ANALYSIS AND OPTIMIZATION OF WEIGHTED ENSEMBLE SAMPLING

DAVID ARISTOFF†

Abstract. We give a mathematical framework for weighted ensemble (WE) sampling, a binning and resampling technique for efficiently computing probabilities in molecular dynamics. We prove that WE sampling is unbiased in a very general setting that includes adaptive binning. We show that when WE is used for stationary calculations in tandem with a Markov state Model (MSM), the MSM can be used to optimize the allocation of replicas in the bins.

Key words. Molecular dynamics, Markov chains, stationary distributions, long time dynamics, coarse graining, resampling, weighted ensemble

AMS subject classifications. 65C05, 65C20, 65C40, 65Y05, 82C80

1. Introduction. In this article we study a type of resampling procedure for Markov chains. The procedure, called weighted ensemble (WE) sampling [4, 5, 10, 15, 16, 20], simulates some replicas of a Markov chain \( Y_t \), resampling from the replicas at certain time intervals. The distinguishing feature of WE sampling, compared to other resampling techniques, is that WE is designed so that the replicas of \( Y_t \) are evenly distributed throughout state space. This is achieved by dividing state space into bins and resampling in each bin so that the number of replicas in a bin remains roughly constant. The replicas carry probabilistic weights so that the resulting statistical distribution is unbiased. This distribution can be used, in principle, to estimate any function of \( Y_t \) at a fixed time [20].

Our interest in WE arises from longstanding problems in computational chemistry. In this setting, \( Y_t \) is a time discretization of some stochastic molecular dynamics (MD). MD simulations have proven useful for understanding many chemical and biological processes; see [14] for an overview. However, such simulations are limited by time scale separation. Many phenomena of interest occur at the laboratory time scale of microseconds, while MD simulations have time steps that correspond to femtoseconds. In this case, straightforward MD simulations are not practical. Many methods exist for extending the time scale of MD simulations; we do not attempt to give a review of them here. However, we mention some related methods, including exact Milestoning [3, 7], NEUS [19], Trajectory Tilting [17], Transition Interface Sampling [18], Forward Flux Sampling [1], and Boxed Molecular Dynamics [8]. See for instance [2, 6] for review and comparison of these methods. We will compare WE to exact Milestoning in Section 6 below.

In this article we give a mathematical framework for WE based on martingale theory. The framework allows us to prove that WE is exact even when the bins and the number of replicas per bin are chosen adaptively, as noted earlier in [20]. While WE can be used with a broad range of stochastic processes, when the process is time homogenous and Markovian – as in many models of MD, such as Langevin dynamics – WE can be used to efficiently compute dynamical quantities like reaction rates without actually simulating the reactions [4, 16]. These computations rely on Hill’s relation [9], which we generalize in Section 6 below. From Hill’s relation, obtaining

*September 2016

Funding: This work was supported by the National Science Foundation via the award NSF-DMS-1522398.

†Colorado State University (aristoff@math.colostate.edu)
reaction rates and probabilities requires a calculation on the stationary distribution of a nonreversible process. To speed up this stationary calculation, WE is combined with a preconditioning step [5, 4] in which a Markov state model (MSM) [12, 13] is used to approximate the stationary distribution. This is sometimes called accelerated WE [5]. Accelerated WE begins with replicas evenly distributed in space, with weights chosen to match the stationary distribution of the MSM. The replicas are then allowed to relax according to their exact dynamics, with WE sampling ensuring that the replicas remain evenly distributed in state space. We show below that information from the MSM can be used to optimize the WE sampling in this relaxation step, in the sense that the variance in the appropriate stationary calculation is minimized. This optimization requires an adaptive number of replicas per bin, in contrast with traditional WE sampling. We show in a simple model that this adaptive sampling can be significantly better than traditional WE sampling.

This article is organized as follows. In Section 2, we introduce notation for the replica processes, and we make assumptions on WE sampling that we assume hold throughout. (We refer to the replica processes as “point and weight” processes, following familiar notation in the mathematical resampling literature.) We prove that WE is unbiased in a general setting that includes adaptive binning procedures. In Section 3, we introduce an explicit algorithm for WE sampling (Algorithm 2) and prove that it is unbiased (Proposition 3 below). We give an exact formula for the mean squared error of this sampling (Proposition 4). Then, using this error formula, we produce a heuristic near-optimal adaptive strategy. In Section 4, we propose a version of accelerated WE which uses this strategy (Algorithm 5). In Section 5, we implement the adaptive strategy on a simple model, and compare with traditional WE sampling as well as naive sampling. In Section 6, we show how to obtain certain dynamical averages, including mean first passage times and hitting probabilities, from the stationary distribution of a nonreversible Markov chain. All proofs are in Section 7.

2. Notation and assumptions. In this section we introduce a general mathematical framework for unbiased WE sampling. Throughout $X_p$ is a Markov chain taking values in a measurable space $(E_p, E_p)$, with transition kernels $K_p$. We write $E$ for the expectation operator of $X_p$. Below we consider sampling schemes for $X_p$ of the following type. Let $\{e^i_p\}_{i=1,\ldots,N_p}$ and $\{\omega^i_p\}_{i=1,\ldots,N_p}$ be a points and weights processes evolving by selection and mutation:

$$
\begin{align*}
\{\xi^i_p\}_{i=1,\ldots,N_p} & \xrightarrow{\text{selection}} \{\hat{\xi}^i_p\}_{i=1,\ldots,N_p} \xrightarrow{\text{mutation}} \{\hat{\xi}^i_{p+1}\}_{i=1,\ldots,N_{p+1}}, \\
\{\omega^i_p\}_{i=1,\ldots,N_p} & \xrightarrow{\text{selection}} \{\hat{\omega}^i_p\}_{i=1,\ldots,N_p} \xrightarrow{\text{mutation}} \{\hat{\omega}^i_{p+1}\}_{i=1,\ldots,N_{p+1}}.
\end{align*}
$$

Here, each $\xi^i_p \in E_p$ and $\omega^i_p \in \mathbb{R}^+$. We refer to $\{\xi^i_p\}_{i=1,\ldots,N_p}$ and $\{\omega^i_p\}_{i=1,\ldots,N_p}$ as the points and weights at time $p$. The expectation operator associated to the points and weights process will be denoted $E^{emp}$. Below we describe rules which guarantee the points and weights process is unbiased. Readers interested only in explicit algorithms can skip to the next section. Before describing this process in more detail, we need some notation. Throughout, $\sim$ indicates denote equality in law of random variables or processes. Consider the information generated by the process up to time $p$, that is, the $\sigma$-algebra generated by the points and weights up to time $p$, prior to the selection step:

$$
\mathcal{F}_p := \sigma \left( \left( \{ e^i_q, \omega^i_q \}_{i=1,\ldots,N_q} \right)_{0 \leq q \leq p} \right),
$$


Consider also the larger \( \sigma \)-algebra obtained when the information from the points and weights in the selection step at time \( p \) is included:

\[
\hat{F}_p := \sigma \left( \left\{ \xi_{q,i}^i, \omega_{q,i}^i \right\}_{i=1}^{N_q}, \left\{ \hat{\xi}_{q,i}^i, \hat{\omega}_{q,i}^i \right\}_{0 \leq q \leq p} \right).
\]

We assume the points and weights process obeys rules (A1)-(A5) below. The rules do not uniquely define the selection and mutation procedure, but they ensure the points and weights process has certain properties, like being unbiased for the law of \( X_p \); see Theorem 1 below.

We first consider the initialization of the process. The initial points and weights have the distribution of \( X_0 \) in the following sense:

(A1) The initial collection of points and weights satisfies \( N_0 > 0 \) w.p. 1 and

\[
\mathbb{E}^{emp} \left[ \sum_{i=1}^{N_0} \omega_0^i h(\xi_0^i) \right] = \mathbb{E}[h(X_0)].
\]

for every bounded measurable \( h : E_0 \to \mathbb{R} \).

