[1] Dirk P Kroese, Tim Brereton, Thomas Taimre, and Zdravko I Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

[2] Masato Shimomura and Yasuhiro Takashima. Application of monte-carlo tree search to traveling-salesman problem. In *The 20th Workshop on Synthesis And System Integration of Mixed Information technologies (SASIMI)*, pages 352–356, 2016.

[3] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[4] Tuen Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

[5] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[6] Christian P Robert and George Casella. The metropolis—hastings algorithm. In *Monte Carlo statistical methods*, pages 231–283. Springer, 1999.

[7] Eric Paquet and Herna L Viktor. Molecular dynamics, monte carlo simulations, and langevin dynamics: a computational review. *BioMed research international*, 2015, 2015.

[8] Donald A McQuarrie. *Statistical mechanics*. Sterling Publishing Company, 2000.

[9] Daniel M Zuckerman and Lillian T Chong. Weighted ensemble simulation: review of methodology, applications, and software. *Annual review of biophysics*, 46:43–57, 2017.

[10] D. Aristoff, T. Leliévre, C. G. Mayne, and I. Teo. Adaptive multilevel splitting in molecular dynamics simulations. *ESAIM*, 48:215–225, 2015.

[11] Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

[12] Gary A Huber and Sangtae Kim. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Biophysical journal*, 70(1):97–110, 1996.

[13] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.

[14] Gerardo Rubino and Bruno Tuffin. *Rare event simulation using Monte Carlo methods*. John Wiley & Sons, 2009.

[15] Atipat Rojnuckarin, Sangtae Kim, and Shankar Subramaniam. Brownian dynamics simulations of protein folding: access to milliseconds time scale and beyond. *Proceedings of the National Academy of Sciences*, 95(8):4288–4292, 1998.

[16] Lopes Laura. *Numerical Methods for Simulating Rare Events in Molecular Dynamics*. PhD thesis, Université Paris-EST, 2019.

[17] Bin W Zhang, David Jasnow, and Daniel M Zuckerman. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of chemical physics*, 132(5):054107, 2010.

[18] Mieczyslaw Torchala, Przemyslaw Chelminiak, Michal Kurzynski, and Paul A Bates. Ratrav: a tool for calculating mean first-passage times on biochemical networks. *BMC systems biology*, 7(1):1–13, 2013.

[19] Frédéric Cérou, Arnaud Guyader, and Mathias Rousset. Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4):043108, 2019.

[20] Divesh Bhatt, Bin W Zhang, and Daniel M Zuckerman. Steady-state simulations using weighted ensemble path sampling. *The Journal of chemical physics*, 133(1):014110, 2010.

[21] Sean P Meyn and Richard L Tweedie. Stability of markovian processes ii: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993.

[22] David Aristoff. An ergodic theorem for the weighted ensemble method. *Journal of Applied Probability*, page 1–15, 2022.

[23] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer, 2015.

[24] Vladimir I Bogachev, Nicolai V Krylov, Michael Röckner, and Stanislav V Shaposhnikov. *Fokker-Planck-Kolmogorov Equations*, volume 207. American Mathematical Soc., 2015.

[25] William Feller. Diffusion processes in one dimension. *Transactions of the American Mathematical Society*, 77(1):1–31, 1954.

[26] David Aristoff and Daniel Zuckerman. Optimizing weighted ensemble sampling of steady state. *Multiscale Modeling and Simulation*, 18:646–673, 2020.

[27] David Aristoff. Analysis and optimization of weighted ensemble sampling. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(4):1219–1238, 2018.

[28] Titus S Van Erp. Dynamical rare event simulation techniques for equilibrium and nonequilibrium systems. *Advances in Chemical Physics*, 151:27, 2012.

[29] Frédéric Cérou, Arnaud Guyader, Tony Lelievre, and David Pommier. A multiple replica approach to simulate reactive trajectories. *The Journal of chemical physics*, 134(5):054108, 2011.

[30] Schilings Claudia, Sprungk Bjorn, and Wacker Philipp. On the convergence of the laplace approximation and noise-level-robustness of laplace-based monte carlo methods for bayesian inverse problems. *Numerische Mathematik*, 145:915–971.

[31] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.