For example, the initial condition could be an average of a fixed number of samples of \( X_0 \), that is, \( N_0 \) is deterministic and \( \omega_1^1 = \ldots = \omega_0^{N_0} = 1/N_0 \), and \( \xi_i^0 \sim X_0 \) for \( i = 1, \ldots, N_0 \). Notice we do not require that the \( \xi_i^0 \)'s are independent, so they can be generated by, for example, Markov chain Monte Carlo or other sequential samplers.

Next we consider the selection step. In WE sampling, points are divided according to “bins” that partition state space. We do not require bins here, but we assume that the collection of points \( \xi_p^i \) at time \( p \) can be divided into \( R_p \) different collections, and that these collections are identifiable at time \( p \). We make the same assumption for the selected points.

(A2) There are partitions \( I_p^1, \ldots, I_p^{R_p} \) of \( \{1, \ldots, N_p\} \) and \( \hat{I}_p^1, \ldots, \hat{I}_p^{R_p} \) of \( \{1, \ldots, \hat{N}_p\} \) that can be determined at time \( p \). That is, \( R_p \) and \( I_p^1, \ldots, I_p^{R_p} \) are \( F_p \)-measurable, and \( \hat{I}_p^1, \ldots, \hat{I}_p^{R_p} \) are \( \hat{F}_p \)-measurable.

Note that the partitions above need not be associated with a partition of state space. All that is required is a classification of the point indices into disjoint collections.

We assume that the selection is unbiased with respect to this classification:

(A3) The selection in unbiased with respect to the partition in (A2), that is,

\[
\mathbb{E}^{emp} \left[ \sum_{i \in I_p^j} \omega_p^i h_p(\xi_p^i) \bigg| F_p \right] = \sum_{i \in I_p^j} \omega_p^i h_p(\xi_p^i)
\]

for every bounded measurable \( h_p : E_p \to \mathbb{R} \).

We now turn to the mutation step. In this step, each selected point evolves to form a point of the next generation, with its weight remaining the same.

(A4) The number of mutated points is the same as the number of selected points:

\[
N_{p+1} = \hat{N}_p,
\]

and moreover the weights do not change upon mutation:

\[
\omega_{p+1}^i = \hat{\omega}_p^i, \quad i = 1, \ldots, N_{p+1}.
\]
Finally, we assume the mutations are independent and follow the law of $X_p$.

(A5) The mutations are obtained by independent draws from $K_p$:

$$
\xi_{p+1}^i \sim K_p(\hat{\xi}_p^i).
$$

At time $p$, the points and weights process defines an empirical distribution

$$
\eta_p = \sum_{i=1}^{N_p} \omega_p^i \delta_{\xi_p^i}.
$$

Consider the first time all the points have been killed,

$$
\tau_{\text{kill}} = \inf \{ p \geq 0 : N_p = 0 \}.
$$

We adopt the convention that

$$
\eta_p \equiv 0, \quad \text{if } p \geq \tau_{\text{kill}}.
$$

We will now show that $\eta_p$ is an unbiased approximation for the law of $X_p$. This allows us to compute unbiased estimates of functions of $X_p$ at a fixed time $p$. See the discussion after Algorithm 2 below.

**Theorem 1.** $\eta_n$ is an unbiased estimator of the law of $X_n$. That is,

$$
\mathbb{E}^{\text{emp}} \left[ \sum_{i=1}^{N_n} \omega_n^i f_n(\xi_n^i) \right] = \mathbb{E}^{\text{emp}}[\eta_n(f_n)] = \mathbb{E}[f_n(X_n)]
$$

for every bounded measurable $f_n : E_n \to \mathbb{R}$.

**3. Sampling.** We now present a WE algorithm that is consistent with the assumptions in Section 2. Compared to the general setting above, our algorithm below has two specific features. First, we use bins to define the partition in (A2) above. And second, the resampling in each bin is designed to produce points in the next generation that all have the same weight (see also [5]). This ensures that the weights do not usually vary by large amounts in small regions of state space.

**Algorithm 2.** Pick initial points $\xi_0^1, \ldots, \xi_0^{N_0}$ and weights $\omega_0^1, \ldots, \omega_0^{N_0}$ as in (A1). Then iterate:

Selection step:

1. Choose bins $B_p^1, \ldots, B_p^{R_p}$ forming a partition of $E_p$, and a target number of points $N_p^r$ for each bin $B_p^r$. The $B_p^r$’s and $N_p^r$’s must be deterministic functions of the points and weights up to and including the current time $p$.

2. For $r = 1, \ldots, R_p$, do the following. Set a target weight in bin $B_p^r$ as

$$
\bar{\omega}_p^r = \frac{\sum_{i \in B_p^r} \omega_p^i}{N_p^r}.
$$

For $i$ such that $\xi_p^i \in B_p^r$, let $C_p^{r,i}$ be a random variable with law

$$
C_p^{r,i} = \begin{cases} 
\lfloor \omega_p^i / \bar{\omega}_p^r \rfloor, & \text{w.p. } 1 - (\omega_p^i / \bar{\omega}_p^r - \lfloor \omega_p^i / \bar{\omega}_p^r \rfloor), \\
\lceil \omega_p^i / \bar{\omega}_p^r \rceil + 1, & \text{w.p. } \omega_p^i \bar{\omega}_p^r - \lfloor \omega_p^i / \bar{\omega}_p^r \rfloor
\end{cases}
$$

For $i$ such that $\xi_p^i \not\in B_p^r$, set $C_p^{r,i} = 0$.
Let the number of selected points equal to $\xi^i_p \in B^r_p$ be $C^{r,i}_p$:

$$\# \{ \xi^i_p : \hat{\xi}^i_p = \xi^i_p \in B^r_p \} = C^{r,i}_p.$$  

Assign the target weight $\bar{\omega}_p^r$ to all the selected points in bin $B^r_p$:

$$\hat{\omega}^i_p = \bar{\omega}_p^r, \quad \text{if } \hat{\xi}^i_p \in B^r_p.$$  

Let $\hat{N}_p$ be the total number of selected points:

$$\hat{N}_p = \sum_{r=1}^{R_p} \sum_{i : \hat{\xi}^i_p \in B^r_p} C^{r,i}_p.$$  

Mutation step:

3. Independently evolve the $\hat{\xi}^i_p$'s to obtain the points at time $p + 1$:

$$\xi^i_{p+1} \sim K_p(\hat{\xi}^i_p, \cdot), \quad i = 1, \ldots, \hat{N}_p.$$  

Set the new weights equal to the selected weights:

$$\omega^i_{p+1} = \hat{\omega}^i_p, \quad i = 1, \ldots, \hat{N}_p.$$  

Set the total number of new particles as $N_{p+1} = \hat{N}_p$. Then update $p \leftarrow p + 1$ and return to 1.

Some remarks are in order.

- Given the points and weights at time $p$, the procedure in Step 2 uniquely defines the law of the selected points and weights, up to index relabeling. From (1), each point $\xi^i_p$ in bin $B^r_p$ is selected on average $\omega^i_p / \bar{\omega}_p^r$ times, that is,

$$\mathbb{E} \left[ C^{r,i}_p \mid \mathcal{F}_p \right] = \frac{\omega^i_p}{\bar{\omega}_p^r},$$

so the expected number of selected points in bin $B^r_p$ is at time $p$ is

$$\mathbb{E} \left[ \sum_{i : \hat{\xi}^i_p \in B^r_p} C^{r,i}_p \mid \mathcal{F}_p \right] = \sum_{i : \hat{\xi}^i_p \in B^r_p} \frac{\omega^i_p}{\bar{\omega}_p^r} = N^r_p.$$  

Thus, the $N^r_p$ can be seen as parameters controlling the average number of points per bin. As we have seen in Theorem 1, these numbers may be chosen adaptively. Indeed, a key feature of Algorithm 2 is that the bins $B^r_p$ and the numbers $N^r_p$ may be chosen adaptively.

- The algorithm generates empirical approximations

$$\eta_p = \sum_{i=1}^{N_p} \omega^i_p \delta_{\xi^i_p}$$

of the distribution of $X_p$. In Proposition 3 below, we show these empirical approximations are unbiased in the sense of Theorem 1.
• Recall that \( \tau_{\text{kill}} \) is the first time that all the points have been killed. If \( n \) is such that \( \mathbb{P}[\tau_{\text{kill}} \leq n] \) is too large, then the variance of the estimator \( \eta_n(f_n) \) of \( \mathbb{E}[f_n(X_n)] \) can be large as well. This is because when \( \mathbb{P}[\tau_{\text{kill}} \leq n] \) is large, in many simulations the value of \( \eta_n(f_n) \) will be 0, which is uninformative. In practice, to avoid this, the values of \( N^r_p \) can be given some lower threshold. See the remarks after Algorithm 5 below.

• In practice, the points may not be resampled at every time step. We have in mind the case where \( X_p \) is obtained from an underlying Markov chain at certain resampling times. That is, \( X_p = Y_{\tau_p} \) where \( Y_t \) is an underlying Markov chain and \( \tau_p \) are some resampling times. The \( \tau_p \) can be deterministic times – for instance \( \tau_p = p \delta t \), where \( \delta t \) is a constant lag time – or they can be certain stopping times for \( Y_t \). For stationary calculations based on Hill’s relation, the resampling times should be chosen so that \( X_p \) is a time homogeneous Markov chain. See the discussion in Section 6 below.

As noted above, Algorithm 2 produces unbiased estimates of the law of \( X_n \):

**Proposition 3.** Algorithm 2 satisfies (A1)-(A5) above when

\[
I_p^r := \{ i : \xi^i_p \in B_p^r \}, \quad \hat{I}_p^r := \{ i : \hat{\xi}_p^i \in B_p^r \}.
\]

In particular, \( \eta_n \) is an unbiased estimator of the law of \( X_n \) in the sense of Theorem 1.

In Proposition 4 below we present an explicit formula for the mean squared error in the estimate \( \eta_n(f_n) \approx \mathbb{E}[f_n(X_n)] \) when \( \eta_n \) is obtained from Algorithm 2. Below Proposition 4, we use the error formula to present an adaptive sampling strategy. This strategy uses some information (for instance, a MSM) to guide the sampling.

**Proposition 4.** Let \( n \) be a fixed time and \( f_n : E_n \to \mathbb{R} \) a bounded measurable function. In Algorithm 2, the mean squared error in the approximation \( \eta_n(f_n) \) of \( \mathbb{E}[f_n(X_n)] \) satisfies

\[
\mathbb{E}^{\text{emp}} \left[ (\eta_n(f_n) - \mathbb{E}[f_n(X_n)])^2 \right] = \mathbb{E}^{\text{emp}}[\eta_0(f_0) - \mathbb{E}[f_n(X_n)])^2]
+ \sum_{p=1}^n \mathbb{E}^{\text{emp}} \left[ (\eta_p(f_p) - \eta_{p-1}(f_{p-1}))^2 \right]
\]

(2)

where, writing \( \sigma_{p,i} = \mathbb{E}^{\text{emp}} \left[ (C_{p,i}^r)^2 \right] \mathcal{F}_p \) and \( g_p = K_pK_{p+1}\ldots K_{n-1}f_n \), we have

\[
\mathbb{E}^{\text{emp}} \left[ (\eta_p(f_{p+1}) - \eta_p(f_p))^2 \right] \mathcal{F}_p
= \sum_{r=1}^{R_p} \sum_{i : \xi^i_p \in B_p^r} \left[ \hat{\omega}_p^{r,i}K_{p+1}(\xi^i_p) \left( \hat{\omega}_p^{r,i} + (\omega_p^{r,i})^2 \right)^2 \sigma_{p,i} \right] g_p(\xi^i_p) \right]^2.
\]

(3)

Given the points and weights at time \( p \), we may try to minimize the contributions to the mean squared error in (3) over choices of \( B_p^r \) and \( N^r_p \). We pursue this idea, for fixed bins but an adaptive target number of points per bin. Suppose for given bins \( B_p^r \) we want to optimize over the allocation \( N^r_p \) of points, subject to the constraint that the average total number of points is fixed at \( N \).

For the remainder of this section, we assume that \( X_p \) is time homogeneous with transition kernel \( K_p \equiv K \), that the bins \( B_p^r \equiv B^r \) do not change in time, with \( R_p \equiv R \).
total bins, and that $f_n \equiv f$. Using the mean squared error formula in Proposition 4, we heuristically derive an adaptive sampling strategy below. Notice that $C_{r,i}^{r,i}$ are defined so that they have minimal variance. We estimate

$$
\sigma_{r,i}^{r,i} \equiv \mathbb{E}^{\text{emp}} \left[ (C_{r,i}^{r,i})^2 | \mathcal{F}_p \right] \approx \mathbb{E}^{\text{emp}} \left[ C_{r,i}^{r,i} | \mathcal{F}_p \right]^2 = \left( \frac{\omega_i^j}{\bar{\omega}_r^p} \right)^2.
$$

Suppose we have approximations $v_r^p$ of $Kg_{p+1}^2 - g_p^2$ inside $B^r$, that is,

$$
v_r^p \approx Kg_{p+1}^2(\xi) - g_p^2(\xi), \quad \text{for } \xi \in B^r.
$$

In this case, (3) rewrites as

$$
\mathbb{E}^{\text{emp}} \left[ (\eta_p(f_{p+1}) - \eta_p(f_p))^2 | \mathcal{F}_p \right]
= \sum_{r=1}^{R_p} \sum_{i: \xi_i^p \in B^r_p} \left[ \bar{\omega}_r^p \omega_i^p Kg_{p+1}^2(\xi_i^p) - (\omega_i^j)^2 \right. 
\left. + (\bar{\omega}_r^p)^2 \sigma_{r,i}^{r,i} \right] g_p(\xi_i^p)^2
\approx \sum_{r=1}^{R_p} \sum_{i: \xi_i^p \in B^r_p} \bar{\omega}_r^p \left[ \omega_i^j Kg_{p+1}^2(\xi_i^p) - \omega_i^j g_p(\xi_i^p)^2 \right]
\approx \sum_{r=1}^{R_p} \bar{\omega}_r^p v_r^p \sum_{i: \xi_i^p \in B^r_p} \omega_i^j
\approx \sum_{r=1}^{R_p} v_r^p \left( \sum_{i: \xi_i^p \in B^r_p} \omega_i^j \right)^2
\approx R_p \sum_{r=1}^{R} N_r^p \left( \sum_{i: \xi_i^p \in B^r_p} \omega_i^j \right)^2.
$$

Minimizing the last expression with the constraint

$$
\sum_{r=1}^{R} N_r^p = N
$$

leads to the formula

$$
N_r^p \approx \frac{N \sqrt{v_p^r \sum_{i: \xi_i^p \in B^r_p} \omega_i^j}}{\sum_{r=1}^{R} \sqrt{v_p^r \sum_{i: \xi_i^p \in B^r_p} \omega_i^j}}.
$$

The approximations $v_r^p$ can be obtained, for instance, by estimating a coarse probability $P_{rs}$ of $X_p$ going from bin $B^r$ to bin $B^s$ in one time step. This essentially amounts to constructing a MSM [12, 13] for $X_p$, with each state corresponding to a bin, and with time step 1. Of course, as with any MSM, there is error arising from the Markov assumption, since the probability $P_{rs}$ for $X_p$ to transition between bins $B^r$ and $B^s$ depends on the initial distribution of $X_p$ in $B^r$. However, the MSM can be seen as a coarse model for $X_p$, and in particular, when $X_p$ has stationary distribution $\pi$, we will use the stationary distribution of the MSM as an initial guess of $\pi$.