[32] Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

[33] Nils Berglund and Barbara Gentz. The eyring-kramers law for potentials with nonquadratic saddles. *arXiv preprint arXiv:0807.1681*, 2008.

[34] Tony Lelievre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.

[35] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

Appendix A

Provided in this appendix is more details for Section 3.2.1 on the development of conditions for powers p, q, r, s, and t such that initial condition importance sampling requirement (3.9) holds. First, recall the transition matrix of interest is

$$K = \begin{pmatrix} 1-\epsilon^p-\epsilon^q & \epsilon^p & \epsilon^q \\ \epsilon^r & 1-\epsilon^r-\epsilon^s & \epsilon^s \\ 1-\epsilon^t & \epsilon^t & 0 \end{pmatrix}$$

which has steady-state distribution, $\boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \mu_3]$, given by

$$\mu_1 = \frac{\epsilon^r + \epsilon^s - \epsilon^{s+t}}{C}, \quad \mu_2 = \frac{\epsilon^p + \epsilon^{q+t}}{C}, \quad \text{and} \quad \mu_3 = \frac{\epsilon^{q+r} + \epsilon^{p+s} + \epsilon^{q+s}}{C} \tag{A.1}$$

where $C = \epsilon^{q+r} + \epsilon^{q+s} + \epsilon^{q+t} + \epsilon^{p+s} - \epsilon^{s+t} + \epsilon^p + \epsilon^r + \epsilon^s$. We are interested in state $B = \{3\}$ so let $\boldsymbol{f} = [0 \ 0 \ 1]^T$. Define $\boldsymbol{v}^2 \coloneqq \lim_{T\to\infty} \text{Var}_K(h_{t+1,T}) = \text{Var}_K(\boldsymbol{h})$ where $\boldsymbol{h} = [h_1 \ h_2 \ h_3]^T$ is the solution to the Poisson equation $(I - K)\boldsymbol{h} = \boldsymbol{f} - \boldsymbol{\mu}(\boldsymbol{f})$ satisfying $\boldsymbol{\mu}(\boldsymbol{h}) = 0$ [34].

Under certain assumption, given in Section 3.2.1, the only variance is from the mutation step which we can approximate, for each bin at step t, by

$$\boldsymbol{\sigma}_t^2 = \left[\sigma_1^2(t) \ \sigma_2^2(t) \ \sigma_3^2(t)\right]^T \coloneqq \left[\frac{\mu_1^2}{N_t(1)}\text{Var}_{K(1,\cdot)}(\boldsymbol{h}) \ \ \frac{\mu_2^2}{N_t(2)}\text{Var}_{K(2,\cdot)}(\boldsymbol{h}) \ \ \frac{\mu_3^2}{N_t(3)}\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})\right]^T.$$

As discussed in Section 3.2.1, for initial condition importance sampling to have significant benefit we want (3.9), which is

$$\mu_1^2\text{Var}_{K(1,\cdot)}(\boldsymbol{h}) + \mu_2^2\text{Var}_{K(2,\cdot)}(\boldsymbol{h}) \leq \mu_3^2\text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h}),$$

to hold. So, we will assume a uniform allocation $N_t(1) = N_t(2) = N_t(3)$ and consequently drop the dependence of $\sigma_1^2(t), \sigma_2^2(t)$, and $\sigma_3^2(t)$ on the particle allocation and time. We write σ_1^2, σ_2^2, and σ_3^2 for the mutation variance terms which are independent of the particle allocation and time. Now

we aim to find conditions on the powers p, q, r, s, and t such that for $k_1, k_2 > 0$

$$\frac{\sigma_1^2}{\sigma_3^2} = \frac{\mu_1^2 \text{Var}_{K(1,\cdot)}(\boldsymbol{h})}{\mu_3^2 \text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})} = \mathcal{O}(\epsilon^{k_1}) \quad \text{and} \quad \frac{\sigma_2^2}{\sigma_3^2} = \frac{\mu_2^2 \text{Var}_{K(2,\cdot)}(\boldsymbol{h})}{\mu_3^2 \text{Var}_{\boldsymbol{\rho}}(\boldsymbol{h})} = \mathcal{O}(\epsilon^{k_2}).$$

Note, using basic linear algebra the mutation variances from bins u_1, u_2, and u_3 are

$$\begin{aligned}
\sigma_1^2 &= \frac{1}{C^4}(\epsilon^r + \epsilon^s - \epsilon^{t+s})^2(\epsilon^{p+2q} + \epsilon^{2p+q} + \epsilon^{p+2s} + \epsilon^{q+2r} + \epsilon^{q+2s} + 2\epsilon^{p+q+r} + 2\epsilon^{q+r+s}) \\
&\quad - \frac{1}{C^4}(\epsilon^r + \epsilon^s - \epsilon^{t+s})^2(\epsilon^{2p+2s} + \epsilon^{2q+2r} + \epsilon^{2q+2s} + 2\epsilon^{p+q+2s} + 2\epsilon^{2q+r+s} + 2\epsilon^{p+q+r+s}), \\
\sigma_2^2 &= \frac{1}{C^4}(\epsilon^{q+t} + \epsilon^p)^2(\epsilon^{2p+s} + \epsilon^{2q+r} + \epsilon^{2q+s} + \epsilon^{r+2s} + \epsilon^{2r+s} + 2\epsilon^{p+q+s} + 2\epsilon^{p+r+s}) \\
&\quad - \frac{1}{C^4}(\epsilon^{q+t} + \epsilon^p)^2(\epsilon^{2p+2s} + \epsilon^{2q+2r} + \epsilon^{2q+2s} + 2\epsilon^{p+q+2s} + 2\epsilon^{2q+r+s} + 2\epsilon^{p+q+r+s}),
\end{aligned}$$

and

$$\sigma_3^2 = \frac{1}{C^4}\epsilon^t(1-\epsilon^t)(\epsilon^q - \epsilon^s)^2(\epsilon^{p+s} + \epsilon^{q+r} + \epsilon^{q+s})^2,$$

respectively. Recall C is the normalization of the steady-state distribution, $\boldsymbol{\mu}$. Since ϵ is a small parameter, we can approximate $C^4\sigma_1^2, C^4\sigma_2^2$, and $C^4\sigma_3^2$ by considering only the smallest powers on ϵ, which gives

$$\begin{aligned}
C^4\sigma_1^2 &\approx (x^r + x^s)^2(x^{p+2q} + x^{2p+q} + x^{p+2s} + x^{q+2r} + x^{q+2s}) \\
C^4\sigma_2^2 &\approx (x^{q+t} + x^p)^2(x^{2p+s} + x^{2q+r} + x^{2q+s} + x^{r+2s} + x^{2r+s})
\end{aligned}$$

and

$$C^4\sigma_3^2 \approx x^t(x^q - x^s)^2(x^{p+s} + x^{q+r} + x^{q+s})^2.$$

Now, we split each of the variance terms $C^4\sigma_1^2, C^4\sigma_2^2$, and $C^4\sigma_3^2$ into specific cases of the lowest power ϵ term when the powers p, q, r, s, and t satisfy certain inequalities. Through simple

comparisons of inequalities for $C^4\sigma_1^2$ this gives the following cases

$$C^4\sigma_1^2 = \begin{cases} \mathcal{O}(\epsilon^{2s+p+2q}) & r \geq s,\ \ q \leq p,\ \ q \leq s,\ \ p+q \leq 2r, \text{ and } p+q \leq 2s \\ \mathcal{O}(\epsilon^{2s+2p+q}) & r \geq s,\ \ p \leq q,\ \ p+q \leq 2s, \text{ and } p \leq s \\ \mathcal{O}(\epsilon^{4s+p}) & r \geq s,\ \ s \leq q,\ \ 2s \leq p+q,\ \ p+2s \leq q+2r, \text{ and } p \leq q \\ \mathcal{O}(\epsilon^{4s+q}) & r \geq s,\ \ s \leq p,\ \ 2s \leq p+q, \text{ and } q \leq p \\ \mathcal{O}(\epsilon^{2r+p+2q}) & r \leq s,\ \ q \leq p,\ \ q \leq s,\ \ p+q \leq 2r, \text{ and } p+q \leq 2s \\ \mathcal{O}(\epsilon^{2r+2p+q}) & r \leq s,\ \ p \leq q,\ \ p+q \leq 2s,\ \ p \leq r, \text{ and } p \leq s \\ \mathcal{O}(\epsilon^{2r+p+2s}) & r \leq s,\ \ s \leq q,\ \ 2s \leq p+q,\ \ p+2s \leq q+2r, \text{ and } p \leq q \\ \mathcal{O}(\epsilon^{4r+q}) & r \leq s,\ \ r \leq p,\ \ 2r \leq p+q, \text{ and } q+2r \leq p+2s. \end{cases}$$