Some comments on the term $Kg_{p+1}^2 - g_p^2$ are in order. It is not hard to see that

$$
Kg_{p+1}^2(\xi) - g_p^2(\xi) = \text{Var} \mathbb{E}^{X_1} [f(X_{n-p-1})|X_0 = \xi].
$$
The heuristic strategy suggests the larger the value of \(Kg^2_{p+1}(\xi) - g^2_p(\xi)\), the more beneficial it is to select \(\xi\). It is also easy to check that for naive sampling — that is, sampling with no selection step — the contributions to the mean squared error are

\[
\mathbb{E}^{\text{emp}} \left[ (\eta_p(f_{p+1}) - \eta_p(f_p))^2 \right] = N \sum_{i=1}^{N} \left[ Kg^2_{p+1}(\xi_i^p) - g^2_p(\xi_i^p) \right],
\]

with \(N\) the (constant) total number of points.

We explore the heuristic strategy above in Section 4 below, showing how it can be combined with existing WE technologies to efficiently sample the stationary distribution of \(X_p\) in the time homogeneous case.

4. Stationary averages. In this section we present a procedure, Algorithm 5 below, for computing stationary averages in the time homogeneous case. In this section, \(X_p\) is a time homogeneous Markov chain with values in a measurable space \((E, \mathcal{E})\), and transition kernel \(K\). We assume \(X_p\) has a unique stationary distribution \(\pi\).

Algorithm 5. Choose a relaxation time \(n\), a sampling measure \(\xi\) on \((E, \mathcal{E})\), and bins \(B^1, \ldots, B^R\) forming a partition of \(E\). Choose a target number of total points \(N \equiv N_0\). Then iterate:

Preconditioning step:
(i) Let \(P_{rs}\) estimate the probability for \(X_p\) to go from \(B^r\) to \(B^s\) in one step:

\[
P_{rs} = \xi(B^r) \int_{B^s} \mathbb{P}[X_{p+1} \in B^s | X_p = x] \xi(dx).
\]

Let \(u_r\) estimate the value of \(f\) inside bin \(B^r\):

\[
u_r = \xi(B^r) \int_{B^r} f(x) \xi(dx).
\]

(ii) Let \(\mu\) be the stationary distribution for the transition matrix \(P = (P_{rs})\). That is, \(\mu = (\mu_r)\) is the left eigenvector of \(P\) corresponding to the eigenvalue 1, normalized so that \(\sum_{r=1}^{R} \mu_r = 1\).

(iii) Let \(v^p_r\) be the \(r\)th entry of the vector \(P(P^{n-p-1}u)^2 - (P^{n-p}u)^2\), where \(u = (u_r)\) is the vector estimate of \(f\) and the squaring operations are entrywise.

Relaxation step:
(i) Choose initial points so that the number of points in \(B^r\) is \(\approx N/R\):

\[
\# \{i : \xi_0^i \in B^r\} \approx N/R.
\]

Choose initial weights so that the total weight inside each bin \(B^r\) is \(\mu_r\):

\[
\omega_0^i = \frac{\mu_r}{\# \{i : \xi_0^i \in B^r\}} \text{, if } \xi_0^i \in B^r.
\]

(ii) Proceed through Algorithm 2 from time \(p = 0\) to \(p = n\), with either

\[
N^p = \frac{(N - \tilde{N}R) \sqrt{v^p_r} \sum_{i : \xi_0^i \in B^r} \omega^i_p}{\sum_{r=1}^{R} \sqrt{v^p_r} \sum_{i : \xi_0^i \in B^r} \omega^i_p} + \tilde{N}
\]
or

\[ N^r_p \equiv N^r = N/R \]

defining the target number of points in bin \( B^r \) at time \( p \).

(iii) The output of the algorithm is \( \eta_n(f) = \sum_{i=1}^{N_p} \omega_i f(\xi_n^i) \).

Some remarks are in order.

- The purpose of the preconditioning step is to obtain a coarse approximation \( \mu \) of \( \pi \), the stationary distribution of \( X_p \). This approximation is then used as the initial condition of the relaxation step. Note that \( P_{rs} \) and \( u_r \) can be estimated by running many short trajectories in parallel. The sampling measure \( \xi \) is the initial distribution for these short trajectories.
- \( P_{rs} \) is essentially the transition matrix of a MSM for \( X_p \) defined so that its states are the bins \( B^1, \ldots, B^r \), and its time step is exactly 1. If \( X_p \) is obtained from an underlying Markov chain \( Y_t \) via \( X_p = Y_{p \delta t} \), then this can be viewed as an MSM for \( Y_t \) with time step \( \delta t \). Under suitable conditions, the vector \( \mu \) is the unique stationary distribution of the MSM. To understand the expression for \( v^r_p \), note that since \( g_p = K^{n-r}f \),

\[ Kg^2_{p+1} - g^2_p = K(K^{n-r-1}f)^2 - (K^{n-r}f)^2. \]

Thus, \( v^r_p \) is defined using (4), where the approximation is obtained by replacing \( K \) with \( P \) and \( f \) with \( u \) on the right hand side of that equation. In particular, the definition of \( v^r_p \) is consistent with the heuristic calculations in Section 3.
- The purpose of the relaxation step is to allow the coarse approximation \( \mu \) to relax to the stationary distribution \( \pi \). When the relaxation time \( n \) is appropriately chosen, \( \eta_n(f) \) is a good approximation of \( \pi(f) \). Since the relaxation step starts at the coarse approximation \( \mu \) of \( \pi \), the relaxation time \( n \) can be much smaller than the typical time to converge from an arbitrary initial distribution.
- When \( N^r = N/R \) as in (7), we recover the accelerated weighted ensemble technique of [4, 5]. On the other hand, when \( N^r \) satisfies (6) the sampling is adaptive. For comparisons of sampling with (6) and (7), see the simulations in Section 5 below. In the adaptive case, the choice of \( N^r_p \) depends on the parameter \( n \). Thus, a simulation at a given value of \( n \) cannot be easily extended to a larger value of \( n \) with the same formula (6) defining \( N^r_p \). Also, in the adaptive case the choice of \( N^r_p \) depends on \( f \), and so separate simulations are likely needed for computing \( E[f(X_n)] \) for different functions \( f \).
- The \( \tilde{N} \) is introduced so that \( \tau_{kill} > n \) with high probability. When \( \tilde{N} \) is too small, it can happen that all the points are killed in a large fraction of simulations. In this situation, the variance can be large. This killing can occur if, for example, \( N^r_p \) is nonzero only in regions where there are no points. On the other hand, if \( \tilde{N} \) is too large, we can lose the benefits of the adaptive sampling.
5. Numerical example. Here we consider an underlying Markov chain $Y_t$ on state space \{1, 2, \ldots, 90\} with transition matrix

$$Q(i, i + 1) = \frac{2}{5} + \frac{m(i)}{5}$$

$$Q(i, i - 1) = \frac{2}{5} + \frac{m(i)}{5},$$

where

$$m(x) := \cos \left( \frac{6\pi x}{90} \right)$$

and $Q(i, i)$ is chosen so that $Q$ is stochastic. We take resampling intervals $\delta t = 4$, so

$$X_p := Y_{4p}$$

and the transition matrix of $X_p$ is $K = Q^4$. The bins will be

$$B'_p \equiv B' = \{3r - 2, 3r - 1, 3r\}, \quad r = 1, \ldots, 30.$$  

Thus, there are $R = 30$ fixed bins. Let $\pi$ be the stationary distribution of $X_p$, and

$$f(i) = \begin{cases} 1, & 28 \leq i \leq 33 \\ 0, & \text{else} \end{cases}$$

We also let $\bar{f} = f/6$ be its normalized version. We use Algorithm 5 to obtain empirical approximations $\eta_n(f)$ of $\pi(f)$ for a fixed relaxation time $n$. Throughout, we use a sampling measure $\xi$ which is uniform: $\xi(i) = 1/90$ for all $i$. In the relaxation step of Algorithm 5, the initial empirical distribution of points and weights can be considered a sample from

$$\nu_0(i) = \frac{\mu_r}{3}, \quad \text{if } i \in B'_r.$$  

We will use three types of sampling. The first type of sampling uses the adaptive allocation of points from equation (6) above. We call this adaptive WE sampling. In (6), we take $N = 150$ and $\tilde{N} = 1$. We found this $\tilde{N}$ to be sufficiently large that $\tau_{kill} > n$ in all our simulations.

In the second type of sampling we used a fixed target number of points per bin. We call this traditional WE sampling. It is the sampling method described in [20, 4]. Here we use equation (7), taking $N'_p = N = 5$. Again, we had $\tau_{kill} > n$ in all our simulations. Notice for both adaptive and traditional WE sampling, we have chosen $N$ so that the target number of points is 150.

The last type of sampling does not use resampling at all. We call this naive sampling. Here, we skip steps (ii) and (iii) of the relaxation step of Algorithm 5. Instead, after the points $\xi_0^1, \ldots, \xi_0^N$ and weights $\omega_0^1, \ldots, \omega_0^N$ are chosen, we simply evolve these points independently until time $n$, without changing the weights. For this sampling we take $N = 150$ total points, to facilitate comparison with the traditional and adaptive sampling described above.

Results comparing the different relaxation techniques are in Figures 1-2. In Figure 1, we plot $\eta_n(f)$ vs. $n$ for various values of $n$, showing convergence to the
stationary value $\pi(f)$. We compute error bars as empirical standard deviations from 1000 independent simulations. We found that the error bars for adaptive WE sampling were significantly smaller than that of traditional WE and naive sampling. In Figure 2, we plot histograms representing the average distribution of the points $\xi_n$ at time $n$. Note that traditional WE sampling distributes the points roughly uniformly in space, as expected, while adaptive WE sampling guides the points towards the region in state space relevant for computing $f$. Meanwhile, naive sampling distributes the points according to the stationary distribution $\pi$. In Figure 3, we plot the estimates $v_p^\tau$ from the adaptive sampling strategy for $p = 0$ and $p = n - 1$ where $n = 30$ is the relaxation time. Note that by time $n - 1$, the sampling is focused near the support of $f$.

When $f$ is a function with large values in regions of low $\pi$ probability, as in this example, naive sampling performs poorly compared to both traditional and adaptive WE sampling. When state space is very large compared to the region $R$ where $f$ has large values (or is non-negligible), we expect adaptive WE sampling to perform much better than traditional WE sampling, due to the fact that traditional WE sampling will distribute the points very thinly throughout space, including in $R$, while adaptive WE sampling will push most of the points towards $R$. 

Fig. 1. Values of $\eta_n(f)$ vs. $n$ from the example in Section 5 from adaptive, traditional and naive sampling. The crosses are exact values corresponding to $n_0K^nf$, and the dotted line is the stationary value $\pi(f)$. Bottom right: empirical standard deviations $\sigma(\eta_n(f))$ of $\eta_n(f)$, obtained from 1000 independent simulations. (These are the error bar radii from the other plots.)
A possible drawback of adaptive WE sampling is that it requires more computations at the resampling times, compared to traditional WE sampling. However, in practice the resampling times may be large enough so that this extra effort contributes little to the overall computational cost.

Finally, we note that the adaptive sampling above can also be used more generally to estimate time marginals of $X_n$, that is, expectations of the form $\mathbb{E}[f(X_n)]$ at fixed finite times $n$, from an arbitrary initial distribution of $X_0$. In this case, a MSM is still required to guide the sampling. One of the advantages of the adaptive sampling in the stationary case is that an MSM has already been computed as part of a preconditioning step.

6. Computing dynamics from stationary averages. In this section we show how to compute certain dynamical averages of time homogeneous Markov chains from stationary calculations on a related Markov chain. The Hill relation [9] shows that a mean first passage time can be reformulated as a certain stationary average. We will generalize this relation below.

Assume that $X_p$ is obtained from an underlying Markov chain $Y_t$ at some resampling times $\tau_p$, that is, $X_p = Y_{\tau_p}$. We assume $X_p$ and $Y_t$ have values in a measurable space $(E, \mathcal{E})$. $Y_t$ can be continuous or discrete in time. For simplicity, and since in simulations time discretization is required anyway, we assume here it is discrete in time. To obtain $X_p$ from $Y_t$, we need to define the resampling times $\tau_p$. We will
assume they satisfy the following:

(B1) $0 = \tau_0 < \tau_1 < \tau_2 < \ldots$ are stopping times for $Y_t$ such that for all $p \geq 0$,

$$\text{Law} \left( (Y_{\tau_p+1})_{t \geq 0}, (\tau_{p+m} - \tau_p)_{m \geq 0} \mid X_{\tau_p} = x \right) = \text{Law} \left( (Y_t)_{t \geq 0}, (\tau_m)_{m \geq 0} \mid X_0 = x \right).$$

Informally, this means $Y_t$ and the stopping time increments $\tau_{p+1} - \tau_p$ regenerate at each stopping time $\tau_p$. For example, we can take $\tau_p$ to be $p$ multiplied by some deterministic lag time. Alternatively, the $\tau_p$ can be a sequence of hitting times of coarse sets in state space, as in exact Milestoning; see the comments below.

To be able to compute dynamical averages from stationary ones, we must introduce a source and sink. Let $I$ and $F$ be disjoint initial and final sets, such that when $Y_t$ arrives in $F$ at one of the resampling times, it immediately returns to $I$. When it returns to $I$, it starts at a fixed source distribution $\rho$ on $I$:

(B2) There are disjoint measurable sets $I, F \subset E$ and a distribution $\rho$ on $I$ so that

$$\mathbb{P} \left[ Y_{t+1} \in A \mid t = \tau_p, \text{ for some } p, Y_{\tau_p} \in F \right] = \int \mathbb{P}^x \left[ Y_1 \in A \right] \rho(dx).$$

We are interested in computing, for example, the mean first passage time of $Y_t$ from $I$ to $F$. We want to compute this from a stationary average, using the source-sink. The following technical assumption guarantees this computation is well defined:

(B3) With $\tau_F := \inf \{ \tau_p > 0 : Y_{\tau_p} \in F \}$ we have

$$\mathbb{E}^x[\tau_F] < \infty, \quad \text{for all } x \in E.$$

We assume (B1)-(B3) hold throughout the remainder of this section. From (B1), it is easy to see $\tau_F$ is a stopping time for $Y_t$. Since $X_p = Y_{\tau_p}$, by the strong Markov prop-
D. ARISTOFF

property, (B1) guarantees $X_p$ is a time homogeneous Markov chain. From (B2)-(B3), $X_p$ has a unique stationary distribution $\pi$. The latter can be proved by standard coupling arguments, based on $X_p$ reaching $F$ in finite expected time and then regenerating at $\rho$. We omit proof.