For $C^4\sigma_2^2$ we get the following cases

$$C^4\sigma_2^2 = \begin{cases} \mathcal{O}(\epsilon^{2q+2t+2q+r}) & p \geq q+t,\ \ 2q+r \leq 2p+s,\ \ r \leq s,\ \ q \leq s, \text{ and } 2q \leq r+s \\ \mathcal{O}(\epsilon^{2q+2t+2q+s}) & p \geq q+t,\ \ s \leq r, \text{ and } 2q \leq r+s \\ \mathcal{O}(\epsilon^{2q+2t+r+2s}) & p \geq q+t,\ \ r+s \leq 2p,\ \ s \leq q,\ \ r+s \leq 2q, \text{ and } s \leq r \\ \mathcal{O}(\epsilon^{2q+2t+2r+s}) & p \geq q+t,\ \ r \leq p,\ \ r \leq q,\ \ r+s \leq 2q, \text{ and } r \leq s \\ \mathcal{O}(\epsilon^{2p+2p+s}) & 2p+s \leq 2q+r,\ \ p \leq q,\ \ 2p \leq r+s, \text{ and } p \leq r \\ \mathcal{O}(\epsilon^{2p+2q+r}) & p \leq q+t,\ \ 2q+r \leq 2p+s,\ \ r \leq s,\ \ q \leq s, \text{ and } 2q \leq r+s \\ \mathcal{O}(\epsilon^{2p+2q+s}) & p \leq q+t,\ \ q \leq p,\ \ s \leq r, \text{ and } 2q \leq r+s \\ \mathcal{O}(\epsilon^{2p+r+2s}) & p \leq q+t,\ \ r+s \leq 2p,\ \ s \leq q,\ \ r+s \leq 2q, \text{ and } s \leq r \\ \mathcal{O}(\epsilon^{2p+2r+s}) & p \leq q+t,\ \ r \leq p,\ \ r \leq q,\ \ r+s \leq 2q, \text{ and } r \leq s. \end{cases}$$

Finally, for $C^4\sigma_3^2$ we get the following cases

$$C^4\sigma_3^2 = \begin{cases} \mathcal{O}(\epsilon^{t+2q+2p+2s}) & q < s,\ \ p+s \leq q+r, \text{ and } p \leq q \\ \mathcal{O}(\epsilon^{t+4q+2r}) & q < s,\ \ q+r \leq p+s, \text{ and } r \leq s \\ \mathcal{O}(\epsilon^{t+4q+2s}) & q < s,\ \ q \leq p, \text{ and } s \leq r \\ \mathcal{O}(\epsilon^{t+2p+4s}) & s < q,\ \ p+s \leq q+r, \text{ and } p \leq q \\ \mathcal{O}(\epsilon^{t+2q+2r+2s}) & s < q,\ \ q+r \leq p+s, \text{ and } r \leq s \\ \mathcal{O}(\epsilon^{t+2q+4s}) & s < q,\ \ q \leq p, \text{ and } s \leq r. \end{cases}$$

Next, combining cases of $C^4\sigma_1^2$ and $C^4\sigma_3^2$ with comparisons of inequalities and removing cases with contradicting inequalities we have

$$\frac{\sigma_1^2}{\sigma_3^2} = \begin{cases} \mathcal{O}(\epsilon^{p-t-2q}) & r \geq s,\ \ q < s,\ \ q \leq p, \quad \text{and} \quad p+q \leq 2s \\ \mathcal{O}(\epsilon^{2s-t-3q}) & r \geq s,\ \ q < s,\ \ q \leq p,\ \ s \leq p, \text{ and } 2s \leq p+q \\ \mathcal{O}(\epsilon^{p-t-2q}) & r \leq s,\ \ q < s,\ \ q \leq p, \quad \text{and} \quad p+q \leq 2r \\ \mathcal{O}(\epsilon^{2r-t-3q}) & r \leq s,\ \ q < s,\ \ r \leq p, \text{ and } 2r \leq p+q. \end{cases}$$

Note that cases where $\sigma_1^2 > \sigma_3^2$ is enforced through the inequalities on powers p, q, r, s, and t are not considered as then (3.9) cannot hold.