For the resampling times $\tau_p$, we could take $\tau_p = p\delta t$ where $\delta t$ is a fixed resampling time interval, as in standard WE sampling. More generally, we could take $\tau_0 = 0$ and $\tau_{p+1} = \tau_p + \tau(X_{\tau_p})$ where $\tau(x)$ is a stopping time for $Y_t$ starting at $Y_0 = x$, as in exact Milestoning. In that setting, $\tau(x)$ is the first time for $Y_t$ to reach a milestone starting at $Y_0 = x$; see [3].

We are now ready to generalize the Hill relation [9].

**Theorem 6.** With $\pi$ the unique stationary distribution of $X_p$, 

$$(8) \quad \mathbb{E}^\rho[g(Y_{\tau_F})] = \frac{\int_F g(y)\pi(dy)}{\pi(F)}$$

and

$$(9) \quad \mathbb{E}^\rho\left[\sum_{t=1}^{\tau_p} g(Y_t)\right] = \frac{\mathbb{E}^\pi\left[\sum_{t=1}^{\tau_1} g(Y_t)\right]}{\pi(F)}.$$ 

In (8), if $F = A \sqcup B$ is the disjoint union of $A$ and $B$ and $g = 1_B$, we have

$$\mathbb{P}^\rho[\tau_B < \tau_A] = \frac{\pi(B)}{\pi(A \sqcup B)}.$$ 

This is the probability to hit $B$ before $A$, starting at $\rho$. In (9), if $g \equiv 1$, 

$$\mathbb{E}^\rho[\tau_F] = \frac{\mathbb{E}^\pi[\tau_1]}{\pi(F)}.$$ 

This is the MFPT from $\rho$ to $F$. Note that when $\tau_1 = \delta t$ is constant, this equation is simply the Hill relation [9]. When $\tau_1$ is random, this is the expression recently proved for exact Milestoning [2, 3]. It is important to note that assumption (B2) does not affect the averages on the left hand side of (8) and (9). Thus, $Y_t$ can be any time homogeneous Markov chain, and in particular it need not have a source and sink. The source and sink in (B2) are only needed to define the stationary distribution $\pi$ used in the sampling.

Since here $\pi$ is the stationary distribution of a time homogeneous Markov chain, Algorithm 5 can be used to compute the right hand side of (8). See the example in Section 4 above. Although it is not obvious how to extend that algorithm to compute the right hand side of (9), in some cases a simple modification may suffice. For instance, in exact Milestoning, one can apply Algorithm 5 to the function $f = g1_M$, where $M$ is the union of the milestones. The empirical distributions $\eta_n$ from the algorithm can be used as the initial distribution for computing the numerator of (9).

7. Proofs. We begin with the following preliminary result.

**Lemma 7.** Suppose $\xi_1, \xi_2, \ldots$ are iid random variables and $N$ is a nonnegative integer valued random variable independent of the $\xi_i$’s. Then

$$\mathbb{E}\left[\sum_{i=1}^N \xi_i\right] = \mathbb{E}[N]\mathbb{E}[\xi_1]$$
and
\[
\mathbb{E} \left[ \left( \sum_{i=1}^{N} \xi_i \right)^2 \right] = (\mathbb{E}[N^2] - \mathbb{E}[N]) \mathbb{E}[\xi_1]^2 + \mathbb{E}[N] \mathbb{E}[\xi_1^2].
\]

We omit the proof, which is straightforward.

We will also use the following result.

**Lemma 8.** Let
\[
g_p = K_p K_{p+1} \ldots K_{n-1} f_n,
\]
where \( g_n := f_n \). Then we have
\[
\mathbb{E}^{emp} \left[ g_{p+1}(\xi_{p+1}) \right] = g_p(\hat{\xi}_p)
\]
and
\[
\mathbb{E}^{emp} \left[ g_{p+1}(\xi_p) g_{p+1}(\xi_{p+1}) \right] = \begin{cases} g_p(\hat{\xi}_p) g_p(\hat{\xi}_p), & i \neq j, \\ K_p g_{p+1}^2(\xi_p), & i = j. \end{cases}
\]

**Proof.** First observe that, by (A5),
\[
\mathbb{E}^{emp} \left[ g_{p+1}(\xi_{p+1}) \right] = \mathbb{E}^{emp} \left[ \int K_{p+1}(\xi_{p+1}, dy_{p+2}) \ldots K_{n-1}(y_{n-1}, dy_n) f(y_n) \left| \hat{F}_p \right. \right]
\]
\[
= \int K_p(\hat{\xi}_p, dy) K_{p+1}(y, dy_{p+2}) \ldots K_{n-1}(y_{n-1}, dy_n) f(y_n) = g_p(\hat{\xi}_p).
\]

When \( i = j \), in the second equation, using (A5) we have
\[
\mathbb{E}^{emp} \left[ g_{p+1}(\xi_p) g_{p+1}(\xi_{p+1}) \right] = \int K_p(\hat{\xi}_p, dy) g_{p+1}(y)^2 = K_p g_{p+1}^2(\hat{\xi}_p).
\]

On the other hand, when \( i \neq j \), conditionally on \( \hat{F}_p \), \( g_{p+1}(\xi_{p+1}) \) and \( g_{p+1}(\xi_p) \) are independent. Thus, appealing to (10) completes the proof.

We now turn to the proof of Theorem 1.

**Proof of Theorem 1.** Let \( g_p \) be as in Lemma 8 and recall that
\[
\eta_p = \sum_{i=1}^{N_p} \omega_p^i \delta_{\xi_p^i}.
\]

To emphasize that \( \eta_p = 0 \) when \( p \geq \tau_{kill} \), we write
\[
\eta_p = \eta_{p, \tau_{kill} > p}.
\]

Let
\[
M_p = \eta_p(g_p) \mathbbm{1}_{\tau_{kill} > p} = \mathbbm{1}_{\tau_{kill} > p} \sum_{i=1}^{N_p} \omega_p^i g_p(\xi_p^i).
\]
Note that $1_{\tau_{kill} > p}$ is $F_p$-measurable and

$$1_{\tau_{kill} > p + 1} = 1_{\tau_{kill} > p} 1_{N_{p+1} > 0}.$$ 

Using also

$$1_{N_{p+1} > 0} = 1 - 1_{N_{p+1} = 0}$$

we find that

$$E_{emp}[M_{p+1} | F_p] = E_{emp}[1_{\tau_{kill} > p + 1} \sum_{i=1}^{N_{p+1}} \omega_{p+1} g_{p+1}(\xi^i_{p+1}) | F_p]$$

$$= 1_{\tau_{kill} > p} E_{emp} \left[ E_{emp} \left[ \sum_{i=1}^{N_{p+1}} \omega_{p+1} g_{p+1}(\xi^i_{p+1}) | \tilde{F}_p \right] | F_p \right]$$

$$= 1_{\tau_{kill} > p} E_{emp} \left[ \sum_{i=1}^{N_p} \tilde{\omega}_i g_p(\tilde{\xi}_i^p) | F_p \right]$$

by (A4)

$$= 1_{\tau_{kill} > p} E_{emp} \left[ \sum_{i \in I_p} \tilde{\omega}_i g_p(\tilde{\xi}_i^p) | F_p \right]$$

by (A5) and Lemma 8

$$= 1_{\tau_{kill} > p} \sum_{r=1}^{R_p} \sum_{i \in I_p} \tilde{\omega}_i g_p(\tilde{\xi}_i^p)$$

by (A3)

$$= 1_{\tau_{kill} > p} \sum_{i=1}^{N_p} \omega_i g_p(\xi^i_p) = M_p$$

by (A2).