Similarly, by combining cases of $C^4\sigma_2^2$ and $C^4\sigma_3^2$ and removing cases with contradicting inequalities we have

$$\frac{\sigma_2^2}{\sigma_3^2} = \begin{cases} \mathcal{O}(\epsilon^{t-r}) & p \geq q+t,\ \ q<s,\ \ r\leq s,\ \ 2q+r\leq 2p+s,\ \ 2q\leq r+s, \text{ and } q+r\leq p+s \\ \mathcal{O}(\epsilon^{t-s}) & p \geq q+t,\ \ q<s,\ \ s\leq r,\ \ 2q\leq r+s, \text{ and } q\leq p \\ \mathcal{O}(\epsilon^{t+2q-3s}) & p \geq q+t,\ \ s<q,\ \ s\leq r,\ \ 2q\leq r+s, \text{ and } q\leq p \\ \mathcal{O}(\epsilon^{t+r-2s}) & p \geq q+t,\ \ s<q,\ \ s\leq r,\ \ r+s\leq 2p, \text{ and } r+s\leq 2q \\ \mathcal{O}(\epsilon^{t+s-2q}) & p \geq q+t,\ \ q<s,\ \ r\leq s,\ \ r\leq q,\ \ r\leq p, \text{ and } r+s\leq 2q \\ \mathcal{O}(\epsilon^{t-s}) & p \geq q+t,\ \ s<q,\ \ r\leq s,\ \ r\leq p, \text{ and } r\leq q \\ \mathcal{O}(\epsilon^{2p-t-3s}) & p \leq q,\ \ s<q,\ \ p\leq r,\ \ 2p+s\leq 2q+r \text{ and } 2p\leq r+s \\ \mathcal{O}(\epsilon^{2p-t-2q-r}) & p \leq q+t,\ \ q<s,\ \ r\leq s,\ \ 2q+r\leq 2p+s,\ \ 2q\leq r+s, \text{ and } q+r\leq p+s \\ \mathcal{O}(\epsilon^{2p-t-2q-s}) & p \leq q+t,\ \ q<s,\ \ s\leq r,\ \ q\leq p, \text{ and } 2q\leq r+s \\ \mathcal{O}(\epsilon^{2p-t-3s}) & p \leq q+t,\ \ s<q,\ \ s\leq r,\ \ q\leq p, \text{ and } 2q\leq r+s \\ \mathcal{O}(\epsilon^{2p+r-t-2q-2s}) & p \leq q+t,\ \ s<q,\ \ s\leq r,\ \ r+s\leq 2p,\ \ r+s\leq 2q, \text{ and } q\leq p \\ \mathcal{O}(\epsilon^{2p+s-t-4q}) & p \leq q+t,\ \ q<s,\ \ r\leq s,\ \ r\leq p,\ \ r\leq q, \text{ and } r+s\leq 2q \\ \mathcal{O}(\epsilon^{2p-t-2q-s}) & p \leq q+t,\ \ s<q,\ \ r\leq s,\ \ r\leq p,\ \ r\leq q, \text{ and } q+r\leq p+s. \end{cases}$$

Again that cases where $\sigma_2^2 > \sigma_3^2$ is enforced through the inequalities on powers p, q, r, s, and t are not considered as then (3.9) will not hold. Finally, we compare cases of σ_1^2/σ_3^2 and σ_2^2/σ_3^2 to ensure, with the corresponding inequality requirements on powers p, q, r, s, and t, that $\sigma_1^2/\sigma_3^2 = O(\epsilon^{k_1})$ and $\sigma_2^2/\sigma_3^2 = O(\epsilon^{k_2})$ for $k_1, k_2 \geq 0$. These final comparisons produces the four cases of inequalities on powers p, q, r, s, and t given in equation (3.12)