It follows that $(M_p)_{0 \leq p \leq n}$ is a martingale, so $E_{emp}[M_0] = E_{emp}[M_n]$ and

$$E[f_n(X_n)] = E[g_0(X_0)]$$

$$= E_{emp}[M_0]$$

by (A1)

$$= E_{emp}[M_n] = E_{emp}[\eta_n(f_n) 1_{\tau_{kill} > n}].$$

\[\square\]

**Proof of Proposition 3.** Clearly, Algorithm 2 satisfies (A1)-(A2) and (A4)-(A5). Conditional on $F_p$, for $j$ such that $\tilde{\xi}_j^p = \xi_i^p$, the $h_p(\tilde{\xi}_j^p)$'s are independent with expected value $h_p(\xi_i^p)$. Conditional on $F_p$, the expected number of such $j$'s is

$$E_{emp}[\# \{ j : \tilde{\xi}_j^p = \xi_i^p \} | F_p] = E_{emp}[C^{r,i} | F_p] = \frac{\omega^i_j}{\tilde{\omega}_p}.$$
Thus, for any bounded measurable \( h_p : E_p \to \mathbb{R} \),

\[
\mathbb{E}^{emp}\left[ \sum_{i \in I_p^e} \omega_i^p h_p(\xi_i^p) \bigg| \mathcal{F}_p \right] = \mathbb{E}^{emp}\left[ \sum_{i \in I_p^e} \omega_i^p h_p(\xi_i^p) \bigg| \mathcal{F}_p \right] = \mathbb{E}^{emp}\left[ \sum_{i \in I_p^e} \omega_i^p h_p(\xi_i^p) \bigg| \mathcal{F}_p \right]
\]

\[
= \sum_{i \in I_p^e} \mathbb{E}^{emp}\left[ \omega_i^p h_p(\xi_i^p) \bigg| \mathcal{F}_p \right] = \sum_{i \in I_p^e} \omega_i^p h_p(\xi_i^p)
\]

by Lemma 7.

**Proof of Proposition 4.** Define \( g_p = K_p K_{p+1} \ldots K_{n-1} f_n \) and

\[
M_p = \eta_p(\xi_p) \mathbb{1}_{\tau_{k,i} > p} = \mathbb{1}_{\tau_{k,i} > p} \sum_{i=1}^{N_n} \omega_i^p g_p(\xi_i^p)
\]

as in the proof of Theorem 1. There it was shown \((M_p)_{0 \leq p \leq n}\) is a martingale, which leads to equation 2. We turn to the expression in (3). Using

\[
\mathbb{1}_{\tau_{k,i} > p+1} = \mathbb{1}_{\tau_{k,i} > p} \mathbb{1}_{N_{p+1} > 0}
\]

we find that

\[
\mathbb{E}^{emp}[M_{p+1}^2 | \mathcal{F}_p]
\]

\[
= \mathbb{E}^{emp}\left[ \mathbb{1}_{\tau_{k,i} > p+1} \left( \sum_{i=1}^{N_{p+1}} \omega_i^{p+1} g_{p+1}(\xi_i^{p+1}) \right)^2 \bigg| \mathcal{F}_p \right]
\]

\[
= \mathbb{1}_{\tau_{k,i} > p} \mathbb{E}^{emp}\left[ \sum_{i=1}^{N_{p+1}} \omega_i^{p+1} g_{p+1}(\xi_i^{p+1}) \bigg| \mathcal{F}_p \right]
\]

\[
= \mathbb{1}_{\tau_{k,i} > p} \sum_{r=1}^{R_p} \sum_{s=1}^{R_p} \mathbb{E}^{emp}\left[ \sum_{i \in I_p^r} \omega_i^p g_{p+1}(\xi_i^p) \bigg| \mathcal{F}_p \right]
\]

\[
= \mathbb{1}_{\tau_{k,i} > p} \sum_{r=1}^{R_p} \sum_{s=1}^{R_p} \mathbb{E}^{emp}\left[ \sum_{i \in I_p^r} \omega_i^p g_{p+1}(\xi_i^p) \bigg| \mathcal{F}_p \right]
\]

From Lemma 8,

\[
\mathbb{E}^{emp}\left[ g_{p+1}(\xi_i^{p+1}) g_{p+1}(\xi_j^{p+1}) \bigg| \mathcal{F}_p \right] = \begin{cases} 
  g_p(\xi_i^{k}) g_p(\xi_j^{\ell}), & k \neq \ell \\
  K_p g_p^2(\xi_i^{k}), & k = \ell
\end{cases}
\]

and recall that

\[
\mathbb{E}^{emp}\left[ C_{p,i}^r | \mathcal{F}_p \right] = \mathbb{E}^{emp}\left[ \# \{ k : \xi_k^p = \xi_i^p \} \bigg| \mathcal{F}_p \right] = \frac{\omega_i^p}{\omega_p^r}.
\]
Thus, for $i \neq j$ such that $\xi^i_p \in B^r_p$ and $\xi^j_p \in B^s_p$, by Lemma 7,

$$E_{\text{emp}} \left[ \sum_{k: \xi^k_p = \xi^i_p} \sum_{\ell: \xi^\ell_p = \xi^j_p} g_{p+1}(\xi^k_p)g_{p+1}(\xi^\ell_p) \mid F_p \right] = \frac{\omega^i_p \omega^j_p}{\bar{\omega}^r_p} g_p(\xi^i_p)g_p(\xi^j_p).$$

On the other hand, when $i = j$ and $\xi^i_p \in B_p^r = B_p^s$, this becomes

$$E_{\text{emp}} \left[ \sum_{k: \xi^k_p = \xi^i_p} \sum_{\ell: \xi^\ell_p = \xi^i_p} g_{p+1}(\xi^k_p)g_{p+1}(\xi^\ell_p) \mid F_p \right] = \left( \sigma^r_{p,i} - \frac{\omega^i_p}{\bar{\omega}^r_p} \right) g_p(\xi^i_p)^2 + \frac{\omega^i_p}{\bar{\omega}^r_p} K_p g_{p+1}(\xi^i_p).$$

where again we used Lemma 7, and we recall that

$$\sigma^r_{p,i} = E_{\text{emp}} \left[ (C^r_{p,i})^2 \mid F_p \right].$$

Combining the above,

$$E_{\text{emp}} [M^2_{p+1} \mid F_p] = \mathbb{1}_{r_{k,i,l} > p} \sum_{r=1}^{R_p} \sum_{s=1}^{R_p} \sum_{i: \xi^i_p \in B^r_p \atop j: \xi^j_p \in B^s_p} \omega^i_p \omega^j_p g_p(\xi^i_p)g_p(\xi^j_p)$$

$$+ \mathbb{1}_{r_{k,i,l} > p} \sum_{r=1}^{R_p} \sum_{i: \xi^i_p \in B^r_p} \left[ \left( \bar{\omega}^r_p \sigma^r_{p,i} - \bar{\omega}^r_p \omega^i_p \right) g_p(\xi^i_p)^2 + \bar{\omega}^r_p \omega^i_p K_p g_{p+1}(\xi^i_p) \right],$$

and so, since

$$M^2_p = \left( \mathbb{1}_{r_{k,i,l} > p} \sum_{r=1}^{R_p} \sum_{i: \xi^i_p \in B^r_p} \omega^i_p g_p(\xi^i_p) \right)^2$$

$$= \mathbb{1}_{r_{k,i,l} > p} \sum_{r=1}^{R_p} \sum_{i: \xi^i_p \in B^r_p} \sum_{j: \xi^j_p \in B^s_p} \omega^i_p \omega^j_p g_p(\xi^i_p)g_p(\xi^j_p),$$

we get

$$E_{\text{emp}} \left[ (M_{p+1} - M_p)^2 \mid F_p \right]$$

$$= E_{\text{emp}} \left[ M^2_{p+1} \mid F_p \right] - M^2_p$$

$$= \mathbb{1}_{r_{k,i,l} > p} \sum_{r=1}^{R_p} \sum_{i: \xi^i_p \in B^r_p} \left( \bar{\omega}^r_p \omega^i_p K_p g_{p+1}(\xi^i_p) - \left( \bar{\omega}^r_p \omega^i_p + (\omega^i_p)^2 \right) \sigma^r_{p,i} \right) g_p(\xi^i_p).$$

\[ \square \]
Proof of Theorem 6. We begin with the proof of (8). Recall $K(x, dy)$ is the transition kernel of $X_p = Y_{\tau_p}$. Conditioning on $Y_{\tau_1}$ and using assumption (B1), the LHS of (8) rewrites

$$\mathbb{E}[g(Y_{\tau_p})] = \int \mathbb{E}[g(Y_{\tau_p}) | Y_{\tau_1} = y] K(x, dy)$$

$$= \int_{E \setminus F} \mathbb{E}^y[g(Y_{\tau_p}) | Y_{\tau_1} = y] K(x, dy) + \int_F g(y) K(x, dy)$$

$$= \int_{E \setminus F} \mathbb{E}^y[g(Y_{\tau_p})] K(x, dy) + \int_F g(y) K(x, dy).$$

Integrating against $\pi$ and using stationarity,

$$\mathbb{E}[g(Y_{\tau_p})] = \int_{E \setminus F} \mathbb{E}^y[g(Y_{\tau_p})] \pi(dy) + \int_F g(y) \pi(dy).$$

Cancelling terms and using the fact that $Y_{\tau_p+1} \sim \rho P$,

$$\int_F g(y) \pi(dy) = \int_F \mathbb{E}^y[g(Y_{\tau_p})] \pi(dy) = \pi(F) \mathbb{E}^y[g(Y_{\tau_p})].$$

Now we turn to the proof of (9). Conditioning on $Y_{\tau_1}$ and using Assumption (B1),

$$\mathbb{E}^x \left[ \sum_{t=1}^{\tau_p} g(Y_t) \right] = \int \mathbb{E}^x \left[ \sum_{t=1}^{\tau_p} g(Y_t) \bigg| Y_{\tau_1} = y \right] K(x, dy)$$

$$= \int \mathbb{E}^x \left[ \sum_{t=1}^{\tau_1} g(Y_t) \bigg| Y_{\tau_1} = y \right] K(x, dy)$$

$$+ \int_{E \setminus F} \mathbb{E}^x \left[ \sum_{t=\tau_1+1}^{\tau_p} g(Y_t) \bigg| Y_{\tau_1} = y \right] K(x, dy)$$

$$= \mathbb{E}^x \left[ \sum_{t=1}^{\tau_1} g(Y_t) \right] + \int_{E \setminus F} \mathbb{E}^y \left[ \sum_{t=1}^{\tau_p} g(Y_t) \right] K(x, dy).$$

Integrating against $\pi$ and using stationarity,

$$\mathbb{E}^x \left[ \sum_{t=1}^{\tau_p} g(Y_t) \right] = \mathbb{E}^x \left[ \sum_{t=1}^{\tau_1} g(Y_t) \right] + \int_{E \setminus F} \mathbb{E}^y \left[ \sum_{t=1}^{\tau_p} g(Y_t) \right] \pi(dy).$$

Cancelling terms and using the fact that $Y_{\tau_p+1} \sim \rho P$,

$$\mathbb{E}^x \left[ \sum_{t=1}^{\tau_1} g(Y_t) \right] = \int_F \mathbb{E}^y \left[ \sum_{t=1}^{\tau_p} g(Y_t) \right] \pi(dy) = \pi(F) \mathbb{E}^y \left[ \sum_{t=1}^{\tau_p} g(Y_t) \right].$$

Acknowledgments. D. Aristoff would like to acknowledge enlightening conversations with Tony Lelièvre, Petr Plecháč, Mathias Rousset, Gideon Simpson, Ting Wang, and Dan Zuckerman. D. Aristoff also gratefully acknowledges support from the National Science Foundation via the award NSF-DMS-1522398.

REFERENCES
[1] R. J. Allen, D. Frenkel, and P. R. ten Wolde. Forward flux sampling-type schemes for simulating rare events: Efficiency analysis, J. Chem. Phys. 124(19), (2006), pp. 463102.
[2] Aristoff, D., Bello-Rivas, J. M., and Elber, R., A mathematical framework for exact milestoning, Multiscale Model. Simul. 14(1), (2016), 301–322.
[3] Bello-Rivas, J. M. and Elber, R., Exact milestoning, The Journal of Chemical Physics, 142 (2015), pp. 094102.
[4] Bhart, D., Zhang, B. W., and Zuckerman, D. M. Steady-state simulations using weighted ensemble path sampling, J. Chem. Phys. 133, (2010), 014110.
[5] Costaouec, R., Feng, H., Izaguirre, J., and Darve, E. Analysis of the accelerated weighted ensemble methodology, Supplement, Discrete and Continuous Dynamical Systems, 2013.
[6] Darve, E. and Ryu, E. Computing reaction rates in bio-molecular systems using discrete macro-states, Innovations in Biomolecular Modeling and Simulations, RSC publishing, 2012.
[7] A. K. Faradjian and R. Elber, Computing time scales from reaction coordinates by Milestoning, J. Chem. Phys. 120 (2004), pp. 10880–10889.
[8] D. R. Glowacki, E. Paci, and D. V. Shalashilin, Boxed Molecular Dynamics: Decorrelation Time Scales and the Kinetic Master Equation, J. Chem. Theory Comput. 7(5) (2011), pp. 1244-1252.
[9] Hill, T. L. Free Energy Transduction and Biochemical Cycle Kinetics, Dover, New York, 1989.
[10] Huber, G. A. and Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions, Biophys. J. 70(1), (1996), 97–110.
[11] Metzner, C., Schütte, C., and Vanden-Eijnden, Transition Path Theory for Markov jump processes, Multiscale Model. Simul., 7(3), (2009), 1192–1219.
[12] Sarich, F. Noé, and C. Schütte, On the approximation quality of Markov State Models, 2010.
[13] Schütte, C. and Sarich, M. Metastability and Markov State Models in Molecular Dynamics, Courant Lecture Notes, 2013.
[14] T. Schlick, Molecular modeling and simulation: an interdisciplinary guide, 2010.
[15] Suárez, E., Lettieri, S., Zwier, M. C., Stringer, C. A., Subramanian, S. R., Chong, L. T., and Zuckerman, D. M. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories, J. Chem. Theory Comput., 10, (2014), 2658–2667.
[16] Suárez, E., Pratt, A. J., Chong, L. T., and Zuckerman, D. M. Estimating first-passage time distributions from weighted ensemble simulations and non-Markovian analyses, Protein Science 25, (2016), 67–78.
[17] E. Vanden-Eijnden and M. Venturoli, Exact rate calculations by trajectory parallelization and tilting, J. Chem. Phys., 131, (2009), pp. 1–7, 0904.3763.
[18] van Erp, T. S., D. Moroni, and P. G. Bolhuis, A novel path sampling method for the calculation of rate constants, The Journal of Chemical Physics, 118(17) (2003), pp. 7762–7774.
[19] A. Warmflash, P. Bhimalapuram, and A. R. Dinner, Umbrella sampling for nonequilibrium processes, The Journal of chemical physics, 127 (2007), pp. 154112.
[20] Zhang, B. W., Jasnow, D., and Zuckerman, D. M. The “weighted ensemble” path sampling method is exact for a broad class of stochastic processes and binning procedures, J. Chem. Phys. 132, (2010), 05417.